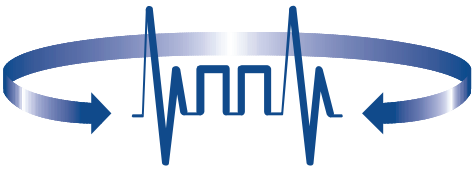
РОССИЙСКАЯ АКАДЕМИЯ НАУК
Институт проблем передачи информации им. А.А. Харкевича



КОНФЕРЕНЦИЯ МОЛОДЫХ УЧЕНЫХ И СПЕЦИАЛИСТОВ

Информационные технологии и системы

ИТиС’08

СБОРНИК ТРУДОВ КОНФЕРЕНЦИИ

ISBN 978-5-901158-08-01

г. Геленджик,
29 сентября – 3 октября 2008 г.

Информационные технологии и системы (ИТиС’08):
сборник трудов конференции. – М.: ИППИ РАН, 2008. – 517с.

Издание содержит труды ежегодной конференции молодых ученых и специалистов «Информационные технологии и системы» (ИТиС'08), традиционно организуемой Советом молодых ученых и специалистов ИППИ РАН. В 2008 году в организации конференции также принимали участие АНО «НИИ СИТРОНИКС» и ИПУ РАН.

Основная цель Конференции ИТиС’08 – дать возможность молодым ученым и специалистам различных подразделений ИППИ РАН, а также студентам, аспирантам и молодым ученым других институтов РАН, отраслевых институтов, университетов и вузов, познакомиться с коллегами и обменяться научными достижениями по основным для ИППИ РАН направлениям научной деятельности: теория передачи и защиты информации; математическая теория информации и управления, многокомпонентные случайные системы; информационно-коммуникационные технологии и их применение в сложных системах и сетях; информационные процессы в живых системах и биоинформатика; компьютерная лингвистика и моделирование естественного языка.

Все включенные в данный сборник работы прошли рецензирование и опубликованы в том виде, в котором они были представлены авторами, среди которых молодые ученые и специалисты ИППИ РАН, ИПУ РАН, СПИИРАН, ИПТМ РАН, ИМБ РАН, ИМГ РАН, ИОГен РАН, МГТУ им. Н.Э. Баумана, МГУ им. М.В. Ломоносова, МИФИ (ГУ), МФТИ (ГУ), ЛГТУ, МИСиС, РУДН и др.

Труды Конференции могут представлять интерес для ученых, студентов и аспирантов, специализирующихся в областях науки, связанных с перечисленными выше научными направлениями.

Конференция проведена при финансовой поддержке ИППИ РАН, АНО «НИИ СИТРОНИКС», Отделения нанотехнологий и информационных технологий РАН, Целевой программы Президиума РАН «Поддержка молодых ученых 2008 г.» и РФФИ (грант 08-07-06047-г)

ISBN 978-5-901158-08-01

© Коллектив авторов, 2008
© Учреждение Российской академии наук Институт проблем передачи информации им. А.А. Харкевича РАН, 2008

Организаторы конференции

Учреждение Российской академии наук
Институт проблем передачи информации им. А.А. Харкевича РАН

Автономная некоммерческая организация
«Научно-исследовательский институт СИТРОНИКС»

Учреждение Российской академии наук
Институт проблем управления им. В.А. Трапезникова РАН

Попечительский совет:

Кулешов А.П., член-корр. РАН, директор ИППИ РАН - председатель
Харченко М.А., директор АНО «НИИ СИТРОНИКС»

Организационный комитет:

Мацнев Д.Н., к.т.н. - председатель
Сафонов А.А. - зам. председателя
Алимова И.В.
Виноградов Д.В.
Грошев Ф.В.
Осипов Д.С.
Ставровская Е.Д.
Чуяшкин М.И.

Программный комитет:

Гельфанд М.С., к.ф.-м.н., д.б.н., проф. - председатель
Кабатянский Г.А., к.ф.-м.н. - зам. председателя
Мацнев Д.Н., к.т.н. - зам. председателя
Апресян Ю.Д., академик РАН
Бастаков В.А., к.б.н.
Богуславский И.М., д.фил.н., проф.
Блиновский В.М., д.ф.-м.н.
Вишневский В.М., д.т.н., проф.
Вьюгин В.В., д.ф.-м.н.
Гитис В.Г., д.т.н.
Зяблов В.В., д.т.н.
Иомдин Л.Л., к.фил.н.
Ляхов А.И., д.т.н.
Новиков Д.А., д.т.н.
Переверзев-Орлов В.С., к.т.н.
Прелов В.В., д.ф.-м.н.
Сагалович Ю.Л. д.т.н., проф.
Цитович И.И., д.ф.-м.н.
Цфасман М.А., д.ф.-м.н.

Содержание

РАЗДЕЛ 1: Технологии, стандарты и протоколы сетей передачи данных 1

Эффективность обобщенного механизма переключения битовой скорости в протоколе IEEE 802.11..... 2
Зарубинский И., Якимов М.

Исследование свойств мультиплексирования потоков в мультисервисных сетях..... 9
Цитович И. , Титов И.

Об интервальных моделях телетрафика мультисервисных сетей..... 14
Цитович И., Догадченко Н.

Пропускная способность широкополосного беспроводного канала точка-точка в режиме насыщения..... 18
Шпилев С.А.

Распределения времени присоединения устройств к сети ECMA-368 (WiMedia)..... 23
Сафонов А., Хоров Е.

Эффективность метрик стоимости соединений стандарта ZigBee 30
Баскаков С.С.

Система поллинга с опросом типа "звезда" и ограниченным обслуживанием..... 35
Семенова О.В., Сергеев П.А.

Анализ характеристик многокилометрового широкополосного беспроводного канала точка-точка с различными функциями управления 39
Фахриев Д.Н., Шпилев С.А.

Организация повторных передач в системах связи с ретрансляционными станциями..... 45
Уханова А.С.

Модель изменения очереди пакетов в беспроводной сети под управлением протокола IEEE 802.16..... 51
Лукин Д.В.

Исследование методов расчета антенн с несколькими облучателями для беспроводных сетей 802.11 56
Камышников И.Г., Воробьев В.М.

Новый алгоритм выбора скорости передачи в сетях стандарта IEEE 802.11..... 62
Адов А.А., Михин Д.Ю., Черкасов М.В.

Моделирование и анализ протокола маршрутизации HWMP стандарта IEEE 802.11s 67
Андреев К., Лаконцев Д., Мазо А.

Эффективность применения MIMO в беспроводной городской сети передачи данных 73
Жилин М.С., Воробьев В.М.

Оптимизация механизма определения занятости среды для mesh-сетей стандарта IEEE 802.11s 79
Коваленко А.В., Лаконцев Д.В., Пешков А.И.

РАЗДЕЛ 2. Информационные технологии и проблемы управления в сложных системах и сетях 85

Система доказательства теорем с автоматической генерацией знаний 86
Раутиаинен А., Стефанюк В.Л.

Размещение фотографий на холсте с определением коэффициентов масштабирования 90
Егорова М.А., Сафонов И.В.

ГеоТайм II - новая геоинформационная технология исследования пространственно-временных процессов 96
Дерендяев А.Б., Метриков П.А., Гитис В.Г.

Анализ эффективности применения методов интерполяции для доопределения законов функционирования сложных систем 104
Епифанов А.С.

Решение задачи нахождения соответствующего эталона методом нейронного кортекса 109
Сергийчук М.Н., Михайлов А.М.

Интерактивные учебные пособия на основе средств виртуальной реальности в CALS – технологиях 115
Ромакин В.А., Балабанов А.В.

Задачи оптимизации и информационного обеспечения при управлении ДПТ 118
Синюков А. В.

Построение стохастической модели прогнозирования состояния вычислительной сети как решение задачи классификации и снижения размерности 121
Еремеев В.Б.

Проектирование структур данных для комплекса сбора оперативной информации о случаях заболеваний 126
Сурпин В.П., Волчков Д.В.

Функциональная модель агента-предприятия в мультиагентной системе управления экологической безопасностью ... 129
Еременко Ю.И., Доронина Е.Г.

Вычисление геометрических параметров поверхности при наличии детерминированных возмущений ... 132
Кочетков С.А.

Оптимизация процесса центрифугирования путем определения текущей массы в барабане ... 138
Ханина Н.Ю.

Разработка и исследование макета отказоустойчивого струйного акселерометра для летательных аппаратов ... 143
Коврыгин П.В.

Методика конкурентного анализа телекоммуникационных компаний ... 147
Сафарли И.И.

Допредельная реализация метода иерархии управлений в системе управления двухзвенным манипулятором ... 152
Уткин А.В.

Введение платы за участие в механизмы активной экспертизы ... 156
Бурков В.Н., Коргин Н.А.

РАЗДЕЛ 3: Передача, защита и обработка информации ... 160

Модель беспроводного канала связи в условиях крупномасштабного замирания и случайного разброса параметров приемопередатчиков ... 161
Баскаков С.С.

Исправление стираний низкоплотностными кодами Галлагера ... 167
Зяблов В.В., Рыбин П.С

Адаптивная подстройка параметров кодирования речевого сигнала в системах IP-телефонии ... 173
Иманкулов Р.А.

Оценивание характеристик канала в системе множественного доступа, основанной на перестроении ортогональных частот ... 177
Осипов Д.С.

Сравнительный анализ помехоустойчивости плетёных свёрточных (woven) и турбо кодов ... 184
Чеклин А.Б., Черкасов М.В., Адов А.А.

Программный комплекс для исследования систем передачи данных ... 187
Грошев Ф.В., Зигангиров Д.К.

Применение вэйвлет-преобразования для анализа спектров EXAFS: исследование причины появления и анализ артефактов на спектрограммах 190
Павлов В.Н., Чукалина М.В.

РАЗДЕЛ 4: Компьютерная лингвистика и моделирование естественного языка 194

Развитие системы построения семантического представления текста с использованием языка-посредника UNL на базе лингвистического процессора ЭТАП-3 195
Диконов В.

Использование конечных автоматов для морфологического анализа и синтеза на основе словарей системы ЭТАП 201
Казенников А.О.

Извлечение информации о сочетаемости лексем из аннотированного корпуса текстов 206
Дяченко П., Фролова Т.

Автоматическое уточнение правил в самозагрузочной системе извлечения информации из текстов на естественном языке 211
Масленников М.

Сравнение качества работы синтаксического анализатора системы ЭТАП и статистического синтаксического анализатора MaltParser на материале текстов из корпуса СинТагРус 219
Сизов В.Г., Валеев Д.Р., Крейдлин Л.Г.

Лексические функции и возможности оптимизации поиска информации в интернете (на материале параметрических слов) 225
Тимошенко С.П.

РАЗДЕЛ 5: Информационные процессы в живых системах 230

Спектрально-временные характеристики акустических сигналов птиц 231
Гладышев А.М.

Различные способы неспецифической активации шагательной активности в условиях разгрузки нижних конечностей у человека 236
Солопова И.А., Селионов В.А.

Наследственные нарушения слуха: дифференцирование 60 случаев ранней глухоты и тугоухости 242
Баталова Д., Божкова В.

Роль структурной организации ЭЭГ-информации при выявлении патогенетического механизма алалии 246
Перемячкина С.Г., Попова С.М., Лукашевич И.П.

Использование записей движений глаз для анализа механизмов зрительной оценки размеров на примере иллюзии Мюллера-Лайера 250
Огнивов В.В., Максимов П.В., Рожкова Г.И.

Движения глаз при восприятии неоднозначных изображений типа куба Неккера..... 256
Желтова М.Ю., Максимов П.В., Бастаков В.А., Панютин А.К.

Компьютерная система для исследования лабильности зрительной системы 259
Грошев Ф.В., Голубцов К.В., Трунов В.Г., Айду Э.А.

РАЗДЕЛ 6: Биоинформатика 262

Positive selection and alternative splicing of human genes 263
Ramensky V., Nurtdinov R., Neverov A., Mironov A., Gelfand M.

SDPfox - Web-сервер для анализа субстратной специфичности белков 266
Мазин П., Калинина О., Рахманинова А., Миронов А., Гельфанд М.

Подтверждение участия давления отбора в поддержании стабильности CpG островов в геноме 268
Перцовская И., Опарина Н., Виноградов Д., Миронов А., Фаворов А.

Comparative Analysis of Transcription Regulation Strategies of Bacteriophages T4 and RB49 271
Pugach K., Severinov K., Gelfand M., Ravcheev D.

Комплексная регуляция дыхания бактерий: исследование методами сравнительной геномики..... 273
Равчеев Д.А.

A Markov Chain Monte Carlo Technique for Identification of Combinations of Allelic Variants Underlying Complex Diseases 277
Favorov A., Sudomoina M., Favorova O., Parmigiani G., Michael F. Ochs

Analyzing Drug-target Interaction from a Comparative Perspective..... 279
Kalinina O., Russell R.

Изучение консервативности межгенных участков ДНК прокариот 280
Пятницкий М.А.

Эволюция под отбором: исследование паттерна замен в сайтах сплайсинга 284
Денисов С., Миронов А., Гельфанд М.

Фиксация точечных мутаций кодирующих участках генов плодовых мушек рода Drosophila 289
Ермакова Е.О., Малько Д.Б., Гельфанд М.С.

Adaptation to the loss of leader protein by cardiovirus: analysis of compensatory mutations 293
Fominykh K., Lidsky P.

Изучение регуляции транскрипции в геномах рода Shewanella путем автоматического поиска регуляторных мотивов в рамках функциональных подсистем 295
Ставровская Е.Д., Сиприано М., Родионов Д.А., Дубчак И.Л., Миронов А.А., Гельфанд М.Г.

Non-5' CpG islands in the human genome: probable involvement in transcription regulation 298
Medvedeva Y.A., Fridman M.V., Oparina N.J., Malko D.B., Ermakova E.O., Kulakovskiy I.V., Makeev V.J.

Изучение системы деградации этаноламина методами сравнительной геномики 300
Цой О.В., Равчеев Д.А., Гельфанд М.С., Мушегян А.

Alternative splicing tends to involve phosphorylation sites 304
Kurmangaliyev Y.Z., Gelfand M.S.

Polymorphism Due to Multiple Amino Acid Substitutions at a Codon Site within Ciona savignyi 306
Donmez N., Bazykin G.A., Brudno M., Kondrashov A.S.

Эволюция недавно дуплицировавшихся паралогов генома человека 308
Панчин А., Артамонова И.И.

CRISPR-системы в метагеномах 314
Сорокин В.А., Артамонова И.И.

ДНК коды, основанные на весовом стебельном сходстве Хэмминга 316
Дьячков А.Г., Воронина А.Н.

Поиск вторичных структур в интронах представителей рода Drosophila 321
Храмеева Е.Е., Миронов А.А., Гельфанд М.С., Первушин Д.Д.

Эволюционный анализ генов семейства SSX 326
Шустрова Е., Артамонова И.

Регуляция дыхания бактерий: транскриптомика и сравнительная геномика 332
Цыганова М., Равчеев Д.

Использование сравнительной геномики для улучшения достоверности аннотации сигнальных пептидов 337
Быкова Н.А., Сутормин Р.А.

Эволюционные изменения структуры новой некодирующей РНК человека 341
Бениаминов А., Вестхоф Э., Крол А.

Исследование корреляций между аминокислотными последовательностями транскрипционных регуляторов и нуклеотидными последовательностями их сайтов связывания с ДНК 342
Коростелев Ю.Д., Лайкова О.Н., Рахманинова А.Б.

Deleterious effect of mitochondrial nucleotide repeats in Chordata 344
Popadin K., Mamirova L., Kuptsov A.

Реконструкция эволюции систем регуляции генов метаболизма гистидина в бактериях 346
Витрещак А.

Предсказание сайтов регуляции рибосомальных генов в бактериях 349
Петрова С., Витрещак А., Гельфанд М.

Построение HMM, основанной на правилах взаиморасположения сайтов связывания транскрипционных факторов, и ее применение для поиска корегулируемых генов в геномах рода Drosophila 352
Никулова А.А., Сутормин Р.А., Миронов А.А., Фаворов А.В.

Функциональная классификация генов полных бактериальных геномов на основе базы данных Clusters of Orthologous Groups (COG) 354
Казанов М.

Корреляции между аминокислотными последовательностями транскрипционных регуляторов суперсемейства CRP-FNR и нуклеотидными последовательностями их операторов 359
Митева Д.С., Степанова В.В.

Эволюция альтернативного сплайсинга генов млекопитающих 361
Нуртдинов Р.Н.

Genome-wide assessment of codon usage correlation 366
Vinogradov D.V., Mironov A.A.

Кислород-зависимая регуляция азотфиксации: исследование методами сравнительной геномики 367
Климова Е.Ю., Равчеев Д.А.

РАЗДЕЛ 7: Математическая теория информации и управления, многокомпонентные случайные системы 370

Моделирование вычислительных систем непрерывными уравнениями 371
Захаров В., Тормасов А.

Моделирование стационарных входных потоков данных для стохастических моделей дискретных систем 376
Шляева А.

Алгоритм прогнозирования функции распределения нестационарного временного ряда 382
Осминин К.П.

Многоэтапные процедуры проверки статистических гипотез и субоптимальность ... 386
Цитович Ф.

Игровые стратегии и адаптивное прогнозирование 391
Вьюгин В.В., Румянцев Е.Л., Цфасман Ф.М.

Уравнение Пенлеве VI и дополнительные особенности фуксовых уравнений 397
Гонцов Р.Р.

О минимаксной асимптотической оптимальности первого порядка в задаче наискорейшего обнаружения изменения интенсивности пуассоновского процесса .. 403
Бурнаев Е.В.

Оптимальное нечеткое агрегирование информации в системах принятия решений .. 406
Алфимцев А.Н., Девятков В.В.

Об одном достаточном условии PSPACE-разрешимости транзитивных модальных логик 411
Шапировский И.

Сложность алгоритмов проверки и поддержания степеней непротиворечивости алгебраических байесовских сетей..... 417
Сироткин А.В.

Барьерно - мультипликативные алгоритмы для двойственных задач 423
Гасникова Е.

Сдвиг инвариантов в алгебрах Пуассона алгебр Ли токов на прямой..... 430
Рыбников Л.Г.

Задача одношагового управления..... 434
Мещерякова О.В.

Применение идемпотентного принципа соответствия для решения задач теории систем массового обслуживания 436
Жбанов С.А.

Байесовские и марковские сети: логико-вероятностный вывод в базах фрагментов знаний с неопределенностью 440
Тулупьев А.Л., Сироткин А.В.

Деформация границы простых множеств под действием растягивающей последовательности отображений..... 457
Комеч С.

Протокол демонстрации K-кратного вхождения строки 459
Федюкович В.Е., Шарапов В.Г.

РАЗДЕЛ 8: Распознавание образов и обработка изображений 467

Локализация семантических блоков на растровых изображениях без предварительного обучения..... 468
Кулешов С.В., Зайцева А.А.

Быстрый алгоритм совмещения изображений документов в произвольной геометрической модели 471
Усилин С., Николаев Д., Постников В.

Использование методов анализа неподвижных изображений для поиска событий в видеопоследовательностях 478
Кондратенков И.Б., Рубанов Л.И.

Обработка данных в системах создания виртуального окружения 482
Торчигин А.В.

Расчет количества текстурных кластеров для сегментации изображений на основе Марковских случайных полей 488
Патана Е.И.

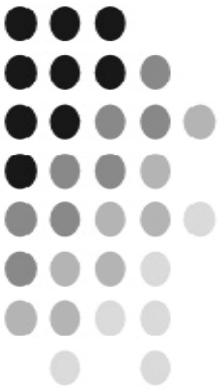
Использование цифрового нейронного кортекса на примере распознавания образов..... 493
Лазаренко А.В., Михайлов А.М.

Метод идентификации типа документа по структуре проекций его изображения на координатные оси..... 498
Безматерных П., Николаев Д., Постников В.

Метод восстановления изображений, подверженных радиальной дисторсии 502
Николаев Д., Гладилин С., Карпенко С.

Раздел 1

Технологии, стандарты и протоколы сетей передачи данных



Эффективность обобщенного механизма переключения битовой скорости в протоколе IEEE 802.11

Зарубинский Игорь, Якимов Михаил
Институт проблем передачи информации им. А.А. Харкевича РАН
E-mails: { zarubinskiy@iitp.ru, yakimov@iitp.ru }

Аннотация

Одним из основных показателей эффективности работы беспроводной сети является ее пропускная способность, которая напрямую зависит от битовой скорости, используемой станциями сети для передачи кадров с данными. В данной работе рассмотрен механизм с обобщенным правилом переключения битовой скорости передачи. Согласно этому правилу, каждая станция сети переключает битовую скорость если числа последовательных успешных и неудачных попыток передачи достигают некоторых пороговых значений, в общем случае различных для разных битовых скоростей. В статье разработана аналитическая модель, позволяющая оценить пропускную способность беспроводной сети типа Hot spot, работающей по протоколу IEEE 802.11 и состоящей из N станций, каждая из которых реализует рассматриваемый механизм переключения скорости. С помощью предложенной модели проведена оптимизация порогов переключения битовой скорости, исследована производительность беспроводной сети при наличии помех и коллизий, а также определено допустимое число активных станций в сети, при котором механизм остается эффективен.

1. Введение

Используя тот или иной механизм, устройства могут автоматически переключать битовую скорость с целью достижения наилучшей пропускной способности для данных условий передачи, т.е. для относительной интенсивности помех (*SNR*) в канале. Таким образом, пропускная способность беспроводной сети зависит от эффективности конкретного механизма переключения скорости, его восприимчивости к коллизиям, а также адекватной оценке значения *SNR* в канале.

Текущая версия стандарта IEEE 802.11 никак не регламентирует механизм выбора битовой скорости, оставляя решения этого вопроса производителям беспроводного оборудования.

2. Базовый механизм переключения битовой скорости

Базовым механизмом переключения скорости передачи в беспроводных сетях был и остается ARF (Auto Rate Fallback) – алгоритм с автоматическим снижением скорости [1]. Изначально он был разработан для первых карт беспроводного доступа WaveLAN-II по протоколу IEEE 802.11 и поддерживал только две битовых скорости передачи – 1 и 2 Мбит/с.

Алгоритм его работы следующий. Для каждого нового соединения в качестве начальной битовой скорости выбирается максимально возможная скорость. Механизм ARF следит за текущим значением битовой скорости и количеством последовательных, т.е. идущих подряд успешных попыток передачи кадра данных. Перед началом очередной попытки передачи кадра данных механизм ARF выбирает битовую скорость согласно следующим правилам:

А) подтверждение о доставке кадра не было получено два раза подряд – снизить битовую скорость;

Б) последовательно прошло 10 удачных передач кадров данных – увеличить битовую скорость;

В) иначе – продолжить работу на текущей битовой скорости.

Основное преимущество алгоритма ARF – его простота. Он хорошо работает в каналах с резкими периодическими затуханиями сигнала, быстро настраивая скорость на максимально возможную, при которой затухание не проявляется. Однако, в канале с постоянным или медленно меняющимся затуханием, а также при высокой вероятности коллизии, алгоритм ARF не всегда может поддерживать пропускную способность на высоком уровне

высоком уровне в силу правила А). В таких случаях ARF настраивается на битовую скорость меньшую оптимальной, не переключаясь на более высокую скорость, для которой пропускная способность может быть выше.

Впоследствии был предложен механизм автоматического переключения скорости AARF (Adaptive ARF), который является улучшением предыдущего алгоритма ARF. Было замечено, что переключение на более высокую битовую скорость, каждый раз после десяти последовательно переданных кадров данных, может заметно снизить производительность. Механизм AARF предусматривает удвоение количества последовательных удачных передач, необходимых для переключения на следующую битовую скорость. Однако, если станция понижает скорость 2 раза подряд, то необходимое количество последовательных успешных передач сбрасывается в начальное значение – 10 кадров.

3. Аналитическая модель

Рассмотрим более универсальный механизм переключения скорости передачи, обобщающий правила выбора битовой скорости ARF и AARF.

Пусть станция имеет I битовых скоростей передачи. Далее будем обозначать их с помощью индекса $i \in \{1,\ldots,I\}$. Пусть также станция имеет два счетчика: счетчик последовательных успешных попыток и счетчик последовательных неудачных попыток передачи, т.е. тех, подтверждение о доставке кадра в которых не было получено. Назовем их счетчиками успехов и неудач соответственно. Если на скорости i счетчик успехов достигает некоторого порогового значения s_i, принимается решение о повышении битовой скорости (индекс скорости увеличивается на единицу) в том случае, если это возможно. Если же счетчик неудач достигает соответствующего порогового значения e_i, то скорость передачи понижается (индекс скорости уменьшается на единицу), если, конечно, это возможно. При изменении индекса скорости передачи в большую или меньшую сторону счетчики успехов и неудач сбрасываются, что соответствует точке обновления всего процесса. Таким образом, при передаче кадра состояние станции определяется значениями счетчиков и номером текущей битовой скорости. Обозначим состояния, попадание в которые произошло с увеличением скорости передачи, либо в которых счетчик успехов отличен от нуля, через $(i, j, +)$, $j \in \{0,\ldots,s-1\}$, а их стационарные вероятности – $R^{+}_{i,j}$. Аналогично, состояния, попадание в которые произошло с уменьшением скорости передачи или счетчик неудач, в которых отличен от нуля, будем обозначать через $(i, j, -)$, $j \in \{0,\ldots,e-1\}$, а их стационарные вероятности – $R^{-}_{i,j}$. Здесь знак показывает, была ли последняя попытка передачи успешной («+») или неудачной («–»), а индекс j в зависимости от знака соответствует значению счетчика последовательных успехов или неудач. Таким образом, процесс передачи кадров можно представить марковским процессом с дискретным

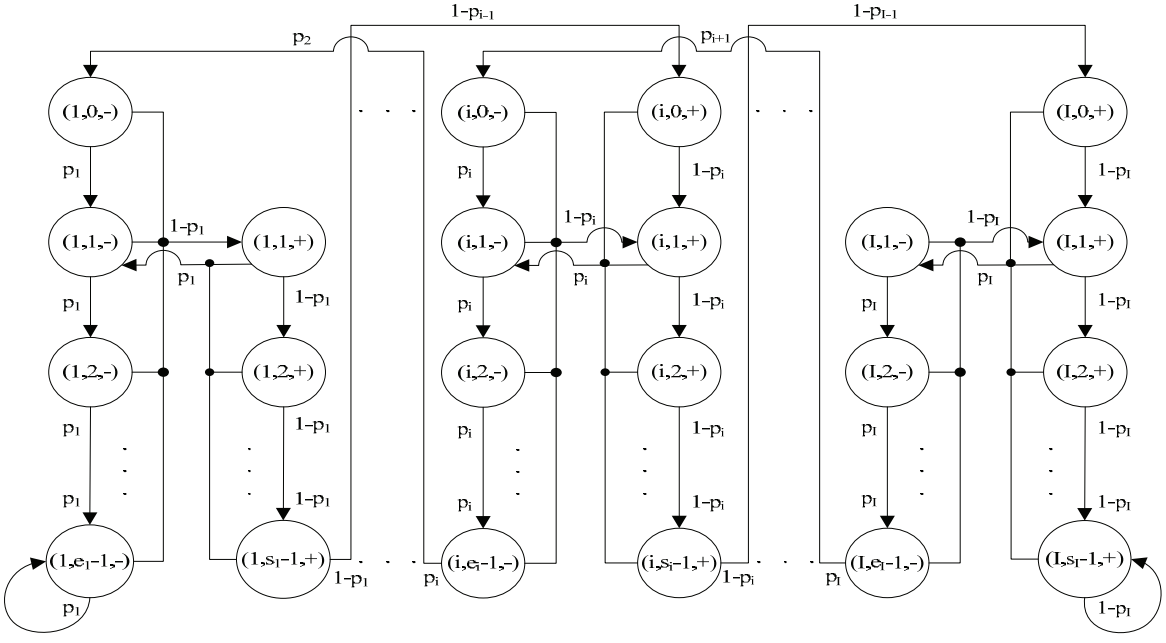


Рис. 1 Марковская цепь процесса передачи кадров

временем, единицей которого является интервал времени между попытками передачи кадра (Рис. 1). Из любого состояния $(i, j, \pm)$ возможно ровно два альтернативных перехода с вероятностями p_i и $1-p_i$, соответствующих неудаче или успеху очередной попытки передачи. На Рис. 1 состояния $(1,0,+)$ и $(I,0,-)$ не показаны, поскольку их стационарные вероятности равны нулю.

Для нахождения стационарных вероятностей $R_{i,j}^{\pm}$ всех состояний $(i, j, \pm)$, преобразуем данную цепь к упрощенному виду (Рис.2), представляющему собой процесс рождения-гибели с I состояниями, в котором i -ое состояние со стационарной вероятностью R_i соответствует передаче станции на i -ой битовой скорости, λ_i – вероятность «рождения», т.е. увеличения битовой скорости, а μ_i – вероятность «гибели», т.е. уменьшения битовой скорости. Уравнения баланса для стационарных вероятностей R_i состояний этой цепи имеют вид:

$$\begin{cases} R_1\lambda_1 = R_2\mu_2, \\ R_i\left(\lambda_i+\mu_i\right) = R_{i-1}\lambda_{i-1} + R_{i+1}\mu_{i+1}, \quad i \in \{2,\ldots,I-1\}, \\ R_I\mu_I = R_{I-1}\lambda_{I-1}; \end{cases}$$

откуда легко получаем следующее соотношение:

$$R_i\mu_i = R_{i-1}\lambda_{i-1} \quad , \quad i \in \{2,\ldots,I\} \quad .$$

(1)

С учетом нормировки $\sum_{i=1}^{I} R_i = 1$ и (1) легко получить выражения:

$$R_1 = \frac{1}{1+\sum_{k=2}^{I}\prod_{i=2}^{k}\frac{\lambda_{i-1}}{\mu_i}},$$

$$R_i = R_1\prod_{k=2}^{i}\frac{\lambda_{k-1}}{\mu_k} \quad , \quad \text{где} \quad i \in \{2,\ldots,I\} \quad .$$

(2)

Далее выразим вероятности переходов λ_i и μ_i, а также R_i через стационарные вероятности состояний исходной марковской цепи на Рис. 1:

$$R_i = \sum_{k=1}^{e_i-1} R_{i,k}^{-} + \sum_{k=1}^{s_i-1} R_{i,k}^{+} + R_{i,0}^{-} + R_{i,0}^{+} \quad ,$$

(3)

$$\lambda_i = \frac{\left(1-p_i\right)R_{i,s_i-1}^{+}}{R_i}, \quad \mu_i = \frac{p_i R_{i,e_i-1}^{-}}{R_i}. \qquad (4),$$

(5)

Уравнения баланса для исходной цепи имеют вид:

$$R_{i,k}^{+} = \left(1-p_i\right)R_{i,k-1}^{+} = \left(1-p_i\right)^{k-1}R_{i,1}^{+},$$

$$k \in \{2,\ldots,s_i-1\}, \qquad (6)$$

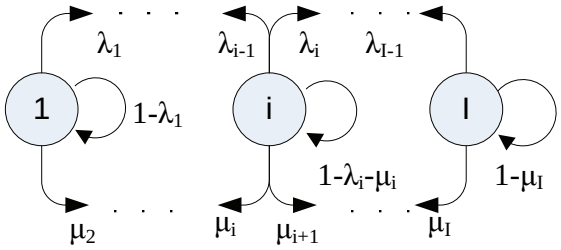


Рис. 2 Упрощенная Марковская цепь процесса переключения скорости

$$R_{i,k}^{-} = p_i R_{i,k-1}^{-} = p_i^{k-1}R_{i,1}^{-} \quad , \quad k \in \{2,\ldots,e_i-1\} \quad ,$$

(7)

$$R_{i,1}^{+} = \left(1-p_i\right)\left(\sum_{k=1}^{e_i-1} R_{i,k}^{-} + R_{i,0}^{-} + R_{i,0}^{+}\right) \quad ,$$

(8)

$$R_{i,1}^{-} = p_i\left(\sum_{k=1}^{s_i-1} R_{i,k}^{+} + R_{i,0}^{-} + R_{i,0}^{+}\right) \quad .$$

(9)

Используя (3), (6) и (8), а также (3), (7) и (9) получаем:

$$R_i = \frac{R_{i,1}^{+}}{1-p_i} + R_{i,1}^{+}\sum_{k=1}^{s_i-1}\left(1-p_i\right)^{k-1} = R_{i,1}^{+}\left(\frac{1-\left(1-p_i\right)^{s_i}}{p_i\left(1-p_i\right)}\right) \quad ,$$

$$R_i = \frac{R_{i,1}^{-}}{p_i} + R_{i,1}^{-}\sum_{k=1}^{e_i-1} p_i^{k-1} = R_{i,1}^{-}\left(\frac{1-p_i^{e_i}}{p_i\left(1-p_i\right)}\right) \quad ,$$

(10) (11)

откуда с учетом (4) и (6), (5) и (7) находим выражение для λ_i и μ_i соответственно:

$$\lambda_i = \frac{p_i\left(1-p_i\right)^{s_i}}{1-\left(1-p_i\right)^{s_i}}, \quad \mu_i = \frac{\left(1-p_i\right)p_i^{e_i}}{1-p_i^{e_i}}.$$

Используя выражения (6) и (10), (7) и (11) легко найти стационарные вероятности $R_{i,j}^{\pm}$, $j \neq 0$, состояний исходной цепи Маркова:

$$R_{i,j}^{+} = \frac{p_i\left(1-p_i\right)^{j}}{1-\left(1-p_i\right)^{s_i}}R_i, \qquad R_{i,j}^{-} = \frac{\left(1-p_i\right)p_i^{j}}{1-p_i^{e_i}}R_i,$$

$$R_{i,0}^{+} = \mu_i R_i, \text{ при } j=0, \quad i \neq 1,$$

$$R_{i,0}^{-} = \lambda_i R_i, \text{ при } j=0, \quad i \neq \mathrm{I}.$$

4. Анализ производительности беспроводной сети

Используя предложенную математическую модель универсального механизма переключения скорости, оценим производительность типовой беспроводной сети с протоколом IEEE 802.11.

Рассмотрим сеть типа Hot-spot, состоящую из $N-1$ оконечных станций и одной базовой

станции, представляющей из себя точку доступа (AP) во внешнюю сеть (Рис.3).

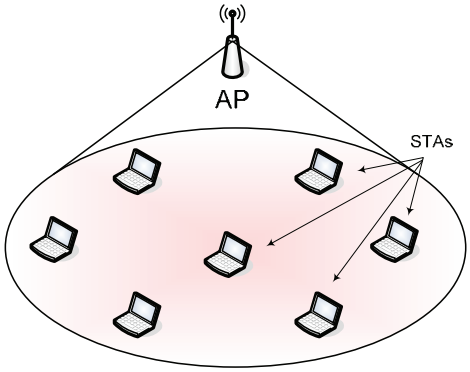


Рис. 3 Беспроводная сеть типа Hot-spot

Пусть все N станций данной сети, находясь в области покрытия друг друга, испытывают воздействие одинаковой шумовой помехи, характеризуемой отношением сигнал/шум SNR, и реализуют метод множественного доступа с прослушиванием несущей и избеганием коллизий CSMA/CA в режиме распределенного управления DCF [2]. Прежде чем начать передачу каждая станция сети, убедившись, что среда свободна, выжидает в течение интервала отсрочки, который формируется из случайного числа слотов постоянной длительности σ. Отсчет слотов отсрочки станция ведет только при свободном канале. Если счетчики двух и более станций одновременно достигли нулевого значения, происходит коллизия, длительность которой определяется кадром максимальной длины из числа вовлеченных в коллизию. Перед переходом в состояние отсрочки после передачи очередного кадра данных станция инициализирует счетчик слотов отсрочки некоторым начальным значением, равновероятно выбираемым из интервала целых чисел $[0,\ldots,CW]$, где величина CW называется конкурентным окном. Если предыдущая передача прошла успешно, то конкурентное окно равно своему минимальному значению $CW_{\min}$. В противном случае, оно определяется числом k неудачных попыток передачи кадра, которое ограничено пределом K, и равно $CW_k = (CW_{\min}+1)2^k - 1$, однако, не может быть больше некоторого значения $CW_{\max} = (CW_{\min}+1)2^m - 1$.

Если кадр данных успешно принят станцией назначения, она по истечении интервала времени *SIFS* отправляет подтверждение – кадр ACK (Acknowledgement). После завершения попытки передачи очередного кадра станция переходит в состояние отсрочки спустя интервал DIFS, если попытка была успешной, или EIFS – при неудачной попытке.

Введем допущение, что кадры искажаются одновременно на всех станциях в сети. Будем также считать, что все станции сети работают в режиме насыщения, т.е. в очередях станций всегда есть пакеты на передачу, а длина их полезной нагрузки L одинакова. Пакеты либо пересылаются от оконечных станций через AP во внешнюю сеть, либо, наоборот, из внешней сети через AP к оконечным станциям.

Для анализа сети будем использовать дискретную целочисленную временную шкалу, моменты времени на которой соответствуют началам следующих друг за другом *виртуальных слотов* [3]. Введем шкалу таким образом, что в начале виртуального слота каждая станция сети меняет значение своего счетчика отсрочки передачи согласно правилам протокола IEEE 802.11 [2] и может приступить к очередной передаче кадра, если значение этого счетчика стало равным нулю. Таким образом, можно выделить три типа виртуальных слотов (см. Рис. 4):

А) Пустой слот со средней длительностью T_e, в котором ни одна из станций сети не передает;

Б) Успешный слот со средней длительностью T_s, в котором передача только одна станция сети;

В) Коллизионный слот со средней длительностью T_c, в котором одновременно начинают передачу две или более станции сети.

Введем допущение, что для любой станции вероятность τ начала передачи в произвольно выбранный виртуальный слот не зависит ни от

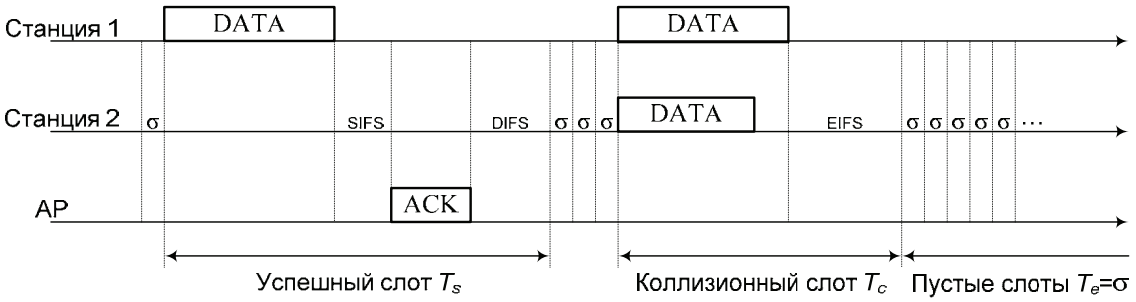


Рис. 4 Типы виртуальных слотов

предыстории, ни от текущих состояний других станций и одинакова для всех станций сети. Тогда вероятности того, что произвольно выбранный виртуальный слот окажется пустым, успешным или коллизионным, будут соответственно равны:

$$p_e = C_0^N \tau^0 (1-\tau)^N = (1-\tau)^N,$$

$$p_s = C_1^N \tau^1 (1-\tau)^{N-1} = N\tau(1-\tau)^{N-1},$$

$$p_c = 1 - p_e - p_s.$$

Принимая во внимание, что длительность пустого слота $T_e = \sigma$ постоянна и жестко определена протоколом, рассчитаем средние длительности успешного и коллизионного слота. При этом усреднение будем проводить с учетом того, что очередная передача кадра происходит на i-ой битовой скорости с вероятностью R_i.

Поскольку в коллизию могут вступать от двух до N станций, то средняя длительность коллизионного слота определяется по формуле:

$$T_c = \frac{1}{p_c}\sum_{k=2}^{N} T_c^{(k)} C_N^k \tau^k (1-\tau)^{N-k},$$

где $T_c^{(k)}$ – средняя длительность коллизионного слота при условии, что в коллизию вступает ровно k станций:

$$T_c^{(k)} = \sum_{i=1}^{I-1} T_i^{DATA} \sum_{m=1}^{k} C_k^m R_i^m \left(\sum_{j=i+1}^{I} R_j \right)^{k-m} + T_I^{DATA} R_I^k + EIFS.$$

Для длительности успешного слота имеем выражение:

$$\begin{aligned} T_s = \sum_{i=1}^{I} R_i \Big[& \left(T_i^{DATA} + SIFS + T_i^{ACK} + DIFS\right) \times \\ & \times \left(1 - P_i^{DATA}\right)\left(1 - P_i^{ACK}\right) + \\ & + \left(T_i^{DATA} + SIFS + T_i^{ACK} + EIFS\right) \times \\ & \times \left(1 - P_i^{DATA}\right) P_i^{ACK} + \\ & + \left(T_i^{DATA} + EIFS\right) P_i^{DATA} \Big], \end{aligned}$$

где T_i^{DATA} и T_i^{ACK} – времена передачи кадра данных и кадра подтверждения соответственно, P_i^{DATA} и P_i^{ACK} – вероятности искажения этих кадров вследствие воздействия шумовых помех.

Среднее количество информации, переданное за один успешный слот, определяется длиной L кадра данных и равно:

$$U = L\sum_{i=1}^{I} R_i \left(1 - P_i^{DATA}\right)\left(1 - P_i^{ACK}\right),$$

поэтому для общей пропускной способности сети можно записать следующее выражение:

$$S_{ARF} = \frac{p_s U}{p_e \sigma + p_s T_s + p_c T_c}. \quad (12)$$

Для вычисления вероятности τ воспользуемся простым соображением, что она равна отношению среднего числа попыток передачи одного кадра $\bar{f}$ к общему числу виртуальных слотов, прошедших с момента начала передачи кадра до успешного её окончания или исчерпания всех попыток передачи. Таким образом, имеет место формула:

$$\tau = \frac{\bar{f}}{\bar{f} + \bar{\omega}}, \quad (13)$$

где $\bar{\omega}$ – среднее число слотов, в которых станция вынуждена отложить передачу текущего кадра. В свою очередь $\bar{f}$ и $\bar{\omega}$ определяются на основе вероятностного распределения числа попыток передачи, получаемого с использованием стационарных вероятностей $R_{i,j}^{\pm}$ состояний цепи Маркова, описывающей процесс передачи кадров (см. Рис. 1).

Вероятность p_i неудачной попытки передачи кадра на скорости i равна:

$$p_i = 1 - \left(1 - p^{(c)}\right)\left(1 - p_i^{(e)}\right),$$

где $p_i^{(e)} = 1 - \left(1 - P_i^{DATA}\right)\left(1 - P_i^{ACK}\right)$ – вероятность неудачи по причине искажения помехами, а $p^{(c)} = 1 - (1-\tau)^{N-1}$ – вероятность неудачи по причине коллизии. Таким образом, окончательно имеем:

$$p_i = 1 - \left(1 - P_i^{DATA}\right)\left(1 - P_i^{ACK}\right)(1-\tau)^{N-1}. \quad (14)$$

Выражения (13) и (14) совместно составляют систему из $I+1$ независимых уравнений относительно τ и p_i, $i = \{1, \ldots, I\}$, решение которой даёт значения этих вероятностей для нахождения средней пропускной способности сети по формуле (12). К сожалению, в явном виде решение такой системы не может быть получено. На практике предлагается использовать итеративные методы вычислительной математики.

5. Численные результаты

Анализ работы предложенного механизмом переключения скорости проведем на примере беспроводной сети с протоколом IEEE 802.11a. Общее количество беспроводных станций в сети N будем брать из набора 2, 3, 4, 5, 10, а отношение сигнал/шум в канале – одним из семи значений SNR =2, 5, 9, 13, 17, 21, 25. Пороги e_i и

S_i , будем считать одинаковыми для каждой битовой скорости передачи и обозначать e и s соответственно.

На Рис. 5 представлены графики зависимостей пропускных способностей при работе на битовых скоростях, максимизирующих пропускную способность для различных отношений сигнал/шум, от числа активных станций в сети. Сплошными линиями показана пропускная способность, полученная при расчете по математической модели, разработанной в среде GPSS World. Видно, что значения пропускных способностей, полученных в результате математического и имитационного моделирования отличаются не более чем на 1%, что говорит о высокой точности разработанной математической модели.

Как и следовало ожидать, с ростом числа активных станций пропускная способность сети постепенно падает по причине возрастания вероятности коллизий. При наличии механизма переключения скорости это падение пропускной способности с ростом числа активных станций становится более резким и зависит от пороговых значений счетчиков s и e последовательных успешных и неудачных попыток. На Рис. 6 показаны графики отношения пропускных способностей сети с механизмом переключения скорости и в случае работы на оптимальной битовой скорости при SNR=13 дБ для некоторых пар порогов s и e. Видим, что начиная с некоторых значений N числа станций это отношение не превышает 50%. Естественно, такое снижение пропускной способности неприемлемо для эффективной и стабильной работы беспроводной сети. Также видно, что при одних значениях порогов s и e пропускная способность остается на высоком уровне в большом диапазоне значений N, в то время как при других – в очень ограниченном. В связи с этим возникает следующая задача: найти такие пары порогов s и e, чтобы пропускная способность S_{ARF} в сети с механизмом переключения скорости отличалась от пропускной способности S при работе на оптимальной скорости не более чем на α процентов при любых отношениях сигнал/шум и как можно более широком диапазоне значений активных числа N станций в сети.

Множество решений этой задачи при $\alpha = 0.1$ приведены на Рис. 7.

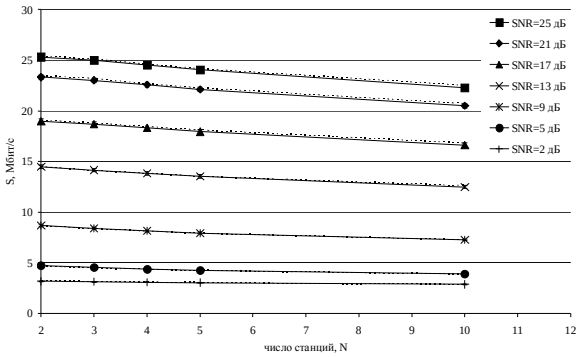


Рис. 5 Работа на оптимальной битовой скорости

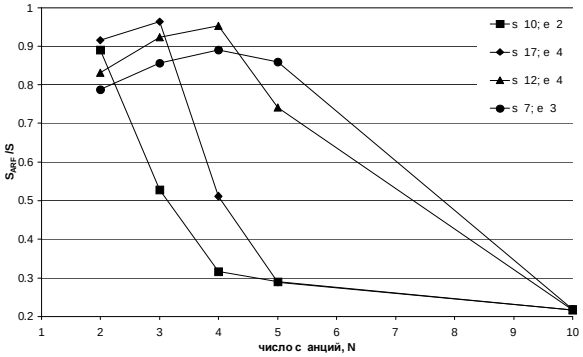


Рис. 6 Относительная пропускная способность сети

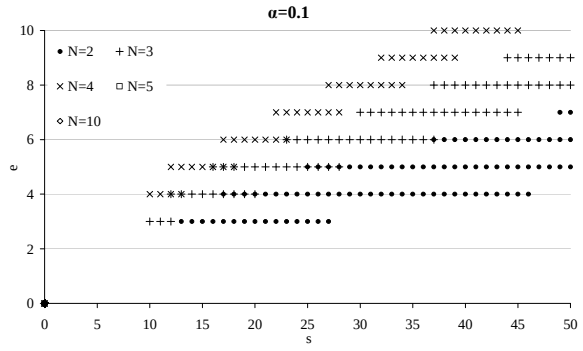


Рис. 7 Набор пар (e, s) при $\alpha = 0.1$

Для нахождения наиболее подходящих пар s и e из этого множества введем понятие *адаптивности* механизма переключения скорости, показывающей насколько быстро он откликается на изменившиеся условия передачи. Понятно, что чем меньше каждый из порогов s и e, тем своевременнее механизм реагирует на изменения отношения сигнал/шум в канале. В связи с этим, естественно рассматривать в качестве меры адаптивности величину, обратную сумме его порогов, т.е. $\frac{1}{s+e}$.

Из графика на Рис. 7 можно заключить, что при $\alpha = 0.1$ наиболее эффективными являются схемы переключения с порогами $s = 17; e = 4$ для диапазона $N = [2;3]$ и с порогами $s = 12; e = 4$ для диапазона $N = [3;4]$. Из проведенного анализа следует, что механизм переключения скорости ARF эффективен лишь при малом (не более четырех) числе активных станций в беспроводной сети. На первый взгляд, это говорит об ограниченности области применения данного механизма. Однако, как показано в работе [4], в случае передачи TCP-трафика по беспроводной сети, а именно такой тип трафика является преобладающим в современных сетях, активными в одно и то же время может быть лишь базовая станция и 1-2 оконечных станций. Отсюда следует вывод об эффективности предложенного механизма переключения скоростей в беспроводных сетях.

7. Заключение

В данной статье разработана аналитическая модель обобщенного механизма переключения скорости передачи. С помощью данной модели проведена оптимизация параметров механизма, исследована его эффективность при наличии радиопомех в канале и коллизий, а также определено предельное число активных станций в сети, при котором данный механизм остается работоспособен.

8. Литература

[1] И.В. Зарубинский, М.Ю. Якимов. "Обзор механизмов автоматического переключения скорости передачи в ШБС." Тр. III Международной конференции по проблемам управления, Москва, 7-9 апреля 2008 г., с. 245-246.

[2] IEEE Std 802.11, 2007 Edition, Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specification, June 2007.

[3] Вишневский В.М., Ляхов А.И, Портной С Л и др., Широкополосные беспроводные сети передачи информации // Москва, Техносфера, 2005, 592 с

[4] R. Bruno, M. Conti, E Gregori, “Throughput Analysis and Measurements in IEEE 802.11 WLANs with TCP and UDP Traffic Flows,” IEEE Transactions on Mobile Computing, Volume 7, Issue 2, Feb. 2008, pp. 171 – 186

Исследование свойств мультиплексирования потоков в мультисервисных сетях

Иван Цитович[1], *Иван Титов*[2]
[1] *Институт проблем передачи информации им. А.А.Харкевича РАН*
[2] *Московский технический университет связи и информатики*
cito@iitp.ru

Аннотация

Основой концепции мультисервисной сети является обслуживание различного по своей природе абонентского трафика (аудио, видео, данные). Достоинством статистического мультиплексирования является то, что при наличии большого числа независимых источников пиковая скорость суммарного (агрегированного) потока будет намного меньше суммы пиковых скоростей входящих потоков, вследствие малой вероятности совпадения по времени пиков скоростей различных потоков. Однако, при ортогональности входных потоков относительно масштаба временной оси эффект компенсации разброса значений не наблюдается.

В работе проводится сравнительный анализ систем обслуживания с различными по природе входными потоками. Показано, что метод управления скоростью передачи, основанный на учёте объёма запрошенной нагрузки, позволяет улучшить параметры выходного потока.

1. Введение

Интеграция в сетях с коммутацией пакетов различного по своей природе трафика, а так же особенности используемых технологий передачи данных для построения сетей связи и повышенная активность пользователей сетей, являются некоторыми из причин проявления самоподобного характера сетевого трафика или, иначе, его фрактальных свойств. Природа возникновения трафика с такими свойствами может быть различной. Одним из факторов является наличие источников нагрузки с пульсирующим трафиком, который характеризуется возможностью возникновения в относительно редкие моменты времени большой нагрузки, которая существенно превосходит средний уровень нагрузки. Математическая модель такого источника нагрузки задается с помощью описания такого источника нагрузки случайной величиной с неограниченной дисперсией. Статистические свойства возникающего потока нагрузки зависят от скорости убывания хвоста вероятностного распределения, задающего распределение интенсивности источника нагрузки. Для эффективного управления ресурсами сети и обеспечения заданных параметров качества обслуживания необходимо управлять значением параметра Херста, задающего соответствующий самоподобный процесс, при этом уменьшение значения этого параметра повышает предсказуемость колебаний трафика и облегчает управление сетью (см. [2], [3]). Для уменьшения значения параметра Херста нужно увеличивать скорость убывания хвоста распределения интенсивности источника нагрузки. В [1] был рассмотрен метод управления скоростью передачи данных сервером для обеспечения заданных значений параметра Херста.

Анализ структуры формируемого такими источниками трафика показал, что эффекты самоподобия трафика возникают из-за того, что различные фрагменты трафика статистически не мультиплексируются. Описанию этих эффектов и методов улучшения свойств формируемого трафика и посвящена работа.

2. Математическая модель

При проведении численных исследований была построена модель сервера данных, обрабатывающего требования с различным временем обслуживания. Пусть источник нагрузки порождает *n* пуассоновских потоков с интенсивностями λ_i, $\Lambda = \sum_{i=1}^{n} \lambda_i$ - суммарная интенсивность поступающих требований, $p_i = \lambda_i / \Lambda$ - вероятность поступления требования из *i*-го потока. Требование *i*-го потока обслуживается в течение a_i единиц времени. Тогда среднее время обслуживание требования составит $A = \sum_{i=1}^{n} p_i \cdot a_i$ единиц времени. В качестве характеристики случайности

источника нагрузки рассматривается среднеквадратическое значение нагрузки $\sigma = \sqrt{D}$, $D = \sum_{i=1}^{n} p_i \cdot a_i^2 - A^2$.

В качестве характеристики, описывающей статистическую однородность нагрузки, рассматривается отношение шума к нагрузке, т.е. величина σ / A. В том случае, когда эта величина принимает небольшие значения, то говорят, что источник обладает дисперсией, если же $\sigma / A >> 1$, то источник не обладает конечной дисперсией и такой источник является объектом наших исследований.

Для простоты положим $\lambda_i = \lambda_1 \cdot k^{-(i-1)}$, $a_i = k^{i-1}$, где k - некоторое достаточно большое число. В этом случае $\Lambda = \frac{k^{-n} - 1}{k^{-1} - 1} \cdot \lambda_1$, $A = \lambda_1 \cdot n / \Lambda$, $A \cdot \Lambda$ задает общую нагрузку в единицу времени и $A \cdot \Lambda = n \cdot \lambda_1$. При таких предположениях все потоки порождают одинаковую нагрузку на ресурс, равную λ_1. Дисперсия такого источника $D = \sum_{i=1}^{n} \frac{\lambda_1 \cdot k^{i-1}}{\Lambda} - \frac{\lambda_1^2 \cdot n^2}{\Lambda^2} \approx k^{n-1} - n^2$ при больших k или n. Поэтому $\sigma / A \approx k^{0.5 \cdot (n-1)} / n \to \infty$ при больших k или n, поэтому при сделанных предположениях мы получаем источник с пульсирующей нагрузкой.

Объем требуемого ресурса для обеспечения пропуска нагрузки источников с $A = 1$ при ограниченной дисперсии источника σ определяется на основании диффузионного приближения модели по формуле $L = \Lambda + u_p \cdot \sqrt{\Lambda} \cdot \sqrt{1 + \sigma^2}$, где u_p - квантиль нормального распределения, определяемая вероятностью потерь P.

Если же потоки не мультиплексируют, то объем требуемого ресурса для обеспечения пропуска нагрузки будет вычисляться по формуле $L = \Lambda + u_p \cdot \sum_{i=1}^{n} \sqrt{\lambda_i} \cdot \sqrt{1 + \sigma^2}$.

2. Модели численных исследований

Для численных исследований использована приведенная выше модель при $n = 3$, $k = 10$ и $\lambda_1 = 33.3$. Поэтому общая нагрузка равна 100.

Для сравнения результатов использовались три модели обслуживания требований. Первая моделирует описанную выше схему, вторая – результат моделирования, если не учитывается структура источника нагрузки, т.е. моделируется пуассоновский поток с интенсивностью Λ и временем обслуживания A. При сделанных предположениях $\sigma = 0$.

Третья модель – это модель из [1], в которой использован метод управления скоростью обслуживания требований, когда время обслуживания увеличивается в a_i^{β} раз, а ширина используемой полосы уменьшается в такое же число раз.

3. Анализ полученных результатов

При проведении моделирования рассчитывался объем используемого ресурса на протяжении 2000 единиц времени. Такая продолжительность должна обеспечить выход процесса на стационарный режим даже при обслуживании требований третьего потока с использованием схемы управления трафиком.

На первой группе рисунков приведены результаты моделирования первой модели. Для второго и третьего рисунков выбирались те временные отрезки, когда средняя нагрузка за период наблюдений была наибольшей.

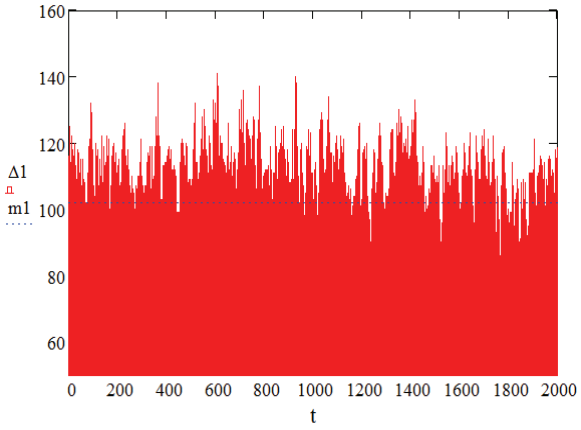


Рис.1. Зависимость объёма занимаемого ресурса за время наблюдений 2000 единиц

На рисунке 1 Δ1 – объём занимаемого в настоящий момент ресурса; m1 – средний за интервал времени 2000 условных единиц объём занимаемого ресурса, m1 = 102.045; t – текущее время.

Максимальное значение Δ1 на данном промежутке равно 141, минимальное 65. Таким образом, максимальное и минимальное отклонение от среднего равны, соответственно, 38.955 и 37.045.

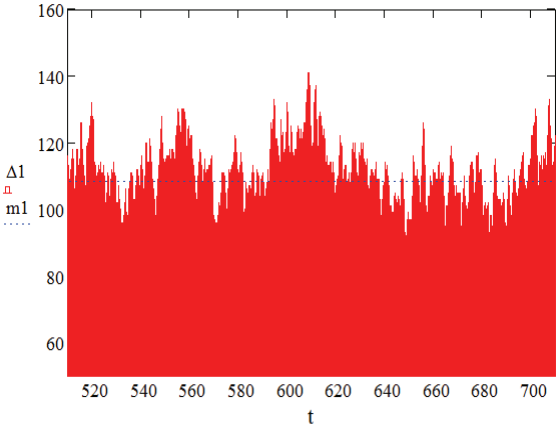


Рис.2. Зависимость объёма занимаемого ресурса за время наблюдений 200 единиц

Средний за интервал времени 200 условных единиц объём занимаемого ресурса равен 108.447. Максимальное значение Δ1 на данном промежутке равно 141, минимальное 86. Таким образом, максимальное и минимальное отклонение от среднего равны, соответственно, 32.553 и 22.447.

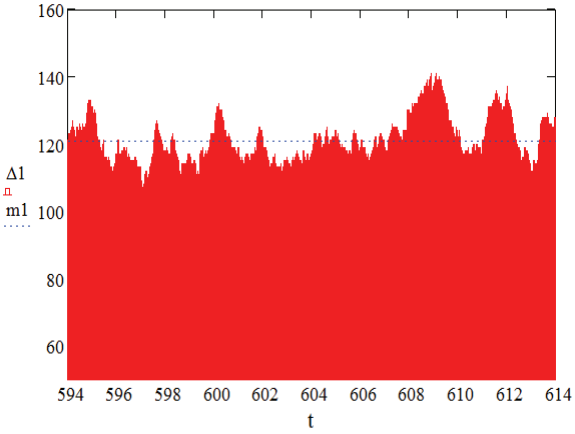


Рис.3. Зависимость объёма занимаемого ресурса за время наблюдений 20 единиц

Средний за интервал времени 20 условных единиц объём занимаемого ресурса равен 120.678. Максимальное значение Δ1 на данном промежутке равно 141, минимальное 105. Таким образом, максимальное и минимальное отклонения от среднего равны, соответственно, 20.322 и 15.678.

На второй группе рисунков приведены результаты моделирования второй системы, когда моделируется чисто пуассоновский процесс поступления требований с фиксированным временем обслуживания, однако требования поступают в те же моменты времени, что и в первом случае.

Для второго и третьего рисунков этой группы так же выбирались те временные отрезки, когда средняя нагрузка за период наблюдений была наибольшей.

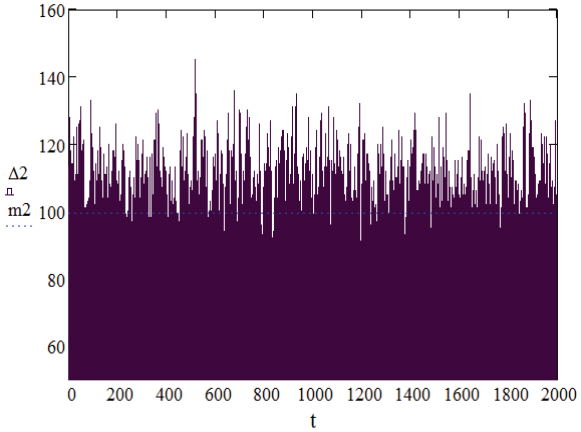


Рис.4. Зависимость объёма занимаемого ресурса за время наблюдений 2000 единиц

Средний за интервал времени 2000 условных единиц объём занимаемого ресурса равен 99.498. Максимальное значение Δ2 на данном промежутке равно 145, минимальное 62. Таким образом, максимальное и минимальное отклонение от среднего равны, соответственно, 45.502 и 37.498.

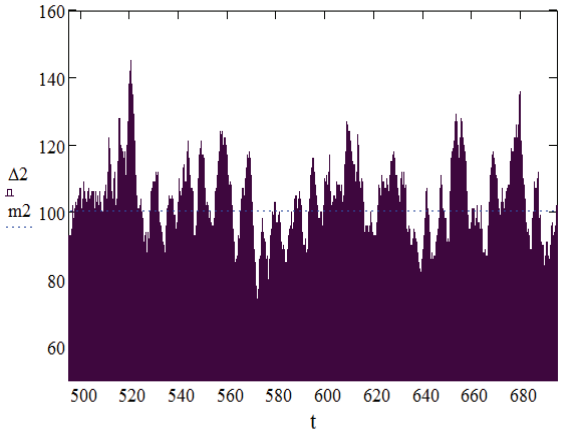


Рис.5. Зависимость объёма занимаемого ресурса за время наблюдений 200 единиц

Средний за интервал времени 200 условных единиц объём занимаемого ресурса равен 100.243. Максимальное значение Δ2 на данном промежутке равно 145, минимальное 68. Таким образом, максимальное и минимальное отклонение от среднего равны, соответственно, 44.934 и 32.066.

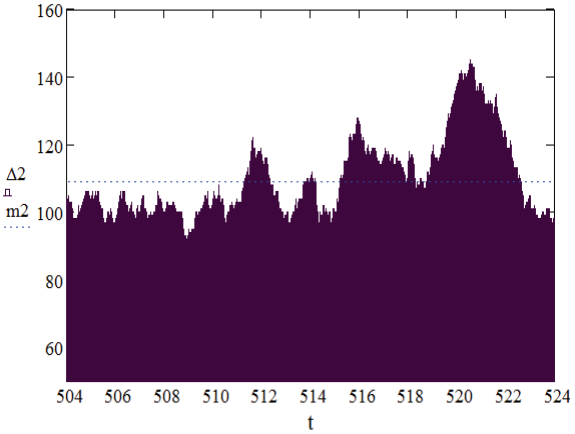


Рис.6. Зависимость объёма занимаемого ресурса за время наблюдений 20 единиц

Средний за интервал времени 20 условных единиц объём занимаемого ресурса равен 109.106. Максимальное значение Δ2 на данном промежутке равно 145, минимальное 90. Таким образом, максимальное и минимальное отклонение от среднего равны, соответственно, 35.894 и 19.106.

На третьей группе рисунков приведены результаты моделирования метода управления скоростью обслуживания требований, потупающих как и в первом случае. Результаты получены при $\beta = 0.5$.

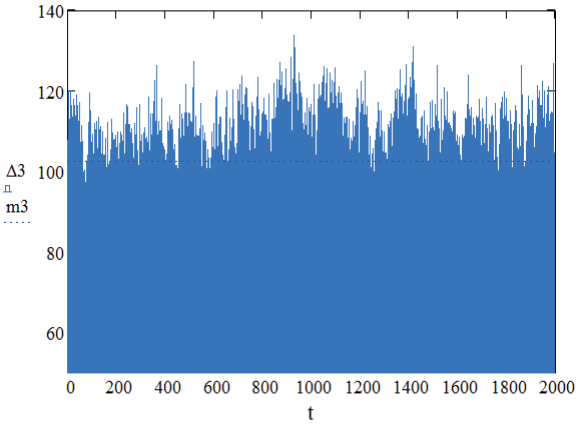


Рис.7. Зависимость объёма занимаемого ресурса за время наблюдений 2000 единиц

Средний за интервал времени 2000 условных единиц объём занимаемого ресурса равен 102.409. Максимальное значение Δ3 на данном промежутке равно 133.88, минимальное 78.34. Таким образом, максимальное и минимальное отклонение от среднего равны, соответственно, 31.471 и 24.069.

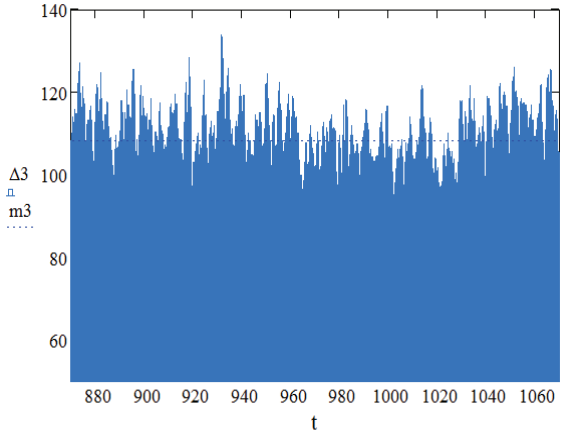


Рис.8. Зависимость объёма занимаемого ресурса за время наблюдений 2000 единиц

Средний за интервал времени 200 условных единиц объём занимаемого ресурса равен 108.263. Максимальное значение Δ3 на данном промежутке равно 133.88, минимальное 88.82. Таким образом, максимальное и минимальное отклонение от среднего равны, соответственно, 25.617 и 19.443.

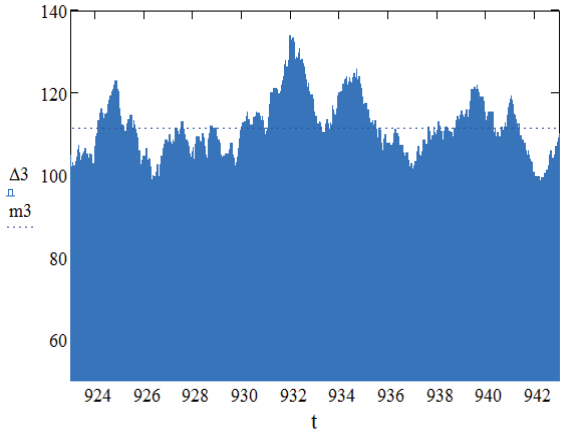


Рис.9. Зависимость объёма занимаемого ресурса за время наблюдений 2000 единиц

Средний за интервал времени 20 условных единиц объём занимаемого ресурса равен 111.258. Максимальное значение Δ3 на данном промежутке равно 133.88, минимальное 96.56. Таким образом, максимальное и минимальное отклонение от среднего равны, соответственно, 22.742 и 14.698.

Сравнение рис. 1 и 4 показывает, что в первом случае имеется значительное количество временных интервалов, когда среднее значение занятого объема ресурса существенно выше средней нагрузки, что соответсвует аналогичным явлениям у самоподобных диффузионных процессов с большим значением параметра Херста.

Таким образом, видно, что игнорирование структуры входящего потока требований может существенно исказить картину обслуживания требований. Хотя на рис.4 имеются значительные отклонения от среднего значения, однако это соответствует большим уклонениям. Если ограничиться уровнем потерь 0.95, то граница отклонений порядка 26, что значительно меньше, чем в первом случае. Как видно из рис. 3, уклонение от среднего значения около 40 единиц, что не может быть объяснено просто случайными отклонениями или большими уклонениями для процессов с независимыми приращениями.

На рис. 2 и 3 приведены временные участки, когда объем используесого ресурса больше среднего значения на 10 и 20 единиц соответственно. Это соответствет отклонением на σ и $2 \cdot \sigma$ соответственно. Этот эффект и соответствует той ситуации, когда отсутствует статистическое мультиплексирование, поскольку на временном интервале, когда обслуживаются требования второго или третьего потоков, успевает поступить значительное число требований первого потока и из-за наличия случайности во времени поступления требований возможны случайные отклонения, которые не могут быть компенсированы требованиями других потоков. Аналогичная картина и при анализе нагрузки второго и третьего потоков. Сравнение рис. 6 и 3 показывает, что во втором случае отклонение в 2 раза больше, чем в первом.

Сравнение результатов моделирования, приведенных на рис. 7-9, с соответствующими результатами рис. 1-3 показывает, что применение схемы из [1] для улучшения статистических свойств трафика оказывается полезным, поскольку случайные колебания нагрузки уменьшаются. Как видно из рис. 3 и 6, среднее значение превышения нагрузки над средним значением уменьшилось почти в 2 раза, хотя структура трафика не изменилась в том смысле, что и возникающий при таком моделировании трафик сохраняет черты самоподобного трафика, но с меньшим значением параметра Херста. Отметим, что эффект достигается за счет применения управления только к 10% от общего числа требований.

4. Выводы

Анализ полученных результатов позволяет сделать следующие выводы.

1. В первой системе обслуживания статистическое мультиплексирование не наблюдается, так как каждый из входных потоков «живёт» в своём временном пространстве. Колебания нагрузки оказываются не скомпенсированными. Так же из рис. 4 видно, что присутствуют длительные промежутки времени, когда средний объём занимаемого ресурса превосходит предполагаемое математическое ожидание. Это является следствием присутствия в системе требований второго и третьего потоков.
2. При использовании управления обслуживанием временные зависимости объёма занимаемого в настоящий момент ресурса выглядят более сглаженными. Максимальное и минимальное отклонение от среднего числа обслуживаемых в системе требований значительно меньше, чем в первой системе обслуживания. Благодаря динамическому управлению скоростью передачи влияние требований второго и третьего потока оказываются не столь сильным. Результаты моделирования показывают, что благодаря данному методу управления скоростью передачи удаётся уменьшить дисперсию нагрузки и улучшить параметры трафика.
3. При исследовании трафика создаваемого пульсирующими источниками целесообразно разделять требования с существенно различным объемом запрашиваемого ресурса и рассматривать его как несколько потоков. При исследовании отдельных групп потоков можно оставаться в рамках классических представлений о виде трафика, а при рассмотрении всей совокупности трафика учитывать отсутствие статистического мультиплексирования. Это позволяет упростить методы расчета пропускной способности каналов и учитывать при этом эффекты, связанные с наличием пульсирующих источников нагрузки.

5. Литература

[1] Цитович И.И., Титов И.Н. Сравнительный анализ эффективности некоторых методов управления скоростью сервера данных // Сборник трудов 30-й конференции молодых учёных и специалистов ИППИ РАН: Информационные технологии и системы ИТиС'07. Москва. 2007. С. 189-192.

[2] Цитович И.И. Устойчивые модели трафика мультисервисных сетей // Труды Российского научно-технического общества радиотехники, электроники и связи имени А.С.Попова. Серия: научная сессия, посвященная Дню радио. Выпуск: LX-2. Москва. 2005. Т2. С. 271-273.

[3] Цитович И.И., Титов И.Н. О методе управления трафиком данных пользователя NGN сетей // Труды Российского научно-технического общества радиотехники, электроники и связи имени А.С.Попова. Серия: научная сессия, посвященная Дню радио. Выпуск: LXII. Москва. 2007. С. 235-236.

Об интервальных моделях телетрафика мультисервисных сетей

Иван Цитович[1], Никита Догадченко[2]
[1] *Институт проблем передачи информации им. А.А. Харкевича РАН*
[2] *Московский технический университет связи и информатики*
cito@iitp.ru, ndogadchenko@gmail.com

Аннотация

Исследование моделей пакетных сетей передачи информации идет по пути их усложнения для обеспечения более точного описания сложных явлений, происходящих в современных сетях связи Вместе с тем, усложнение модели приводит к возрастанию числа параметров, описывающих функционирование такой модели В связи с этим возникает задача обеспечения компромисса между сложностью модели и ее точностью для обеспечения заданных параметров точности в определении важных характеристик функционирования сети В работе проводится анализ точности получаемых значений пропускной способности звена сети, основанных на использовании интервальных моделей ее описания, и влияние усложнения модели на точность определения окончательных характеристик модели

1. Введение

Для обеспечения большей точности и адекватности исследование моделей современных пакетных сетей передачи информации идет по пути их усложнения Усложнение происходит по многим направлениям: введение большого числа потоков, обладающих различными статистическими свойствами, уточнение поведения потока за счет усложнения модели его описания, введение управления потоками для учета приоритета в обслуживании и обеспечения заданных параметров качества обслуживания

Вместе с тем, усложнение модели приводит к возрастанию числа параметров, описывающих функционирование такой модели как правило, численные значения параметров модели получаются на основании статистических измерений, которые могут результат с ошибкой При этом необходимо отметить [1], что чем сложнее вероятностная модель обслуживаемого потока, тем труднее получить оценки его параметров с надлежащим качеством, поэтому погрешность в параметрах модели становится значительной

Даже в простейших случаях, когда делаются предположения о независимости поступающих потоков требований, при их большом числе и большой емкости сети число определяемых параметров (стационарных вероятностей модели) столь велико, что приходится прибегать к приближенным методам расчета [2], [3] Таким образом, результат вычислений даже при предположении о точном знании параметров модели и об ее адекватности будет приближенным

Введение в модель элементов управления, что совершенно необходимо при наличии приоритетов в обслуживании и/или для обеспечения заданных параметров трафика [5], существенно усложняет модель и предъявляет дополнительные требования к точности определения характеристик модели, поскольку их значения влияют на управляющее воздействие

В связи с этим возникает задача обеспечения компромисса между сложностью модели и ее точностью с целью обеспечения заданных параметров в определении важных характеристик функционирования сети Некоторые из возникающих при этом проблем и является предметом исследования

2. Интервальный анализ

Интервальный анализ это раздел, предметом которого является решение задач с интервальными (или, более общо, ограниченными) неопределенностями и неоднозначностями в данных, возникающими либо в постановке задачи, либо на промежуточных стадия процесса решения, чьей характеристической особенностью является рассмотрение множеств неопределчнности как самостоятельных целостных объектов, посредством установления арифметических, аналитических и т п операций и отношений между

ними

Интервальный анализ и его специфичные методы имеют, таким образом, наивысшую ценность в задачах, где неопределенности и неоднозначности возникают с самого начала и являются неотъемлемой частью постановки задачи Хотя это никоим образом не исключает других плодотворных применений интервального анализа, в частности, в задачах, формулируемых вообще без привлечения понятия интервала Например, в последние десятилетия интервальный анализ получил широчайшее распространение в качестве основы для так называемых доказательных (достоверных, надежных) вычислений на ЭВМ, вычислений с гарантированной точностью и т п , несмотря на то, что в этих приложениях интервальные методы являются всего лишь вспомогательным средством для решения задач неинтервальных по своей природе Интервальная идея по своей сути алгоритмична и требует реализации на вычислительной машине, и потому неудивительно, что в докомпьютерную эпоху развитие интервального анализа не состоялось Но уже в середине прошлого века в связи с развитием методов приближчнных вычислений и, особенно, с появлением и распространением первых ЭВМ, потребность в интервальных методах и оценках стала ощущаться столь остро, что пионерские работы по интервальному анализу появились практически одновременно и независимо в Советском Союзе, США, Японии и Польше

Интервальный анализ и возникшая практически одновременно с ним теория нечетких множеств явились ответом на вызов бурно развивающейся практики, которая требовала развития аппарата для учета неопределенностей нестатистической (или, в общем случае, неизвестной) природы При этом интервальный анализ способен исследовать содержательные модели, которые основываются на наиболее скудных априорных допущениях о характере неопределчнности, когда относительно рассматриваемых величин ничего не известно, кроме их свойства принимать значения из заданных ограниченных множеств

Характерная черта исследований, в которых интервальный анализ используется для доказательных вычислений на ЭВМ с конечной разрядной сеткой (т е для получения математически гарантированных результатов с уччтом ошибок округлений) - допущение о малости интервалов изменений "входных" данных, позволяющее во многих случаях осуществлять асимптотический анализ и т п Но погрешности вычислений необходимо учитывать при этом во всех без исключения операциях на ЭВМ, формирующих окончательный результат

Существенное влияние на работы по этой тематике оказывают конкретные особенности вычислительных машин и процессоров, их архитектура, языки программирования и пр Напротив, в тех работах, где интервальный анализ служит средством для исследования ограниченных неопределенностей, опираться на малость возмущений уже нельзя, размеры "входных" интервалов потенциально могут быть сколь угодно велики, но зато часто предполагается, что все арифметические операции как с точечными (неинтервальными) величинами, так и с интервалами выполняются абсолютно точно

3. Интервальные системы линейных уравнений

Одним из стандартных методов описания сетей передачи информации является применение случайных процессов, для которых ищутся стационарные вероятности В этом случае необходимо решать систему линейных уравнений для нахождения этих вероятностей В том случае, когда параметры системы точно неизвестны, возникает задача решения систем линейных уравнений с интервальными коэффициентами [4] В этом случае нужно решать системы линейных алгебраических уравнений

$$\sum_{j=1}^{n} a_{ij} x_j = b_i, \quad i = 1, 2, \ldots, n,$$

с неопределенными коэффициентами a_{ij}, b_i:

$$a_{ij}^- \le a_{ij} \le a_{ij}^+, \quad b_i^- \le b_i \le b_i^+,$$
$$i = 1, 2, \ldots, m, \quad j = 1, 2, \ldots, n,$$

где $a_{ij}^-, a_{ij}^+, \underline{b}_i, \bar{b}_i$, заданные вещественные числа, задающие интервалы изменения коэффициентов системы уравнений, x_j искомые неизвестные

Составим стандартным образом из коэффициентов и неизвестных $n \times n$-матрицы A, A^+, A^-, n-векторы b, b^+, b^-, x и условимся понимать матричные и векторные неравенства поэлементно Тогда исходные уравнения и неравенства для коэффициентов удобно записать в виде

$$Ax = b, \tag{1}$$
$$A^- \le A \le A^+, \quad b^- \le b \le b^+. \tag{2}$$

Пары A, b, удовлетворяющие требованиям (2), будем называть *допустимыми*

Введем в рассмотрение неотрицательный n-вектор ε, характеризующий точность выпол-

нения равенства (1) Вектор x называется *ε-решением* уравнения (1), если неравенства

$$-\varepsilon \le Ax - b \le \varepsilon \tag{3}$$

выполняются для любых допустимых пар A, b За *универсальное* решение уравнения (1) примем ε-решение с минимальным в некоторой норме вектором ε Отвечающий универсальному решению x вектор ε называется *минимальной невязкой* уравнения (1) Тогда по определению пара x, ε удовлетворяет неравенству $|Ax - b| \le \varepsilon$ для любых допустимых A, b и среди всех пар $\tilde{x}, \tilde{\varepsilon}$ с тем же свойством выделяется условием $\|\varepsilon\| \le \|\tilde{\varepsilon}\|$

Задача нахождения ε-решений и универсальных решений оказывается весьма трудной с вычислительной точки зрения, поскольку для этого необходимо решать задачу линейного программирования, размерность которой больше размерности исходной задачи Эти трудности особенно значительны при больших значениях n В связи с этим в [4] введено понятие субоптимального решения, которое близко к универсальному решению

Пусть

$$A^0 = 0,5(A^- + A^+), \quad \Delta A = 0,5(A^+ - A^-),$$
$$b^0 = 0,5(b^- + b^+), \quad \Delta b = 0,5(b^+ - b^-),$$
$$A_0^+ = (A^0)'(A^0(A^0)')^{-1},$$

где A' транспонированная матрица A, а

$$\hat{x} = A_0^+ b^0. \tag{4}$$

Решение (4) называется субуниверсальным решением уравнения (1) Для этого решения получена оценка его погрешности:

$$\hat{\varepsilon} = \varepsilon(\hat{x}) = \Delta A|\hat{x}| + \Delta b \tag{5}$$

и $\hat{x}$ есть $\hat{\varepsilon}$-решение уравнения (1) с наименьшей для него невязкой

Для отыскания значений характеристики сети связи по стационарным вероятностям вычисляют значения характеристик как средние значения некоторой величины неотрицательной $c = (c_1, \ldots, c_n)$, где c_i значение интересующей нас характеристики, когда система находится в состоянии i Из (5) получаем, что погрешность при вычислении среднего значения c составит величину

$$\Delta c = cA|\hat{x}| + (c, \Delta b_i) = \\ = \sum_{j=1}^{n}\sum_{i=1}^{n} c_i \Delta a_{ij}|\hat{x}_j| + \sum_{i=1}^{n} c_i \Delta b_i. \tag{6}$$

В простейшем случае, когда все погрешности примерно одинаковы, т е , например, $\Delta a_{ij} \le \delta_a$, $\Delta b_i \le \delta_b$, $c_i \le c_{max}$, то (6) принимает вид

$$\Delta c \le n c_{max} \left(\delta_a \sum_{j=1}^{n} |\hat{x}_j| + \delta_b \right), \tag{7}$$

т е погрешность растет с ростом числа состояний

4. Анализ погрешности интервальных моделей при исследовании мультисервисных сетей

Наиболее простые вероятностные модели функционирования современных сетей связи получаются при использовании процессов рождения и гибели Основными характеристиками таких процессов оказываются вероятность потерь требований и/или средние значения нагрузки и т п Все они будут неотрицательными функционалами от стационарных вероятностей соответствующего процесса рождения и гибели

Поскольку в этом случае x это вектор стационарных вероятностей, а вектор $\hat{x}$ должен быть близок к x, то (7)

$$\Delta c \le n c_{max} \left(K\delta_a + \delta_b \right), \tag{8}$$

т е пропорциональна числу состояний, причем в случае неограниченного функционала c (например, если вычисляется среднее значение занятого ресурса) граница погрешности становится еще больше

Простейшие модели мультисервисных сетей строятся в предположении, что имеется несколько независимых потоков, описываемых процессами рождения и гибели ([2], [3]) Каждый поток задается интенсивностью поступления требования, интенсивностью обслуживания и требуемой полосой обслуживания Если L число рассматриваемых потоков, то общее число состояний оказывается порядка n^L и если пользоваться оценкой (8) для определения погрешности вычислений, то результат становится неприемлемым с практической точки зрения даже при весьма точном знании параметров потоков Следует отметить, что и сама модель является приближенной, поскольку предположения о статистических свойствах обслуживаемых потоков и времени обслуживания обычно достаточно далеки от реальных, и методы вычислений даже в этом простейшем случае приходится использовать приближенные при небольших значениях L

Если проанализировать протоколы, например, Wi-Max-сетей, то необходимо отметить, что

в этом случае предусмотрено 5 дисциплин обслуживания, а число регистрируемых потоков L может быть порядка миллиона Понятно, что применение стандартных методов расчета становится неприемлемым даже при предположениях о независимости потоков

В таких случаях представляется целесообразным огрублять модель аналогично алгоритму Кауфмана Робертса, когда состояние задается емкостью занятого ресурса В этом случае можно оценить погрешность, которая возникает в определении интенсивностей перехода между состояниями из-за того, то в укрупненное состояние входят состояния с различными интенсивностями перехода, однако при таком подходе колебания могут быть относительно небольшими

Другой подход состоит в том, что для описания эффекта наличия зависимости между моментами поступления требования в систему используют уложения в описании статистических свойств входного потока Один из таких подходов состоит во введении марковских процессов более высокого порядка Это приводит к увеличению пространства состояний: если изначально было n состояний, то при рассмотрении марковской цепи l-го порядка получим n^l состояний Это, в соответствии с (8), приведет к росту погрешности в вычислениях Дополнительным фактором здесь является то, что относительная точность определения переходных вероятностей для марковской цепи более высокого порядка будет значительно ниже, чем для обычной марковской цепи (первого порядка), т е значение δ_a будет существенно больше

Таким образом, возникает задача достижения компромисса при выборе порядка марковского процесса, описывающего входной поток Может оказаться, что существенно лучший результат с точки зрения погрешности будет получен в том случае, когда используется процесс более низкого порядка, однако в (1) вносятся поправки, связанные с огрублением модели В этом случае возрастает величина δ_b, однако существенно уменьшается число состояний, поэтому погрешность вычислений будет значительно меньше

5. Выводы

1 Свойство "гарантированности" ответов, обеспечиваемое методами интервального анализа часто является избыточным, а соответствующие интервальные оценки (даже оптимальные) недостижимыми и излишне "пессимистичными" Это во многом связано с тем, что не оптимизирована исходная модель для расчетов с точки зрения ее устойчивости относительно ошибок в определении параметров модели

2 В некоторых интервальных задачах ослабление требований может, на наш взгляд, привести к упрощению постановок и/или созданию более эффективных алгоритмов для их решения

3 Формула определения погрешности (5) является неустойчивой относительно эквивалентных преобразований (1), сохраняющих множество ее решений В связи с этим возникает дополнительная задача задания системы (1) таким образом, чтобы погрешность в определении решения была минимальной с точки зрения определения характеристик модели

4 При решении задач с управлением потоками или обслуживанием, как, например, следует из [6], имеется некоторый диапазон изменения параметров системы, когда результат управляющего воздействия слабо зависит от значений параметров В этом случае, при определении порогов выбора управляющего воздействия можно применять интервальные модели с заранее известным требованием к точности решения задачи; при этом точное решение задачи не требуется

6. Литература

1. *Tsitovich I., Naumova E., Tsitovich F.* Markov chain order estimating and its applications in bioinformatics // XII-th International summer conference on probability and statistics. Bulgaria, Sozopol. 2006. P. 38.

2. *Лагутин В.С., Степанов С.Н.* Телетрафик мультисервисных сетей связи. М.: Радио и связь, 2000.

3. *Наумов В.А., Самуйлов К.Е., Яркина Н.В.* Теория телетрафика мультисервисных сетей. М.: РУДН, 2007.

4. *Ашепков Л.Т.* Универсальные решения интервальных задач оптимизации и управления. М.: Наука, 2006.

5. *Цитович И.И.* Устойчивые модели трафика мультисервисных сетей // Труды Российского научно-технического общества радиотехники, электроники и связи имени А.С.Попова. Серия: научная сессия, посвященная Дню радио. Выпуск: LX-2. М., 2005. Т2. С.271-273.

6. *Цитович И.И., Чернушевич А.В.* Влияние гистерезиса на управление приоритетами в телекоммуникационной сети // Обозрение прикладной и промышленной математики. Т. 15. № 4. 2008.

Пропускная способность широкополосного беспроводного канала точка-точка в режиме насыщения[1]

Шпилев С. А.
Институт проблем передачи информации им. А.А. Харкевича РАН, Москва
santt@iitp.ru

Аннотация

Характеристики широкополосных беспроводных сетей под управлением протокола IEEE 802.11 с функцией распределенного управления (DCF, Distributed Coordination Function) в отсутствии шумов в режиме насыщения, особенно сетей с топологией точка-точка, хорошо изучены как аналитически, так и имитационно. Однако характеристики многокилометрового канала точка-точка с присущими ему особенностями с функциями распределенного, централизованного (PCF, Point Coordination Function) и гибридного (HCCA, Hybrid Controlled Channel Access) управления пока не исследованы. В данной работе аналитическая модель беспроводной сети с функцией распределенного управления в режиме насыщения, предложенная В.М. Вишневским и А.И. Ляховым [7], применяется для моделирования обычного и модифицируется для случая многокилометрового канала точка-точка. Данная модель применяется для вычисления пропускной способности указанных сетей. Представлен сравнительный анализ результатов аналитического и имитационного (на языке GPSS World [6]) моделирования. Также для сравнения приводится аналитически полученная пропускная способность данных каналов в режиме насыщения в случае использования централизованной или гибридной функции управления.

1. Введение

Режим насыщения для сетей с функцией DCF в отсутствии скрытых станций и шумов хорошо описан аналитически в многочисленных работах: смотри, например, [2], [3], [7]. Однако, во всех этих работах рассматривается локальная сеть, то есть расстояние между станциями считается достаточно малым (порядка ста метров), а современный стандарт IEEE 802.11 [4] предусматривает расположение станций и на большом расстоянии друг от друга (до 115 км). При большом расстоянии между станциями, поведение сети существенно отличается от рассмотренного в указанных работах. Простейшим и применяемым на практике случаем такой сети является многокилометровый канал точка-точка. Результаты имитационного исследования такого многокилометрового канала приведены в [1], однако методов его аналитического исследования до сих пор не существует. В данной работе, модель разработанная в [7], модифицируется для случая многокилометрового канала.

В случае, когда расстояние между станциями превышает несколько километров, время распространения сигнала от одной из них до другой (δ) становится настолько велико, что может даже превышать время передачи самого пакета. Слоты отсрочки призваны снизить вероятность коллизии, то есть если одна станция начала передачу в данном слоте, то другая станция, при условии, что она сама не начала передачу в нем, должна узнать о том, что канал занят, до окончания данного слота. Так как о начале передачи одной станцией другая узнает только через время δ, то виртуальные слоты двух станций смещены друг относительно друга на это время. Однако, если начнет передавать станция, слот которой начинается позже на δ, то до другой станции этот сигнал дойдет через время 2δ после начала ее слота отсрочки. Таким образом, слот отсрочки должен превышать время 2δ, то есть увеличиваться при увеличении расстояния между станциями, что и определено протоколом [4]. Межкадровые интервалы DIFS и EIFS, включают в себя несколько слотов, поэтому тоже увеличиваются с увеличением данного расстояния. Все это вместе приводит к тому, что пропускная способность сети сильно снижается. Искажается и само понятие коллизии ввиду того, что при

[1] Работа выполнена при поддержке Российского Фонда Фундаментальных Исследований, грант № 08-07-90102-Мол_а.

определенном соотношении времени передачи пакета и времени распространения сигнала коллизия для одной станции вовсе не означает коллизию для второй (рис. 2). Влияние всех указанных факторов учтено при построении аналитической модели и получении численных результатов.

2. Модель

Согласно подходу Бьянки (Bianchi) [2] в модели используется следующая дискретная целочисленная шкала времени: t и $t+1$ соответствуют началу двух последовательных виртуальных слотов. Виртуальным слотом называется интервал времени между двумя последовательными изменениями величины интервала отсрочки. Виртуальные слоты неодинаковы. Каждый из них может быть: i) "пустым" слотом отсрочки σ, когда ни одна станция не передает, ii) "успешным" слотом, когда одна и только одна станция передает и iii) "коллизионным" слотом, когда две или более станций пробуют передать одновременно.

Рассмотрим сеть состоящую из n узлов. Согласно модели Бьянки в начале каждого слота все станции сети имеют одинаковую вероятность τ начать отправку пакета. В модели Кали (Calì) [3], время отсрочки b предполагается независящим от количества попыток послать текущий пакет n_c и значения $b = 0, 1, 2, \ldots$ выбираются из геометрической прогрессии с вероятностями $\tau, \tau(1-\tau), \tau(1-\tau)^2, \ldots$ соответственно. Однако τ выбирается с учетом реального правила отскрочки и вероятности коллизии. Как показано в [7], используя данные предположения легко получить вероятности возникновения "пустого" (p_e), "успешного" (p_s), и "колизионного" (p_c) слотов:

$$p_e = (1-\tau)^n, \;\; p_s = n\tau(1-\tau)^{n-1},$$
$$p_c = 1 - (1-\tau)^n - n\tau(1-\tau)^{n-1}, \qquad (1)$$

где n – это количество станций.

Для упрощения дальнейшего анализа предположим, что используется только базовый механизм доступа к среде (без использования механизма RTS/CTS) и будем считать, что все пакеты в сети имеют одинаковый размер. Случай, когда размеры пакетов различны и используется маханизм RTS/CTS, представлен в [7]. Обозначим длительности слотов в случае короткого канала точка-точка верхним индексом 0; в случае многокилометрового канала точка-точка при условии, что $H + P \le 2\delta$ – индексом s (short packet, короткий пакет); и при условии, что $H + P > 2\delta$ – индексом l (long packet, длинный пакет). Здесь P и H – времена необходимые для передачи данных и загаловка пакета данных соответственно. Разделение случая многокилометрового канала на два типа связано с тем, что при коротком пакете в случае коллизии одна станция все же успешно передает пакет (рис. 2), а при длинном пакете коллизия происходит практически также, как в случае короткого канала точка-точка (рис. 3). Время передачи пакета для данных двух типов не различается, поэтому оно отображено только на рис. 2.

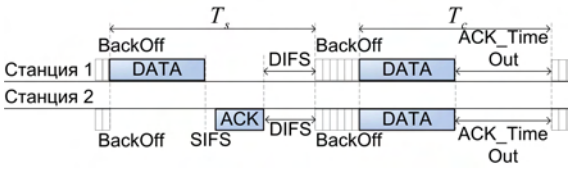


Рис. 1. Стандартный канал, режим DCF

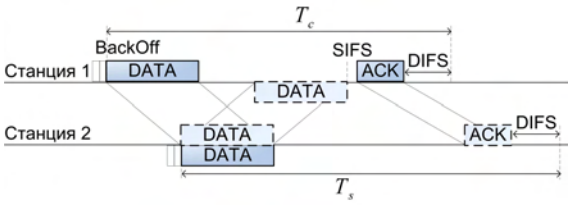


Рис. 2. Многокилометровый канал, режим DCF, короткие пакеты

Как ясно из рис. 1, 2 и 3, длительности виртуальных слотов не всегда совпадают для обеих станций, однако под длительностью слота будем подразумевать большую из них. Выбор большей из величин объясняется тем, что начала слотов станций расходятся на время распространения сигнала, и то слоты первой отстают от второй, то наоборот. Таким образом, дли-

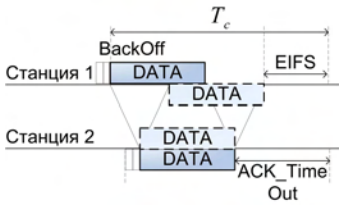


Рис. 3. Многокилометровый канал, режим DCF, длинные пакеты, коллизия

тельности "успешного" и "коллизионного" слотов – T_s и T_c задаются выражениями

$$T_s^0 = H + P + t_{ACK} + \text{SIFS} + \text{DIFS}, \quad (2)$$

$$T_c^0 = H + P + \text{ACK_TimeOut}, \quad (3)$$

$$T_s^s = T_c^s = H + P + t_{ACK} + 2\delta + \\ + \text{SIFS} + \text{DIFS}, \quad (4)$$

$$T_s^l = T_s^s, \quad (5)$$

$$T_c^l = H + P + 2\delta + \text{EIFS}, \quad (6)$$

где t_{ACK} – время необходимое для передачи ACK пакета. Таким образом, определены размеры всех типов виртуальных слотов.

Так же, как это сделано в [3], рассмотрим временной интервал t_v между двумя последовательными успешными передачами, который будем называть виртуальным временем передачи. Тогда пропускная способность в режиме насыщения S определяется формулой

$$S = \frac{V_c P}{E[t_v]}, \quad (7)$$

где $E[t_v]$ – среднее значение t_v, а V_c – номинальная скорость передачи данных в канале.

В случае обычного канала, либо многокилометрового канала при условии $H + P > 2\delta$ виртуальное время передачи может состоять из $l = 1, 2, \ldots$ виртуальных слотов, где последний слот "успешный", $k = 0, \ldots, l-1$ слотов "коллизионные" и $l - 1 - k$ слотов "пустые", то есть $t_v^0 = T_s^0 + kT_c^0 + (l - 1 - k)\sigma$. Таки образом

$$E[t_v^0] = p_s \sum_{l=1}^{\infty} \sum_{k=0}^{l-1} [T_s^0 + kT_c^0 + (l - 1 - k)\sigma] \times \\ \times \binom{l-1}{k} p_c^k p_e^{l-1-k}.$$

Преобразуя выражение в правой части, получаем

$$E[t_v^0] = T_s^0 + \frac{p_c}{p_s} T_c^0 + \frac{p_e}{p_s}\sigma. \quad (8)$$

Аналогичное выражение получено в [3] и [7].

Однако, в случае многокилометрового канала при условии $H + P \leq 2\delta$ в случае возникновения коллизии одна из станций, тем не менее, успешно передает свой пакет (рис. 2). Поэтому виртуальное время передачи может состоять из $l = 1, 2, \ldots$ виртуальных слотов, где последний слот "успешный" или "коллизионный", а оставшиеся $l - 1$ слотов "пустые", то есть $t_v^s = T_s^s + (l - 1)\sigma$ или $t_v^s = T_c^s + (l - 1)\sigma$. Таки образом,

$$E[t_v^s] = p_s \sum_{l=1}^{\infty} [T_s^s + (l - 1)\sigma] p_e^{l-1} + \\ + p_c \sum_{l=1}^{\infty} [T_c^s + (l - 1)\sigma] p_e^{l-1}.$$

Преобразуя выражение в правой части, получаем

$$E[t_v^s] = \frac{T_s^s p_s + T_c^s p_c}{1 - p_e} + \frac{\sigma p_e (p_c + p_s)}{(1 - p_e)^2}. \quad (9)$$

Таким образом, чтобы определить пропускную способность S в режиме насыщения, остается найти τ. Как и в [7], конкурентное окно при i-ой посылке пакета задается выражением

$$W_i = \begin{cases} CW_{min} 2^i, & \text{при } i < m; \\ CW_{max} = CW_{min} 2^m, & \text{при } i \geq m. \end{cases}$$

Воспользуемся формулой (9) из [7]

$$\tau = 2/(E[w] + 1) \quad (10)$$

и леммой 1 из [7], которая гласит: если количество повторных посылок пакета ограничено числом R, то среднее конкурентное окно определяется следующими формулами: при $R \leq m$

$$E[w] = \left\{ \sum_{i=1}^{R} p^{i-1} W_i + \frac{p^R W_R - W_0}{1 - p} \right\} \times \\ \times \frac{(1 - p)^2}{1 - p^R} \quad (11)$$

и

$$E[w] = \left\{ \sum_{i=1}^{m} p^{i-1} W_i + W_m \sum_{i=m}^{R-1} (i - m + 2) p^i + \right. \\ \left. + \frac{(R - m + 1) W_m p^R - W_0}{1 - p} \right\} \frac{(1 - p)^2}{1 - p^R}, \quad (12)$$

в противном случае. Где $p = 1 - (1 - \tau)^{n-1}$ – вероятность того, что данная попытка передачи оказалась неуспешной. Следовательно в случае канала точка-точка $p = \tau$.

Решая численно систему уравнений (10)-(12), последующей подстановкой результата в (1)-(9), получаем пропускную способность канала точка-точка во всех описанных случаях.

В случае использования функции централизованного или гибридного управления в режиме насыщения, осуществляется неконкурентный доступ к среде, а в этом случае нет коллизий и отсрочки посылки пакетов. На рис. 5 вместо пакетов данных и подтверждений координатор использует пакет $D + A + P$ – данные с подтверждением и опросом (Poll), а вторая станция использует пакет $D + A$ – данные с подтверждением, что разрешено стандартом [4], для минимизации накладных расходов.

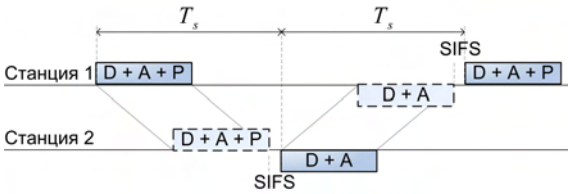


Рис. 4. Режим HCCA

При этом размеры этих пакетов равны размеру обычного пакета данных. Из рис. 5 видно, что в этом случае виртуальное время передачи задается формулой $t_v = H + P + \delta + \text{SIFS}$, и следовательно

$$S = \frac{V_c P}{H + P + \delta + \text{SIFS}}. \quad (13)$$

Здесь стоит оговориться, что согласно стандарту при использовании PCF или HCCA функций в течении некоторого времени все же осуществляется конкурентный доступ к среде, однако длительность данного интервала может быть сколь угодно мала, поэтому в данной работе затраты на переход из одного состояния в другое и период конкурентного доступа не рассматриваются.

3. Результаты

В данном разделе приведены результаты численного решения аналитических моделей совместно с результатами имитационного моделирования и приведено их расхождение (таблицы 1 – 4) для функции DCF. Приведены также результаты численного решения аналитической модели для режима HCCA (таблица 5), ввиду простоты модели и отсутствия в ней приближений, в экспериментальных данных, в этом случае, необходимости нет.

В представленных экспериментах для короткого и многокилометрового канала точка-точка меняется минимальное конкурентное окно. Для многокилометрового канала расстояние между станциями выбрано равным 50 км, и время распространения сигнала δ на такое расстояние равно 166,5 мкс. Размер пакета данных выбирается равным одному килобайту – общепринятый размер пакета, и 2312 байт – максимальный допустимый в стандарте IEEE 802.11 размер пакета. Эти два размера пакета, при выбранном растоянии между станциями в случае многокилометрового канала, соответствуют двум описанным в работе случаям – $H + P \leq 2\delta$ и $H + P > 2\delta$, соответственно. Максимальное количество повторных посылок пакета $R = 7$, а максимальное конкурентное окно $CW_{max} = 1024$, согласно стандарту IEEE 802.11 [4]. Номинальная скорость передачи в канале (V_c) выбрана максимально возможной согласно стандарту IEEE 802.11a [5], то есть 54 Мбит/с, и прочие параметры взяты из него же: $\sigma = 9$ мкс $+ 2\delta$, SIFS = 16 мкс, DIFS = SIFS+2σ, ACK_Timeout = SIFS+σ+25 мкс, EIFS = SIFS + DIFS + 44 мкс. Времена, необходимые для передачи пакета данных и пакета подтверждения на данной скорости, также легко вычисляются: $t_{ACK} = 24$ мкс, а $H + P = 180$ мкс при размере пакета 1024 байт и 372 мкс, при размере пакета 2312 байт.

Таблица 1. Пропускная способность (расстояние 50 км, пакет данных 1024 байт)

CW_{min}	Теория (Мбит/с)	Эксперимент (Мбит/с)	Δ (%)
4	5,008	5,173	3,24
8	4,172	4,335	3,83
16	3,108	3,21	3,23
32	2,050	2,092	2,03
64	1,217	1,233	1,31
128	0,672	0,676	0,59

Таблица 2. Пропускная способность (расстояние 50 км, пакет данных 2312 байт)

CW_{min}	Теория (Мбит/с)	Эксперимент (Мбит/с)	Δ (%)
6	8,503	8,903	4,60
8	7,734	7,792	0,75
16	6,178	6,196	0,29
32	4,285	4,292	0,16
64	2,632	2,631	0,04
128	1,481	1,481	0,00

Таблица 3. Пропускная способность (расстояние $\leq$ 100 м, пакет данных 1024 байт)

CW_{min}	Теория (Мбит/с)	Эксперимент (Мбит/с)	Δ (%)
4	26,493	27,678	4,38
8	27,334	26,686	2,40
16	26,776	26,232	2,05
32	24,429	24,052	1,56
64	20,284	20,043	1,20
128	15,002	14,874	0,86

Таблица 4. Пропускная способность (расстояние $\leq$ 100 м, пакет данных 2312 байт)

CW_{min}	Теория (Мбит/с)	Эксперимент (Мбит/с)	Δ (%)
4	34,423	36,296	5,30
8	36,087	35,537	1,54
16	36,326	35,846	1,33
32	34,692	34,343	1,01
64	30,888	30,636	0,82
128	25,010	24,851	0,64

Таблица 5. Пропускная способность (HCCA)

Расстояние (км)	Пакет (байт)	Пропускная способность (Мбит/с)
0,1	1024	41,80
0,1	2312	47,67
50	1024	22,60
50	2312	33,36

Из таблиц видно, что результаты аналитической модели тем лучше сходятся с результатами имитационной, чем больше минимальное конкурентное окно. Данная особенность легко объясняется тем, что в модели не учтен эффект захвата, который описан в [7], а также предполагается, что τ не зависит от того, сколько раз пакет уже пытались передать, и оба указанных эффекта тем более заметны, чем больше вероятность коллизии, то есть чем меньше конкурентное окно. Однако, несмотря на данное приближение, при $CW_{min} \geq 8$ расхождение результатов не превышает 3,9%, что является весьма хорошим показателем. А в случае многокилометрового канала и пакета данных, такого что $H + P > 2\delta$, это расхождение и вовсе не превышает 0,8%.

Последняя таблица наглядно демонстрирует насколько эффективнее применение механизмов централизованного управления в канале точка-точка в режиме насыщения. Оно позволяет поднять пропускную способность обычного канала более чем на 30%, а в случае 50-и километрового канала и вовсе в 3,7 – 4,6 раз, в зависимости от размера пакета.

4. Заключение

В данной работе предложена модификация аналитической модели, представленной в [7], позволяющая довольно точно оценить максимальную пропускную способность канала точка-точка с функцией DCF в отсутствии шумов при любых расстояниях между станциями и прочих параметрах сети. Предложена также простейшая модель, позволяющая оценить максимальную пропускную способность канала точка-точка при использовании функции HCCA или PCF также при любых расстояниях между станциями. При помощи имитационного моделирования проверена точность аналитической модели и установлены рамки ее применимости. Продемонстрирован ощутимый выигрыш в пропускной способности при использовании механизмов централизованного управления при построении обычного канала точка-точка и многократный выигрыш в случае многокилометрового канала.

Список литературы

[1] С.А. Шпилев и Д.Н. Фахриев, "Анализ характеристик широкополосных беспроводных каналов, под управлением протокола IEEE 802.11", *Третья всероссийская молодежная научная конференция по проблемам управления (ВМКПУ-2008)*, 2008.

[2] G. Bianchi, "Performance Analysis of the IEEE 802.11 Distributed Coordination Function", *IEEE Journal on Selected Areas in Communications*, 2000, Vol. 18, No. 3, pp. 535–547.

[3] F. Calì, M. Conti, and E. Gregory, "IEEE 802.11 Wireless LAN: Capacity Analysis and Protocol Enhancement", *Proceedings of INFOCOM'98*, 1998, pp. 142–149.

[4] *Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, IEEE Std 802.11-2007, June 2007.

[5] *Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: High-speed Physical Layer in the 5 GHz Band*, IEEE Std 802.11a-1999, 1999.

[6] T.J. Schriber, *Simulation using GPSS*, John Wiley & Sons, 1974.

[7] V.M. Vishnevsky and A.I. Lyakhov, "IEEE 802.11 Wireless LAN: Saturation Throughput Analysis with Seizing Effect Consideration", *Cluster Computing*, 2002, Vol. 5, No. 2, pp. 133–144.

Распределения времени присоединения устройств к сети ECMA-368 (WiMedia)

Александр Сафонов
ИППИ РАН
E-mail: safa@iitp.ru

Евгений Хоров
МФТИ (ГУ)
E-mail: horov@frtk.ru

Аннотация

Одним из перспективных стандартов высокоскоростной беспроводной персональной сети является разработка альянса крупнейших телекоммуникационных компаний мира – WiMedia (WiMedia Alliance, http://wimedia.org, ранее известного как Multiband OFDM Alliance, MBOA). В данной работе исследуется эффективность механизмов синхронизации в беспроводных персональных сетях с распределенным управлением. Механизмы синхронизации обеспечивают согласованную работу и особенно важны, когда в сети происходят изменения: присоединение новых устройств, отключение устройств от сети, смена рабочей частоты, изменение топологии, при возникновении помех, интерференции нескольких пикосетей и т. д. Основное внимание в работе уделено изучению одношаговых сетей в случае одновременного присоединения к сети нескольких устройств, что нередко случается в описанных выше случаях. Исследуется распределение времен присоединения устройств к сети. Для этого построены оптимистическая и консервативная аналитические модели процесса присоединения. На основании результатов численного моделирования составлены рекомендации по изменению значений параметров стандарта ECMA-368, которые призваны существенно уменьшить задержки.

Введение

В сети с распределенным управлением на MAC-уровне базовой единицей является устройство. Так как координатора не существует, то каждое устройство формирует свое видение топологии сети, загруженности каналов, конфликтов между устройствами, интерференции с другими устройствами и т д.

Базовой временной единицей пикосети является суперкадр (англ.: superframe). Для выполнения функции синхронизации используются биконы. Биконы посылаются каждый суперкадр. Все биконы передаются в начале (в младших слотах) суперкадра в течение интервала времени, называемого «бикон-периодом» . Бикон каждого устройства занимает фиксированный промежуток времени, называемый бикон-слотом. Продолжительность 1 бикон-слота составляет 85 мкс, т. е. в 3 раза короче обычного слота для передачи данных. В дальнейшем, говоря «слот», будем подразумевать именно «бикон-слот». Первые 2 бикон-слота являются сигнальными и заполняются устройством, бикон которого занимает следующий за ими слот. В работе, если не оговорено иначе, мы для простоты будем нумеровать слоты, без учета сигнальных. Длина бикон-периода варьирует в зависимости от количества устройств в сети. Его максимальная длина не может превосходить значения $MaxBP$ (англ.: Maximum Beacon Period – максимальный бикон-период), определяемого в спецификации WiMedia (в текущей версии – 96 с учетом сигнальных, то есть 94 без них).

Иногда (например, при подсоединении устройств к сети или при рассогласовании часов) возможны ситуации, когда в одном слоте будут размещаться 2 или более бикона. Если коллизия произошла при подключении устройства, то мы будем говорить что, устройство не подключилось успешно к сети.

В бикон каждое устройство помимо остальной информации помещает и некую маску. Поля маски отражают состояние каждого бикон-слота с точки зрения того устройства, которое включило маску в свой бикон.

Чтобы снизить вероятность ложной тревоги, устройство А должно получить от одного и того же устройства В подряд U биконов, сигнализирующих о коллизии, чтобы принять окончательное решение о смене бикон-слота. Таким образом, если в течение U суперкадров подряд устройство получает сигналы о бикон-коллизии, оно, согласно текущей спецификации WiMedia, выбирает новый бикон-слот случайным образом в определяемом стандартом окне размером *R*, начиная от HOBS предыдущего шага. Аналогичным образом слот выбирается и при подключении нового устройства.

Самый старший слот, в котором находится бикон хотя бы одного устройства, будем обозначать $HOBS$ (англ.: High Occupied Beacon Slot – старший занятый бикон-слот). По аналогии, самый старший слот, в котором находится ровно 1 устройство, будем обозначать

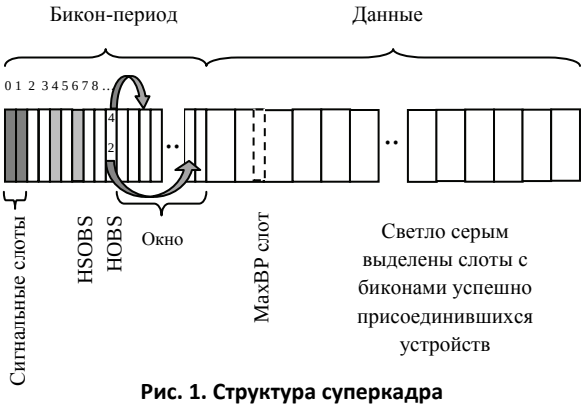


Рис. 1. Структура суперкадра

$HSOBS$ (англ.: High Successfully Occupied Beacon Slot – старший успешно занятый бикон-слот).

Алгоритм выбора слота приводит к тому, что занятые слоты чередуются со свободными. В то же время эффективным размещением, будет такое, что биконы заполняют младшие слоты подряд без пропусков. Для реализации эффективного размещения применяется процедура сжатия. Суть ее состоит в том, что устройство, занимающее не коллизионный слот $HOBS$ на протяжении последних U суперкадров, может на $(U+1)$-м переместить свой бикон в младший свободный слот, если он был таковым (младшим свободным) также на протяжении последних U суперкадров.

Также возможна ситуация, когда $HOBS = HSOBS$, при этом существуют устройства, которые успешно не присоединились. Алгоритм, описанный в спецификации, приводит к тупикам (англ.: deadlock), поэтому рассмотрим более совершенный, предложенный в [1]. Его идея заключается в том, что устройство по окончанию U суперкадров совсем покидает сеть на W суперкадров (для одношаговых сетей достаточно $W \geq U+2$). За это время происходит сжатие, которое, в общем случае, может уменьшить $HOBS$.

Описание задачи

В работе рассмотрен сценарий, когда некое устройство А сформировало пикосеть, заняв первый бикон-слот в бикон-периоде суперкадра, и остальные k_0 устройств начинают одновременное присоединение к созданной сети. Состояние системы будем обозначать S, начальное состояние (а также значения всех его параметров) пометим индексом «$_0$»: S_0.

Нас будет интересовать: **распределение времени подключения всех k_0 устройств,** при заданной функции $R(M)$ – функции общего вида от величины M, где $M = MaxBP - HOBS$.

Целью исследования является поиск такой $R(M)$, что распределение имеет наилучший вид (обеспечивается высокая вероятность присоединения за малое время и при этом практически нулевая вероятность неприсоединения за время ~1-2 секунд). При проведении численных расчетов ограничимся 2 видами функций $R(M)$:

- $R(M) = \min\{D,\ M\}$, где D – некоторая константа. Эта функция используется в стандарте, причем в стандарте $D = 8$.
- $R(M) = ceil(\alpha M),\ \alpha < 1,\ ceil(x)$ – функция, возвращающая минимальное целое число, большее или равное x.

Успешными состояниями (**успехом**) будем считать ситуации, в которых все устройства присоединились. Остальные ситуации будем считать **неудачей**.

Для построения аналитической модели рассмотрен марковский процесс $J(S,t)$, где t – время, измеряемое в шагах. Процесс прекращается, когда он пришел в успешное состояние. Под шагом мы будем понимать промежуток времени, начинающийся с очередного разыгрывания слотов, и продолжающийся до следующего разыгрывания (в случае прихода в неудачное состояние) или до окончания процесса (в случае успеха). Заметим, что согласно этому определению сжатие является частью шага, и шаг имеет переменную длину, выраженную в секундах.

Обозначения

Введем некоторые обозначения и соглашения, общие для рассмотренных ниже моделей. Ψ - множество состояний, за которыми последует сжатие. ($k > 0,\ HOBS = MaxBP$). Обозначим $\varphi(S)$ – вероятность того, что из состояния S система перейдет в успешное состояние для данной задачи. $\pi(S \mid < условие >)$ – условная вероятность того, что из состояния S система за 1 шаг перейдет в неуспешное состояние, при этом сам переход (шаг) будет описываться <*условием*>. Набор параметров, входящих в <*условие*> специфичен для каждой модели и для конкретного типа шага. c_t –количество устройств, вошедших в коллизию на очередном шаге. $z_t = HOBS_t - HOBS_{t-1}$ – максимальный номер занятого (не обязательно успешно) на очередном шаге слота, отсчитываемый от предыдущего $HOBS$. Поскольку шаги имеют разную длину, и интересует распределение времени присоединения в суперкадрах (секундах) введем $T(S`,S)$ – продолжительность шага из состояния $S`$ в состояние S в суперкадрах. $p(\tau)$ – распределение вероятности присоединения за время τ. $q(\tau) = 1 - p(\tau)$ – вероятность неприсоединения за время τ.

Оптимистическая модель (M, k)

Если система находится в каком-либо состоянии S, таком что $HOBS < MaxBP$, то для того чтобы определить состояние, в которое придет система на следующем шаге, не надо знать расположение биконов в занятой области, а достаточно только размера свободной области M и количества устройств, которые не подсоединились к сети k. Таким образом состояние S описывается 2 компонентами: M и k. Определить состояние системы после сжатия данная модель не позволяет, поэтому вводится предположение, что после того, как произойдет сжатие, все не подключившиеся на тот момент устройства подключатся без коллизий. Пусть система находится в состоянии $S`=(M`,k`)$ (все параметры, относящиеся к

этому состоянию, будем помечать «штрихом»). Тогда возможны следующие варианты:

1) С вероятностью $\varphi(S`)$ система перейдет в удачное состояние (длительность шага 1 суперкадр)

2) С вероятностью $\pi(S`|\, z, c)$, система перейдет в состояние $S$, такое что, $HOBS = HOBS`+z$, при этом с устройств попадут в коллизии (длительность шага $U+1$ суперкадр). Как следствие: $M = M`\text{-}z;\ k = c \le k`$.

3) С вероятностью $\pi(S`|\, z = M`, c)$, будет занят последний слот и состояние будет неудачным, при этом с устройств попадут в коллизии. Таким образом, длительность шага: $U+1+W$. Согласно предположению модели после сжатия система перейдет в состояние, из которого она перейдет в успешное с вероятностью. $\varphi(S)=1$.

Конкретный вид функций $\varphi(S)$ и $\pi(S\,|< условие >)$ будет приведен далее.

Пусть τ_w – минимальный возможный промежуток времени от начала процесса, такой что вероятность выполнения условий $HOBS(\tau_w) = MaxBP$ и $k(\tau_w) > 0$ отлична от нуля. Поскольку за время $\tau \le \tau_w$ сжатия не происходят, то модель (M, k) дает абсолютно точно распределение вероятности присоединения для времен $\tau \le \tau_w$, а затем завышает (по сравнению с симуляцией) вероятность успеха, что для расчетов сетей критичных к возможным длительным задержкам является недопустимым.

Точная модель (M, k, L)

Введем вектор $\vec{L}$ переменной длины, описывающий, сколько слотов освободятся в результате сжатий. Состояние в этой модели имеет вид : $S = (M; k; \vec{L})$

На рис. 2 видно, что после первого сжатия освобится 2+3 слота, а после второго – еще 1. Между сжатиями ни одно устройство успешно не присоединилось. После второго сжатия бикон-период заполнен эффективно, поэтому если и после второго сжатия будет занят $HOBS$ и ни одно устройство не присоединится, то третьего сжатия не последует.

Компоненты вектора $\vec{L}$ имеют следующие значения: l_0 - количество слотов от $HSOBS$ до $HOBS$, l_j - количество слотов, которое освободится за счет перехода бикона в младший слот во время j-того сжатия. Таким образом, $\sum_{j=0}^{i} l_j$ – суммарное количество слотов, которые освободятся в результате i сжатий без успешных присоединений. Для состояния системы, приведенного на рис. 2: $l_0 = 2,\ l_1 = 3,\ l_2 = 1.$ l_0, а также все ненулевые координаты вектора будем называть ведущими.

Количество ведущих координат вектора ограничено сверху количеством подключившихся устройств $(k - k_0 + 1)$. При подсоединении 1 или нескольких устройств компоненты вектора пересчитываются, и количество ведущих координат увеличивается либо не изменяется. Фактически в начала вектора добавляются новые координаты, а старые пересчитываются, причем на значение i-той компоненты влияние оказывают младшие. При совершении сжатия, количество ведущих координат вектора уменьшается на единицу (происходит «выталкивание» младшей координаты, остальные пересчитываются), при этом l_0 становится равным 0.

Для того чтобы пересчитать вектор $\vec{L}$, необходимо знать как именно распределены биконы успешно присоединившихся устройств на отрезке $[HOBS_{t\text{-}1}+1,\ HOBS_t]$. Для этого пронумеруем слоты с успешно присоединенными устройствами, на этом интервале, в «обратном порядке» (то есть первым слотом будет $HSOBS_t$), и введем вектор $\vec{D}$, координатам которого придадим следующий смысл:

$d_0 = HOBS_t - $«*номер слота первого устройства*»

$= HOBS_t - HSOBS_t$

остальные

$d_i = $«*номер слота i-того устройства*»$ - $«*номер слота (i+1)-того устройства*»

Количество компонент вектора $\vec{D}$ ограничено сверху количеством успешно занятых слотов на рассматриваемом промежутке.

Модель $(M; k; \vec{L})$ точно описывает реальный процесс присоединения, но неприменима для практических вычислений. Количество элементов в пространстве $(M; k; \vec{L})$ чрезвычайно велико. В то же время координата l_2 учитывается лишь в случаях, когда после первого сжатия ни одно из устройств не присоединилось, вероятность чего предельно мала. Эта особенность будет использовано в консервативной модели

Рис. 2. Сжатие бикон-периода

Консервативная модель (M, k, l_0, l_1)

Ограничим вектор $\vec{L}$ двумя первыми компонентами. Эта упрощенная модель дает абсолютно точный результат, когда происходит не более 2 сжатий подряд без успешных присоединений устройств между ними. И лишь в случае, если после первого сжатия ни одно устройство не смогло присоединиться, и в слоте $MaxBP$ снова произошла коллизия, мы должны каким-то образом определить недостающую информацию, чтобы определить состояние после второго сжатия.

Действительно, после пусть после первого сжатия $k(t+1)$ устройств выбирают слоты среди $M(t+1) = l_0(t) + l_1(t)$ слотов ($l_0(t+1) = 0$, $l_1(t+1)$ - неизвестно). Если все $k(t+1)$ устройств попали в коллизии, слот $MaxBP$ оказался опять занят, и произошло сжатие, то тогда $M(t+2) = l_0(t+1) + l_1(t+1)$. Мы не знаем значение $l_1(t+1)$, но мы можем его оценить следующим образом:

Если все биконы расположены «плотно» в младших слотах, (наличие свободных слотов младше $HSOBS$ не трудно определить, зная количество подсоединившихся устройств и $HSOBS$), то, следовательно, $l_1(t+1) = 0$.

Если есть только 1 свободный слот левее $HSOBS$, то $l_1(t+1) = 1$.

Наконец, в остальных случаях будем считать: $l_1(t+1) = 1$, рассматривая наихудший вариант (на самом деле $l_1(t+1) \geq 1$). Тем самым мы уменьшаем вероятность присоединения устройства (и это предположение дает название этой модели – консервативная, в противовес оптимистичной модели, описанной ранее).

Аналогично, ограничиваем младшими координатами вектор $\vec{D}$.

Переходы между состояниями в консервативной модели.

Пусть система находится в состоянии $S` = (M`, k`, l`_0, l`_1)$ в момент времени $t`$(все параметры, относящиеся к этому состоянию, будем помечать «штрихом»). Нас интересует, в какое состояние $S = (M, k, l_0, l_1)$ и с какой вероятностью может перейти система за 1 шаг. Для простоты рассуждений предлагается разделить возможные переходы на несколько групп.

Рассмотрим переходы между состояниями в отсутствии сжатия

Пусть рассматриваемый шаг не содержит операцию сжатия. Тогда его продолжительность $T(S`, S) = U + 1$.

Минимальное возможное значение M для этого случая обозначим буквой μ, $\mu = \max(1, M` - R(M`))$.

1. Рассмотрим случай, когда все устройства попали в коллизию ($c = k` \geq 2$). В новом состоянии $M \in [\mu, M` - 1]$. Поскольку устройства не подсоединись, то $k = k`$, $l_1 = l_1`$ и $l_0 = l_0` + \Delta M$. Здесь и далее мы будем полагать $\Delta M = M - M`$ (и вообще для любой величины X: $\Delta X = X - X`$). Вероятность данного перехода вероятность – $\pi(S` \mid z = -\Delta M)$.

2. Пусть только одно устройство поместило свой бикон в слот без коллизии. (Это возможно при $c = k` - 1$, $k` \geq 3$). В этом случае новое значение $M \in [\mu, M` - 2]$. Оставшееся количество не присоединенных устройств: $k = c = k` - 1$. $l_0 \in [0, -\Delta M - 1]$.

Если в состоянии $S`$ есть хотя бы 1 пустой слот (это утверждение эквивалентно тому, что $M(0) - M(t`) > k(0) - k(t`)$), то $l_1 = -\Delta M - \Delta l_0$; в противном случае $l_1 = -\Delta M - \Delta l_0 - 1$.

Вероятность такого перехода: $\pi(S` \mid z = -\Delta M, c = k, d_0 = l_0, d_1 = -\Delta M - l_0 - 1)$

3. Наконец, рассмотрим случай, когда $2 \leq c < k` - 1$. Поскольку 2 устройства должны успешно подсоединиться и еще как минимум в 1 слоте будет коллизия $M \in [\mu, M` - 3]$. Количество биконов в коллизиях с может быть любым из диапазона $[\max(2, k` + \Delta M + 1), k` - 2]$.

$d_0 \in [0, -\Delta M + \Delta k]$, $l_0 = d_0$. В случае, если не существует пустых слотов левее слота с номером $z - d_0 - 1 - d_1 - 1$, в том числе и среди слотов левее предыдущего $HOBS_{t-1}$, новое значение $l_1 = d_1$, где $d_1 \in [0, -\Delta M + \Delta k - l_0]$. Если же есть хотя бы 1 пустой слот, то $l_1 = d_1 + 1$. Вероятность этого перехода: $\pi(S` \mid z = -\Delta M, c = k, d_0, d_1)$

Рассмотрим переходы между состояниями при наличии сжатия.

Если сжатие произошло, будем (для удобства) рассматривать 3 состояния: $S`$ – начальное состояние, промежуточное состояние $S^* \in \Psi$ (не принадлежит процессу и вводится во вспомогательных целях) и S – состояние после сжатия. Таким образом, шаг имеет вид:

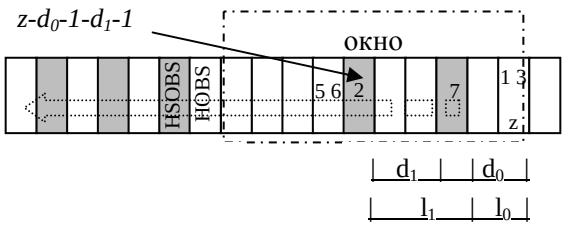


Рис. 3. Переход при c>2

$S` \to S^* \to S$. Полная длительность шага: $U+W+1$.

Рассмотрим 4 подгруппы переходов:

4. Если ни одно устройство не присоединилось успешно $c = k`$(необходимо $k` \geq 2$), то $l_0 = 0$ и $M = M` - l_0` - l_1`$, $l_1$ – неизвестно. В начале части 5 было показано, как в этом случае определять значение $l_1$ ($l_1$ принимает значение 0 или 1). Вероятность перехода $S` \to S$: $\pi(S` \mid z = M`)$.

5. При $c = k`-1$ получаем, параметр промежуточного состояния $l_0^* = d_0 \in [0, M`-1]$. Новое значение $M = M`+d_0$, если $d_1>0$ (тогда при сжатии бикон из слота $HSOBS$ сможет перейти в младший слот); иначе $M = M`+ d_0 - 1$. Если $l_1` > 0$, то $l_1 = l_1`-1$, иначе $l_1 = 0$. Вероятность перехода $\pi(S` \mid z = M, c = k`-1, d_0, d_1 = M - d_0 - 1)$.

Пусть $2 \le c < k`-1$, тогда $k \in \left[\max(2, k`-M`+1),\ k`-2\right]$, $d_0 \in \left[0,\ M`-k`+c\right]$, $d_1 \in \left[0, M`-d_0-k`+k\right]$.

6. Выделим случай $k - c = 2$. Его вероятность - $\pi(S` \mid z = M`, c = k`-2, d_0,\ d_1)$ $M = d_1 + d_0 + \omega$, где $\omega = 1$, когда существуют свободные слоты левее второго слота с биконом успешно присоединенного устройства $(M(0) - k(0) + k(t) - d_1^* - d_0^* > 0)$, иначе 0. На рисунке, этот слот помечен стрелкой. Будем считать зоной I слоты от начала ВР до третьего с конца ВР слота с биконом успешно присоединенного устройства (при $k - c = 2$, это – $HSOBS$), а зоной II – слоты между вторым и третьим слотом с биконами успешно присоединенных устройств. Зона II состоит из пустых слотов (либо коллизионных, которых мы будем считать здесь пустыми). Если в I зоне нет свободных слотов и «*ширина зоны II*» положительна, то после сжатия она уменьшится на единицу. $d_1 = $«*ширина зоны II после сжатия*»$ + \omega$, где ω принимает значение 0, если зона I после сжатия не содержит свободных слотов, иначе 1.

7. Аналогичные рассуждения справедливы и при $k - c > 2$: $d_2 \in [0, M`-d_0-d_1+\Delta k]$. Вероятность перехода: $\pi\left(S` \mid z = M`, c = k, d_0, d_1, d_2\right)$. Новое значение M вычисляется аналогично: если в первой зоне нет свободных слотов и «*ширина зоны II*» положительна, то после сжатия она уменьшится на единицу; $l_1 = $«*ширина зоны II после сжатия*»$ + \omega$, где ω принимает значение 0, если зона I после сжатия не содержит свободных слотов, иначе 1.

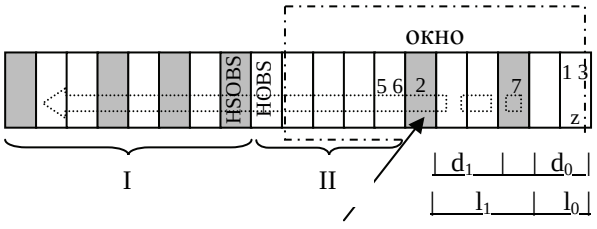


Рис. 4. Переход при c=2 и последующем сжатии

Основные математические формулы

(За ограничением размера статьи ниже приводятся формулы без доказательств.)

Пусть $V(v,c)$ - количество вариантов разместить c биконов в v слотах, так чтобы в каждом слоте было по крайней мере 2 устройства. При $v > \frac{c}{2}$, $V(v,c) = 0$, в противном случае

$$V(v,c) = v^c - \sum_{y=1}^{v-1} C_y^v V(y,c) - \sum_{u=1}^{v-1} C_u^v A_u^c \sum_{y=1}^{v-u} C_y^{v-u} V(y, c-u)$$

Обозначим $\mathrm{U}(\mathrm{v},\mathrm{c})$ - количество вариантов размещения с биконов в v слотах так, чтобы все биконы попали в коллизии. (В каждом слоте либо 0, либо ≥2 биконов.) Соответственно $\mathrm{U}`(\mathrm{v},\mathrm{c})$ - количество вариантов размещения с биконов в v слотах так, чтобы все биконы попали в коллизии и при этом был занят последний слот.

$$U'(k,z) = 1 + \sum_{c=2}^{k-2} C_c^k U(z-1, c)$$

Обозначим $Y(z,k)|_g$ функцию, которая тождественно равна $\mathrm{U}(\mathrm{z},\mathrm{k})$ при $g = 0$, иначе $\mathrm{U}`(\mathrm{z},\mathrm{k})$.

Пусть система находится в состоянии S с параметрами M и k. Тогда вероятность успеха

$$\varphi(S) = \frac{A_k^R}{R^k}$$

Здесь и далее полагается $R = R(M)$.

Вероятность того, что ни одно устройство не подсоединится $\pi(S \mid z) = \dfrac{U'(k,z)}{R(M)^k}$

Для оптимистической модели в случае $k\text{-}c \le z$ при $c=0$ и $k\text{-}c<z$ при $c>0$

$$\pi(S,z,c) = \frac{C_c^k (A_{k-c}^{z-1} U'(z-k+c, c) + (k-c) A_{k-c-1}^{z-1} U(z-k+c, c))}{R(M)^k}$$

Если без коллизии подсоединилось только одно устройство, то есть при c = k-1

$$\pi_A(S \mid z, k-1, d_0) = \frac{k * Y(z-1, k-1)|_{d_0}}{R^k}$$

При c<k-1

$$\pi_A\left(S|z,c,d_0,d_1\right)=\frac{C_c^k A_2^{k-c} A_{k-c-2}^{z-d_0-d_1-2} Y\left(z-k+c,c\right)|_{d_0}}{R\left(M\right)^k}$$

При k-c>2

$$\pi_A\left(S\mid z,c,d_0,d_1,d_2\right)=$$

$$\frac{C_c^k A_3^{k-c} A_{k-c-3}^{z-d_0-d_1-d_2-3} Y\left(z-k+c,c\right)|_{d_0}}{R\left(M\right)^k}$$

Нахождения распределения

Использование единицы времени «шаг» удобно при описании процесса и переходов из одного состояние в другое, но оказывается совершенно не удобным при практических вычислениях. Поэтому прибегнем к следующему способу.

Вначале ($\tau_0=0$) состояние системы – S_0, Пусть точке $\tau`$ временной оси соответствует $n`=n(\tau`)$ состояний: $S`_{\tau`,1},\ldots, S`_{\tau`,n`}$; их вероятности $r`_{\tau`,1},\ldots, r`_{\tau`,n`}$ соответственно. Вероятность успеха при присоединении устройств в момент времени $\tau=\tau`+1$: $\sum_{i=1}^{n`}\varphi(S`_{\tau`,i})r_{\tau`,i}$. Таким образом, распределение принимает вид:

$$p(\tau)=p(\tau-1)+\sum_{i=1}^{n(\tau-1)}\varphi(S`_{\tau-1,i})r_{\tau-1,i}=\sum_{\tau`=0}^{\tau-1}\sum_{i=1}^{n(\tau`)}\varphi(S`_{\tau`,i})r_{\tau`,i}$$

.

Численные результаты

Начальное состояние системы определяется параметрами M_0, k_0, а также l_0, l_1 в консервативной модели. Учитывая спецификацию, $M_0=94-1=93$, так как из 94 бикон-слотов первый занят устройством, создавшим сеть. k_0 принимает значение 4, 8, 12, 18, что соответствует 5, 9, 13, 19 устройствам в сети. $l_0=l_1=0$.

При вычислениях использовалась в основном

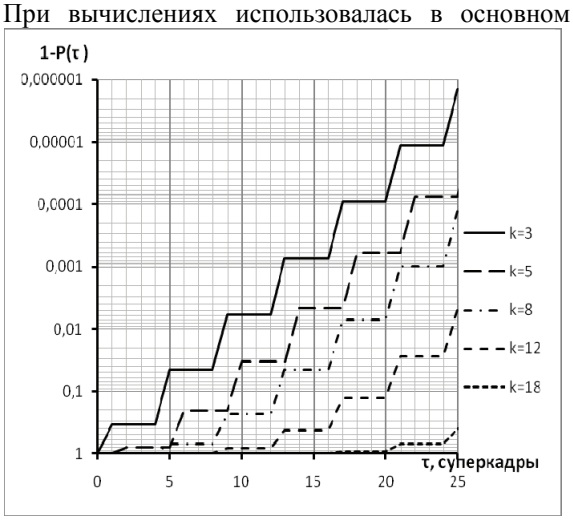


Рис. 5. R(M)=min(M,8)

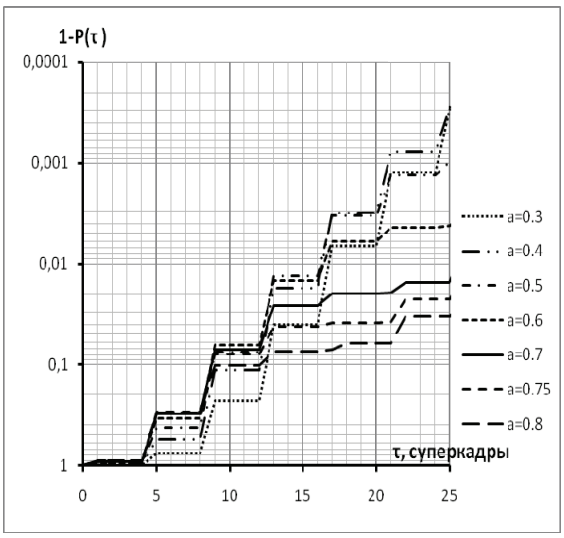


Рис. 6. Вероятность неприсоединения при R(M)=ceil(αM)

консервативная модель. Однако для времен $\tau<\tau_w$ результаты обеих моделей совпадают, поэтому, при малых α и D использовалась оптимистическая модель (как более простая и быстрая).

Выше (рис. 5) приведены результаты при $R(M)=8$ (по стандарту) для разных количеств устройств. Результаты обеих моделей совпадают, поскольку слот $MaxBP$ не достигается за время 25 суперкадров.

Из графиков видно, что предложенный в спецификации размер окна приводит к низкой вероятности подключения всех устройств за короткое время.

Исследование показало что функция $R(M)$ вида $R(M)=ceil(\alpha M)$, при $\alpha=0.6$ обеспечивает наилучшее распределение для различного количества устройств, давая почти максимальную вероятность присоединения для всех $\tau \le 25$ (на рис. 6 приведен график для k_0=18).

Выводы

В данной работе исследуется эффективность механизмов синхронизации в беспроводных персональных сетях с распределенным (децентрализованным) управлением на MAC-уровне.

Алгоритмы, приведенные в спецификации WiMedia, могут быть изменены таким образом, что работа сети станет гораздо более эффективной. В данной работе оценивается эффективность алгоритмов присоединения к сети, не приводящих к dead-lock`ам. Разработаны аналитические модели, которые позволяют выбрать наилучшие параметры предложенных алгоритмов. Сравнение результатов имитационного и аналитического моделирования позволяет судить о высокой точности последнего.

Библиография

[1] ECMA-368. High Rate Ultra Wideband PHY and MAC Standard. 2nd Edition - December 2007

[2] Vishnevsky V.M., Lyakhov A.I., Safonov A.A, Mo S.S., Gelman A.D., "Study of Beaconing in Multi-Hop Wireless PAN with Distributed Control", *IEEE Transactions on Mobile Computing, Volume 7, issue 1, Jan 2008, pp.113-126*

[3] Alexander Safonov, Andrey Lyakhov, Evgeny Khorov. ”Channel Switch Time Distribution in ECMA-368 Networks”, *Proc. of IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC'08)*, France, 2008

[4] В.М. Вишневский, А.И. Ляхов, С.Л. Портной, И.Л. Шахнович. Широкополосные беспроводные сети передачи информации. *М.: Техносфера, 2005. 592 с*

Эффективность метрик стоимости соединений стандарта ZigBee

Баскаков С.С.
Московский государственный технический университет имени Н.Э. Баумана
sergey.baskakov@mail.ru

Аннотация

В работе с помощью имитационного моделирования выполнена приближенная к реальным условиям эксплуатации оценка эффективности протокола маршрутизации стандарта ZigBee при типе трафика «многие-к-одному». Показано, что учет качества связи обеспечивает более эффективное использование пропускной способности канала связи и более длительный срок эксплуатации сети, но предложенные в стандарте варианты метрик стоимости соединений не отражают в полной мере особенности маломощных радиоканалов.

1. Введение

В последние годы технологии беспроводных сенсорных сетей находят все больше практических применений, связанных с распределенным сбором информации по радиочастотному каналу. Среди стандартизованных решений наибольшее распространение получил стандарт ZigBee, однако эффективность таких сетей в реальных условиях эксплуатации недостаточно исследована.

В настоящей работе рассматривается эффективность механизма маршрутизации ZigBee, поскольку маршрутизация пакетов является одной из наиболее важных составляющих любого стека сетевых протоколов, оказывая значительное влияние на показатели качества обслуживания сети. Рассматриваемый тип трафика «многие-к-одному», при котором все устройства сети передают пакеты в одну точку сбора (базовая станция, шлюз и т.п.), характерен для многих задач распределенного сбора информации (например, системы мониторинга). Поэтому оценка эффективности и совершенствование механизмов маршрутизации в данном режиме имеет важное практическое значение.

Исследование выполнено с помощью среды имитационного моделирования OMNeT++ версии 3.2 и пакета моделей Mobility Framework версии 2.0 при следующем сценарии работы сети. Множество узлов и одна точка сбора (PAN-координатор) случайно распределены на двумерной плоскости, сеть имеет многоячейковую топологию. При этом все узлы являются маршрутизаторами и работают на автономном источнике питания с одинаковой начальной емкостью, а точка сбора имеет стационарное питание. После первоначальной конфигурации сети (обнаружение соседей и т п.) все узлы начинают периодическую передачу пакетов данных в точку сбора. Предполагаем, что в процессе функционирования сети положение узлов и параметры среды окружения не изменяются.

Под стандартом ZigBee подразумевается версия 2006 года [1] в сочетании с наиболее полным и функциональным дополнением «ZigBee PRO Feature Set», включающем в себя поддержку трафика типа «многие-к-одному» для простого поиска маршрутов с минимальной стоимостью и учет асимметрии качества связи между узлами.

2. Модель канала связи

В большинстве работ протоколы маршрутизации рассматриваются с использованием бинарной модели беспроводного канала связи, согласно которой пакет успешно принимается только при условии, что расстояние между передатчиком и приемником меньше некоторого фиксированного радиуса радиосвязи. Однако многочисленные исследования характеристик маломощных беспроводных каналов связи показали, что в реальных системах наблюдается значительная вариация и асимметрия качества связи между узлами. Поэтому в настоящей работе используется модель физического уровня стандарта IEEE 802.15.4 в условиях крупномасштабного замирания и при наличии случайных отклонений параметров приемопередатчиков от номинальных значений, которая отражает особенности маломощных радиоканалов.

Значения выходной мощности P_t и уровня шума P_n реальных приемопередатчиков интегрального исполнения имеют случайные отклонения от номинальных значений и подчиняются нормальному закону распределения вида

$$\begin{bmatrix} P_t \\ P_n \end{bmatrix} \sim N\left(\begin{bmatrix} \overline{P}_t \\ \overline{P}_n \end{bmatrix}, \Sigma_{hw} \right), \quad (1)$$

где

$\overline{P}_t$ - среднее (номинальное) значение выходной мощности передатчика, дБм;

$\overline{P}_n$ - среднее (номинальное) значение уровня шума приемника, дБм;

Σ_{hw} - ковариационная матрица вида

$$\Sigma_{hw} = \begin{bmatrix} \sigma_t^2 & \mathrm{cov}(P_t, P_n) \\ \mathrm{cov}(P_n, P_t) & \sigma_n^2 \end{bmatrix}.$$

Для учета влияния затенения используем модель логарифмически-нормального распределения потерь в тракте:

$$PL(d) = PL(d_0) + 10\alpha \lg\left(\frac{d}{d_0}\right) + X_{ch}, \text{ дБ}, \quad (2)$$

где

d - расстояние между передатчиком и приемником, м;

d_0 - эталонное расстояние, м;

α - показатель степени потерь в тракте;

X_{ch} - случайная гауссова переменная с нулевым средним и дисперсией σ_{ch}^2, дБ.

Тогда при расстоянии d между передатчиком и приемником отношение сигнал-шум (ОСШ) на входе приемника равно

$$\gamma(d) = P_t - PL(d) - P_n, \text{ дБ}. \quad (3)$$

Для расчета вероятности битовой ошибки в зависимости от значения ОСШ используем выражение из спецификации стандарта IEEE 802.15.4 [2]:

$$\beta(\gamma) = \frac{1}{30}\sum_{k=2}^{16}(-1)^k \binom{16}{k} \exp\left[20\left(\frac{1}{k}-1\right)10^{\gamma/10}\right]. \quad (4)$$

Тогда вероятность успешного приема пакета длиной L байт равна

$$\Psi(\gamma, L) = \left[1-\beta(\gamma)\right]^{8L}. \quad (5)$$

Таким образом, выражения (1–5) представляют собой модель физического уровня стандарта IEEE 802.15.4, которая позволяет рассчитать вероятность успешного приема пакета в условиях медленного замирания и при наличии отклонений параметров приемопередатчиков от номинальных значений. На рис. 1 приведен пример зависимости вероятности успешного приема пакета для двух моделей канала связи.

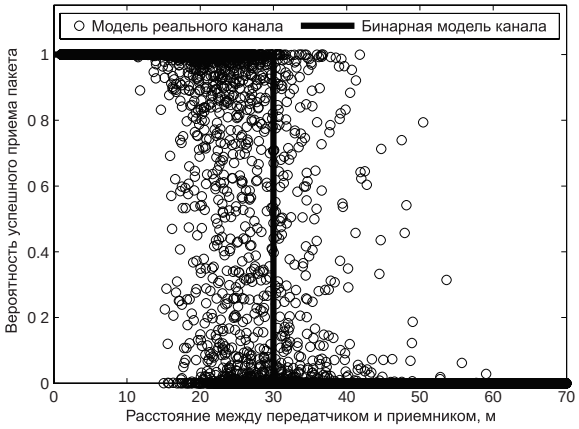


Рис. 1. Модели беспроводного канала связи

3. Модель энергопотребления узла

В целом среднее потребление узла зависит от потребляемой мощности в различных режимах работы и проведенного в них времени. Основными режимами узла являются передача и прием пакетов, прослушивание канала, обработка данных (например, опрос датчиков, фильтрация и т.п.) и дежурный режим.

Как правило, беспроводной сенсорный узел большую часть времени находится в «спящем» режиме, поэтому среднее энергопотребление узла в основном определяется затратами на передачу/прием пакетов и прослушивание канала. В данной работе объектом исследования являются только механизмы маршрутизации, поэтому особенности реализации MAC-уровня не учитываются и используется упрощенная модель энергопотребления узла, согласно которой среднее потребление узла складывается только из затрат энергии непосредственно на передачу и прием пакетов данных и подтверждения, а широковещательные пакеты не учитываются.

В стандарте IEEE 802.15.4 наиболее коротким является пакет подтверждения суммарной длины $L_{ack} = 11$ байт (с учетом преамбулы и служебных полей). Поэтому энергию, необходимую для его передачи/приема, примем в качестве ее единицы измерения, т.е. $e_{ack} = 1$. Тогда затраты энергии на передачу/прием пакетов данных длиной L_{data} (без шифрования и с 16-разрядной адресацией) равны

$$e_{data}(n) = \frac{L_{data}(n)}{L_{ack}} = \frac{17+n}{11},$$

где n - размер поля данных, байт.

При этом считаем, что токи потребления в режиме приема и передачи равны, что верно для большинства приемопередатчиков стандарта

IEEE 802.15.4 при выходной мощности порядка 1 мВт.

Таким образом, остаточный запас энергии узла в момент времени t равен

$$E(t) = E(0) - e_{data}(n)\left[T_{data}^{tx}(t) + T_{data}^{rx}(t) + \right.$$
$$\left. + T_{edata}^{rx}(t)\right] - e_{ack}\left[T_{ack}^{tx}(t) + T_{ack}^{rx}(t) + T_{eack}^{rx}(t)\right],$$

где

$E(0)$ - начальный запас энергии;

$T_{data}^{tx}(t)$, $T_{data}^{rx}(t)$, $T_{edata}^{rx}(t)$ - количество переданных, принятых без ошибок и принятых с ошибками пакетов данных на момент времени t;

$T_{ack}^{tx}(t)$, $T_{ack}^{rx}(t)$, $T_{eack}^{rx}(t)$ - количество переданных, принятых без ошибок и принятых с ошибками пакетов подтверждения на момент времени t.

При $E(t) \le 0$ узел прекращает функционировать из-за истощения источника энергии.

4. Метрики стоимости соединения

В ZigBee при типе трафика «многие-к-одному» выполняется построение дерева, корнем которого является узел-назначение (точка сбора), а вершинами – остальные узлы сети. Задача протокола маршрутизации заключается в обнаружении маршрута между узлом-отправителем и точкой сбора с минимальной стоимостью. Стоимость маршрута равна сумме стоимостей соединений между узлами, через которые этот путь проходит.

В простейшем случае соединению присваивается постоянная стоимость, т.е.

$C_{const} = 7$.

Но беспроводные соединения могут значительно отличаться надежностью, что никак не учитывается метрикой C_{const}, поэтому при ее использовании могут быть выбраны пути с низким качеством связи, вызывая снижение эффективности протокола маршрутизации. Поэтому в ZigBee предусмотрена метрика с учетом надежности соединения между соседними узлами:

$$C_{sym} = \min\left(7, \frac{1}{p_{RX}^4}\right), \quad (6)$$

где p_{RX} - оценка вероятности успешного приема пакета от соседнего узла.

Однако метрика C_{sym} не учитывает возможную асимметрию качества связи, поэтому в [3] предложена модифицированная метрика

$$C_{asym} = \max\left[\min\left(7, \frac{1}{p_{RX}^4}\right), \min\left(7, \frac{1}{p_{TX}^4}\right)\right],$$

где p_{TX} - оценка вероятности успешной передачи пакета соседнему узлу.

Таким образом, в настоящей работе фактически исследуется эффективность принятых в ZigBee метрик стоимости соединения в условиях реального канала связи.

5. Результаты моделирования

Для сравнения эффективности метрик выбраны следующие критерии оценки:

- эффективность по трафику η_T – отношение объема полезного трафика к общему объему переданного трафика в сети;
- эффективность по энергии η_E – среднее число принятых точкой сбора пакетов данных от каждого узла до первого выхода из строя какого-либо узла по причине истощения источника энергии.

Эффективность по трафику рассчитывается по формуле

$$\eta_T = \frac{L_{data} T}{L_{data} T_{data}^{\Sigma} + L_{ack} T_{ack}^{\Sigma}},$$

где

L_{data} и L_{ack} - длина пакетов данных и подтверждения соответственно, байт;

T - общее количество пакетов данных, принятых точкой сбора;

T_{data}^{Σ} и T_{ack}^{Σ} - суммарное количество переданных в сети пакетов данных и подтверждения соответственно.

Чем больше значение η_T, тем более эффективно протокол маршрутизации использует пропускную способность канала связи. Для простоты полагаем, что все пакеты данных и подтверждения имеют фиксированные размеры L_{data} и L_{ack}, которые задаются в параметрах модели сети. Величина T отражает только количество «полезных» пакетов данных, т.е. без учета пакетов, которые повторно принимаются из-за потерь пакетов подтверждения. В T_{data}^{Σ} и T_{ack}^{Σ} входят «полезные» пакеты, пакеты ретрансляции, а также пакеты, переданные при повторных попытках доставки. Предполагаем, что широковещательные пакеты используются только для передачи служебной информации при конфигурации сети и ее объем пренебрежимо мал.

Эффективность по энергии равна

$$\eta_E = \frac{T}{N}$$

(где N - количество узлов в сети) и ее можно интерпретировать как срок службы узла до начала разрушения сети по причине разряда элементов питания. Многие приложения требуют длительного срока эксплуатации сети при автономных источниках питания узлов, поэтому энергетическая эффективность является одним из ключевых критериев оценки протоколов маршрутизации для систем данного класса.

В таблице 1 приведены значения параметров, использованных при имитационном моделировании.

Таблица 1. Параметры моделирования

Эталонное расстояние d_0	1 м
Затухание на эталонном расстоянии $PL(d_0)$	55 дБ
Показатель степени потерь в тракте α	3
Среднеквадратическое отклонение σ_{ch}	3 дБ
Средняя выходная мощность передатчика $\overline{P}_t$	0 дБм
Средний уровень шума приемника $\overline{P}_n$	-95,5 дБм
Дисперсия σ_t^2	3 дБ^2
Дисперсия σ_n^2	3 дБ^2
Ковариация $\mathrm{cov}(P_t, P_n)$	-2 дБ^2
Начальный запас энергии $E(0)$	100000

На рис. 2 и 3 представлены результаты моделирования для сети из 200 узлов при различной длине пакетов данных. Как и следовало ожидать, увеличение длины пакетов данных приводит к росту затрат энергии на их передачу и прием, поэтому сокращается количество пакетов, которые могут переданы до начала истощения источников энергии узлов, но при этом повышается эффективность использования пропускной способности канала.

При коротких пакетах данных (L_{data}=26 байт) получен вполне ожидаемый результат: наилучшие характеристики обеспечивает метрика C_{asym}, а далее следует C_{sym}. Однако, если качество исходящей связи значительно ниже качества входящей ($p_{TX} < p_{RX}$), предположение о симметричности соединений метрики C_{sym} может привести к худшему результату, чем при выборе маршрутов просто по количеству промежуточных узлов (метрика C_{const}). Этот эффект проявляется больше с увеличением длины пакетов данных.

Учитывающая асимметрию метрика дает несколько лучший результат, но она не позволяет в полной мере использовать информацию о качестве связи по следующей причине. Если хотя бы один из показателей качества связи (p_{TX} или p_{RX}) принимает значение меньше $7^{-1/4} \approx 0{,}62$, то эта метрика становится эквивалентной простой метрике постоянной стоимости C_{const}. В реальных каналах возможны значительные вариации надежности связи (см. рис. 1), поэтому неравенства $p_{TX} \geq 0{,}62$ и $p_{RX} \geq 0{,}62$ выполняются только на малом интервале расстояний между передатчиком и приемником, а на большей части области приема имеют место существенные колебания надежности. Поэтому метрика C_{asym} дает преимущества относительно C_{const} только для небольшой части беспроводных соединений, а в остальных случаях они эквивалентны.

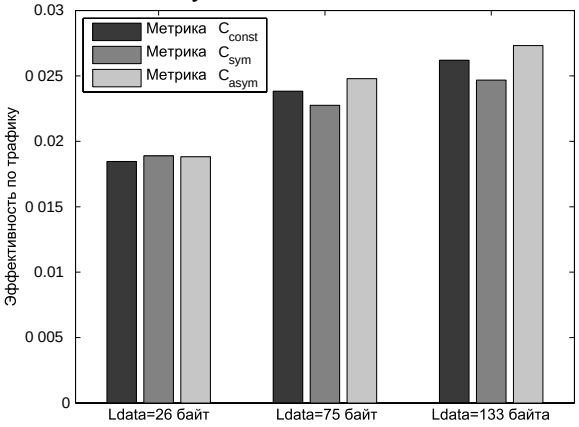


Рис. 2. Эффективность по трафику при переменной длине пакетов данных

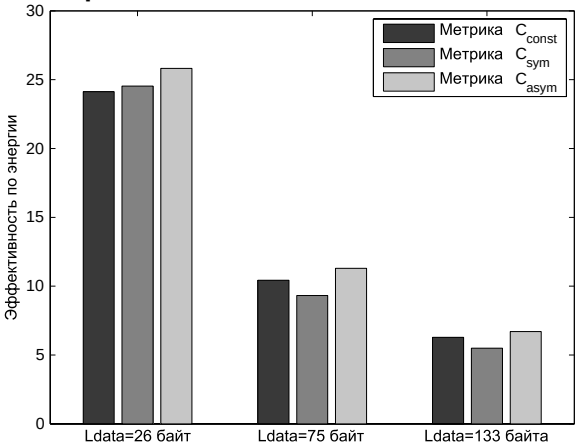


Рис. 3. Эффективность по энергии при переменной длине пакетов данных

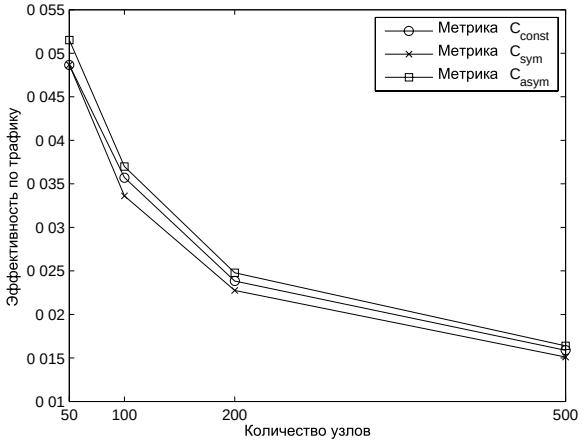


Рис. 4. Эффективность по трафику при переменном количестве узлов

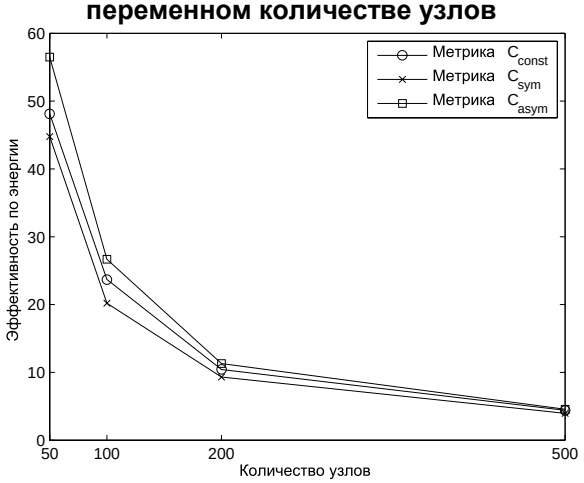


Рис. 5. Эффективность по энергии при переменном количестве узлов

На рис. 4 и 5 приведены результаты моделирования для сетей с переменным количеством узлов и фиксированной плотностью размещения (среднее количество соседних узлов) равной 10 узлам. При этом длина пакетов данных равна 75 байтам.

С увеличением количества узлов при постоянной плотности возрастает диаметр сети, поэтому пути становятся длиннее и требуется больше пакетов ретрансляции для доставки «полезных» пакетов данных. В результате снижается эффективность по трафику протокола маршрутизации. Время жизни сети также сокращается из-за увеличения сетевой нагрузки на узлы, расположенные вблизи точки сбора.

Видно, что при всех масштабах сети соотношение между метриками сохраняется: наилучшей является метрика C_{asym}, а наихудшей - C_{sym}. Тем не менее, вид метрики стоимости соединения не оказывает значительного влияния на эффективность протокола маршрутизации стандарта ZigBee. В среднем метрика C_{asym} обеспечивает время жизни узла только на 10% большее, чем простейшая метрика C_{const}, при этом метрика C_{sym} на 11% хуже C_{const}. По эффективности использования пропускной способности канала связи разница еще меньше – около 4% как в лучшую сторону (для C_{asym}), так и в худшую сторону (для C_{sym}) относительно метрики C_{const}.

6. Заключение

В работе представлены результаты исследования эффективности протокола маршрутизации стандарта ZigBee при типе трафика «многие-к-одному». При этом использовалась модель физического уровня стандарта IEEE 802.15.4 в условиях влияния затенения и случайного разброса параметров приемопередатчиков.

Результаты имитационного моделирования показали, что предложенные в стандарте метрики стоимости соединений не обеспечивают высокой эффективности в реальных условиях эксплуатации. Более того, метрика, основанная на предположении о симметричности качества связи, дает хуже результат, чем простейшая метрика постоянной стоимости.

Предложенная в [3] метрика имеет преимущества относительно стандартных метрик, но выигрыш незначителен.

Следовательно, необходимо разработать новые варианты метрик стоимости соединений, которые будут более полно отражать их параметры и обеспечивать более высокую эффективность протокола маршрутизации.

7. Список источников и литературы

[1] ZigBee specification. – ZigBee Alliance, 1 December 2006.

[2] IEEE Std 802.15.4–2006. IEEE standard for local and metropolitan area networks. Part 15.4: wireless medium access control (MAC) and physical layer (PHY) specifications for low rate wireless personal area networks (LR-WPANs). – IEEE, 8 September 2006.

[3] EmberZNet application developer's guide. Ember Corporation. 17 July 2007.

Система поллинга с опросом типа "звезда" и ограниченным обслуживанием[1]

Семенова О.В., Сергеев П.А.
Институт проблем передачи информации им. А.А. Харкевича РАН, Москва
olgasmnv@iitp.ru, peter.mailorama@gmail.com

Аннотация

В работе исследуется система поллинга с N очередями и опросом типа "звезда", то есть очереди обслуживаются в порядке 1, N, 2, N, ..., $N-1$, N. Обслуживание очередей – 1-ограниченное, за одно посещение очереди сервер обслуживает не более одной заявки. Предложен алгоритм расчета стационарных вероятностей состояний системы путем их полиномиального разложения. Получены основные характеристики модели. Результаты проиллюстрированы численными примерами.

1. Введение

Интерес к моделям систем поллинга в последние годы значительно возрос в связи с интенсивным развитием широкополосных беспроводных сетей передачи информации с централизованным механизмом управления [1]. Анализ моделей поллинга для исследования характеристик персональных и локальных беспроводных сетей проведен в [9]; для широкополосных беспроводных региональных сетей с централизованным управлением в [2, 10]. Классификацию моделей поллинга, методы и результаты их исследования можно найти в [3, 4, 7].

Системы поллинга, или системы упорядоченного опроса, являются разновидностью систем массового обслуживания с несколькими очередями и общим обслуживающим прибором (сервером) или несколькими приборами. В каждую очередь поступает свой поток заявок. Обслуживающий прибор по определенному правилу посещает очереди и обслуживает находящиеся в них заявки.

Правило, следуя которому сервер выбирает очередь для обслуживания, называется порядком обслуживания. Примерами такого правила могут служить циклический опрос очередей, когда сервер посещает очереди от первой до последней и вновь возвращается к первой очереди, или случайный порядок, при котором очередь на обслуживание выбирается случайно.

Очереди системы поллинга обслуживаются согласно заданной дисциплине обслуживания. Она характеризуется числом заявок, которое может обслужить сервер за одно посещение очереди. Наиболее распространенными дисциплинами обслуживания очередей являются исчерпывающая дисциплина, при которой сервер обслуживает заявки до тех пор, пока очередь не опустеет, шлюзовая дисциплина, при которой сервер обслуживает лишь те заявки, которые находились в очереди в момент подключения к ней сервера, и ограниченная дисциплина, при которой число заявок, которое может обслужить сервер, ограничено.

В данной работе рассматривается система поллинга, состоящая из N очередей, с порядком опроса типа "звезда". Такой порядок опроса представляет собой один из способов присвоения приоритета одной из очередей (когда в этом возникает необходимость). Системы поллинга с таким опросом исследовались в работе [8].

Мы предлагаем алгоритм расчета стационарных вероятностей состояний системы поллинга с 1-ограниченным обслуживанием очередей. За одно посещение очереди сервер обслуживает не более одной заявки. Алгоритм основан на вычислении вероятностей через их полиномиальное разложение. Алгоритмы такого рода представлены в работах [5, 6].

2. Модель

Рассматривается система поллинга с одним обслуживающим прибором (сервером) и N очередями. Очереди имеют неограниченное число мест для ожидания. Сервер посещает очереди в порядке 1, N, 2, N, ..., $N-1$, N, то есть очередь N обладает приоритетом по сравнению с другими очередями.

Обслуживание очередей – 1-ограниченное.

[1]Работа выполнена при поддержке гранта РФФИ № 08-07-90102-Мол_а

Поток заявок в i-ю очередь – стационарный пуассоновский с параметром λ_i. Времена обслуживания имеют экспоненциальное распределение с параметром μ_i. Перед обслуживанием очереди сервер тратит время на подключение, экспоненциально распределенное с параметром s_i, $i = \overline{1,N}$.

Для рассматриваемой системы условие существования стационарного режима имеет следующий вид:

$$\rho + \max_{i=\overline{1,N}} \frac{\lambda_i}{\delta_i} S < 1,$$

где $S = (N-1)\frac{1}{s_N} + \sum_{i=1}^{N-1} \frac{1}{s_i}$ – совокупное среднее время переключения сервера в цикле,

$$\delta_i = \begin{cases} 1, & i \neq N, \\ N-1, & i = N, \end{cases}$$

– среднее максимальное число заявок, которое может обслужить сервер в i-й очереди за один цикл, $\rho = \sum_{i=1}^{N} \rho_i$, $\rho_i = \frac{\lambda_i}{\mu_i}$, $i = \overline{1,N}$.

3. Стационарное распределение вероятностей состояний системы

Представим порядок опроса очередей 1, N, 2, N, ..., $N-1$, N как таблицу поллинга длины $M = 2(N-1)$. Обозначим через $T(k)$ – номер очереди, соответствующей позиции h в таблице поллинга, $l = \overline{1,M}$, $h = \overline{1,N}$; r_i число заявок в очереди i в произвольный момент времени, $i = \overline{1,N}$. Пусть также $\mathbf{r} = (r_1, \ldots, r_N)$.

Предположим, что система функционирует в стационарном режиме. Обозначим через $p(\mathbf{r}, h, m)$ – стационарные вероятности того, что в произвольный момент времени число заявок в системе представлено вектором $\mathbf{r}$, сервер обслуживает ($m = 1$) очередь либо подключается ($m = 0$) к очереди, соответствующей позиции h в таблице поллинга, $h = \overline{1,M}$.

Стационарные вероятности удовлетворяют следующей системе уравнений:

$$\begin{aligned} p(\mathbf{r},h,0)\left(\lambda + s_{T(h)}\right) = \sum_{j=1}^{N} \lambda_j p(\mathbf{r}-\mathbf{e}_j,h,0) I_{\{r_j>0\}} + \\ +\mu_{T(h-1)} p(\mathbf{r}+\mathbf{e}_{T(h-1)},h-1,1) + \\ +s_{T(h-1)} p(\mathbf{r},h-1,0) I_{\{r_{T(h-1)}=0\}}, \\ r_j \geq 0, j = \overline{1,N}, h = \overline{1,M}, \end{aligned} \quad (1)$$

$$\begin{aligned} p(\mathbf{r},h,1)\left(\lambda + \mu_{T(h)}\right) = \sum_{j=1}^{N} \lambda_j p(\mathbf{r}-\mathbf{e}_j,h,0) I_{\{r_j>0\}} + \\ +\mu_{T(h)} p(\mathbf{r},h,0), r_j \geq 0, j \neq h, r_h > 0, \end{aligned} \quad (2)$$

где $\lambda = \sum_{j=1}^{N} \lambda_j$, $\mathbf{e}_j$ – вектор-строка, все элементы которого равны нулю, за исключением j-го, равного 1, $I_{\{A\}}$ – функция-индикатор события A, $I_{\{A\}} = 1$, если событие A произошло, и равно нулю в противном случае.

Далее изложим алгоритм вычисления стационарных вероятностей $p(\mathbf{r}, h, m)$, основанном на подходе, подробно изложенном в работах [5, 6] для систем поллинга с очередями типа $M/M/1$. Данный подход использует полиномиальное представление стационарных вероятностей состояний системы, коэффициенты которого являются функциями от величины загрузки системы ρ.

Обозначим через

$$\theta = \rho + \max_{i=\overline{1,N}} \frac{\lambda_i}{\delta_i} S \quad (3)$$

и представим интенсивности входных потоков заявок λ_i как

$$\lambda_i = a_i \theta, \quad i = \overline{1,N}. \quad (4)$$

Далее, для построения алгоритма вычисления стационарных вероятностей используем следующие разложения:

$$p(\mathbf{r},h,m) = \theta^{|\mathbf{r}|} \sum_{k=0}^{\infty} \theta^k b(k,\mathbf{r},h,m), \quad (5)$$

где $|\mathbf{r}| = \sum_{i=1}^{N} r_i$, $\mathbf{r}_i \geq 0$, $h = \overline{1,M}$, $m = \overline{0,1}$.

Заменяя λ_i в равенствах (1)–(2) на выражения (4), а затем приравнивая коэффициенты при соответствующих степенях θ, получим следующие равенства:

$$\begin{aligned} s_{T(h)} b(k,\mathbf{r},h,0) = \sum_{j=1}^{N} a_j \big[b(k,\mathbf{r}-\mathbf{e}_j,h,0) I_{\{r_j>0\}} - \\ -b(k-1,\mathbf{r},h,0) I_{\{k>0\}} \big] + \\ +\mu_{T(h-1)} b(k-1,\mathbf{r}+\mathbf{e}_{T(h-1)},h-1,1) + \\ +s_{T(h-1)} b(k,\mathbf{r},h-1,0) I_{\{r_{T(h-1)}=0\}}, \\ k \geq 0, r_j \geq 0, j = \overline{1,N}, h = \overline{1,M}, \end{aligned} \quad (6)$$

$$\begin{aligned} \mu_{T(h)} b(k,\mathbf{r},h,1) = \sum_{j=1}^{N} \lambda_j \big[b(k,\mathbf{r}-\mathbf{e}_j,h,0) I_{\{r_j>0\}} - \\ -b(k-1,\mathbf{r},h,1) I_{\{k>0\}} \big] + \mu_{T(h)} b(k,\mathbf{r},h,0), \\ r_j \geq 0, j \neq h, r_h > 0, \end{aligned} \quad (7)$$

Для дополнения рекуррентной схемы (6)–(7) используем условие нормировки

$$\sum_{r_1=0}^{\infty} \sum_{r_2=0}^{\infty} \cdots \sum_{r_N=0}^{\infty} \sum_{h=1}^{M} (p(\mathbf{r},h,0) + p(\mathbf{r},h,1)) = 1,$$

которое с учетом разложений (5) может быть представлено в виде

$$\sum_{r_1=0}^{\infty} \cdots \sum_{r_N=0}^{\infty} \sum_{h=1}^{M} \sum_{m=0}^{1} \sum_{k=0}^{\infty} \theta^{|\mathbf{r}|+k} b(k,\mathbf{r},h,m) = 1. \quad (8)$$

Из равенства (8) получаем дополнительные соотношения для коэффициентов $b(k, \mathbf{r}, h, m)$:

$$b(0, \mathbf{r}, h, 0) = \frac{1}{s_{T(h)} S}, \tag{9}$$

$$\sum_{|\mathbf{r}|+k=l} \sum_{h=1}^{M} (b(k, \mathbf{r}, h, 0) + b(k, \mathbf{r}, h, 1)) = 0. \tag{10}$$

В частности, из равенства (10) получаем формулу для вычисления $b(k, \mathbf{0}, 1, 0)$ через величины $b(k-1, \mathbf{r}, h, 0)$, где $\mathbf{0} = (0, \ldots, 0)$,

$$b(k, \mathbf{0}, 1, 0) = -\frac{1}{s_{T(1)} S} \times \tag{11}$$

$$\times \Bigg[\sum_{h=2}^{M} \Bigg(\mu_{T(h-1)} b(k-1, \mathbf{e}_{T(h-1)}, h-1, 1) -$$

$$- \sum_{j=1}^{N} a_j b(k-1, \mathbf{0}, h, 0) \Bigg) \sum_{j=h}^{M} \frac{1}{s_{T(j)}} +$$

$$+ \sum_{0<r_1+\cdots+r_N \le k} \cdots \sum \sum_{h=1}^{M} \sum_{i=0}^{1} b(k-|\mathbf{r}|, \mathbf{r}, h, i) \Bigg].$$

4. Алгоритм вычисления стационарного распределения вероятностей состояний системы

Алгоритм построен на основе равенств (6)–(10) для вычисления коэффициентов $b(k, \mathbf{r}, h, m)$ полиномиальных разложений (5) и описывается следующим образом.

1. Вычисляем $b(0, \mathbf{r}, h, 0)$, $h = \overline{1, M}$, из (9).
2. Далее для каждого уровня l ($l \ge 1$) вычисляем коэффициенты $b(k, \mathbf{r}, h, m)$ таким образом, чтобы $k + |\mathbf{r}| = l$. Полагаем $l = 1$.
3. (a) $k = 0$.
   (b) Из формулы (6) вычисляем коэффициенты $b(k, \mathbf{r}, h, 0)$ ($h = \overline{1, M}$) для всех возможных значений $\mathbf{r}$ таких, что $|\mathbf{r}| = l - k$, изменяя элементы вектора $\mathbf{r}$ в лексикографическомпорядке от $(0, \ldots, 0, l-k)$ до $(l-k, 0, \ldots, 0)$.
   (c) Аналогичным образом из формулы (6) вычисляем $b(k, \mathbf{r}, h, 1)$ для всех $\mathbf{r}$ таких, что $|\mathbf{r}| = l - k$, и $h = \overline{1, M}$.
   (d) Полагаем $k = k + 1$.
   (e) Повторяем шаги 3b–3d до тех пор, пока k не станет равным l.
   (f) Из формулы (11) вычисляем $b(l, \mathbf{0}, 1, 0)$. Далее определяем коэффициенты $b(l, \mathbf{0}, h, 0)$ для всех $h = \overline{2, M}$ из равенств (6).
4. Увеличиваем значение l на единицу и возвращаемся на шаг 3 алгоритма.

Критерием остановки алгоритма может служить расхождение коэффициентов $b(k, \mathbf{r}, h, m)$ для $k = l$ и $k = l+1$ не более чем на допустимую величину ε.

После вычисления величин $b(k, \mathbf{r}, h, m)$, находим стационарное распределение вероятностей состояний системы $p(\mathbf{r}, h, m)$, $\mathbf{r} \ge \mathbf{0}$, $h = \overline{1, M}$, $m = 0, 1$, с помощью равенств (3) и (5). Далее, можно найти характеристики производительности системы, в частности, среднее число заявок в очереди j момент, когда сервер подключается к i-й очереди,

$$L_i^j = \sum_{r_1=0}^{\infty} \cdots \sum_{r_N=0}^{\infty} \sum_{h=1}^{M} r_j p(\mathbf{r}, h, 0) I_{\{T(h)=i\}},$$

а также среднее число заявок в очереди j момент, когда сервер обслуживает i-ю очередь, $i, j = \overline{1, N}$,

$$T_i^j = \sum_{r_1=0}^{\infty} \cdots \sum_{r_N=0}^{\infty} \sum_{h=1}^{M} r_j p(\mathbf{r}, h, 1) I_{\{T(h)=i\}}.$$

5. Численные примеры

В качестве примера, иллюстрирующего полученные результаты, рассмотрим симметричную систему поллинга с тремя очередями.

Пусть параметры, описывающие систему, одинаковы для всех очередей и имеют следующие значения: $\lambda_i = 4$, $\mu_i = 50$, $s_i = 100$, $i = \overline{1, 3}$. Загрузка системы $\rho = 0,24$. Величина θ из равенства (3) имеет значение 0,4. Погрешность, принимаемая для алгоритма, $\varepsilon = 10^{-3}$.

В таблице 1 приведены некоторые значения вероятностей $p(r_1, r_2, r_3, h, 0)$, вычисленные по формулам (5) и соответствующие периоду переключения сервера между очередями. В первом столбце перечислены состояния $(r_1, r_2, r_3, h, 0)$, во втором – соответствующие им вероятности.

В таблице 2 приведены некоторые значения вероятностей, соответствующие обслуживанию сервером очереди.

Средние длины первой и второй очередей имеют значения 0,271, а третьей – 0,141. Коэффициенты разложения (5) рассчитывались до уровня $l = r_1 + r_2 + r_3 + k = 10$.

Далее для краткости в качестве характеристик системы будем приводить лишь средние длины очередей. Пусть теперь в симметричной системе $\lambda_i = 5$, $\mu_i = 50$, $s_i = 200$. Загрузка системы $\rho = 0,3$, $\theta = 0,4$. Средние длины первой и второй очередей равны 0,285, а третьей очереди – 0,161.

Таблица 1. Значения вероятностей $p(r_1, r_2, r_3, h, 0)$

$(\mathbf{r}, h, 0)$	$p(\mathbf{r}, h, 0)$	$(\mathbf{r}, h, 0)$	$p(\mathbf{r}, h, 0)$
(0,0,0,1,0)	0,142	(0,1,0,1,0)	0,018
(0,0,0,2,0)	0,135	(0,1,0,2,0)	0,024
(0,0,0,3,0)	0,138	(0,1,0,3,0)	0,008
(0,0,0,4,0)	0,130	(0,1,0,4,0)	0,011
(0,0,1,1,0)	0,009	(0,1,1,1,0)	0,002
(0,0,1,2,0)	0,014	(0,1,1,2,0)	0,003
(0,0,1,3,0)	0,021	(0,1,1,3,0)	0,001
(0,0,1,4,0)	0,026	(0,1,1,4,0)	0,003
(0,0,2,1,0)	0,001	(0,1,2,1,0)	$2{,}97{\cdot}10^{-4}$
(0,0,2,2,0)	0,001	(0,1,2,2,0)	$7{,}06\cdot10^{-4}$
(0,0,2,3,0)	0,002	(0,1,2,3,0)	$2{,}89\cdot10^{-4}$
(0,0,2,4,0)	0,004	(0,1,2,4,0)	$8{,}34\cdot10^{-4}$
(0,0,3,1,0)	$1{,}51\cdot10^{-4}$	(0,1,3,1,0)	$3{,}95\cdot10^{-5}$

Таблица 2. Значения вероятностей $p(r_1, r_2, r_3, h, 1)$

$(\mathbf{r}, h, 1)$	$p(\mathbf{r}, h, 1)$	$(\mathbf{r}, h, 1)$	$p(\mathbf{r}, h, 1)$
(0,0,1,4,1)	0,043	(0,2,2,4,1)	$6{,}14\cdot10^{-4}$
(0,0,2,4,1)	0,009	(0,2,3,2,1)	$2{,}09\cdot10^{-4}$
(0,0,3,4,1)	0,001	(0,2,3,4,1)	$2{,}10\cdot10^{-4}$
(0,0,4,4,1)	$2{,}99\cdot10^{-4}$	(0,2,4,2,1)	$6{,}21\cdot10^{-5}$
(0,0,5,4,1)	$5{,}54\cdot10^{-5}$	(0,2,4,4,1)	$6{,}95\cdot10^{-5}$
(0,0,6,4,1)	$1{,}07\cdot10^{-5}$	(0,2,5,2,1)	$1{,}75\cdot10^{-5}$
(0,1,2,2,1)	0,001	(0,2,5,4,1)	$2{,}14\cdot10^{-5}$
(0,1,2,4,1)	0,002	(0,2,6,2,1)	$4{,}36\cdot10^{-6}$
(0,1,3,2,1)	$3{,}94\cdot10^{-4}$	(0,2,6,4,1)	$5{,}69\cdot10^{-6}$
(0,1,3,4,1)	$6{,}05\cdot10^{-4}$	(0,2,7,2,1)	$8{,}66\cdot10^{-7}$
(0,1,4,2,1)	$9{,}34\cdot10^{-5}$	(0,2,7,4,1)	$1{,}19\cdot10^{-6}$
(0,2,1,4,1)	0,001	(0,2,8,2,1)	$1{,}05\cdot10^{-7}$
(0,2,2,2,1)	$7{,}24\cdot10^{-4}$	(0,2,8,4,1)	$1{,}49\cdot10^{-7}$

Рассмотрим несимметричную систему, в которой $\lambda_1 = 5$, $\lambda_2 = 2$, $\lambda_3 = 1$, $\mu_1 = 70$, $\mu_2 = 40$, $\mu_3 = 30$, $s_1 = 50$, $s_2 = 75$, $s_3 = 100$. Загрузка системы равна $\rho = 0,154$, величина $\theta = 0,313$. В этом случае средние длины очередей равны, соответственно, 0,232, 0,175 и 0,090.

6. Заключение

В работе рассмотрена система поллинга с опросом типа "звезда" и 1-ограниченным обслуживанием. Разработан алгоритм расчета стационарных вероятностей состояний системы путем их полиномиального разложения. Результаты работы алгоритма проиллюстрированы численными примерами. Стационарное распределение вероятностей далее может быть использовано для нахождения времен ожидания обслуживания в очередях.

Список литературы

[1] В.М. Вишневский, А.И. Ляхов, С.Л. Портной, И.В. Шахнович, *Широкополосные беспроводные сети передачи информации*, М.: Техносфера, 2005.

[2] В.М. Вишневский, А.И. Ляхов, Н.Н. Гузаков, "Оценка максимальной производительности беспроводного доступа в Интернет", *Автоматика и телемеханика*, 2004, , No. 9, pp. 52–70.

[3] В.М. Вишневский, О.В. Семенова, "Математические методы исследования систем поллинга", *Автоматика и телемеханика*, 2006, , No. 2, pp. 3–56.

[4] В.М. Вишневский, О.В. Семенова, *Системы поллинга: теория и применение в широкополосных беспроводных сетях*, М.: Техносфера, 2007.

[5] J. P. C. Blanc, "The power-series algorithm for polling systems with time limits", *Probability in the Engineering and Informational Sciences*, 1998, Vol. 12, pp. 221–237.

[6] J.P.C. Blanc, "A numerical approach to cyclic-service queueing models", *Queueing Systems*, 1990, Vol. 6, No. 2, pp. 173–178.

[7] S.C. Borst, *Polling systems*, Amsterdam: Stichting Mathematisch Centrum, 1996.

[8] N.P. Giannakouros, A. Laloux, "Waiting-time approximations for service systems with star pollingsequence and mixed service strategies", *IEEE Transactions on Communications*, 1991, Vol. 39, No. 7, pp. 1041–1045.

[9] V. M. Vishnevsky, A. I. Lyakhov, "Adaptive features of IEEE 802.11 protocol: utilization, tuning and modifications", *Proc. of 8th HP-OVUA Conf.*, 2001.

[10] V. M. Vishnevsky, A. I. Lyakhov, N. N. Guzakov, "An adaptive polling strategy for IEEE 802.11 PCF", *Proc. 7th Int. Symp. on Wireless Personal Multimedia Communications (WPMC'04)*, 2004, Vol. 1, pp. 87–91.

Анализ характеристик многокилометрового широкополосного беспроводного канала точка-точка с различными функциями управления[1]

Фахриев Д.Н., Шпилев С.А.
Институт проблем передачи информации им. А.А. Харкевича РАН, Москва
fahriev.denis@gmail.com, santt@iitp.ru

Аннотация

В данной работе исследуются зависимость скорости передачи от нагрузки и пропускная способность многокилометрового широкополосного беспроводного канала точка-точка, работающего под управлением протокола IEEE 802.11. Рассматриваются и сравниваются между собой два случая. В первом используется режим распределенного управления DCF (Distributed Coordination Function), в котором реализуется механизм множественного доступа с контролем несущей и предотвращением коллизий (Carrier Sense Multiple Access with Collision Avoidance - CSMA/CA). Во втором одна из двух станций исполняет роль координатора (Hybrid Coordinator, HC), и канал работает в режиме гибридного управления HCF (Hybrid Coordination Function) в соответствии с механизмом HCCA (HCF Controlled Channel Access). Для каждого из случаев исследуется зависимость параметров канала от размеров пакета и минимального конкурентного окна. Механизм RTS/CTS не используется, так как предполагается, что шумы в канале отсутствуют, вследствие чего искажений кадров не происходит. Все исследования проведены с использованием имитационного моделирования на языке GPSS World.

1. Введение

Наиболее часто протокол IEEE 802.11 используется для построения локальных компьютерных сетей, однако он предусматривает возможность расположения станций на большом расстоянии друг от друга (до 115 км), что позволяет строить на его основе транспортные сети регионального масштаба. Реализуется данная возможность путем введения в стандарте понятия зоны покрытия (coverage class). В заголовке каждого пакета находится специальное поле (Coverage Class field) размером 1 байт, значение которого (Coverage Class value) имеет простой физический смысл и вычисляется следующим образом:

T_0 = (Coverage Class value) × 3

где T_0 – время распространения сигнала от одной станции до другой и обратно.

Для локальных сетей, характерный размер которых порядка 100 м, время T_0 пренебрежимо мало (≈ 1 мкс), и можно считать, что сигнал мгновенно доходит от одной станции до другой. В случае же большого расстояния между устройствами поведение сети изменяется, так как время распространения становится весьма ощутимым. Так, например, время распространения сигнала при расстоянии между станциями 50 км составляет 166.5 мкс, что сравнимо со временем передачи самого пакета (для пакета с информационным полем 1 кбайт время его передачи при скорости 54 Мбит/с равняется 180 мкс).

2. Режим распределенного управления

Режим распределенного управления DCF является основным для протокола IEEE 802.11. Сеть в данном режиме работает в соответствии с механизмом множественного доступа с контролем несущей и предотвращением коллизий (carrier sense multiple access with collision avoidance, CSMA/CA). Перед началом передачи устройство слушает эфир и дожидается, когда канал освободится. После того, как это произойдет, станция ждет еще в течение специального отрезка времени, который называется межкадровым интервалом IFS (interframe space). Он бывает двух типов: DIFS (DCF interframe space) – в случае если предыдущая передача была успешной (т.е

[1] Работа выполнена при поддержке Российского Фонда Фундаментальных Исследований, грант № 08-07-90102-Мол_а.

коллизия отсутствовала, и все кадры, относящиеся к этому пакету, были переданы корректно) или EIFS (extended interframe space) – при неудачной попытке (рис. 1). Если в течение этого времени канал продолжал оставаться свободным, то устройство переходит в состояние отсрочки (BackOff), при этом специальный счетчик времени нахождения в этом состоянии устанавливается в начальное значение b. Это значение называется временем отсрочки, измеряется в слотах σ (slot time) и равновероятно выбирается из множества (0,...,CW-1). Значение CW, называемое конкурентным окном (contention window), зависит от значения числа n_r сделанных попыток передачи текущего пакета:

$CW = CW_{min} 2^{n_r}$ при $n_r < m$,
$CW = CW_{max}$ при $n_r \geq m$,

где CW_{min} и CW_{max} соответственно минимальное и максимальное конкурентные окна.

В состоянии отсрочки устройство продолжает прослушивать среду и, пока та остается свободной, декрементирует значение счетчика времени отсрочки. Как только его значение обнулится, устройство сможет начать передачу пакета. Если до обнуления счетчика канал занимает другое устройство, счет останавливается, сохраняя достигнутое значение. При следующей попытке отсчет начнется с сохраненной величины. В результате неуспевшая в этот раз занять среду станция получает больше шансов занять ее при следующей попытке.

Если посланный передатчиком пакет был предназначен конкретному устройству (не широковещательная и не многоадресная передача), то приемник, приняв пакет, спустя интервал времени SIFS (short interframe space) посылает обратно короткий кадр подтверждения получения ACK (Acknowledgment). Если передатчик не принял ACK, он считает пакет утерянным и повторяет процедуру его передачи. Так как предполагается, что шумы в канале отсутствуют, потери пакетов могут происходить только в случае возникновения коллизий при передаче. Если станция сразу после передачи пакета обнаруживает (рис. 2), что среда занята (например, пакетом от другой станции), она автоматически считает, что при передаче возникла коллизия, ждет освобождения канала,

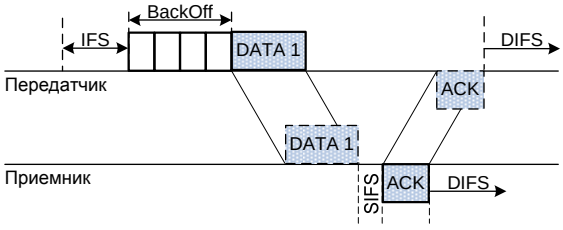


Рис. 1. Режим DCF

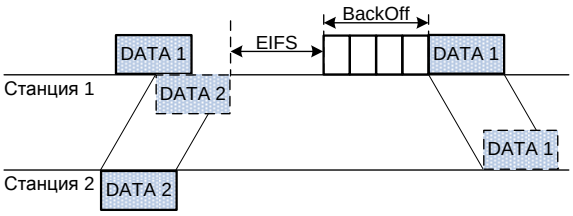


Рис. 2. Коллизия в режиме DCF

после чего повторяет процедуру его передачи с IFS = EIFS (рис. 2). В случае если количество попыток передачи данного пакета превысило R раз, станция отбрасывает его и передает следующий пакет с межкадровым интервалом IFS = DIFS. Если сразу после передачи пакета канал оказался свободным, станция переходит в состояние ожидания кадра ACK на время ACK_Timeout. Если в течение этого времени ничего не получено, устройство считает, что произошла коллизия, и повторяет процедуру передачи, уходя в состояние отсрочки без интервала IFS (рис. 3).

Если, находясь в состоянии ожидания ACK, станция получает ACK, она считает передачу успешной. Если же она принимает пакет данных, то сама отвечает на него кадром ACK, после чего повторяет процедуру передачи с IFS = DIFS (рис. 4).

К нежелательным последствиям временных задержек, возникающих при распространении сигнала на большие расстояния, относятся случаи избыточной посылки одного и того же пакета. Так, например, в случае, изображенном на рис. 4, Станция 1 успешно принимает пакет DATA 2 и отвечает кадром ACK. В соответствии с описанным выше алгоритмом, Станция 2, обнаружив, что после передачи пакета канал занят, считает, что произошла коллизия. Поэтому, не обращая внимания на полученный

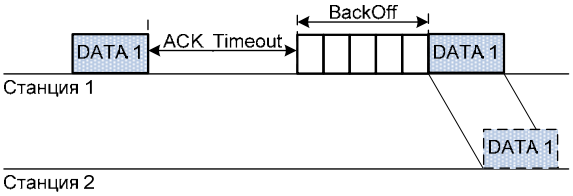


Рис. 3. ACK_Timeout

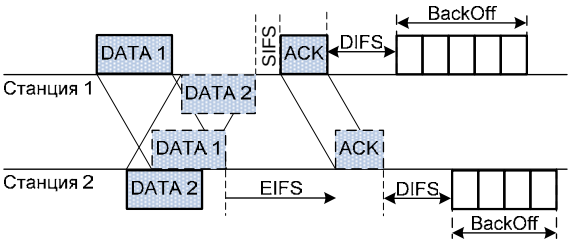


Рис. 4. Случай избыточной посылки пакета

позже ACK, она повторяет процедуру передачи посланного успешно пакета DATA 2.

3. Режим гибридного управления

Одним из главных недостатков режима DCF является отсутствие гарантий качества обслуживания QoS (Quality of Service). Поэтому в случаях, когда необходимо обеспечить дифференцированное качество обслуживания, в Wi-Fi сетях используются другие режимы управления доступом к среде: PCF (point coordination function) и HCF (hybrid coordination function).

В данной статье рассматривается режим гибридного управления HCF с механизмом HCCA (HCF controlled channel access). В этом режиме одна из станций, называемая гибридным координатором HC (hybrid coordinator), управляет работой всех остальных станций сети, рассылая специальные управляющие кадры.

В механизме HCCA определены два различных временных периода, которые в процессе работы сети постоянно сменяют друг друга. Первый из них – это период неконкурентного доступа CFP (contention free period), в течение которого HC полностью контролирует доступ к среде, а второй – конкурентного доступа CP (contention period), когда сеть работает в режиме распределенного управления. Длительности двух этих периодов не являются фиксированными величинами и могут динамически настраиваться на HC.

Необходимо отметить, что HC может управлять работой сети и в течение периода конкурентного доступа, так как имеет наивысший приоритет в распределенном режиме управления. В отличие от всех остальных станций сети, координатор, дождавшись освобождения среды, ждет уменьшенный межкадровый интервал PIFS (PCF interframe space), после чего, если канал все еще не занят, начинает передачу пакета. Для того чтобы отдать приоритет некоторой станции сети, HC должна послать ей специальный управляющий CF-Poll пакет, в заголовке которого содержится специальное 8-битное поле TXOP (transmission opportunity) Limit. Значение этого поля представляет собой время, измеряемое в слотах по 32 мкс, на которое станция получает канал в свое распоряжение. В течение TXOP Limit она может осуществлять бесконкурентный доступ к среде и передавать необходимое количество пакетов(рис. 5). Если значение TXOP Limit равно нулю, то, спустя интервал времени SIFS после приема CF-Poll, станция должна послать в ответ либо один информационный пакет, либо, если у станции нет данных для передачи, специальный пакет Null.

Для рассматриваемого в данной работе многокилометрового канала точка-точка использование механизма HCCA сводится к следующему: одна из двух станций исполняет роль гибридного координатора HC; в качестве рабочего выбирается период конкурентного доступа CP (длительность неконкурентного периода CFP можно сделать сколь угодно малой и не учитывать при анализе работы сети). Тогда, если у HC есть данные для передачи, и канал свободен в течение межкадрового интервала времени PIFS, она посылает другой станции пакет типа DATA+CF-Poll, с нулевым значением поля TXOP Limit (рис. 6). В соответствии с описанным выше алгоритмом, станция, получив этот пакет, должна ответить координатору одним пакетом DATA+CF-Ack (либо CF-Ack (no data) при отсутствии данных для передачи), а тот в свою очередь посылает ей пакет DATA+CF-Ack+CF-Poll (обозначенный на рисунке 6 как D+A+CF-Poll) и так далее. Этот процесс будет продолжаться до тех пор, пока данные для передачи не закончатся у обеих станций.

Видно, что гибридный режим управления HCF с механизмом HCCA позволяет значительно улучшить характеристики канала точка-точка ввиду того, что паузы между посылками пакетов практически отсутствуют, и канал используется с наибольшей эффективностью.

4. Результаты

В данном разделе статьи приведены полученные с помощью имитационного моделирования характеристики канала точка-точка в режимах DCF и HCF.

В качестве примера, в работе рассматривается канал точка-точка длины 50 км при максимально возможной (согласно протоколу IEEE 802.11a)

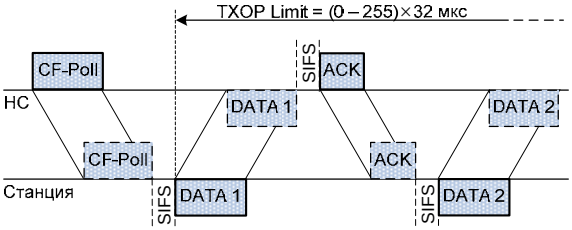


Рис. 5. TXOP Limit в режиме HCCA

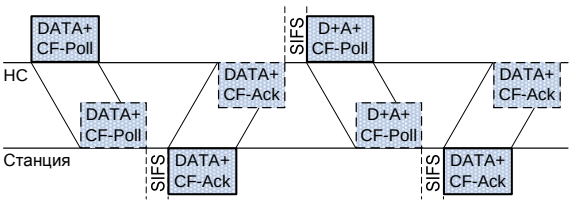


Рис. 6. Режим HCCA для канала точка-точка

номинальной скорости передачи 54 Мбит/с. Входящий поток пакетов для каждой из станций является пуассоновским с интенсивностью λ. Максимальное количество повторных посылок пакета R = 7, максимальное конкурентное окно CW_{max} = 1024. При таком расстоянии между станциями время δ распространения сигнала составляет 166.5 мкс, что позволяет определить все остальные временные параметры:
σ = 9 мкс + 2δ = 342 мкс,
SIFS = 16 мкс,
PIFS = SIFS + σ = 358 мкс,
DIFS = SIFS + 2σ = 700 мкс,
EIFS = SIFS + DIFS + 44 мкс = 760 мкс,
ACK_Timeout = SIFS + σ + 25 мкс = 383 мкс.

На рисунке 7 представлены графики зависимости скорости передачи V от среднего интервала времени между моментами поступления пакетов 1/λ для режима распределенного управления DCF. Рассматриваются два случая: в первом все пакеты имеют размер информационного поля 1 кбайт (общепринятый размер пакета), во втором – 2312 байт (максимально допустимый в стандарте IEEE 802.11 размер пакета). Время, необходимое для передачи этих пакетов, равно соответственно 180 и 372 мкс, а для передачи ACK пакета – 24 мкс. Для каждого из случаев проведено по четыре серии экспериментов с различным значением минимального конкурентного окна CW_{min}.

В таблице 1 приведены полученные с помощью имитационного моделирования значения пропускной способности канала. Видно, что увеличение размера пакета влечет за собой повышение скорости передачи, так как за за одну посылку пакета передается большее количество информации при тех же накладных расходах. Так же видно, что при уменьшении CW_{min} пропускная способность канала растет. Это объясняется тем, что в среднем станции находятся в состоянии отсрочки меньшее время, и пакеты начинают посылаться значительно чаще.

При постепенном увеличении нагрузки количество коллизий так же увеличивается, и в определенный момент достигает некоторого

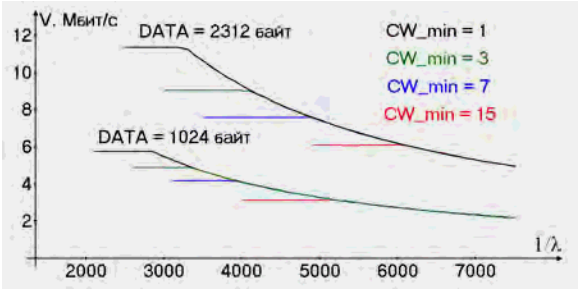


Рис. 7. Скорость передачи в режиме DCF

Таблица 1. Пропускная способность в режиме DCF

CW_{min}	DATA=2312 байт (Мбит/с)	DATA=1024 байт (Мбит/с)
1	11.365	5.768
3	8,978	4,864
7	7,547	4,148
15	6,048	3,122

максимального значения (рис. 8). После чего для случаев CW_{min} = 1 и CW_{min} = 3 количество коллизий резко падает, что объясняется влиянием эффекта захвата.

Рассмотрим эффект захвата среды для канала точка-точка при CW_{min} = 1. Пусть после успешной передачи пакета станцией 1, обе станции выбрали одинаковые времена отсрочки b_1 и b_2 , тогда станция 1 успешно передаст пакет, а станция 2 попадет в коллизию (рис. 9). Поэтому при следующем выборе времени отсрочки, вторая станция увеличит значение CW до трех. Предположим, что станция 2 выбрала $b_2 \neq 0$ (как будет показано позже, это единственный случай, при котором станция 1 может потерять контроль над каналом). Пусть для определенности b_2 = 3 (рис. 10).

Видно, что в зависимости от значения b_1 время отсрочки второй станции либо уменьшается на единицу, либо не изменяется вообще, в то время как станция 1 успешно посылает пакет.

Система будет находиться в этом состоянии до тех пор, пока b_2 не уменьшиться до единицы, а станция 1 после очередной посылки пакета не выберет b_1 = 1 (рис. 11). Видно, что система после этого вернулась в исходное состояние,

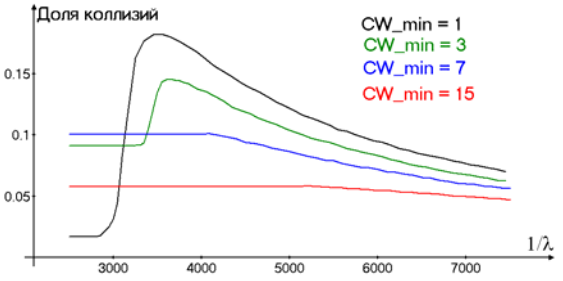


Рис. 8. Доля коллизий в режиме DCF (DATA = 1024 байт)

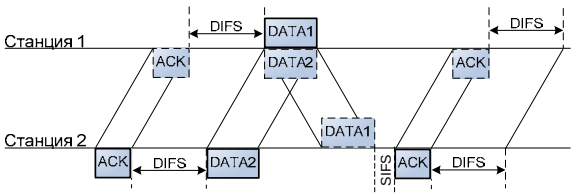


Рис. 9. Эффект захвата. Состояние 1

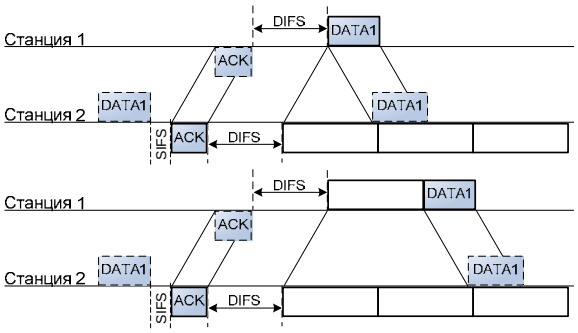


Рис. 10. Эффект захвата. Состояние 2

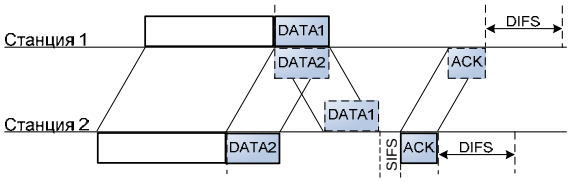


Рис. 11. Эффект захвата. Состояние 3

изображенное на рисунке 9, с той лишь разницей, что CW_2 стало больше, а поэтому весь цикл повторится еще раз только с большим временем отсрочки у второй станции. Таким образом станция 1 захватила канал и свободно передает пакеты, в то время как станция 2 постоянно попадает в коллизию.

Как упоминалось выше, случай, когда после очередной коллизии b_2 выбирается равным нулю, является единственным, при котором станция 1 может потерять контроль над каналом. Если станция 1 выбрала $b_1 = 1$, то в этом случае станция 2 успешно передаст пакет данных и, так как b_1 останется равным единице, сама захватит канал (рис. 12).

Таким образом эффект захвата сводиться к тому, что станции по очереди шлют друг другу несколько пакетов. Хотя это и уменьшает количество коллизий в канале (рис. 8), но приводит к крайне нежелательной ситуации, когда у каждой из станций возникают длительные паузы в передаче пакетов.

На рисунке 13 представлена зависимость скорости передачи от среднего интервала времени между моментами поступления пакетов $1/\lambda$ для режима гибридного управления HCF. Как было показано в пункте 3, в режиме насыщения станции по очереди шлют друг другу пакеты данных через короткий межкадровый интервал

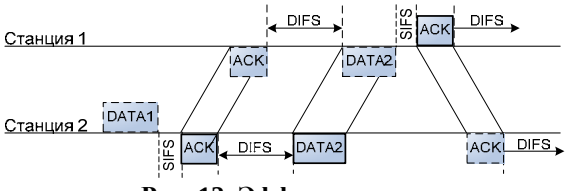


Рис. 12. Эффект захвата.
Потеря станцией 1 контроля над каналом

SIFS, и пропускная способность канала не зависит от значения минимального конкурентного окна.

Значение пропускной способности легко вычислить аналитически. Если обозначить время передачи пакета как T_p, а размер информационного поля как D, то, как видно из рисунка 6, скорость передачи в режиме насыщения составляет

$$V_{\max} = \frac{D}{T_p + \delta + SIFS}$$

В таблице 2 приведены полученные с помощью имитационного моделирования значения пропускной способности канала для пакетов различной величины.

Приведенные результаты наглядно демонстрируют эффективность применения механизмов централизованного управления в канале точка-точка в режиме насыщения. Использование гибридного режима HCF позволяет повысить пропускную способность в 3-4 раза, в зависимости от размера пакета.

5. Заключение

В данной работе приведен анализ работы многокилометрового беспроводного канала точка-точка в режимах распределенного и гибридного управления. При помощи имитационного моделирования получена зависимость скорости передачи от нагрузки канала. Продемонстрирован значительный выигрыш в пропускной способности, получаемый при использовании в канале механизмов

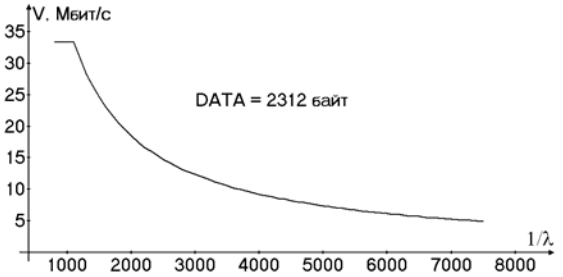


Рис. 13. Скорость передачи в режиме HCF (HCCA)

Таблица 2. Пропускная способность в режиме HCF (HCCA)

Пакет (байт)	Пропускная способность (Мбит/с)
1024	22,6
1500	27,9
2000	31,6
2312	33,4

централизованного управления.

6. Список литературы

[1] С.А. Шпилев, Д.Н. Фахриев, “Анализ характеристик широкополосных беспроводных каналов, под управлением протокола IEEE 802.11”, Третья всероссийская молодежная научная конференция по проблемам управления (ВМКПУ-2008), 2008.

[2] В.М. Вишневский, А.И. Ляхов, С.Л Портной и И.В. Шахнович, *Широкополосные беспроводные сети передачи информации*, Техносфера, Москва, 2005.

[3] Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, IEEE Std 802.11-2007, June 2007.

[4] Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: High-speed Physical Layer in the 5 GHz Band, IEEE Std 802.11a-1999, 1999.

[5] T.J. Schriber, Simulation using GPSS, John Wiley & Sons, 1974.

Организация повторных передач в системах связи с ретрансляционными станциями

А.С. Уханова
Санкт-Петербургский Государственный Университет Аэрокосмического Приборостроения
ann.ukhanova@gmail.com

Аннотация

При построении систем с ретрансляционными станциями возникает большое число задач. В данной работе из этого большого числа задач решается только часть из них связанных с передачей данных от одного абонента через одну промежуточную ретрансляционную станцию.
При передаче сообщения могут искажаться, поэтому с одной стороны целесообразно разбивать сообщение на более мелкие части и осуществлять передачу по сегментам, а с другой стороны при разбиении на сегменты возникают дополнительные издержки, например, на контрольную сумму. Данную оптимизационную задачу необходимо решить в предположении, что передача идет через ретрансляционную станцию. Наличие промежуточной станции накладывает специфику на эту оптимизационную задачу.

1. Введение

В последнее время огромными темпами развивается мобильная беспроводная индустрия. На сегодняшний момент более чем миллиард людей пользуются мобильной связью и в течение ближайших лет эта цифра будет только увеличиваться.

Впервые концепция мобильного телефона была разработана в 80х годах. Концепция сотовой связи представляла собой набор базовых станций расположенных неподалеку друг от друга и протокола передачи управления между двумя соседними станциями при наличии движущегося абонента. На тот момент все системы строились на основе аналоговой передачи данных и были достаточно громоздкие. Впоследствии эти системы стали называть системами сотовой связи первого поколения.

Второе поколение сотовой связи появилось в начале 90х вместе с разработкой первой цифровой системы в GSM (Global Systems for Mobile Communications). Размер и вес этой системы значительно снизился, что способствовало проникновению мобильных технологий в ежедневную жизнь. GSM был создан в основном для передачи голоса, и немногим позже эволюционировал в EDGE (Enhanced Data Rates for GSM Evolution) и GPRS (General Packet Radio System).. Однако требования к сервисам мультимедиа продолжали расти, и это подталкивало к развитию новых технологий. На смену сотовой связи второго поколения пришли такие технологии как UMTS, которые впоследствии стали называться технологиями третьего поколения. Мобильные технологии третьего и четвертого поколения ориентируются на нужды будущего. Все сервисы, включая интернет, мультимедиа сервисы, должны быть реализованы на высоких скоростях с высокой пропускной способностью. Другим аспектом связи четвертого поколения является то, что целях уменьшения энергопотребления сужается диапазон покрытия базовой станции. Для решения этой проблемы используются ретрансляционные станции.

2. Постановка задачи и система допущений

Чтобы решить проблему оптимального разбиения сообщения на сегменты при передаче данных от базовой станции к ретрансляционной и от ретрансляционной станции к абоненту, необходимо уметь решать эту задачу хотя бы для одного участка этой цепи, например, от ретрансляционной станции к абоненту.

Для решения этой задачи введем следующие допущения[1]:

1)На базовой станции имеется непрерывная последовательность сообщений каждое состоящее из k информационных бит,

которое может быть разбито на n сегментов. В одном сегменте не могут содержаться фрагменты разных сообщений.

2) К информационным битам добавляется контрольная сумма, состоящая из r бит. Таким образом, если сообщение разбивается на n сегментов, то при передаче одного сегмента на физический уровень передается $k/n+r$ бит.

3) Контрольная сумма позволяет обнаруживать ошибки. Предполагается, что все ошибки обнаруживаются.

4)В прямом канале каждый бит может быть искажен с вероятностью p. Ошибки в разных битах статистически независимы.

5)Получатель отправляет положительную квитанцию по обратному каналу, если сообщение принято без ошибок, и отрицательную в обратном случае. Важно отметить, что в обратном канале ошибки не происходят.

6) Отправитель получает квитанции через постоянное время τ после отправки сообщения. В данной работе единицы задержки считаются во времени передачи одного сообщения.

3. Передача без использования ретрансляционной станции

Коэффициент использования канала одного участка зависит от следующих факторов:

- вероятность ошибки на бит;
- размер сообщения;
- количество сегментов в сообщении;
- длина контрольной суммы;
- величина задержки в канале;
- тип используемого алгоритма и число сегментов, которое может храниться на приемной стороне, и размер буфера.

Минимальный размер буфера при котором может функционировать система равен одному сообщению. Необходимо также отметить, что несмотря на то, что сообщение передается вместе с контрольной суммой, хранить ее на приемной стороне нет необходимости.

В общем случае коэффициент использования канала рассчитывается по следующей формуле:

$$\eta = AB, \qquad (1)$$

Где A – коэффициент, не зависящий от свойств используемого алгоритма повторных передач, а B - коэффициент, зависящий от конкретного алгоритма повторных передач.

Укажем способ, как можно вычислить множитель для разных типов алгоритмов повторных передач и объема буфера.

Формула будет рассчитываться в два этапа. Первым множителем будет являться следующий коэффициент

$$A = \frac{k/N}{k/N+r}, \qquad (2)$$

где k - длина исходного сообщения,

N - количество сегментов в сообщении,

r - длина контрольной суммы.

Второй множитель будет учитывать особенности используемого алгоритма. Для расчета понадобятся значения задержки в канале y (рассчитываемой в зависимости от длины сообщения) и вероятности ошибки на сообщение q:

$$y = \left\lceil \frac{\tau(k+r)}{k/N+r} \right\rceil, \qquad (3)$$

$$q = 1-(1-p)^{k/N+r}, \qquad (4)$$

где τ - исходная величина задержки в канале,

k - длина исходного сообщения,

N - число сегментов, на которое разбивается сообщение,

r - длина контрольной суммы,

p - вероятность ошибки на бит.

Расчет задержки происходит следующим образом. Берем первоначальную задержку, выраженную в битовом эквиваленте – это величина, которая показывает, через какое время придет квитанция. Далее рассчитываем, сколько сегментов новой длины мы можем передать за это время. На деле квитанция придет в момент окончания передачи последнего сообщения, если в задержку укладывается целое число сообщений, либо в момент передачи последнего сообщения, и тогда мы округляем количество сегментов вверх.

При расчете значения вероятности ошибки на сегмент мы исходим из допущения, что вероятность ошибки на бит не зависит от длины сегмента.

Для алгоритма с возвратом второй множитель будет равен

$$B = \frac{1-q}{1+yq} \qquad (5)$$

Для алгоритма с селективным повторением с буфером на один сегмент этот множитель будет равен

$$B = \frac{1}{1+\dfrac{q(y+1)(1+q)}{1+q-2q^2}} \qquad (6)$$

Для алгоритма с селективным повторением, когда на приемной стороне хранится более одного сегмента, значение

множителя, зависящего от алгоритма повторных передач, может быть примерно оценено или определено с помощью имитационного моделирования.

Если нет никаких ограничений на размер буфера и используется алгоритм с селективным повторением, то вычисления можно производить по следующей формуле

$$B = 1 - q \qquad (7)$$

Для алгоритма с возвратом и алгоритма с селективным повторением с неограниченным буфером формулы расчета коэффициента использования канала можно найти во многих работах, например, [2].

Пусть у нас есть некая функция $B = f(q, \tau, b \cdot m - 1)$. Программа для выбора числа сегментов способна при следующих входных параметрах таких как вероятность ошибки на пакет q, задержки в системе τ, размера буфера $b \cdot m - 1$ выдавать значение оптимального числа сегментов в этом случае.

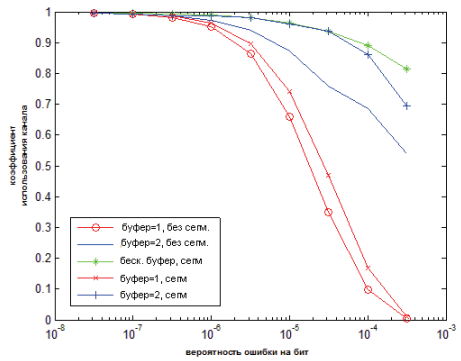


Рис. 1 – Коэффициент использования канала в зависимости от вероятности ошибки на бит и типа используемого алгоритма

Зафиксируем вероятность ошибки и покажем, как меняется коэффициент использования канала от размера буфера при сегментации и без нее.

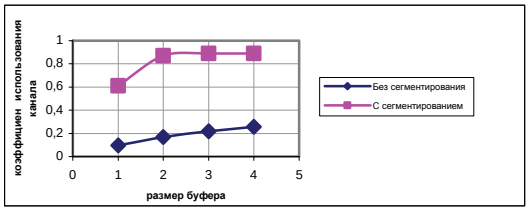


Рис. 2 - Коэффициент использования канала с сегментированием и без сегментирования (вероятность ошибки в канале = 10^{-4})

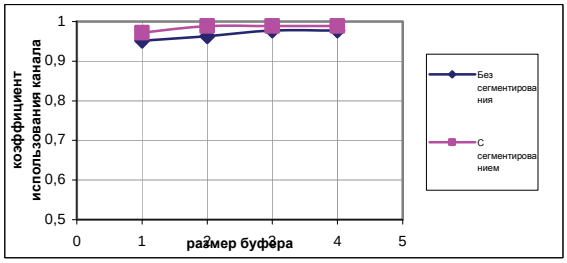


Рис. 3 - Коэффициент использования канала с сегментированием и без сегментирования (вероятность ошибки в канале = 10^{-6})

При увеличении объема буфера коэффициент использования канала возрастает, но эффект который дает сегментирование становится меньше.

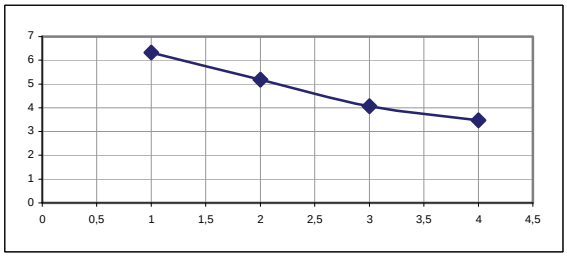


Рис. 4 - Отношение коэффициента использования канала для случая с сегментированием к коэффициенту использования канала для случая без сегментирования для вероятности 10^{-4}

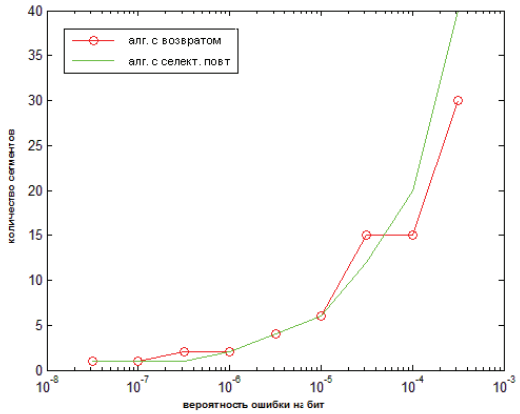


Рис. 5 – Зависимость оптимального числа сегмента от вероятности ошибки на бит

Оптимальное число сегментов увеличивается с ростом вероятности ошибки в канале для всех видов алгоритмов (см. рис. 5).

В работе предложена методика, позволяющая для любых входных параметрах – вероятность ошибки в канале, задержка, объем

буфера - выбрать оптимальный алгоритм и оптимальное число сегментов.

Как было сказано выше, все расчеты проводились при допущении, что вероятность ошибки на бит не зависит от длины сегмента. На практике вероятность ошибки на бит уменьшается с увеличением длины сегмента. Однако даже при этом допущении расчеты показывают, что при ограниченном размере буфера нет необходимости производить дополнительную сегментацию на ретрансляционной станции. Однако, если бы передавалось только одно сообщение, и требовалось бы минимизировать задержку на время передачи одного сообщения, то ресегментацию надо было бы делать всегда. Сначала выбирается оптимальный размер сегмента для первого участка, потом оптимальный размер для второго участка. Еще раз подчеркнем, что если же идет непрерывный поток сообщений или информационных бит, то ресегментация не нужна и размер сегмента определяется исходя из вероятностей ошибок и задержки в двух участках канала.

4. Передача с использованием ретрансляционной станции

4.1. Алгоритм с селективным повторением и неограниченным размером буфера

Рассмотрим случай алгоритма с селективным повторением с неограниченным буфером на ретрансляционной станции.

Схематично модель изображена на рис. 6. Аналогичная модель описана в работе [1]. За p_1 обозначим вероятность ошибки на бит на участке от базовой станции к ретрансляционной, а за p_2 – вероятность ошибки на бит при передаче от ретрансляционной станции к абонентской.

С учетом ранее приведенных допущений очевидно, что все определяется каналом с наихудшей пропускной способностью, но мы дадим достаточно строгое обоснование этого утверждения.

Покажем, каким образом можно выбрать оптимальную длину сегмента. С учетом введенных допущений на передачу одного сегмента на первом и втором участках тратится

$$T_{обсл.1} = \frac{1}{1-(1-p_1)^{\frac{k}{n}+r}} \quad (8)$$

и

$$T_{обсл.2} = \frac{1}{1-(1-p_2)^{\frac{k}{n}+r}} \quad (9)$$

единиц времени соответственно. За единицу времени выбрано время передачи одного сегмента.

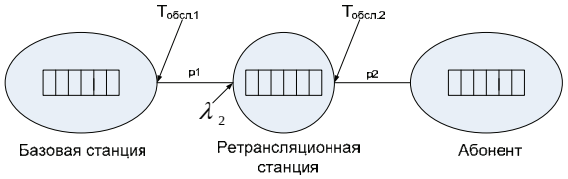


Рис. 6 - Схема передачи данных с ретрансляционной станцией.

Рассмотрим два случая:

1) Пусть отношение сигнал/шум (SNR) на первом участке меньше, чем на втором ($SNR_1 < SNR_2$), то есть $p_1 > p_2$, тогда из формул (8) и (9) следует, что $\overline{T}_{обсл.1} > \overline{T}_{обсл.2}$.

В этом случае ретрансляционную станцию можно описать как систему массового обслуживания с одним обслуживающим прибором, интенсивностью входного потока $\lambda_2 = \frac{1}{T_{обсл.1}}$ и с временем обслуживания равным $T_{обсл.2}$. При этом $\lambda_2 \cdot T_{обсл.2} = \frac{T_{обсл.2}}{T_{обсл.1}} < 1$. Если $p_1 > p_2$, то при любой длине сегмента у ретрансляционной станции не будет бесконечных очередей, поэтому мы можем выбирать размер сегмента так, чтобы оптимизировать коэффициент использования канала на первом участке.

2) Пусть $SNR_1 > SNR_2$, то есть $p_1 < p_2$, тогда $\overline{T}_{обсл.1} < \overline{T}_{обсл.2}$.

При этом $\lambda_2 \cdot T_{обсл.2} = \frac{T_{обсл.2}}{T_{обсл.1}} > 1$, следовательно система будет работать неустойчиво и очередь будет переполняться. В этом случае ретрансляционной станции придется отсылать отрицательные квитанции, и время работы будет определяться худшим каналом.

Таким образом, оптимальная длина сегмента может быть выбрана с помощью следующей формулы:

$$\frac{k/n}{k/n+r}(1-\max(p_1, p_2))^{k/n+r} \quad (10)$$

Оптимальным является такое число n, при котором максимизируется значение вышеуказанной формулы.

4.2. Алгоритм с минимально возможным буфером (алгоритм с возвратом)

Для рассмотренного выше случая с неограниченным размером буфера задержка в канале не влияет на коэффициент использования канала для алгоритма с селективным повторением. Теперь обобщим вышеприведенные рассуждения на случай ограниченного объема буфера.

Так как ретрансляционная станция должна быть относительно дешевой по сравнению с базовой станцией, то на ней имеется буфер ограниченного размера. Однако это не означает что пакеты будут теряться, так как если ретрансляционная станция не может принять данные из-за того, что буфер заполнен, то она пересылает это базовой станции и она не посылает сегменты до тех пор пока ретрансляционная станция не освободит свой буфер.

Пусть на ретрансляционной станции и на абоненте имеется буфер ограниченного объема и для передачи используется алгоритм с возвратом. Задержка от момента окончания передачи сообщения базовой станцией до момента получения квитанции от ретрансляционной станции равняется τ_1(то есть до момента получения квитанции по каналу может быть передано τ_1 сообщений). Задержка от момента окончания передачи сообщения ретрансляционной станцией до момента получения квитанции от абонента равняется τ_2.

Для простоты рассуждений ограничимся случаем, когда отношение сигнал/шум (SNR) на участке от базовой станции к ретрансляционной больше, чем на участке от ретрансляционной станции к абоненту, то есть вероятность ошибки на бит на первом участке меньше, чем на втором ($p_1 < p_2$).

С учетом введенных допущений время для передачи сообщения на первом участке $T_{обсл.1} = \frac{1+y_1 q_1}{1-q_1}$. И также на втором участке

$$T_{обсл.2} = \frac{1+y_2 q_2}{1-q_2}$$

Повторим рассуждения, основанные на описании ретрансляционной станции в виде системы массового обслуживания.

Рассмотрим три ситуации.

1) $\tau_1 = \tau_2$

Докажем, что при равных значениях τ все будет определяться вторым участком.

$$T_{обсл.1} = \frac{1+y_1 q_1}{1-q_1} = \frac{1+\left\lceil \frac{\tau_1(k+r)}{k/N+r} \right\rceil \left(1-(1-p_1)^{k/N+r}\right)}{(1-p_1)^{k/N+r}} =$$

$$\frac{1}{(1-p_1)^{k/N+r}} + \frac{\left\lceil \frac{\tau_1(k+r)}{k/N+r} \right\rceil}{(1-p_1)^{k/N+r}} - \left\lceil \frac{\tau_1(k+r)}{k/N+r} \right\rceil$$

$$T_{обсл.2} = \frac{1}{(1-p_2)^{k/N+r}} + \frac{\left\lceil \frac{\tau_2(k+r)}{k/N+r} \right\rceil}{(1-p_2)^{k/N+r}} - \left\lceil \frac{\tau_2(k+r)}{k/N+r} \right\rceil$$

Пусть $T_{обсл.1} < T_{обсл.2}$, тогда:

$$\frac{1}{(1-p_1)^{k/N+r}} + \frac{\left\lceil \frac{\tau_1(k+r)}{k/N+r} \right\rceil}{(1-p_1)^{k/N+r}} - \left\lceil \frac{\tau_1(k+r)}{k/N+r} \right\rceil <$$

$$\frac{1}{(1-p_2)^{k/N+r}} + \frac{\left\lceil \frac{\tau_2(k+r)}{k/N+r} \right\rceil}{(1-p_2)^{k/N+r}} - \left\lceil \frac{\tau_2(k+r)}{k/N+r} \right\rceil$$

Так как $\tau_1 = \tau_2$, то

$$\frac{1}{(1-p_1)^{k/N+r}} + \frac{\left\lceil \frac{\tau(k+r)}{k/N+r} \right\rceil}{(1-p_1)^{k/N+r}} <$$

$$\frac{1}{(1-p_2)^{k/N+r}} + \frac{\left\lceil \frac{\tau(k+r)}{k/N+r} \right\rceil}{(1-p_2)^{k/N+r}}$$

Так как $p_1 < p_2$, то это утверждение верно, следовательно, верно и предположение, что $T_{обсл.1} < T_{обсл.2}$, следовательно, все будет определяться наихудшим участком, то есть вторым.

2)Если $\tau_1 < \tau_2$, то повторяя аналогичные рассуждения можно доказать что $T_{обсл.1} < T_{обсл.2}$, следовательно, все по-прежнему определяется наихудшим участком канала.

3)Если $\tau_1 > \tau_2$, то при каких-то значениях T_1 может оказаться больше чем T_2 и следовательно при выборе размера сегмента надо будет ориентироваться на канал с большим значением сигнал/шум, то есть на канал от базовой станции к ретрансляционной.

Выясним, при каких значениях p_1, p_2, τ_1, τ_2 такое возможно.

Рассмотрим конкретный пример.

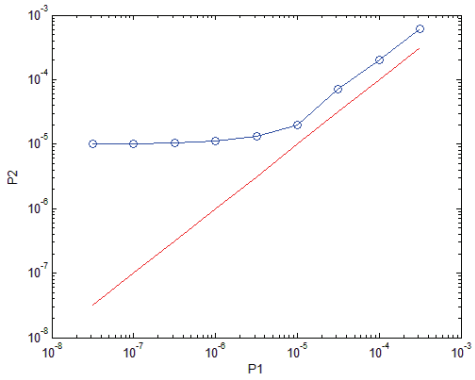


Рис. 7 - Области для выбора значений для определения числа сегментов при $\tau_1 = 2\tau_2$

Рассмотрим приведенный выше график. Нижняя прямая соответствует случаю бесконечного буфера. Участок под этой прямой нас не интересует, так как мы рассматриваем случай, когда $p_1 < p_2$. Участок между двумя прямыми относится к ситуации, когда оптимальное число сегментов определяется первым участком цепи от базовой станции к ретрансляционной. Чем больше разница между τ_1 и τ_2, тем больше эта площадь между прямыми. Площадь выше верхней прямой относится к случаю, когда оптимальный размер сегмента определяется вторым участком цепи, то есть от ретрансляционной станции к абоненту.

5. Заключение

В данной работе была разработана методика определения оптимального числа сегментов при передаче данных от абонента к базовой станции через ретрансляционную с ограниченными ресурсами на объем запоминающего устройства. Разработанная методика позволяет определять оптимальное число сегментов с учетом таких параметров физического уровня как вероятность ошибки на бит и задержка в канале связи и физическом уровне.

Данные результаты были получены для идеализированного случая, когда вероятность ошибки на бит, которую обеспечивает физический уровень, не зависит от размера сегмента. Возможным направлением развития работы является обобщение разработанной методики с учетом особенностей реализации реальных физических уровней.

6. Список используемой литературы

[1] A. Ukhanova, J.Luo, A. Turlikov “Optimal segment size for transmission in multi-hop system”, International Workshop on Multiple Access Communications

[2] J. A. C. Martins and J.C. Alves. ARQ Protocols with Adaptive Block Size Perform Better Over a Wide Range of Bit Error Rates, IEEE Transactions on Communications, Vol. 38, No. 6, June 1990

Модель изменения очереди пакетов в беспроводной сети под управлением протокола IEEE 802.16

Лукин Д.В.
Московский физико-технический институт
(государственный университет)
ldvrt@mail.ru

Аннотация

В работе рассмотрен процесс передачи данных в централизованной беспроводной сети под управлением протокола IEEE 802.16, в котором группа абонентских станций использует общий канал для передачи данных базовой станции. Исследован процесс передачи пакетов. Разработана аналитическая модель, описывающую процесс изменения суммы длин очередей зарегистрированных пакетов всех оконечных станций, и позволяющая определить вероятность пустой очереди у конкретной станции. Проведено сравнение аналитических результатов с результатами имитационного моделерования для суммарной длины очередей зарегестрированных пакетов.

1. Введение

Построение новых беспроводных сетей, является основным направлением развития сетей в настоящее время, причем развитие получили не только беспроводные локальные сети, но и городские беспроводные сети (Wireless Metropolitan Access Network, WMAN). Беспроводные городские сети призваны обеспечивать пользователям доступ в глобальные сети, то есть решать проблему последней мили, заменяя традиционные, но уже устаревшие аналоговые модемы и другие кабельные решения.

Стандарт 802.16 "Air Interface for Fixed Broadband Wireless Access Systems" (воздушный интерфейс для фиксированных систем с широкополосным беспроводным доступом), разработанный в 2001 году Институтом инженеров по электротехнике и электронике (IEEE), представляет собой рассчитанную на внедрение в городских беспроводных сетях (WMAN) технологию широкополосной связи.

Стандарт IEEE 802.16 разрабатывался с целью удовлетворения требований корпоративных пользователей в предоставлении стабильного качества обслуживания в беспроводной широкополосной сети, а также потребности домашних пользователей при работе с популярными широкополосными приложениями (в том числе и мультимедийными). Стандарт позволяет дифференцировать уровни предоставляемых услуг, что позволяет обеспечивать стабильность качества сервиса приложениям, работающим как не в реальном времени, так и в реальном времени (передача звука, видео).

Время работы сети разбивается на фиксированные интервалы, называемые кадрами, причем каждый кадр содержит восходящий и нисходящий подкадры для передачи соответствующих потоков информации.

В централизованной сети под управлением протокола 802.16 Оконечные Станции (ОС) разделяют восходящий канал к Базовой станции (БС) на основе запросов полосы пропускания (ЗПП). Это означает, что ОС информируют базовую станцию о необходимости выделения полосы пропускания в последующих кадрах с помощью отправки ЗПП. Получая ЗПП и учитывая количество данных для передачи в нисходящем канале, БС выделяет время для передачи данных в восходящем кадре каждой из станций.

Запрашиваемая полоса пропускания определяется количеством бит, необходимых для передачи MAC заголовков и пакетов с данными, без дополнительной служебной информации, добавляемой на физическом уровне (PHY Overhead).

В протоколе предусмотрены два основных метода отправки ЗПП: выделение отдельных слотов для отправки ЗПП или прикрепление ЗПП к данным.

Активными станциями считаются станции, для которых выделено время для передачи данных. Активные станции встраивают ЗПП в пакеты с данными в восходящем канале. Данный

механизм будем называть поллингом с прикреплением (piggy-backing).

Неактивные станции (станции, не имеющие слотов для передачи данных), но имеющие пакеты с данными в очереди для передачи, передают ЗПП, либо путем их одноадресного поллинга БС, либо методом конкурентного опроса. При конкурентном опросе, который смоделированв имитационной модели, для организации доступа группы абонентов к общему каналу связи применяются алгоритмы случайного множественного доступа (СМД). Базовая станция опрашивает оконечные станции, определяя временные интервалы, в которые ОС могут передавать ЗПП. Каждый кадр содержит один такой интервал, состоящий из *K* слотов, причем *K* не меняется в процессе передачи. За время одного слота возможна передача ровно одного ЗПП.

Стандартизованным в IEEE 802.16 алгоритмом СМД является “двоичный экспоненциальный откат”. После успешной отправки ЗПП на один или несколько пакетов эти пакеты считаются зарегистрированными, то есть БС может планировать их передачу. Время задержки на этапе регистрации пакетов на основе алгоритма “двоичный экспоненциальный отката” оценивалось в [3].

Так как базовая станция посылает подтверждение (ACK) на пришедшие в предыдущем кадре пакеты в управляющей секции в самом начале следующего кадра, то пренебрежем промежутком времени на посылку подтверждения и считаем моментом окончания обслуживания конец кадра, в котором пакет был получен БС. Будем говорить, что БС регистрирует пакет при получении соответствующего ЗПП. Будем считать моментом регистрации пакета конец кадра, в котором БС получила соответствующий ЗПП.

Цель данной работы состоит в разработке модели изменения очередей как пакетов так и пачек, а также в оценке времени обслуживания пакета, отсчитываемого с момента регистрации пакета в очередь до прихода к оконечной станции подтверждения (ACK) от БС о получении пакета.

Далее проведено сравнение части аналитических результатов с результатами, полученными на основе имитационного моделирования.

2. Модель изменения очереди пакетов

Рассмотрим систему массового обслуживания, описывающую процесс изменения суммы длин очередей зарегистрированных пакетов всех ОС. На вход которой подается групповой поток интенсивности $N\lambda$, а размер группы описывается геометрическим распределением с параметром q. Пусть i(t) - число пакетов в системе в момент времени t.

2.1 Общая модель изменения очереди пакетов

В момент t_v окончания кадра v число пакетов в системе претерпевает скачок, для определенности будем считать $i(t_v) = i(t_v + 0)$, то есть пакеты, переданные за этот кадр, уже не учитываются. Далее на обслуживание выбираются следующие S пакетов, которые покинут систему в конце следующего кадра t_{v+1} по приходу к ОС подтверждения (ACK) от БС о получении пакетов. Подтверждение на посланные пакеты приходит к ОС по стандарту в начале следуюшего кадра (а не в конце текущего), но этим пренебрегаем, так как в рассматриваемой модели помехи отсутствуют, поэтому время окончания обслуживания принимается за конец текущего кадра. Для того чтобы в момент t_{v+1} в системе осталось j пакетов, при условии, что в момент t_v было i пакетов, необходимо, чтобы за интервал времени (t_v, t_{v+1}) в систему поступило j-i+S пакетов.

Следовательно, переходные вероятности $p_{i,j}, i \geq S, j \geq i - S$ определяются формулой:

$p_{i,j} = f_{j-i+S}, i \geq S, j \geq i - S$, где f_k - вероятность поступления в систему k пакетов за произвольно выбранный кадр.

Пусть в момент t_v число запросов в системе $i(t_v)$ меньше S. Все эти запросы будут обработаны к моменту $i(t_{v+1})$, а вновь пришедшие запросы будут ожидать начала следующего кадра для начала обслуживания. Следовательно, чтобы в системе в момент t_{v+1} в системе осталось j запросов, за период (t_v, t_{v+1}) в систему должно прийти ровно j запросов.

$$p_{i,j} = f_j, 0 \leq i < S, j \geq i - S$$

Таким образом определены все ненулевые элементы матрицы одношаговых переходов.

$$\begin{cases} p_{i,j} = f_j, 0 \leq i < S, j \geq i - S \\ p_{i,j} = f_{j-i+s}, i \geq S, j \geq i - S \end{cases}$$

Запишем уравнения равновесия и уравнение нормировки.

$$\begin{cases} \pi_j = f_j \sum_{i=0}^{S-1} \pi_i + \sum_{i=S}^{S+j} \pi_i f_{j-i+S} \quad j \geq 0 \\ \sum_{j=0}^{\infty} \pi_j = 1 \end{cases} \quad (1)$$

Решение данной системы уравнений позволит определить среднюю длину очереди

$$L_Q = \sum_{j=0}^{\infty} j\pi_j \quad (2)$$

Найдем вероятность P_0 отсутствия зарегистрированных пакетов у данной станции в конце произвольно выбранного кадра. Для этого определим вероятность $\varphi(m \mid i)$ того, что ровно m станций являются активными при наличии i пакетов во всех очередях ОС.

Определим количество активных станций при условии того, что в очереди находится ровно i пакетов. Рассмотрим последовательность n испытаний Бернулли с вероятностью успеха p=1-q.

Определим количество испытаний, предшествующих r-му успеху. Вероятность того, что r-ый успех случится при r+k-м испытании, обозначим $\psi(k; r; p)$. Она равна вероятности того, что r-му успеху предшествует ровно k неудач.

Это событие осуществляется тогда и только тогда, когда среди r+k-1 испытаний ровно k привели к неудаче, а следующее (r+k)-е испытание привело к успеху. Соответствующие вероятности равны $C_{r+k-1}^{k} p^{r-1} q^{k}$ и p.

Таким образом, $\psi(k; r; p)$ имеет отрицательное биномиальное распределение

$$\psi(k; r; p) = C_{r+k-1}^{k} p^{r} q^{k}$$

Согласно принятой модели нагрузки, количество пакетов j в пачке имеет геометрическое распределение pq^{j-1}, то есть количество пакетов разыгрывается в испытаниях Бернулли до первого успеха, соответствующего окончанию пачки. Суммарное количество пакетов в k пачках будет равно (k+r), где r - количество неуспехов, т.е. пакетов, не являющихся последними в пачках. Заметим, что один успех (т.е. одна пачка) всегда присутствует и нужно определить вероятность k-1 успехов в i-1 испытании, которая равна $C_{r+k-1}^{k} p^{k-1} q^{r}$. Так как, k+r=i, , вероятность k пачек при наличии i пакетов равна

$$g(k \mid i) = C_{i-1}^{k-1} p^{k-1} q^{i-k}$$

Определим вероятность m активных станций при условии, что в очереди k пачек, пачки неразличимы и распределяются между станциями равновероятно:

$$h(m \mid k) = C_N^m C_{k-1}^{m-1} \Big/ C_{N+k-1}^{k}.$$

Следовательно вероятность m активных станций при условии i пакетов в очередях равна

$$\varphi(m \mid i) = \sum_{k=1}^{i} h(m \mid k) g(k \mid i) =$$

$$= C_N^m \sum_{k=1}^{i} C_{k-1}^{m-1} C_{i-1}^{k-1} p^{k-1} q^{i-k} \Big/ C_{N+k-1}^{k} \quad (3)$$

Вероятность того, что конкретная станция является неактивной при наличии m активных станций равна $\frac{N-m}{N}$. Поэтому вероятность пустой очереди конкретной станции при условии общей очереди длиной j определяется формулой

$$p(0 \mid j) = \sum_{m=1}^{j} \frac{(N-m)\varphi(m \mid j)}{N}$$

Следовательно,

$$P_0 = \sum_{j=0}^{\infty} p(0 \mid j)\pi_j = \pi_0 + \sum_{j=1}^{\infty} \sum_{m=1}^{j} \frac{(N-m)\varphi(m \mid j)\pi_j}{N}$$

Наконец, найдем среднее значение суммы длин очередей пачек зарегистрированных пакетов:

$$L_Q^b = \sum_{j=1}^{\infty} \sum_{k=1}^{j} k\pi_j g(k \mid j) =$$

$$= \sum_{j=1}^{\infty} \sum_{k=1}^{j} k\pi_j C_{i-1}^{k-1} p^{k-1} q^{i-k} \quad (4)$$

2.2 Модель изменения очереди пакетов для TDD OFDM режима.

Учтем, что среди S слотов, выделенных для передачи данных, часть слотов используются для синхронизации. ОС начинает передачу данных с посылки синхронизационной преамбулы длительностью в одни слот. Наложим ограничение, что количество слотов для передачи данных S должно быть больше удвоенного количества станций N, для того чтобы все станции имели возможность послать минимальный пакет с данными в произвольном суперкадре.

$$S \geq 2N$$

Так как каждой из m активных станций выделяется по одному слоту для отсылки преамбулы, то количество слотов для передачи данных составляет *S-m*.

Количество запросов, пришедших в систему за интервал времени (t_v, t_{v+1}) равно *j-i+s-m*, причем условная вероятность $\varphi(m \mid i)$ определяется в соотвествии с (3).

Следовательно уравнение глобального баланса, учитывая $S \geq 2N$, будет

$$\pi_j = f_j \sum_{i=0}^{S-1} \pi_i \sum_{m=1}^{\min\{i,N\}} \varphi(m \mid i) + \\ + \sum_{i=S}^{S+j} \pi_i \sum_{m=1}^{N} f_{j-i+S-m} \varphi(m \mid i)$$

Дополним систему уравнением нормировки

$$\begin{cases} \pi_j = f_j \sum_{i=0}^{S-1} \pi_i \sum_{m=1}^{\min\{i,N\}} \varphi(m \mid i) \\ + \sum_{i=S}^{S+j} \pi_i \sum_{m=1}^{N} f_{j-i+S-m} \varphi(m \mid i) \\ \sum_{j=0}^{\infty} \pi_j = 1 \end{cases}$$

Решение данной системы уравнений позволит определить среднюю длину очереди

$$L_Q = \sum_{j=0}^{\infty} j\pi_j$$

Таким образом. решив полученную систему уравнений, можно определить среднюю длину общей очереди L_Q, что позволяет оценить среднее время отправки пакета.

3. Оценка времени обслуживания зарегистрированных пакетов

Воспользовавшись формулой на Литтла находим средние времена отправки зарегестрированных пакетов $T^p{}_{omnp}$ и пачек $T^b{}_{omnp}$

$$T^p{}_{omnp} = \frac{L_Q(1-q)}{N\lambda},$$

$$T^b{}_{omnp} = \frac{L_Q^b}{N\lambda},$$

где L_Q и L_Q^b - соответственно средние длины суммарных очередей зарегистрированных пакетов и пачек пакетов на всех ОС, которые были получены в предыдущих пунктах.

4. Численные результаты

В ходе экспериментов с разработанной аналитической моделью, а также с имитационной моделью, написанной на языке GPSS, предполагалось, что размер восходящего канала S, выделенного для передачи данных, а наличие помех и, как следствие, потери пакетов не рассматриваются.

Базовая станция опрашивает оконечные станции, применяя широковещательные запросы, при которых в выделенный временной интервал ЗПП могут передавать группы оконечных станций. Количество слотов, выделенных на отправку ЗПП в конкурентном периоде передачи, равно K и не меняется в процессе передачи. Слоты для передачи ЗПП выделяются в конце суперкадра. Модель конкурентного доступа с использованием алгоритма двоичного экспоненциального отката была более подробно рассмотрена в [3]. Активные станции передают ЗПП в режиме с прикреплением (piggy-backing).

На основе численного решения системы уравнений (1) и рассчета длины суммарной очереди пакетов по формуле (2) получены графики зависимости (рис. 1) длины суммарной очереди пакетов от интенсивности группового потока.

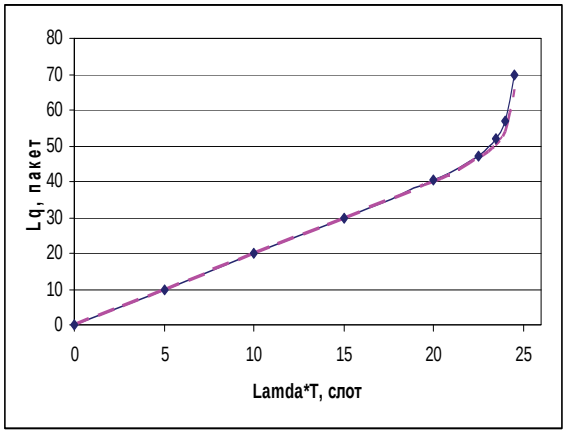


Рис. 1. Зависимость длины суммарной очереди пакетов от интенсивности группового потока.

$(N = 10,\ q = 0.5,\ S = 50)$.

—♦— – аналитическая модель

— — – имитационная модель

Сравнение с вероятностями, полученными имитационно, показало хорошую точность модели.

5. Заключение

В работе разработана аналитическая модель сети под управлением протокола IEEE 802.16, описывающая состояние суммарной очереди зарегистрированных пакетов оконечных станций . Модель позволяет определить средние длины суммарных очередей зарегистрированных пакетов и пачек пакетов. Результаты моделирования показали достаточно высокую точность предлагаемой модели.

6. Литература

[1] IEEE Std 802.16-2004 (Revision of IEEE Std 802.16-2001), IEEE Standard for Local and metropolitan area networks, part 16: Air Interface for Fixed Broadband Wireless Access Systems

[2] Вишневский В.М., Ляхов А.И., Портной С.Л., Шахнович И.В. Широкополосные беспроводные сети передачи информации. М.: Техносфера, 2005

[3] Ляхов А.И., Лукин Д.В. Оценка производительности конкурентого доступа в сети IEEE 802.16. М.: Труды семинара Распределенные Компьютерные и Телекомуникационные Сети: теория и приложения (DCCN-2007), 2007

[4] A. Vinel, Y. Zhang, Q. Ni, A. Lyakhov, “Efficient Request Mechanism Usage in IEEE 802.16,” Proc. 49th IEEE Global Telecommunications Conference (GLOBECOM 2006), San Francisco, California, USA, November 27 - December 1, 2006.

Исследование методов расчета антенн с несколькими облучателями для беспроводных сетей 802.11

И.Г. Камышников, *МФТИ*
В.М. Воробьев, *ИППИ РАН*
camish@list.ru

Аннотация

Работа посвящена исследованию методов расчета параболической антенны с несколькими облучателями как одного из элементов базовой станции с переключением диаграммы направленности во времени в сети фиксированной широкополосной беспроводной передачи данных. Приведены результаты расчета влияния на диаграмму антенны облучателей и смещения облучателей от оси симметрии.

1. Введение

В условиях дефицита частотного ресурса, роста абонентской базы и появления новых сервисов, требующих все больших полос пропускания, актуальной задачей является внедрение новейших антенных технологий.

Для базовой станции беспроводной сети передачи данных предлагается иметь только один луч в каждый момент времени приема/передачи сигнала, направленный на абонентскую станцию. Это дает возможность спроектировать недорогую антенну с диаграммой направленности, обеспечивающей широкую зону покрытия.

В настоящее время большая часть беспроводных сетей передачи данных построена на оборудовании, выполненном в соответствии с протоколами IEEE 802.11 (сети Wi-Fi). Начинают появляться сети на основе протокола IEEE 802.16 - Wi-MAX. Согласно этим протоколам, в каждый момент времени базовая станция работает (на прием или передачу) с одним из приписанных к ней абонентов. Приняв или передав ему серию пакетов, станция тут же адресуется к другому абоненту, затем к третьему и т.д. Смена абонентов в сетях 802.11 происходит путем смены заголовка (адреса) очередного пакета. При таком разделении абонентов диаграмма направленности антенны должна охватывать всех приписанных к станции абонентов. В зависимости от их расположения применяются секторные или с круговой диаграммой направленности антенны. Совершенно очевидно, что в этом случае только незначительная доля мощности сигнала, адресованного абоненту, пойдет в его направлении, остальная рассеивается впустую и создает помехи другим радиоустройствам. При приеме ситуация даже хуже: вместо того, чтобы усиливать слабый сигнал с нужного направления, антенна собирает помехи с других направлений. Таким образом, необходимо иметь антенну с узким лучом, направление которого можно задать хотя бы в азимутальной плоскости в зависимости от того, кому предназначен (или от кого ожидается) очередной пакет информации.

В настоящей работе рассматривается способ расчета антенн на примере параболической антенны, который является наиболее точным и универсальным на сегодняшний день. Параболические антенны представляют большой интерес, поскольку, с их помощью, можно реализовать наиболее дешевую, как показывают исследования, антенную систему, которая в заданном азимутальном секторе может формировать несколько узконаправленных лучей, частично перекрывающихся между собой.

2. Уравнение электрического поля антенны

На сегодняшний день из всех подходов к вычислению диаграммы направленности зеркальных антенн наиболее универсальным является использование интегральных уравнений, которые разрешаются с помощью метода «моментов» (Method of Moments). В настоящей работе рассматривается интегральное уравнение электрического поля (EFIE).

Пусть S - поверхность идеально проводящей антенны, $\mathbf{E}^i$ - поле, создаваемое некоторым источником в пространстве в отсутствии антенны. Это поле индуцирует на поверхности S поверхностный ток $\mathbf{J}$. Тогда поле $\mathbf{E}^s$, создаваемое антенной, может быть вычислено по формуле:

$$\mathbf{E}^s(\mathbf{r}) = -\frac{1}{j\omega\varepsilon}(D\mathbf{J}) \tag{1}$$

Здесь D - интегральный оператор:

$$\begin{aligned}(D\mathbf{J})(\mathbf{r}) &= \nabla\int_S G(\mathbf{r},\mathbf{r}')\nabla'_s\cdot\mathbf{J}(\mathbf{r}')dS' \\ &+k^2\int_S G(\mathbf{r},\mathbf{r}')\mathbf{J}(\mathbf{r}')dS',\end{aligned} \tag{2}$$

где $G(\mathbf{r},\mathbf{r}') = \dfrac{e^{jkR}}{4\pi R}$ - функции Грина, $R = |\mathbf{r}-\mathbf{r}'|$, $k = \omega\sqrt{\varepsilon\mu} = 2\pi/\lambda$, где λ - длина волны.

Используя формулу (1) и граничные условия на поверхности проводника $\hat{\mathbf{n}}\times(\mathbf{E}^i + \mathbf{E}^s) = \mathbf{0}$, получаем интегральное уравнение относительно $\mathbf{J}$:

$$\left(-\frac{1}{j\omega\varepsilon}(D\mathbf{J}) = -\mathbf{E}^i\right)_{\tan} \quad (3)$$

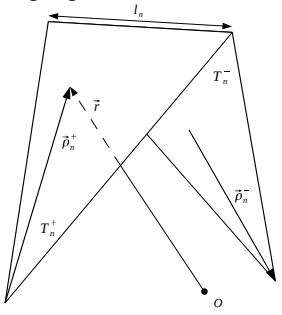
Это и есть интегральное уравнение электрического поля. Чтобы его решить методом моментов, необходимо представить неизвестный ток $\mathbf{J}$ в виде суммы известных функций с некоторыми коэффициентами. Эти функции в методе моментов называются базисными функциями:

$$\mathbf{J} = \sum_n \alpha_n \mathrm{f}_\mathrm{n} \quad (4)$$

В качестве базисных функций можно выбирать любые функции, сумма которых может приблизить решение уравнения, т.е. в нашем случае ток на поверхности проводника.

3. Метод расчета

Одним из самых главных изобретений в области расчета поля антенн и являются разработанные S. Rao, D. Wilton, A. Glisson базисные RWG функции [1-10]. Для того, чтобы использовать эти функции, поверхность проводящего тела необходимо разбить на элементарные площадки – треугольники. Каждая функция определяется на паре треугольников, имеющих общее ребро:



$$\mathbf{f_n}(\mathbf{r}) = \begin{cases} \dfrac{l_n}{2A_n^+}\rho_n^+ & \mathbf{r}\in T_n^+ \\ \dfrac{l_n}{2A_n^-}\rho_n^- & \mathbf{r}\in T_n^- \\ 0 & \text{в остальных случаях} \end{cases} \quad (5)$$

Здесь l_n – длина общего n-ого ребра треугольников T_n^+ и T_n^-, A_n^+ и A_n^- - их площади, ρ_n^+ - вектор, направленный из свободной вершины треугольника T_n^+ до произвольной точки этого треугольника, ρ_n^- - вектор, направленный к свободной вершине из произвольной точки треугольника T_n^-. Здесь можно заметить, что существует разделение треугольников на «положительные» и «отрицательные». Это деление произвольное при условии, что за полный цикл расчета антенны оно меняться не будет.

Подставляя (4) в уравнение (3) и используя определение RWG функции $\mathbf{f}_n$, получаем уравнение, в котором неизвестными становятся коэффициенты α_n. Для того, чтобы получить однозначное решение, в методе моментов проводят вторую процедуру – *testing procedure*. В качестве пробных функций будем использовать все те же RWG функции. Определив операцию умножения как

$$\langle \mathbf{f},\mathbf{g}\rangle = \int_S \mathbf{f}\cdot\mathbf{g}dS,$$

и умножив слева обе части уравнения (3) поочередно на функции $\mathbf{f}_m$, можно перейти к системе линейных уравнений относительно коэффициентов α_n:

$$-\frac{1}{j\omega e}\begin{pmatrix} \langle \mathbf{f}_1,(D\mathbf{f}_1)\rangle & \cdots & \langle \mathbf{f}_1,(D\mathbf{f}_N)\rangle \\ \vdots & \langle \mathbf{f}_m,(D\mathbf{f}_n)\rangle & \vdots \\ \langle \mathbf{f}_M,(D\mathbf{f}_1)\rangle & \cdots & \langle \mathbf{f}_M,(D\mathbf{f}_N)\rangle \end{pmatrix}\begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \\ \vdots \\ \alpha_N \end{pmatrix} = -\begin{pmatrix} \langle \mathbf{f}_1,\mathbf{E}^i\rangle \\ \vdots \\ \langle \mathbf{f}_m,\mathbf{E}^i\rangle \\ \vdots \\ \langle \mathbf{f}_M,\mathbf{E}^i\rangle \end{pmatrix} \quad (6)$$

Для краткости введем новые обозначения и перепишем (6) в более коротком виде:

$$-\frac{1}{j\omega\varepsilon}Z_{mn}V_n = -I_m \quad (7)$$

Разрешив эту систему линейных уравнений относительно неизвестных коэффициентов α_n столбца V_n, можно найти неизвестный ток $\mathbf{J}$ на поверхности проводящего тела, вызываемый полем $\mathbf{E}^i$ по формуле (4). Зная этот ток, затем по формуле (1) находится поле, рассеиваемое проводящим телом (или в нашем случае – антенной).

Таким образом, полный цикл расчета зеркальной антенны включает следующие этапы:

1. Приближение формы антенны (или антенной системы) набором треугольников
2. Формирование RWG функций по общим ребрам (нахождение общих ребер треугольников)
3. Расчет элементов Z_{mn} и I_m
4. Нахождение V_n
5. Вычисление поля антенны

Рассмотрим каждый из перечисленных этапов подробнее.

3.1 Триангуляция

Разбиение поверхности на треугольники (patch'и) следует проводить с разумной точностью. Чем

меньше характерный размер треугольников, тем точнее будет приближаемый индуцированный ток $\mathbf{J}$, а следовательно, и поле $\mathbf{E}^s$. К сожалению, выбор очень мелких треугольников неизбежно ведет к увеличению их количества, а значит, и к увеличению общих ребер и RWG функций.

Это означает, что в системе уравнений (7) растет количество неизвестных V_n и элементов матрицы Z_{mn}. Например, для частоты 2.4 ГГц ($\lambda = 12$ см) и для квадратной проводящей поверхности размерами $1\text{м} \times 1\text{м}$ требуется минимум 550 треугольников, тогда как для частоты 5 ГГц – уже более 2000. Тем не менее, нынешние персональные компьютеры могут позволить рассчитать такие объемы данных.

3.2 Расчет элементов Z_{mn}

Для расчета элементов Z_{mn} и I_m требуется уметь вычислять три типа интегралов:

$$\begin{aligned} I_1 &= \int_{S_m} \mathbf{g}_m(\mathbf{r}) \cdot \nabla \int_{S_n} G(\mathbf{r},\mathbf{r}') \nabla_S' \cdot \mathbf{f}_n(\mathbf{r}') dS' dS \\ I_2 &= \int_{S_m} \mathbf{g}_m(\mathbf{r}) \cdot \int_{S_n} G(\mathbf{r},\mathbf{r}') \cdot \mathbf{f}_n(\mathbf{r}') dS' dS \\ I_2 &= \int_{S_m} \mathbf{g}_m(\mathbf{r}) \cdot \mathbf{E}(\mathbf{r}) dS \end{aligned} \qquad (8)$$

В случае, когда точка $\mathbf{r}$ находится далеко (расстояние больше λ) от точки $\mathbf{r}'$, численное интегрирование проводится несложно с использованием Гауссовской квадратуры. В случаях же, когда $\mathbf{f}_m = \mathbf{f}_n$ или когда векторы $\mathbf{r}$ и $\mathbf{r}'$ находятся близко, в знаменателе функции Грина появляются особые точки $|\mathbf{r} - \mathbf{r}'| \to 0$, что делает обычное численное интегрирование невозможным. Такие ситуации представляют собой отдельную задачу, которая успешно решается с помощью разложения функции Грина в ряд Тейлора и использования аналитических формул [3] для интегралов вида

$$\begin{aligned} K_1^n &= \int_T R^n dS' \\ K_2^n &= \int_T R^n (\mathbf{r}' - \mathbf{q}) dS' \end{aligned}, \qquad (9)$$

на сумму которых распадаются интегралы I_1 и I_2. Кроме того, сама операция вычисления элементов Z_{mn} занимает значительную часть времени полного цикла расчета антенны.

Для того, чтобы ускорить расчет оставшихся элементов, в настоящей работе была использована система распределенных вычислений вместе с базой данных, используемой для хранения элементов Z_{mn}. Использование такой системы дает значительный выигрыш во времени расчета Z_{mn}, поскольку для вычисления интегралов (8) необходимо знать лишь координаты двух пар треугольников (всего 8 вершин), что не нагружает локальную сеть при распределении задач между компьютерами системы распределенных вычислений, тогда, как сама операция вычисления (8) занимает более продолжительное время, чем пересылка исходных данных (особенно при увеличении точности вычисляемых интегралов).

3.3 Нахождение V_n

Нахождение вектора коэффициентов поверхностных токов антенны связано с решением системы линейных уравнений (7). Одним из способов ее решения является нахождение обратной матрицы для Z_{mn}.

Вычисление матрицы Z_{mn}^{-1} осуществляется известным алгоритмом Гаусса. К сожалению, его прямое использование на больших матрицах ограничено оперативной памятью компьютера. Такое ограничение можно снять, если попытаться вычислять обратную матрицу с использованием системы распределенных вычислений, в которой элементы матрицы хранятся в базе данных.

Операции алгоритма Гаусса нахождения обратной матрицы (вычитание одной строки матрицы из другой и умножение строки матрицы на общий коэффициент) не требуют много памяти и могут быть выполнены параллельно. Например, пусть исходная матрица Z_{mn} имеет вид:

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \qquad (10)$$

Первым шагом алгоритма Гаусса является преобразование первой строки матрицы к виду $(1 \ \cdots \ \beta_{1n} \ \cdots \ \beta_{1N})$. Выполнив этот шаг, необходимо преобразовать остальные строки матрицы вычитанием из них первой строки с нужным коэффициентом таким образом, чтобы в первом столбце остались нули. Такие преобразования можно выполнять параллельно для каждой строки (или группы строк). Кроме того, выполнив преобразование второй строки таким способом, для нее можно уже выполнить первый шаг алгоритма Гаусса.

Однако, несмотря на то, что задача легко раскладывается на независимые подзадачи и удобна для использования в системе распределенных вычислений, исследования показали, что такой способ неэффективен, если в качестве места хранения элементов матрицы использовать базу данных. В самом деле, для того, чтобы полностью преобразовать один столбец матрицы к виду

$$\begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \qquad (11)$$

необходимо считать из базы все элементы, вычислить новые и сохранить их обратно. Когда речь идет о миллионах элементов, большинство баз данных не могут справиться с поиском и сохранением такого объема результатов за разумное время. К этому еще можно добавить, что после выборки необходимых элементов их необходимо пересылать на соответствующий узел (node) системы распределенных вычислений, что создает ненужную нагрузку на сеть. Иными словами видно, что сами операции преобразования строк занимают гораздо меньше времени, чем их выборка из базы данных и пересылка по сети, и еще меньше, если сравнивать со временем сохранения результата обратно в базу. В этом смысле задача о нахождении обратной матрицы представляет из себя полную противоположность задаче о вычислении элементов матрицы Z_{mn}.

Таким образом, при расчете обратной матрицы методом Гаусса стоит отказаться от попыток ускорения вычислений за счет их распределения на различные узлы системы распределенных вычислений, использующей общую базу данных для хранения элементов матрицы (если, конечно, речь не идет о специально разработанных для хранения матриц базах данных).

Систему линейных уравнений можно решить и без поиска обратной матрицы. Для этого используют итерационные методы решения систем линейных уравнений. Минусом такого подхода считается то, что ими придется пользоваться для каждого отдельного вектора I_m, т.е. для каждого нового исследования результирующего поля одной и той же антенны, находящейся в различных полях $\mathbf{E}^i$. Нахождение же обратной матрицы надо выполнить один раз, что, например, в задаче о нахождении поля зеркальной антенны позволяет свободно перемещать облучатель в любую точку пространства без дополнительных объемных вычислений.

3.4 Вычисление поля антенны

После нахождения коэффициентов V_n, поле антенны в любой точке $\mathbf{r}$ простанства вычисляется по формуле (1), в которой нам известны координаты всех RWG функций и все коэффициенты V_n. Принимая во внимание то, что эти числа остаются постоянными и что их количество гораздо меньше, чем количество элементов матрицы Z_{mn}, существует возможность использования системы распределенных вычислений для расчета поля антенны. Самым простым способом разделения зачад по узлам является разделение по параметру $\mathbf{r}$, т.е. каждый узел считает поле антенны в заданной конкретно для него точке пространства.

4. Примеры расчета поля зеркальный антенны.

Параметры антенны:

- Рабочая частота – 2.4 ГГц (длина волны $\lambda = 12$ см)
- Диаметр зеркала – 91 см
- Фокус зеркала – 56.6 см

В качестве облучателя антенны будем использовать диполь, поле которого вычисляется по формулам:

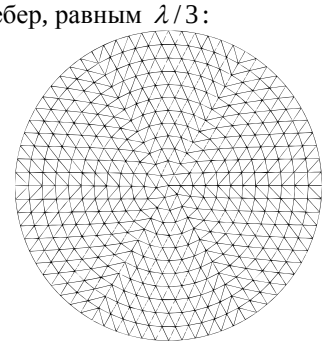
$$E_r = \eta \frac{I_0 l \cos\theta}{2\pi r^2}\left[1 + \frac{1}{jkr}\right] e^{-jkr}$$

$$E_\theta = j\eta \frac{kI_0 l \sin\theta}{4\pi r}\left[1 + \frac{1}{jkr} - \frac{1}{(kr)^2}\right] e^{-jkr}$$

$$E_\phi = 0$$

4.1 Параболическая антенна с диполем в качестве облучателя

Поверхность антенны, изображенной на рисунке выше, была разбита на 864 треугольника, со средним размером ребер, равным $\lambda/3$:

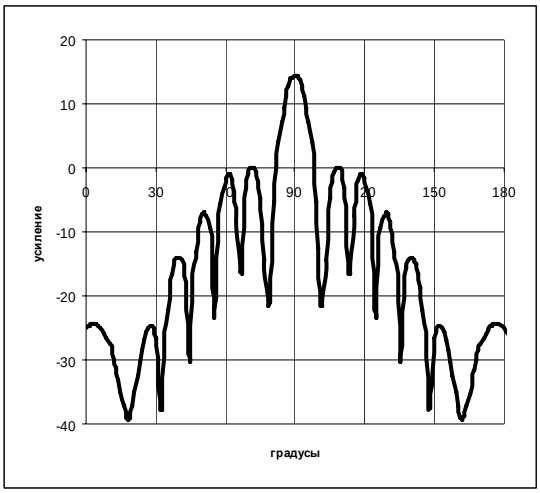


Треугольники такого разбиения имеют 1260 общих ребер. Таким образом, матрица Z_{mn} содержит 1587600 элементов. Расчет половины из них разработанным программным комплексом с использованием 3-х персональных компьютеров Pentium Core 2 Duo и 3-х Pentium IV занимает около 3.5 часов при условии, что вычисление внешних интегралов (8) происходит по сетке с размерами 10x10, а внутренних – 9x9. Иными словами, для вычисления внешнего интеграла требуется взять значение подынтегральной функции в 100 точках, каждая из которых требует вычисления значений внутренней подынтегральной функции в 81 точке. Таким образом, для вычисления одного интеграла (8) с указанной выше точностью требуется 8100 операций, не считая операций сложения и других

накладных расходов, связанных с системными вызовами выделения памяти и др. При увеличении количества точек интегрирования для достижения большей точности время вычисления сильно возрастает. Сетки с размерами 16x16 требуют уже 65536 операций, что в 8 раз больше по сравнению с указанными выше параметрами. Однако, как ранее было описано, время можно сильно сократить, добавив в распределенную систему дополнительные вычислительные мощности.

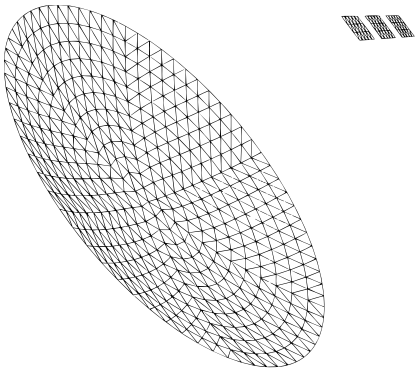
Время вычисления обратной матрицы с размерами 1260x1260 с использованием базы данных и предложенного распределенного алгоритма оценивалось в 15 дней, что объясняется огромными потерями в производительности базы данных, содержащей миллионы записей. Поскольку размеры матрицы все же не так велики, вычисление обратной матрицы производилось на одной машине и заняло менее десяти минут.

В результате после прохождения всех этапов вычислений, было рассчитано поле параболической антенны с диполем в качестве облучателя, установленным в ее фокусе.



На графике представлена зависимость усиления поля диполя в зависимости от угла, под которым видна антенна в горизонтальной плоскости. Нулевому уровню соответствует сила поля диполя без использования антенны. Видно, что «зеркало» придает полю диполя направленные свойства. Такой результат хорошо согласуется с другими методами расчета параболической антенны.

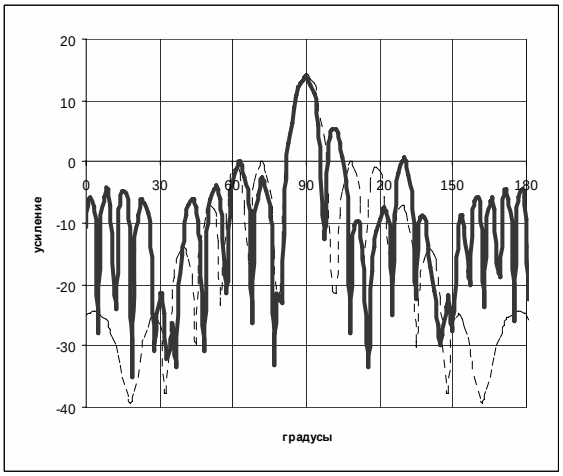
4.2 Параболическая антенна, облучаемая диполем с установленными в фокальной плоскости препятствиями



На рисунке изображена та же самая антенна, у которой в фокусе находится диполь, а рядом в фокальной плоскости расположены три дополнительные металлические пластинки размерами 6 см x 10 см. Размер пластин выбран исходя из реальных размеров облучателей – ячеек, плоских антенн. Каждая пластинка состоит из 30 треугольников. Такая система разбита на 954 треугольников, которые образуют 1371 RWG функцию.

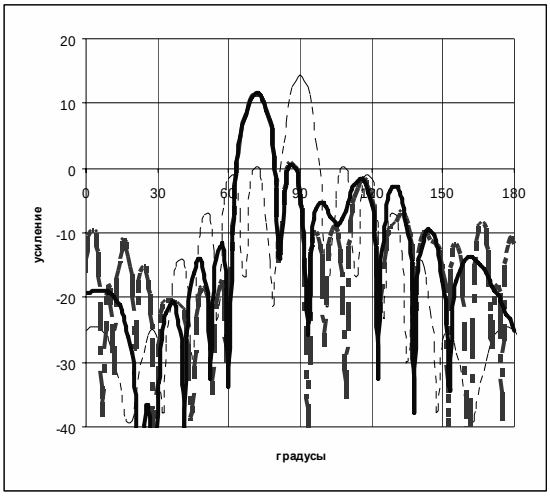
Можно заметить, что матрицу Z_{mn} предыдущей задачи достаточно модифицировать, чтобы не рассчитывать заново все элементы новой матрицы. При сохранении всех старых элементов, матрица Z_{mn} расширилась на 111 столбцов и строк (1371-1260=111), т.е. в нее добавились еще $111 \cdot 111 + 2 \cdot 111 \cdot 1260 = 292041$ элементов. Расчет их занимает менее часа при тех же настройках точности, что и в первой задаче.

На графике хорошо видно, что поле антенны изменилось. Во-первых, ослабло усиление в главном лепестке, во-вторых появилась ассиметрия поля и появились новые боковые лепестки. Такой результат показывает, что при нахождении в фокальной плоскости нескольких облучателей диаграмма направленности антенны будет видоизменяться: усиление будет падать, а боковые помехи – возрастать.



4.3 Смещение облучателя из фокуса антенны

В этом разделе рассматриваются антенны, из предыдущих параграфов, но со смещенным из фокуса облучателем на 20 см (соответствует повороту диаграммы направленности на 20 градусов)



На этом графике представлены:

- Пунктирная линия – усиление поля антенной без помех в фокальной плоскости (для сравнения)
- Сплошная линия – антенна из первого примера, но со смещенным облучателем
- Штрих-пунктирная линия – антенна из второго примера, но со смещенным облучателем

График показывает, что разницы между антеннами в главном лепестке почти нет: антенна с помехами в фокально плоскости совсем немного уступает по усилению простой зеркальной антенне; вся разница проявляется в боковых лепестках: антенна из второго примера испускает больше помех на острых углах обзора апертуры антенны.

5. Выводы

В результате исследования применения метода моментов для вычисления диаграммы направленности параболической антенны была написана программа на языке Java, позволяющая рассчитывать диаграмму направленности антенн, работающих по принципу отражения волн, в частности, параболических зеркальных антенн. При этом учитываются взаимодействия всех элементов антенны не только с внешним полем, но и между собой, что значительно повышает точность расчета поля антенны.

При реализации метода моментов были опробованы различные схемы оптимизации и повышения скорости расчета, а также системы хранения большого количества данных и их поиска: MySQL, PostgreSQL, Oracle. При расчетах применяются: Java Parallel Processing Framework для организации распределенных вычислений и Oracle в качестве базы данных, в которой хранятся все параметры системы и результаты. Использование базы данных повышает надежность разработанной программы, поскольку на уровне базы данных поддерживаются транзакционные механизмы, которые позволяют безболезненно останавливать работающую программу в любой момент без потери уже посчитанных данных.

[1] Sophocles J. Orfanidis, Electromagnetic Waves and Antennas – http://www.ece.rutgers.edu/~orfanidi/ewa, June, 2004.
[2] Constantine A. Balanis, Antenna theory: analysis and design, 2nd ed. – John Wiley & Songs, Inc, 1997 – 941 c.
[3] Pasi Yla-Oijala, Matti Taskinen, Calculation of CFIE Impendance Matrix Elements With RWG and $n \times$ RWG Functions, IEEE Trans. Antennas Propagat., vol. 51, NO 8, August 2003
[4] Sadasiva M. Rao, Donald R. Wilton, Allen W. Glisson, Electromagnetic Scattering by Surfaces of Arbitrary Shape, IEEE Trans. Antennas Propagat., vol. AP-30, NO. 3, May 1982
[5] Roger F. Harrington, Field Computation by Moment Methods – IEEE Press, 1993.
[6] W. Ewe, Le-Wei Li, Q. Wu, M. Leong, Analysis of Reflector and Horn Antennas Using Adaptive Integral Method, IEICE TRANS. COMMUN., vol.E88-B, NO.6 June 2005 P. Yla-Oijala, M. Taskinen, J. Sarvas, Surface Integral Equation Method for General Composite Metallic and Dielectric Structures with Junctions, Progress In Electromagnetics Research, PIER 52, 81-108, 2005
[7] H.T. Rathod, K.V. Nagaraja, B. Venkatesudu, N.L. Ramesh, Gauss Legendre quadrature over a triangle, J. Indian Inst. Sci., Sept.-Oct. 2004, 84, 183-88
[8] W.Cai, Y. Yu, X.C. Yuan, Singularity treatment and high-order RWG basis functions for integral equations of electromagnetic scattering, International Journal for Numerical methods in engineering, 2002, 53; 31-47
[9] Mark Vrancken, Full wave integral equation based electromagnetic modelling of 3D metallic structures in planar stratified media, December 2002
[10] Anuja Ape, Simulation on patch antennas on arbitrary dielectric substrates – RWG basis functions, May 2003

Новый алгоритм выбора скорости передачи в сетях стандарта IEEE 802.11

А.А. Адов (ИППИ РАН)
Д.Ю. Михин (ИППИ РАН)
М.В. Черкасов (ИППИ РАН)

adov@iitp.ru; mikhin@iitp.ru; max@iitp.ru

Аннотация

В данной работе описан новый, перспективный алгоритм выбора скорости в сетях стандарта IEEE 802.11, использующий данные о длинах передаваемых в сети кадров, а также учитывающий свойства радиоканала. Алгоритм был успешно реализован и применяется в реальном беспроводном оборудовании.

1. Введение

Эффективность работы беспроводной сети во многом зависит от того, как много кадров теряется при передаче. Эти потери происходят, прежде всего, по двум причинам: возникновение коллизии и искажение кадра при передаче в результате воздействия радиопомех. Вероятность искажения кадра зависит от уровня радиопомех в эфире. Таким образом, выбор эффективной скорости передачи позволяет значительно улучшить качество передачи.

2. Существующие алгоритмы

2.1. ARF

Алгоритм ARF (Auto Rate Fallback) – алгоритм с автоматическим откатом скорости. Был разработан для первых карт стандарта 802.11.

Для каждого соединения ARF следит за текущим значением скорости и количеством последовательных передач без переповторов. Алгоритм для каждого нового соединения выбирает начальную скорость максимально возможной. Дальше ARF выбирает скорость на основе следующих правил:

- снизить скорость, если подтверждение о доставке кадра не было получено
- увеличить скорость, если произошло 10 удачных передач подряд без переповторов
- иначе продолжить работу на текущей скорости

Основное преимущество этого алгоритма – его простота. ARF хорошо работает на линиях с резким затуханием. Этот алгоритм также может быстро реагировать на меняющиеся условия в линии.

2.2. AARF

AARF (Adaptive ARF [7], [8]) – адаптивный ARF, является улучшением предыдущего алгоритма. Переключение на более высокую битовую скорость после 10 удачно отправленных кадров может заметно снизить производительность, если на более высоких скоростях по линии кадры не будут передаваться совсем.
Для снижения таких задержек, в AARF каждый раз удваивается количество нужных удачных передач подряд для переключения на следующую битовую скорость, когда происходит откат скорости назад. Если же откат произошел 2 раза подряд, то значение сбрасывается в начальное значение 10 кадров.

2.3. RBAR

RBAR (Receiver-Based Auto-Rate [9]) – выбирает скорость на основе соотношения сигнал-шум. Выбирается наибольшая скорость, для которой $BER < 10^{-5}$ (Bit Error Rate).

2.4. SampleRate

SampleRate [4] – основан на выборе скорости, для которой время отправки кадра будет минимально. Для этого алгоритм периодически посылает кадры на разных скоростях, для

подсчета среднего времени отправки. Переключение на новую битовую скорость происходит, если ожидаемое время отправки будет ниже. Он перестаёт использовать битовую скорость, если произойдёт несколько неуспешных передач подряд.

3. Предпосылки для создания алгоритма

В новом алгоритме нужно учесть следующие обстоятельства:

1) Для разных длин кадров, максимальную пропускную способность канала могут обеспечивать разные скорости
2) Алгоритм должен работать хорошо при всех поведениях линии

Для разных устройств значения $rssi$ будут разными при одинаковых условиях в линии

4. Эмулятор

Для всех описанных выше случаев был написан эмулятор передачи кадров по беспроводной среде. Перед посылкой, вызывается алгоритм выбора оптимальной скорости, для кадра длины l. Результатом работы алгоритма будет некоторая битовая скорость для кадра v. Далее генерируется значение $rssi$ текущего состояния беспроводной среды, затем считается

$$BER(v) = Q_v(rssi) \quad (1)$$

где

$$Q_{v_l} = erfc\left(\sqrt{10^{\frac{rssi_v}{10}}}\right) \quad (2)$$

$$erfc(z) = \frac{2}{\sqrt{\pi}}\int_z^{\infty} e^{-t^2}dt \quad (3)$$

где $rssi_v$ выбирается для скорости v так, чтобы обеспечить одно из исследуемых поведений линии. Затем вычисляется вероятность возникновения ошибки в кадре:

$$PER(l,v) = 1-(1-BER(v))^l \quad (4)$$

На основе этой вероятности, создается биномиальная последовательность с вероятностью появления единицы равной $PER(l,v)$. Выпадение единицы означает, что кадр не был успешно отправлен (произошла ошибка в передаче и ack-кадр не был получен), выпадение нуля означает, что передача прошла успешно. На основе этой последовательности вычисляется сколько произошло переповторов, для того чтобы кадр был успешно доставлен. Параметры отправки $rssi$ и количество переповторов передаются в алгоритм выбора скорости. Затем операция повторяется для следующего кадра.

5. Новый алгоритм выбора скорости

Для достижения максимальной пропускной способности линии, алгоритм будет выбирать битовую скорость, минимизирующую время отправки кадра. Основным моментом, является адекватное измерение времени отправки кадра. При работе в режиме DCF, время отправки можно посчитать так:

$$t = difs + avgbackoff(r) + (r+1)\cdot\left(sifs + ack + header + \frac{l\cdot 8}{b}\right) \quad (6)$$

где r - количество переповторов, l - длина кадра, b - битовая скорость передачи. $avgbackoff$ - усредненное значение суммарного $backoff$ периода, предусмотренного перед отправкой кадра, можно вычислить как сумму средних значений периода. Среднее значение после i-ого переповтора задается формулой [5]:

$$backoff(i) = \begin{cases} \frac{2^i\cdot(CWmin+1)-1}{2} aSlotTime, & 0 \le i \le 6 \\ \frac{CWmax}{2} aSlotTime, & i \ge 6 \end{cases} \quad (7)$$

тогда

$$avgbackoff(r) = \sum_{i=0}^{r} backoff(i) \quad (8)$$

$difs$, $sifs$, ack, $header$, $CWmin$, $CWmax$, $aSlotTime$ являются константами физического уровня протокола 802.11, их значения перечислены в Таблице 1.

А при работе в режиме PCF, время отправки считается по формуле:

$$t = pifs + (r+1)\cdot\left(sifs + ack + header + \frac{l\cdot 8}{b}\right) \quad (9)$$

Параметр	801.11 b	802.11a/ g
$difs$, мкс	50	28
$sifs$, мкс	9	10
$pifs$, мкс	30	25/19
ack, мкс	304	200

header , мкс	192	96
aSlotTime , мкс	20	9
CWmin	31	15
CWmax	1023	1023

Таблица 1. Значения констант физического уровня 802.11

Для того чтобы воспользоваться формулой (6) или (9) мы должны знать число требуемых переповторов r. Их значение можно оценить из следующей формулы:

$$r = \frac{PER(l,v)}{1 - PER(l,v)} \qquad (10)$$

Значение $PER(l,v)$ мы можем измерять, по данным, которые передаются в алгоритм, после отправки кадра

$$PER(l,v) = \frac{N_r(l,v)}{N_t(l,v)} \quad (11)$$

где $N_r(l,v_l)$ - количество переповторов, произошедших при отправке кадра длины l, на скорости v, а $N_t(l,v) = N_r(l,v) + N_s(l,v)$ - суммарное количество передач (сумма количества переповторов и успешных передач).

В данном случае мы имеем дело с биномиальным распределением. Чтобы измерить $PER(l,v)$ с нужной точность, воспользуемся доверительным интервалом для биномиального распределения. Обозначим $\tilde{p} = PER(l,v)$. Тогда классический доверительный интервал для p - успеха в одном испытании имеет вид:

$$\tilde{p} - u_{1-a/2}\sqrt{\frac{\tilde{p}\cdot(1-\tilde{p})}{N_t}} \le p \le \tilde{p} + u_{1-a/2}\sqrt{\frac{\tilde{p}\cdot(1-\tilde{p})}{N_t}} \quad (12)$$

где

$$u_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2),$$

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}dt$$

- функция распределения нормальной случайной величины при $\mu = 0$, $\sigma = 1$. При $\alpha = 0.95$, $u_{1-\alpha/2} \approx 1.96$. Основным недостатком этого интервала является его неприменимость при $\tilde{p} \to 0$ и $\tilde{p} \to 1$, а также реальная частота попадания в интервал даже при больших N_t редко превосходит α [6]. Так как для различных l, v значение $PER(l,v)$ может стремиться и к 0 и к 1, то эта интервальная оценка нам не подходит. В нашем случае хорошо подходит интервал Агрести-Коула (АК), благодаря своей простоте вычисления и адекватном поведении при $\tilde{p} \to 0$ и $\tilde{p} \to 1$[6].

АК интервал имеет такой же вид, как стандартный:

$$\tilde{p} - u_{1-a/2}\sqrt{\frac{\tilde{p}\cdot(1-\tilde{p})}{\tilde{N}_t}} \le p \le \tilde{p} + u_{1-a/2}\sqrt{\frac{\tilde{p}\cdot(1-\tilde{p})}{\tilde{N}_t}} \quad (13)$$

где $\tilde{N}_t = N_t + u_{1-\alpha/2}^2$, а $\tilde{p} = \frac{N_r + \frac{u_{1-\alpha/2}^2}{2}}{\tilde{N}_t}$. Можно считать, что PER будет посчитан с нужной точностью, когда выполняется неравенство:

$$u_{1-a/2}\sqrt{\frac{\tilde{p}\cdot(1-\tilde{p})}{\tilde{N}_t}} \le \varepsilon \quad (14)$$

В алгоритме значения $\alpha = 0.95$, $\varepsilon = 0.05$.

Далее, чтобы алгоритм выбора минимального времени начал работать, необходимо чтобы у нас уже были посчитаны с нужной точностью значения $PER(l,v_1)$, $PER(l,v_2)$, $PER(l,v_3)$ и т.д. Где l - длина кадра, v_1, v_2, v_3 - различные битовые скорости. Для ускорения набора такой статистики, во-первых, можно группировать кадры с близкой длиной. То есть разбить все возможные длины кадров на K интервалов, и считать что кадры, попадающие в один интервал имеют равную длину. Во-вторых, посылка кадров на скоростях, на которых передача заведомо не возможна, приведет к бесполезной занятости среды. Но с другой стороны следует проверить эти скорости, чтобы убедиться, что они совсем не подходят. Для разрешения этой ситуации можно использовать следующую схему:

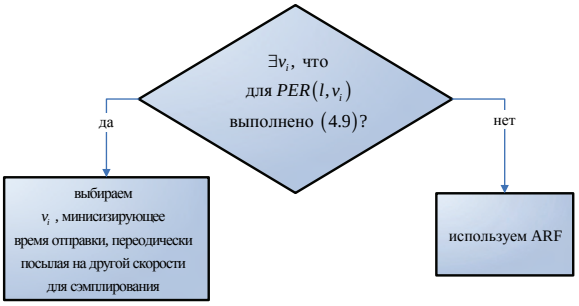


Схема 1.

Это обеспечивает, что в начале набора статистики новый алгоритм будет работать не хуже чем ARF. Затем, когда статистика будет набрана, алгоритм будет обеспечивать максимальную пропускную способность.

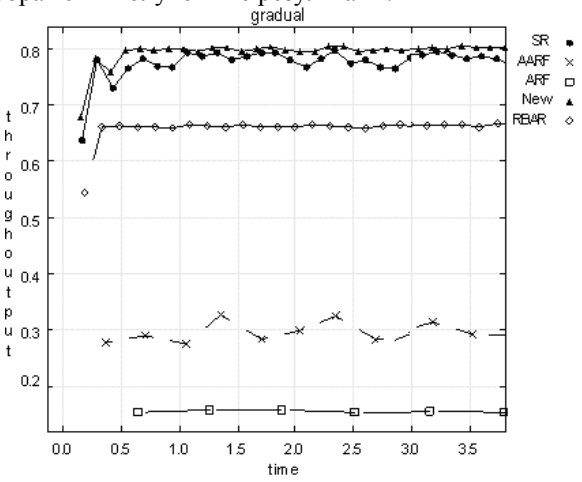
Далее при изменяющихся условиях среды (изменяется $rssi$), нужно пересчитать набор $PER(l,v)$ по Схеме 1 для новых значений $rssi$. При каждом вызове алгоритма, мы не знаем текущее $rssi$ (значение $rssi$ мы узнаём после отправки). Поэтому необходим механизм предсказания поведения среды. В качестве такового был выбран метод экспоненциально-взвешенных скользящих средних – ewma (exponentially weighted moving average), в котором предсказанное значение на k-ом шаге считается по следующей формуле:

$$predictedRssi_k = \lambda \cdot lastRssi + (1-\lambda) \cdot predictedRssi_{k-1} \quad (15)$$

где *lastRssi* - значение $rssi$ при предыдущей отправке. Для возможности быстро среагировать на изменение $rssi$ значение λ взято равное 0.3.

6. Заключение

Для получения результатов, в эмуляторе посылалось 1000000 пакетов, измерялось время, затраченное на отправку и средняя скорость для 20000 пакетов. На следующих рисунках отображены полученные результаты:



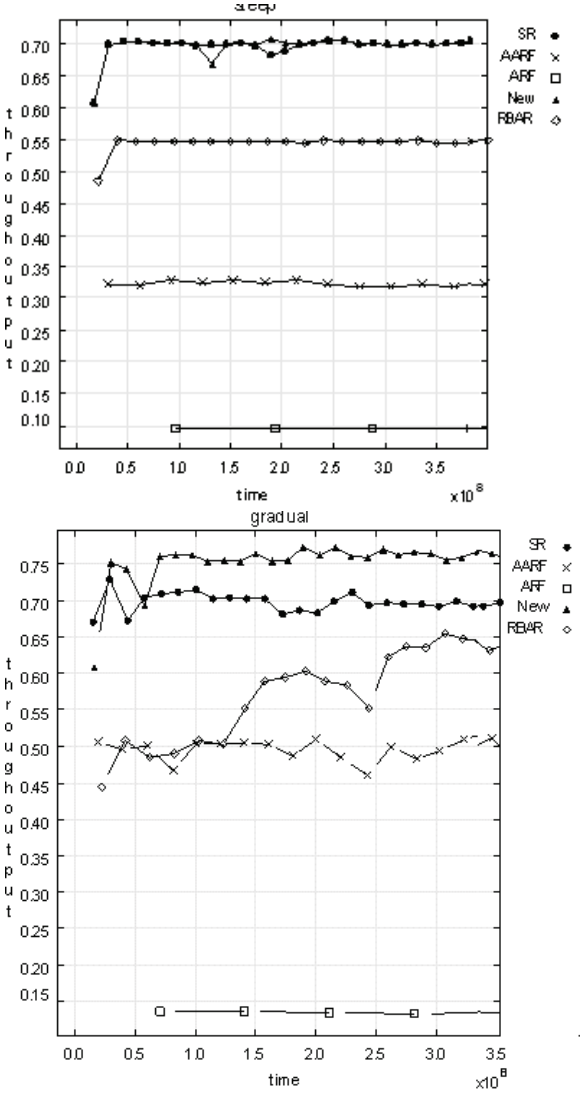


Рисунок 1. Графики производительности, при работе различных алгоритмов выбора скорости, на линии с постепенным затуханием.

Полученные результаты и практическая реализация алгоритма в драйвере беспроводного устройства показало его эффективность над существующими алгоритмами.

6. Список литературы

[1] Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification. IEEE Standard 802.11. IEEE Press, 1997.

[2] А.И.Ляхов, Д.Н.Манцев, Д.В.Лаконцев, О.Н.Шелихов «Сбор и анализ характеристик функционирования действующей беспроводной сети на основе протокола IEEE 802.11», Москва, Россия.

[3] Д.В.Лаконцев «Исследование эффективности механизмов адаптации протокола IEEE 802.11 к реальному беспроводному трафику,

и методы оптимизации работы радиосети.», Москва, 2002г.

[4] John Bicket «Bit-rate Selection in Wireless Networks». Master's thesis, Massachusetts Institute of Technology, February 2005.

[5] D. Qiao and S. Choi, «Goodput Enhancement of IEEE 802.11a Wireless LAN via Link Adaptation» in ICC, June 2001.

[6] Lawrence D. Brown, T. Tony Cai and Anirban DasGupta «Interval Estimation for a Binomial Proportion», Statistical Science 16, iss. 2 (2001), 101–133

[7] Mohamed Naimi, Michele Zorzi «CROSS LAYER INTERACTIONS FOR ADAPTIVECOMMUNICATIONS IN IEEE 802.11 WIRELESS LANS», France, INRIA 2005

[8] Lacage, Manshaei, Turletti «IEEE 802.11 Rate Adaptation: A Practical Approach», France, INRIA, 2004

[9] Vaidya, Bahl, «A Rate-Adaptive MAC Protocol for Wireless Networks» - Holland, 2000

[10] Douglas S. J. De Couto, Daniel Aguayo, John Bicket, and Robert Morris «A high-throughput path metric for multi-hop wireless routing» In Proceedings of the 9th ACM International Conference on Mobile Computing and Networking (MobiCom '03), San Diego, California, September 2003

[11] Mark Yarvis, W. Steven Conner, Lakshman Krishnamurthy, Jasmeet Chhabra, Brent Elliott, and Alan Mainwaring «Real-world experiences with an interactive ad hoc sensor network» In Proceedings of the International Workshop on Ad Hoc Networking, August 2002

Моделирование и анализ протокола маршрутизации HWMP стандарта IEEE 802.11s

Кирилл Андреев
ИППИ РАН
kirillano@yandex.ru

Дмитрий Лаконцев
ИППИ РАН
lakontsev@iitp.ru

Андрей Мазо
ИППИ РАН
ahippo@yandex.ru

Аннотация

Рассматривается беспроводная mesh-сеть, построенная на базе стандарта IEEE 802.11s (Mesh Networking). Сеть состоит из однородных узлов, каждый из которых выполняет функции маршрутизатора; сеть не имеет какого-либо фиксированного транспортного ядра. Основная цель работы – разработка имитационной модели протокола маршрутизации HWMP (Hybrid Wireless Mesh Protocol) и исследование на её основе различных режимов работы протокола при различных размерах топологии. В результате исследования получены численные оценки показателей производительности работы mesh-сети в различных режимах.

1. Введение

В связи с разработкой стандарта IEEE 802.11s для беспроводных mesh-сетей моделирование протокола Hybrid Wireless Mesh Protocol (HWMP), являющегося обязательным для реализации согласно этому стандарту, представляет большой интерес.

На данный момент публично-доступных реализаций протокола HWMP для среды имитационного моделирования Network Simulator версии 2 (NS-2) нет. Точнее, нам известна одна реализация протокола HWMP, но она является фактически модификацией реализации протокола AODV (Ad-hoc On-Demand Distance Vector Routing) и выполнена на третьем уровне модели OSI, вместо второго. Кроме того, её статус неизвестен.

В связи с вышесказанным, было принято решение о разработке реализации протокола HWMP для среды имитационного моделирования NS-2 и исследование на её основе различных режимов работы протокола HWMP.

2. Имитационная модель

2.1. Допущения, касающиеся MAC-уровня 802.11s.

Основной составляющей реализации mesh-сетей является организация логических каналов передачи данных (peer-link), в то время как реализация уровня доступа к среде в среде имитационного моделирования NS-2 позволяет организовать лишь структуру ad-hoc сети без логических каналов.

Вместе с этим базовая скорость и скорость, на которой передаются одноадресные пакеты, различаются, потому метрики могут оказаться резко несимметричными по каналу. А в случае, когда в сети передаются одновременно одноадресные и многоадресные маршрутные пакеты (что наиболее характерно для реактивного режима с DO = 0), одноадресные PREP с большей вероятностью терпят коллизию, что существенно ухудшает время доставки PREP-сообщений, поскольку время передачи широковещательного сообщения существенно возрастает по сравнению со временем передачи одноадресного сообщения.

Передача данных на MAC-уровне осуществляется посредством механизма RTS/CTS (Ready-to-send/Clear-to-send), что также несколько ограничивало возможности и вносило коррективы в результаты симуляции.

Поскольку нет возможности выяснить долю ошибок в передаваемых пакетах на MAC-уровне, то в метрике, предложенной стандартом, ошибки передачи не рассматриваются. Поскольку между всеми станциями скорости одинаковы, то в метрике фигурируют только константы. Более того, при только одном потоке данных и механизме RTS/CTS коллизии сводятся к минимуму.

Стандарт 802.11s предполагает 6-адресную схему для фреймов, в то время как реализация MAC-уровня беспроводных сетей в среде моделирования NS-2 предполагает 3-х адресную схему.

2.2. Параметры и допущения в модели HWMP

В модели HWMP максимальное число попыток установить маршрут равняется 10. Вместо экспоненциального двоичного отката используется линейное увеличение интервала между двумя последующими запросами после неудачной попытки установить маршрут. Один временной слот выбирался равным 0.1 секунды, однако реально этот параметр практически не влиял на производительность сети, поскольку на четвертом уровне использовался протокол TCP. В случае, когда на верхних уровнях происходит генерация трафика с постоянной скоростью, этот параметр может существенно повлиять на работу сети.

В связи с использованием статических топологий при моделировании, реализация протокола HWMP в среде NS обошла стороной пункт стандарта IEEE802.11s о рассылке сообщений о потере маршрута Path Error (PERR), в которых не было необходимости. Потому потеря маршрутной информации является следствием исключительно ее устаревания.

В рамках данной модели интересно проследить поведение проактивного режима протокола HWMP при наличии нескольких корневых станций. Так как в черновике стандарта IEEE 802.11s такой режим вообще не оговаривается, пришлось сделать несколько дополнительных предположений о поведении станций в таком режиме. Основным предположением являются условия приёма проактивных Path Request PREQ от нескольких корневых станций и их обработка. В результате был выбран способ полностью игнорирования и отбрасывания проактивных PREQ от второй и последующих корневых станций.

2.3. Итоговые допущения

Подытожим вышесказанное относительно параметров и допущений построенной модели:

- механизм RTS/CTS
- скорость передачи данных одинаковая для всех станций
- отсутствие 6-адресной схемы
- метрика не учитывает потери пакетов
- отсутствие механизма Root Announcement (RANN) в реактивном режиме
- отсутствие сообщений Path Error (PERR)
- приём Path Request (PREQ) только от первой корневой станции

3. Методика моделирования

3.1. Используемые топологии

При моделировании использовались топологии вида, изображённого на рис. 1. Передача данных производилась между станциями, отмеченными жирными точками. Область слышимости показана возле левой нижней станции.

Корневой станцией была станция, обведённая в кружочек. Для режима с двумя корневыми элементами вторая корневая станция выбиралась симметрично относительно центра.

Расстояния между станциями выбирались равными примерно 200 метров

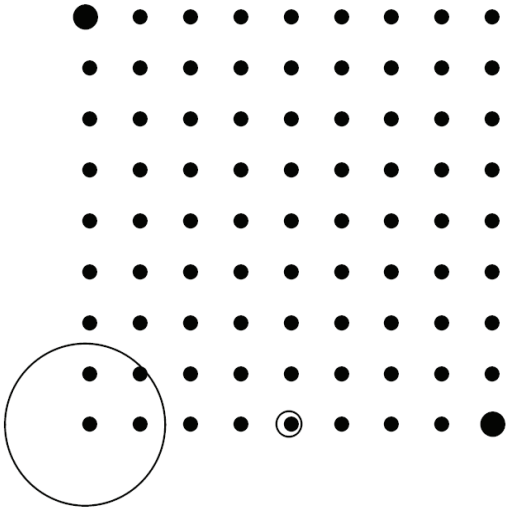


Рис. 1. Топология сети

3.2. Варьируемые параметры

При моделировании общая структура топологии и модель трафика не менялись, а менялись лишь некоторые их параметры:

- размер торологии
- режим работы протокола HWMP
  - проактивный с 1 корнем
  - проактивный с 2 корнями
  - реактивный с DO = 0, RF = 0
  - реактивный с DO = 0, RF = 1
  - реактивный с DO = 1
- базовая скорость
  - равная скорости передачи данных (54 Мбит/с)
  - неравная скорости передачи данных (6 Мбит/с)

3.3. Измеряемые параметры

В качестве основных параметров производительности были выбраны следующие величины:

- средняя скорость передачи данных
- среднее время доставки сообщения
- средняя длина маршрута
- число широковещательных служебных сообщений (точнее, PREQ) в расчёте на одну станцию

4. Результаты моделирования

Для повышения наглядности представления результатов, все сравниваемые режимы разделены на 2 группы:

- проактивные режимы, реактивный режим с DO = 0 RF = 0
- реактивные режимы

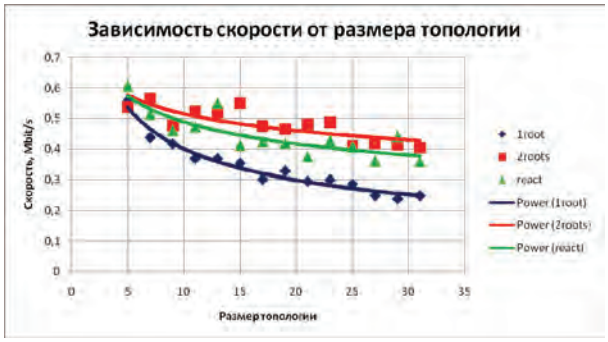


Рис. 2. Сравнение скорости для реактивного и проактивных режимов.

Из Рис. 2 видно, что наилучшую скорость для всех имеет экспериментальный режим с двумя корневыми станциями, за ним следует реактивный режим и худшую скорость имеет обычный проактивный режим.

Такое поведение режима с двумя корневыми станциями можно объяснить прохождением пакетов RTS/CTS в прямом и обратном направлениях разными путями, что значительно освобождает среду для полезных данных. (заметим, что обратный поток RTS/CTS вызван наличием подтверждений (ACK, acknowledgment) в протоколе TCP, переносящем полезные данные).

Реактивный режим оказывается лучше по сравнению с обычным проактивным, так как довольно заметно отличие средних приведённых длин маршрута, что видно из Рис. 6. Помимо этого в проактивном режиме периодически инициируется волна проактивных PREQ, на которую через некоторое время приходят PREP, причём примерно в одно время. Из-за этого есть периоды времени, когда корневая станция занимается практически только приёмом PREP и скорость полезных данных падает почти до нуля.

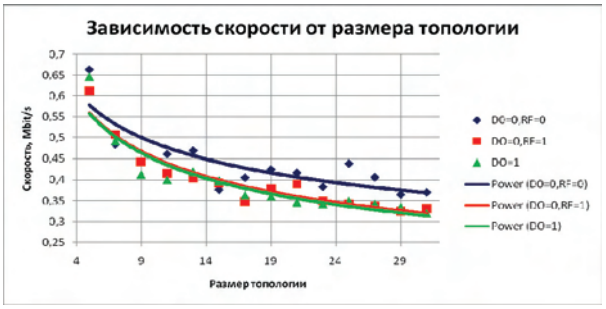


Рис. 3. Сравнение скорости для различных реактивных режимов.

Из Рис. 3 видно, что наилучшую скорость имеет режим с DO = 0 RF = 0, а различия между режимами DO = 0 RF = 1 и DO = 1 нет.

Преимущество режима DO = 0 RF = 0 объясняется тем, что на выяснение маршрута тратися существенно меньше времени, чем в других режимах, что также видно из Рис. 5.

Отсутствие различия между двумя другими режимами можно объяснить тем, что PREQ с DO = 0 RF = 1 переходит в PREQ с DO = 1 после первой станции, знающей маршрут до станции назначения.

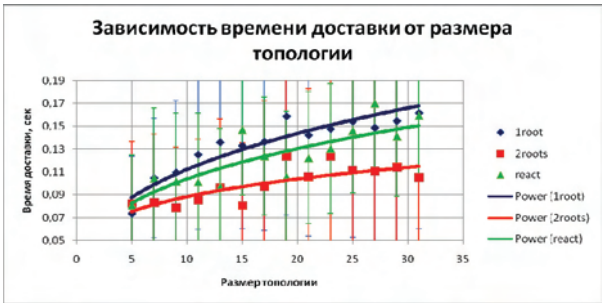


Рис. 4. Сравнение времени доставки сообщения для реактивного и проактивных режимов.

Из Рис. 4 видно, что время доставки сообщения минимально для проактивного режима с двумя корневыми станциями.

Это связано с меньшим временем установления маршрута по сравнению с другими режимами, а также эффектом, связанным с RTS/CTS, обсуждавшемся выше.

Выигрыш реактивного режима связан с лучшим выбираемым маршрутом по сравнению с проактивным режимом с одним корневым элементом, что видно на Рис. 6.

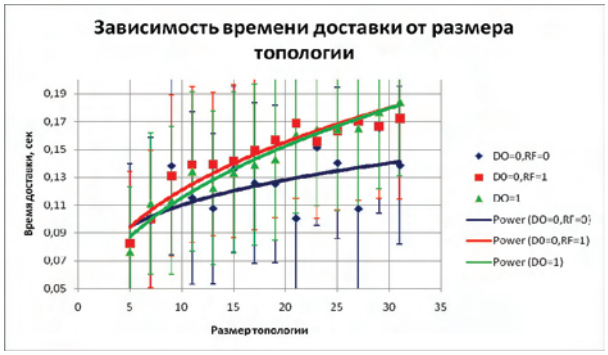


Рис. 5. Сравнение времени доставки сообщения для различных реактивных режимов.

Из Рис. 5 виден выигрыш режима DO = 0 RF = 0 и отсутствия различий между DO = 0 RF = 1 и DO = 1.

Это объясняется тем, что нахождение маршрута для режима DO = 0 RF = 0 происходит быстрее всего.

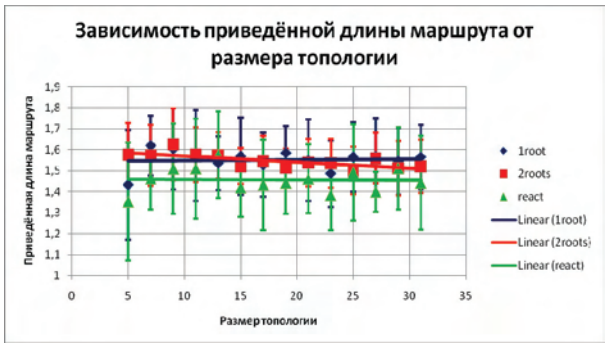


Рис. 6. Сравнение длины маршрута для реактивного и проактивных режимов.

Из Рис. 6 видно, что реактивный режим выбирает маршрут, наиболее близкий к идеальному, причём для топологии любого размера. Оба проактивных режима выбирают маршрут примерно одинаково отличающийся от идеального. Это связано с тем, что часть пути всегда идёт по дереву, а не по идеальному пути.

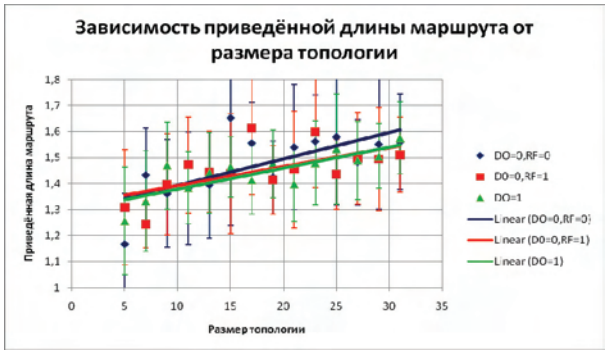


Рис. 7. Сравнение длины маршрута для различных реактивных режимов.

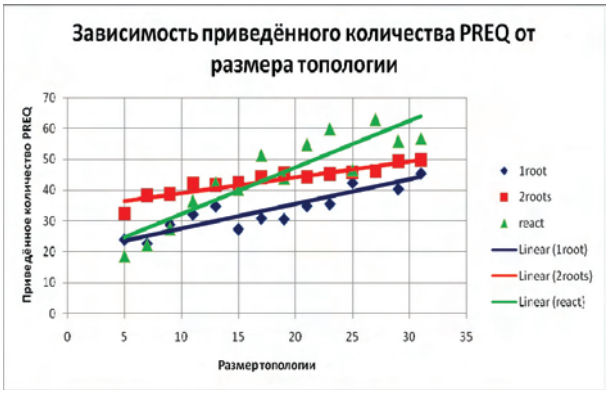


Рис. 8. Сравнение количества служебных сообщений для реактиного и проактивных режимов.

Из Рис. 8 видно отличие в поведении рассмотренных режимов для больших и маленьких топологий.

Сравним сначала режимы на небольших топологиях. Как видно, количество PREQ в реактивном режиме оказывается меньше, чем в проактивном, и уж тем более в проактивном с двумя корневыми станциями. Это связано с тем, что проактивные PREQ рассылаются принудительно по таймеру, а реактивные рассылаются только по мере надобности. В режиме с двумя корневыми станциями принудительно рассылается ещё больше проактивных PREQ, так как это делает каждая из двух корневых станций.

На больших топологиях ситуация иная. Количество PREQ в режиме с двумя корневыми станциями несильно увеличилось по сравнению с небольшими топологиями, так как количество принудительно рассылаемых PREQ осталось таким же (однако возросло количество реактивных PREQ в этом режиме, необходимых для соединения двух деревьев).

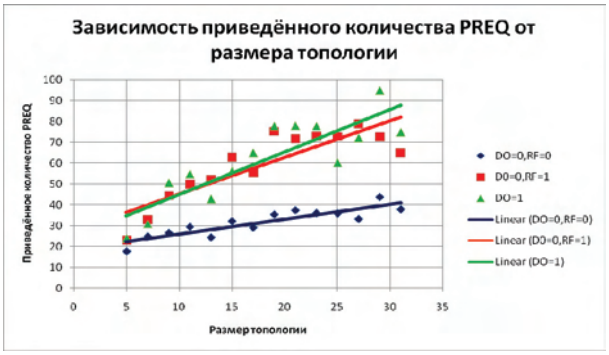


Рис. 9. Сравнение количества служебных сообщений для различных реактивных режимов.

Из Рис. 9 видно практически двукратное увеличение количества PREQ для режимов DO = 0 RF = 1 и DO = 1 по сравнению с DO = 0 RF = 0 для всех размеров топологий.

Рассмотрим теперь влияние изменения базовой скорости на измеряемые параметры. Уменьшение базовой скорости с 54 Мбит/сек до 6 Мбит/сек повлекло за собой незначительное снижение средней скорости передачи данных (хотя, казалось бы, увеличение времени выяснения маршрута вряд ли должно было бы повлиять на скорость передачи, поскольку оно мало по сравнению со временем устаревания маршрутной информации). Приведённое число узлов и количество служебных широковещательных запросов изменилось незначительно, но все же в пользу одинаковых скоростей (обратим внимание, что эти два параметра не зависят явно от скорости передачи данных). Однако понятно, что использование одинаковых скоростей для всех маршрутных фреймов более предпочтительно.

5. Пояснение аппроксимаций

5.1. Скорость передачи данных — степенная зависимость

Ожидаемая зависимость средней скорости передачи данных TCP-соединения от размера топологии при условии неизменного размера окна — обратно пропорциональная, поскольку чем больше число узлов, тем меньше скорость.

Связано это с тем, что при неизменном размере окна следующий пакет не будет отправлен, пока не придет подтверждение от одного из предыдущих (какого именно – зависит от размера окна), а время распространения одного пакета пропорционально длине маршрута. Следовательно, ожидаемая зависимость — линейная.

Однако размер окна не является постоянным, в сети одновременно могут присутствовать несколько пакетов, между которыми “расстояние” по числу узлов определяется областью слышимости и механизмом RTS/CTS. Поскольку этот механизм одинаков для всех участков сети, размер окна может увеличиваться пропорционально максимальному допустимому числу пакетов на маршруте. В таком случае скорость практически не должна будет изменяться.

Исходя из данных предположений, при аппроксимации выбрана степенная зависимость. Понятно, что показатель степени должен быть по модулю меньше единицы. Измеренные значения показателей находятся около значения -0.25. Объяснить полученную зависимость довольно тяжело, поскольку мы ничего не можем сказать о реализации TCP в NS-2.

5.2. Время доставки сообщения — степенная зависимость

Ожидаемая зависимость среднего времени доставки сообщения от размера топологии – линейная. Однако, из общего вида графиков зависимости видно, что рост времени доставки сообщения в зависимости от размера топологии происходит медленнее. Вероятно, это связано опять же с использованием TCP — размер окна увеличивает “расстояние”, выраженное в числе узлов, между двумя пакетами в сети, а потому пакет с меньшей вероятностью будет остановлен на пути следования из-за занятости среды. Потому также возможно отклонение от линейной зависимости в лучшую сторону. Надо отметить, что незначительное увеличение “расстояния” может привести к существенному снижению вероятности блокировки пакета. Исходя из вышесказанного, выбрана степенная зависимость. Показатели степеней оказались в промежутке от 0.25 до 0.45.

5.3. Приведённая длина маршрута — линейная зависимость

Ожидаемая зависимость приведенной длины маршрута от размера топологии — линейная, поскольку ожидается, что по мере увеличения топологии среднее число узлов, отнесенное к максимальному должно изменяться незначительно. Потому при линейной аппроксимации все коэффициенты наклона оказались небольшими. В данном случае аппроксимация является весьма условной, поскольку дисперсия числа узлов довольно высока, и избавиться от этого можно только посредством введения метрики, учитывающей прошлые маршруты.

5.4. Приведённое количество служебных сообщений — линейная зависимость

Ожидаемая зависимость общего количества широковещательных служебных сообщений на одну станцию от размера топологии — линейная.

Связано это с тем, что количество пакетов, находящихся в сети в одно время, пропорционально размеру топологии с незначительными отклонениями. Поскольку маршрутная информация устаревает одновременно для всех маршрутов, установленных в результате одной процедуры установления маршрута (вспомним, что время жизни таблицы записано в PREQ или PREP – является общим для всех), то в один момент времени происходит генерация запросов в

количестве, пропорциональном размеру топологии. Поскольку топологии статические, то устаревание происходит только по таймауту, который общий для большинства станций. Далее каждая инициализация запроса приводит примерно к одному PREQ от каждой станции.

6. Заключение

В рамках данной работы построена имитационная модель протокола маршрутизации HWMP в среде имитационного моделирования NS-2. В дальнейшем возможно усовершенствование данной модели, включение в неё всех сознательно пропущенных на данном этапе возможностей, перенос модели на готовящуюся к выходу среду имитационного моделирования NS-3. Отметим также, что построенная модель будет свободно доступна под лицензией GPLv2 (GNU Public License) (или совместимой с ней) и скорее всего войдёт в будущие версии среды имитационного моделирования NS.

Проанализированные режимы работы протокола маршрутизации HWMP для статических топологий явно показали, что определение времени жизни маршрутной информации по одной общей метки некорректно, поскольку одновременное устаревание маршрутной информации сразу на всех станциях существенно усложняет процесс установления маршрута.

7. Список литературы

[1] Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications. ANSI/IEEE Std. 802.11, 1999 Edition

[2] Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: Mesh Networking. IEEE Draft Std. 802.11s/D1.07, September 2007

[3] *K. Fall* (editor), *K. Varadhan* (editor), "The *ns* Manual (*ns* Notes and Documentation)", The VINT Project, January 2008

[4] *E. Altman, T. Jimenez*, "NS Simulator for beginners", Lecture notes, 2003-2004, Univ. de Los Andes, Merida, Venezuela and ESSI, Sophia-Antipolis, France, December 2003

[5] *F. J. Ros, P. M. Ruiz*, "Implementing a New Manet Unicast Routing Protocol in NS2", Dept. of Information and Communications Engineering, University of Murcia, December 2004

[6] *В. Вишневский, Д. Лаконцев, А. Сафонов, В. Шпилев*, "Mesh-сети. В ожидании стандарта 802.11s": Первая Миля, 2-3 2008

Эффективность применения MIMO в беспроводной городской сети передачи данных

Жилин М.С., *МФТИ*
Воробьев В.М., *ИППИ*
E-mail: mizhka@gmail.com

Аннотация

В работе рассмотрены вопросы определения производительности MIMO каналов передачи информации. Разработан алгоритм для определения параметров MIMO канала беспроводной сети передачи данных на базе детерминистической модели в условиях городской среды.

1. Введение

MIMO-система независимой обработки переотраженных декоррелированных сигналов - технология, нацеленная увеличить пиковую скорость передачи трафика, среднюю скорость передачу данных и пропускную способность сот в широкополосных беспроводных сетях. В концепции MIMO подразумевается использование нескольких приемо-передающих антенн.

Основная идея MIMO - разбиение потока транслируемых данных между независимыми приемопередатчиками, обеспечивающими связь для одного и того же абонента на одной и той же частоте. В этом случае, по сравнению с обычной системой, совокупный поток данных может быть увеличен. Поскольку для потоков данных используется одна и та же частота в одном географическом регионе, потоки необходимо декоррелировать. Для этого используется либо пространственное разнесение антенн, либо использование разных поляризаций антенны для разных трактов. Соответственно, требуется наличие более одного (обычно двух) приемо-передающих трактов в составе абонентского оборудования. Преимущества от использования MIMO:

- улучшается работа нисходящих линий связи, что проявляется в виде более стабильной и более надежной передачи данных в условиях сильного рассеивания радиоволн. Для восходящих линий связи улучшений нет.
- увеличивается пропускная способность MIMO за счет возможности использования схем модуляции более высокого уровня при одновременной передаче нескольких независимых потоков,
- увеличивается стабильность рабочих характеристик MIMO систем, особенно при перемещениях абонента.

2. Распространение радиосигнала

Принятый сигнал является суперпозицией полезного сигнала и шума. Для того, чтобы можно было отделить полезный сигнал от шума, требуется, чтобы принятый сигнал имел мощность большую заранее заданной TH_{Rx}. Затухания сигнала в среде L_{Path} определяются параметрами среды и расположением источника и приемника. Источник излучает электромагнитные волны с эффективной изотропно излучаемой мощностью EIRP. Антенна позволяет усилить сигнал в определенном направлении с коэффициентом усиления G_{Rx}. Все упомянутые характеристики удовлетворяют следующему уравнению баланса [1]:

$$\mathrm{EIRP} - L_{path} + G_{Rx} - \mathrm{TH}_{Rx} = LM \qquad (1)$$

EIRP - эффективная изотропно излучаемая мощность (в dBW или dBm),
L_{Path} - суммарные потери при прохождение сигнала от антенны к антенне.,
G_{Rx} - коэффициент усиления антенны,
TH_{Rx} - минимальный уровень сигнала,
LM - запас мощности, часть которого определяет соотношение сигнал-шум.

3. Двунаправленная модель распространения

Двунаправленная модель необходима для рассмотрения радиоканалов с пространственно-

временным кодированием. Радиоволны делятся на отдельные группы, каждая из которых проходит отдельный путь, отличный от другой группы. Каждая группа волн характеризируется двумя пространственными углами: приема и передачи. Данная модель позволяет отделить влияние антенн от влияния среды на характеристики радиочастотного канала [2]. Рассмотрим канал между передатчиком в точке r_{Tx} и приемником в r_{Rx}. Передаточная функция канала зависит от значений угла передачи ϕ и угла приема ψ. Она представима в виде суммы отдельных передаточных функций для отдельных групп.

$$SNR = SNR(EIRP) \quad (2)$$

В радиочастотном канале можно менять только мощность излучаемую, поскольку все остальные параметры задаются либо средой, либо характеристиками оборудования. Таким образом, соотношение сигнал-шум является монотонно возрастающей функцией от значения EIRP.

$$\mathrm{h}(\mathrm{r}_{\mathrm{Tx}}, \mathrm{r}_{\mathrm{Rx}}, \tau, \varphi, \psi) = \sum \mathrm{h}_1 (\mathrm{r}_{\mathrm{Tx}}, \mathrm{r}_{\mathrm{Rx}}, \tau, \varphi, \psi) \quad (3)$$

В случае плоской волны передаточная функция выглядит следующим образом:

$$\mathrm{h} = \mathrm{a}_1 \delta(\tau - \tau_\lambda)\delta(\varphi - \varphi_1)\mathrm{d}(\psi - \psi_1) \quad (4)$$

При учете поляризации передаточную функцию следует описывать как матрицу 2x2 [3]:

$$H = \begin{Vmatrix} h_{HH} & h_{HV} \\ h_{VH} & h_{VV} \end{Vmatrix} \quad (5)$$

4. Математическая модель

Радиочастотный канал может быть представлен как линейная система с входом и выходом. Вход и выход могут быть как дискретными, так и непрерывными. Передаточная функция зависит от времени и частоты.

$$y = h(t, \tau, f) * x \quad (6)$$

5. MIMO канал

В отличие от классических радиоканалов MIMO канал оборудован несколькими антеннами на приемной и передающей сторонах. Каждая пара антенн описывается своим линейным коэффициентом. Поэтому канал можно описать матрицей соответствующих коэффициентов и векторами входного и выходного сигналов.

$$H(t, \tau, f) = \begin{Vmatrix} h_{11}(t, \tau, f) & \cdots & h_{1n}(t, \tau, f) \\ \vdots & \ddots & \vdots \\ h_{m1}(t, \tau, f) & \cdots & h_{mn}(t, \tau, f) \end{Vmatrix} \quad (7)$$

6. Классификация моделей

Модели MIMO можно разделить на три основных типа[3]: аналитические, стохастические и детерминированные. Детерминированные модели основаны на распространение радиоволн в определенной местности. Они позволяют предсказать такие физические параметры как углы приема сигнала, задержки и т.д. Стохастические методы также позволяют предсказать физические параметры, но уже не на основе отдельно выбранной местности, а на основе случайно выбранном разложении рассеивающих объектов или, же рассматривая параметры как случайные величины с определенными функциями распределения. Аналитические методы описывают матрицу канала без учета физических параметров.

7. Распространение волн

В этом разделе рассматриваются факторы влияния среды на распространения волн. Во-первых, отражение сигнала вызывает затухание и изменение направления сигнала. Часть сигнала поглощается поверхность, часть отражается, оставшаяся рассеивается с некоторым угловым распределением. Вторым фактором является рассеивание как при падении волны на поверхность, так и при прохождение волн через препятствия. Третьи фактором является дифракция на гранях предметов. Дифракция позволяет заполнить затененное пространство электромагнитным полем. Влияние дифракции и отражение на распространение сигнала может быть определено благодаря детерминистическим методам. В более сложных моделях [4] учитывается влияние рассеивания на неровных поверхностях, таких как стены и фасады зданий.

7.1. Отражение

Одним из самых наглядных примеров распространения волн рассматривается отражение от земли. В этом случаи на приемную сторону приходит 2 луча: отраженный и прямой. Математически это выглядит следующим образом:

$$y = A\frac{\exp(-i2\pi kr)}{r} + BR\frac{\exp(-i2\pi kr')}{r'}$$

Коэффициенты A и B включают в себя зависимости от частоты и от диаграммы направленности антенн. Коэффициент отражения R определяется поляризацией сигнала, диэлектрической проницаемостью поверхности и углом падения волны. В случаи горизонтальной поляризации (параллельной поверхности Земли)

существует критический угол (угол Брюстера), при котором отражение равно 0.

$$R_H = \frac{\sqrt{\varepsilon^2 - \sin^2(\varphi)} - \cos(\varphi)}{\sqrt{\varepsilon^2 - \sin^2(\varphi)} + \cos(\varphi)} \quad (8)$$

$$R_V = \frac{\varepsilon\cos(\varphi) - \sqrt{\varepsilon^2 - \sin^2(\varphi)}}{\varepsilon\cos(\varphi) + \sqrt{\varepsilon^2 - \sin^2(\varphi)}} \quad (9)$$

$$\varphi_{Br} = \arctan(\varepsilon) \quad (10)$$

8. Пропускная способность

Пропускная способность радиочастотного канала ограничена Шенноновской границей:

$$C = B\log\left(1 + \frac{P_S}{P_N}\right) \quad (11)$$

B - ширина частотного спектра канала,
P_s - принятая мощность полезного сигнала ,
P_n - мощность шума.

8.1. Граница Шеннона для MIMO

В случаи MIMO канала сигналы представляются как вектора, поэтому применить формулу границы Шеннона для SISO канала в этом случаи нельзя. Используя сингулярное разложение матрицы канала и замену переменных, зависимость между векторами:

$$\vec{y} = H\vec{x} + \vec{n} \quad (12)$$

можно привести к системе независимых уравнений:

$$\tilde{y}_i = \sigma_i\tilde{x}_i + \tilde{n}_i \quad (13)$$

Таким образом MIMO канал можно представить как набор SISO каналов с передаточным коэффициентом σ_i. Суммарная пропускная способность подканалов равна пропускной способности MIMO канала и равна:

$$C = \sum\log\left(1 + SNR\sigma_i^{\,2}\right) \quad (14)$$

8.2. Нормализация матрицы передачи

Рассмотрим ситуацию, в которой шум η характеризуется случайным вектором, компоненты которого нормально распределены с нулевым средним и ковариацией $\sigma^2_\eta\,\mathbf{I}$. Средняя мощность шума на каждом приемнике, таким образом, равна σ^2_η. Средняя же мощность сигнала определяется по всем приемным антеннам как [10, параграф 1.1.2]:

$$P_R = \frac{1}{m}tr\left\{E\left\{x^{(k)H}x^{(k)}\right\}E\left\{H^{(k)H}H^{(k)}\right\}\right\} \quad (15)$$

С учетом независимости компонент вектора x получаем мощность входного сигнала равной $(P/N)I$. Поставляя в выражение для приемной мощности, мы получаем взаимосвязь между мощностью приема и матрицей передачи.

$$P_R = \frac{P}{mn}tr\left\{E\left\{H^{(k)H}H^{(k)}\right\}\right\} \quad (16)$$

$$H^{(k)H}H^{(k)} = \sum h_{ij}^{\,2} \quad (17)$$

$$SNR = \frac{P_T}{\sigma_n^{\,2}}E\left\{\frac{\sum h_{ij}^{\,2}}{mn}\right\} \quad (18)$$

Можно заметить, что соотношение сигнал-шум равно среднему значению SNR для отдельных пар антенн, если бы мощность на передающей антенне равнялась бы P_T. Отсюда следует, что рассмотрение матрицы H имеет смысл только для ранее заданной величины соотношения сигнал-шум. Поскольку это соотношение задает из ограничений по мощности в разных участках радиочастотного тракта, то имеет смысл ввести норму на H:

$$\frac{\sum h_{ij}^{\,2}}{n} = 1 \quad (19)$$

8.3. Взаимосвязь с физическими параметрами

Выше мы описали два модели канала: физическую и информационную. Между двумя моделями существует взаимосвязь. Зная параметры физической модели, параметры информационной модели вычисляются однозначно[3]:

$$h_{ij}(t,\tau) = \iiint h(\tau,\varphi,\psi)G(\varphi)\,f(\tau-\tau')d\tau'\,d\varphi d\psi \quad (20)$$

8.4. Корреляция и разновидности каналов

Формулу для пропускной способности можно записать в следующем виде:

$$C = \log_2\left(\det\left[I + \frac{SNR}{n}HH^*\right]\right) \quad (21)$$

где H - матрица канала, а SNR - соотношение сигнал-шум. Произведение HH^* называется корреляционной матрице, т.к. каждый коэффициент матрицы характеризует взаимосвязь двух пар передающих и принимающих антенн. Собственные числа матрицы корреляции равны квадратам сингулярных чисел матрицы H. То есть

выражение для пропускной способности может принять вид:

$$C = \sum \log_2 \left(1 + \frac{SNR}{n} \lambda_i \right) \quad (22)$$

Распределение собственных чисел матрицы определяется ее рангом [5].Поэтому введем классификацию каналов:

1. некоррелированные высокоранговые каналы
2. некоррелированные низкоранговые каналы
3. коррелированные низкоранговые каналы

9. MIMO канал в городской среде

9.1. Эффект "Зеленого поля" или Green Field

Рассмотрим идеализированный случай без отражений и затуханий в среде (случай свободного пространства). Благодаря идеализированности случая, физические параметры канала могут быть определены законами распространения волн. Необходимым критерием некоррелированной высокорангового канала в этом случаи является [6]:

$$\frac{d_t d_r}{R} > \frac{\lambda}{N} \quad (23)$$

9.2. Городская среда: эффект замочной скважины или keyhole

Другим частным случаем является keyhole эффект. Эффектом замочной скважины называется случай, когда некоррелированная матрица канала имеет низкий ранг. К примеру, если матрица имеет следующий вид:

$$\mathrm{H} = \begin{Vmatrix} a_1 b_1 & a_2 b_1 \\ a_1 b_2 & a_2 b_2 \end{Vmatrix} \quad (23)$$

то det(H) = 0. Таким образом, канал обладает только одной степенью свободы вместо двух. Во внутриофисном окружении данный эффект может наблюдаться в коридорах и узких проходах. В случаи внеофисной связи туннели, узкие улицы и дифракция поверх крыш домов могут приводить к низкому рангу некоррелированных каналов. Эффект редко встречается на практике Использование горизонтальной поляризации может уменьшать влияние данного эффекта.

10. Моделирование

10.1. Отражение от земли

10.1.1. Программный комплекс – SciLab

В данной части работы моделирование строится на бесплатном математическом продукте SciLab, альтернативном к коммерческому продукту MatLab. Язык SciLab очень похож на MatLab. Более подробную информацию о программе можно посмотреть [7].

10.1.2. Описание MIMO системы

Рассмотрим MIMO(2,2) систему, состоящую из 4 ненаправленных антенн, расположенных на высоте h на двух зданиях, расстояние между которыми l. Между домами находится абсолютно пустое пространство. Дома представляют собой параллелепипеды. Отражение от поверхности Земли считается идеальным. Диэлектрическая проницаемость равна 3 (близкое значение к значению для асфальта). Антенны будет считать изотропными.
Расстояние между антенн одинаково на каждой стороне: δ. Частота радиосигнала: f.

10.1.3. Геометрические параметры

Определим коэффициенты матрицы по физическим параметрам канала. Сигнал на приемной стороне является суперпозицией прямого и отраженного сигнала. Также предположим некоторую идеализацию модели, что разница в задержки прямого и отраженного сигнала незначительна по сравнению со временами сигнала. Таким образом, H(τ) = Hδ(τ). В этом случаи ненормализованную матрицу H можно описать без учета постоянных коэффициентов следующим образом:

$$h_{ij} = \frac{\exp(-i2\pi f\, l/c)}{l} + R\frac{\exp(-i2\pi f\, l'/c)}{l'} \quad (24)$$

$$l_{11} = L, l'_{11} = \sqrt{L^2 + 4h^2} \quad (25)$$

$$l_{12} = \sqrt{L^2 + {d_r}^2}, l'_{11} = \sqrt{L^2 + 4h^2 + {d_r}^2} \quad (26)$$

$$l_{21} = \sqrt{L^2 + {d_r}^2}, l'_{21} = \sqrt{L^2 + 4h^2 + {d_r}^2} \quad (27)$$

$$l_{22} = L, l'_{22} = \sqrt{L^2 + 4h^2} \quad (28)$$

10.1.4. Вычисления

Рассмотрим частоты f_1 = 2.4 GHz, f_2 = 5.6 GHz.

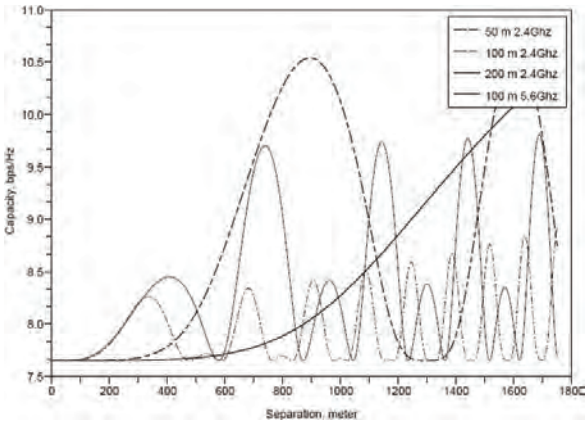


Рис.1

Вначале изучим зависимость удельной пропускной способности на рис. 1 от расстояния между антеннами для различных пар частоты и расстояния между узлами тракта. Самый левый максимум зависимости является оптимальным для рассмотрения, поскольку все последующие пики несильно отличаются от первого. Поэтому будем рассматривать именно этот оптимум. Расстояние между антеннами, необходимое для достижения максимальной удельной пропускной способности, растет с ростом расстояния между антеннами, что является логичным фактом, поскольку, чем более дальнюю связь мы хотим обеспечить, тем все размеры системы должны увеличиваться, в том и расстояние между антеннами. С ростом частоты расстояние сильно не изменяется, но уменьшается удельная пропускная способность.

При близких расстояниях между антеннами можно наблюдать некоррелированный высоко-ранговый канал. Максимальное расстояние сравнимо с расстоянием между антеннами в массиве. Некоррелированный высоко-ранговый канал (green field) имеет пропускную способность в 2 раза большую, чем SISO канал. Таким образом, использование N антенн на стороне приема и передачи позволяет увеличить в N раз пропускную способность. В случаи нечетного количества антенн большее количество антенн целесообразно располагать на стороне приема, т.к. коэффициент увеличения пропускной способности не может превышать количества антенн на каждой стороне, а соотношение сигнал-шум растет с добавлением еще одной антенны.

На дальних расстояниях вклад отражения от земной поверхности начинает расти, когда вклад свободного пространства равен нулю. Величина вклада пропорциональна расстоянию между антеннами в массиве. С увеличением частоты вклад падает.

В городской среде вклад отражения от земли является наиболее важным на больших площадях. На широких улицах отражение от земли становится равнозначным с отражениями от зданий. Чем уже улицы или больше количество отражений от зданий, тем менее значимым становится эффект отражения от земной поверхности.

10.1.5. Городская среда: 2D Ray Tracing

Для исследования влияния отражения от зданий на MIMO канал разработана программа на языке Java, основанная на детерминистическом методе трассировки луча. Программа базируется на платформе Eclipse и реализована как plugin. Математический аппарат основывается на Java Math пакете и для вычисления сингулярных чисел матрица используется пакет Jama.

Алгоритм трассировки луча состоит из двух этапов: построение дерева видимости объектов и обратной трассировки луча. Построение дерева видимости было реализовано на машине состояния и пошагового углового обзора. Для начала ищутся все объекты, видимые из точки излучения. Следующий шаг базируется уже на зеркальном отражении сигнала относительно видимого участка стены здания. Видимый участок формирует угол обозрения относительно новой точки обзора. И после этого процедура повторяется. Максимальное количество отражений луча задается заранее программно. Луч считается сформированным, если последним элементом пути является приемная антенна. Количество листьев дерева растет экспоненциально с увеличением максимального количества отражений.

Процедура обратной трассировки заключается в обработки дерева от листьев до верхушки дерева. После процедуры трассировки остается только список лучей. Далее для каждого луча вычисляются его углы приема и передачи, задержка сигнала, суммарные потери. Учитывая диаграмму направленности, вычисляются коэффициенты матрицы канала и пропускная способность.

10.1.6. Параметры модели

Источники: изотропные
Соотношение сигнал-шум: 20 dB
Максимальное количество отражений луча: 20
Количество зданий: 2-4
Количество антенн на передающей стороне: 2
Количество антенн на приемной стороне: 2
Расстояние между антеннами: 1 метр
Расстояние между массивами: 150 метров
Сценарии:

- Узкая улица длиной 200 метров и шириной 20 метров, массивы располагаются случайно на расстоянии 150 метров

- Широкая улица длиной 200 метров и шириной 35 метров, массивы располагаются случайно на расстоянии 150 метров

- Четыре здания разного размера, расположенные по кругу с радиусом 100 метров, массивы располагаются случайно на расстоянии 150 метров

- Сценарий аналогичен первому. Добавляется здание на конце улицы.

10.1.7. Результаты

В первом случае количество лучей между двумя антеннами разбросано от 20 до 60. Разница между пропускными способностями MIMO и SISO каналов составляет от 10^{-2} до 1 процента. Канал является высоко-коррелированным низкого ранга. Узкая улица не создает различия между точками передачи.
Во втором случаи количество лучей между двумя антеннами разбросано от 20 до 60. Разница между пропускными способностями MIMO и SISO каналов составляет от 10^{-1} до 1.2 процента. Канал является высоко-коррелированным низкого ранга. Широкая улица не создает значительного различия между точками передачи, но производительность MIMO канала выше, чем в случае узкой улицы.
В третьем случаи количество лучей между двумя антеннам порядка 2000. Процентное преимущество MIMO канала составляет от 15 до 25 процентов. Канал не является высоко-коррелированным, ранг матрицы равен 2. Канал можно считать близким к некоррелированному. Данный результат сравним с измерения в городской среде в [8].
При увеличении плотности зданий количество лучей остается того же порядка, а корреляция уменьшается. Это обусловлено тем что, длины лучей уменьшается и импульсная характеристика канала становится более равномерной, а отражения становятся сравнимы с вкладом прямой видимости.
В четвертом случаи количество лучей между двумя антеннами разбросано от 100 до 300. Разница между пропускными способностями MIMO и SISO каналов составляет от 10^{-1} до 1.2 процента. Канал является высоко-коррелированным низкого ранга. Дополнительные отражения не отразились значительно на увеличении пропускной способности.

11. Заключение

В работе рассмотрен вопрос определения производительности MIMO каналов в условиях городской среды. Реализована программа для определения физических параметров канала на основе законов отражения и метода трассировки луча. Изучены различные сценарии реализации канала. Рассмотрены зависимости типа матрицы от расположения и плотности зданий.
Исследовано влияние отражения от земной поверхности на производительность MIMO каналов. Приведены данные для оценки параметров MIMO системы для максимальной пропускной способности. Рассмотрено влияние факторов среды на пропускную способность канала. В работе нет учета дифракции и поляризации для трассировки луча. При построении больших деревьев требуется использование баз данных.

12. Литература

[1] John S.Seybold “Introduction to RF Propagation", 2005

[2] Martin Steinbauer, Andreas F. Molisch, Ernst Bonek "The Double-Directional Radio Channel", Aug. 2001

[3] P. Almers, E. Bonek, A. Burr, N. Czink, M. Debbah, V. Degli-Esposti, H. Hofstetter, P. Kyosti,D. Laurenson, G. Matz, A. F. Molisch, C. Oestges, H. Ozcelik "Survey of Channel and Radio Propagation Models for Wireless MIMO Systems", May 2006

[4] Enrico Maria Vitucci "MULTIDIMENSIONAL CHARACTERIZATION OF THE MIMO RADIO CHANNEL", 2006

[5] A.B. Gershman, N.D. Sidiropoulos "Space-time processing for MIMO communications", 2005

[6] David Gesbert, Helmut Bolcskei, Dhananjay A. Gore, Arogyaswami J. Paulraj Öutdoor MIMO Wireless Channels: Models and Performance Prediction"

[7] SciLab. "The open source platform for numerical computation", http://www.scilab.org/

[8] 8 J. Ling, D. Chizhik, P. Wolniansky, R. Valenzuela, N. Costa and K. Huber "Multiple transmit multiple receive (MTMR) capacity survey in Manhattan"

Оптимизация механизма определения занятости среды для mesh-сетей стандарта IEEE 802.11s

Коваленко А.В.
ИППИ РАН
kovalenko@iitp.ru

Лаконцев Д.В.
ИППИ РАН
lakontsev@iitp.ru

Пешков А.И.
ИППИ РАН
peshkov@iitp.ru

Аннотация

The в работе предложен метод оптимизации определения занятости среды для mesh-сетей стандарта IEEE 802.11s, на основе интерференционной модели.

1. Введение

Одной из задач возникающей при разработке оборудования для беспроводных mesh-сетей является повышение общей пропускной способности самоорганизующейся сети с учетом возможных произвольных изменений топологии во времени. Сложность заключается в интерференции сигналов одновременно передающих станций, что способствует потере информации принимающим узлом сети. Это, в свою очередь, приводит к повторным отправкам пакетов, что в значительной степени снижает общую пропускную способность. Для решения проблем, связанных с интерференцией, стандарт IEEE 802.11 предполагает использование протоколами MAC механизма проверки физического доступа к среде (physical carrier sensing – PCS). В частности, PCS использует протокол множественного доступа CSMA/CA, который регулирует порядок доступа к среде для передачи данных. Принцип работы заключается в следующем: каждая станция прослушивает среду для определения возможных передач в текущий момент времени. Передача возможна только тогда, когда уровень принимаемого сигнала меньше порога занятости среды (carrier sense threshold). В свою очередь, порог занятости связан с понятием зоны занятости среды D, которая определяет минимальное расстояние между одновременными передатчиками. В случае CSMA/CA это расстояние является фиксированным. В текущей работе мы рассмотрим вопрос увеличения общей пропускной способности беспроводной mesh-сети посредством оптимизации механизма PCS.

2. Интерференционная модель

Мы проанализируем возможность увеличения общей пропускной способности беспроводной mesh-сети в случае идеального MAC, т.е. будем пренебрегать всеми накладными расходами MAC-уровня, возникающими при пересылке служебной информации. Это ограничение значительно упростит поиск оптимального значения порога занятости среды передающей станции, а в дальнейшем подобные рассуждения помогут в переходе к рассмотрению реальной ситуации.

В дальнейших рассуждениях мы будем использовать модель зависимости уровня сигнала от расстояния:

$$P_{rx} = \frac{P_{TX}}{r^{\theta}} \quad (1)$$

где P_{rx} - уровень радиосигнала на расстоянии r от источника, излучающего с мощностью P_{TX} (дБ), а величина θ является показателем спада уровня сигнала, значение которого варьируется от 2 для свободного пространства до 4 для случая распространения сигнала в закрытых помещениях [1].

Поскольку каждая станция способна безошибочно воспринимать сигнал только в пределах ограниченного радиуса слышимости L, то расстояние от передатчика до приемника не должно превышать это значение. Иными словами, уровень сигнала на стороне принимающей станции не должен быть ниже порога слышимости P_L этой станции:

$$P_{rx} \geq P_L \quad (2)$$

Величина P_L, таким образом, характеризует чувствительность станции.

Однако для уверенного приема передаваемого сигнала этого условия мало, поскольку необходимо учесть помехи, связанные с интерференцией, что обеспечивается следующим соотношением:

$$\frac{P_{rx}}{N_{bg} + \sum_i \frac{P_{TX}}{r_i^\theta}} \geq S_0 \quad (3)$$

Левое выражение – это величина сигнал-интерференция-шум (signal-interference-noise ratio - SINR) принимающей станции, включающая величину N_{bg} стороннего шума и интерференцию $\sum_i \frac{P_{TX}}{r_i^\theta}$ прилежащих источников. S_0 - величина порога SINR, которая тем выше, чем выше скорость передачи данных.

Рассмотрим фрагмент беспроводной mesh-сети, включающий передатчик TX, приемник RX и четыре соседних с ними узла (A, B, C и E).

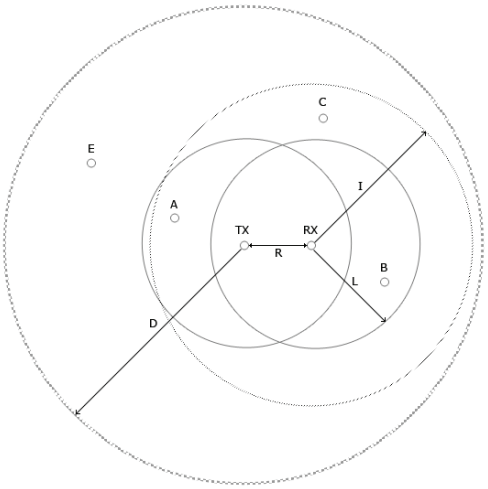


Рис.1. Фрагмент mesh-сети для описания интерференционной модели

Величина R - расстояние между TX и RX, определяемая (1):

$$P_{RX} = \frac{P_{TX}}{R^\theta} \quad (4)$$

где P_{RX} - уровень сигнала на стороне принимающей станции.

I - радиус области интенсивной интерференции. Граница этой области определяется соотношением (3). Иными словами, станции за пределами этой области не оказывают существенного влияния на передачу данных от TX к RX.

D - радиус области блокировки (или точнее области занятости среды), определяемый порогом занятости P_C:

$$P_C = \frac{P_{TX}}{D^\theta} \quad (5)$$

Теперь кратко рассмотрим ограничения, накладываемые на PCS MAC уровнем протокола IEEE 802.11.

3. Ограничения, накладываемые на PCS MAC уровнем протокола IEEE 802.11.

Стандартный PCS-механизм 802.11 предполагает использование фиксированного значения P_C, как уже отмечалось ранее. Более того, это значение в большинстве случаев является заниженным, что блокирует передачу даже существенно удаленных станций. К тому же, не предусматривается динамическая подстройка этого значения, что могло бы существенно увеличить общую пропускную способность мобильной mesh-сети. VCS, в свою очередь, не является эффективным механизмом в абсолютном большинстве случаев, тем более в мобильных сетях.

Гораздо более эффективным является возможность использования адаптационного PCS-механизма, когда значение P_C автоматически подстраивается для каждой конкретной ситуации, принимая оптимальное значение (или, по крайней мере, близкое к нему).

4. Оптимизация значения порога занятости среды

PCS позволяет рабочим станциям проверять уровень доступа к каналу связи. Каждый узел осуществляет проверку уровня входящего сигнала, и если он ниже порога P_C, то станция вправе начать передачу данных. Определителем того, будет ли переданный сигнал успешно принят другой станцией, является условие (3), т.е. значение SINR принимающей станции должно превышать пороговую величину S_0. Поэтому целью настройки порога занятости является предотвращение одновременных передач данных, которые могут провоцировать коллизии пакетов,

в то время как уровень SINR принимающей станции должен соответствовать условиям успешного приема.

Как значение порога P_C может регулировать величину области занятости D, определяет соотношение (5). Наглядно это показано на рисунке 2:

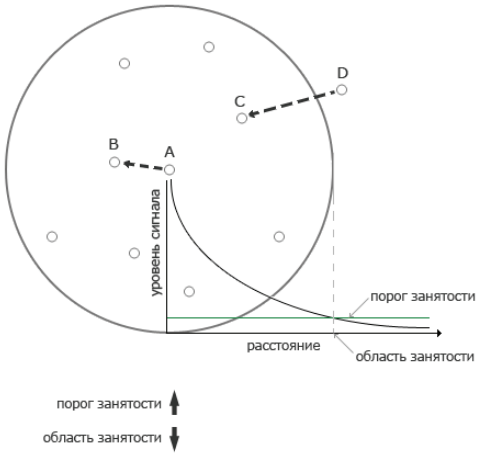


Рис.2. Зависимость размеров зоны занятости среды от величины порога P_C

Таким образом, чем выше поднимается значение P_C, тем уже становится область занятости среды и тем меньше прилежащих станций оказываются в заблокированном состоянии, и наоборот. Как это влияет на пропускную способность сети, продемонстрировано на рис. 3.

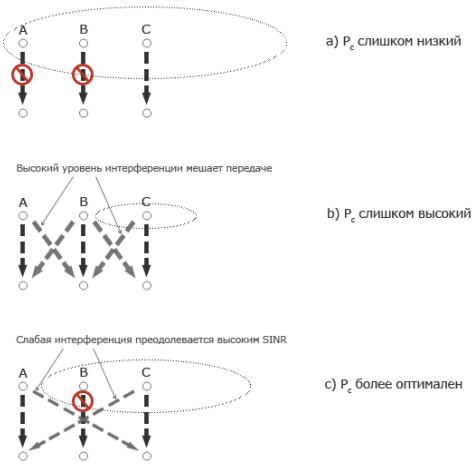


Рис.3. Влияние величины P_C на общую пропускную способность сети

Как мы видим, слишком малое значение порога занятости приводит к простою обеих станций A и B в момент передачи C (рис. 6а). Однако станция A могла бы осуществлять одновременную передачу, не оказывая существенного отрицательного влияния на текущую коммуникацию. В то же время, слишком большое значение P_C приводит к возможности одновременных передач всех трех станций A, B и C, что приведет к высокому уровню интерференции и потере данных (рис. 6б). Когда же порог занятости настроен оптимальным образом (рис. 6в), осуществима одновременная передача станциями A и C, в то время как B будет вынуждена дождаться окончания коммуникаций, что предотвратит вероятные коллизии.

Отметим также, что PCS с возможностью подстройки P_C является более надежным механизмом по сравнению с VCS, поскольку нет необходимости в приеме и корректном декодировании контрольных пакетов данных. Более того, PCS более гибок вследствие элементарной доступности регулирования диапазона занятости среды.

Каково же оптимальное значение порога P_C? Чтобы это понять, рассмотрим фрагмент сети, изображенный на рисунке 4. Опираясь на введенные ранее величины, можно заключить, что условие (3) имеет место быть только лишь в случае справедливости соотношения (6):

$$D \geq R + I \quad (6)$$

Т.е. область интенсивной интерференции I, оказывающая отрицательное влияние на передачу, полностью охватывается областью занятости D. При этом узел E также оказывается в области занятости передающей станции TX (проблема незащищенного узла). В то же время чем больше значение D, тем большее количество точек может быть блокировано таким образом. Естественно, подобный эффект является нежелательным и должен быть сведен к минимуму, поэтому условием оптимального значения D (а значит и P_C) является соотношение (7):

$$D = R + I \quad (7)$$

Графически эта ситуация показана на рис. 4, когда область интенсивной интерференции RX «касается» области занятости станции TX.

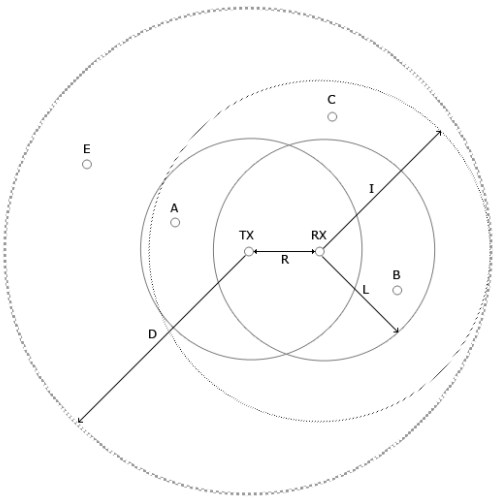


Рис.4. Оптимальное значение

Покрытие области интенсивной интерференции I способствует устранению проблемы скрытых точек. В то же время вероятность появления проблемы незащищенного терминала по-прежнему имеет место быть (точка E на рис. 7). Попробуем оценить масштабность этой проблемы в самом консервативном случае, когда интерференция вызвана одним единственным источником. В таком случае из (3), пренебрегая значением стороннего шума N_{bg}, имеем:

$$I = RS_0^{\frac{1}{\theta}}$$

Тогда учитывая (7), получаем соотношение для D:

$$D = R + I = R\left(1 + S_0^{\frac{1}{\theta}}\right)$$

Пусть величина ρ характеризует отношение площади локализации незащищенных точек ко всей области занятости среды:

$$\rho = \frac{\pi D^2 - \pi R^2}{\pi D^2} = 1 - \left(\frac{I}{D}\right)^2$$

Тогда учитывая соотношения для I и D, получим:

$$\rho = 1 - \left(\frac{S_0^{\frac{1}{\theta}}}{1 + S_0^{\frac{1}{\theta}}}\right)^2$$

Поэтому в случае, когда $S_0 \gg 1$, что, как правило, соответствует действительности [1][2], мы получаем $\rho \to 0$. Таким образом, проблемой незащищенных узлов можно почти полностью пренебречь

5. Общая пропускная способность сети

Попробуем оценить величину общей пропускной способности сети и ее зависимость от других параметров. Для этого воспользуемся теоремой Шеннона-Хартли для определения верхней границы скорости передачи данных через канал TX-RX (рис. 4) с аддитивным белым гаусовским шумом:

$$ChannelRate = W \log_2\left(1 + SINR\right) \quad (8)$$

Здесь W представляет собой полосу пропускания канала (в герцах), а величина $SINR$, как и ранее, определяется соотношением (3). Тогда общая пропускная способность сети Th будет зависеть от величины (8) и общего количества возможных одновременных передач данных:

$$Th = W \sum_{i=1}^{N} \log_2\left(1 + SINR_i\right) \quad (9)$$

Величина N определяет количество одновременных коммуникаций, а $SINR_i$ характеризует уровень сигнал-интерференция-шум i-ой станции. Очевидно, что величины $SINR_i$ не одинаковы для всех станций в общем случае, однако в дальнейших рассуждениях мы будем использовать наихудший случай интерференции, о котором немало упоминалось в [2][3][4]. Вариант топологии, воспроизводящей такой случай, приведен на рисунке 5.

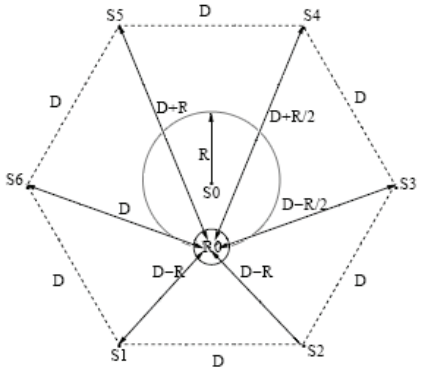


Рис.5. Топология с максимальной интерференцией

Поскольку принимающая станция R_0 расположена на краю области возможной передачи данных станции S_0, то соответствующий принимаемый сигнал будет иметь минимальный уровень мощности. Тогда величина SINR в наихудшем случае принимает вид:

$$SINR = \frac{\frac{P}{R^{\theta}}}{\frac{2P}{(D-R)^{\theta}} + \frac{P}{\left(D-\frac{R}{2}\right)^{\theta}} + \frac{P}{D^{\theta}} + \frac{P}{\left(D+\frac{R}{2}\right)^{\theta}} + \frac{P}{(D+R)^{\theta}}} \quad (10)$$

Обозначим отношение величины области занятости станции S_0 к расстоянию до приемника за X:

$$X = \frac{D}{R} \quad (11)$$

С учетом этого выражение (10) перепишется в виде:

$$SINR = \frac{1}{\frac{2}{(X-1)^{\theta}} + \frac{1}{\left(X-\frac{1}{2}\right)^{\theta}} + \frac{1}{X^{\theta}} + \frac{1}{\left(X+\frac{1}{2}\right)^{\theta}} + \frac{1}{(X+1)^{\theta}}} \quad (12)$$

Тогда выражение (9) для общей пропускной способности примет вид:

$$Th = N \log_2 (1 + SINR) \quad (13)$$

Остается оценить количество возможных одновременных передач N. В этом нам поможет рисунок 6.

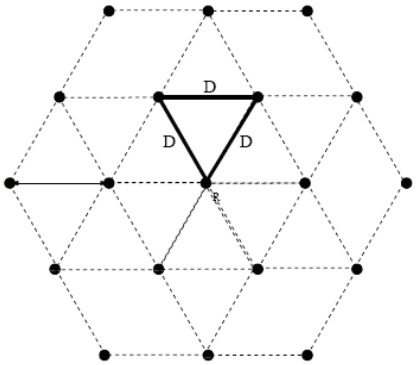


Рис.6. Количество одновременных передач

Поскольку минимальное расстояние между передатчиками определяется величиной D, то существует некоторое значение, характеризующее площадь, занимаемую одной транзакцией. Треугольник, показанный на рисунке 9, содержится в 3х областях одновременных передач данных. В то же время, каждая такая область содержит по 6 таких же треугольников. В связи с этим, площадь одной транзакции определяется величиной $\sqrt{3}D^2/2$. Тогда окончательное выражение для пропускной способности сети примет вид:

$$Th = C\frac{\ln(1+SINR)}{X^2}, \quad (14)$$

где $C = \frac{W * A}{\sqrt{3}/2 * \ln 2 * R^2}$, а величина A характеризует общую площадь охвата сети.

Полученное соотношение показывает, что величина общей пропускной способности есть функция X. Когда величина SINR очень мала, $\ln(1+SINR)$ возрастает почти линейно с ростом SINR, и общая пропускная способность растет. Когда же SINR принимает достаточно большие значения, логарифмическая функция продолжает расти, однако величина X^2 играет доминирующее значение, поэтому общая пропускная способность падает. Поэтому существует некое оптимальное значение X, характеризующее максимальную пропускную способность в некоторой своей δ-окрестности. Эту ситуацию, как мы выяснили ранее, иллюстрирует рисунок 4.

Оптимальные значения X, максимизирующие выражение (14), были получены для всего спектра значений θ.

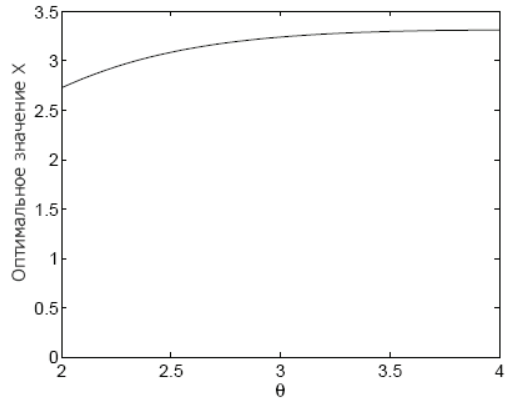


Рис.7. Зависимость оптимального значения X от величины θ

Таким образом, формальное значение общей пропускной способности беспроводной mesh-сети, полученное в наиболее экстремальном случае интерференционных помех, подтверждает наличие оптимального значения величины X, а значит и величины порога занятости P_C. Поэтому соотношение (7), определяющее оптимальное значение диапазона занятости среды (рис. 4), можно считать вполне оправданным.

Естественно, реальная ситуация может в значительной степени отличаться от приведенной экстремальной модели. Поэтому задача автоматической подстройки величины P_C в сторону оптимального значения не должна сводиться к использованию общих выражений вида (14). Наиболее разумным вариантом было бы использование алгоритма адаптации, проводимой каждой станцией в отдельности на основании текущей ситуации вокруг нее. Оценка такой этой ситуации может производиться с определенной долей периодичности, когда станция производит самостоятельную оценку необходимых величин или получает их в сообщениях от своих соседей.

Проведенный анализ предполагал использование идеального MAC-уровня, не обремененного передачей служебной информации. Тем не менее, в следующем разделе будет показано, что подобное пренебрежение может иногда существенным образом сказаться на отрицательной производительности сети.

6. Заключение

В работе проведен анализ проблем, связанных с интерференцией сигналов беспроводных mesh-сетей и рассмотрели возможный вариант оптимизации общей пропускной способности сетей за счет подстройки порога занятости среды. Было показано, что такая настройка может быть оптимальной, что позволит добиться максимальной пропускной способности в рамках предложенной модели..

[1] J. Kivinen, X. Zhao, and P. Vainikainen. “Empirical characterization of wideband indoor radio channel at 5.3 ghz”.

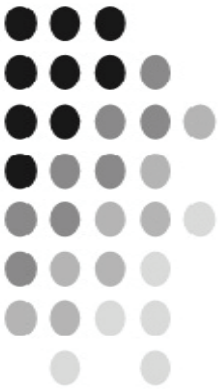
[2] Xue Yang and Nitin H. Vaidya. “On the Physical Carrier Sense in wireless Ad Hoc Networks”.

[3] Tae-Suk Kim, Hyuk Lim, Jennifer C. Hou. “Improving Spatial Reuse through Tuning Transmit Power, Carrier Sense Threshold, and Data Rate in Multihop Wireless Networks”.

[4] Matthias Grossglauser and David Tse. “Mobility increases the capacity of ad-hoc wireless networks”.

Раздел 2

Информационные технологии и проблемы управления в сложных системах и сетях



Система доказательства теорем с автоматической генерацией знаний

Раутиаинен А.
Российский университет дружбы народов
antti.rautiainen@gmail.com

Стефанюк В.Л.
Институт проблем передачи информации РАН
stefanuk@iitp.ru

Аннотация

В докладе решен ряд вопросов, связанных с созданием системы искусственного интеллекта с генерацией логического знания. Найден способ, позволяющий в ряде случаев обойти проблему полуразрешимости математической логики, и показан принцип создания генератора логических высказываний, который удовлетворяет некоторым простым, но обязательным исходным требованиям. В качестве примера приведена система доказательства теорем, в которой используется автоматическая генерация знаний.

1. Введение

Автоматической генерации знаний уже давно отводилась центральная роль в искусственном интеллекте. В первых системах искусственного интеллекта речь шла только об автоматическом использовании знаний, а сами знания вводились в систему «в ручную». Однако уже скоро системы стали настолько сложными, что возникла необходимость автоматической генерации знания. В попытках создания машинного разума для игр сталкивались с большим числом неодинаковых, но похожих друг с другом комбинаций — в итоге генерация тактики для разных комбинаций автоматическим образом часто оказывалась самым подходящим решением. Данный подход использовали в самой успешной в истории программе для игры в шашки «Chinook»[1], для компьютерной игры «Го» [2], для кубика Рубика [3] и многих других задач в области игр.

Пример совершенно другого способа автоматической генерации правил – американская система тактической симуляции для воздушных войск «TacAir – Soar». В нем правила создаются автоматическим образом прямо из наблюдения сенсорами поведения пилотов. Создатели «TacAir-Soar» могут гордиться тем, что с 7500 правилами поведения их экспертная система – одна из крупнейших в мире систем, находящихся в практическом применении. Понятно, что создание подобной базы вручную уже не целесообразно – необходима автоматизация [4].

В большинстве этих систем знание представлено не в логическом виде, а в форме, которая зависит от типа задачи (в случае экспертных систем – продукция, в случае игр – последовательность ходов и т.д.) [5]. Пока вопрос об автоматической генерации именно логического знания мало исследован, несмотря на то, что логическая форма представления знания самая общая, и допускает самое широкое применение.

Прежде всего, логическая форма представления знания используется в области автоматического доказательства теорем, но понятно, что она может пригодиться и для других областей искусственного интеллекта. Логическое знание подходит именно для тех систем автоматического доказательства, которые основаны на использовании знания.

Подход к доказательству теорем, основанный на знании, предусматривает методы, в которых в ходе доказательства новых теорем используются знания об уже доказанных теоремах. Такой подход несложно обосновать – новое знание всегда основано на накопленном старом. Однако в области машинного доказательства теорем до сих пор наибольшее внимание уделялось попыткам создания универсальных алгоритмов для доказательства любых теорем без первоначальных знаний, в первую очередь методом резолюции.

Во всех приложениях, как было показано многими авторами, знание играет фундаментальную роль. Но обычно речь идет о конкретных для данной предметной области знаниях. Нам кажется важным вопрос о построении всех возможных знаний, минимально ограничивая предметную область.

Формализация предметной области – это множества алфавитов, которые означают понятия в области, и множества аксиом, которые дают характеристику понятий в области. Но необходимо иметь некоторые аксиомы в самом начале. Знание – это теоремы о предметной области. Генератор знания состоит из генератора предложений (то есть, формул логики без свободных переменных) и системы доказательства теорем, которые пытаются выявить истинность предложений.

В подходе, основанном на знании, доказательство одной теоремы неотделимо от процесса накопления знаний в системе в целом. Система, в которой немного знаний (минимум — множества аксиом в теории, над которой мы работаем), не способна доказывать сложные теоремы. И каждая новая доказанная теорема может пригодиться при доказательстве других теорем. Создаваемая нами система исключительна в том отношении, что у других существующих систем логическое знание не генерируются автоматическим образом.

В существующих системах используется база данных из теорем, но эти теоремы создаются ”в ручную”, они основаны на интуиции программиста или пользователя о том, какие результаты могут быть полезными при доказательстве новых теорем. То есть программист или пользователь должен обладать большим запасом знаний о предметной области, а кроме того он должен тратить довольно много времени, если приходится адаптировать систему для новой предметной области, в которой необходимо доказать леммы. Этих затрат можно было бы избежать, если система в состоянии самостоятельно создавать базы данных из теорем для каждой предметной области.

Проблема генерации логического знания эквивалентна проблеме нумерации всех синтаксически правильных предложений в заданном формальном языке [6]. Действительно, логическое знание – это некоторая совокупность предложений, и для поиска предложений, которые могут быть полезными, нам нужен способ полного перебора всех предложений. То есть в системе, которая автоматическим образом генерирует логическое знание, должен присутствовать генератор предложений, который должен удовлетворять следующим 3 обязательным требованиям:

1. Он должен генерировать только предложения, т.е. замкнутые предложения, правильные конструкции в заранее указанном синтаксисе.
2. Он должен генерировать все возможные предложения, то есть обеспечивать полный перебор всех предложений.
3. Он не должен дважды генерировать одно и то же предложение.

Генератор, который выполняет все 3 требования, мы называем *правильным генератором*.

Естественно, сами предложения пока не являются знанием — знанием могут являться только истинные предложения. То есть система должна совершить попытку доказать или опровергнуть каждое предложение, и те предложения, которые системе удалось доказать, сохраняются в базе знаний.

2. Полуразрешаемость логики предикатов первого порядка

Свойство полуразрешаемости означает, что доказательство для некоторой совокупности предложений найдется (рано или поздно), если оно вообще существует. Однако если доказательства не существует, не факт, что система сможет это обнаружить и прервать процесс доказательства.

Но в тех случаях, когда либо у теоремы, либо у её отрицания есть доказательство, можно обойти проблему возможного бесконечного процесса доказательства путем одновременной генерации предложения и его отрицания. Это не полное решение проблемы полуразрешаемости, поскольку в неполных теориях теорема может быть верной, но не доказуемой в заданной аксиоматической системе. Однако данный метод помогает определить неистинность леммы в тех случаях, когда доказательство обратного ему утверждения – не слишком затруднительно.

То есть генератор должен генерировать одновременно и предложение, и его отрицание.

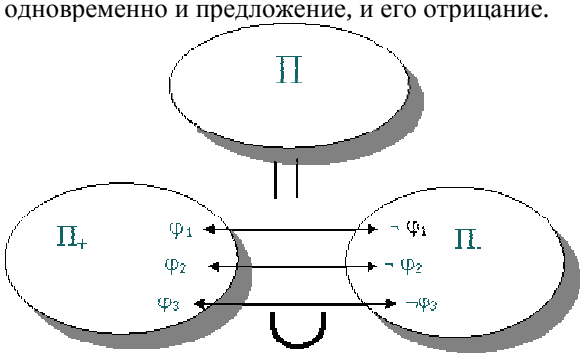


Рис. 1 Разделение множества предложений

Теперь, чтобы выполнялось требование 3) правильного генератора, мы должны позаботиться о том, чтобы отрицание, которое мы генерировали, не было бы сгенерировано повторно. То есть мы должны разделить множества предложений П на два, непересекающихся множества $П_+$ и $П_-$, с таким свойством, что у каждого предложения $φ ∈ П_+$ есть эквивалентный его отрицанию элемент в множестве $П_-$ (см. рис. 1). Естественно, существует бесконечно много способов реализовать подобные разделения, но мы выберем один из них.

3. Рекурсивная нумерация множеств

Рассмотрим сначала теорию нумерации, с помощью которой мы можем формализовать задачи полного перебора и создание генератора предложений в частности.

В аксиоматической теории множеств доказано, что у каждого множества s имеется изоморфный кардинал |s|. Соответственно, если дано конечное или счетное множество s, то каждому натуральному числу i < |s| соответствует единственный элемент x ∈ s

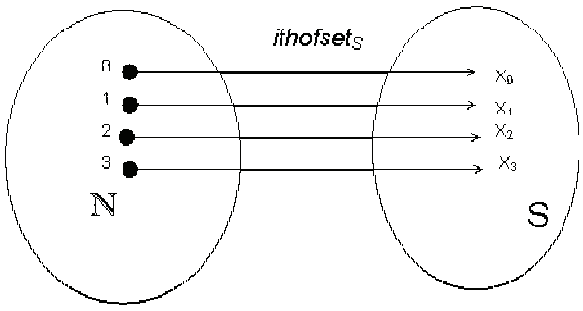


Рис. 2 Изоморфный кардинал множества S

Обозначим через *ithofset*(i,s) = x, определенный i-ый элемент множества s. Если фиксировано s, то это – функция одной переменной x = *ithofset*$_S$(i).

Данная функция *ithofset*$_S$ всегда существует, когда их имеется несколько (как в случае, когда |s| > 1), можно ограничиться одной из них. Но простого знания, что функция существует, нам недостаточно – мы должны уметь ее вычислять, то есть найти рекурсивную конструкцию, которая вычислит функцию *ithofset*$_S$.

Рекурсивную конструкцию, вычисляющую *ithofset*$_s$, обозначим через ithofset$_S$ и назовём «рекурсивной нумерацией множества s». Проблема нумерации – это проблема нахождения рекурсивной конструкции ithofset$_S$, если известен способ построения множества s. Мы используем обозначения ithofset(i,s) = ithofset$_S$(i) для рекурсивной нумерации произвольного множества s, в случае, когда она определена.

Если удалось найти нумерацию множества s, то можно осуществить полный перебор множества s, проверяя все элементы

ithofset$_S$(0),
ithofset$_S$(1),
ithofset$_S$(2),
…

в порядке их нумерации.

Если удалось найти нумерацию множества всех предложений ithofset$_П$, она также будет генератором предложений, который удовлетворяет всем 3 обязательным требованиям, указанным выше, то есть является правильным генератором.

Теория нумерации располагает теоремами, позволяющими построить нумерацию множества S ithofset$_S$, исходя из составляющих множества S.

Для нахождения нумерации всех положительных предложении $П_+$, необходимо представить множество с помощью таких операций теории множества, как X (картезианское произведение), ⊕ (объединение без пересечений) и.т.д. Такая конструкция слишком сложна, чтобы представить её в рамках данной статьи, но она подробно изложена в дипломной работе автора.

4. Система доказательства теорем с автоматической генерацией знаний

В системе должен обязательно присутствовать компонент «разделитель», который будет отделять отрицательные формулы от эквивалентных положительных, и отправлять их в разные процессы доказательства. Разделитель должен также уметь собирать результаты доказательств, и прерывать второй процесс при успешном завершении первого.

Как правило, все алгоритмы доказательства теорем имеют экспоненциальную сложность. Если невозможно распределить любое количество системных ресурсов в процессе доказательства, которое может длиться часами (или годами), то можно задать максимальное время, после которого процесс прерывается и предложения загружаются в базы данных предложений (вместе с предложением, эквивалентным его отрицанию) с пометкой «неопределенное предложение».

Также необходимо, чтобы система выполняла параллельно не только одно предложение и его отрицание, а также разные предложения одновременно. То есть человек может заниматься

одной проблемой, которую компьютеру не удалось решить, и компьютер может одновременно стараться доказать множество других предложений (вместе с отрицаниями). Параллелизм задает синергию, в которой развитие в одном параллельном процессе может помогать и остальных процессах, которые находятся в действии.

Итак, в системе должно быть предусмотрено ветвление к разным, параллельным процессам, и существовать компонент «распределитель», который будет управлять распределением задач по разным подпроцессам. Распределитель должен знать, какой из процессов занят, и какой доказывает какие предложения. Распределитель должен уметь передать процессу команду о прерывании, которую ему отправил разделитель в случае удачного завершения доказательства предложения, эквивалентного его отрицанию. Всего системе требуется два распределителя для отрицательных и для положительных предложений.

5. Реализация системы доказательства теорем с автоматической генерацией знаний

Все компоненты системы являются асинхронными процессами в самой операционной системе, чтобы архитектура системы была как можно более независимой от архитектуры компонентов.

В качестве системы доказательства теорем для данной работы выбрана система Isabelle. Она позволяет использовать множество разных логик. Она использует слабую теорию типов в качестве металогики, и реализует разные системы логики, в том числе аксиоматическую теорию множеств ZFC, в которой предложения генерируемы [7]. Isabelle может действовать как интерактивно, так и в автоматическом режиме. Главный метод доказательства в системе Isabelle – это резолюция для логики предикатов высшего порядка, но в ней есть также схемы переписывания термов и система автоматического доказательства методами аналитических таблиц. Преимущества Isabelle с точки зрения подхода, основанного на накоплении знаний, в том, что в ней есть вложенная концепция «теории», которая поддерживает идею о постепенно расширяющихся знаниях. Предшественник системы Isabelle, HOL, является одной из наиболее популярных систем в области верификации в различных приложениях. В самой системе Isabelle доказано немало нетривиальных математических теорем (теорема неполноты Геделя, теорема Геделя о непротиворечивости аксиомы выбора), и также ряд прикладных задач (о корректности протокола безопасности и свойств семантики языков программирования).

6. Заключение

В данном исследовании был решен ряд вопросов, связанных с созданием системы искусственного интеллекта с генерацией логического знания. Был продемонстрирован способ обходить проблему полуразрешаемости логики в ряде случаев, и показан математический принцип создания генератора, который выполняет ряд простых, но нужных условий.

Были также описаны требуемые компоненты системы, и в качестве примера приведена система автоматического доказательства теорем, которая использует принцип автоматической генерации знаний.

12. Литература

[1] J. Schaeffer, Y. Bjornsson, N. Burch, A. Kishimoto, M. Muller, R. Lake, P. Lu and S. Sutphen, “Solving Checkers”, *International Joint Conference on Artificial Intelligence (IJCAI),*, 2005,pp. 292-297

[2] B. Bouzy , T. Cazenave, “Computer Go: An AI oriented survey”, *Artificial Intelligence* Volume 132, Issue 1, 2001

[3] R. Korf, “Finding optimal solutions to Rubik's Cube using pattern databases”, *Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI-97)*, Providence, RI, July, 1997

[4] G. Taylor., B. Stensrud , S. Eitelman,,C. Dunham and E. Harger, “Toward Automating Airspace Management.” *Computational Intelligence for Security and Defense Applications (CISDA)*, Honolulu, HI: IEEE Press.J. ,2007

[5] В.Л.Стефанюк “Локальная организация систем – модели и приложения” М.: Физматлит 2004.

[6] А. Раутиаинен, “Алгоритм нумерации объединения конечных множеств без пересечении”, *XLII Всероссийская конференция по проблемам математики, информатики, физики и химии 17-21 апреля 2006 года: Труды конференции. Секции «программные исследования»*. – М.:Издательство РУДН, 2006, стр. 122-130.

[7] M.M. Wentzel, *Isabelle/Isar - a versatile environment for human-readable formal proof documents* Institut für Informatik der Technischen Universität München 2001.

Размещение фотографий на холсте с определением коэффициентов масштабирования

М.А.Егорова, И.В.Сафонов
Московский инженерно-физический институт (государственный университет)
marta.egorova at yahoo.com, ilia.safonov at gmail.com

Аннотация

Рассматривается задача оптимального размещения *N* прямоугольных фотографий произвольного размера на изображении-холсте произвольного размера с определением коэффициентов масштабирования фотографий. Критериями оптимальности могут быть минимум площади перекрытий фотографий, максимум покрытия площади холста, минимум разности между масштабированными изображениями максимальной и минимальной площади, нахождение коэффициентов масштабирования в заданном диапазоне, а также комбинация данных критериев.

Существующие алгоритмы размещения и упаковки не позволяют определить коэффициенты масштабирования, поэтому предложено оригинальное решение, включающее три основных этапа: составление всех возможных схем размещение для заданного количества изображений, нахождения для каждого размещения, а также заданных размеров фотографий и холста, коэффициентов масштабирования с помощью МНК, выбор оптимального размещения.

Обсуждается вычислительная сложность данного алгоритма и предлагается его эффективная реализация. Предлагается фрактальный подход для решения задачи в случае большого *N*. Приведены примеры работы предложенного алгоритма для различных критериев оптимальности размещения.

1. Постановка задачи

В настоящее время существует множество прикладных задач, в которых требуется разместить несколько изображений в пределах прямоугольной области – холста, при этом определив коэффициенты масштабирования размещаемых изображений. Примерами таких задач являются размещение фотографий на страницах фотоальбома и фото-книги, составление коллажей [1], создание различных интеллектуальных интерфейсов просмотра и навигации графической информации.

Задача размещения *N* прямоугольных фотографий произвольного размера на изображении-холсте произвольного размера с определением коэффициентов масштабирования фотографий формулируется следующим образом:

Имеется *N* изображений произвольного размера (w_i, h_i), $i=1..N$ и холст I_C заданного размера (w_C, h_C). Представим изображение I_i как множество пикселов *p*, положение каждого задается парой чисел (x_P, y_P): $I_i = \{p: (x_P, y_P) \mid x_P \in [1, w_i], y_P \in [1, h_i]\}$. Изображения после масштабирования будем обозначать индексом *n*.

Необходимо разместить все изображения в заданной области при следующих условиях:

1) $w_{in}/w_i = h_{in}/h_i = k_i$, $i=1..N$, где k_i - коэффициент масштабирования *i-го* изображения, w_{in} и h_{in} габариты масштабированного изображения;

2) $\forall i \; I_{in} \cap I_C \neq 0$;

3) $\forall i, \quad j: \; I_{in} \cap I_{jn} >= 0$.

Первое условие является условием пропорционального масштабирования изображений, второе описывает присутствие на холсте всех масштабированных изображений, третье определяет возможность их частичного перекрытия.

2. Задачи размещения и упаковки

Сходными с решаемой задачей являются задачи размещения и упаковки, применяемые в различных областях жизнедеятельности [2]:

- раскрой тканей; резка стекла, дерева, металла; размещение различных компонент на платах;
- упаковка предметов в контейнеры;
- планирование времени, распределение задач;
- составление бюджета и т.д.

Постановки наиболее известных из них приведены в табл. 1 [3].

Таблица 1. Постановка задач размещения и упаковки

Задача	Исходные данные	Постановка задачи
двумерной упаковки	n прямоугольных предметов с заданными размерами	Найти упаковку с минимальной площадью окаймляющего прямоугольника
упаковки в полосу	n прямоугольных предметов с заданными размерами и полубесконечная полоса шириной W	Найти упаковку предметов с минимальной длиной занимаемой полосы
о рюкзаке для прямоугольников	n прямоугольных предметов с заданными размерами и весами c_{ij} и большой прямоугольник с заданными размерами	Найти подмножество предметов максимального веса, которые можно вырезать из большого прямоугольника
упаковки в контейнеры для прямоугольников	n прямоугольных предметов с заданными размерами и неограниченное число одинаковых заготовок заданным размером	Найти упаковку предметов с минимальным числом используемых заготовок
прямоугольного раскроя для серийного производства	n прямоугольных предметов с заданными размерами и потребностью k_i штук	Найти план раскроя заготовок, позволяющий получить все предметы в требуемых количествах и использующий минимальное число заготовок

Все перечисленные задачи являются NP-полными, т.е. для них не найдено решения, которое бы решало задачу за полиномиальное время. Также, в явном виде ни одна из представленных задач размещения и упаковки не соответствует поставленной задаче размещения изображений в заданном размере с одновременным определением коэффициентов пропорционального масштабирования. Поэтому предложен алгоритм, позволяющий размещать изображения оптимальным образом на холсте заданного размера за приемлемое время.

3. Описание алгоритма

3.1. Общие положения

Введем понятия топологии и матрицы размещения. Под топологией будем понимать схему размещения некоторого количества изображений в прямоугольнике. Матрицей размещения будем называть матрицу из 0 и 1, где 1 показывает, что данное изображение образует измерение w (ширину) или h (высоту). Алгоритм состоит из трех этапов, показанных на рис. 1.

На первом шаге находятся все топологии для заданного количества изображений. На следующем этапе определяется оптимальный способ размещения, с учетом размеров изображений и размера холста, а также находятся коэффициенты для пропорционального масштабирования. На последнем этапе изображения масштабируются и размещаются согласно найденной топологии.

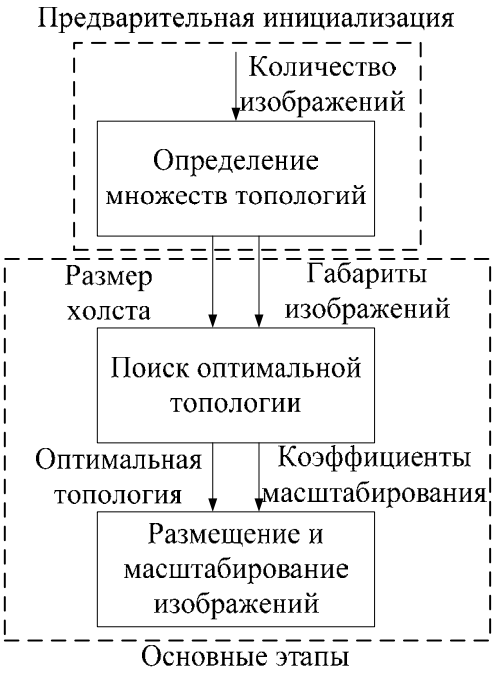


Рис. 1. Этапы алгоритма размещения изображений в заданном размере

3.2. Определение множеств топологий

Представим результирующее изображение как квадрат, разделенный по вертикали и горизонтали на столько частей, сколько входных изображений. Например, число входных изображений равно 5, тогда результирующее изображение представляется в виде квадрата 5x5. Значение каждой ячейки данного квадрата есть число от 1 до 5, которое означает, что в данном месте результирующего изображения

располагается входное изображение с таким номером.

В общем случае, одно изображение будет в левом верхнем углу (первое), а одно в правом нижнем углу (второе). Поэтому сначала в любой топологии устанавливаются значения для первого изображения, потом для второго и потом все остальные, по принципу слева направо и сверху вниз, пока есть свободное место. После того как получена некоторая матрица топологий, делается проверка: существует ли подобная топология в ранее отобранных. На рис. 2 показаны все топологии для случая, когда входных изображений три.

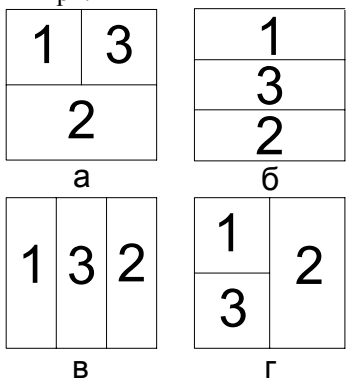


Рис. 2 Топологии для 3-х изображений

Проверка единственности топологии осуществляется по количеству соседей по вертикали и горизонтали для каждого изображения.

Рассмотрим систему координат, показанную на рис. 3, связанную с холстом для случая размещения рис. 2г.

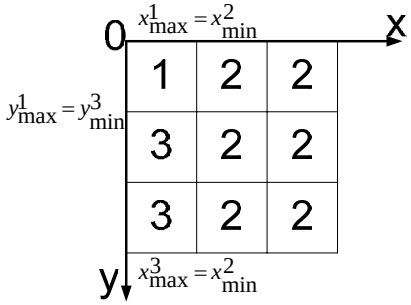


Рис. 3 Система координат холста

Два изображения *i* и *k* на холсте являются соседними по горизонтали, если $x^i_{\max} = x^k_{\min}$, аналогично, два изображения *i* и *k* на холсте являются соседними по вертикали, если $y^i_{\max} = y^k_{\min}$. Обозначим A^i_x и A^i_y количество соседей *i-го* изображения по горизонтали и вертикали, соответственно. Тогда две топологии *T1* и *T2* будем считать идентичными, если $\forall i$ $\exists k$ такое что выполняются условия:

1. $\begin{cases} \left(A^i_x\right)_{T1} = \left(A^k_x\right)_{T2} \\ \left(A^i_y\right)_{T1} = \left(A^k_y\right)_{T2} \end{cases}$
2. $k \notin \{M\}$, где M – множество значений номеров изображений, для которых условия 1 и 2 были выполнены ранее. Первоначально $M=\{\varnothing\}$.

Рассмотрим более подробно механизм проверки. Пусть матрицы топологий имеют вид как на рис. 4, где матрица топологий а) находится в списке топологий, а матрица топологий б) претендует на то, что является новой топологий, которой еще нет в списке отобранных.

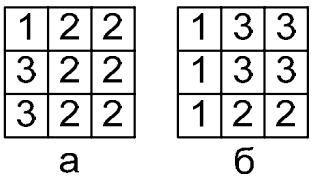


Рис. 4. Пример одинаковых топологий
а) Отобранная топология
б) Топология, претендующая стать отдельной

Для топологий, показанных на рис. 4а и 4б данные о количестве соседей представлены в табл. 2 и табл. 3, соответственно.

Таблица 2 Количество соседних изображений для топологии рисунка 4а

i	**1**	**2**	**3**
A^i_x	1	2	1
A^i_y	1	0	1

Таблица 3 Количество соседних изображений для топологии рисунка 4б

i	**1**	**2**	**3**
A^i_x	2	1	1
A^i_y	0	1	1

Поскольку топология определяется количеством соседей для каждого изображения, но без привязки к конкретному изображению, то данные две топологии идентичны и поэтому топология б) отобрана не будет.

После того, как топология отбирается в результирующий набор, для данной топологии находится матрица размещений. Первый столбец $g^{,0}$ матрицы размещения обозначается либо *w* либо *h* в зависимости от измерения, которое образуется следующими элементами строки.

Остальные элементы матрицы имеют значение либо 0, либо 1. Каждый следующий столбец отвечает за номер соответствующего изображения топологии (например. третий столбец матрицы размещений отвечает за второе изображение топологии). Если изображение

образует данное измерение, то в его столбце ставится 1, если нет то 0. Так для топологии, представленной на рис. 4а матрица размещений будет выглядеть следующим образом:

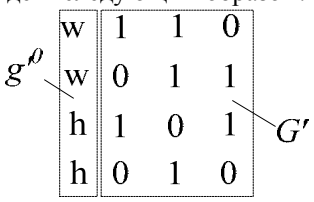


Рис. 5 Пример матрицы размещения

или в виде системы уравнений:

$$\begin{cases} w_c = w_1 + w_2; \\ w_c = w_2 + w_3; \\ h_c = h_1 + h_3; \\ h_c = h_2. \end{cases} \quad (1)$$

3.3. Определение коэффициентов масштабирования изображений

Для определения коэффициентов масштабирования используется метод наименьших квадратов, описанный в [4, 5].

Рассмотрим работу метода в терминах сформулированной нами ранее задачи. Имеется явная зависимость между размерами изображений и размером холста. По матрице размещений составим матрицу габаритов изображений G_{ij} и вектор-столбец габаритов холста g^0:

$$G_{ij} = \begin{cases} G'_{ij} w_i, & \text{если} \quad g'^0_i = w \\ G'_{ij} h_i, & \text{если} \quad g'^0_i = h \end{cases}, \quad (2)$$

$$g^0 = \begin{cases} w_C, & \text{если} \quad g'^0_i = w \\ h_C, & \text{если} \quad g'^0_i = h \end{cases}, \quad (3)$$

где $j = \overline{1, N}$, $i = \overline{1, m}$, m – количество строк в матрице размещения.

Требуется найти функцию ψ от набора значений g_j, которая бы по возможности наиболее точно выражала значение g_j^0, т.е. значение $\psi(g_j)$ было бы близко к g_j^0 для всех $j = \overline{1, n}$.

Представим возможные аппроксимирующие функции в виде:

$$\psi(g_j) = \sum_{i=1}^{m} \alpha_i g_j^i, \quad (4)$$

где α_i, $i = \overline{1, m}$ - набор вещественных коэффициентов, образующих вектор α. Его будем назвать вектором коэффициентов пропорционального масштабирования. При разных значениях вектора α будут получаться разные функции ψ и соответственно разные значения величин $\psi(g_j)$. Требуется подобрать такой вектор параметров α, при котором для каждого j величина $\psi(g_j)$ была бы близка к g_j^0.

Степень «близости» характеризуется абсолютными значениями разности между величинами g_j^0 и $\psi(g_j)$. Эту разность будем называть невязкой или погрешностью описания функцией ψ связи между размерами входных изображений и размером холста. Невязки образуют вектор y, с компонентами

$$y_j = \psi(g_j) - g_j^0, \; j = \overline{1, n}. \quad (5)$$

Учитывая (4), получаем, что вектор невязок можно представить в виде линейной вектор-функции от вектора искомых коэффициентов:

$$y(\alpha) = G\alpha - g^0. \quad (6)$$

При различных значениях вектора α образуются различные векторы невязок, множество которых составляет линейное многообразие $Y = \{y(\alpha): \; \alpha \in E^m\}$.

Итак, задача состоит в выборе такого вектора α, при котором абсолютные значения величин $y_j(\alpha)$, $j = \overline{1, n}$, были бы минимальны.

Решение данной задачи методом наименьших квадратов состоит в минимизации функции:

$$f(y) = \sum_{j=1}^{n} (y_j)^2 \quad (7)$$

на линейном многообразии Y. Выразив вектор невязок в виде линейной функции от вектора параметров α, можем перейти к задаче безусловной минимизации функции:

$$\tilde{f}(\alpha) = (G\alpha - g^0)^T (G\alpha - g^0). \quad (8)$$

Функция $\tilde{f}$ является выпуклой и дифференцируемой. Ее градиент имеет вид:

$$\nabla \tilde{f}(\alpha) = 2G^T (G\alpha - g^0). \quad (9)$$

Приравняв градиент нулевому вектору, получим систему линейных уравнений:

$$(G^T G)\alpha = G^T g^0. \quad (10)$$

Решение этой системы $\bar{\alpha}$ будет требуемым значением вектора коэффициентов масштабирования. Решение системы можно записать следующим образом:

$$\bar{\alpha} = (G^T G)^{-1} G^T g^0. \quad (11)$$

Например, рассмотрим вычисление коэффициентов масштабирования для матрицы размещений, представленной на рис. 5. Пусть $w_C = 900$ и $h_C = 600$, размеры трех входных изображений равны соответственно (200, 300), (400, 500), (250, 270).

Тогда матрица габаритов изображений и вектор габаритов холста:

$$G = \begin{bmatrix} 200 & 400 & 0 \\ 0 & 400 & 250 \\ 300 & 0 & 270 \\ 0 & 500 & 0 \end{bmatrix} \qquad g^0 = \begin{bmatrix} 900 \\ 900 \\ 600 \\ 600 \end{bmatrix}$$

По формуле (10) определяется вектор коэффициентов масштабирования $\overline{\alpha}$.

3.4. Выбор оптимального размещения

В качестве условия оптимальности применимы следующие критерии и их комбинация:

1. $\min(S_{\max} - S_{\min})$, где S_{max}, S_{min} – соответственно максимальная и минимальная площади масштабированных изображений, то есть размер размещенных изображений не должен сильно различаться;
2. $\min\{S_c - \bigcup_{i=1}^{N} S_{in}\}$, где S_c - площадь холста, S_{in} – площадь, занимаемая i-ым отмасштабированным изображением на холсте, то есть площадь непокрытых изображениями областей на холсте должна быть минимальна.
3. k_i: $\forall i \quad k_{min}<k_i<k_{max}$, где k_{min}, k_{max} – соответственно нижняя и верхняя граница допустимых значений коэффициентов масштабирования, то есть коэффициенты масштабирования могут быть ограничены, например, изображения можно только уменьшать и нельзя увеличивать.

4. Время работы

Оценка по времени проводилась для первого этапа (предварительная инициализация данных) и второго этапа (размещение). Инициализация, т.е. построение топологий и матриц размещений проводится один раз. Данные о времени вычисления этих данных приведены в табл. 4.

Таблица 4. Данные о времени построения матриц топологий и матриц размещений

Количество входных изображений	3	4	5	6	7	8
Время поиска решений, с	<0.1	<0.1	~0.2	~5	~222	~2533
Количество топологий	4	16	37	116	385	1211

Время измерялось на компьютере с характеристиками: процессор Pentium 4 1,9 ГГц, 224 Мб ОЗУ.

На рис. 6 показан график зависимости времени поиска оптимального размещения от количества размещаемых изображений.

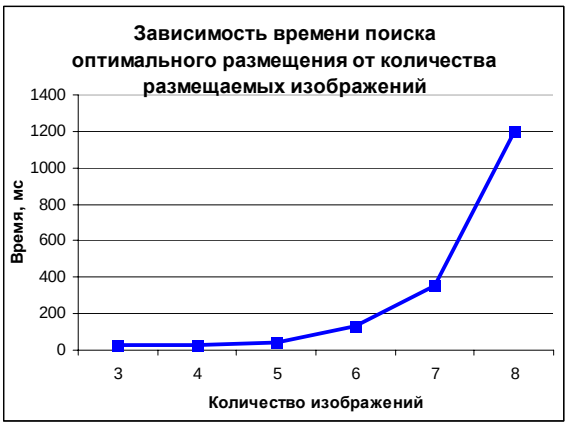


Рис. 6 График зависимости времени поиска оптимального размещения от количества размещаемых изображений

По графику видно, что чем больше количество изображений, которые необходимо расположить на холсте, тем длительнее поиск оптимального размещения. Эта зависимость обусловлена значительным увеличением числа топологий при увеличении числа входных изображений.

5. Фрактальный подход

Для использования алгоритма размещения в задачах большой размерности предложен подход, который является фрактальным в смысле подобия фрагмента размещения, полученного рекуррентной процедурой, всему множеству размещений.

Пусть необходимо разместить N изображений (N>8), тогда на первом шаге выбирается K изображений (K<=8) и они размещаются на холсте. Затем определяется наибольшая по площади область холста и формируется множество неразмещенных изображений. При этом изображение, занимающее наибольшую площадь на холсте, возвращается к множеству неразмещенных изображений R и участвует на следующем шаге размещения. Процесс повторяется до тех пор пока количество неразмещенных изображений R не станет меньше или равным Z. Оставшиеся неразмещенные изображения располагаются в последней максимальной по площади области. Шаги алгоритма показаны на рис. 7.

Рис. 7 Алгоритм фрактального подхода размещения изображений

При выборе изображений для размещения в максимальной области возможны следующие варианты реализации:

1. Размещать одинаковое количество изображений ($K = const$).
2. Размещать разное количество изображений и выбирать размещение, отвечающее наперед заданным критериям оптимальности как в п. 3.3.
3. Выбирать случайное количество для размещения.
4. Уменьшать с каждым шагом количество размещаемых изображений и т.д.

Для размещения изображений внутри максимальной области используется алгоритм, описанный в п.3

Данный подход позволяет значительно сократить время размещения при $N>8$. Например, время размещения 20 изображений несколько секунд вместо нескольких дней.

6. Примеры размещения

На рис. 8 и 9 показаны результаты размещения, соответственно, 6 и 23 изображений. На рис. 9 красным показаны области на первом шаге размещения изображений. Для выбора оптимального размещения использовался комбинированный критерий:

$$\min(2*(S_C - \bigcup_{i=1}^{N} S_{in}) + \bigcap_{i=1}^{N} S_{in} + (S_{\max} - S_{\min})) \quad (12)$$

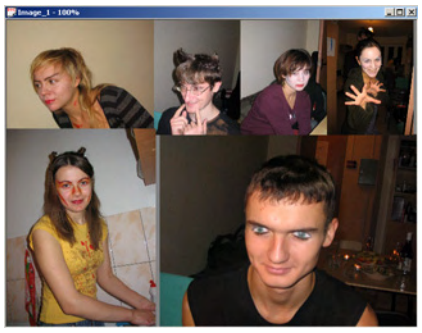


Рис. 8 Пример размещения 6-ти изображений

При размещении с помощью фрактального подхода использованы константы K=7, Z=6.

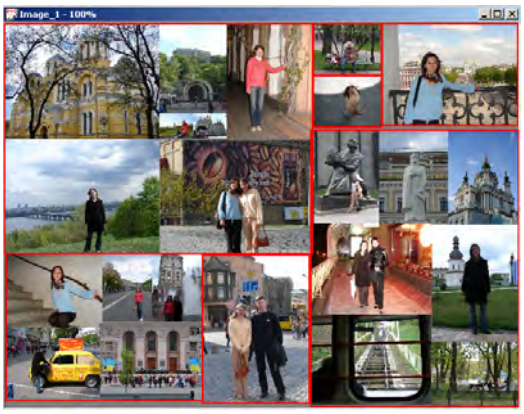


Рис. 9 Пример размещения 23-х изображений

7. Заключение

Предложен алгоритм размещения изображений произвольных размеров на холсте заданного размера с определением коэффициентов масштабирования. Алгоритм использует метод наименьших квадратов и может быть использован для решения различных задач в зависимости от используемого критерия оптимального размещения. Также рассмотрен фрактальный подход размещения большого количества изображений, который позволяет значительно сократить время обработки. Алгоритмы реализованы в демо-программе.

8. Список литературы

[1] Carsten Rother, Lucas Bordeaux, Youssef Hamadi, Andrew Blake. “AutoCollage”. SIGGRAPH, 2006

[2] Janne Karelahti, “Solving the cutting stock problem in the steel industry”, Master’s thesis, Espoo, 25.11.2002.

[3] Ю.А.Кочетов, Курс лекций «Теория принятия решений», НГУ.

[4] Метод наименьших квадратов: геометрические свойства/ В.И.Зоркальцев.- Новосибирск: ВО «Наука». Сибирская издательская фирма, 1995. - 220с.

[5] Т.Кормен, Ч. Лейзерсон, Р.Ривест. Алгоритмы: построение и анализ/Пер. с англ. Под ред. А.Шеня. – М.: ЦНМО, 2002. – 960с.:263 ил.

ГеоТайм II - новая геоинформационная технология исследования пространственно-временных процессов

А.Б. Дерендяев, П.А. Метриков, В.Г. Гитис
ИППИ РАН
wintsa@gmail.com, metpavel@mtu-net.ru, gitis@iitp.ru

Abstract

Обсуждается геоинформационная технология анализа пространственно-временных процессов и ее реализация в распределенной сетевой ГИС ГеоТайм II. ГИС ориентирована на решение двух типов задач: (1) Визуализация пространственной и пространственно-временной географической информации (ГИ), оценивание связей между ее компонентами и (2) Нахождение зависимостей в ГИ, прогнозирование, обнаружение и распознавание целевых стационарных и динамических свойств изучаемой среды. Решение этих задач требует разработки геоинформационной технологии, которая поддерживает комплексный анализ ГИ, обеспечивает высокую интерактивность анализа, позволяет интегрировать данные, распределенные на сетевых серверах и на ПК пользователя. Дополнительное требование связано с многообразием предметных областей анализа: система должна иметь открытую архитектуру и состоять из базового ядра и распределенных плагинов.

1. Введение

В настоящее время увеличивается количество новых источников цифровой ГИ, которая имеет и пространственную и временную компоненты. Это связано в первую очередь с развитием средств мониторинга природных и социально-экономических процессов. Одной из важнейших инициатив в этой области является рассчитанный на предстоящие десятилетия Проект по созданию Глобальной системы систем наблюдений за Землей (GEOSS), которая начинает разрабатываться в целях понимания и решения глобальных проблем окружающей среды и экономики (http://www.epa.gov/geoss/). Эти тенденции требуют своевременного создания новых геоинформационных технологий хранения, передачи и обработки огромных массивов данных, представляющих исследуемые процессы в пространстве и во времени. В сетевой распределенной ГИС ГеоТайм II сделаны первые шаги в направлении развития методов и средств исследования этой информации.

Аналитической геоинформационной системой (ГИС) называется универсальная совокупность средств сбора, хранения, обработки и исследования географической информации (ГИ) о пространственных и пространственно-временных объектах, явлениях и процессах. Аналитические геоинформационные технологии направлены на исследование пространственных и пространственно-временных данных и извлечение из них существенной географической информации и знаний (*Openshow and Openshow, 1997, Pyle, 1999, Malczewski, 1999, Miller and Han, editors, 2001, Klosgen and Zytkow, editors, 2002*).

ГИС ГеоТайм II предназначена для интерактивного представления, анализа и моделирования векторно-растровой географической информации о пространственно-временных процессах, которые описываются моделями с локальным взаимодействием. Характеристики таких процессов изменяются под воздействием прямых топологических или функциональных связей между компонентами пространственной структуры. В геоинформационных моделях этими компонентами являются, в частности, географические объекты или соседние элементы регулярной координатной сетки. Этим моделям соответствуют многие природные и техногенные процессы, такие, как сейсмотектонические взаимодействия, распространение водного стока и загрязнений в окружающей среде, подготовка природных катастроф, а также ряд социально-экономических процессов.

Система реализована в виде программы на языке Java. Для запуска апплета ГеоТайм II необходимо, чтобы была установлена виртуальная Java машина версии 1.6.

2. Данные

ГИС ГеоТайм II поддерживает операции с 2D, 3D и 4D сеточными и векторными данными. Обычно (но необязательно) данные

подготавливаются в виде ГИС-*проектов*. Проект представляет набор данных по некоторой пространственной области. Каждый проект удобно размещать в отдельном каталоге в виде набора файлов данных и метаданных. Для представления метаданных используется конфигурационный файл в формате XML.

Для загрузки (и сохранения) данных в ГИС ГеоТайм II используются наиболее распространенные форматы представления геопространственной информации. Для представления векторных данных используется разработанный ESRI формат SHP. Внешним форматом 2D и 3D сеточных данных является текстовый формат FLT, который является аналогом соответствующего формата ESRI.

3. Архитектура

Систему ГеоТайм II условно можно разбить на три подсистемы: хранения, обработки и визуализации данных.

Подсистема хранения данных ГИС ГеоТайм II разделена на пять частей. Пространственно-временные свойства исследуемой среды представлены в разделе *Исходная и обработанная ГИ*. Раздел поддерживает хранение ГИ в векторном и сеточном форматах. Векторные данные могут сопровождаться атрибутивной информацией и структурированы в виде информационных слоев, содержащихся в SHP-файлах. Сеточные данные представлены в виде прямоугольных 2D, 3D и 4D матриц. При этом каждый файл сеточных данных сопровождается XML-файлом метаданных и файлом палитры. Общий формат сеточных данных представляет четырехмерные массивы точек. В частном случае, когда требуется представить данные только в двух или трех координатах, четырехмерный сеточный массив вырождается в соответствующий двухмерный или трехмерный.

Процессор *компоновки ГИС-проекта* интерпретирует конфигурационный XML-файл метаданных проекта и обеспечивает динамическую загрузку данных и плагинов в любой момент, пока загружено ядро системы. Процессор *сохранения ГИС-проекта* обеспечивает сохранение на компьютере пользователя измененных конфигурационных файлов ГИС-проектов и новых полученных результатов.

Средства геоинформационного анализа пространственно-временных процессов, ориентированные на предметные области, представлены в разделе *Загруженные плагины*. Плагины представляют собой откомпилированные Java-классы, реализующие интерфейс GPPlugin, и подключаются динамически на этапе исполнения программы при помощи технологии Reflection (при этом не требуется ни изменять программный код самой системы, ни перекомпилировать её). Интерфейс определяет функции, реализуемые модулем: типы обрабатываемых модулем слоев (векторные или сеточные), имя модуля, метаданные модуля (описание функций модуля и алгоритм), вызов функции обработки. Модули могут располагаться как на стороне сервера, так и на стороне пользователя.

Подсистема визуализации данных поддерживает интерактивное изменение области визуализации и масштаба изображения, изменение порядка географических слоев, изменение параметров визуализации 2D, 3D и 4D сеточных и векторных слоев, анимационное и интерактивно управляемое представление сеточных слоев в виде динамически меняющихся кадров в любых двумерных проекциях, выполнение картографических измерений сеточных характеристик географических объектов, интерактивный просмотр и редактирование сеточных метаданных и векторной атрибутивной информации.

Подсистема обработки данных позволяет выполнять операции комплексной обработки и анализа данных, алгоритмы которых реализованы в виде плагинов. Аналитические преобразования реализуют функции применения заранее

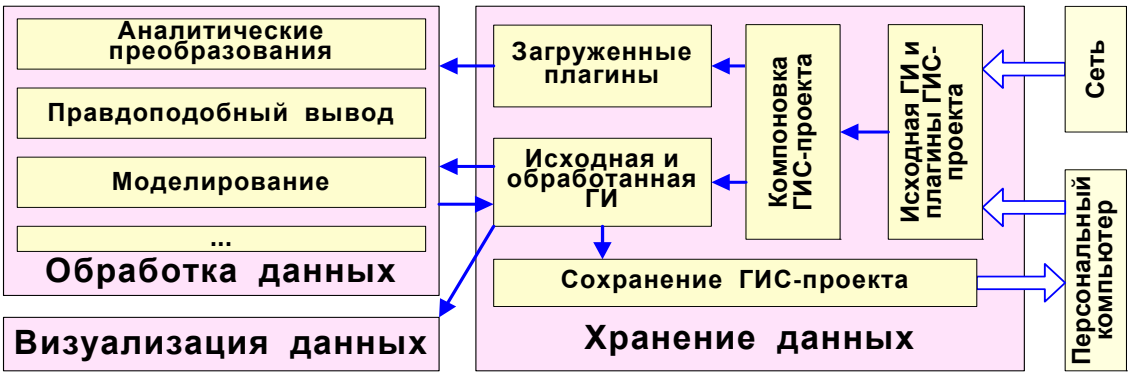


Рис.1. Схема ГИС ГеоТайм II

заданных операторов к имеющимся географическим данным для вычисления нового представления и выявления новых свойств географической информации. Правдоподобный вывод состоит в нахождении заранее неизвестного оператора аналитического преобразования по данным и знаниям, хранящимся в динамической памяти.

4. Сетевые функции

ГИС ГеоТайм II поддерживает возможность интеграции данных и плагинов, которые могут быть распределены на сетевых серверах и компьютере пользователя. Подключение распределенных данных и плагинов может быть заранее прописано в файле конфигурации системы или может осуществляться динамически с помощью диалогового окна. ГИС ГеоТайм II имеет следующие сетевые функции:

- Распределенные ГИС-проекты: возможность чтения ГИС-проектов размещенных на сетевых серверах и на компьютере пользователя.
- Распределенные ГИС-данные: возможность чтения данных размещенных на сетевых серверах и на компьютере пользователя и сборка ГИС-проекта на клиенте.
- Распределенные плагины: возможность динамического подключения Java модулей, размещенных на сетевых серверах и на компьютере пользователя.
- Сохранение ГИС-проекта: сохранение данных и метаданных на компьютере пользователя.

5. Аналитические возможности

Операции визуального исследования

ГИС ГеоТайм II запускается с подготовленными ГИС-проектами ИППИ РАН либо с портала «Аналитические ГИС online», http://www.geo.iitp.ru/ либо по запросу http://www.geo.iitp.ru/geotime/index.htm. Кроме того возможен режим запуска системы с ГИС-проектом, находящимся на компьютере пользователя. После запуска на экране возникает окно ГИС-проекта которое содержит одно или несколько *окон карты*, *окно геоинформационных слоев*, *окно атрибутов* слоя, *панель управления визуализаций и панель управления операциями*. Расположение и размер окон и панелей управления может изменяться пользователем.

Размер выводимого фрагмента карты совпадает с размерами *окна карты*. Внутри окна карты можно изменять масштаб изображения и перемещать увеличенный фрагмент карты. В *окне геоинформационных слоев* представлен список всех исходных и вычисленных слоев ГИС-проекта. Возможными операциями являются: выбор слоя для нанесения на карту, активизация слоя для выполнения над ним операций визуального анализа и изменения порядка следования слоев. При перемещении курсора в *окне карты* выводятся, координаты точки карты, совпадающие с положением курсора, и соответствующие значения активного сеточного слоя. *Окно атрибутов* предназначено для управления операциями картографического представления слоев.

Для визуального исследования сеточных слоев весьма важны операции изменения параметров закраски (прозрачность слоя, диапазон визуализируемых значений, интервалы закраски и цветовая закраска интервалов, выбор линейной, логарифмической или экспоненциальной шкалы изменения цветов). К операциям визуализации относится операция построения модели освещенности сеточного слоя, которая напоминает объемное представление поверхности. Операции визуализации 3D сеточного слоя поддерживают анимационное представление 2D срезов данных в трех плоскостях. Для пространственно-временных слоев анимационная визуализация возможна для срезов в трех плоскостях: долгота-широта XY (изменяется время T), долгота-время XT (изменяется широта X) и широта-время YT (изменяется долгота Y). Карта с анимационной визуализацией может быть сохранена в GIF-файле. Аналогичные операции поддерживаются для пространственных 3D сеточных слоев в плоскостях XY, XZ и YZ.

Параметрами визуализации векторных данных являются цвет, толщина линий и размер точечных объектов. При этом может быть выбран диапазон визуализируемых значений точек (например, диапазон магнитуд землетрясений), могут быть выбраны размеры, цвет заполнения и границы круговых пиктограмм, а также сдвиги положения надписей. Часто точечный слой представляет каталог землетрясений. Поэтому для точечного слоя реализована анимационная визуализации, при которой отображаются только те события, которые попадают в скользящее временное окно. Для землетрясений обычно указываются четыре координаты: долгота эпицентра X, широта Y, глубина Z и время события T. Для исследования этих сейсмических данных поддерживается по два варианта анимационной визуализации во всех возможных плоскостях XY, XZ, XT, YZ, YT, ZT. Например, при визуализации сейсмического процесса в плоскости XY можно задать интервалы глубин событий ΔZ и временной интервал ΔT. Далее, перемещая во времени интервал ΔT, можно анимационно визуализировать развитие процесса на интервале глубин ΔZ или, перемещая по

глубине интервал ΔZ, анимационно визуализировать изменение сейсмичности в заданном временном интервале по глубине. На Рис. 2. показан пример визуализации каталога землетрясений Камчатки с 1962 по 2004 гг. (http://www.geo.iitp.ru/geotime/kam4atka.html). На Рис. 3 показаны эпицентры землетрясений с глубиной от 50 до 100 км. В верхней части рисунка показано окно управления визуализацией по глубине. При визуализации перемещение интервала глубин 0 до 652 км сопровождается отображением соответствующего подмножества эпицентров. В данном случае с увеличением глубины «полоса» эпицентров при анимационной визуализации перемещается к северо-востоку. На Рис. 4 показаны эпицентры землетрясений в широтной области от 50º до 52º в проекциях XY и XZ.

При отображении слоя линий атрибутами визуализации являются толщина и цвет линий слоя. Для градусной сетки атрибутами визуализации являются цвет, параметры шрифтов и шаг сетки.

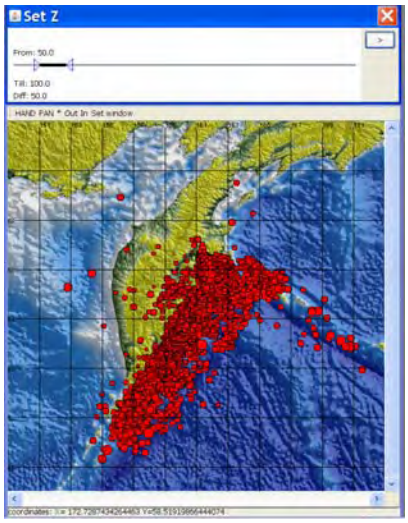


Рис. 2. Эпицентры землетрясений Камчатки с 1962 по 2004 гг.

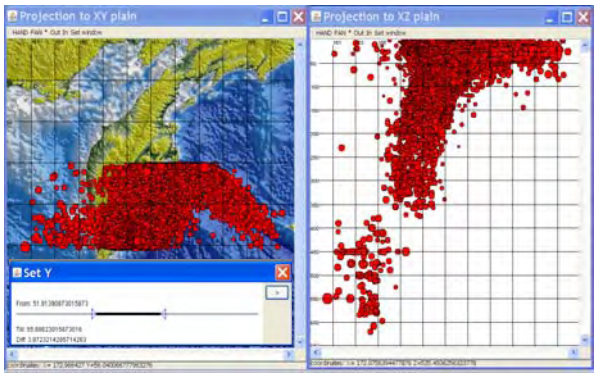


Рис. 3. Эпицентры землетрясений с 1962 по 2004 гг. с глубинами от 50 до 100 км.

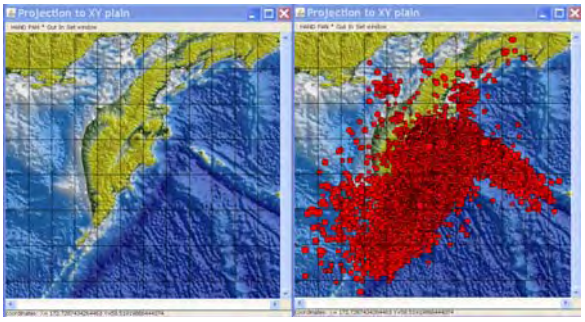


Рис. 4. Эпицентры землетрясений с 1962 по 2004 гг. в широтах от 50º до 52º в проекциях XY и XZ.

Операции обработки данных

Аналитические преобразования позволяют выявить новые свойства пространственных и пространственно-временных данных. В ГИС ГеоТайм II поддерживаются 3 типа аналитических преобразований: *векторный слой ⇨ сеточный слой, сеточный слой ⇨ сеточный слой и сеточный слой ⇨ векторный слой.*

Операции *векторный слой ⇨ сеточный слой*. Эти операции реализуют вычисления сеточного слоя по векторным слоям линий и точек. Линейные слои в ГИС ГеоТайм II всегда заданы на плоскости. Поэтому для линейных слоев могут вычисляться только сеточные 2D слои расстояний от узлов сетки до линий и близости узлов сетки к линиям. Точечные слои, например, каталог землетрясений, обычно заданы в координатах X, Y, Z, T. Для точечных слоев, вычисляются 2D, 3D, 4D сеточные слои расстояний от узлов сетки до множества объектов точечного слоя и близости узлов сетки к точкам слоя.

Для каталогов землетрясений поддерживается вычисление 3D слоев, которые представляют важнейшие характеристики сейсмического процесса в пространстве и во времени: минимальные представительные магнитуды, значения наклона графика повторяемости, сейсмическую активность и различные модификации критерия RTL [Соболев]. На Рис. 4 показаны карты параметров сейсмичности Камчатки на 19 июля 1990 г.: карта минимальной представительной магнитуды, карта значений наклона графика повторяемости и карта сейсмической активности для энергетического класса k=10. Карты построены с окном 150 * 150 км^2 и на временном интервале 5 лет. Вверху показана шкала закраски, внизу - панель управления анимацией, с помощью которой выбран соответствующий временной срез 3D сеточного слоя.

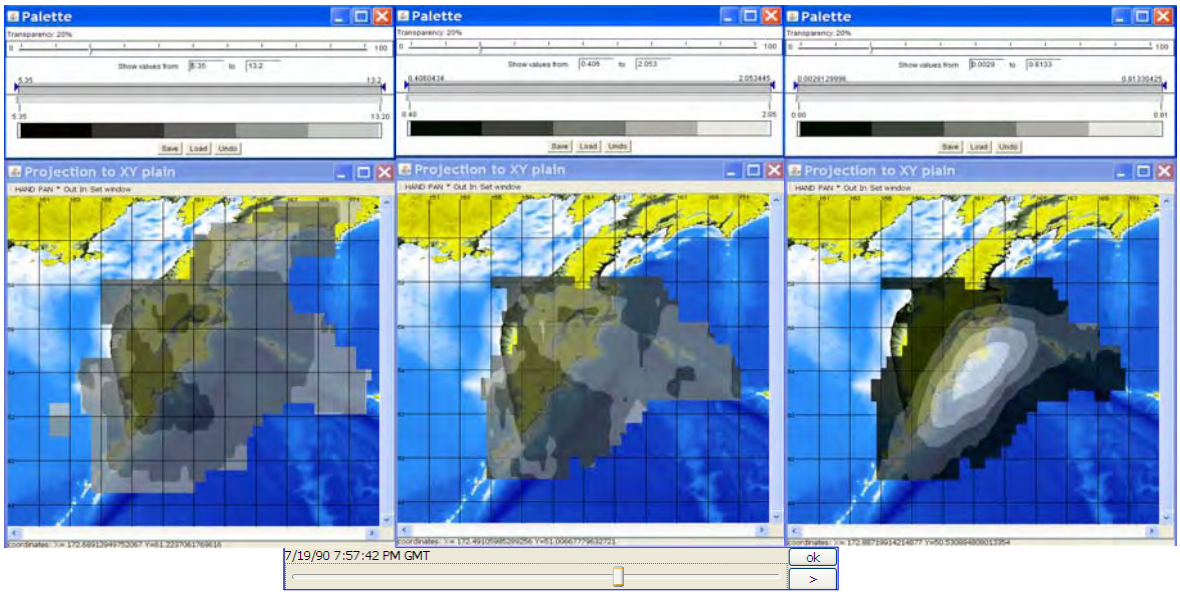


Рис. 5. Параметры сейсмичности Камчатки на 19 июля 1990 г.

(Слева-направо: карты минимальной представительной магнитуды, значений наклона графика повторяемости и сейсмической активности для энергетического класса k=10)

Операции *сеточный слой ⇨ сеточный слой*. Рассмотрим *операцию вычисления* 2D, 3D, 4D *сеточного слоя как функции от нескольких сеточных слоев* поддерживают генерирование 2D, 3D и 4D сеточного слоя с помощью оператора от нескольких исходных сеточных слоев. Оператор конструируется пользователем из элементарных функций с использованием алгебраических и логических операций. В одной операции могут быть использованы сеточные слои с разной размерностью 2D, 3D и 4D и с разным шагом сетки. При вычислениях все сеточные слои, выбранные в качестве аргументов, автоматически приводятся к одинаковой размерности и единой координатной сетке. По умолчанию размерность берется равной максимальной размерности слоя, взятого в качестве аргумента. При этом к слоям с меньшей размерностью добавляется соответствующая координата, по которой значения слоя просто повторяются на каждом шаге сетки. По умолчанию шаг единой сетки выбирается по каждой координате отдельно и равен минимальному шагу сеток используемых в преобразовании слоев. Изменение шага сетки слоев выполняется с помощью билинейной интерполяции. ГИС ГеоТайм II использует для конструирования преобразования следующие элементарные функции: абсолютная величина значений сеточного слоя, арктангенс, косинус, день, экспонента, квадратный корень, логическое выражение, кусочно-линейная функция, натуральные и десятичные логарифмы, максимальная или минимальная величины значений сеточных слоев, добавление к сеточному слою случайного шума и т.д.

Операции *фильтрации* 2D, 3D и 4D сеточных слоев выполняются для 2D срезов в любой из заданной пользователем плоскости XY, XT, YT XY, XZ, YZ. Эти операции позволяют вычислить статистические и геометрические характеристики сеточного слоя в выбранной плоскости среза. Комбинируя операции фильтрации и операции вычисления функций от нескольких сеточных слоев, можно реализовать более сложные сеточные преобразования. Часть операций фильтрации производится по значениям сеточного слоя внутри скользящего прямоугольного окна произвольного радиуса. Результат вычислений приписывается центру окна. Радиус окна 2D среза по координате в градусной сетке задается в км., по координате в декартовой сетке - в принятых в ГИС-проекте значениях координатной шкалы. В скользящем окне вычисляются среднее значение, медиана, среднеквадратичное отклонение, разность между значением в центре окна и средним значением в окне, максимум, минимум и разность между максимумом и минимумом, модуль и азимут градиента 2D среза, кривизны изолинии 2D среза. На Рис. 5. показана карта кривизны высот рельефа региона Камчатки. Красным цветом показаны выделенные с помощью данного преобразования хребты, а синим – тальвеги. Слева в качестве фона использованы карта высот рельефа с моделью освещенности, справа фоном является только модель освещенности.

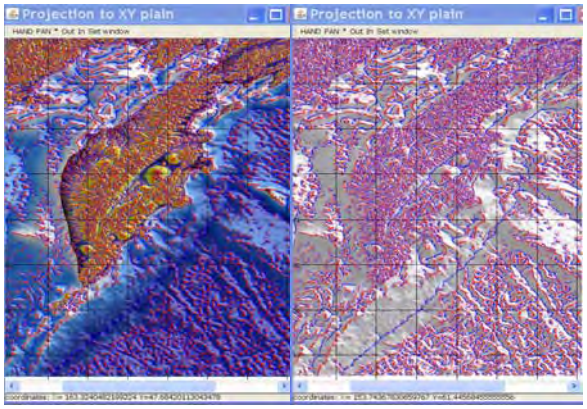


Рис. 6. Карта хребтов и тальвегов рельефа региона Камчатки

Следующее сеточное преобразование поддерживает генерирование сеточного слоя, равного значениям коэффициента корреляции двух 2D, 3D или 4D сеточных слоев в скользящем 2D окне, относящимся к 2D срезам в заданной плоскости. Например, пусть $\mathbf{A}(x,y,t)$ и $\mathbf{B}(x,y,t)$ – два 3D слоя, для которых требуется вычислить корреляцию по срезам (x,y). Результатом операции будет 3D слой $\mathbf{K}(x,y,t)$, который состоит из срезов $\mathbf{k}(x,y,t_n)$, значения которых равны коэффициентам корреляции соответствующих срезов $\mathbf{a}(x,y,t_n)$ и $\mathbf{b}(x,y,t_n)$. При вычислениях сеточные слои, автоматически приводятся к единой размерности и координатной сетке. По умолчанию шаг единой сетки выбирается равным минимальным шагам сеток используемых слоев. Преобразование слоев к более детальной сетке выполняется с помощью квадролинейной интерполяции. По умолчанию результирующий слой также представляется в наиболее детальной сетке.

Для обнаружения аномальных изменений значений 3D сеточных полей X,Y,T по координате T используются статистические методы обнаружения разладки для нескольких вероятностных моделей сеточных полей.

Операции сеточный слой ⇨ векторный слой. Данная операция используется для оценивания параметров ожидаемой катастрофы, в частности, землетрясения по 3D сеточному полю пространственно-временных аномалий. Этими параметрами являются пространственно-временные координаты зоны подготовки ожидаемого события и амплитуда предвестника события.

Прочие операции. Помимо базовых операций, на сервере доступны дополнительные модули: плагин для выделения кластеров сейсмических событий, плагин анализа последовательностей событий (GeoSeries), плагин моделирования поверхностного стока.

6. Пример анализа пространственно-временных предвестников Сусамырского землетрясения

Для анализа использовался каталог землетрясений со снятыми афтершоками землетрясений. Каталог покрывает территорию 40.4° - 44° СШ и 71° - 81° ВД и относится к периоду 1980 - 2001 гг. В каталоге представлено 16329 событии энергетического класса K от 7 до 17. Сусамырское землетрясение произошло 19.08.1992 г. энергетического класса K=17, координаты эпицентра λ=73.63° по долготе, и φ=42.06° по широте.

Попытки получения оценок параметров сейсмического режима, активности, наклона графика повторяемости и фрактальной размерности, оказались неудачными из-за недостаточного объема сейсмостатистики по отношению к временной протяженности искомой аномалии. Поэтому мы использовали наиболее простой параметр сейсмического потока – плотность эпицентров землетрясений, равный отношению числа землетрясений в окне оценивания к произведению площади окна на его временной интервал. При оценивании использовались следующие параметры окна: радиус $R_{\min}$=100 км, временной интервал $T_{\min}$=10 дней. Трехмерное сеточное поле плотности эпицентров построено в сетке с шагом по долготе $\Delta\lambda$=0.3°, по широте $\Delta\varphi$=0.2° и по времени ΔT=10 дней. Поле аномалий оценивалось для Гауссовской модели (9.3) с параметрами: T_1=1440 дней (10 дней, умноженных на 144 временных срезов динамического поля), T_2=30 дней (10 дней, умноженных на 3 временных среза).

На Рис. 7. на фоне модели освещенности рельефа показаны аномалии плотности эпицентров со значениями от -5.5 до -3.5 и относящиеся к 12 срезам динамического поля от 111 до 1 дня перед землетрясением. Отрицательные значения аномалии на рисунке свидетельствуют о том, что плотность землетрясений в окне T_2 меньше, чем в окне T_1. Можно видеть, что аномалии группируются только в окрестности эпицентра готовящегося землетрясения. Аномалия со значением менее -3.5 появляется за 101 день перед землетрясением. Площадь аномалии и численное значение монотонно увеличиваются и достигают максимума за 31 день до землетрясения. Затем аномалия убывает.

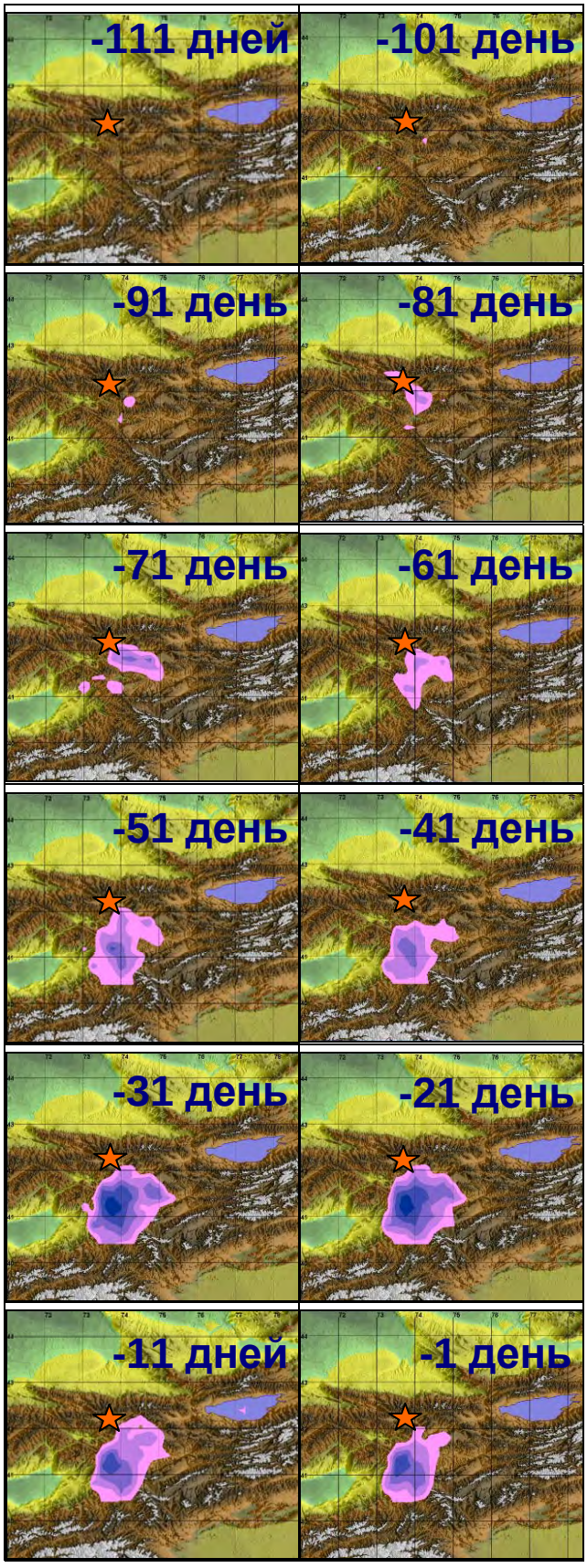


Рис. 7. Пространственно-временная аномалия плотности эпицентров землетрясений, предшествующая Сусамырскому землетрясению. Значения аномалии меняются от -5.5 до -3.5, положение эпицентра показано звездочкой.

7. Заключение

Аналитические сетевые ГИС технологии имеют широкий диапазон применений, начиная от визуального исследования и картографического измерения данных и кончая решением сложных задач прогнозирования пространственно-временных процессов. Сетевые ГИС характеризуются рядом свойств. Наиболее существенными из них являются типы обрабатываемых данных, глубина комплексного анализа ГИ, степень интерактивности операций анализа, степень доступности операций анализа для пользователя, не являющегося специалистом в информационных технологиях, возможность подключения данных, распределенных на серверах сети Интернет и в локальной сети, открытость архитектуры для подключения плагинов с целью настройки ГИС на предметную область и возможность сохранения пользователем результатов анализа.

ГИС ГеоТайм II используется в качестве одного из основных информационных ресурсов геоинформационной среды «Электронная Земля». ГИС предназначена для геоинформационного исследования пространственно-временных процессов. Анализ пространственно-временных процессов проводится с целью решения фундаментальных и прикладных задач исследуемой предметной области. К этим задачам относятся оценивание зависимостей и отношений между свойствами изучаемых процессов, прогнозирование заранее неизвестных целевых свойств, обнаружение целевых объектов и явлений, прогнозирование развития пространственно-временных процессов.

Система состоит из ядра и из подключаемых пользователем предметно-ориентированных плагинов. Она поддерживает на этапе выполнения программы интеграцию распределенных информационных ресурсов и с помощью подключения плагинов может настраиваться на соответствующую предметную область. Средствами анимационной визуализации ГИС поддерживает визуальное исследование 3D сеточных слоев и 4D каталогов точечных событий. Система обладает развитыми функциями аналитических преобразований. На примере Сусамырского землетрясения показано, что, средствами ГИС ГеоТайм II по каталогу землетрясений, можно с высокой достоверностью обнаружить пространственно-временную аномалию, связанную с процессом подготовки землетрясения.

Работа поддержана грантами РФФИ 04-07-90155, 06-07-89139, 07-07-12019 и программой Президиума РАН «Разработка фундаментальных основ создания научной распределенной информационно-вычислительной среды на основе технологий GRID», направление:

«Электронная Земля: научные информационные ресурсы и информационно-коммуникационные технологии».

8. Список литературы

1. *Гитис В.Г., Юрков Е.Ф. , Пирогов С А., Дерендяев А.Б., Метриков П А. 2008.* Комплексный анализ пространственно-временных процессов (в печати)
2. *Гитис В.Г., Юрков Е.Ф. , Пирогов С А., Дерендяев А.Б., Метриков П А. 2008.* Пространственно-временной анализ процессов подготовки геологических катастроф (в печати)
3. *Соболев Г А. 1993.* Основы прогноза землетрясений // М.: Наука. 313 с.
4. *Соболев Г А., Тюпкин Ю.С. 1996.* Аномалии в режиме слабой сейсмичности перед сильными землетрясениями Камчатки // Вулканология и сейсмология. №4. С. 64-74.
5. *Klosgen, W., Zytkow, J.M., editors. 2002.* Handbook of Data mining and Knowledge Discovery. // Oxford university press, 908 p.
6. *Malczewski, J. 1999.* GIS and Multicriteria decision analysis // Jon Willey &Sons. INC, 389 p.
7. *Miller, H.J., Han, J., editors. 2001.* Geographical data mining and knowledge discovery // Taylor & Francis, 367 p.
8. *Openshow, S., Openshow, C. 1997.* Artificial Intelligence in Geography. John Wiley&Sons. 329 p.
9. *Pyle, D. 1999.* Data preparation for data mining // Morgan Kaufman Published, Inc., San Francisco, California, 540 p.

Анализ эффективности применения методов интерполяции для доопределения законов функционирования сложных систем

Епифанов А.С.
Институт Проблем Точной Механики и Управления РАН, г.Саратов
desimalorlimewax@rambler.ru

Аннотация

Преобразование фазовых картин в геометрические образы законов функционирования автоматов позволило представлять фазовые картины едиными математическими структурами – ломаными линиями с числовыми координатами точек. Такое представление позволяет рассматривать не полностью определенные фазовые картины и доопределять их на основе классических методов интерполяции. В статье исследуется эффективность применения классических методов интерполяции по отношению к частично заданным геометрическим образам автоматов из подклассов класса (4,2,2) – автоматов в зависимости от выбранных узлов интерполяции.

.

1. Введение

Методы решения задач технического диагностирования, управления и синтеза систем используют задание законов функционирования систем. В фазовых картинах систематизированы и представлены фазовые траектории, определяющие поведение систем. Большое разнообразие фазовых картин Твердохлебовым В.А. в работах [4-6] сведено к единой форме – символьным и числовым графикам, точки которых размещаются на геометрических кривых линиях. Это позволяет не только сводить фазовые картины к удобной стандартной форме, но и использовать для постановок и решения задач мощные средства непрерывной математики: актуальную бесконечность, непрерывность, бесконечно малые величины, предельные переходы, суммирование бесконечных рядов и т.п. Существенной оказывается возможность доопределения фазовых картин, представленных частично заданными геометрическими образами, с использованием классических методов интерполяции.

2. Символьный и числовой геометрические образы законов функционирования автоматов

Базовой моделью является дискретная детерминированная динамическая система в виде автомата $A_S = (S, X, Y, \delta, \lambda, s)$, где S – множество состояний, X и Y – множества входных и выходных сигналов, $\delta : S \times X \to S$ – функция переходов, а $\lambda : S \times X \to Y$ – функция выходов. Уравнения динамики $s(t+1) = \delta(s(t), x(t))$ и $y(t) = \lambda(s(t), x(t))$ задают правила размещения на абстрактной целочисленной положительной оси времени сигналов и состояний. При начальном состоянии $s \in S$ поведение автомата A_S представимо автоматным отображением $\rho_S = \bigcup_{p \in X^*} \{(p, \lambda'(s, p))\}$ с расширением функции δ и λ до функций вида $\delta' : S \times X^* \to S$ и $\lambda' : S \times X^* \to Y$, где X^* и Y^* - множества всех конечных последовательностей элементов из множеств X и Y, по правилам $(\forall s \in S)(\forall x \in X)(\forall p \in X^*)\ \{\delta'(s, xp) = \delta'(\delta(s,x), p)\}$ и $(\forall x \in X)(\forall s \in S)(\forall p \in X^*)\ \{\lambda'(s, px) = \lambda(\delta(s,p), x)\}$.

Автоматное отображение ρ_S (множество пар) упорядочивается линейным порядком ω, определенным на основе порядка ω_1 на X^* и заданным следующими правилами (см., например, [4-5]):

Правило 1. На множестве X вводится некоторый линейный порядок ω_1 (который будем также обозначать $\prec$).

Правило 2. Порядок ω_1 на X распространим до линейного порядка на множестве X^*

- $(\forall p_1, p_2 \in X^*)\{|p_1| < |p_2| \to p_1 \prec p_2\}$;
- для любых $p_1, p_2 \in X^*$, для которых $|p_1| = |p_2|$ & $p_1 \neq p_2$, их отношение по порядку ω_1

повторяет отношение ближайших слева несовпадающих букв в словах p_1 и p_2. Аналогично определяется порядок ω_2 на множестве слов Y^*.

Линейно упорядоченное множество (ρ_s, ω) с учетом порядка ω_2 на Y естественно представляется графиком $G(\rho_s, \omega(\omega_1), \omega_2)$ (см. рис.1), который будем называть геометрическим образом законов функционирования автомата A_s и обозначать γ_s^A или γ_s.

На основании линейного порядка ω_1 на множестве X* и существовании наименьшего по порядку элемента $x_1 \in X^*$ каждый элемент $p \in X^*$ имеет номер $r_1(p) \in N^+$. Это позволяет геометрический образ размещать как в символьной системе координат с осями (X^*, ω_1) и (Y^*, ω_2), так и в прямоугольной декартовой системе координат на плоскости. Преобразование символьных координат точек геометрического образа определяется формулой

$$r(p) = \sum_{j=1}^{k} r(x_{i_j}) \cdot |X|^{j-1} - 1,$$

где $r(p)$ – номер слова p по порядку ω_1 на X^* (первая координата точки), а вторая координата определяется номером элемента $y \in Y$ по порядку ω_2.

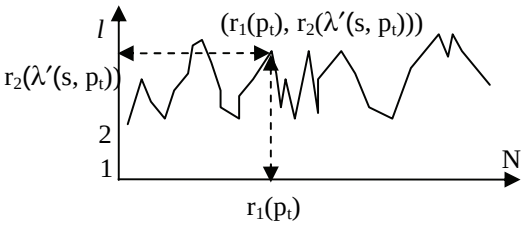


Рис.1. Числовые координаты точек геометрического образа законов функционирования автомата

Геометрический образ законов функционирования автомата является фазовой картиной, а отдельные фазовые траектории представлены сечениями геометрического образа (см.рис.2.).

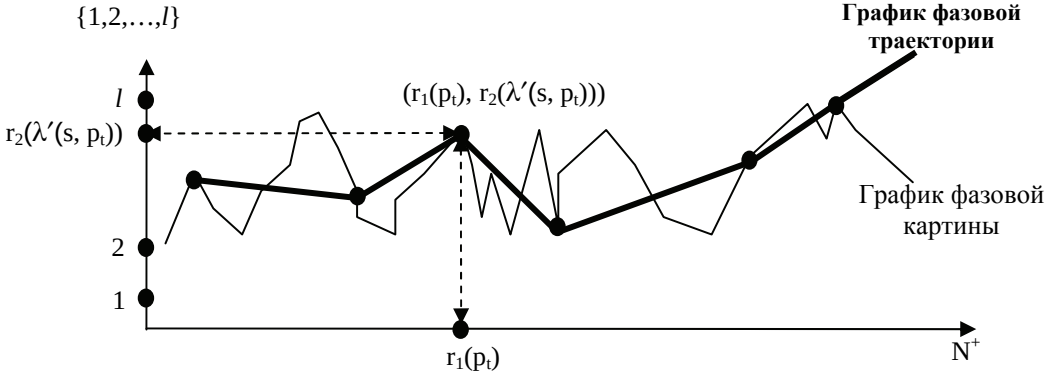


Рис.2. Числовые графики фазовой картины и фазовой траектории, построенные по геометрическому образу законов функционирования автомата A = (S, X, Y, δ, λ, s_0).

Геометрический образ конечного детерминированного автомата представляет собой числовую структуру – ломаную линию. Первая координата $r_1(p)$ точки геометрического образа $(r_1(p), r_2(\lambda'(s, p)))$ является номером соответствующего этой точке входного слова *p* по введенному линейному порядку ω_1 на множестве входных последовательностей X^*. Вторая координата $r_2(\lambda'(s,p)))$ точки $(r_1(p), r_2(\lambda'(s, p)))$ геометрического образа представляет собой номер последнего выходного сигнала $\lambda'(s,p)$ по порядку ω_2 на множестве Y в реакции $\lambda(s,p)$ инициального автомата на входную последовательность p. При зафиксированном порядке на множестве входных последовательностей специфика законов функционирования заключена только в последовательности вторых координат точек геометрического образа.

3. Эквивалентность задания законов функционирования геометрическими образами и числовыми последовательностями

В работе [4] приведены следующие теоремы, указывающие на эквивалентность задания законов функционирования автоматов с помощью геометрического образа и задания законов числовой последовательностью, совмещенной с любым конечным непустым множеством и любым линейным порядком на нем.

Теорема 1. Пусть $A_s = (S, X, Y, \delta, \lambda, s)$ - конечный детерминированный автомат, $X=\{x_1, x_2, \ldots, x_m\}$ и $\xi = \langle y_{i_1}, y_{i_2}, \ldots, y_{i_t} \rangle$ - последовательность вторых координат геометрического образа γ_s. Тогда последовательность ξ, множество X и ограничение порядка ω_1 на X*, использованное при построении γ_s, однозначно определяют t первых точек геометрического образа γ_s.

Доказательство. Пусть выполнены посылки теоремы. Тогда для заданных ограничения линейного порядка ω_1 и множества X однозначно определяется расположение первых по порядку ω_1 t элементов линейно упорядоченного множества (X, ω_1), где $t = |\xi|$. Вторые координаты первых t элементов линейно упорядоченного порядком ω множества (ρ_s,ω) однозначно определяются взаиморасположением элементов в последовательности ξ.

Теорема 2. Пусть Y – конечное непустое множество. Тогда для любых $\xi \in Y^*$, конечного непустого множества X и линейного порядка ω_1 на X* однозначно определяются t, где $t = |\xi|$, первых точек геометрического образа γ_s для частично определенного конечного детерминированного автомата с множеством входных сигналов X множеством выходных сигналов Y.

Доказательство. В достаточно очевидной теореме 1 утверждается, что для заданных автомата, линейного порядка на множестве его входных сигналов и известной последовательности ξ вторых координат точек геометрического образа однозначно определяются t первых элементов геометрического образа. В теореме 2 показывается, что для произвольных непустых конечных множеств X и Y, любого линейного порядка на X и любой последовательности $\xi \in Y^*$ однозначно определяются первые t, где $t = |\xi|$, элементов частично заданного конечного детерминированного автомата. При выполнении посылок теоремы 2 на основе заданного порядка ω_1 на X и его расширения по правилам 1 и 2 на множество X* однозначно определяются первые t элементов линейно упорядоченного множества (X^*,ω_1). Эти элементы полагаются первыми координатами первых t точек геометрического образа автомата. Вторые координаты точек однозначно определяются последовательностью ξ и порядком элементов в ней.

Из теорем 1 и 2 следует принципиально важный вывод: любая конечная последовательность элементов из конечного множества, совмещенная с любым конечным непустым множеством и любым линейным порядком на нем, однозначно определяет частично заданный конечный детерминированный автомат. Предложенный и разработанный Твердохлебовым В.А. новый геометрический подход к заданию законов функционирования дискретных детерминированных динамических систем позволяет рассматривать произвольную последовательность элементов из конечного множества с автоматной интерпретацией, т.е. произвольную последовательность можно рассматривать как последовательность вторых координат точек геометрического образа автомата. Для этого множеством состояний автомата полагается подмножество множества $\{s_p\}_{p \in X^*}$ и функция δ определяется равенством: $\delta(s_p, x) = s_{px}$ для любых $p \in X^*$ и $x \in X$. Каждая имеющаяся в геометрическом образе автомата точка (px, y), где $p \in X^*$, $x \in X$ и $y \in Y$, определяет равенство $\lambda(s_p,x) = y$.

Следовательно, совмещение законов функционирования δ' и λ' автомата $A' = \langle \{s_p\}_{p \in X^*}, X, Y, \delta', \lambda', s_\varepsilon \rangle$ с геометрическими фигурами (кривыми линиями на плоскости) может быть осуществлено на основе:

- выбора точек на геометрической фигуре, выбора направления обхода этих точек и построения множества Y вторых координат для выбранных точек;
- построения последовательности $y_{i_1}, y_{i_2}, \ldots, y_{i_t}, \ldots$ соответствующей обходу выбранных на геометрической фигуре точек;
- выбора целого положительного числа m, множества $X=\{x_1, x_2, \ldots, x_m\}$ и начального линейного порядка $x_1 \prec x_2 \prec \ldots \prec x_m$ для определения линейного порядка ω_1 на X^*;
- построение функций δ и λ с использованием указанных выше правил построения функций по точкам геометрического образа.

Таким образом, к математической структуре в форме последовательности сведены задания законов функционирования автоматов, соответствующих им фазовых картин и фазовых траекторий, представление процессов произвольной природы как последовательностей.

4. Интерполяция частично заданных геометрических образов конечных автоматов

Реальные технические системы представляют собой сложные системы, для которых

математическая модель определена частично. Определение изменения таких систем возможно связать с изменением их законов функционирования, которые допускают представление фазовой картиной, т.е. развитие таких систем в ряде случаев возможно изобразить изменением фазовой картины. Таким образом, изменение некоторых компонент в начальном фрагменте фазовой картины определяет преобразование фазовой картины. Полная фазовая картина или фрагмент фазовой картины на выбранном временном интервале может быть получена на основе использования классических методов интерполяции частичных геометрических образов законов функционирования и последующего преобразования образов в фазовую картину. Поведение системы в виде фазовой картины на сколь угодно удаленных от начала временных интервалах может быть получено на основе использования методов экстраполяции. Ввиду этого для эффективного представления результатов развития системы новыми законами функционирования системы требуется удачный выбор методов интерполяции и экстраполяции. В данной статье осуществляется исследование эффективности применения классических методов интерполяции и экстраполяции по отношению к частично заданным геометрическим образом автоматов из подклассов класса (4,2,2) автоматов, т.е. автоматов, имеющих 4 состояния, 2 входных и 2 выходных сигнала. Класс (4,2,2) автоматов содержит 16777216 автоматов. Выделение для анализа подклассов автоматов осуществляется на основе новой классификации конечных детерминированных автоматов, предложенной в работе [7]. Используемая классификация базируется на использовании для характеристики функций переходов δ и выходов λ автомата свойств пяти замечательных классов Поста (см.[7]), а также сочетаний этих свойств. Проведенное исследование анализа эффективности применения методов интерполяции и экстраполяции имеет следующую схему. Осуществлено построение начальных отрезков геометрических образов длины 62 (на словах до длины 5 включительно). Для методов интерполяции Ньютона и Лагранжа были выбраны 10 узлов интерполяции, соответствующих автономным подавтоматам, т.е. первым координатам таких узлов интерполяции соответствуют входные слова 0, 1, 00, 11, 000, 111, 0000, 1111, 00000, 11111. Для методов интерполяции Гаусса 11 узлов интерполяции были выбраны с постоянным шагом 6, начиная с 1 точки геометрического образа. Такой выбор узлов интерполяции для метода Гаусса обоснован тем, что он применим только в случае равноотстоящих узлов. При таком способе частичного задания геометрических образов были применены методы интерполяции. Полученные в результате интерполяции геометрические образы сравнивались с исходными, вычислялось количество правильно восстановленных точек. Эффективность методов интерполяции по отношению к выделенным подклассам автоматов рассчитывалась как отношение количества правильно восстановленных точек к общему числу точек в анализируемых начальных отрезках геометрических образов. Анализ полученных данных показывает, что применение методов Ньютона и Лагранжа к рассматриваемым классам автоматов дает различные результаты. Используя полученные данные возможно осуществить выбор наилучшего метода интерполяции по количеству правильных «восстановленных» точек в последовательностях вторых координат точек геометрических образов автоматов для каждого из рассмотренных классов автоматов. Недостатком метода интерполяции Гаусса в отличие от рассмотренных методов интерполяции Ньютона и Лагранжа является ограничение возможности его использования только для случая равноотстоящих узлов интерполяции.

5. Интерполяция частичных фазовых картин, представленных числовыми последовательностями

В данной части статьи осуществляется исследование эффективности применения методов интерполяции Ньютона и Лагранжа по отношению к частично заданным фазовым картинам, представленным в виде частично заданных геометрических образов автоматов. Последовательностями вторых координат точек исследуемых геометрических образов полагаются последовательности, задающие приближения фундаментальных математических величин (π, e, $\sqrt{2}$, φ и т.д.). Рассматривается множество последовательностей Ψ, состоящее из 7000 последовательностей длины 62, выбранная длина соответствует начальному отрезку геометрического образа автомата из класса (4,2,2) автоматов на словах до длины 5 включительно. Для методов интерполяции Ньютона и Лагранжа были проведены вычислительные эксперименты, анализирующие эффективность метода в зависимости от выбранных узлов интерполяции. Были проанализированы 3 различных набора узлов интерполяции:

1) 10 узлов, соответствующих автономным подавтоматам, т.е. первым координатам таких

узлов интерполяции соответствуют входные слова 0, 1, 00, 11, 000, 111, 0000, 1111, 00000, 11111;

2) 10 узлов, номерами первых координат которых являются числа 1,7,13,19,25,31,37,43,49,55, т.е. узлы выбраны с постоянным шагом;

3) 10 узлов, номерами первых координат которых являются числа 1,2,3,4,5,6,7,8,9,10 (т.е. данные узлы формально могут считаться набором узлов для экстраполяции).

В результате анализа полученных данных отмечено, что в целом по отношению ко всему рассматриваемому множеству последовательностей Ψ метод интерполяции Лагранжа дает одинаковые результаты при 1 и 2 наборах узлов. Число правильно восстановленных точек для всех последовательностей множества Ψ с помощью метода Лагранжа при 1 и 2 наборах узлов равно 28%, при третьем наборе узлов интерполяции метод правильно восстанавливает лишь 17% точек. Метод интерполяции Ньютона также по отношению к двум рассматриваемым наборам узлов интерполяции дал одинаковые результаты (разница составляет менее 0,1%) при 1 и 2 наборах узлов. В целом по отношению ко всему множеству Ψ метод Ньютона при 1 и 2 наборах узлов оказался эффективнее метода Лагранжа на 4%, т.е. количество правильно восстановленных точек 32%. В результате интерполяции методом Лагранжа правильно восстановлено всего 14 последовательностей из множества Ψ при первом наборе узлов интерполяции, 6 при втором наборе узлов интерполяции и 2 при третьем наборе.

6. Краткие выводы

Имеющиеся неопределенности в математических моделях сложных систем в ряде случаев могут быть устранены методами интерполяции. Для этого могут быть использованы частично заданные геометрические образы автоматов с числовыми координатами точек. Возможность и эффективность интерполяции частично заданных геометрических образов исследованы не достаточно. В данной статье на примере класса (4,2,2) автоматов, его подклассов, выделенных на основе свойств Поста для функций переходов и выходов, показано, что для рассматриваемых подклассов автоматов метод интерполяции Гаусса дает лучшие результаты, чем методы интерполяции Ньютона и Лагранжа. Для каждого из выделенных 15 непустых подклассов класса (4,2,2) автоматов из 2-х методов интерполяции (Ньютона и Лагранжа) определен наиболее эффективный метод.

На примере фундаментальных математических величин, в предположении, что они задают законы функционирования исследована эффективность методов интерполяции Ньютона и Лагранжа.

Таким образом, показано, что классические методы интерполяции применимы для доопределения законов функционирования дискретных детерминированных автоматов, возможен выбор эффективного метода интерполяции и модели сложных систем могут доопределяться классическими методами интерполяции.

7. Литература

[1] Елисеев К.В., А.С.Епифанов “Интерполяция в доопределении законов функционирования больших систем”, *2-ая школа-семинар молодых ученых «Управление большими системами»,Сборник трудов 2 конференции*, том 2 , Воронеж, Научная книга,2007.

[2] А.С.Епифанов “Интерполяция фазовых картин дискретных детерминированных систем”, *Радіоелектронні і комп’ютерні системи*, Харків, 2008, №5, С. 128-132.

[3] В.А. Твердохлебов “Геометрические образы конечных детерминированных автоматов”, *Известия Саратовского ун-та (Новая серия)*, Т.5. Вып.1,Саратов, 2005, С.141-153.

[4] В.А. Твердохлебов “Спектры числовых характеристик фазовых картин объектов диагностирования”, *Радіоелектронні і комп’ютерні системи*, Харків, 2008, №5, С. 148-155.

[5] В.А. Твердохлебов “Методы интерполяции в техническом диагностировании”,*«Проблемы управления»*, Москва, №2, 2007, С.28-34.

[6] В.А. Твердохлебов “Специфика технического диагностирования сложных дискретных систем ”, *Материалы международной конф. «Автоматизация проектирования дискретных систем»*,Т.2, Минск, 2007, с.207

[7] В.А. Твердохлебов, А.В. Пономаренко “Классификация конечных автоматов по свойствам функций переходов и выходов”, *Сб. научных трудов ИПТМУ РАН "Проблемы точной механики и управления"*,Саратов, 2004, С. 16-25.

Решение задачи нахождения соответствующего эталона методом нейронного кортекса

Сергийчук М.Н.
МФТИ
sergiuchuckmisha@gmail.com

Михайлов А.М.
ИПУ РАН
alx_mikh@yahoo.com

Аннотация

В докладе представлен алгоритм нейронного кортекса для решения задачи нахождения соответствующего эталона в случае данных большой размерности. Метод нейронного кортекса реализует идею снижения размерности данных, используемых при решении задачи поиска, путем выделения из всей базы запомненных шаблонов актуальной ее части. Метод основан на моделировании когнитивных информационных процессов, происходящих в коре головного мозга. В докладе также представлены результаты сравнения быстродействия алгоритма нейронного кортекса и базового алгоритма "Full Search". Показано что в случае больших размерностей рассматриваемых данных, а также большого числа запомненных шаблонов алгоритм нейронного кортекса значительно эффективнее (реализовано ускорение более чем в 200 раз) по сравнению с алгоритмом "Full Search" и, в зависимости от управляющих параметров, результат может быть еще эффективнее.

Введение

Задача нахождения соответствующего эталона является актуальной проблемой, находящей практическое применение во многих задачах компьютерного зрения и, в частности, в видеонаблюдении.

Известно, что человек решает задачу нахождения соответствующего эталона значительно лучше, чем современные алгоритмы, которые с ней справляются лишь в сотни раз быстрее чем Full Search алгоритм поэлементного сравнения.

Целью данной работы является исследование возможности практического применения запоминания информации в индексной форме [2] для решения задачи нахождения соответствующего эталона.

Существует немало алгоритмов решения задачи нахождения соответствующего эталона. Например, в [6] предложена модификация Full Search подхода связанная с использованием результатов рассмотрения предыдущих шаблонов.

Традиционно производительность алгоритмов решения задачи нахождения соответствующего эталона сравнивается с Full Search подходом. В [6] получено ускорение до 200 раз в зависимости от исходных данных. Соответствующая функция из Open Computer Vision Library дает ускорение до 80 раз [6].

Таким образом, можно считать успехом, если удастся получить ускорение порядка сотен раз.

Также существуют методы снижения размерности данных такие как хэширование или кодирование, среди которых можно выделить метод вэйвлетов, можно рассматривать как способы решения задачи нахождения признаков, и, признавая их эффективность, можно рассматривать их не как конкурентов а как партнеров в том смысле что можно выделять признаки объектов, например, с помощью вэйвлетов, а распознавать полученный вектор признаков с помощью индексной памяти [2].

Используемый алгоритм решения задачи нахождения соответствующего эталона построен исходя из концепции индексной памяти, предложенной А.М.Михайловым в статьях [1-3]. При чем эта концепция основана на некоторых принципах работы неокортекса.

1. Постановка задачи нахождения соответствующего эталона

Пусть у нас есть база запомненных шаблонов, при чем каждый **шаблон характеризуется вектором свойств**. Примером таких свойств могут быть цвета пикселей изображения. Тогда длина вектора свойств равна количеству пикселей, а цвет пикселя – это значение соответствующего свойства. Идентифицируемый **объект** также

рассматривается как вектор тех же **свойств**. Для идентифицируемого объекта требуется найти наиболее *похожий* на него шаблон, а также меру их соответствия друг другу (смотри рисунок 1). При чем термин “похожий шаблон” мы будем понимать как шаблон, у которого количество свойств, которые отличаются от соответствующих свойств идентифицируемого объекта, не больше чем на заранее определенную величину **r** (**r** – это один из управляющих параметров) (в дальнейшем будем говорить, что соответствующие свойства должны совпадать).

2. Теоретические основы индексного метода решения задачи нахождения соответствующего эталона

2.1. Формальное определение индекса

Термин “индекс” может обозначать разные вещи: численную величину, отношение, палец или алфавитный список слов с указанием страниц где они упоминаются. Нас интересует именно последнее значение: книжный индекс. Книжный индекс состоит из списка слов, причем каждому слову соответствует список страниц (Рис.2).

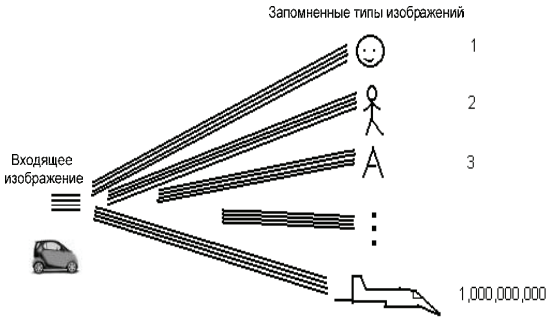


Рис.1. Задача нахождения соответствующего эталона

			277	115		11
170	222		206	90	25	34
90	170	153	191	72	23	24
Cycle	Levy	Lick	Light	Life	Log	Logic

Рис.2. Книжный индекс

Формальное определение книжного индекса и обозначение$\{x\}_z$ приведено в [1]. Пусть мы запомнили X шаблонов, при чем каждый шаблон характеризуется вектором из Y свойств и **динамический диапазон** изменения каждого признака равен Z. Тогда для заданного свойства y $\{x\}_z$ означает множество всех запомненных шаблонов у которых значение y-свойства равно z. В соответствии с [2] такое множество мы будем называть *колонкой* с адресом (y,z). Содержимое колонки являет собой множество. В отличие от матриц, которые характеризуются двумя размерностями (числом строк и столбцов), индекс характеризуется только одной размерностью – числом столбцов. Вертикальная размерность неопределена, так как мощность множеств может быть произвольная.

Индекс – средство запомининия множеств и подчиняется соответствующим ограничениям. Например, поскольку каждая колонка – множество, то она не может содержать одинаковые элементы.

2.2. Концепция индексной памяти

Индексная память предполагает схему хранения запомненных шаблонов отличную от стандартного записывания шаблонов подряд. Данные хранятся в виде *индексов*.

Рисунок 3 показывает сравнение традиционного метода хранения шаблонов и индексного подхода. Считаем, что **горизонтальная** плоскость являет собой используемое **адресное пространство**.

Слева показан традиционный и естественный способ хранения запомненных шаблонов. В памяти подряд записаны наборы признаков для каждого шаблона. При этом пара (номер шаблона, номер признака) однозначно определяет адрес, по которому хранится значение соответствующего признака.

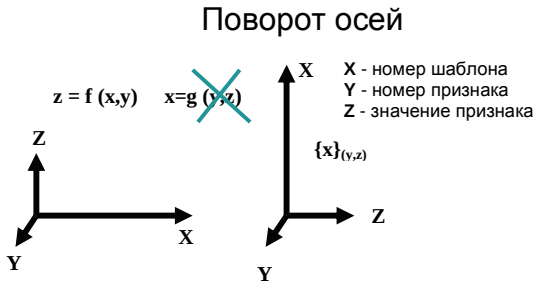


Рис.3. Поворот осей

Справа на рисунке 2 проиллюстрирован индексный способ хранения данных. Теперь адресом служит пара (номер признака, значение признака), при чем по этому адресу хранится *множество* всех шаблонов, у которых значение соответствующего признака совпадает с адресным.

Следует обратить внимание, что в индексной памяти по соответствующему адресу хранится *множество*, а не конкретное значение (не функция, что отражено на рис.3), при чем

оказывается, что наиболее удобной структурой хранения множеств является список. При чем, процедура добавления очередного шаблона в базу данных выполняется естественно и не требует пересмотра всей базы.

2.3. Индексный метод решения задачи нахождения соответствующего эталона

Пусть мы запомнили X шаблонов, при чем каждый шаблон характеризуется вектором из Y свойств и **динамический диапазон** изменения каждого признака равен Z.

Процедура решения задачи нахождения соответствующего эталона, использующая свойства индексной памяти, следующая:

Для каждого свойства выписываются все номера запомненных шаблонов, у которых значение соответствующего свойства совпадает с значением этого свойства у идентифицируемого объекта. Таким образом выписано Y множеств (Y колонок). Тогда шаблон, который встречается чаше всего среди выписанных номеров, и есть решение задачи, поставленной в пункте 1.

Обращаю внимание на схожесть предложенного метода и преобразования Радона.

2.4. Экономия вычислений

Оценим совокупный объем выписанных множеств. Если считать что признаки распределены равномерно по всему динамическому диапазону их изменения, то мощность одной колонки примерно равна X/Z и количество выписанных номеров шаблонов примерно равно XY/Z. Теперь вспомним что общий объем запомненной информации равен XY (X шаблонов по Y свойств в каждом) и что при Full Search подходе для решения задачи требуется перебрать всю запомненную информацию. Таким образом получается что для решения задачи нахождения соответствующего эталона индексным методом используется в Z раз меньше информации чем при Full Search подходе.

Также показано [13], (Приложение 1), что в случае если распределение значений каждого признака **y** у шаблонов и у идентифицируемого объекта существуют и совпадают, то доля используемой информации может быть оценена по формуле (1),

$$k = \overline{k_y} \qquad (1)$$

где черта сверху означает усреднение по всем признакам, а $\mathbf{k_y}$ определяется формулой (2)

$$k_y = \frac{\left[\sum_z \left[n_y(z)\right]^2\right]}{\left[\sum_z n_y(z)\right]^2} \qquad (2)$$

где $\mathbf{n_y(z)}$ – число запомненных шаблонов у которых значение признака с номером **y** соответствует **z**.

Величину, обратную коэффициенту (1) естетсвенно назвать *эффективным динамическим диапазоном*.

Оказывается, что эффективный динамичесий диапазон не больше чем теоретический Z, а обычно, существенно меньше. То есть в реальности дела обстоят хуже, чем в теории ☺

2.5. Декомпозиция многомерных индексов.

Можно доказать теорему о том, что многомерный индекс $\{x\}_{(a,\ b,\ldots,z)}$ (индекс, построенный для многомерного признака) эквивалентен объединению соответствующих одномерных индексов [2].

$$\{x\}_{(a,\ b,\ldots,z)} = \{y\}_a \cap \{y\}_b \cap \ldots \cap \{y\}_z \qquad (3)$$

Это дает возможность изменять динамический диапазон изменения признаков изменяя размерность рассматриваемых признаков. К примеру, если объединить признаки попарно и рассматривать пары как двумерные признаки, то количество признаков сократится вдвое до Y/2, а динамический диапазон двумерного признака составит Z^2 с соответствующей экономией вычислений.

4. Результаты численных экспериментов

Были поставлены численные эксперименты для сравнения производительности индексного алгоритма и Full Search подхода.

2.1. Используемые данные

В качестве объектов использовались машины, идентифицированные на фотографиях уличного движения (рис.4). Обсуждение метода идентификации объектов выходит за рамки доклада. В качестве свойств использовались вертикальные и горизонтальные проекции [11] (рис.5). То есть, вычислялись суммы яркостей пикселей по столбцам и рядам. Таким образом объект рассматривался как 32 вертикальные плюс 32 горизонтальные проекции (всего **Y=64**). Значение каждой проекции округлялось до динамического диапазона **Z=256**. Была

сформирована база запомненных шаблонов из **X=700000** объектов.

2.1. Результаты

Задача нахождения соответствующего эталона решалась в 2 режимах:

1. Объект рассматривался как 64 признака, динамический диапазон изменения признака Z=256;
2. Объект рассматривался как 32 двумерных признака (смотри пункт 2.5), динамический диапазон изменения признака $Z=256^2$.

Рис.4. Пример используемых данных

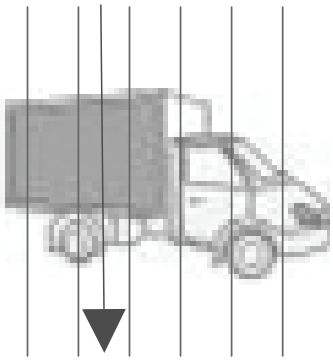


Рис.5. Пример вертикальной проекции

Также соответствующая задача решалась с помощью Full Search подхода и сравнивалась производительность.

Полученные результаты отражены в таблице 1. Приведены теоретический динамический диапазон (256 и 256^2), измеренное ускорение по сравнению с Full Search подходом, величина, обратная измеренной доле использованной при решении задачи информации, которую естественно назвать *реальным динамическим диапазоном*, оцененный *эффективный динамический диапазон* (приводится величина обратная коэффициенту (1)).

Следует отметить, что реальное ускорение оказалось значительно ниже, чем экономия обрабатываемой информации.

Для лучшего понимания смысла концепции эффективного динамического диапазона и формулы (1) на рисунке 6 представлена гистограмма распределения совокупности всех признаков (показана сумма всех гистограмм, показывающих количество шаблонов с данным значением соответствующего признака), а также, голубой линией, показан оцененный эффективный динамический диапазон (величина обратная коэффициенту (1), если его вычислять для соответствующего индекса). В данном случае он оказался равным 98.8 что значительно больше чем эффективный динамический диапазон одного признака, поскольку распределение совокупности шире, чем распределение составляющих.

Если гистограмму рассматривать как плотность вероятности (после соответствующей нормировки) то эффективный динамический диапазон даст оценку размеров области, в которой сосредоточено распределение.

Табл.1. Результаты численных экспериментов

динамический диапазон	ускорение	реальный динамический диапазон	эффективный динамический диапазон
256	2,6	72,3	71,7
65536	210,2	5547,8	4983,2

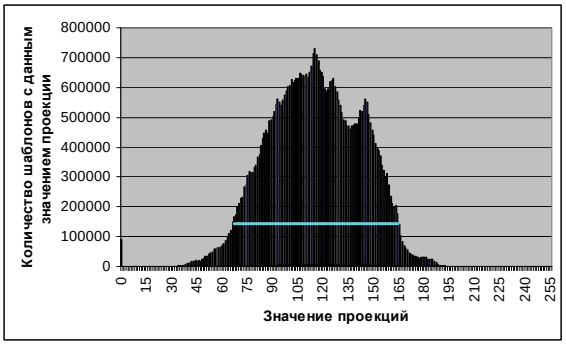


Рис.6. Гистограмма совокупного распределения признаков

Выводы

- Программно реализована предложенная А.М.Михайловым концепция индексной памяти.
- Показано что применение индексной памяти может привести к ускорению более чем 200 раз по сравнению с Full Search подходом.
- Показано что реальная доля используемой информации значительно выше, чем теоретически предсказано в [2] и она хорошо описывается предложенной формулой (1).

Подводя итог, хочу сказать, что концепция индексной памяти является эффективным методом решения задачи нахождения соответствующего эталона, хотя для каждой конкретной задачи требуется выбрать оптимальный набор признаков, а также управляющих параметров.

Приложение 1. Доказательство формулы (1)

Коэффициент эффективности – это доля информации, которую требуется обработать для решения задачи нахождения соответствующего эталона индексным методом.

Для начала решим следующую задачу. Пусть в индексной памяти запомнено некоторое количество шаблонов **X**. Переменной **z** будем обозначать значение признака. Динамический диапазон каждого признака равен **Z**. Рассмотрим некоторый признак **y**. Пусть $\mathbf{n_y(z)}$ – количество запомненных шаблонов, у которых этот признак равен **z**.

Допустим, мы распознаем не неизвестные объекты, а один за другим все уже запомненные шаблоны. Подсчитаем сколько номеров шаблонов придется выписать при решении задачи нахождения соответствующего шаблона индексным методом и только с помощью рассматриваемого признака. Рассмотрим некоторое значение **z**. Для всех шаблонов, у которых значение рассматриваемого признака равно **z,** нам придется выписать **n(z)** номеров шаблонов. А всего таких шаблонов **n(z)**. Таким образом, для каждого **z** нам придется выписать **n(z)*n(z)** номеров шаблонов. Соответственно всего нам придется выписать $\sum_z \left[n_y(z)\right]^2$ шаблонов, при чем суммирование ведется по всему динамическому диапазону изменения признака. Соответственно в среднем для распознавания одного шаблона из индексной памяти потребуется считать $\frac{\sum_z \left[n_y(z)\right]^2}{X}$ номеров шаблонов. Теперь обратим внимание, что

$$\sum_z n_y(z) = X \qquad (4)$$

так как каждый номер должен попасть в некоторую колонку, слева записана суммарная мощность всех колонок, а справа весь объем памяти.

Таким образом, получаем, что в среднем для решения задачи нахождения соответствующего эталона индексным методом потребуется $\frac{\sum_z \left[n_y(z)\right]^2}{\sum_z n_y(z)}$ обращений к памяти. А это количество обращений к памяти можно получить, умножив всю память X (напомню, что пока что рассматривается единственный признак **y**) на

$$\frac{\frac{\sum_z \left[n_y(z)\right]^2}{\sum_z n_y(z)}}{X} = \frac{\left[\sum_z n_y(z)\right]^2}{\sum_z \left[n_y(z)\right]^2}$$

число равное

То есть можно записать что

$$\mathbf{V_{index}=V_{FullSearch}*k} \qquad (5)$$

где $\mathbf{V_{index}}$ - это объем обработанной памяти при решении задачи нахождения соответствующего эталона индексным методом, $\mathbf{V_{FullSearch}}$ – это вся память, равная **X*Y** запомненных значений признаков, $\mathbf{k}$ – это коэффициент эффективности, показывающий во сколько раз меньше памяти мы используем при индексном подходе. Если рассматривать только один признак, то получилось что

$$k_y = \frac{\left[\sum_z \left[n_y(z)\right]^2\right]}{\left[\sum_z n_y(z)\right]^2} \qquad (2)$$

До сих пор мы рассматривали только один признак. Посмотрим, что изменится, если решать задачу нахождения соответствующего диапазона, учитывая **Y** число признаков. Суммирование по **y** означает суммирование по всем признакам.

Опять мы последовательно один за другим распознаем каждый из запомненных шаблонов. Количество выписанных номеров шаблонов найдем, просто просуммировав по всем признакам уже полученное число шаблонов для одного признака (2) и получим $\sum_y \sum_z \left[n_y(z)\right]^2$

Учитывая, что теперь объем всей памяти равен $\sum_y \sum_z n_y(z) = X \cdot Y$ получаем, что среднее количество выписанных номеров для распознавания одного шаблона равно

$\frac{\left[\sum_y \sum_z \left[n_y(z)\right]^2\right]}{\left[\sum_y \sum_z n_y(z)\right]}$, и чтобы его получить, нужно весь объем памяти умножить на коэффициент эффективности

$$k = \frac{\left[\sum_y \sum_z \left[n_y(z)\right]^2\right]}{\left[\sum_y \sum_z n_y(z)\right]^2} \quad (6)$$

Формулу (6) можно переписать в виде

$$k = \frac{\overline{\left[\sum_z (n(z))^2\right]}}{\left(\sum_z n(z)\right)^2} \quad (7)$$

где черта сверху означает усреднение по всем признакам, или в виде

$$k = \overline{k_y} \quad (1)$$

где $\mathbf{k_y}$ определяется формулой (2), а черта сверху тоже означает усреднение по всем признакам.

Наконец, если допустить, что распределение каждого признака идентифицируемого объекта существует и это распределение равно распределению соответствующего признака, порожденному сохраненной базой шаблонов как генеральной совокупностью, то матожидание доли информации, которую требуется обработать для решения задачи нахождения соответствующего эталона будет по определению совпадать с коэффициентом (1).

Литература

[1] A.M. Mikhailov, “Real-time target tracking using neural cortex technology”, Proceedings of the Artificial Neural Network in Enginering Conference, 2001, Vol.11.

[2] A.M. Mikhailov, Y.M. Pok “Artificial Neural Cortex”, Proceedings of the artificial Neural Network in Enginering Conference,2001, Vol.11, pp.113-120.

[3] A.M. Mikhailov, “Biologically Inspired Digital Cortex Architecture and its Formalism”, Singapore, 2006.

[4] J.Hawkins “On Intelligence”, Times Books, 2004, ISBN 0-8050-7456-2

[5] D.Specht, R.LAMM, “Real-time target reinforcement learning for target tracking”, Proceedings of the Artificial Neural Network in Enginering Conference, 2001, Vol.11, pp.57-62.

[6] F.Tombari, S.Mattoccia, L. Di Stefano, “*Template Matching Based on the L_p Norm Using Sufficient Conditions with Incremental Approximation”*, Video and Signal Based Surveillance, AVSS IEEE International Conference, 2006, p.20. ISBN: 0-7695-2688-8.

[7] T. Kaneko, O. Hori, *”Feature selection for reliable tracking using template matching”* Computer Vision and Pattern Recognition, Proceedings, 2003 IEEE Computer Society Conference, Volume: 1, On page(s): I-796- I-802 vol.1ISSN: 1063-6919 ISBN: 0-7695-1900-8.

[8] A. Zhen Jia Balasuriya, S. Challa, *“Target tracking with Bayesian fusion based template matching”* Image Processing, 2005. ICIP 2005. IEEE International Conference Volume: 2, On page(s): II- 826-9 ISBN: 0-7803-9134-9.

[9] K.Tsunoda, Y.Yamane, *“Complex Objects are Represented in Macaque Inferotemporal Cortex by the Combination of Feature Columns”* Nature, Neuroscience; 2001, 4(8); p 832-838.

[10] D.J.Felleman, Van Essen, *“Distributed hierarchical processing in primate cerebral cortex”* Cerebral Cortex, 1991, 1:1-47.

[11] A.Specht, R.Lamm, *“Real-Time Reinforsment Learning for Target Tracking”* Proceedings of the Artificial Neural Networks in Engineering Conference, 2001, Vol. 11, pp. 57- 62.

[12] B. Vernon Mountcastle, “*An Organizing Principle for Cerebral Function: The Unit Model and the Distributed System”* MA: MIT Press, Cambridge, 1978.

[13] М.Н.Сергийчук, “Индексный подход и его применение к проблеме слежения за движущимися объектами”, МФТИ, 2008.

Интерактивные учебные пособия на основе средств виртуальной реальности в CALS – технологиях

Ромакин В.А., Балабанов А.В.
Институт проблем управления РАН, Москва
insight.ru@gmail.com, fca07@mail.ru

Аннотация

Целью настоящей статьи является рассмотрение проблем, связанных с эффективным применением средств виртуальной реальности для поддержки продукции на стадии эксплуатации.

Интерактивные учебные пособия на основе средств виртуальной реальности (ИУП ВР) являются эффективным средством обучения. Преимущества использования и особенности разработки ИУП ВР, а также примеры их применения в сфере образования подробно описаны в [1], [2], [3]. В настоящей статье рассматриваются проблемы применения ИУП ВР в технологиях непрерывной поддержки жизненного цикла продукции (CALS-технологиях).

Ввод в эксплуатацию и техническое обслуживание в процессе эксплуатации являются важными стадиями жизненного цикла продукции. Для эффективного осуществления этих стадий необходимы специалисты, имеющие соответствующую квалификацию. Как было показано в [1], [2], использование ИУП ВР в процессе обучения может значительно повысить его эффективность. Однако одним из основных условий эффективного использования ИУП ВР является своевременное обновление базы данных виртуальной реальности (БД ВР). Это условие можно выполнить, включив ИУП ВР в качестве составляющего элемента в интегрированную информационную среду (ИИС), которая является своеобразным фундаментом для внедрения CALS-технологий. Информация, которая появляется на той или иной стадии жизненного цикла продукции, помещается в ИИС и может быть использована на всех последующих стадиях жизненного цикла.

Рассмотрим стадию конструкторской подготовки производства, на которой создаются объемные геометрические модели разрабатываемой продукции.

При создании моделей, описывающие их данные могут быть представлены в различной форме. Например, при создании моделей в различных системах автоматизированного проектирования (САПР), таких как SolidWorks, ProEngineer, T-Flex, ADEM, AutoCAD, CATIA и т.д., ручным способом, используются форматы *.asm, *.prt, *.dwg, *.cgr, *.emn, *.brd, *idb, *.rpt, *.3dm; при полуавтоматическом способе с использованием 3D-сканирования наряду с указанными также могут использоваться форматы *.xml, *.nxm, *.3ds, *.obj, *.stl, *.wrl, *.ply, *.ply2, *.xyz, *.asc, *.vda, *.jds.

На стадии эксплуатации продукции, при использовании ИУП ВР объемные геометрические модели, как было показано в [1] и [3], целесообразно сохранять в БД ВР в формате *.wrl.

При таком многообразии форм представления данных неминуемо возникает вопрос о выборе форм представления, которые являются оптимальными для той или иной стадии жизненного цикла и их промежуточных этапов. Эта задача решается индивидуально для каждого конкретного случая, поскольку методы и средства проектирования, а, следовательно, и формы представления данных, сильно зависят от вида выпускаемой продукции, особенностей производства, технического оснащения производства, внедренных на предприятии информационных технологий и т.д. Поэтому далее процесс выбора оптимальных форм представления данных будем рассматривать на примере автомобильной промышленности с учетом использования на стадии эксплуатации ИУП ВР.

Проектировщики автомобилей достаточно широко используют в своей работе метод 3D-сканирования, что значительно упрощает и ускоряет процесс создания сложных геометрических моделей. Одним из нескольких форматов, используемых при 3D-сканировании, является формат *.wrl; этот же формат применяется для представления данных, описывающих объемные геометрические модели, в БД ВР.

Поэтому, в целях унификации используемых при проектировании форматов и программного обеспечения, для представления данных при 3D-сканировании будем использовать формат *.wrl.

При 3D-сканировании формируется файл с облаками точек. Затем этот файл, при помощи соответствующего программного обеспечения (ПО), например SolidWorks, преобразуется в объемную геометрическую модель (формат файла *.prt). Поскольку облака точек, сформированные при 3D-сканировании, не обеспечивают идеальную геометрию поверхности модели, то модель необходимо отредактировать. Для исключения лишних операций преобразования форм представления данных, редактирование объемной геометрической модели целесообразно произвести при помощи того же ПО, которое использовалось для ее формирования из облаков точек. После редактирования, текущий формат файла модели (*.prt) можно, в зависимости от поставленной цели, преобразовывать в другие форматы, например в чертежный формат *.drw, в формат *.stl с целью выполнения прототипирования и т.д.

Напомним, что в настоящей статье рассматривается проблема своевременного обновления БД ВР с целью обеспечения соответствия изучаемых изделий, в частности автомобилей, современному уровню развития техники. Для этой цели, файл объемной геометрической модели необходимо преобразовать из формата *.prt в формат *.wrl, поместить в ИИС и использовать для обновления (актуализации) БД ВР. Такое преобразование также можно осуществить при помощи SolidWorks. Таким образом, все необходимые преобразования форм представления данных можно произвести в рамках одной САПР (как в SolidWorks, так и в других указанных выше САПР).

Отразим описанные выше преобразования структур данных на графе (см. рисунок 1).

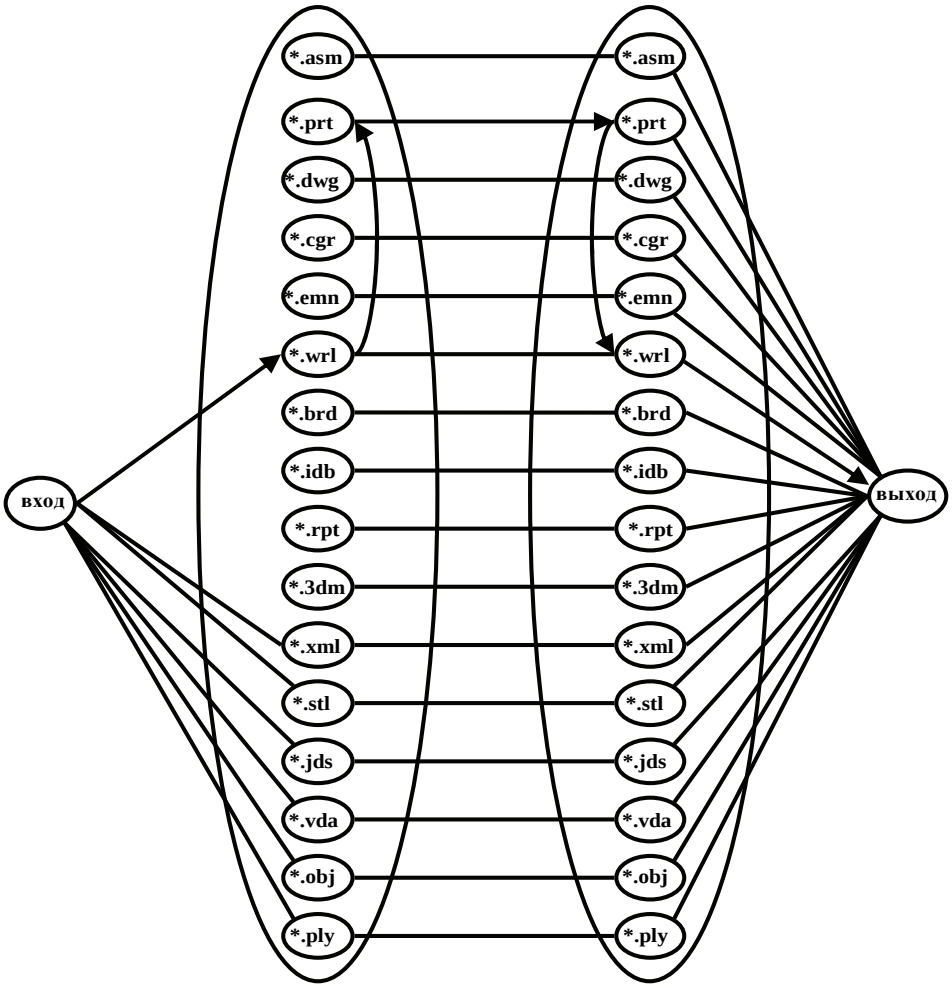


Рисунок 1. Граф преобразований структур данных для процесса формирования базы данных виртуальной реальности

Узлами графа являются различные формы представления данных, использование которых возможно на соответствующем этапе проектирования. Ребрами графа являются различные варианты реализации операций преобразования форматов данных. Эллипсом обозначены преобразователи форматов, обеспечивающие преобразование любого из имеющихся форматов во все остальные и в себя. Для определения пути на графе, который обеспечивает минимальную стоимость преобразования форматов данных, необходимо установить критерии для оценки различных вариантов реализации преобразования.

Критерии устанавливаются индивидуально для каждого конкретного случая. В качестве универсальных критериев могут выступать сложность кодирования, точность представления данных в памяти компьютера, минимальное количество операций преобразования форматов данных, минимальное время на произведение операций преобразования форматов данных. Для решения проблемы обновления БД ВР можно ввести дополнительный критерий в виде унификации программных и аппаратных средств разработчика продукции и пользователя ИУП ВР. Этот критерий был учтен при использовании формата *.wrl при 3D-сканировании, поскольку тот же формат *.wrl используется и для представления данных в БД ВР [1].

В результате анализа на основе заданных критериев всех отраженных на графе вариантов реализации операций преобразования форматов данных, был выбран оптимальный вариант. Этот вариант обозначен стрелками на рисунке 1. На первом шаге (вход – *.wrl) производится 3D-сканирование и формирование wrl-файла, содержащего облака точек, описывающих сканированную поверхность. На втором шаге (*.wrl – *.prt) облака точек преобразуются в объемную геометрическую модель при помощи соответствующих команд SolidWorks. На третьем шаге (*.prt – *.prt) производится редактирование модели без изменения формата представления данных. На четвертом шаге (*.prt – *.wrl) отредактированный файл модели преобразуется в формат *.wrl, посредством которого данные, описывающие объемные геометрические модели, представляются в БД ВР. На заключительном пятом шаге (*.wrl – выход) файл модели в формате *.wrl помещается в ИИС, где может быть доступен пользователю ИУП ВР при необходимости обновления БД ВР.

Произведенные выкладки легли в основу проводимой авторами настоящей статьи работы по внедрению ИУП ВР в ряде московских учебных заведений, которые осуществляют подготовку специалистов по ремонту и обслуживанию автомобилей.

Автомобилестроение является одной из отраслей, которые имеют дело с производством сложных механизмов при довольно частом обновлении модельного ряда выпускаемой продукции. Новые модели автомобилей должны обязательно иметь поддержку на стадии эксплуатации, поскольку продукция, для которой отсутствует возможность проведения качественного технического обслуживания и ремонта, вряд ли будет иметь успех на рынке. Соответственно, учебные заведения, осуществляющие подготовку специалистов по ремонту и техническому обслуживанию автомобилей и использующие ИУП ВР в процессе обучения, часто будут сталкиваться с необходимостью обновления (актуализации) БД ВР. В этом случае, своевременное обновление БД ВР новыми виртуальными моделями возможно лишь при взаимодействии учебного заведения с разработчиком (изготовителем) новой модели автомобиля, который, с одной стороны, уже имеет в наличии необходимые виртуальные модели, а с другой – заинтересован в поддержке своей продукции на стадии эксплуатации и, соответственно, в повышении ее конкурентоспособности.

[1] А.В. Балабанов, В.А. Ромакин, *Разработка интерактивных учебных пособий на основе виртуальной реальности*, Материалы III всероссийской молодежной конф. по проблемам управления (ВМКПУ'2008), ИПУ РАН, Москва, 2008, С. 222-223.

[2] А.В. Балабанов, В.А. Ромакин, *Использование средств виртуальной реальности при разработке интерактивных учебных пособий*, Материалы IV Всероссийской школы–семинара молодых ученых «Проблемы управления и информационные технологии ПУИТ'08», Изд-во Казан. гос. техн. ун-та, Казань, 2008, С. 12-15.

[3] Е.И. Артамонов, А.В. Балабанов, В.А. Ромакин, М.Ю. Щегольков, *Принципы создания интерактивных технических руководств*, Материалы 35-й междунар. конф. «Информационные технологии в науке, образовании, телекоммуникации и бизнесе (IT+S&E*08)», Открытое образование, 2008, 2 с.

Задачи оптимизации и информационного обеспечения при управлении ДПТ

Синюков А. В.
(Институт проблем управления РАН, Москва)
sinyukov@ipu.rssi.ru

Аннотация

В работе предложены методы синтеза оптимального управления для двигателя постоянного тока (ДПТ), с использованием процедуры декомпозиции основанной на скользящих режимах. Для информационного обеспечения алгоритмов управления применяются наблюдатели состояния на скользящих режимах, позволяющие оценить скорость вращения вала двигателя и неизвестный момент сопротивления.

1. Введение

В работе предложены методы синтеза оптимального управления и информационного обеспечения для двигателя постоянного тока (ДПТ), позволяющие решить ряд задач за счет применения новых подходов к решению проблемы. Стандартные подходы к решению задач оптимизации нелинейных объектов высокой размерности приводят к вычислительным сложностям и не позволяют выписать решение в аналитическом виде. К тому же известные трудности использования методов оптимизации связаны с не грубостью алгоритмов решения оптимальных задач к неточности задания начальных условий, параметрическим неопределенностям модели объекта управления и влиянию внешних неконтролируемых возмущений. Основной подход состоит в использовании процедуры декомпозиции на скользящих режимах [1], использование которой позволяет разбить задачу для нелинейного объекта высокой размерности на две независимо решаемые подзадачи: линейную при движении в скользящем режиме меньшей размерности и задачу по быстродействию также для подсистемы меньшей размерности. Для синтеза оптимального управления используется два критерия, позволяющие решить задачу позиционирования за заданное время с минимальной затрачиваемой энергией и выхода на заданную частоту вращения вала двигателя за минимальное время. Для информационного обеспечения алгоритмов управления применяются наблюдатели состояния на скользящих режимах, позволяющие оценить скорость вращения вала двигателя и неизвестный момент сопротивления.

2. Постановка задачи

Объект управления (ДПТ) описывается системой дифференциальных уравнений четвертого порядка:

$$\begin{cases} \dot{x}_1 = x_2, \\ \dot{x}_2 = a_{21}(x_3 x_4 - m_L(t)), \\ \dot{x}_3 = a_{32}(u_1 - x_2 x_4 - a_{31} x_3), \\ \dot{x}_4 = a_{41}(u_2 - f(x_4)), \\ \dot{m}_L = 0, \end{cases} \qquad (1)$$

где x_1 - угол поворота вала двигателя; x_2 - угловая скорость; x_3 - ток якоря; x_4 - магнитный поток; u_1, u_2 - напряжение якоря и обмотки возбуждения; m_L - момент нагрузки, $f(x_4)$ - кривая намагничивания. Измерению доступны угол поворота вала двигателя x_1 и токи x_3, x_4 в обмотках якоря и возбуждения.

Требуется выбрать управляющие воздействия u_1, u_2, чтобы минимизировать критерии

$$I_1 = \int_{t_0}^{t_f} (q_1 x_1^2 + q_2 x_2^2 + r((x_3 x_4) - m_L)^2) dt, q \geq 0,$$

$$r > 0;$$

$$I_2 = \int_0^T dt \to \min.$$

Ограничения наложены на управления u_1, u_2

$$|u_1| \le U_1; \quad |u_2| \le U_2.$$

3. Синтез алгоритма управления

Рассмотрим механическую подсистему

$$\begin{cases} \dot{x}_1 = x_2, \\ \dot{x}_2 = a_{21}\nu, \end{cases}$$

где $\nu = x_3 x_4 - m_L$ рассматривается в качестве нового управления.

Для синтеза фиктивного управления, минимизирующего критерий I_1 воспользуемся методом динамического программирования Беллмана[2].

Уравнение Беллмана

$$q_1 x_1^2 + q_2 x_2^2 + r\nu^2 + \sum_{i=1}^{2} \frac{\partial V}{\partial x_i}\dot{x}_i = -\frac{\partial V}{\partial t},$$

где V – функция Беллмана.

Для линейной стационарной системы функция Беллмана выбирается в виде квадратичной формы $V = x^T P x = P_{11}x_1^2 + 2P_{12}x_1x_2 + P_{22}x_2^2$, где P – симметричная положительно определённая матрица.

Подставляя выражения для производных вектора состояния в последнее уравнение, получим

$$q_1 x_1^2 + q_2 x_2^2 + r\nu^2 + 2P_{11}x_1x_2 + 2P_{12}x_2^2 + (2P_{12}x_1 + 2P_{22}x_2)a_{21}(\nu - m_L) = 0,$$

Минимум данного выражения по ν

$$2r\nu + (2P_{12}x_1 + 2P_{22}x_2)a_{21} = 0,$$

достигается при $\nu = -r^{-1}a_{21}(P_{12}x_1 + P_{22}x_2)$.

Подставляя данное выражение в уравнение Беллмана, получим уравнение Риккати относительно неизвестной матрицы P

$$q_1x_1^2 + q_2x_2^2 - r^{-1}a_{21}^{-2}(P_{12}x_1 + P_{22}x_2)^2 + 2P_{11}x_1x_2 + 2P_{12}x_2^2 + 2r^{-1}a_{21}^2(P_{12}x_1 + P_{22}x_2)^2 = 0,$$

решая которое получим выражение для оптимального управления

$$\nu_{опт} = -k_1x_1 - k_2x_2,$$

$$k_1 = \sqrt{\frac{q_1}{r}}, k_2 = a_{21}^{-1}\left(q_2 + \sqrt{\frac{q_1 r}{a_{21}^2}}\right). \qquad (2)$$

Таким образом, получим

$$\nu = x_3x_4 - m_L = -k_1x_1 - k_2x_2.$$

Согласно процедуре декомпозиции на скользящих режимах [1] выберем поверхность скольжения S_1 в виде

$$S_1 = x_3x_4 - m_L + k_1x_1 + k_2x_2.$$

Информация о переменных x_2, m_L может быть получена с помощью наблюдателей состояния на скользящих режимах.

Попадание на поверхность скольжения обеспечивается в нижней электрической подсистеме

$$\begin{cases} \dot{x}_3 = a_{32}(u_1 - x_2x_4 - a_{31}x_3), \\ \dot{x}_4 = a_{41}(u_2 - f(x_4)), \end{cases}$$

с помощью управляющих воздействий u_1, u_2 равных

$$\begin{cases} u_1 = -M_1 sign S_1 \\ u_2 = -M_2 sign S_2, \end{cases}$$

где M_1 и M_2 – амплитуды переключения, $S_2 = x_4 - x_{4d}$, x_{4d} – номинальный поток возбуждения.

Отметим, что управляющее воздействие (2) является оптимальным по критерию I_1 только после попадания траекторий системы на поверхность S_1. Чем меньше это время, тем быстрее достигается оптимальный режим с минимумом затрачиваемой энергии. Поэтому амплитуды u_1, u_2 выбираются так, чтобы время попадания на поверхность S_1 было наименьшим, т.е.

$$I_2 = \int_0^T dt \to \min.$$

Известно [2], что данная задача решается выбором u_1, u_2

в виде

$$\begin{cases} u_1 = -U_1 sign S_1 \\ u_2 = -U_2 sign S_2. \end{cases} \qquad (3)$$

4. Задача наблюдения на скользящих режимах

Проблема информационного обеспечения алгоритмов управления решается с помощью каскадного наблюдателя на скользящих режимах[3,4]

$$\begin{cases} \dot{\hat{x}}_1 = \hat{x}_2 + \nu_1, \\ \dot{\hat{x}}_2 = a_{21}(v - \hat{m}_L) + \nu_2, \\ \dot{\hat{m}}_L = \nu_3, \end{cases} \qquad (4)$$

где $\hat{x}_1, \hat{x}_2, \hat{m}_L$ – оценки переменных x_1, x_2, m_L; ν_1, ν_2, ν_3 – корректирующие воздействия.

Запишем выражения относительно невязок $\overline{x}_1 = x_1 - \hat{x}_1$, $\overline{x}_2 = x_2 - \hat{x}_2$, $\overline{m}_L = m_L - \hat{m}_L$

$$\begin{cases} \dot{\overline{x}}_1 = \overline{x}_2 - \nu_1, \\ \dot{\overline{x}}_2 = -a_{21}\overline{m}_L - \nu_2, \\ \dot{\overline{m}}_L = -\nu_3. \end{cases}$$

Выберем $\nu_1 = M sign \overline{x}_1$. Тогда при $M > \overline{x}_2$ на поверхности $\overline{x}_1 = 0$ возникает скользящий режим. При этом эквивалентное управление $\nu_{1eq} = \overline{x}_2$. ν_{1eq} может быть получено как выход низкочастотного фильтра с малой постоянной времени $\mu > 0$

$$\mu \dot{z}_1 = -z_1 + \nu_1,$$
$$z_1(t) = \nu_{1eq}(t) + O_1(\mu, t), \qquad \lim_{\mu \to 0} O_1(\mu, t) = 0.$$

Величину ν_{1eq} используем для формирования корректирующего воздействия

$$\nu_2 = M sign \nu_{1eq} = M sign z_1.$$

При $M > -a_{21}\overline{m}_L$ на поверхности $\overline{x}_2 = 0$ возникает скользящий режим. В этом случае $\nu_{2eq} = -a_{21}\overline{m}_L$.

На последнем шаге выбираем управление ν_3 в виде

$$\nu_3 = -M sign \nu_{2eq} = -M sign z_2,$$
$$\mu \dot{z}_2 = -z_2 + \nu_2,$$

которое обеспечивает сведение невязки $\overline{m}_L$ к нулю.

Окончательный алгоритм управления (3) примет вид

$$\begin{cases} u_1 = -U_1 sign S_1 = x_3 x_4 - \hat{m}_L + k_1 x_1 + k_2 \hat{x}_2, \\ u_2 = -U_2 sign S_2 = -U_2 sign(x_4 - x_{4d}), \end{cases}$$

где $\hat{x}_2$, $\hat{m}_L$ получаются с помощью наблюдателей (4).

5. Литература

[1] Уткин В. И. “Скользящие режимы в задачах в задачах оптимизации и управления.”, *Book Title*, М.: Наука, главная редакция физико-математической литературы, 1981, 116 - 132 с.

[2] “Современная прикладная теория управления: Новые классы регуляторов технических систем/под ред. А.А. Колесникова.”, *Book Title*, Таганрог: Изд-во ТРТУ, 2000.

[3] Уткин В.А., Нгуен Куанг Хынг. “Задачи управления двигателем постоянного тока”, *Journal*, АиТ. №5.2006. С. 102-118.

[4] Краснова С. А., Уткин В. А. “Каскадный синтез наблюдателей состояний динамических систем.” ,*Book Title*, М.: Наука, 2006, 32 - 51 с.

Построение стохастической модели прогнозирования состояния вычислительной сети как решение задачи классификации и снижения размерности

Еремеев В.Б.
МИКТ, г. Липецк
Failure_vt@mail.ru

Аннотация

Представлена модель вероятностного прогнозирования состояния наблюдаемых сетевых характеристик, в качестве которых выступают переменные MIB. Данная модель может быть использована в создании программного обеспечения активного мониторинга вычислительной сети, ориентированного на обнаружение аномалий. Построение модели представлено в контексте решения задач классификации и снижения размерности.

1. Введение

Для анализа состояния компьютерной сети: определения значений характеристик качества транспортных услуг сети, выявления узких мест, прогнозирования последствий изменений в топологии, выявления аномалий и проч., - необходим мониторинг – сбор статистических данных с привязкой к моменту времени.

В последнее время создан целый ряд продуктов для сбора сетевой статистики, анализа и управления. Но универсальность использования данных средств ограничена необходимостью заблаговременно определять различные параметры, такие, как, например, пороговые значения величин для конкретных соединений или статическое описание так называемых нормальных профилей работы сети, а также постоянного наблюдения за системой. Очевидно, что чем точнее произведена настройка параметров, тем эффективнее будет работа данных программных средств, что опять-таки напрямую зависит от опыта работы администратора с данной сетью и знания особенностей ее поведения.

Проблема разработки эффективных методов и средств выявления аномальных состояний сети является весьма актуальной. Различные исследования [5-8] сосредоточены на том, как наиболее точно определить и оценить возникающие отклонения от нормального режима работы сети.

2. Предшествующие работы

Предлагаемая схема активного мониторинга сети была представлена в работе [4]. Основные её положения заключаются в следующем: вводятся адреса наблюдаемых сетевых устройств и OID-идентификаторы интересующих переменных. В качестве наблюдаемых переменных используются переменные MIB, значения которых снимаются с сетевых устройств посредством протокола SNMP. Для переменных задаётся общее время мониторинга $T_{мон}$ и шаг накопления Δt – временной интервал, через который происходит считывание значений MIB-переменных $Var_i(t)$ и преобразование к виду $Var_i*(\Delta t_j)$ на протяжении $T_{мон}$.

Ключевым моментом является возможность проведения произвольной периодизация данных с любым интересующим временным интервалом $T_{набл}$ после накопления некоторого объёма данных, что даёт возможность представить наблюдаемую переменную как совокупность реализаций случайной величины [3]. Таким образом, получаем ансамбль из K реализаций случайной нестационарной величины, каждая из которых состоит из N шагов на $T_{набл}$.

Второй важный момент - обучение системы на основе спериодизированных данных, устраняющее недостатки моделей прогнозирования, используемых в других работах [5-8]. Данные недостатки обусловлены:

- старением данных;
- нестационарностью процесса;
- наличием памяти (автокорелляции) процесса.

Обучение подразумевает:

- переход от пространства значений наблюдаемых переменных к пространству состояний;
- определение оптимальных глубин обращения в прошлое для каждого шага;

– построение вероятностной модели прогнозирования состояний наблюдаемых переменных.

Автором разработаны алгоритмы автоматического выполнения всех названных пунктов.

В работе [2] было показано, что если процесс является $(d+1)$-связным, т.е., если на состояние системы $s^{(j+1)}_{m(j+1)}$ на шаге Δt_{j+1} основное влияние оказывают результаты $(d+1)$ предыдущих испытаний (рис.1), то вероятность нахождения системы в состоянии $s^{(j+1)}_{m(j+1)}$ на шаге Δt_{j+1} может быть определена по формуле:

$$p(s^{(1)}_{m1}, s^{(2)}_{m2},.., s^{(j+1)}_{m(j+1)}) = p(s^{(j-d)}_{m(j-d)}) \times \times \prod_{i=n-d+1}^{j+1} p(s^{(i)}_{mi} \mid p(s^{(j-d)}_{m(j-d)},..., s^{(i-1)}_{m(i-1)}) \quad (1)$$

В общем случае, глубина обращения в прошлое будет своя для каждого наблюдаемого шага.

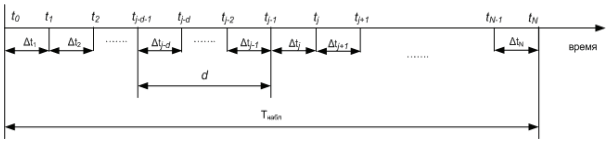


Рис.1. Значимая глубина d для шага $j+1$

Выражение (1) позволяет осуществить статистическое прогнозирование состояния сети на основе представления прогнозируемой сетевой характеристики как совокупности реализаций случайного процесса.

Построенная модель стохастического прогнозирования состояний наблюдаемых сетевых переменных реализует выражение (1) и представляет собой решение задачи классификации и снижения размерности.

3. Предпосылки снижения размерности

Имеется, по крайней мере, три основных типа принципиальных предпосылок, обусловливающих возможность перехода от большого числа p исходных показателей состояния (поведения, эффективности функционирования) анализируемой системы к существенно меньшему числу p' наиболее информативных переменных [1]. Это, во-первых, дублирование информации, доставляемой сильно взаимосвязанными признаками; во-вторых, неинформативность признаков, мало меняющихся при переходе от одного объекта к другому (малая «вариабельность» признаков); в-третьих, возможность агрегирования, т. е. простого или «взвешенного» суммирования, по некоторым признакам.

Формально задача перехода (с наименьшими потерями в информативности) к новому набору признаков $\tilde{z}^{(1)}$, $\tilde{z}^{(2)}$,..., $\tilde{z}^{(p')}$ может быть описана следующим образом. Пусть $Z = Z(X)$ - некоторая p-мерная вектор-функция исходных переменных $x^{(1)}$, $x^{(2)}$,..., $x^{(p)}$ ($p' < p$) и пусть $I_{p'}(Z(X))$ - определенным образом заданная мера информативности p'-мерной системы признаков $Z(X)=(z^{(1)}(X),\ldots,z^{(p')}(X))$. Конкретный выбор функционала $I_{p'}(Z)$ зависит от специфики решаемой реальной задачи и опирается на один из возможных критериев: критерий автоинформативности, нацеленный на максимальное сохранение информации, содержащейся в исходном массиве $\{X_i\}$, $i=1,..,n$ относительно самих исходных признаков; и критерий внешней информативности, нацеленный на максимальное «выжимание» из $\{X_i\}$ информации, содержащейся в этом массиве относительно некоторых других (внешних) показателей.

Задача заключается в определении такого набора признаков $\tilde{Z}$, найденного в классе F допустимых преобразований исходных показателей $x^{(1)}$, $x^{(2)}$,..., $x^{(p)}$, что [1]:

$$I_{p'}(\tilde{Z}(X)) = \max_{Z \in F} \{I_{p'}(Z(X))\} \quad (2)$$

4. Построение вероятностной модели

После выполнения периодизации и перехода к пространству состояний зафиксированные значения наблюдаемой переменной можно представить в виде массива X_{Si} состояний на шагах для различных реализаций (таблица 1).

Для каждого из шагов $j>d^*$, где d^* – оптимальная глубина обращения к значениям прошлых шагов, предшествующих шагу j, может быть составлена матрица зарегистрированных состояний $A_{i(j+1)}$ размером $(d^*+2)\times K$, где K – число реализаций, полученное в результате периодизации, получаемая “вырезанием” и последующим транспонированием части массива X_{Si}. Элементы матрицы представляют собой значения состояний наблюдаемой случайной величины на шагах $(j-d^*)$, $(j-d^*+1)$,..., $(j+1)$ для всех полученных реализаций 1, ..., K.

Для шагов $j \le d^*$ размерность матрицы будет составлять $(j+1)\times K$.

Матрицу $A_{i(j+1)}$ можно записать в виде массива $A^*_{i(j+1)}$ размерностью $2\times K$, где первый столбец содержит векторы предшествующих состояний – соответствующие строки матрицы $A_{i(j+1)}$, а второй – значения на последующем $(j+1)$ шаге, т.е., элементы последнего столбца матрицы $A_{i(j+1)}$.

Структуры массивов $A_{i(j+1)}$ и $A^*_{i(j+1)}$ приведены в таблицах 2а и 2б соответственно.

Массив $A^*_{i(j+1)}$ может содержать равные векторы предшествующих состояний для разных реализаций:

$$[S^{(j)}]^l = [S^{(j)}]^m; l \neq m, l = \overline{1,K}, m = \overline{1,K} \quad (3)$$

Составим массив $B_{i(j+1)}$ неповторяющихся сочетаний состояний, предшествующих шагу ($j+1$), размерностью $2\times c$, где c – общее число таких сочетаний. Таким образом, построение массива $B_{i(j+1)}$ представляет собой решение задачи снижения размерности массива $A^*_{i(j+1)}$ на основе соотношения (3), т.е., на основе дублирования информации. Первый столбец содержит векторы предшествующих состояний из массива $A^*_{i(j+1)}$ без повторов $[S_B^{(j)}]^i$ (i = 1,..,c), а второй – число повторов конкретного состояния $[S_B^{(j)}]^i _ count$ в массиве $A^*_{i(j+1)}$;

$$\sum_{i=1}^{c} [S_B^{(j)}]^i _ count = K \quad (4)$$

Структура массива $B_{i(j+1)}$ приведена в таблице 3.

Для вероятностного прогнозирования состояния случайной величины на шаге ($j+1$) по данным массивов $A^*_{i(j+1)}$ и $B_{i(j+1)}$ составим массив $C_{i(j+1)}$, содержащий c строк, соответствующих предшествующим сочетаниям массива $B_{i(j+1)}$ и переменное число столбцов в соответствии со структурой таблицы 4. Первый столбец массива $C_{i(j+1)}$ содержит число состояний на шаге ($j+1$) без повторов для соответствующего сочетания массива $B_{i(j+1)}$.

Следует заметить, что

$$\sum_{i=1}^{sk_n} [s_{combn}^{(j+1)}]^i _ count = [S_B^{(j)}]^n _ count;$$

$$n = \overline{1,c}; \quad (5)$$

Вероятность появления на шаге ($j+1$) состояния $[s_{combn}^{(j+1)}]^i$ при условии того, что на предыдущих d шагах система находилась в состояниях, дающих сочетание $[S_B^{(j)}]^n$ может быть определена как:

$$p([s_{combn}^{(j+1)}]^i) = \frac{[s_{combn}^{(j+1)}]^i _ count}{[S_B^{(j)}]^n _ count};$$

$$i = \overline{1, sk_n}; n = \overline{1,c}; \quad (6)$$

Следует заметить, что каждая из строк массива $C_{i(j+1)}$ отражает полную группу событий, т.е., справедливо соотношение:

$$\sum_{i=1}^{sk_n} p([s_{\text{combn}}^{(j+1)}]^i) = 1;\ n = \overline{1,c}; \quad (7)$$

Массивы A, A^*, B, C представляют собой вероятностную модель прогнозирования состояния случайной величины на шаге ($j+1$), реализующую нахождение значения выражения (1).

Таблица 1. Массив X_{Si} состояний на шагах для различных реализаций

Шаг	Реализация 1	Реализация 2	……	Реализация k	……	Реализация K
1	$[s^{(1)}]^1$	$[s^{(1)}]^2$	……	$[s^{(1)}]^k$	……	$[s^{(1)}]^K$
2	$[s^{(2)}]^1$	$[s^{(2)}]^2$	……	$[s^{(2)}]^k$	……	$[s^{(2)}]^K$
……	……	……	……	……	……	……
N	$[s^{(N)}]^1$	$[s^{(N)}]^2$	……	$[s^{(N)}]^k$	……	$[s^{(N)}]^K$

Таблица 2а. Матрица $A_{i(j+1)}$ зарегистрированных состояний

Реализация	Состояния на шагах				
	Δt_{j-d}	Δt_{j-d+1}	……	Δt_j	Δt_{j+1}
1	$[s^{(j-d)}]^1$	$[s^{(j-d+1)}]^1$	……	$[s^{(j)}]^1$	$[s^{(j+1)}]^1$
2	$[s^{(j-d)}]^2$	$[s^{(j-d+1)}]^2$	……	$[s^{(j)}]^2$	$[s^{(j+1)}]^2$
……	……	……	……	……	……
K	$[s^{(j-d)}]^K$	$[s^{(j-d+1)}]^K$	……	$[s^{(j)}]^K$	$[s^{(j+1)}]^K$

Таблица 2б. Массив $A^*_{i(j+1)}$ зарегистрированных состояний

Реализация	Состояния на шагах	
	Предшествующие состояния	Δt_{j+1}
1	$[S^{(j)}]^1$	$[s^{(j+1)}]^1$
2	$[S^{(j)}]^2$	$[s^{(j+1)}]^2$
……	……	……
K	$[S^{(j)}]^K$	$[s^{(j+1)}]^K$

Таблица 3. Массив $B_{i(j+1)}$ неповторяющихся сочетаний

Сочетание	Предшествующие неповтор. состояния	Число сочетаний
1	$[S_B^{(j)}]^1$	$[S_B^{(j)}]^1_count$
2	$[S_B^{(j)}]^2$	$[S_B^{(j)}]^2_count$
……	……	……
c	$[S_B^{(j)}]^c$	$[S_B^{(j)}]^c_count$

Таблица 4. Массив $C_{i(j+1)}$ неповторяющихся состояний на шагах (*j+1*) для неповторяющихся предшествующих сочетаний

Соч.	Число неп. сост. для сочет.	Неп. сост. для неп. сочет.	Число неп. сост.	Вер-ть неп. сост. для неп. сочет.	Неп сост. для неп. соч.	...	Вер-ть неп. сост. для неп. сочет.
1	sk_1	$[s_{comb1}^{(j+1)}]^1$	$[s_{comb1}^{(j+1)}]^1_count$	$p([s_{comb1}^{(j+1)}]^1)$	$[s_{comb1}^{(j+1)}]^2$	...	$p([s_{comb1}^{(j+1)}]^{sk_1})$
2	sk_2	$[s_{comb2}^{(j+1)}]^1$	$[s_{comb2}^{(j+1)}]^1_count$	$p([s_{comb2}^{(j+1)}]^1)$	$[s_{comb2}^{(j+1)}]^2$	...	……
…	……	……	……	……	……	...	……
c	sk_c	$[s_{combc}^{(j+1)}]^1$	$[s_{combc}^{(j+1)}]^1_count$	$p([s_{combc}^{(j+1)}]^1)$	$[s_{combc}^{(j+1)}]^2$	...	……

5. Заключение

Представленная модель будет использована в разрабатываемом программном обеспечении активного мониторинга вычислительной сети, предназначенного для обнаружения аномалий в её работе. Данное программное обеспечение будет свободно от недостатков аналогичных средств, названных в пункте 2.

Разрабатываемое ПО представляет собой инструмент сетевого инженера, дополняющий функциональность систем, используемых в настоящее время в вычислительных сетях для решения вопросов диагностики и мониторинга.

Обобщённые правила использования разрабатываемого программного обеспечения могут быть сформулированы следующим образом:

1) Анализ значений и состояний наблюдаемых переменных сетевых устройств и данных о степени их аномальности: вероятность появления зафиксированного состояния с учётом наблюдаемого шага и предшествующих состояний на значимой глубине памяти, код аномальности наблюдаемого состояния и его расшифровка.

2) Фиксация наличия аномалии в работе конкретного наблюдаемого сетевого устройства.

3) Определение причины аномального состояния, которое может быть осуществлено посредством анализа журнала наблюдаемого сетевого устройства, событий сетевого приёмника trap-сообщений, данных систем обнаружения вторжений и пр.; в случае необходимости, при наличии аномалии на каком-либо интерфейсе, следует произвести захват и анализ трафика с данного интерфейса.

4) В случае если аномалия является недопустимой, - принять меры для устранения причин и, возможно, последствий её возникновения.

Область применения разрабатываемого инструмента активного мониторинга вычислительной сети представляет собой наблюдение сетевых устройств, характеризующихся определённой (сменной или суточной) периодичностью своей работы. К таким устройствам относится оборудование магистрали сети предприятия, сетевые фермы, сами сервера.

6. Литература

[1] Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Классификация и снижение размерности. – М.: Финансы и статистика, 1989. – 607 с.

[2] Еремеев В.Б. Мониторинг и статистическое прогнозирование состояния компьютерной сети // Информационные технологии моделирования и управления, №5(39), Воронеж: Научная книга. - 2007. - С.566-574.

[3] Еремеев В.Б. Подход к мониторингу сети как рассмотрению совокупностей реализаций случайных величин // Компьютерные технологии в технике и экономике. – Воронеж: Международный институт компьютерных технологий. – 2007. – С. 23-26.

[4] Еремеев В.Б. Создание методики активного мониторинга вычислительной сети // Всероссийская молодёжная конференция по проблемам управления (ВМКПУ'2008): Труды. – Москва: ИПУ РАН. – 2008. – С.240-242.

[5] Brutlag J. Aberrant behavior detection in time series for network monitoring / Proceedings of the USENIX Fourteenth System Administration Conference LISA XIV, New Orleans, LA, December 2000.

[6] Hood C., Ji C. Proactive network fault detection / Proceedings of IEEE INFOCOM '97, Kobe, Japan, April 1997, INFOCOM, pp.342-351.

[7] Thottan M., Ji C. Adaptive Thresholding for Proactive Network Problem Detection / Third IEEE International Workshop on Systems Management, pp.108-116, Newport, Rhode Island, April, 1998.

[8] Ye N. A markov chain model of temporal behavior for anomaly detection / Workshop on Information Assurance and Security, West Point, NY, June 2000.

Проектирование структур данных для комплекса сбора оперативной информации о случаях заболеваний

Сурпин В.П., Волчков Д.В.
(*Институт проблем передачи информации РАН, Москва*)
vadim@iitp.ru, vol@iitp.ru

Аннотация

Работа связана с проектом, который выполняется в институте по заказу министерства здравоохранения РФ. Целью проекта является создание автоматизированного комплекса сбора и анализа информации по инфекционным заболеваниям. В данной работе рассматриваются вопросы проектирования структур данных для комплекса сбора оперативной информации о случаях заболеваний на основании исследований предметной области, произведенной в [1].

Для эффективной обработки больших объемов информации в распределенных информационных системах требуется спроектировать структуру данных наиболее оптимальным образом. Критериями эффективности работы системы являются быстрота выполнения поисковых запросов и простота представления информации в удобном для пользователя виде. Разработка структуры базы данных тесно связана с исследованием объектной структуры предметной области.

В данной работе рассматривается метод использования объектно-ориентированного подхода для исследования предметной области с целью формирования структуры базы данных на примере проекта «Проектирование комплекса сбора оперативной информации о случаях заболеваний».

В работе [1] был произведен анализ требований, предъявляемых к системе сбора оперативной информации о случаях заболеваний. В процессе исследований были выявлены сценарии использования системы, приведенные на Рис. 1.

В овалах указаны наименования процессов, пользователи системы обозначены фигурками

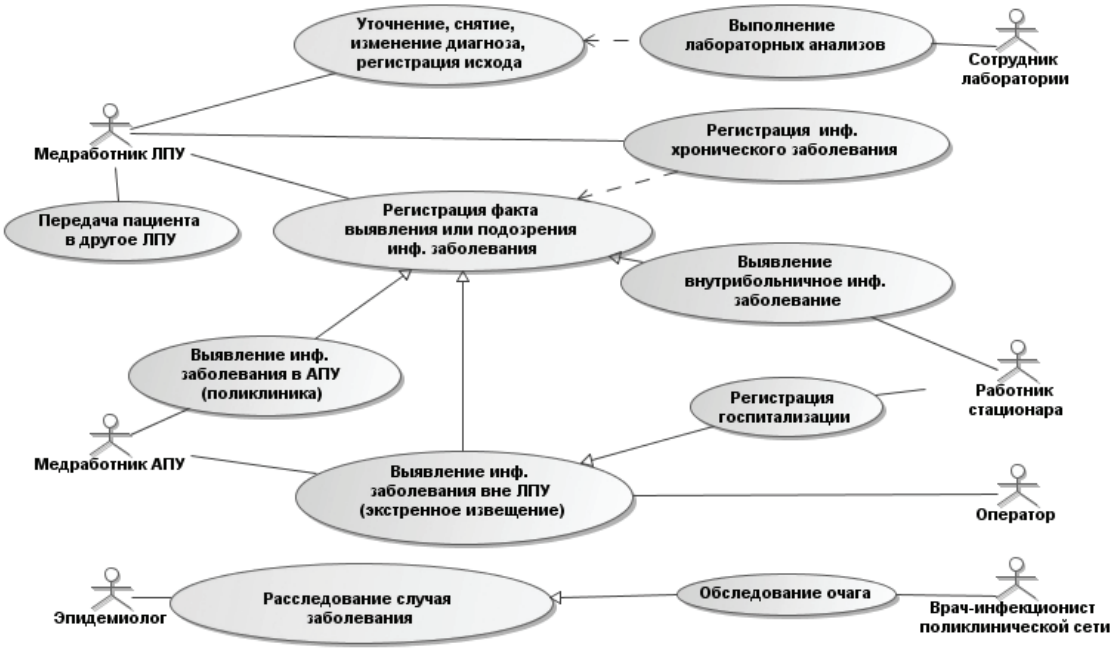


Рис. 1. Процессы режима регистрации

людей, линии на схеме показывают связи между ними.

Среди выявленных процессов первоочередным для автоматизации является «Выявление инфекционного заболевания» т.к. он связан с непосредственным сбором информации из медицинских учреждений. Без его реализации невозможна дальнейшая работа модулей статистики и аналитики т к. введенные на данном этапе данные являются исходными для работы указанных блоков системы. Схема алгоритма данного процесса представлена на Рис. 2.

Алгоритм имеет две точки входа – события, инициирующие действия, необходимые для выполнения указанного выше процесса:

- выявление подозрения на острое заболевание;
- выявление подозрения на внутрибольничную инфекцию.

Далее в систему вводятся данные для оформления экстренного извещения, в процессе которого выполняется взаимодействие с пользователем, верификация введенных данных, после чего данные заносятся в базу и отправляются уведомления заинтересованным лицам.

Согласно принципам объектно-ориентированного проектирования, на основании модели Use-Case (Рис. 1) и схемы алгоритма (Рис. 2) строится модель классов предметной области, которая обычно отражается на диаграмме классов, но поскольку современные системы объектно-ориентированного отражения позволяют установить практически однозначное соответствие между классами модели предметной области и таблицами базы, то приведем лишь структуру БД (Рис. 3).

Важной особенностью задачи регистрации инфекционного заболевания является то, что процесс обслуживания заявки (экстренного извещения) может проходить несколько этапов и в различных учреждениях здравоохранения: поликлинике, лаборатории, в кабинете эпидемиолога, разделённых как территориально, так и по времени работы. Таким образом, процесс обслуживания экстренного извещения представляет собой сложную схему со многими параллельными ветвями обработки заявки, каждая из которых представляет собой схему наподобие Рис. 1.

Пользователи, взаимодействующие с системой, могут выполнять вход с интервалом в несколько дней, поэтому важно сохранять не только информацию, указанную ими, но также и состояние процесса на тот момент, когда работа с ним была приостановлена.

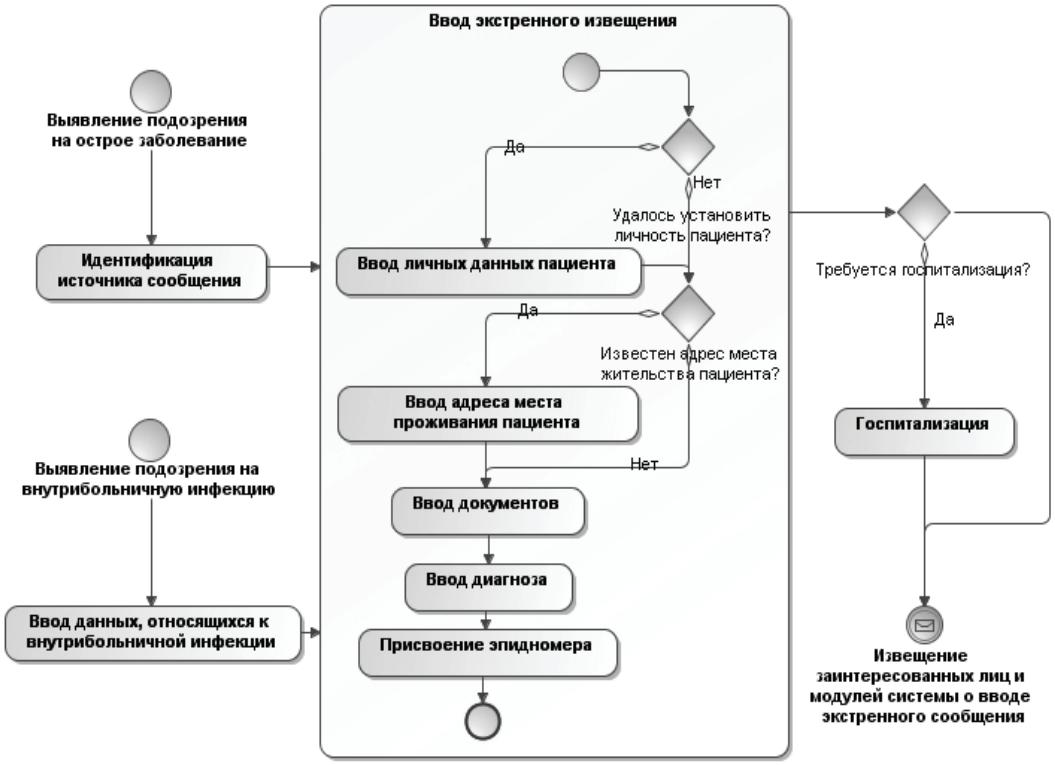


Рис. 2. Процесс регистрации экстренного извещения

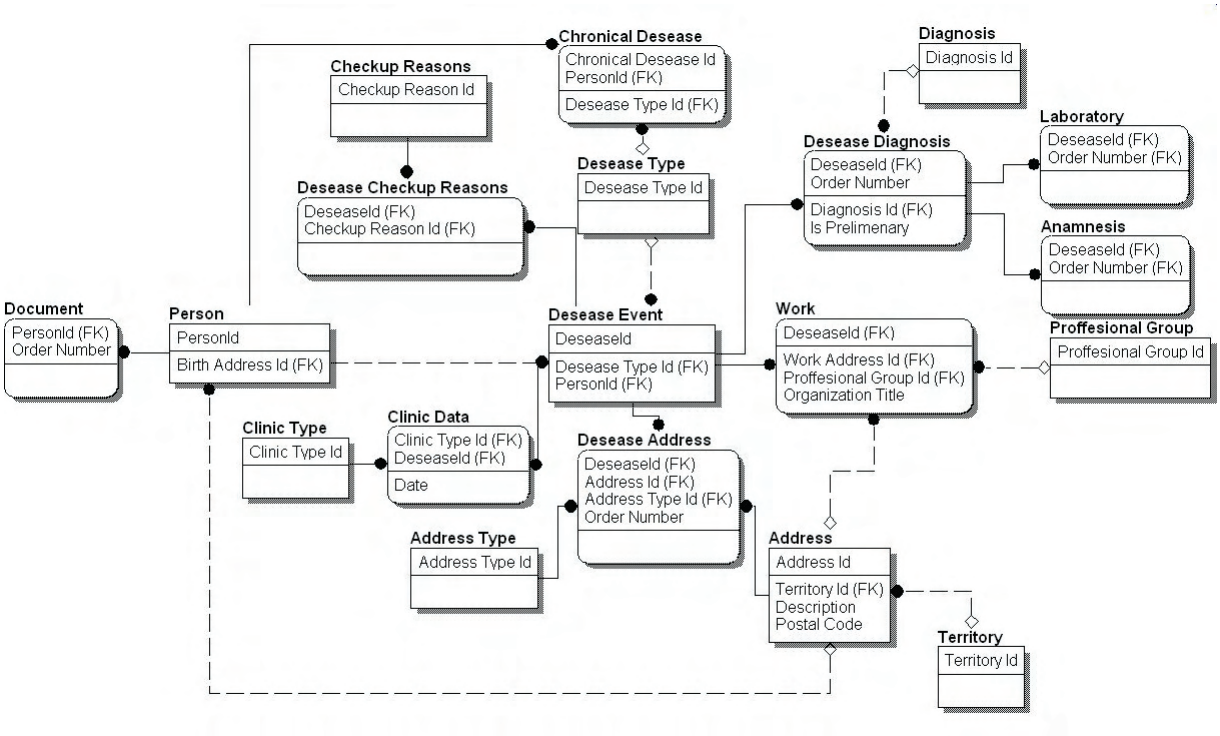


Рис. 3. Структура базы данных

Следовательно, в базе данных должна храниться не только информация, указанная пользователем, но также и некоторая служебная информация, которая позволит восстановить процесс обработки (workflow) с той точки, где он был остановлен. Приведённая на Рис. 3 схема базы данных описывает лишь модель предметной области. Структуры данных, необходимые для сохранения состояний рабочего процесса на данной схеме не отражены. Управление процессами обработки заявки возлагается на специализированные workflow-реализации, создающие собственную структуру объектов БД.

Правильное спроектированная модель предметной области и структура данных значительно сокращают время разработки всего приложения в целом и являются важнейшим предметом исследования.

На них будет опираться вся функциональность системы, такая как намеченные для реализации в дальнейшем блоки статистического анализа, географического выявления очагов, и т.д.

Литература

1. СУРПИН В.П., ВОЛЧКОВ Д.В. Проектирование комплекса сбора оперативной информации о случаях заболеваний. III Всероссийская молодежная конференция по проблемам управления (ВМКПУ'2008):Труды
2. БАКАНОВА Н.Б., ЦАПАЕВА Ю.А., ГУРВИЦ А.Л., СУРПИН В.П., Создание корпоративной системы сбора и анализа информации на основе Web-технологий. Международный семинар «Распределенные компьютерные телекоммуникационные сети». М.: Техносфера, 2005.
3. ЛАРМАН К. *Применение UML 2.0 и шаблонов проектирования. 3-е издание*. М.: Вильямс, 2007. – 736 с.
4. Нейбург Эрик Дж., Максимчук Роберт А. Проектирование баз данных с помощью UML. - М.: Вильямс, 2002. – 288 с.
5. Чекалов А.П. Базы данных: от проектирования до разработки приложений. – С-Пб.: БХВ-Петербург, 2003. – 384 с.

Функциональная модель агента-предприятия в мультиагентной системе управления экологической безопасностью

Ю.И. Еременко, Е.Г. Доронина
katerina_komova@mail.ru

Аннотация

Для решения задач экологической безопасности предлагается использовать систему, основанную на технологии мультиагентных систем. В частности, в статье разработана функциональная модель агента-предприятия мультиагентной системы управления экологической безопасностью.

Вне зависимости от типа и характера производства любое предприятие связано с окружающей его природной средой. На всех этапах хозяйственной деятельности происходит обмен веществом, энергией и информацией с окружающей средой. Предприятие выступает первым и наиважнейшим элементом в хозяйственной деятельности человека, влияющим на загрязнение и деградацию окружающей природной среды, то есть деятельность каждого предприятия региона приводит к снижению уровня экологической безопасности региона. Для того, чтобы свести к минимуму отрицательное воздействие на окружающую среду, необходимо выработать механизмы управления – это совокупность управленческих, технологических, финансово-экономических мероприятий, направленных на снижение экологической нагрузки на окружающую природную среду. При этом важно учесть, что эти мероприятия должны проводиться в рамках действующего производства. Ответственность же за безопасность региона возложена на органы власти, которые располагают полномочиями применять различные экономические и административные механизмы, направленные на увеличение или поддержания необходимого уровня безопасности.

Для управления экологической безопасностью региона предлагается использовать технологию мультиагентных систем (МАС), которая позволяет построить и промоделировать модели поведения предприятий и органов власти региона. Данное моделирование включает в себя основные параметры, характеризующие действующие лица, и, прежде всего мотивы экономической активности, ее цели и средства достижения этих целей. Определяющей при создании МАС является парадигма интеллектуальных агентов. Агент обладает набором таких свойств, как автономность, реактивность, активность и коммуникативность, которые наилучшим образом соответствуют необходимым условиям функционирования МАС и позволят реализовать мультидисциплинарный подход и модульные решения для сбора, обработки данных и принятия решений для управления экологической безопасностью. В МАС экологической безопасности (МАСЭБ) были выделены следующие типы агентов: агент - ресурса, агент - предприятия, агент-супервизор, агент - интерфейса.

Агент - предприятия отстаивает интересы своего предприятия, преследующего свои частные цели, часто расходящиеся с объективно существующими целями устойчивого и безопасного развития региона. Но каждое предприятие должно соответствовать действующему природоохранному законодательству, установленному агентом-супервизором уровню экологической безопасности, а также другим требованиям природоохранного характера. Создание агента-предприятия позволит обеспечить создание системы оценивания экологической эффективности (ОЭЭ) предприятия – это внутренний процесс и инструмент управления, предназначенный для обеспечения руководства информацией о том, соответствует ли экологическая эффективность организации заданным критериям. В общем можно дать следующее определение: ОЭЭ – это постоянный процесс сбора и оценки данных, а также информации для обеспечения текущего оценивания экологической эффективности и тенденций ее изменения со временем. ОЭЭ позволяет сравнить прошлую и настоящую экологическую эффективность организации с критериями этой эффективности[1].

По полученной информации от агента – ресурса агент – предприятия может выявить места высоких выбросов в ОС и провести переговоры с агентом – супервизором (рис.1). На основе этих переговоров производится выбор

системы экономических механизмов, что может включать как существенное изменение типа применяемых механизмов (например, замена механизма платы за выбросы механизмом ограничений на выбросы), так и изменение параметров механизма без изменения его типа (изменение принципов распределения квот, платежей и т.д.)[2].

Процесс переговоров агентов МАСЭБ характеризуется возможностью учета интересов партнеров, формирование планов совместных действий, выявление конфликтов и формирование правил поведения. Система может быть описана множеством состояний $s \in S$, в которых могут находиться агенты с вероятность $P(S)$. Каждый агент - предприятия имеет набор возможных действий - A_i. В какой-то момент времени $t \in [0;T]$ агент находится в состоянии $s \in S$ и предпринимает действие $a_t \in A$ и получает некоторое вознаграждение. Поведение агента - предприятия определено ограничениями, накладываемыми центральным агентом-супервизором. Как правило, этим ограничением является уровень безопасности предприятия (к которому предприятие должно стремиться), ограничения на выбросы загрязняющих веществ и другие административные ограничения. Управление экологической безопасностью в виде итерационного процесса на предприятии позволяет реализовать метод уменьшающихся рисков, разработанный А.А. Быковым[3]. Пусть в начальный момент времени предприятию известны значения экологического риска R_0, ущерба Y_0 и цены экологического риска G_0. Пусть мероприятия по снижению риска и ущерба в сумме составят:

$$Z_{don}^{R_0} + Z_{don}^{Y_0} = Z_{don}^{G_0} \qquad (1)$$

Если эта величина в сумме с новым значением G_1 окажется меньше первоначального значения G_0, то затраты, направленные на снижение риска привели к положительным результатам.

Выбор того или иного действия в момент времени $t \in [0;T]$ осуществляется агентом – предприятия, исходя из его экономических интересов, как правило, стремящегося к максимизации прибыли [4], полученной в результате хозяйственной деятельности, за вычетом затрат, связанных с очисткой выбросов на предприятиях и платой за сброс загрязняющих веществ:

$$f = \int_0^T (D(u,t) - R(u,t) - Z(u,p,t) - N_m(t) - N_l(t) - N_s(t))dt \to \max_p \qquad (2)$$

где t – временная координата; T – момент времени, до которого ведется рассмотрение; $D(u,t)$ – прибыль предприятия от реализации продукции объемом u в момент времени t; $R(u,t)$- затраты предприятия на производство продукции объема u в момент времени t; $Z(u,p,t)$ – затраты предприятия на очистку отходящих газов, на вывоз и утилизацию твердых отходов и на очистку сточных вод. Данные затраты зависят от объемов производства предприятия и доли загрязняющих веществ, удаляемых на предприятии (p) в момент времени t. $N_m(t)$, $N_l(t)$, $N_s(t)$ – отчисляемые платежи за негативное воздействие на окружающую среду в пределах нормы, сверхнормативные и сверхлимитные выбросы в момент времени t.

Задача (2) решается при ограничениях:

$q \le p$, $0 \le t \le T$

где q – минимально допустимая степень очистки отходящих газов и сточных вод.

Рис. 1. Обобщенный алгоритм поведения агента – предприятия

Максимизация прибыли возможна при минимизации затрат предприятия на природоохранные мероприятия или, другими

словами, затрат предприятия на достижение и поддержание необходимого уровня безопасности предприятия. При этом будем считать, что уровень безопасности предприятия - *y* зависит от объема выпускаемой продукции - *u* и средств - *v*, направляемых на природоохранные мероприятия[5]:

$$y = y(u, v) \quad (3)$$

$$\frac{\partial y(u,v)}{\partial u} \le 0, \frac{\partial y(u,v)}{\partial v} \ge 0 \quad (4)$$

Для агента - предприятия эти условия означают следующие правила оценки безопасности предприятия:

- с ростом объема выпуска продукции при неизменных затратах на природоохранные мероприятия уровень безопасности убывает;

- при увеличении затраченных средств на природоохранные мероприятия при постоянных объемах выпуска продукции уровень безопасности не убывает.

Предложенный в работе подход к моделированию агента-предприятия МАСЭБ, позволяет создавать автономного программного агента, способного осуществлять как интеллектуальное, так и реактивное поведение (своевременно реагировать на появление новой информации, изменения в существующих данных и т. д.). Создание агента-предприятия обеспечит создание системы экологической эффективности предприятия, осуществлять непрерывный контроль деятельности предприятия и определить экологические аспекты деятельности предприятия; оценить, какие аспекты являются для нее наиболее важными, задать критерии экологической эффективности и оценить соответствие своей экологической эффективности этим критериям.

[1] Масленникова И.С., Кузнецов Л.М., Пшенин В.Н. *Экологический менеджмент. Учебное пособие*, СПбГИЭУ, 2005 г., 200с.

[2] Бурков В.Н., Щепкин А.В. *Экологическая безопасность*. М.:ИПУ РАН, 2003, 92 с.

[3] Быков А.А. *Моделирование природоохранной деятельности. Учебное пособие*. Москва: НУМЦ Госкомэкологии России, 1998 г.,182 с.

[4] Угольницкий Г.А., Усов А.Б. Математическая формализация методов иерархического управления эколого-экономическими системами.// *Проблемы управления*, 2007, №4.

[5] Бурков В.Н., Новиков Д.А., Щепкин А.В. *Механизмы управления эколого-экономическими системами*. М.: Издательство физико-математической литературы, 2008. – 244с.

Вычисление геометрических параметров поверхности при наличии детерминированных возмущений

Кочетков С.А.
Институт проблем управления им. В.А. Трапезникова РАН
E-mail: ser711@yandex.ru

Abstract

В данной работе рассматривается алгоритм идентификации для систем, измеряющих геометрические характеристики поверхности, при наличии детерминированных возмущений, действующих на измерительный датчик. Решение данной проблемы разбивается на две подзадачи: синтез алгоритма управления с использованием нелинейного компенсатора и синтез алгоритма идентификации с использованием принципов динамической компенсации. Приводятся результаты компьютерного моделирования, показывающие эффективность предложенного подхода.

1. Введение

Большая часть размерного контроля деталей в машиностроении выполняется контактным методом с помощью индуктивных приборов [1]. В таких измерительных системах используется свойство катушки индуктивности изменять своё реактивное сопротивление при изменении положения чувствительного элемента (якоря), касающегося исследуемой поверхности. Обмотки первичного преобразователя питаются от генератора синусоидального сигнала высокой частоты. Сигнал на выходе датчика определяется величиной индуктивностей катушек и, в конечном счете, является линейной функцией перемещения якоря, пропорционального контролируемому параметру. В дальнейшем сигнал усиливается, фильтруется, детектируется подобно процессу обработки сигнала в тракте радиопередачи. Этот или похожий алгоритм используется в подавляющем большинстве приборов рассматриваемого типа. Как видно, такой подход имеет существенный недостаток. Аддитивные и параметрические возмущения (например, магнитные поля) в силу линейности тракта передачи искажают полезный сигнал. Поэтому необходимы схемы компенсации внешних помех.

В данной работе рассматривается проблема подавления внешних детерминированных возмущений для описанных выше измерительных систем. Решение данной задачи разбивается на две подзадачи: синтез алгоритма управления с использованием интегральной нелинейной обратной связи и синтез алгоритма идентификации с использованием принципов динамической компенсации.

Во втором разделе ставится задача исследования, а также обсуждаются некоторые особенности решаемой проблемы.

Далее рассматривается синтез простейшего нелинейного динамического компенсатора, в первом приближении оценивается ошибка компенсации.

В четвёртом разделе рассматривается алгоритм управления с использованием предложенного нелинейного компенсатора, а также алгоритм идентификации с компенсацией возмущений в наблюдателе.

Далее приводятся результаты компьютерного моделирования, демонстрирующие свойство грубости, которое придаёт исходной системе предложенная нелинейная интегральная обратная связь. Приводятся также результаты оценивания параметров с помощью предложенных алгоритмов.

2. Постановка задачи

Рассмотрим конструкцию первичного преобразователя (рисунок 1).

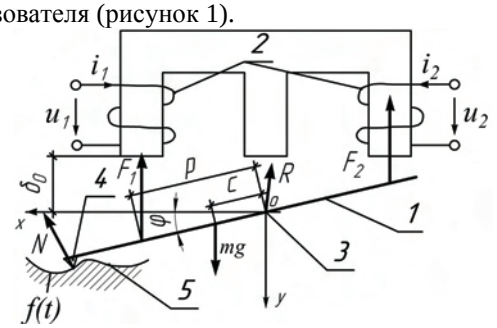


Рисунок 1 – конструкция датчика

Пусть управляемый объект описывается уравнениями [2]

$$\dot{\varphi} = \omega;$$

$$\dot{\omega} = \frac{mgc}{J_0} + \frac{pk}{2J_0}\left(\frac{i_2^2}{(\delta_0 - p\varphi)^2} - \frac{i_1^2}{(\delta_0 + p\varphi)^2}\right) + \frac{M(t)}{J_0}; \quad (1)$$

$$\frac{di_1}{dt} = \frac{pi_1\omega}{\delta_0 + p\varphi} - \frac{i_1 R_1(\delta_0 + p\varphi)}{k} + \frac{u_1(\delta_0 + p\varphi)}{k} + \xi_1;$$

$$\frac{di_2}{dt} = -\frac{pi_2\omega}{\delta_0 - p\varphi} - \frac{i_2 R_2(\delta_0 - p\varphi)}{k} + \frac{u_2(\delta_0 - p\varphi)}{k} + \xi_2,$$

где m – масса якоря; R_1, R_2 – активные сопротивления обмоток якоря; δ_0 – номинальный зазор между сердечником и якорем; φ, ω – угловое положение и угловая скорость якоря; i_1, i_2 – токи в обмотках якоря, измеряемые точно; u_1, u_2 – управляющие воздействия; k, p, c – конструктивные коэффициенты; J_0 – момент инерции якоря относительно точки подвеса; $M(t)$ – момент реакции опоры поверхности; ξ_1, ξ_2 – внешние детерминированные возмущения.

Ставится задача идентификации величины φ, пропорциональной контролируемому параметру, по результатам измерения токов i_1, i_2.

Основное допущение относительно ξ_1, ξ_2 состоит в предположении, что они и их первые производные ограничены. Кроме того, спектр ξ_1, ξ_2 принадлежит к области низких частот. Отметим также, что поскольку обмотки датчика находятся примерно в одинаковых условиях, то величины ξ_1, ξ_2 приблизительно равны. Поэтому динамическую компенсацию внешних возмущений будем осуществлять только в уравнении для суммы токов. Необходимость этого обусловлена следующими соображениями. На установившиеся значения переменных x_1 и x_2. влияют не только напряжения u_1, u_2, но и внешние возмущения. В этом случае использование x_1 в качестве опорного сигнала для решения задачи идентификации может привести к неустойчивости наблюдателя [3] из-за смещений, вызванных ξ_1, ξ_2. Для уменьшения влияния возмущений на ошибку вычисления параметра, характеризующего контролируемый профиль, предлагается динамическая компенсация возмущений в наблюдателе.

В завершение данного раздела ещё раз отметим, что современные приборы для измерения геометрических параметров строятся на принципе амплитудной модуляции. Данный подход хорошо зарекомендовал себя на протяжении многих лет. Поэтому предлагается взять за основу режим автоколебаний. Однако задачу параметрической идентификации предлагается решить современными методами [4].

3. Динамический компенсатор

Модель объектов управления, как правило, не точна. Неточности могут быть следствием допущений сделанных относительно теоретических положений о процессе, следствием линеаризации уравнений нелинейной модели, могут возникать из-за ошибок в задании параметров процесса и по другим причинам. Кроме того, управляемые процессы часто подвержены влиянию возмущений, которые нельзя измерить. Одним из наиболее эффективных методов обеспечения инвариантности к внешним детерминированным возмущениям является метод динамической компенсации [5, 6]. Основная его идея заключается во введении в замкнутый контур системы управления динамического звена, имеющего структуру модели возмущений. Если в такой системе подогнать начальные условия динамического компенсатора и модели возмущений, то появляется возможность компенсировать воздействие возмущений с помощью выходного сигнала компенсатора. Таким образом, задача обеспечения инвариантности сводится к задаче стабилизации расширенной системы, включающей модель объекта управления и динамический компенсатор возмущений. Отметим, что полная инвариантность в системе достигается лишь после совпадения выходных сигналов модели возмущений и динамического компенсатора и, поскольку в линейных системах эти сигналы сходятся асимптотически, такую инвариантность называют асимптотической.

Рассмотрим построение динамического компенсатора возмущений, действующих на объект управления.

Перепишем (1) относительно суммы $x_1 = i_1 + i_2$ и разности токов $x_2 = i_1 - i_2$

$$\begin{cases} \dot{x}_1 = -\alpha_1 x_1 - \alpha_2 x_2 + b_1 \Sigma u + b_2 \Delta u + \xi_1^*; \\ \dot{x}_2 = -\alpha_2 x_1 - \alpha_1 x_2 + b_1 \Delta u + b_2 \Sigma u + \xi_1^*, \end{cases} \quad (2)$$

где $\alpha_1 = (a_{11} + a_{22})/2$, $\alpha_2 = (a_{11} - a_{22})/2$;

$a_{11} = -p\omega/(\delta_0 + p\varphi) + R_1(\delta_0 + p\varphi)/k$;

$a_{11} = p\omega/(\delta_0 - p\varphi) + R_2(\delta_0 - p\varphi)/k$;

$\Sigma u = u_1 + u_2$; $b_1 = \delta_0/k$; $b_2 = p\varphi/k$;

$\Delta u = u_1 - u_2$; $\xi_1^* = \xi_1 + \xi_2$; $\xi_2^* = \xi_1 - \xi_2$.

Как уже описывалось, компенсацию возмущений будем осуществлять только для переменной x_1.

Рассмотрим первое уравнение системы (2)

$$\dot{x}_1 = -\alpha_1 x_1 + b_1 \Sigma u + f,$$

где $f = -\alpha_2 x_2 + b_1 \Sigma u + b_2 \Delta u + \xi_1^*$.

Введём простейшую модель возмущений в виде $\dot{f} = 0$.

Следуя идее динамической компенсации, расширим пространство состояний введением новой переменной

$$\begin{cases} \dot{x}_1 = -\alpha_1 x_1 + b_1 \Sigma u + f; \\ \dot{z}_1 = \nu_1, \end{cases}$$

где ν_1 – корректирующее воздействие, позволяющее подогнать начальные условия динамического компенсатора и модели возмущений.

В рамках классического подхода [5] управляющие воздействия u_1, u_2, ν_1 выбираются в виде

$$\Sigma u = -l_2 z_1; \nu_1 = l_1 x_1.$$

Таким образом, замкнутая система приобретает вид

$$\begin{cases} \dot{x}_1 = -\alpha_1 x_1 - b_1 l_2 z_1 + f; \\ \dot{z}_1 = l_1 x_1. \end{cases}$$

Установившиеся значения переменных x_1, z_1

$$\begin{cases} x_1 = 0; \\ z_1 = \dfrac{f}{b_1 l_2}. \end{cases}$$

Коэффициенты l_1, l_2 определяют скорость сходимости переменной x_1 к началу координат.

Данный алгоритм управления реализован в широко распространённых в технике ПИ-регуляторах. Описанный подход является негрубым и предполагает, что модель возмущений известна, что не всегда имеет место на практике. Кроме того, хорошо известно, что ПИД-регуляторы плохо справляются со своими задачами в случае, когда внешнее возмущение имеет существенно нелинейный характер. В данной работе предлагается ввести нелинейный динамический компенсатор с использованием разрывных воздействий. Рассмотрим один из вариантов такого компенсатора.

Устремим коэффициент усиления l_2 к бесконечности и учтём, что физически можно реализовать управление лишь конечной амплитуды. Тогда получим управляющие воздействия u_1, u_2, ν_1 в виде

$$\Sigma u = -M_1 sign(z_1); \quad \nu_1 = l_1 x_1.$$

Выбор управлений в виде релейной функции придаёт системе свойство колебательности.

Рассмотрим функцию Ляпунова в виде

$$V = M_1 b_1 |z_1| / l_1 + x_1^2 / 2 - z_1 f / l_1.$$

Её производная в силу системы с учётом управляющих воздействий

$$\dot{V} = -\alpha_1 x_1^2 < 0.$$

Таким образом, получим, что в асимптотике при $t \to \infty$

$x_1 = 0; z_1 = 0; M_1 sign(z_1) = f$.

Известен также, так называемый twisting-алгоритм[6], который обеспечивает движение в скользящем режиме по пересечению поверхностей $x_1 = 0, z_1 = 0$. При этом указанные соотношения выполняются за конечное время. При использовании этого алгоритма управляющие воздействия u_1, u_2, ν_1 выбираются в виде

$$\Sigma u = -M_1 sign(z_1) - M_2 sign(x_1); \nu_1 = l_1 x_1.$$

Отметим, что описанные регуляторы обеспечивают подавление переменных во времени, ограниченных возмущений. Однако со временем частота переключений растёт неограниченно до бесконечности, и реализовать такие регуляторы представляется невозможным. Необходимо получить практически реализуемый алгоритм управления с конечной частотой переключения. Соответственно возникает задача возбуждения колебаний в системе (3) с помощью управляющих воздействий u_1, u_2, ν_1.

Пусть

$$\Sigma u = -M_1 sign(z_1) + u'_1; \nu_1 = l_1 x_1,$$

где u'_1 – корректирующее воздействие, позволяющее компенсировать диссипации энергии в системе, обусловленные компонентой $-\alpha_1 x_1$ и обеспечить устойчивые автоколебания с заданной амплитудой. Можно выбрать u'_1 различными способами. Примем

$$\begin{cases} \Sigma u = -M_1 sign(z_1) + M_2 sign(x_1); \\ \nu_1 = l_1 x_1. \end{cases} \qquad (3)$$

Вновь, рассматривая положительно определённую функцию

$$V = M_1 b_1 |z_1| / l_1 + x_1^2 / 2 - z_1 f / l_1.$$

и вычисляя её производную, получим

$$\dot{V} = |x_1| \left(-\alpha_1 |x_1| + M_2 \right).$$

Видно, что если в среднем на периоде колебаний выполняется соотношение $|x_1| = M_2 / \alpha_1$, то $\dot{V} = 0$ и, следовательно $V = const$, что означает возникновение предельного цикла в системе (2) при управляющих воздействиях (3). Найдём с помощью метода гармонического баланса приблизительные соотношения, определяющие конечную амплитуду и частоту колебаний.

Представим переменные x_1, z_1 в виде

$$x_1 = x_{10} + X cos(\omega t), z_1 = z_{10} + Z sin(\omega t),$$

где x_{10}, z_{10} и X, Z – амплитуды медленно меняющихся и быстро меняющихся составляющих; ω – частота автоколебаний.

Заменим нелинейные функции их приближением с использованием ряда Фурье. Как известно, коэффициенты этого ряда для постоянной составляющей и первой гармоники определяются соотношениями

$$\begin{cases} M_2 sign(x_1) \approx M_2\ arcsin\left(\dfrac{x_{10}}{X}\right) + \dfrac{4M_2}{\pi}\sqrt{1-\left(\dfrac{x_{10}}{X}\right)^2}\ cos(\omega t); \\ M_1 sign(z_1) \approx M_1\ arcsin\left(\dfrac{z_{10}}{Z}\right) + \dfrac{4M_1}{\pi}\sqrt{1-\left(\dfrac{z_{10}}{Z}\right)^2}\ sin(\omega t). \end{cases}$$

Запишем уравнения для медленно меняющихся составляющих

$$\begin{cases} \dot{x}_{10} = -\alpha_1 x_{10} - M_1\ arcsin\left(\dfrac{z_{10}}{z}\right) + M_2\ arcsin\left(\dfrac{x_{10}}{x}\right) + f; \\ \dot{z}_{10} = l_1 x_{10}. \end{cases}$$

Как видно, при $f = const$ в установившемся режиме

$z_{10} = Z\ sin(f / M_1);\quad x_{10} = 0$.

Теперь запишем выражения для амплитуд и частоты быстро меняющихся составляющих

$$\begin{cases} -\omega X\ sin(\omega t) = -\alpha_1 X\ cos(\omega t) - \dfrac{4M_1}{\pi} cos\left(\dfrac{f}{M_1}\right) sin(\omega t) + \\ + \dfrac{4M_2}{\pi} cos\ cos(\omega t);\ Z\omega\ cos(\omega t) = l_1 X\ cos(\omega t). \end{cases}$$

Непосредственно вычисляя, получим

$$\begin{cases} X = \dfrac{4M_2}{\pi\alpha_1}; Z = \dfrac{4l_1 M_2^2}{\pi\alpha_1^2 M_1\ cos(f / M_1)}; \\ \omega = \dfrac{M_1\alpha_1}{M_2} cos(f / M_1). \end{cases}$$

Таким образом, с помощью коэффициентов l_1, M_1, M_2 можем задать необходимый нам установившийся режим.

Рассмотрим теперь качественно поведение переменной x_1 в случае, когда $\dot{f} \neq 0$. При $M_1 > |f|$ функции $M_1\ arcsin(z_{10} / Z), M_2\ arcsin(x_{10} / X)$ достаточно точно аппроксимируются выражениями

$$\begin{cases} M_1\ arcsin(z_{10} / Z) \approx \dfrac{\pi\alpha_1^2 M_1^2 \sqrt{1-(z_{10} / Z)^2}}{4l_1 M_2^2} z_{10}; \\ M_2\ arcsin(x_{10} / X) \approx \dfrac{M_2}{X(x_{10})} x_{10}. \end{cases}$$

Переходя снова к системе для медленных составляющих, получим

$$\begin{cases} \dot{x}_{10} = -\left(\alpha_1 - \dfrac{M_2}{X(x_{10})}\right) x_{10} - \dfrac{\pi\alpha_1^2 M_1^2 \sqrt{1-(z_{10} / Z)^2}}{4l_1 M_2^2} z_{10} + f; \\ \dot{z}_{10} = l_1 x_{10}. \end{cases}$$

Как видно, в общем случае система является нелинейной. Предположим, однако, что система устойчива. Применяя процедуру линеаризации относительно некоторого x_{10}^* и принимая

$M_1\ arcsin(z_{10} / Z) \approx \pi\alpha_1^2 M_1^2\ cos(f / M_1) / (4l_1 M_2^2) z_{10}$

,

получим

$\dot{x}_{10} = -\alpha_1^* x_{10} - l_2 z_{10} + f; \dot{z}_{10} = l_1 x_{10}$,

где $l_2 = \pi\alpha_1^2 M_1^2\ cos(f / M_1) / (4l_1 M_2^2)$, $\alpha_1^* > 0$.

Характеристическое уравнение данной системы

$p^2 + \alpha_1 p + l_1 l_2 = 0.$

Выберем коэффициенты M_1, M_2 так, чтобы выполнялось условие $\pi M_1^2\ cos(f / M_1) / M_2^2 >> 1.$

Тогда корни характеристического уравнения будут комплексно сопряжёнными с частотой ω_0.

Найдём переходную функцию для переменной x_{10}, описываемой уравнениями

$$\ddot{x}_{10} + \alpha_1^* \dot{x}_{10} + l_1 l_2 x_{10} = \dot{f}. \qquad (4)$$

Импульсная переходная функция

$$g(t) = -e^{-\frac{\alpha_1^*}{2}t}\left(\alpha_1^{*2} + 4\omega_0^2\right) / (4\omega_0 l_1)\ sin(\omega_0 t).$$

Тогда решение линеаризованного уравнения (4) при нулевых начальных условиях

$$x_{10}(t) = \left(\frac{\alpha_1^{*2} + 4\omega_0^2}{4\omega_0 l_1 l_2}\right)\int_0^t e^{-\frac{\alpha_1^{*2}}{2}(t-\tau)}\ sin(\omega_0(t-\tau))\dot{f}(\tau)d\tau.$$

Введём ограничения $|\dot{f}(\tau)| \le C$. Тогда

$$|x_{10}(t)| \le \left(\frac{\alpha_1^{*2} + 4\omega_0^2}{4\omega_0 l_1 l_2}\right)\int_0^t e^{-\frac{\alpha_1^*}{2}(t-\tau)}\ sin(\omega_0(t-\tau))\,|\dot{f}(\tau)|d\tau \le$$

$$\le \left(\frac{\alpha_1^{*2} + 4\omega_0^2}{4\omega_0 l_1 l_2}\right) C\int_0^t e^{-\frac{\alpha_1^{*2}}{2}(t-\tau)}\ sin(\omega_0(t-\tau))d\tau.$$

Вычисляя интеграл, при $t \to \infty$ получим

$$|x_{10}(t)| \le \frac{C}{l_1 l_2} = \frac{4M_2^2 C}{\pi\alpha_1^2 M_1^2\ cos(f / M_1)}.$$

Как видно выбором параметров M_1, M_2 можно сделать величину x_{10} сколь угодно малой, что означает высокую робастность рассмотренного нелинейного динамического компенсатора по отношению к внешним детерминированным возмущениям. Однако не следует забывать, что отношение M_1 / M_2 определяет частоту автоколебаний, а она ограничена сверху возможностью технической реализации.

4. Выбор алгоритма управления и идентификации

В предыдущем разделе было показано, что в установившемся режиме существует устойчивый предельный цикл. В данном случае этот эффект является полезным, поскольку мы получаем пару линейно независимых опорных сигналов, которые могут быть использованы для построения идентификатора на скользящих режимах или с большими коэффициентами. Выберем управляющие воздействия в виде (3)

$$\begin{cases} \Sigma\, u = -M_1 sign(z_1) + M_2 sign(x_1); \\ \Delta\, u = 0; \\ \nu_1 = l_1 x_1 \end{cases} \quad (5)$$

Параметры M_1, M_2 выбираются так, чтобы задать номинальный режим, на который рассчитан датчик.

Используем принцип динамической компенсации [7] для решения задачи идентификации при наличии внешних возмущений. Введём модель внешних возмущений

$$\begin{cases} \dot{\xi}_1^* = H_1; \\ \dot{H}_1 = 0. \end{cases} \quad \begin{cases} \dot{\xi}_2^* = H_2; \\ \dot{H}_2 = 0. \end{cases} \quad (6)$$

Уравнения идентификаторов выбираются в виде аналогичном (2), (6)

$$\begin{cases} \dot{\hat{x}}_1 = -\hat{\alpha}_1 x_1 + \hat{b}_1 \Sigma\, u + \hat{\xi}_1^* + L\bar{x}_1; \\ \dot{\hat{\alpha}}_1 = -F_1 L\bar{x}_1 x_1; \\ \dot{\hat{b}}_1 = F_1 L\bar{x}_1 x_1 \\ \dot{\hat{\xi}}_1^* = \hat{H}_1 + F_1 L\bar{x}_1; \\ \dot{\hat{H}}_1 = F_2 L\bar{x}_1; \end{cases} \quad (7.1)$$

$$\begin{cases} \dot{\hat{x}}_2 = -\hat{\alpha}_2 x_1 - \hat{\alpha}_1 x_2 + \hat{b}_2 \Sigma\, u + \hat{\xi}_2^* + L\bar{x}_2; \\ \dot{\hat{\alpha}}_2 = -F_1 L\bar{x}_2 x_2; \\ \dot{\hat{b}}_2 = F_1 L\bar{x}_2 x_2 \\ \dot{\hat{\xi}}_2^* = \hat{H}_2 + F_1 L\bar{x}_2; \\ \dot{\hat{H}}_2 = F_2 L\bar{x}_2, \end{cases} \quad (7.2)$$

где $F_1 < F_2$, L – положительные коэффициенты; $\bar{x}_1 = x_1 - \hat{x}_1$, $\bar{x}_2 = x_2 - \hat{x}_2$; $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\xi}_1^*$, $\hat{\xi}_2^*$, $\hat{V}_1$, $\hat{b}_1$, $\hat{b}_2$ – оценки величин α_1, α_2, ξ_1^*, ξ_2^*, V_1, b_1, b_2. Отметим, что в уравнении для $\hat{x}_1$ мы преднамеренно пренебрегли величиной $\hat{\alpha}_2 x_2$ в силу её малости, потому что коэффициент α_2 близок к нулю, поскольку величины сопротивлений R_1, R_2 приблизительно равны, величина ЭДС движения пренебрежительно мала, а сумма x_1 и разность токов и разности токов x_2 различаются также как параметры δ_0 и $p\varphi$. Компонента $\hat{\alpha}_1 x_2$ в уравнении для $\hat{x}_2$ вводится с помощью оценки $\hat{\alpha}_1$ из первого идентификатора.

Запишем уравнения относительно невязок (производные неизвестных параметров принимаются приблизительно равными нулю)

$$\dot{\bar{x}}_1 = -\bar{\alpha}_1 x_1 - \alpha_2 x_2 + \bar{b}_1 (u_1 + u_2) + \bar{\xi}_1^* - L\bar{x}_1;$$
$$\dot{\bar{\alpha}}_1 = F_1 L\bar{x}_1 x_1;$$
$$\dot{\bar{b}}_1 = -F_1 L\bar{x}_1 x_1;$$
$$\dot{\bar{\xi}}_1^* = \bar{H}_1 - F_1 L\bar{x}_1;$$
$$\dot{\bar{H}}_1 = -F_2 L\bar{x}_1.$$

Рассмотрим первую подсистему. При $L \to \infty$ получим

$$L\bar{x}_1 \to -\bar{\alpha}_1 x_1 - \alpha_2 x_2 + \bar{b}_1 \Sigma\, u + \bar{\xi}_1^*;$$
$$\dot{\bar{\alpha}}_1 = F_1 \left(-\bar{\alpha}_1 x_1 - \alpha_2 x_2 + \bar{b}_1 \Sigma\, u + \bar{\xi}_1^* \right) x_1;$$
$$\dot{\bar{\xi}}_1^* = \bar{H}_1 - F_1 \left(-\bar{\alpha}_1 x_1 - \alpha_2 x_2 + \bar{b}_1 \Sigma\, u + \bar{\xi}_1^* \right);$$
$$\dot{\bar{b}}_1 = -F_1 \left(-\bar{\alpha}_1 x_1 - \alpha_2 x_2 + \bar{b}_1 \Sigma\, u + \bar{\xi}_1^* \right) \Sigma\, u;$$
$$\dot{\bar{H}}_1 = -F_2 \left(-\bar{\alpha}_1 x_1 - \alpha_2 x_2 + \bar{b}_1 \Sigma\, u + \bar{\xi}_1^* \right)$$

Выберем функцию Ляпунова в виде

$$V_1 = \frac{1}{2} \left(\bar{\alpha}_1^2 + \bar{b}_1^2 + \bar{\xi}_1^{*2} + \bar{H}_1^2 \right).$$

Отметим, что при достаточно большом F_1:

$$\bar{\xi}_1^* \to \bar{H}_1 / F_1 + \bar{\alpha}_1 x_1 + \alpha_2 x_2 - \bar{b}_1 \Sigma\, u.$$

Подставляя данное выражение в уравнение для $\bar{H}_1$ получим

$$\dot{\bar{H}}_1 = -F_2 / F_1 \bar{H}_1.$$

Следовательно, $\bar{H}_1 \to 0$ асимптотически.

Учитывая данные выражения, найдём производную V_1

$$\dot{V}_1 = -F_1 \left(\bar{\alpha}_1 x_1 - \bar{b}_1 \Sigma\, u - \bar{\xi}_1^* \right)^2 + \chi_1.$$

где χ_1 – достаточно малая величина, отражающая влияние компоненты $\alpha_2 x_2$. Как видно, при $\chi_1 = 0$ функция $\dot{V}_1$ является отрицательно определённой. Следовательно, V_1 сходится к области начала координат, размеры которой определяются величиной χ_1. Аналогичным образом доказывается сходимость к некоторой области невязок для (7.2).

5. Результаты моделирования

Моделирование проводилось при следующих значениях параметров и условиях:

Омические сопротивления катушек, Ом

$R_1 = 90$, $R_2 = 80$;

Номинальный зазор, м

$\delta_0 = 0{,}003$;

Конструктивный параметр, м/Гн

$k = 0{,}24 \cdot 10^{-3}$;

Конструктивный параметр, м

$p = 0{,}02$;

Параметры управляющих воздействий

$M_1 = 10;\quad M_2 = 0{,}08;\qquad l_1 = 1000$.

Начальные условия

$i_1(0) = 0{,}001$, $i_2(0) = -0{,}007$.

Внешнее возмущение изменяется по синусоидальному закону с частотой напряжения сети 50 Гц $\xi_1^* = 50\sin(100\pi t)$; $f \approx \xi_1^*$.

Переходные процессы в системе (2) с управляющими воздействиями (4) приведены на рисунке 2.

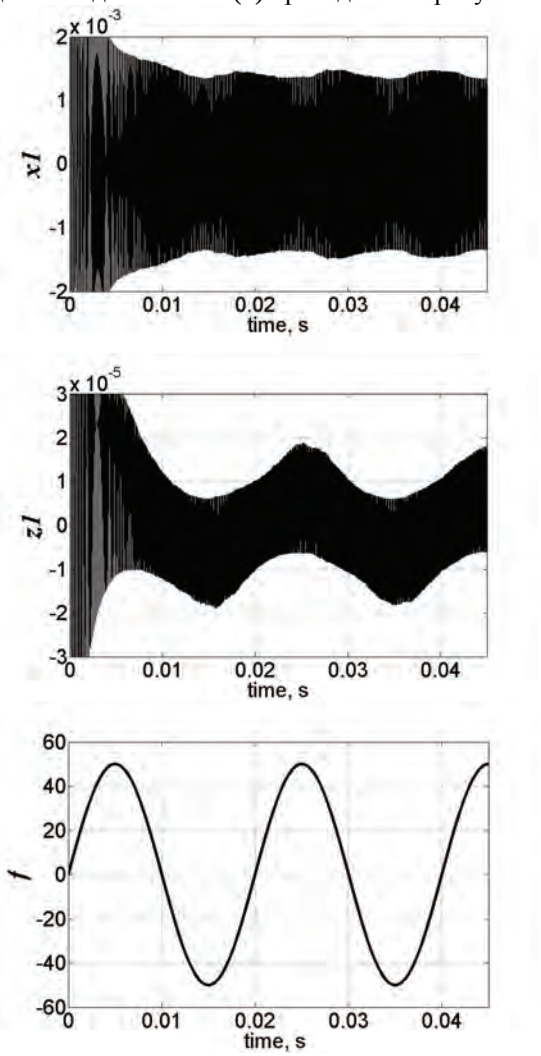


Рисунок 2 – Моделирование системы (2) с управлениями (4)

Как видно, $x_1(t)$ является «хорошим» опорным сигналом для решения задачи параметрической идентификации, поскольку его амплитуда слабо зависит от возмущения f.

Пусть переменные ξ_2^* и φ изменяются по законам

$\xi_2^*(t) = 0{,}125\sin(100\pi t)$;

$\varphi(t) = 2\cdot 10^{-6}\sin(500t)$.

Результаты моделирования работы наблюдателей (7.2) приведены на рисунке 3. Отметим, что оценка параметра $\hat{\varphi}(t)$ производится согласно выражению

$$\hat{\varphi}(t) = \frac{\hat{b}_2 k}{p}.$$

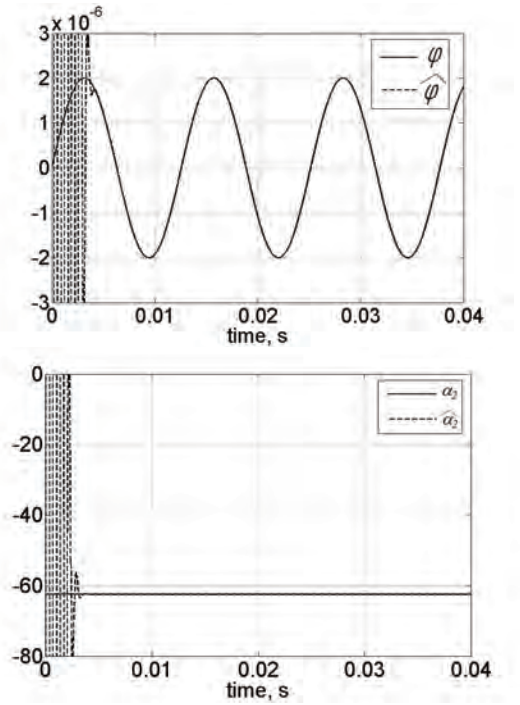


Рисунок 3 – Моделирование работы наблюдателей (7.2)

6. Список литературы

[1] Г.С. Бирюков, А Л. Серко, *Измерение геометрических величин и их метрологическое обеспечение*, Издво стандартов, Москва, 1987.

[2] Ю. Н. Журавлёв. *Активные магнитные подшипники: Теория, расчёт, применение*, Политехника, Санкт-Петербург, 2003.

[3] Б.Н. Петров, С.Д. Земляков, В.Ю. Рутковский, *Адаптивное координатно-параметрическое управление нестационарными объектами*, Наука, Москва, 1980.

[4] В.И. Уткин, "Принципы идентификации на скользящих режимах", *ДАН СССР*, Наука, Москва, 1981. Т. 257, № 3. С. 558-561.

[5] Ю.Н. Андреев, *Управления конечномерными линейными объектами*, Наука, Москва, 1976.

[6] A. Levant, "Higher order sliding modes, differentiation and output feedback control", *International Journal of Control*, Taylor&Francis, London, 2003, 76(9-10), pp. 924-941.

[7] V.A. Utkin, "The method of dynamic compensation in a problem of estimation of external disturbances", *In proc. SICPRO'06*, IPU RAS, Moscow, 2006, pp. 754-771.

Оптимизация процесса центрифугирования путем определения текущей массы в барабане

Ханина Н. Ю.
Липецкий государственный технический университет
lisky@lipetsk.ru

Аннотация

Для увеличения производительности сахарной центрифуги необходимо знать текущую массу суспензии в барабане. Например, при загрузке слишком большая или маленькая загруженная масса ведёт к неоправданно высокому расходу энергии.

Центрифуга представляет собой электромеханическую систему "электродвигатель – барабан с загружаемой массой". Во время вращения эта система описывается основным уравнением движения электропривода, используя которое, датчик вычисляет массу, находящуюся в барабане в каждый момент времени.

Но в уравнении движения не одна, а две неизвестные величины. Помимо массы, присутствует ещё статический момент на валу электродвигателя. Его строгое определение не представляется возможным из-за сложных гидромеханических процессов, происходящих во время центрифугирования. Поэтому в использующихся на производстве датчиках предполагается, что статический момент постоянен. Однако в реальных условиях это не так, что приводит к погрешностям в определении массы и к снижению качества продукта.

В работе на основании экспериментальных данных определяется вид зависимости статического момента от времени в течение центрифугирования.

1. Введение

Принцип работы существующих в настоящее время устройств автоматической загрузки основан на решении основного уравнения движения электропривода

$$M - M_c = J\frac{dw}{dt} + \frac{w}{2}\frac{dJ}{dt}, \qquad (1)$$

или, в более наглядной форме,

$$J' = -\frac{2w'}{w}J + \frac{2}{w}(M - M_c), \qquad (2)$$

относительно момента инерции загружаемого барабана центрифуги J. В этом уравнении значения силы тока $I(t)$ и скорости вращения барабана $w(t)$ снимаются датчиком через определенные промежутки времени. Момент на валу двигателя M прямо пропорционален силе тока I. Коэффициент пропорциональности известен – это постоянная двигателя KF_n.

Используя полученное из (1) значение J и размеры барабана, датчик вычисляет текущую массу суспензии, находящуюся в барабане. По значениям массы определяется момент прекращения загрузки. Проблема состоит в том, что в уравнении (1), кроме J, есть еще одна неизвестная. Это M_c – статический момент на валу электропривода, возникающий из-за разного рода сил трения. В предыдущих устройствах M_c считался постоянным в пределах одной загрузки. Однако на самом деле это не совсем так, хотя бы потому, что длина «жгута» вливаемой в барабан центрифуги суспензии во время загрузки уменьшается, уменьшается и M_c. Эта не учитываемая особенность приводит к дополнительным ошибкам всего устройства в целом.

Но можно предположить, что вид зависимости $M_c = f(t)$ не сильно изменяется от загрузки к загрузке. Эту зависимость можно определить на основании опытных данных.

Для того, чтобы найти зависимость $M_c = f(t)$, был проведен ряд экспериментов, позволяющих определять время и степень загрузки. Из всего многообразия опытов были выбраны четыре, в которых реологические свойства загружаемой массы наиболее отличались. По экспериментальным данным (ток двигателя и скорость барабана) определялись

момент инерции загруженной суспензии $J(t_z)$ и время загрузки t_z.

Загрузку в барабан центрифуги производят с постоянным напуском K_m, то есть загружаемая масса изменяется по закону:

$$m(t) = K_m t . \quad (3)$$

Загружаемая в барабан масса во вращающемся барабане образует цилиндр. Закон изменения момента инерции цилиндра имеет вид:

$$J(t) = J_0 + m(t)\ R_0^{\ 2} - \frac{m(t)^2}{2\pi\rho H}, \quad (4)$$

где $J_0 = 433 кг \cdot м^3$ – момент инерции пустого барабана, $\rho = 1500 кг/м^3$ – плотность суспензии, $R_0 = 0{,}685 м$ – радиус барабана, $H = 1{,}06 м$ – высота барабана центрифуги.

Для момента окончания загрузки формулу (4) можно записать в виде (с учетом формулы (3)):

$$J(t_z) = J_0 + K_m R_0^{\ 2} t_z - \frac{K_m^{\ 2} t_z^{\ 2}}{2\pi\rho H}. \quad (5)$$

Решая квадратное уравнение (5) относительно K_m, находим темп напуска при загрузке для каждого опыта. Таким образом, выражение (3) становится определенным в любой момент времени загрузки и его можно подставлять в уравнение движения электропривода (1). Вычислить присутствующую в уравнении движения производную момента инерции не составляет труда:

$$\frac{dJ(t)}{dt} = K_m R_0^{\ 2} - \frac{K_m^{\ 2} t}{\pi\rho H} . \quad (6)$$

Теперь в уравнении движения электропривода (1) осталась одна неизвестная переменная – статический момент M_c, который легко найти.

Для построения кривой зависимости $M_c(t)$ во время загрузки был использован математический аппарат MathCAD 11. В пакет MathCAD включены средства, предназначенные для обработки физического эксперимента. Он содержит программы интерполирования, регрессии, сглаживания данных. Конкретные применения методов обработки данных к решаемой задаче описаны ниже.

2. Решение задачи оптимизации загрузки

Подробно опишем решение задачи по определению зависимости $M_c = f(t)$ при загрузке на примере одного опыта. Опытные данные представляют собой массивы значений тока двигателя и скорости барабана, снятые датчиком через 0,1 секунды. В среде математического пакета MathCAD находим регрессионные полиномы $w(t)$ и $I(t)$ для массивов значений тока и скорости. Регрессия необходима для устранения погрешности, возникающей при снятии экспериментальных данных [3]. Для скорости были получены полиномы девятой степени.

Помимо скорости, в уравнении движения электропривода присутствует ее производная $\frac{dw(t)}{dt}$, которую мы находим, дифференцируя полученный регрессионный полином.

Для тока также были получены полиномы девятой степени.

Регрессионные многочлены описывают исходные кривые достаточно точно. Коэффициенты корреляции были не менее 0.999.

Для определения момента инерции $J(t)$ нужно найти темп напуска K_m. Темп напуска можно найти из уравнения (5), зная момент инерции загруженного барабана $J(t_z)$. Время загрузки известно для каждого опыта.

Для определения момента инерции загруженной массы были созданы специальные условия сразу после загрузки. На одну секунду принудительно отключался ток. При этом принималось допущение, что загруженная масса не успевает значительно измениться за этот малый промежуток времени.

Таким образом, были получены два участка движения (до отключения и после отключения тока), описываемыми двумя линейно независимыми уравнениями движения электропривода:

$$\begin{cases} M_1 - M_{c1} = J_1 w_1' + \frac{1}{2} w_1 J_1' \\ M_2 - M_{c2} = J_2 w_2' + \frac{1}{2} w_2 J_2' \end{cases} .$$

За небольшой отрезок времени, на который ток был отключен, статический момент не мог сильно измениться, поэтому $M_{c1} = M_{c2}$. Изменение массы, а, следовательно, и момента инерции за это же время тоже пренебрежимо мало: $J_1 = J_2 = J_z$. То же можно сказать про производную момента инерции и скорость $J'_1 = J'_2$, $w_1 = w_2$. Учитывая это и вычитая из первого уравнения второе, получим простую формулу для расчета момента инерции:

$$J_z = \frac{M_1 - M_2}{w_1' - w_2'},$$

или для токов:

$$J_z = \frac{KF_n(I_1 - I_2)}{w_1' - w_2'},$$

где I_1 – средний ток первого участка,

I_2 – средний ток второго участка,

KF_n – постоянная двигателя,

w_1' – ускорение на первом участке,

w_2' – ускорение на втором участке.

Так рассчитывается момент инерции барабана после загрузки для каждого опыта.

Теперь из уравнения (5) можно определить темп напуска для конкретного опыта.

Подставляя полученное значение темпа напуска K_m в уравнения (3), (4) и (6), находим текущие значения момента инерции и его производной при загрузке.

Теперь, зная все функции, присутствующие в уравнении движения электропривода (1), можно определить искомую зависимость $M_c = f(t)$ во время загрузки. Графики зависимостей $I_c = f(t)$ для каждого опыта были построены с использованием «MathCAD 11». Напомним, что M_c и I_c (ток статический) связаны постоянным коэффициентом KF_n:

$$M_c = KF_n \cdot I_c .$$

Для полученной зависимости $I_c = f(t)$ найдем регрессионную кривую в виде $y(t) = I_{cp1}(ce^{-t} + at + b)$.

Здесь I_{cp1} – средний ток двигателя в течение первой секунды загрузки, общий параметр, характеризующий реологические свойства загружаемой суспензии. Так как реологические свойства меняются от загрузки к загрузке, коэффициенты регрессионных кривых для разных опытов будут значительно отличаться. Чтобы уменьшить эту разницу, вводится общий параметр I_{cp1}.

Коэффициенты регрессионных кривых были определены как среднее арифметическое четырех рассмотренных опытов.

Эмпирически выведенная формула тока статического во время загрузки принимает вид:

$$I_c(t) = I_{cp1}(0.88e^{-t} - 0.015t + 0.94) \qquad (7)$$

А для момента статического это будет:

$$M_c(t) = KF_n I_{cp1}(0.88e^{-t} - 0.015t + 0.94) \qquad (8)$$

3. Описание задачи оптимизации оттека

После загрузки сахарная центрифуга периодического действия начинает набирать обороты (разгоняться). При этом жидкая фаза загруженной массы (оттек) под действием центробежных сил отходит через отверстия в барабане, а твердая (кристаллы сахара) остаются. Согласно технологии центрифугирования, в момент окончания оттека необходимо подать горячую воду, чтобы смыть остатки межкристальной жидкости. Момент подачи воды, так же как и ее количество, очень важные характеристики процесса центрифугирования, влияющие на производительность центрифуги и качество сахара. Более ранняя подача будет неэффективна из-за того, что не весь оттек еще отошел и потребуется большее количество воды, а при запоздалой придется смывать засохшую пленку с кристаллов сахара, что также ведет к увеличению подаваемой воды. В обоих случаях налицо еще и прямые потери, так как излишки воды размоют часть кристалла. Момент окончания оттека, необходимое количество воды для промывки кристаллов, и пара для пробелки можно определить, зная текущую массу [4].

Исходя из вышесказанного, очень важно контролировать текущую массу в барабане центрифуги не только во время загрузки, но и в дальнейших этапах центрифугирования.

Определить текущую массу в барабане при оттеке можно так же, как и при загрузке, решая относительно момента инерции J уравнение (1).

В этом уравнении значения силы тока $I(t)$ и скорости вращения барабана $w(t)$ также снимаются датчиком (через определенные промежутки времени). Момент на валу двигателя M прямо пропорционален силе тока I. Коэффициент пропорциональности известен – это постоянная двигателя KF_n. Найти момент инерции J (а значит, и массу в барабане) можно, применив к уравнению какой-либо численный метод решения дифференциальных уравнений. Но прежде чем определять J, необходимо, как и при загрузке, найти зависимость $M_c = f(t)$ при оттеке. Для этого по опытным данным вычисляются загруженная масса, масса после оттека, темп оттека, а затем и момент инерции J. В уравнении (1) остается единственная неизвестная – M_c. С использованием уравнения (1) строится график ее зависимости от времени. Для получения формулы $M_c = f(t)$

существуют различные методы построения приближенных формул по табличным данным.

Для того, чтобы получить закон изменения момента инерции J во время оттека, определим закон изменения массы. Так как во время оттека темп оттека K_m нельзя принимать постоянным, в отличие от темпа загрузки, закон изменения массы имеет вид:

$$m(t) = m_z + \int_{t_1}^{t_2} K_m(t)dt \qquad (9)$$

где m_z – загруженная масса.

Зависимость темпа напуска от времени при оттеке представляет собой параболу. Действительно, в начале оттека выше скорость, больше центробежные силы, и K_m должен расти. Однако со временем количество межкристальной жидкости иссякает и K_m падает. Коэффициенты параболы можно определить приближенно, зная массу в момент окончания оттека. Масса в момент окончания оттека вычисляется по опытным данным, так же, как и масса после загрузки.

Моменты инерции цилиндра загруженной массы при оттеке и загрузке аналогичны и равны:

$$J(t) = J_0 + m(t)\ R_0{}^2 - \frac{m(t)^2}{2\pi\rho H} \qquad (10)$$

Исключение составляет лишь J_0. При загрузке это момент инерции пустого барабана перед загрузкой, а при оттеке – момент инерции полностью загруженного барабана после загрузки.

Производная момента инерции:

$$\frac{dJ(t)}{dt} = R_0^2 \cdot \frac{dm(t)}{dt} - \frac{1}{\pi\rho H} \cdot \frac{dm(t)}{dt} \qquad (11)$$

где $\frac{dm(t)}{dt}$ находим, дифференцируя (9), с учетом выведенной формулы для темпа напуска.

Теперь можно решить уравнение (1), найти изменение $M_c(t)$ во время оттека, и построить график зависимости $M_c(t)$. Это было сделано с использованием MathCAD 11.

4. Решение задачи оптимизации оттека

Подробно опишем вывод формулы для статического момента и расчет текущей массы на примере одного из опытов. При выводе эмпирической формулы $M_c = f(t)$ для того, чтобы описать зависимость тока и скорости от времени, в п. 2 (при расчете загрузки) вычислялись интерполяционные многочлены. При решении второй задачи, чтобы уменьшить погрешности датчиков, решено было применить методы сглаживания данных – алгоритм бегущих медиан и алгоритм Гауссова ядра [1]. С их помощью были получены кривые, достаточно точно описывающие исходные зависимости.

Производные скорости по времени, присутствующие в уравнении движения электропривода (1), сначала вычислялись с использованием формулы численного дифференцирования по пяти ординатам [2]. Затем полученные значения еще раз сглаживались по методу Гауссова ядра.

Момент инерции во время оттека вычислялся по формуле (10). Перед этим по опытным рассчитывался момент инерции барабана после оттека. Применялся метод, аналогичный методу расчета момента инерции барабана после загрузки. По значению момента инерции барабана после оттека вычислялась масса, находящаяся в этот момент в барабане. Затем определялась зависимость темпа оттека от времени $K_m(t)$, представляющая собой параболу. Коэффициенты параболы подбирались так, чтобы значения загруженной массы и массы после оттека при данном $K_m(t)$ совпадали с вычисленными значениями массы.

Теперь, зная $J(t)$, можно вывести значения $M_c(t) = f(t)$ (или $I_c = f(t)$, так как эти величины связаны соотношением $M_c = KF_n \cdot I_c$) на график.

По точкам графика, используя полиномиальную регрессию, определяем коэффициенты полинома второй степени, который наиболее точно описывает данную кривую. Таким образом, мы получаем искомую зависимость $M_c(t) = P_2(t)$.

Очевидно, что для разных опытов коэффициенты многочлена будут отличаться в той же степени, в которой отличаются реологические свойства загруженной массы. Однако, к счастью, нами уже рассчитан во время загрузки параметр, зависящий от реологических свойств суспензии. Это темп напуска при загрузке K_{mz}. Коэффициенты в функции $Y = \frac{P_2(t)}{K_{mz}}$ уже не будут сильно отличаться от опыта к опыту, и можно вывести некоторый полином $Y_{cp}(t)$ со средними коэффициентами, а эмпирическую зависимость статического момента представлять в виде

$$M_c(t) = K_{mz} Y_{cp}(t).$$

В качестве решения примем регрессионную кривую вида:

$$P_2(t) = K_{mz}(a_{0cp} + a_{1cp}t + a_{2cp}t^2),$$

где коэффициенты определены как среднее арифметическое четырех опытов.

Эмпирически выведенная формула тока статического во время оттека принимает вид:

$$I_c(t) = K_{mz}(0.39 + 0.066t - 3.12 \cdot 10^{-3} \cdot t^2).$$

А для момента статического это будет:

$$M_c(t) = KF_n K_{mz}(0.39 + 0.066t - 3.12 \cdot 10^{-3} \cdot t^2) \quad (12)$$

Для того, чтобы убедиться в достаточной точности выведенных формул статического момента во время загрузки и оттека, необходимо провести проверку для всех четырех опытов и рассчитать ошибки для каждого из них. Для этого в исходное уравнение (1) подставляются отфильтрованные массивы входных данных w_i, I_i, а вместо статического момента M_c значения, вычисленные в точках t_i по выведенным формулам для загрузки и оттека. Производную скорости определяется методом численного дифференцирования с интерполяцией при пяти ординатах. Таким образом, получаем массивы всех переменных, присутствующих в (1). Теперь его можно решить относительно J во всех точках массива t_i. Сделаем это методом Рунге-Кутта четвертого порядка. Полученные точки J_i для всех опытов сравним со значениями J, полученными исходя из формулы (4). Разница между этими значениями будет абсолютной ошибкой метода в данной точке. Ошибки были рассчитаны для всех четырех опытов. Они не превышали 5%. Это означает, что предложенные формулы статического момента во время загрузки и оттека достаточно точны. Их можно использовать при определении текущей массы в барабане центрифуги.

5. Список литературы

[1] Вадутов, О.С. *Математические основы обработки сигналов.* Томск: Изд-во Томского политехнического университета, 2007. – 100 с.

[2] Дьяконов, В.П. *Справочник по расчетам на микрокалькуляторах*. М.: Наука, 1989. – 464 с.

[3] Самарский, А.А. *Введение в численные методы*. М.: Наука, 1986. – 270 с.

[4] Терешин, Б. Н. *Центрифуги в сахарной промышленности.* М.: Пищевая промышленность, 1975. – 600 с.

Разработка и исследование макета отказоустойчивого струйного акселерометра для летательных аппаратов

Коврыгин П.В.
Институт проблем управления РАН, Москва
kovrigin-pavel@ya.ru

Аннотация

Проведены исследования по построению акселерометра на пневматических струйных элементах. Рассмотрены возможные принципиальные схемы реализации последовательного преобразования ускорения в пневматический аналоговый сигнал давления, далее в частотный сигнал и интегрирования его для вычисления сигнала скорости. Проведенные исследования направлены на создание макета акселерометра на носителе энергии неэлектрической природы, на сигналах скорости (расходов) потоков воздуха (газа) и их давлениях. Разработаны макеты струйных акселерометров двух типов: со струйным преобразователем ускорения в частоту и компенсационный с пропорциональным преобразованием ускорения в частоту. По второй схеме макет выполняет операцию интегрирования выходной частоты с преобразованием её в скорость по направлению действия ускорения.

1. Введение

В последнее время в связи с появлением развитых высокотехнологических объектов резко возрастает сложность их систем управления (СУ), к живучести которых при действии дестабилизирующих факторов (ДФ) предъявляются повышенные требования. Действительно, любой сбой в СУ летательных аппаратов (ЛА), атомных станций, океанских лайнеров, пассажирских аэробусов и поездов может привести к непредсказуемым последствиям и катастрофам. А это сопряжено с огромными человеческими и экономическими потерями. Поэтому повышение живучести СУ таких объектов представляется сегодня наиболее актуальным. Одной из существенных причин, влияющих на устойчивость работы СУ, являются внешние воздействия - ДФ, которые могут быть как природными (естественными), так и специально разработанными целенаправленными (искусственными) средствами. В частности, под внешними ДФ будем понимать как естественные (грозовые разряды, радиационные потоки космического пространства, электромагнитные поля и т. п.), так и искусственные (ионизирующие, сверхвысокочастотные (СВЧ) и лазерные излучения, а также пучки нейтральных частиц).

В настоящей работе рассмотрен подход к разработке акселерометра (одного из важнейших приборов СУ), характеристики которого должны помочь позволить реализовать измерение главных параметров полета СУ устойчивой к ДФ.

Избыточные электронные высокопроизводительные СУ ЛА обеспечивают достаточную работоспособность в жестком реальном времени (РВ) при проявлении собственных, распределенных во времени и пространстве, физических неисправностей (отказов и сбоев) отдельных элементов [1]. Для этих систем хорошо развиты и методы выявления проектных ошибок в прикладном и системном математическом обеспечении (МО) [2]. Поэтому они широко используются в качестве основных СУ ЛА. Однако, даже самые изощрённые электронные избыточные структуры [3] не могут предохранить их от внешних ДФ, которые на некоторое время целиком и полностью выводят их из строя, что в жестком РВ означает недопустимое прерывание процесса управления. Следствием этого является гибель ЛА. Для обеспечения непрерывности управления в момент проявления внешнего воздействия необходимо использовать дополнительный канал СУ неэлектрической природы, на работоспособность которого не влияют ДФ.

2. Принципиальная схема

Акселерометр это прибор, измеряющий силу, развиваемую инерционной массой при её движении с ускорением. В зависимости от

способа измерения силы различают акселерометры пружинные и компенсационные.

Выходными сигналами в компенсационных акселерометрах в зависимости от рода энергии, на которой работает измерительный преобразователь, являются сила тока или давление пропорциональные ускорению. Структурная схема таких акселерометров обязательно содержит позиционную отрицательную обратную связь.

Наиболее распространенная технология выполнения струйных устройств - планарная. Струйные элементы и функциональные модули в виде печатных пластин с коммуникационными каналами собираются в многослойные узлы, выполняющие заданные законы управления.

В таких конструкциях параметром, определяющим быстродействие, является линейный размер, поскольку лишь от него зависит время прохождения сигнала по цепи устройства. Поэтому далее будем рассматривать только его влияние на быстродействие струйных элементов.

Критерий динамического подобия для неустановившихся движений жидкости или газа выражается безразмерным числом Струхаля:

$$St = \frac{VT}{L} \quad , \qquad (1)$$

где V— скорость течения газа; T - характерное время (для неустановившихся периодических процессов - период); L - характерный линейный размер.

Если величины V, T и L принимают различные значения, отмечаемые ниже индексами 1, 2, ... n, то для подобных аэродинамических процессов

$$\frac{V_1 T_1}{L_1} = \frac{V_n T_n}{L_n} . \qquad (2)$$

Если средняя скорость течения не меняется (для струйных элементов это будет при постоянном давлении питания), то

$$\frac{T_1}{L_1} = \ldots = \frac{T_n}{L_n} ,$$

$$\frac{L_1}{L_2} = \ldots = \frac{L_{n-1}}{L_n} . \qquad (3)$$

Таким образом, время протекания процесса прямо пропорционально линейному размеру элемента. Отсюда следует, что уменьшение линейных размеров канала питания (проходного сечения) струйного элемента снижает расход рабочей среды пропорционально квадрату размера. Во столько же раз уменьшается и мощность потока питания. Этот показатель чрезвычайно важен для ЛА, так как от него зависят габаритно-массовые характеристики (ГМХ) СУ

$$\frac{G_1}{G_2} = \sqrt{\frac{\rho_1}{\rho_2}} \sqrt{\frac{\Delta p_1}{\Delta p_2}} \quad . \qquad (4)$$

Физические свойства и параметры рабочего газа определяют газодинамические процессы в струйных элементах: давление питания, с которым подводится газ к струйным устройствам; давление окружающей среды, в которую сбрасывается прошедший через струйные устройства газ; сам газ с его плотностью и скоростью распространения в нём звуковой волны.

Рассмотрим влияние указанных параметров рабочей среды на быстродействие струйных элементов. Для простоты сравнения принимаем, что коэффициент расхода рабочей среды из сопла струйного элемента сохраняет постоянное значение при изменении давления питания, плотности рабочей среды и абсолютного давления окружающей среды.

Весовой расход газа через сопло для докритических скоростей истечения может быть подсчитан [3, с. 259] по формуле:

$$G = \varepsilon f g \sqrt{2\rho_0 p_0 \frac{k}{k-1}\left[\left(\frac{p_1}{p_0}\right)^{\frac{2}{k}} - \left(\frac{p_1}{p_0}\right)^{\frac{k+1}{k}}\right]} , \qquad (5)$$

где ε - коэффициент расхода, f- площадь поперечного сечения сопла питания струйного элемента, g - ускорение силы тяжести, p_0 и p_1 - абсолютные давления газа перед и после сопла питания, k - показатель изэнтропы, ρ_0 - плотность газа перед соплом питания.

Так как плотность газа ρ за соплом питания описывается следующим уравнением:

$$\rho = \frac{\rho_0}{gRT} \quad , \qquad (6)$$

где R - газовая постоянная, T - абсолютная температура по шкале Кельвина, выражение (1) можно представить в форме:

$$G = \varepsilon f g \sqrt{\frac{2g}{RT}\frac{k}{k-1}\left[\left(\frac{p_1}{p_0}\right)^{\frac{2}{k}} - \left(\frac{p_1}{p_0}\right)^{\frac{k+1}{k}}\right]} . \qquad (7)$$

Согласно формуле (2), при заданной абсолютной температуре газа весовой расход через сопло питания является функцией абсолютного давления на входе в сопло и отношения к этому давлению абсолютного давления газа за соплом (окружающей среды). Расчеты по формуле (2) проводятся с помощью таблиц газодинамических функций. Однако для сравнительного анализа точного определения расхода не требуется. Достаточно определения качественного характера зависимости

быстродействия от названных параметров в диапазоне соблюдения автомодельности газодинамических процессов. Поэтому можно воспользоваться формулой, предложенной в работе [2], по которой данные расчетов близки полученным по формуле (2) для всего диапазона докритических режимов истечения, если плотность газа считать по давлению p_1

$$G = \varepsilon f g \sqrt{2\rho(p_0 - p_1)} \quad , \qquad (8)$$

где Δp - разность давлений до p_0 и после сопла p_1. Соответственно, для объёмного расхода

$$Q = \varepsilon f g \sqrt{\frac{2\Delta p}{\rho}} \quad . \qquad (9)$$

Для одного и того же сечения сопла при различных условиях истечения соотношения весовых расходов можно представить

$$\frac{G_1}{G_2} = \sqrt{\frac{\rho_1}{\rho_2}} \sqrt{\frac{\Delta p_1}{\Delta p_2}} \quad , \qquad (10)$$

а для объёмных расходов

$$\frac{Q_1}{Q_2} = \sqrt{\frac{\rho_2}{\rho_1}} \sqrt{\frac{\Delta p_1}{\Delta p_2}} \quad . \qquad (11)$$

Для СУ ЛА важнейшие показатели струйных элементов - быстродействие, весовой расход рабочей среды и потребляемая мощность. Они практически и определяют ГМХ СУ ЛА.

Рассмотрим зависимости указанных показателей от давления питания, выбора рабочей среды и давления окружающего пространства, куда она сбрасывается.

Для удобства выявления характера этих зависимостей в струйных элементах введем относительные коэффициенты и сравним с известными, принятыми за исходные.

За исходные принимаем известные значения для струйного активного элемента, имеющего ширину канала сопла питания 0,4 мм, глубину 0,8 мм, работающего на воздухе с избыточным давлением питания 1 кПа при абсолютном давлении окружающей среды 100 кПа. При этих условиях объёмный расход Q = 40 нл/ч. Удельный вес воздуха при этих условиях равен 1,29 кг/м^3, а весовой расход - G = 0,05 кг/ч. Мощность питания элемента с округлением примем равной 10 мВт.

Введем следующие относительные коэффициенты.

Относительный коэффициент быстродействия λV соответствующий отношению скорости истечения через сопло питания при работе на выбранных рабочем газе, давлениях питания рабочей и окружающей сред к скорости истечения через сопло при вышеуказанных исходных условиях; при расчете величины λV отношение скоростей можно заменить отношением объёмных расходов при тех же условиях.

Относительные коэффициенты λG и λN определяют соответствующие отношения весовых расходов и мощностей к тем же исходным значениям, причем для самих этих условий значения указанных коэффициентов равны единице.

По формулам (3)-(6) были рассчитаны величины этих коэффициентов для случаев работы струйного элемента на различных рабочих средах (воздух, гелий и водород), при разных избыточных давлениях питания (0,1; 1,0; 10 и 100 кПа) и абсолютных давлениях окружающей среды (100; 10 и 1 кПа). Данные этих расчетов сведены в таблицу.

Во всех случаях истечение рабочей среды является докритическим. При расчетах считалось, что течение является изотермическим, и принималось, что плотность рабочих сред пропорциональна абсолютному давлению окружающей среды.

3.Результаты экспериментов

Основным фактором, определяющим коэффициенты λV, λG и λN при заданном избыточном давлении питания, является плотность рабочей среды независимо от того, на каком газе работают струйные элементы.

Быстродействие струйных элементов существенно увеличивается за счет снижения плотности рабочего газа и снижения абсолютного давления окружающей среды, при этом имеет место энергетический выигрыш в сравнении с тем, который достигается только повышением давления питания.

Например, чтобы повысить быстродействие элементов при работе на воздухе с давлением питания 1 кПа при нормальных условиях в три раза, необходимо увеличить давление питания в 10 раз (до 10 кПа). При этом потребляемая мощность возрастет в 30 раз. При работе на гелии с тем же давлением питания, быстродействие увеличивается в 8,4 раза, мощность - в 30 раз.

При работе на водороде этот выигрыш ещё заметнее: мощность увеличивается в 30 раз, быстродействие повышается в 12 раз. Но если снизить абсолютное давление окружающей среды со 100 кПа до 10 кПа, получим то же приращение быстродействия, но без увеличения давления питания. При этом мощность увеличится только в 3 раза, а весовой расход уменьшится в 3 раза. При одновременном снижении абсолютного давления с использованием легких газов можно получить быстродействие струйных элементов выше 30 кГц.

Таблица 1. Величины полученных относительных коэффициентов

Рабочая среда	Избыт. давление питания, кПа	Абсолютное давление окружающей среды								
		100кПа			10кПа			1кПа		
		λ_V	λ_G	λ_N	λ_V	λ_G	λ_N	λ_V	λ_G	λ_N
Воздух	0,1 1,0 10 100	0,316 1,0 3,16 10	0,316 1,0 3,16 10	0,316 1,0 31,6 1000	1,0 3,16 10 -	0,1 0,316 1 -	0,1 3,16 100 -	3,16 10 - -	0,0316 0,1 - -	0,316 10 - -
Гелий	0,1 1,0 10	0,84 2,65 8,4	0,12 0,376 1,2	0,084 2,65 84	2,65 8,4 26,5	0,038 0,12 0,38	0,265 8,4 265	8,4 26,5	0,012 0,038 -	0,84 26,5 -
Водород	0,1 1,0 10	1,2 3,76 12	0,084 0,265 0,84	0,12 3,76 120	3,8 12 38	0,026 0,084 0,265	0,38 12 380	0,12 37,6 - -	0,0084 0,0026 -	1,2 37,6 -

Все вышеупомянутые струйные элементы были опробованы на воздухе, гелии и водороде в газогидродинамической лаборатории Института Проблем Управления.

Обсуждаемые данные относятся к струйным элементам с шириной канала питания 0,4 мм.

Существующая сегодня технология изготовления струйных элементов позволяет получать каналы с размерами 0,1 мм. Это увеличит рост быстродействия еще в 4 раза, что позволит достичь быстродействия свыше 100 кГц. Дальнейшее увеличение быстродействия (до 200 кГц) возможно благодаря выполнению размеров каналов питания струйных элементов меньшими 0,1 мм и повышению скоростей истечения гелия до 800 м/с, а водорода - свыше 1000 м/с (скорость звука в гелии 965 м/с, в водороде - 1284 м/с).

Важное обстоятельство - снижение коэффициента λG с переходом на легкие газы и пониженное абсолютное давление. Весовой расход газа при этом снижается во столько раз, во сколько увеличивается быстродействие λV. Этот фактор - важнейший для ЛА при питании от баллонов с газом, так как уменьшается весовой заряд при том же объёмном эквиваленте.

3.Заключение

Инерционный метод можно использовать в интегрирующих акселерометрах, выходной сигнал которых пропорционален интегралу от линейного ускорения за некоторое время [1].

Угол поворота магнита пропорционален интегралу по времени от измеряемого ускорения [2]

$$\varphi = \int_0^t \dot{\varphi} dt = \frac{ml}{k} \int \ddot{x} dt \qquad (12)$$

Это позволяет получить непосредственно путевую скорость по действительному ускорению с момента начала движения [3].

Разработана схема струйного акселерометра, изготовлен макет с использованием струйных элементов СМСТ-2 и ВОЛГА и проведены его экспериментальные исследования.

Исследования показали оригинальность, и работоспособность макета в пределах значений измеряемых ускорений по выбранной принципиальной схеме, а также что для точного измерения усилий может быть использован частотный датчик с колебательным элементом в виде струны с возбуждением и считыванием колебаний с помощью газоструйных технических средств.

4. Литература

[1] М.М БЕЛЯЕВ, А.А. ХИТРОВО *"Способ возбуждения и измерения частоты колебаний колебательного звена резонансных датчиков"*, Патент РФ № 2269385.

[2] Д.С. БРАСЛАВСКИЙ *"Приборы и датчики летательных аппаратов"*, Машиностроение, М., 1970, с. 354 -500.

[3] R.S.SIDELL, P.P. RIGHARGSON *"A stepbearing accelerometer with pneumatic signal output"*, Paper. Amer. Soc. Mech. Eng., № LubS-40, 1968.

Методика конкурентного анализа телекоммуникационных компаний

И.И. Сафарли

Институт кибернетики НАН Азербайджана, Баку
e–mail: isafarli@yahoo.ca

Аннотация

Предлагается методика конкурентного анализа компаний – поставщиков телекоммуникационных продуктов. Методика может быть использована при стратегическом планировании деятельности компаний. В отличие от традиционных методик конкурентного анализа, используемых в задачах стратегического планирования предприятий, предлагаемая методика учитывает специфику рынка телекоммуникационных продуктов и фактор глобализации телекоммуникационного бизнеса.

1. Введение

В условиях глобализации телекоммуникационного бизнеса поиск устойчивого конкурентного преимущества компаний, работающих в этом секторе экономики, требует принципиально новых методов конкурентного анализа [1, с. 115]. Традиционные подходы, предлагаемые классической теорией конкурентного анализа [2, с. 97; 3, с.53] и ориентирующие менеджеров на изучение целого комплекса параметров компании (качество продуктов компании, репутация компании, компетентность управленческого звена, производственные характеристики, технологический потенциал, маркетинг и реклама, сбытовая сеть, уровень общих издержек и т.д.), оказываются малоэффективными. В специфических условиях телекоммуникационного бизнеса наиболее приемлемым инструментом для анализа конкурентных позиций компании становятся показатели, характеризующие потребительские свойства самого продукта, важные с точки зрения потенциальных потребителей [4, с. 71].

Эти показатели обладают двумя важными особенностями. Во-первых, они объективны, так как отражают реальные запросы потребителей безотносительно к компаниям, предлагающим продукт. Во-вторых, они значительно более доступны для изучения, чем внутренние факторы конкурентного преимущества, используемые в традиционных методиках конкурентного анализа.

2. Постановка задачи

В соответствии с этим рассматриваемая в статье методика конкурентного анализа была направлена на решение трех вопросов:

1. Построение модели продукта, как объекта конкурентного анализа.
2. Разработка на основе модели некоторого общего показателя конкурентоспособности продукта, отражающего структуру потребительских запросов.
3. Анализ аналогичных продуктов, предлагаемых конкурентами, и разработка по результатам анализа некоторого «норматива конкурентоспособности» компании. Как известно, существует традиционный способ решения данного вопроса. Это – сравнение с *отраслевыми нормативами.* Такое сравнение обычно полезно при определении конкурентной позиции компании. Однако в условиях высокой динамики телекоммуникационного бизнеса существует риск упустить из внимания то, что вся отрасль может развиваться неудовлетворительно и в течение срока анализа она может подвергнуться атаке зарубежных, более успешных конкурентов.

Очевидно, что норматив конкурентоспособности не должен побуждать компанию производить продукт с максимально высоким качеством. Он должен в ясной и доступной форме определять интегральное качество продукта, которое, будучи лучшим, чем у конкурентов, может гарантировать компании долгосрочное конкурентное преимущество.

3. Методика конкурентного анализа

Предлагаемая методика является достаточно универсальной и может быть использована в качестве инструмента первичного конкурентного анализа компании. В этом качестве методика позволяет получить общее представление о конкурентоспособности компании относительно ее потенциальных конкурентов.

В соответствии с поставленной целью в методике предусмотрена следующая последовательность шагов.

Шаг 1. Уточнение сегмента рынка, на котором работает или предполагает работать компания. Производится по известным методикам стратегической сегментации рынка [5, с. 75–85].

Шаг 2. Установление продукта, который должен рассматриваться как объект конкурентного анализа.

Шаг 3. Выявление перечня потребительских свойств продукта важных для потенциальных потребителей.

В качестве базовой используется 6–позиционная «модель мнений потребителя» фирмы IBM Rochester [6]. Адаптация модели к конкретному продукту производится путем опроса ведущих сотрудников компании и/или потенциальных пользователей.

Шаг 4. Построение сводного показателя конкурентоспособности продукта компании, а также продуктов, предлагаемых компаниями–конкурентами.

Известны математические подходы к решению задачи построения сводного показателя конкурентоспособности. Однако они требуют точной количественной информации, которая на ранних стадиях стратегического планирования, как правило, отсутствует. Для такого рода случаев теория принятия решений рекомендует достаточно простой и общедоступный «метод взвешенных сумм», широко используемый в управленческой практике [7, с. 81–83,127–129]. Сводный показатель в данном методе рассчитывается по формуле

$$U = \sum_{i=1}^{N} w_i x_i \ , (I = 1, 2, 3, \ldots, p),$$

где w_i – положительные или отрицательные весовые коэффициенты потребительских свойств; положительные при тех свойствах x_i, которые желательно максимизировать; отрицательные – при тех, которые желательно минимизировать; $\sum_{i=1}^{N} w_i = 1$; x_i – оценка продукта по i-му потребительскому свойству.

Логическим обоснованием метода взвешенной суммы является представление об общей конкурентоспособности продукта как о сумме оценок нескольких независимых его потребительских свойств. Коэффициенты w_i выражают относительную важность свойств.

Метод позволяет учитывать и оценки отдельных потребительских свойств и оценки их весовых коэффициентов, отражающих их вклад в обеспечение конкурентоспособности компаний. При практическом применении следует иметь в виду два важных момента. Первый – данный метод оценки сводного показателя корректен, когда свойства носят независимый характер. Второй – метод позволяет установить только порядок предпочтения при сравнении продуктов, но не абсолютную оценку конкурентоспособности продукта.

Шаг 5. Ранжирование потребительских свойств продукта, отражающее относительную важность отдельных свойств с точки зрения потребителей.

Оценка относительной важности, т.е. весовых коэффициентов w_i, потребительских свойств, может быть выполнена с помощью методов экспертных оценок [8, с. 154]. Однако на практике при большом числе свойств эксперты затрудняются решать эту задачу на множестве предъявляемых свойств. Психологически им проще решать задачу путем парных сравнений этих свойств. Такой способ оценки отражает естественный ход человеческих рассуждений и положен в основу «метода анализа иерархий» Т.Саати [9, с. 123].

Вопрос ранжирования потребительских свойств с помощью этого метода рассмотрен нами в работе [10].

Шаг 6. Выбор шкал для оценки потребительских свойств продукта.

Вопрос выбора шкал для оценки потребительских свойств носит нетривиальный характер, так как свойства продукта имеют различный физический смысл, различную размерность и различный диапазон изменения. Поэтому желательно представить их в каком–то безразмерном (относительном) виде. В [11] показано, что в случае использования для оценки обобщенного показателя метода взвешенной суммы допустимо применение так

называемых «шкал разностей», определенных с точностью до сдвига. Таким требованиям отвечают балльные оценки, основанные на понятии едва заметного различия. В нашем случае используется 10–ти балльная оценка свойств.

Шаг 7. Выявление потенциальных компаний–конкурентов. Изучение и балльная оценка значений потребительских свойств продуктов, поставляемых компанией и предлагаемых конкурентами.

Шаг 8. Построение конкурентного профиля продукта компании и продуктов, предлагаемых компаниями–конкурентами.

Конкурентный профиль строится в графической форме, наглядно отражающей результаты конкурентного анализа. Конкурентный профиль может быть представлен в виде лепестковой диаграммы или столбиковой гистограммы, отражающей полученные оценки в развернутом виде.

Шаг 9. Анализ конкурентных профилей продуктов компании и продуктов ее конкурентов. Установление «норматива конкурентоспособности» компании на установленный срок конкурентного анализа.

При построении конкурентного профиля продуктов компаний–конкурентов могут быть использованы следующие способы:

(а) составление «досье» на каждого потенциального конкурента и

(б) использование элементов бенчмаркинга (сравнение с конкурентами по контрольным узлам) [12, с. 47]. При этом информация о конкурентах может быть получена из источников, перечисленных в табл. 1.

Таблица 1. Источники информации о конкурентах

Первичная информация (необработанные данные)	Вторичная информация (данные, прошедшие обработку, цели которой не всегда совпадают с целями анализа)
Каналы распределения и сбытовая сеть; Поставщики; Потребители; Рекламные агентства, торговые агенты; Торговые и профессиональные ассоциации; Персонал фирмы-конкурента; Специальные аналитические службы, промышленный шпионаж и т.д.	Бухгалтерская отчетность; Аналитические материалы в прессе; Справочные издания о рынке; Нормативно-правовые акты, затрагивающие деятельность конкурента; Рекламная компания конкурента.

Необходимо еще раз отметить, что ни определенные экспертами балльные оценки свойств, ни общий показатель конкурентоспособности продукта не являются числами. Это – порядковые величины, поэтому, например, по поводу оценок конкурентоспособности «4» и «2» продуктов двух различных компаний нельзя сказать, что конкурентоспособность продукта у одной из компании в два раза меньше, чем у другой.

4. Пример

Ниже в качестве примера приводятся результаты конкурентного анализа, выполненного с помощью предлагаемой методики для компании, поставляющей системы CTi-телефонии [13].

Компьютерная телефония (Computer& Telephony Integration) является перспективной ИКТ–технологией, которая дает возможность создавать различные продукты как для корпоративных клиентов, так и для операторов связи: информационные системы, голосовая почта, универсальная почта, системы удаленного голосования, телелотереи, автоматизации выставления счетов за пользование каналами связи, автосекретарь и др. Продуктом является программно–аппаратный комплекс, конфигурация которого формируется в соответствие с требованиями клиента из существующих функциональных блоков. Комплекс сдается клиенту «под ключ», либо клиенту оказываются услуги по установке (автономно или в интеграции с существующими системами). После установки клиентам оказываются услуги по послепродажному обслуживанию в течение гарантийного срока.

В настоящее время возможны два варианта реализации систем CTi–телефонии: аппаратная (АР) и программно–аппаратная (ПАР). Аппаратная реализация

осуществляется на базе комплектующих, поставляемых такими компаниями, как AT&T, Ericsson, Motorola, др. Системы этого типа имеют высокую стоимость, которая в несколько раз, иногда на несколько порядков, превышает стоимость программно-аппаратных систем. Программно–аппаратные системы реализуются в форме PC–совместимого компьютера, платы компьютерной телефонии Dialogic (компании Dialogic Corp., USA) и программного обеспечения. Стоимость их значительно ниже, меньше и сроки внедрения. Однако, вследствие использования нестандартного программного обеспечения, повышенные требования предъявляются к их отказоустойчивости, особенно на участках операторов связи по предоставлению платных услуг.

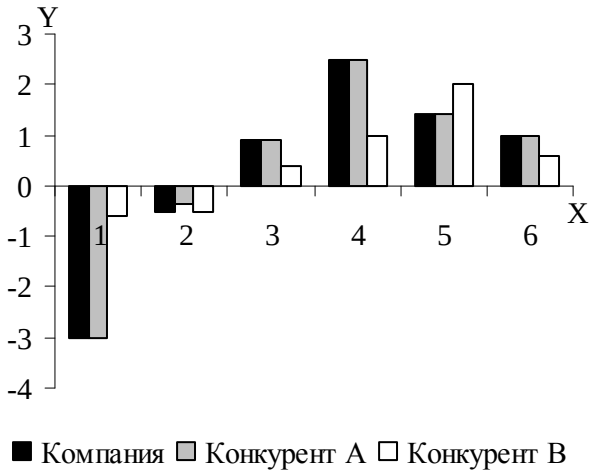
Таким образом, ряд важных потребительских свойств УИС, построенных на основе систем компьютерной телефонии, оказываются зависящими от принятого варианта реализации и от внутренних возможностей компании.

При выявлении потенциальных конкурентов в качестве последних рассматривались компании А и Б, способные использовать технологии мировых лидеров в области компьютерной телефонии и обладающие всеми необходимыми ресурсами и бизнес-процессами для эффективной конкуренции на рынке.

В табл.2 даны результаты конкурентного анализа компаний. На рис.1 представлены конкурентные профили компаний.

Таблица 2. Результаты конкурентного анализа компаний

Потребительские свойства УИС CTi-телефонии	Вес свойства	Значения свойств по 10-ти бальной шкале		
		Компания (АР)	Конкурент А (АР)	Конкурент В (ПАР)
1.Цена (–)	0,30	10	10	2
2.Срок внедрения (–)	0,05	10	7	10
3.Простота установки (+)	0,10	9	9	4
4.Отказоустойчивость (+)	0,25	10	10	4
5.Масштабируемость (+)	0,20	7	7	10
6.Сертификация продукта (+)	0,10	10	10	6
Общая оценка конкурентоспособности	1,00	**2,3**	**2,45**	**2,9**



X – свойства УИС системы, Y – взвешенные оценки свойств.

Рис. Конкурентные профили компании и ее конкурентов А и В.

Заключение

Полученные общие оценки конкурентоспособности и конкурентные профили компаний дают исходную базу для выбора «норматива конкурентоспособности» телекоммуникационной компании по поставляемым ею системам CPi-телефонии. Имея общую оценку конкурентоспособности по данному продукту, оценки весов и значений потребительских свойств продукта, руководство компании получает ориентиры для выбора наиболее приемлемых направлений улучшения конкурентоспо–собности продукта, исходя из имеющихся у компании возможностей. Если в какой–то области вырисовывается возможная проблема, то руководство компании может создать рабочую группу для углубленного анализа проблемы. При этом могут привлекаться все доступные методы контроля и управления конкурентоспособностью телекоммуникаци–онного продукта (дополнительные опросы, контрольные листки, диаграммы Исикавы, Парето–анализ, матрицы коэффициентов удовлетворения потребителей, методы TQM (Total Quality Management)). В конечном итоге методика конкурентного анализа и выбора «норматива конкурентоспособности» помогают компании решать главный вопрос сегодняшнего телекоммуникационного бизнеса: не производить и обслуживать свои продукты, а удовлетворять потребителя лучше, чем компании–конкуренты. Следовательно – использовать удовлетворение потребителя для развития бизнеса.

Литература

[1] C.K. Prahalad, G. Hamel. The core competence of the corporation. *Harvard business review*. 1990, № 5 P.79-91 (рус. перевод: Прахалад К.К., Хамел Г. Стержневые компетенции корпорации. В кн.: Стратегический процесс. Под ред. Г. Минцберг, Дж. Б. Куинн, С. Гошал «Питер», СПб, 2001. С. 112-124).

[2] М. Портер. *Международная конкуренция*. Пер. с англ. Под ред. В. Д. Щетинина. Международные отношения, Москва, 1993, 567 с.

[3] Г.Л. Азоев. *Конкуренция: Анализ, стратегия и практика*. ИНФОТЕСТ, Москва, 1996, 296 с.

[4] А..М. Аббасов, Р.А. Караев, И.И. Сафарли. Метод идентификации ключевых управленческих компетенций телекоммуникационных компаний. *Известия НАН Азербайджана. Серия матем. и физико–техн. наук*. № 3. 2006, с. 15–19.

[5] И. Ансофф. *Стратегическое управление*. Экономика, Москва, 1978, 520 с.

[6] С. Джордж, А. Ваймерскирх. *Всеобщее управление качеством: стратегии и технологии, применяемые сегодня в самых успешных компаниях. (TQM)*. «Виктория плюс», СПб, 2002. 256 с.

[7] Щ.И. Ларичев. *Наука и искусство принятия решений*. Наука, Москва, 1979. 254 с.

[8] И.М. Макаров, Т.М. Виноградская, А.А. Рубчинский, В.Б. Соколов. *Теория выбора и принятия решений*. Наука, Москва, 1982, 328 с.

[9] Т. Саати. *Принятие решений. Метод анализа иерархий*. Радио и связь, Москва, 1993, 320 с.

[10] И.И. Сафарли. Применение метода анализа иерархий для ранжирования потребительских свойств продуктов компаний-поставщиков телекоммуникационных услуг. *Материалы международной конференции «Проблемы кибернетики и информатики», PCI'2006*. том 3. Баку. 2006, С.115-118.

[11] Ю.В. Любатов. О средних взвешенных оценках качества. *Стандарты и качество*. 1972, с. 48–50.

[12] Н.Н. Бек. *Методы стратегического анализа. Материалы к лекциям и практическим занятиям*. Часть 3. М.: Государственный Университет – Высшая Школа Экономики, 2003. 342 с.

[13] Архив: *Компьютерная телефония* http://osp.aanet.ru/nets/archive/1.htm

Допредельная реализация метода иерархии управлений в системе управления двухзвенным манипулятором

Уткин А.В.
Институт проблем управления им В.А. Трапезникова РАН, Москва
autkin@ipu.rssi.ru

Введение

Рассматривается задача прямого управления конечной точкой двухзвенного манипулятора. Разработан метод допредельной иерархии управлений, позволяющий решить задачу слежения за заданными траекториями конечной точки с заданной точностью, в условиях неопределённости входных каналов управления.

1. Постановка задачи

Рассматривается система управления двухзвенным плоскостным манипулятором с двумя степенями свободы, представленная следующей математической моделью

$$\begin{aligned} &\dot q_1 = q_2,\\ &\dot q_2 = H^{-1}(q_1)[\upsilon - C(q_1,q_2)q_2 - G(q_1) + \eta(t)], \end{aligned} \quad (1)$$

$q_1 \in Q_1 \subset R^2$ – вектор угловых положений, $q_2 \in Q_2 \subset R^2$ – вектор угловых скоростей; $H_{2\times 2}(q_1) = (H_{ij}) > 0$ – нелинейная симметрическая матрица инерции, $H^{-1}(q_1) = (H_{ij}^{-1}) > 0$; $C_{2\times 2}(q_1,q_2)$ – матрица центростремительных и кориолисовых сил; $G(q_1)$ – вектор гравитационных сил; вектор $\eta(t)$ – внешние неконтролируемые возмущения; $\upsilon \in R^2$ – моменты, развиваемые исполнительными устройствами.

Составляющие оператора системы (1) имеют вид [1]

$$H_{11} = m_1 l_{c1}^2 + I_1 + m_2(l_1^2 + l_{c1}^2 + 2l_1 l_{c2}\cos q_{12}) + I_2$$
$$H_{22} = m_2 l_{c2}^2 + I_2,$$
$$H_{12} = H_{21} = m_2 l_1 l_{c2}\cos q_{12} + m_2 l_{c2}^2 + I_2;$$

$$C_{11} = -m_2 l_1 l_{c2}\sin q_{12} q_{22},$$
$$C_{12} = -m_2 l_1 l_{c2}\sin q_{12} q_{22} - m_2 l_1 l_{c2}\sin q_{12} q_{21},$$

$$C_{21} = m_2 l_1 l_{c2}\sin q_{12} q_{21}, C_{22} = 0,$$

$$G = \mathrm{col}(G_1, G_2),$$

$$G_1 = m_1 l_{c1}^2 g\cos q_{11} + m_2 g[l_{c2}\cos(q_{11}+q_{12}) + l_1\cos q_{11}],\ G_2 = m_2 l_{c2} g\cos(q_{11}+q_{12}),$$

где I_1, I_2 – приведенные моменты инерции, m_1, m_2 – массы звеньев, l_1, l_2 – длины звеньев, l_{c1}, l_{c2} – расстояния от начала до центров тяжести звеньев.

Положение конечной точки манипулятора определяется точкой $P(y_{11}, y_{12})$ на плоскости, координаты которой однозначно выражаются через угловые положения звеньев:

$$\begin{aligned} y_{11} &= h_1(q_1) = l_1\cos q_{11} + l_2\cos(q_{11}+q_{12}),\\ y_{12} &= h_2(q_2) = l_1\sin q_{11} + l_2\sin(q_{11}+q_{12}), \end{aligned} \quad (2)$$

где $J_{2\times 2} = \{\partial h/\partial q_1\}$, $\det J = l_1 l_2 \sin q_{12} \neq 0$ при $q_{12}^* \neq \pm k\pi$ – вырожденные конфигурации, соответствующие вытянутой или сложенной руке манипулятора.

Ставится задача синтеза управления движением конечной точки в предположении, что для измерения доступны только её координаты $y_1(t) = col(y_{11}, y_{12})$. Задача сводится к стабилизации невязок между фактическими и заданными траекториями движения конечной точки $e_{1i} = y_{1i} - y_{1id}$, $i = \overline{1,2}$.

Как правило, решение поставленной задачи требует решения в реальном времени обратных задач динамики. В данной работе иной подход [2], разработанный для общего вида модели робота-манипулятора, позволяющий избежать обратных преобразований, применён для конкретной модели двухзвенного манипулятора с определёнными параметрами.

2. Блочное представление

Решение задачи управления конечной точкой

манипулятора возможно на основе выходного отображения, заключающегося в последовательном дифференцировании вектора выходных переменных (2) в силу исходной модели (1). С учётом ряда упрощающих переобозначений выходное отображение исходной системы можно записать в следующем каноническом виде

$$\dot{y}_{11} = y_{21}, \dot{y}_{12} = y_{22},$$
$$\dot{y}_{21} = f_1 + b_{11}\upsilon_1 + b_{12}\upsilon_2,$$
$$\dot{y}_{22} = f_2 + b_{21}\upsilon_1 + b_{22}\upsilon_2,$$

где

$$f_1 = y_{12}\frac{1}{\Delta}(H_{22}G_1^* - H_{21}G_1^* - H_{12}G_2^* + H_{11}G_2^*) -$$
$$- y_{22}(\dot{q}_1 + \dot{q}_2) + l_1(\cos q_1\dot{q}_1\dot{q}_2 + \sin q_1\frac{1}{\Delta}H_{21}G_1^* +$$
$$+ \sin q_1\frac{1}{\Delta}H_{11}G_2^*),$$
$$f_2 = y_{11}\frac{1}{\Delta}(H_{21}G_1^* - H_{22}G_1^* + H_{12}G_2^* - H_{11}G_2^*) +$$
$$+ y_{21}(\dot{q}_1 + \dot{q}_2) + l_1(\sin q_1\dot{q}_1\dot{q}_2 + \cos q_1\frac{1}{\Delta}H_{11}G_2^* -$$
$$- \cos q_1\frac{1}{\Delta}H_{21}G_1^*),$$
$$\Delta = H_{11}H_{22} - H_{21}H_{12},$$
$$G_1^* = (c_{11}\dot{q}_1 + c_{12}\dot{q}_2) + G_1(q) - \eta_1,$$
$$G_2^* = (c_{21}\dot{q}_1 + c_{22}\dot{q}_2) + G_2(q) - \eta_2,$$
$$b_{11} = y_{12}\frac{1}{\Delta}(H_{21} - H_{22}) - l_1\sin q_1\frac{1}{\Delta}H_{21},$$
$$b_{12} = y_{12}\frac{1}{\Delta}(H_{12} - H_{11}) + l_1\sin q_1\frac{1}{\Delta}H_{11}$$
$$b_{21} = y_{11}\frac{1}{\Delta}(H_{22} - H_{21}) + l_1\cos q_1\frac{1}{\Delta}H_{21},$$
$$b_{22} = y_{11}\frac{1}{\Delta}(H_{11} - H_{12}) - l_1\cos q_1\frac{1}{\Delta}H_{11}.$$

В рамках решения поставленной задачи слежения для выходного отображения требуется обеспечить стабилизацию невязок $e_{1i} = y_{1i} - y_{1id} \to 0, i = \overline{1,2}$, поведение которых описывается следующими уравнениями $\dot{e}_{1i} = \dot{y}_{1i} - \dot{y}_{1id} = y_{2i} - \dot{y}_{1id}$, в которых y_{2i} трактуется как фиктивное управление, размерность которого совпадает с размерностью регулируемой переменной e_{1i}.

На первом шаге выберем фиктивное управление в виде $y_{2i} = e_{2i} - k_{1i}e_{1i} + \dot{y}_{1id}$. Далее для обеспечения выбранного управления требуется решить задачу стабилизации невязки $e_{2i} = y_{2i} + k_{1i}e_{1i} - \dot{y}_{1id}$.

Дифференциальное уравнение относительно невязки e_{1i} становится первым блоком: $\dot{e}_{1i} = -k_{1i}e_{1i} + e_{2i}$.

На втором шаге составим дифференциальное уравнение невязок $\dot{e}_{2i} = \dot{y}_{2i} + k_{1i}\dot{e}_{1i} - \ddot{y}_{1id}$ и с учетом результатов первого шага и полученного раннее выходного отображения получим следующее блочное представление в невязках

$$\dot{e}_{11} = -k_{11}e_{11} + e_{21}, \quad \dot{e}_{12} = -k_{12}e_{12} + e_{22},$$
$$\dot{e}_{21} = -k_{11}^2e_{11} + k_{11}e_{21} + \varphi_1 + b_{11}\upsilon_1 + b_{12}\upsilon_2,$$
$$\dot{e}_{22} = -k_{12}^2e_{12} + k_{12}e_{22} + \varphi_2 + b_{21}\upsilon_1 + b_{22}\upsilon_2,$$

(3)

где $\varphi_1 = f_1 - \ddot{y}_{11d}$, $\varphi_2 = f_2 - \ddot{y}_{12d}$.

В системе (3) для реализации стабилизирующей обратной связи требуются вычисления в реальном времени матрицы перед управляющими воздействиями, зависящей от угловых положений звеньев маниулятора. В условиях же отсутствия информации об угловых положениях звеньев, т.е. в условиях неопределённости входных каналов управления для системы (3) разработан метод допредельной иерархии управлений, как развитие известного метода иерархии для систем с разрывными управлениями [2].

2. Метод допредельной иерархии управлений

В данной работе для системы (3) предлагается использовать обратную связь в виде непрерывной функции $\upsilon_i = -arctg(k_{2i}e_{2i})$. Подобный вид обратной связи является допредельной реализацией разрывных управлений, которые часто применяются при управлении манипуляторами с учётом динамики моделей исполнительных устройств [4]. Применительно к системе (1) разрывная обратная связь технически не реализуема, ввиду того, что механические моменты υ на валах электродвигателей физически не могут подлежать разрыву с большой частотой. Допредельная реализация разрывных управлений (4) в виде непрерывной функции, предложенная в данной работе для решения задачи слежения с заданной точностью имеет ряд преимуществ:

1) ограниченность $\left|\mathrm{arctg}(k_{2i}e_{2i})\right| < \frac{\pi}{2}$,;

2) эквивалентность линейной функции (рис.1): $\mathrm{arctg}(k_{2i}e_{2i}) \underset{e_{2i}\to 0}{\sim} k_{2i}e_{2i}$;

3) допредельная гладкая реализация разрывной

функции (рис.2):

$$\operatorname{arctg}(k_{2i}e_{2i}) \underset{k_{2i}\to\infty}{\to} \tfrac{\pi}{2}\operatorname{sign} e_{2i}$$

(*)

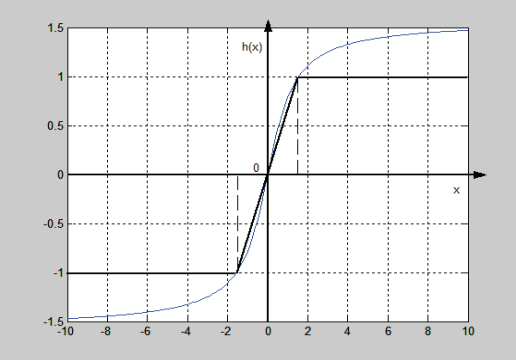


Рис. 1

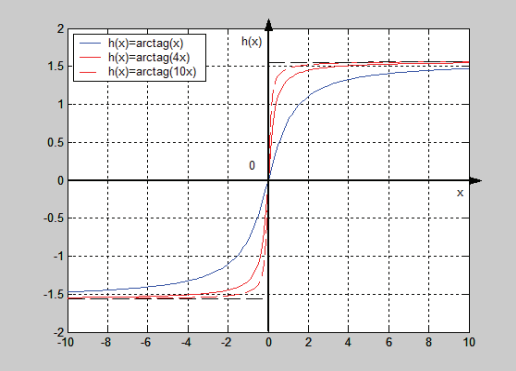


Рис. 2

Метод иерархии управлений заключается в установлении иерархии, которая определяет очерёдность возникновения скользящих режимов на соответствующих поверхностях скольжения, что позволяет привести матрицу перед управлениями к верхнему треугольному виду с преобладающей диагональю, а затем использовать обратную связь в виде непрерывной функции.

В блочном представлении (3) в силу предположений $\operatorname{sign} b_{11} = \text{var}$, $\operatorname{sign} b_{12} = \text{const}$ задаётся следующая иерархия управлений $\upsilon_2 \to |e_{21}| \le \delta_{21} \to \upsilon_1 \to |e_{22}| \le \delta_{22}$. Исследование оператора объекта управления позволило получить диапазоны изменения элементов матрицы перед управлениями $B_{(2\times 2)}(q_1)$:

$-0{,}08 \le b_{11} \le 0{,}1$, $-1{,}2 \le b_{12} \le -1{,}1$,

$0{,}37 \le b_{21} \le 0{,}44$, $-0{,}37 \le \ \le b_{22} \le -0{,}2$.

Процедура допредельной реализации метода иерархии управлений и настройки коэффициентов $M_{(2\times 2)} = \operatorname{diag}(m_i)$, для системы (3) состоит из двух шагов.

Шаг 1. В третьем уравнении системы (3) выберем управление в виде

$$\upsilon_2 = -m_2 \operatorname{sign} b_{12} \operatorname{arctg}(k_2 e_{21}),$$

(4)

Для системы (3) составим функцию Ляпунова вида $V = V_1 + V_2 = (\frac{1}{2}e_{11}^2 + \frac{1}{2}e_{12}^2) + (\frac{1}{2}e_{21}^2 + \frac{1}{2}e_{22}^2)$

Далее подставив в (3) выбранное управление вида (4) имеем нижнюю оценку для выбора коэффициента обратной связи

$$\dot V_{21} = e_{21}\dot e_{21} < 0 \Rightarrow$$

$$(2k_1^2\delta_1 + |\varphi_1| + \overline{\overline{b}}_{11}\overline{\overline{\upsilon}}_1)/(\overline{b}_{12}(\tfrac{\pi}{2} - \varsigma)) < m_2$$

(5)

Для регуляризации задачи введено новое управление $\tilde\upsilon_2$. Допредельная реализация метода эквивалентного управления: в асимптотике $k_2 \to +\infty$ выполняются соотношение (*), $\dot e_{21} \to 0$ и $\tilde\upsilon_2 \to \upsilon_{2\text{eq}}$. Из уравнения статики $\dot e_{21} \to 0$, которое выполняется с точностью до б.м., найдем значение эквивалентного управления $\tilde\upsilon_2 = (k_1^2 e_{11} - \ -k_1 e_{21} - \varphi_1 - b_{11}\upsilon_1)/b_{12}$.

Шаг2. Подставляя полученное значение эквивалентного управления в последнее уравнение системы (3) получим

$$\dot e_{22} = -k_1^2 e_{12} + k_1 e_{22} + \varphi_2 + b_{21}\upsilon_1 + b_{22}\tilde\upsilon_2 =$$

$$= k_1^2(e_{11}b - e_{12}) + k_1(e_{22} - e_{21}b) + (\varphi_2 - \varphi_1 b) + \upsilon_1 b_{21}^1,$$

где $b_{21}^1 = b_{21} - b_{11}b_{22}/b_{12}$, $b = b_{22}/b_{12}$,

$\overline{b}_{21}^1 = \overline{b}_{21} - \overline{\overline{b}}\,\overline{\overline{b}}_{11} = 0{,}34$, $\overline{\overline{b}} = \overline{\overline{b}}_{22}/\overline{b}_{12} = 0{,}36$.

Выберем управления во второй системе (3) в виде $\upsilon_1 = -m_1 \operatorname{sign} b_{21}^1 \operatorname{arctg}(k_2 e_{22})$ и получим нижнюю оценку для выбора коэффициента обратной связи

$$\dot V_{22} = e_{22}\dot e_{22} < 0 \Rightarrow (2k_1^2\delta_1(1 + \overline{\overline{b}}) + \quad (6)$$

$$+ |\varphi_2 - \varphi_1 b|)/\overline{b}_{21}^1(\tfrac{\pi}{2} - \varsigma) \le 63.1 \le m_1$$

Подставив полученную оценку (6) в неравенство (5), имеем $m_2 \ge 146$. Выбранные на основе полученных оценок параметры обеспечивают следующую последовательность сходимости невязок $|e_{21}| \le \delta_2 \Rightarrow |e_{22}| \le \delta_2 \ \Rightarrow |e_{1i}| \le \delta_1$. Существенно, что анализ разрешимости задачи и указанные построения проводятся на этапе исследования, что позволяет сократить объём вычислений, выполняемых в реальном времени.

На рисунках показаны результаты моделирования с законом управления (4).

На рис. 3-4 проиллюстрированы ошибки слежения за заданными положениями конечной точки. На рис. 5-6 – выход и движение конечной точки по заданным траториям: окружности и прямой соответственно.

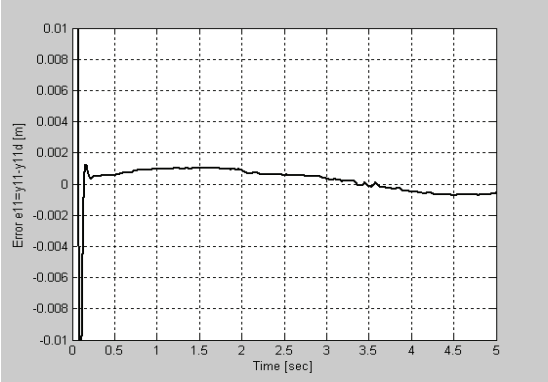


Рис.3

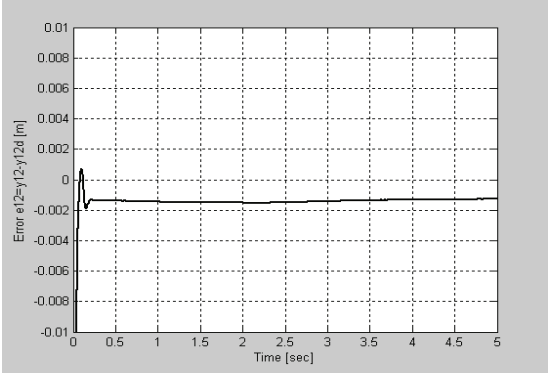


Рис. 4

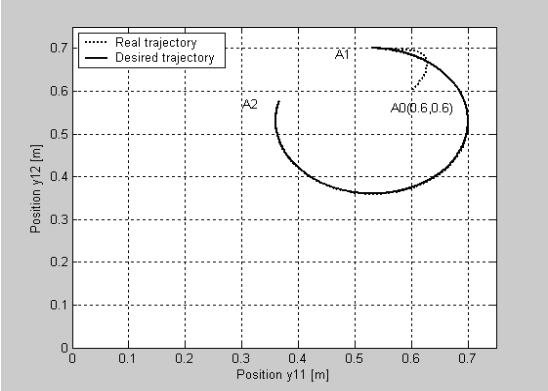


Рис.5

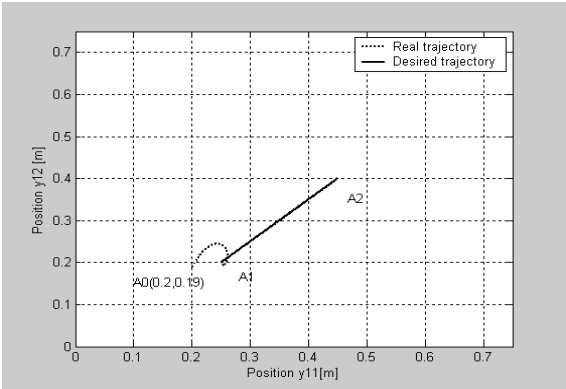


Рис.6

Заключение

В работе рассмотрены методы синтеза задачи управления движением конечной точки двухзвенного робота манипулятора при условии наличия информации только о выходных координатах. Предложено решение задачи слежения в условии параметрической неопределённости входных каналов управления с использованием метода допредельной иерархии управлений. Приведены результаты моделирования, подтверждающие работоспособность предложенных алгоритмов.

Список литературы

[1] Колесников А.А., “Современна прикладная теория управления ”, ТРТУ, 2000.

[2] Краснова С.А., Уткин В.А., Уткин А.В., Нгуен Тхань Тиен, “Прямой метод синтеза системы управления рабочим органом манипулятора при неполных измерениях”, *Проблемы управления*, №1, 2008, С.10–18.

[3] Краснова С.А., “Каскадный синтез системы управления манипулятором с учетом динамики электроприводов”, *AuT*, №11, 2001,С.51–72.

[4] Уткин В.И., “Скользящие режимы в задачах оптимизации и управления”, Наука, Москва, 1987.

Введение платы за участие в механизмы активной экспертизы

Бурков В.Н.,
Коргин Н.А.
ИПУ РАН
nkorgin@ipu.ru

Аннотация

Исследуется возможность введения трнсферов в механизм активной экспертизы для улучшения эффективности. Неманипулируемый механизм активной экспертизы с с платой за участие строится с помощью методов mechanism design – теоремы Гровса-Грина-Лафона. Приводится альтернативный метод построения, позволяющий существенно упростит алгоритм получения результатов.

1. Введение

Большое количество управленческих решений в современных условиях принимаются на основании выводов экспертизы. Нередко складывается ситуация, когда привлекаемые эксперты являются лицами, заинтересованными в результате, которые формируется на основании экспертизы. Что создает предпосылки для оппортунистического поведения со стороны экспертов – искажения собственных заявок с целью влияния на результат экспертизы в собственных интересах. В теории организационных систем подобные задачи называются задачами активной экспертизы [1]. Механизмы активной экспертизы (правила определения итоговой оценки экспертов на основании их заявок) называются неманипулируемыми, если они побуждают экспертов сообщать истинные значения собственных оценок. В «традиционных» постановках задачи экспертизы взаимодействие организатора экспертизы (центра в терминологии теории управления организационными системами) с экспертами ограничивается лишь сбором заявок от экспертов, на основании которых определяется результат экспертизы [1]. Интересными представляется анализ расширения возможностей центра по взаимодействию с экспертами. В рамках данного доклада рассматривается модель, в которой центр в дополнение к сбору заявок от экспертов (на основании которых он определяет результат экспертизы) может брать с экспертов «плату» за участие в экспертизе, размер которой определяется на основании заявки эксперта и итогового результат экспертизы. Данную плату можно назвать «штрафом» за отклонения мнения эксперта от результата. В подобной постановке задача активной экспертизы может быть рассмотрена как задача выбора уровня общественного блага с трансфертами [2,3]. Исследованию подобных задач уделяется значительное внимание в отрасли микроэкономике, которая называется *mechanism design* [4].

2. Постановка задачи

Рассматривается базовая постановка задачи активной экспертизы [1]. Организационная система состоит из одного элемента верхнего уровня – организатор экспертизы или центра (Ц) и конечного множество агентов (экспертов) $N = \{1, 2,..., n\}$.

Центр определяет значение некоторого параметра x на основании сообщений экспертов $s = \{s_1, ..., s_n\}$ о значении данного параметра на основании процедуры экспертизы $x = \pi(s)$. Область допустимых значений результата экспертизы и заявок экспертов – отрезок из R^1: $s_i \in A \subset \mathrm{R}^1,\ \forall i \in N$

У каждого агента $i \in N$ есть предпочтения относительно результата экспертизы, зависящие от *типа* r_i - наиболее предпочтительного для эксперта результата экспертизы. В данной работе рассматриваются *квадратичные* функции предпочтения экспертов

$$\upsilon_i(x, r_i) = -(x - r_i)^2, \quad (1)$$

а типы всех экспертов принадлежат тому же множеству, что и их заявки: $r_i \in A \subset \mathrm{R}^1,\ \forall i \in N$.

В отличии от «классической» задачи активной экспертизы, центр имеет возможность назначать экспертам плату за участие в экспертизе, которая может зависеть от заявок всех экспертов $t: A^n \to \mathrm{R}^n$,
которая входит линейно в функцию полезности экспертов

$$u_i = -(\pi(s) - r_i)^2 + t_i(s)$$

Задачей центра является построения механизма активной экспертизы $g(s) = \{\pi(s), t(s)\}$ (механизма общественного выбора), который позволит получать в качестве результата экспертизы среднее всех истинных типов агентов

$$f:\ x = \frac{1}{n}\sum_i r_i \qquad (2)$$

при любом возможном профиле (наборе) типов экспертов

$$\forall r \in A^n \quad g(s(r)) = f(r). \qquad (3)$$

3. Решение задачи методами mechanism design

В теории организационных систем доказано, что в случае отсутствия трансфертов для любого механизма активной экспертизы возможно предложить прямой неманипулируемый механизм $h(r)$ (механизм в котором эксперты сообщают непосредственное значение собственного типа и сообщение истинного значения является наилучшей стратегией для любого эксперта не зависимо от сообщений остальных) не меньшей эффективности. Однако задача (3) решена быть не может. Возможно, лишь найти прямой неманипулируемый механизм, наилучший по критерию минимума максимальной ошибки

$$K(h) = \max_{r \in A^n} \left| h(r) - f(r) \right|.$$

Причем для функции усреднения (2) оптимальным по данному критерию оказывается прямой неманипулируемый механизм $h(r)$, эквивалентный процедуре экспертизы описываемой (2):

$$\forall r \in A^n,\ h^*(r) = f(s^*(r)),$$

где $s^*(r)$ - равновесные по Нэшу сообщения экспертов [5].

В *mechanism design* важным свойством функции общественного выбора является ее *ex-post* эффективность:

$$\forall r \in A^n,\ \forall f' \in A$$

$$\sum_i \upsilon_i(f(r), r_i) \geq \sum_i \upsilon_i(f', r_i)$$

Для поставленной задачи можно доказать следующую лемму.

Лемма 1. Функция общественного выбора (3) ex post эффективна для функций предпочтения экспертов, описываемых (1).

Доказательство. Данное утверждение доказывается достаточно легко, т.к.

$$\arg\max_{x \in A}\left(\sum_i -(x - r_i)^2\right) = \frac{1}{n}\sum_i r_i$$

для всех $r \in A^n$ и $f' \in A$ •

Поэтому, для описанной модели применима следующая теорема (Гровса – Грина – Лаффона)[2,4]:

(I) Пусть *f* - ex post эффективная функция общественного выбора, *t* - функция трансферов и для каждого $i \in N$ существует такая функция $x_i : \times_{j \neq i} A \to \mathrm{R}$,
что

$$t_i(r) = x_i(r_{-i}) + \sum_{j \neq i} \upsilon(f(r), r_j). \qquad (4)$$

Тогда прямой механизм (*f, t*) реализует *f* в доминантных стратегиях.

(II) Обратно, пусть *f* - ex post эффективная функция общественного выбора, прямой механизм (*f, t*) реализует *f* в доминантных стратегиях и, кроме того, $\{\upsilon_i(\cdot, r_i) \mid r_i \in A\} = \{\upsilon : A \to \mathrm{R}\}$ для каждого $i \in N$. Тогда для каждого $i \in N$ существует функция $x_i : \times_{j \neq i} \Theta_j \to \mathrm{R}$

удовлетворяющая [4].

Принципиальным моментом в данной теореме является тот факт, что трансфер, назначаемый агенту, не может зависеть от заявки агента о собственном типе. Результатом данной теоремы является механизм Гровса-Кларка [2], который носит так же название решающего голоса. В нем функции $x_i(r_{-i})$ для агента *i* определяются как ex-post эффективное решение поставленной задач в его отсутствии

$$x_i(r_{-i}) = -\max_{a \in A}\left(\sum_{j \neq i} \upsilon_j(a, r_j)\right),$$

а сам трансфер (налог Кларка) записывается следующим образом:

$$t_i(\theta) = \sum_{j \neq i} \upsilon_j(f(\theta), \theta_j) - \max_{x \in A}\left(\sum_{j \neq i} \upsilon_j(x, \theta_j)\right). \qquad (5)$$

Для поставленной нами задачи, получаем, что искомый нами прямой и неманипулируемый механизм $g(r) = \{\pi(r), t(r)\}$:

$\pi(r) = f(r)$,

$$t_i(r) = \sum_{j \neq i} [(\frac{1}{n-1} \sum_{l \neq i} r_l - r_j)^2 - (\frac{1}{n} \sum_l r_l - r_j)^2] . \quad (6)$$

Минусом данного механизма является его *несбалансированность*:

$$\sum_i t_i(r) \neq 0,$$

т.е. полученные от экспертов платежи должны быть удалены из системы. В [3] приводится алгоритм модификации данного механизма приводящий к сбалансированности трансферам – механизм «*квадратичного правительства*», однако здесь он опускается в силу большой сложности его получения. Далее мы приведем альтернативный принцип построения механизма Гровса-Кларка, позволяющий достаточно легко перейти к сбалансированным трансферам.

4. Решение задачи методом штрафа за отклонение от общего мнения

Рассмотрим следующий принцип определения платы для эксперта за участие в экспертизе:

$$t : t_i(r'_i) = -\alpha \upsilon_i(f(r'_i, r_{-i}) - r'_i). \quad (7)$$

Иными словами плата для каждого эксперта пропорциональна с некоторым коэффициентом α значению его функции предпочтения при условии, что сообщенный им собственный тип r'_i является истинным. Т.е. плата, назначаемая эксперту зависит напрямую от сообщаемой им информации. Тогда для механизма $g^*(r) = \{f(r), t(r))\}$, где $f(r)$ определяется из (2), а $t(r)$ - из (7) можно доказать следующие результаты.

Утверждение 1. Для функции предпочтения экспертов, определяемых (1), механизм $g^*(r)$ является прямым и неманипулируемым при $\alpha = (n-1)^{-1}$.

Доказательство. Условие, что сообщение правды для каждого эксперта является равновесием в доминантных стратегиях записывается следующим образом

$$\arg\max_{s_i \in A} (\upsilon_i(f(s_i, s_{-i}), r_i) + t(s_i, s_{-i})) = r_i$$

для $\forall i \in N, \forall r \in A^n$. В условиях утверждения получаем, что

$$\arg\max_{s'_i \in A} (-(f(s_i, s_{-i}) - r_i)^2 - \frac{1}{n-1}(f(s'_i, s_{-i}) - s'_i)^2) = r_i \quad \bullet$$

Более того, нетрудно получить, что записи (6) и (7) при $\alpha = (n-1)^{-1}$ эквивалентны. Т.е. механизм Гровса-Кларка получен исходя из гораздо более простых предположений.

Так же, представляется возможным решит рассматриваемую задачу сбалансировав трансферы следующим образом:

$$t : t_i(r_i) = \hat{t}_i(r_i) - \frac{1}{n} \sum_j \hat{t}_j(r_j), \quad (8)$$

где $\hat{t}_i(\theta'_i)$ определяется из (7) . Иными словами, общая сумма платежей, собранная со всех экспертов, делится между ними поровну. Очевидно, что $t(r)$, определяемая (8), удовлетворяет условию $\sum_i t_i(r) = 0$,

т.е. трансферы являются сбалансирванными. Тогда для механизма $g^{**}(r) = \{f(r), t(r))\}$, где $f(r)$ определяется из (2), а $t(r)$ - из (8) можно доказать следующие результаты.

Утверждение 2. Для функции предпочтения экспертов, определяемых (1), механизм $g^{**}(r)$ является прямым и неманипулируемым при $\alpha = (n-2)^{-1}$.

Доказательство. Доказательство аналогично доказательству утверждения 1.

Полученный нами механизм $g^{**}(r)$ является механизмом «квадратичного правительства», полученным в [3]. Но, в нашем подходе плата, назначаемая эксперту, основывается на отклонении его заявки от результата экспертизы, в то время как механизм Гровса-Кларка и механизм «квадратичного правительства» исходя из условия, что штраф, назначаемый эксперты определяется на основании влияния заявки эксперта на результат экспертизы. Следует отметить, механизмы, получаемые на основании этих двух подходов совпадают далеко не всегда. Причиной этого совпадения является тот факт, что процедура экспертизы (2) ex-post эффективной для функций предпочтения экспертов, определяемых (1).

5. Заключение

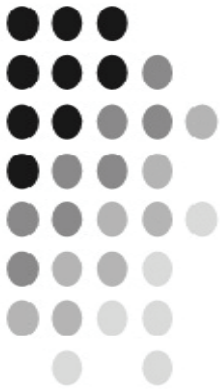
Введение платы за участие в экспертизе позволяет с помощью методов mechanism design построить неманипулируемые механизмы активной экспертизы, реализующие максимально достоверный результат, что не возможно при применение механизмов без платы за участие. Предложен подход к определению платы за участие, позволяющий упростить процесс разработки механизмов активной экспертизы с платой за участие по сравнению с подходами mechanism design.

6. Библиография

1. БУРКОВ В.Н., НОВИКОВ Д.А. *Как управлять организациями.* М.: Синтег, 2004. – 400 с.
2. GROVES T., Efficient Collective Choice when Compensation is Possible // *The Review of Economic Studies*, Vol. 46, No. 2. (Apr., 1979), pp. 227-241.
3. GROVES T., Ledyard J., The Existence of Efficient and Incentive Compatible Equilibria with Public Goods // *Econometrica*, Vol. 48, No. 6. (Sep., 1980), pp. 1487-1506.
4. JACKSON M., Mechanism Theory // *Optimization and Operations Research*, edited by Ulrich Derigs, in Encyclopedia of Life Support Systems (EOLSS), Developed under the auspices of the UNESCO, Eolss Publishers, Oxford, UK, 2003 http://www.eolss.net
5. БУРКОВ В.Н., ДАНЕВ Б., ЕНАЛЕЕВ А.К. и др. *Большие системы: моделирование организационных механизмов*. М.: Наука, 1989. - 245 с.

Раздел 3

Передача, защита и обработка информации



Модель беспроводного канала связи в условиях крупномасштабного замирания и случайного разброса параметров приемопередатчиков

Баскаков С.С.
Московский государственный технический университет имени Н.Э. Баумана
sergey.baskakov@mail.ru

Аннотация

В работе представлена аналитическая модель маломощного радиочастотного канала связи в условиях крупномасштабного замирания и при наличии случайных отклонений параметров приемопередатчиков от номинальных значений. Она позволяет создавать более адекватные имитационные модели беспроводных сетей и получать более приближенную к реальным условиям эксплуатации оценку их характеристик.

1. Введение

В последнее время активное развитие получили технологии беспроводных сенсорных сетей (БСС), которые имеют множество практических применений, связанных с распределенным сбором, анализом и передачей информации. Маршрутизация пакетов является одной из наиболее актуальных задач в области БСС, но в большинстве работ протоколы маршрутизации рассматриваются с использованием сильно упрощенной модели беспроводного канала связи, согласно которой пакет успешно принимается только при условии, что расстояние между передатчиком и приемником меньше некоторого фиксированного радиуса радиосвязи, в противном случае пакет игнорируется. Однако многочисленные экспериментальные исследования (например, [1,2]) характеристик маломощных беспроводных каналов связи показали, что диапазон приема условно разделен на 3 области: область надежного приема, переходная область и область отсутствия связи.

В области надежного приема соединения обеспечивают высокое и стабильное качество связи, в то время как переходная область отличается значительной вариацией и асимметрией качества связи. При этом большая относительная протяженность переходной области приводит к тому, что указанные эффекты проявляются в большинстве беспроводных соединений.

Вариация надежности соединений в различных направлениях передачи вызвана неравномерностью затухания сигнала, причинами которой являются:

- крупномасштабные и мелкомасштабные замирания;
- неравномерность диаграмм направленности антенн;
- препятствия на пути прохождения сигнала, вызывающие различное поглощение сигнала.

Асимметрия качества связи означает, что вероятность успешного приема узлом v пакета от узла w не равна вероятности успешной передачи в обратном направлении. Асимметрия соединений обусловлена тем, что узлы имеют различные значения выходной мощности и уровня шума. При этом отклонения параметров приемопередатчиков от номинальных значений описываются случайным законом распределения, характеристики которого зависят от технологических особенностей производства приемопередатчиков и аппаратной реализации узлов.

Переходная область может оказать существенное негативное влияние на характеристики протоколов верхних уровней, в первую очередь - на эффективность протоколов маршрутизации.

Таким образом, крайне важно использовать адекватную модель беспроводного канала связи на этапе разработки и анализа протоколов маршрутизации. Кроме того, применение при имитационном моделировании реалистичной модели радиоканала позволит более точно предсказывать свойства и показатели качества обслуживания сети, которые будут иметь место при развертывании системы в полевых условиях.

Множество исследований было проведено с целью создания методов моделирования и инструментов проектирования систем сотовой связи и сетей передачи данных (например, проекты COST 207 и COST 259). Однако в данных проектах предметом изучения являются строго определенные частотные диапазоны, методы модуляции, способы кодирования и т.д. При этом критерии проектирования и условия эксплуатации систем персональной радиосвязи таковы, что эффекты переходной области не оказывают на них существенного влияния. Следовательно, эти модели

неприменимы для исследования беспроводных сенсорных сетей.

В работах [2-4] предлагаются различные эмпирические модели описания характеристик маломощного беспроводного канала, основанные на экспериментальных данных, но они не предлагают фундаментального объяснения базовых причин наблюдаемых явлений. Системный анализ перечисленных эффектов содержится в работе [5], в которой с помощью аналитических выражений описана зависимость надежности и асимметричности соединений от параметров приемопередатчика и среды распространения радиоволн.

В настоящей работе предлагается модель беспроводного канала связи, основанная на [5], но расширенная рядом дополнительных параметров. Кроме того, получены более точные выражения для математического ожидания и дисперсии вероятности успешного приема пакета, а также приведены формулы для оценки величины асимметрии качества связи.

Предложенная модель может использоваться для исследования цифровых беспроводных каналов связи с различными типами манипуляции и кодирования двоичных символов. В настоящей работе в качестве примера выбран физический уровень стандарта IEEE 802.15.4.

2. Модель потерь в тракте

В большинстве практических приложений сенсорные сети используются в условиях многолучевого распространения сигналов. Одной из наиболее часто используемых моделей каналов с крупномасштабными замираниями является логнормальная модель, согласно которой средние потери на расстоянии d равны

$$\overline{PL}(d) = PL(d_0) + 10\alpha \lg\left(\frac{d}{d_0}\right), \text{ дБ}, \qquad (1)$$

где

d - расстояние между передатчиком и приемником, м;

d_0 - эталонное расстояние, м;

α - показатель степени потерь в тракте.

Для учета влияния крупномасштабного замирания в формулу (1) вводится случайная составляющая, распределенная по нормальному закону с нулевым средним и дисперсией σ_{ch}^2, т.е.

$$PL(d) = \overline{PL}(d) + X_{ch}, \text{ дБ}. \qquad (2)$$

Для учета анизотропии диаграмм направленности антенн используем модель нерегулярности приемопередатчиков RIM [3], в которой введено понятие степени нерегулярности DOI (degree of irregularity) как максимальное отклонение (в процентах) величины затухания при единичном изменении направления передачи. Предполагая размещение узлов на плоскости и принимая в качестве единицы изменения направления передачи 1 угловой градус, можно задать 360 коэффициентов $\{K_i\}_{i=1}^{359}$, отражающих отличия потерь в различных направлениях.

Для i-ого направления коэффициент равен

$$K_i = \begin{cases} 1 & при \quad i = 0, \\ K_{i-1} \pm R_w \times DOI & при \quad 0 < i < 360, \end{cases}$$

где R_w - случайная величина, распределенная по закону Вейбулла.

Влияние известных препятствий (например, стен) между узлами может быть учтено введением в выражение (2) дополнительного слагаемого Ω_{vw}, характеризующего суммарное затухание (в дБ), которое вносят препятствия на пути прохождения сигнала между узлами v и w.

Таким образом, общая формула для вычисления потерь в тракте между узлами v и w имеет следующий вид

$$PL_{vw}(d) = K_{v,i} K_{w,j} \left[\overline{PL}(d) + \Omega_{vw}\right] + X_{ch}, \text{ дБ}. \qquad (3)$$

3. Модель приемопередатчика

Параметры реальных приемопередатчиков имеют случайные отклонения от номинальных значений. Эти отклонения вызваны как технологическим процессом производства микросхем, так и особенностями аппаратной реализации узлов.

Разброс значений выходной мощности P_t и уровня шума P_n приемопередатчиков подчиняются нормальному закону распределения, при этом в общем случае P_t и P_n являются зависимыми случайными величинами для конкретного экземпляра приемопередатчика интегрального исполнения. Тогда

$$\begin{bmatrix} P_t \\ P_n \end{bmatrix} \sim N\left(\begin{bmatrix} \overline{P}_t \\ \overline{P}_n \end{bmatrix}, \Sigma_{hw}\right), \qquad (4)$$

где

$\overline{P}_t$ - среднее (номинальное) значение выходной мощности передатчика, дБм;

$\overline{P}_n$ - среднее (номинальное) значение уровня шума приемника, дБм;

Σ_{hw} - ковариационная матрица вида

$$\Sigma_{hw} = \begin{bmatrix} \sigma_t^2 & \mathrm{cov}(P_t, P_n) \\ \mathrm{cov}(P_n, P_t) & \sigma_n^2 \end{bmatrix}.$$

При передаче данных от узла v к узлу w на расстояние d отношение сигнал-шум (ОСШ) на входе приемника узла w равно

$$\gamma_{vw}(d) = P_{v,t} - PL_{vw}(d) - P_{w,n}, \text{ дБ}, \quad (5)$$

где

$P_{v,t}$ - выходная мощность передатчика узла v, дБм;

$P_{w,n}$ - уровень шума приемника узла w, дБм.

Заметим, что в приведенном выражении для расчета ОСШ мы не учитываем влияние внутриполосных помех, т.к. предполагается, что для работы сети выбран частотный диапазон свободный от других систем связи. А также из-за малого объема трафика вероятность возникновения коллизий пренебрежимо мала, поэтому они также не включены в модель канала. Однако, в процессе имитационного моделирования мощность помехи, воздействующей на конкретный узел, и уровни сигналов от конкурирующих узлов могут использоваться при расчете мгновенного значения ОСШ для каждого принятого пакета.

В формуле (5) используются значения выходной мощности и уровня шума для разных узлов, поэтому они являются независимыми относительно друг друга случайными величинами с распределениями $\mathrm{N}(\overline{P}_t, \sigma_t^2)$ и $\mathrm{N}(\overline{P}_n, \sigma_n^2)$ соответственно. Объединяя выражения (3) и (5), а также фиксируя значения $K_{v,i}$, $K_{w,j}$ и Ω_{vw}, получаем, что случайная величина $\gamma_{vw}(d)$ подчиняется нормальному закону распределения с параметрами

$$\mu_{\gamma,vw}(d) = \overline{P}_t - K_{v,i}K_{w,j}\left[\overline{PL}(d) + \Omega_{vw}\right] - \overline{P}_n \quad (6)$$

$$\sigma_\gamma^2 = \sigma_t^2 + \sigma_n^2 + \sigma_{ch}^2. \quad (7)$$

Зависимость вероятности битовой ошибки (bit error rate) от значения ОСШ γ определяется типом манипуляции. В настоящей работе для примера рассматривается физический уровень стандарта IEEE 802.15.4, поэтому

$$\beta(\gamma) = \frac{1}{30}\sum_{k=2}^{16}(-1)^k \binom{16}{k} \exp\left[20\left(\frac{1}{k} - 1\right)10^{\gamma/10}\right]. \quad (8)$$

Тогда вероятность успешного приема пакета (packet reception rate) длиной L байт равна

$$\Psi(\gamma, L) = \left[1 - \beta(\gamma)\right]^{8L}. \quad (9)$$

Выражение (9) записано в предположении об отсутствии специального кодирования двоичных символов, что верно для приемопередатчиков стандарта IEEE 802.15.4. Если используется какая-либо схема кодирования, то зависимость пакетной ошибки от вероятности битовой ошибки может иметь другой вид.

4. Оценка границ областей приема

Ранее было сказано, что в распределении вероятностей успешного приема пакетов можно выделить 3 области диапазона приема: область надежной связи, переходная область и область отсутствия связи. Определим границу между ними следующим образом.

Началом переходной области (окончанием области надежного приема) является максимальное значение расстояния между передатчиком и приемником d_b, при котором с вероятностью p_b вероятность успешного приема пакета больше заданной величины ψ_b (при некотором фиксированном размере пакета L). Тогда из (8) и (9) следует, что значение $d_{vw,b}$ для соединения между узлами v и w определяется формулой

$$p_b = \Pr(\Psi > \psi_b) = \Pr\left(\gamma_{vw}(d_{vw,b}) > \gamma_b\right),$$

где $\gamma_b = \Psi^{-1}(\psi_b, L)$ - пороговое значение ОСШ, соответствующее уровню вероятности успешного приема пакета ψ_b.

Следовательно, имеет место равенство

$$p_b = 1 - \Phi\left(\frac{\gamma_b - \mu_{\gamma,vw}(d_{vw,b})}{\sigma_\gamma}\right)$$

(где $\Phi(x)$ - функция распределения стандартной нормальной случайной величины), из которого получаем искомое значение

$$d_{vw,b} = d_0 10^{D_{vw,b}/10\alpha}, \text{ м}, \quad (10)$$

где

$$D_{vw,b} = \frac{\overline{P}_t - \overline{P}_n - \gamma_b + \sigma_\gamma u_{1-p_b}}{K_{v,i}K_{w,j}} - PL(d_0) - \Omega_{vw};$$

u_p - квантиль порядка p стандартного нормального распределения.

Окончанием переходной области (началом области отсутствия связи) является минимальное значение расстояния между передатчиком и приемником d_e, при котором с вероятностью p_e вероятность успешного приема пакета меньше заданной величины ψ_e, т.е. $p_e = \Pr(\Psi < \psi_e)$.

Выполнив аналогичные действия, получаем формулу для вычисления значения $d_{vw,e}$:

$$d_{vw,e} = d_0 10^{D_{vw,e}/10\alpha}, \text{ м}, \quad (11)$$

где

$$D_{vw,e} = \frac{\overline{P}_t - \overline{P}_n - \gamma_e + \sigma_\gamma u_{p_e}}{K_{v,i} K_{w,j}} - PL(d_0) - \Omega_{vw};$$

$\gamma_e = \Psi^{-1}(\psi_e, L)$ - пороговое значение ОСШ, соответствующее вероятности успешного приема пакета ψ_e.

Выражения (10) и (11) используются для оценки параметров переходной области для каждого конкретного соединения между заданными узлами v и w. Для получения средних оценок для всей сети положим $K_{v,i}K_{w,j} = 1$ и $\Omega_{vw} = 0$. Тогда

$d_b = d_0 10^{D_b/10\alpha}$ и $d_e = d_0 10^{D_e/10\alpha}$,

где

$$D_b = \overline{P}_t - PL(d_0) - \overline{P}_n - \gamma_b + \sigma_\gamma u_{1-p_b};$$

$$D_e = \overline{P}_t - PL(d_0) - \overline{P}_n - \gamma_e + \sigma_\gamma u_{p_e}.$$

Коэффициент переходной области Γ определен как отношение размера переходной области к размеру области надежного приема, т.е.

$$\Gamma = \frac{d_e - d_b}{d_b} = 10^{\frac{\gamma_b - \gamma_e + \sigma_\gamma [u_{Pe} - u_{1-Pb}]}{10\alpha}} - 1.$$

Полученные выражения показывают, что границы начала и окончания областей приема зависят как от параметров приемопередатчиков, так и от условий среды распространения. Но коэффициент переходной области Γ не зависит от номинальных выходной мощности передатчика и уровня шума приемника, а зависит, в частности, от показателя степени потерь в тракте α и общего среднеквадратического отклонения ОСШ σ_γ. Это означает, что, например, увеличение выходной мощности передатчика приведет к увеличению как области надежной связи, так и переходной области, при этом соотношение их размеров останется прежним. Усиление эффектов крупномасштабного замирания и увеличение отклонения параметров приемопередатчиков от номинальных значений приводит к относительному расширению переходной области, в то время как большие значения α уменьшают Γ, т.к. сигнал затухает быстрее.

Отметим, что выбранные значения p_b и p_e можно рассматривать как доверительные вероятности оценивания границ областей приема. Чем выше значения p_b и p_e, тем выше точность оценки, но при этом переходная область расширяется, что продемонстрировано на рис. 1.

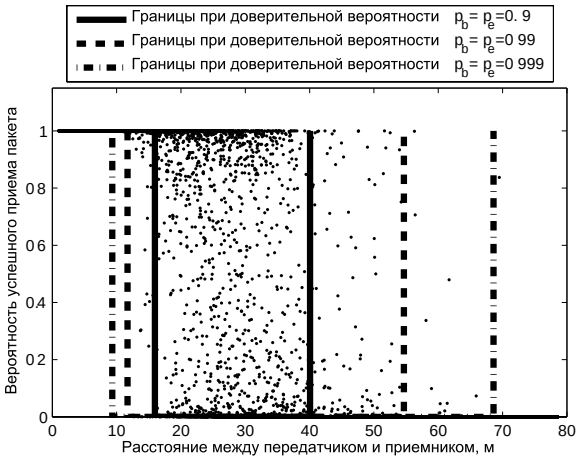


Рис. 1. Границы областей приема при различных доверительных вероятностях оценивания

5. Распределение вероятности успешного приема пакета

Отношение сигнал-шум на входе приемника является нормальной случайной величиной с математическим ожиданием μ_γ (6) и дисперсией σ_γ^2 (7), поэтому математическое ожидание значения коэффициента принятых пакетов определяется выражением

$$\mathbf{M}[\Psi] = \int_{-\infty}^{\infty} \Psi(\gamma) f(\gamma, \mu_\gamma, \sigma_\gamma) d\gamma, \quad (12)$$

где $f(x, \mu, \sigma)$ - функция плотности вероятности нормального распределения $\mathrm{N}(\mu, \sigma^2)$.

Аналитическое вычисление данного интеграла затруднительно из-за сложного характера зависимости вероятности успешного приема пакета от значения ОСШ (см. выражения (8) и (9)), поэтому предлагается аппроксимировать $\Psi(\gamma)$ кусочно-линейной функцией (рис. 2)

$$\Psi_L(\gamma) = \begin{cases} 0, & при \quad \gamma \le \gamma_{e1}, \\ k_e\gamma + b_e, & при \quad \gamma_{e1} < \gamma < \gamma_{e2}, \\ 1, & при \quad \gamma \ge \gamma_{e2}. \end{cases}$$

Параметры функции $\Psi_L(\gamma)$ вычисляются по выражениям:

$$\gamma_{e1} = \Psi^{-1}(\psi_1) \qquad \gamma_{e2} = \Psi^{-1}(\psi_2)$$

$$k_e = \frac{\psi_2 - \psi_1}{\gamma_{e2} - \gamma_{e1}} \quad b_e = \frac{\psi_1\gamma_{e2} - \psi_2\gamma_{e1}}{\gamma_{e2} - \gamma_{e1}},$$

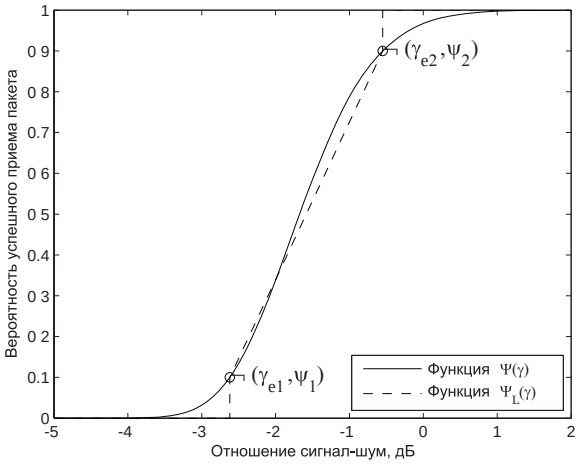


Рис. 2. Кусочно-линейная аппроксимация функций $\Psi(\gamma)$

где ψ_1 и ψ_2 - выбранные пороговые значения функции $\Psi(\gamma)$ (например, 0.1 и 0.9 соответственно).

Тогда приближенное значение математического ожидания коэффициента принятых пакетов равно

$$M_\Psi \approx \int_{-\infty}^{\infty} \Psi_L(\gamma) f(\gamma,\mu_\gamma,\sigma_\gamma)\,d\gamma = \\ = I(\gamma_{e1},\gamma_{e2},k_e,b_e,\mu_\gamma,\sigma_\gamma), \quad (13)$$

где

$$I(x_1,x_2,k,b,\mu,\sigma) = k\sigma^2\left[f(x_1,\mu,\sigma) - \right. \\ \left. - f(x_2,\mu,\sigma)\right] + (k\mu+b-1)\Phi\left(\frac{x_2-\mu}{\sigma}\right) - \\ -(k\mu+b)\Phi\left(\frac{x_1-\mu}{\sigma}\right)+1.$$

Для вычисления дисперсии необходимо определить математическое ожидание квадрата случайной величины Ψ

$$\mathbf{M}\left[\Psi^2\right] = \int_{-\infty}^{\infty} \Psi^2(\gamma) f(\gamma,\mu_\gamma,\sigma_\gamma)\,d\gamma. \quad (14)$$

Представив функцию $\Psi^2(\gamma)$ в виде кусочно-линейной аппроксимации по аналогии с $\Psi_L(\gamma)$, получаем

$$D_\Psi = \mathbf{D}[\Psi] \approx I(\gamma_{v1},\gamma_{v2},k_v,b_v,\mu_\gamma,\sigma_\gamma) - M_\Psi^2. \quad (15)$$

В [5] также приведены выражения для оценки среднего значения и дисперсии коэффициента принятых пакетов, но при их выводе использовались линейные аппроксимации как функций $\Psi(\gamma)$ и $\Psi^2(\gamma)$, так и функции плотности вероятности $f(\gamma,\mu_\gamma,\sigma_\gamma)$ на интервалах $(\gamma_{e1};\gamma_{e2})$ и $(\gamma_{v1};\gamma_{v2})$. В настоящей же работе приближенное представление было выполнено только для $\Psi(\gamma)$ и $\Psi^2(\gamma)$, поэтому выражения (13) и (15) представляют собой более точные оценки.

На рис. 3 показан пример зависимости от расстояния M_Ψ и D_Ψ, при этом для сравнения приведены значения, полученные при численном решении интегралов (12) и (14). Графики наглядно демонстрируют сказанное выше относительно характеристик областей приема (см. рис. 1 и 3): с увеличением расстояния среднее качество связи монотонно убывает, в то время как дисперсия имеет колоколообразный вид, принимая максимальное значение в переходной области. В области надежного приема коэффициент принятых пакетов стабильно высокий и $D_\Psi \approx 0$, но по мере движения в переходную область дисперсия увеличивается, приводя к значительным колебаниям качества связи, а в области отсутствия связи оба параметра распределения убывают до 0.

Очевидно, что выражения (13) и (15) могут использоваться при произвольных типах манипуляции и схемах кодирования. Необходимо только вычислить параметры кусочно-линейных аппроксимаций для рассматриваемого сочетания $\beta(\gamma)$ и $\Psi(\gamma,L)$ при заданной длине пакетов L.

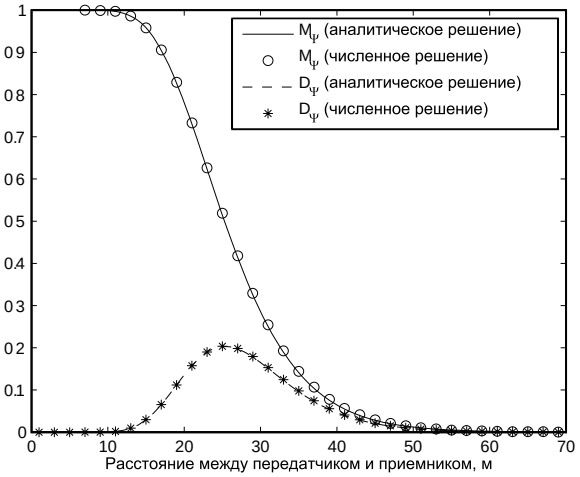


Рис. 3. Параметры распределения вероятности успешного приема пакета

6. Распределение асимметрии вероятности успешного приема пакета

Определим величину асимметрии ОСШ $\Delta\gamma$ как разность ОСШ пар узлов. Тогда из формулы

(5) следует, что величина $\Delta\gamma_{vw}$ не зависит от затухания на пути распространения сигнала между узлами и равна

$$\Delta\gamma_{vw} = \gamma_{vw} - \gamma_{wv} = \left(P_{v,t} + P_{v,n}\right) - \left(P_{w,t} + P_{w,n}\right).$$

Следовательно, $\Delta\gamma$ является нормально распределенной случайной величиной с нулевым средним и дисперсией равной

$$\sigma^2_{\Delta\gamma} = 2\left[\sigma_t^2 + \sigma_n^2 + 2\operatorname{cov}\left(P_t, P_n\right)\right].$$

Определим параметры распределения величины асимметрии вероятности успешного приема пакета, измеренную как разность качества входящей и исходящей связей, т.е.

$$\Delta\Psi\left(\gamma\right) = \Psi\left(\gamma\right) - \Psi\left(\gamma + \Delta\gamma\right).$$

Снова используя кусочно-линейные приближения, получаем

$$M_{\Delta\Psi} \approx \Psi\left(\gamma\right) - I\left(\gamma_{e1}, \gamma_{e2}, k_e, b_e, \gamma, \sigma_{\Delta\gamma}\right),$$

$$D_{\Delta\Psi} \approx \Psi^2\left(\gamma\right) - 2\Psi\left(\gamma\right) I\left(\gamma_{e1}, \gamma_{e2}, k_e, b_e, \gamma, \sigma_{\Delta\gamma}\right) +$$

$$+ I\left(\gamma_{v1}, \gamma_{v2}, k_v, b_v, \gamma, \sigma_{\Delta\gamma}\right) - M^2_{\Delta\Psi}.$$

На рис. 4 показан пример зависимости параметров $M_{\Delta\Psi}$ и $D_{\Delta\Psi}$ от ОСШ, а также приведены результаты численного решения интегралов без использования линейной аппроксимации.

Допустим, что γ представляет собой значение ОСШ на входе приемника узла v при приеме сигнала от узла w. Тогда для узла v $\Psi\left(\gamma\right)$ является показателем качества входящей связи, а $\Psi\left(\gamma + \Delta\gamma\right)$ - показателем качества исходящей связи.

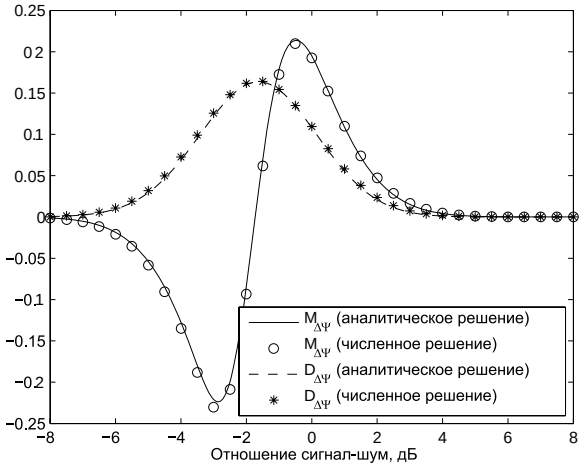


Рис. 4. Параметры распределения асимметрии вероятности успешного приема пакета

При малых ОСШ $\Psi\left(\gamma\right) \approx 0$, но тем не менее передача данных от v к w возможна с ненулевой вероятностью. Дальнейшее повышение ОСШ приводит к еще большему увеличению асимметрии в пользу исходящей связи, но достигнув экстремума она убывает, принимая нулевое значение при $\Psi\left(\gamma\right) \approx 0.5$, что означает равновероятное отклонение как в сторону качества исходящей связи, так и в обратную сторону, при этом дисперсия принимает максимальное значение. Далее наоборот начинает доминировать входящая связь, т.е. для большей части соединений имеет место неравенство $\Psi\left(\gamma\right) > \Psi\left(\gamma + \Delta\gamma\right)$.

7. Заключение

В данной работе предложена аналитическая модель для описания цифрового беспроводного канала связи в условиях крупномасштабного замирания и при наличии случайных отклонений параметров приемопередатчиков от номинальных значений.

Модель позволяет учесть такие особенности маломощных радиоканалов, как вариация и асимметрия качества связи при разработке и исследовании протоколов для БСС, что повысит их надежность в реальных условиях эксплуатации, а также увеличит точность оценок, полученных при имитационном моделировании сетей.

8. Список литературы

[1] J. Zhao and R. Govindan, “Understanding packet delivery performance in dense wireless sensor networks”, In *Proceedings of the 1st International Conference on Embedded Networked Sensor Systems*, 2003.

[2] A. Woo, T. Tong, and D. Culler, “Taming the underlying challenges of reliable multihop routing in sensor networks”, In *Proceedings of the 1st International Conference on Embedded Networked Sensor Systems*, 2003.

[3] G. Zhou, T. He, S. Krishnamurthy, J. Stankovic, “Impact of radio irregularity on wireless sensor networks”, In *Proceedings of the 2nd International Conference on Mobile Systems, Applications, and Services*, 2004.

[4] A. Cerpa, J.L. Wong, L. Kuang, M. Potkonjak, and D. Estrin, “Statistical model of lossy links in wireless sensor networks”, In *Proceedings of the 4th International Symposium on Information Processing in Sensor Networks*, 2005.

[5] M. Zuniga and B. Krishnamachari, “An analysis of unreliability and asymmetry in low-power wireless links”, *ACM Transactions on Sensor Networks*, vol. 3, no. 2, 2007.

Исправление стираний низкоплотностными кодами Галлагера

Зяблов В.В.
ИППИ РАН
zyablov@iitp.ru

Рыбин П.С
ИППИ РАН
prybin@iitp.ru

Аннотация

В статье рассмотрены низкоплотностные коды с кодом-компонентом с проверкой на четность и их корректирующие свойства в канале со стираниями. Показано, что существует такой низкоплотностный код, который при просто реализуемом итеративном алгоритме декодировании исправляет долю стираний, растущую линейно с длиной кода. При этом количество итераций декодирования, необходимых для исправления стираний, равно по порядку логарифму от длины кода. Рассмотрен метод расчета доли исправимых стираний. Проведен сравнительный анализ численных результатов для некоторых параметров кода, представленных в этой статье и полученных ранее.

1. Введение

Низкоплотностный код с проверкой на честность Галлагера [1] (LDPC) определяется разреженной проверочной матрицей $\mathbf{H}$, на строках которой стоят проверочные матрицы кода с проверкой на четность $\mathbf{H}_0$ для символов кодов-компонентов. Проверочная матрица кода с проверкой на четность представляет собой $\mathbf{H}_0 = \underbrace{(111...1)}_{n_0}$, где n_0 -длина кода-компонента.

Как известно, кодовое расстояние кода с проверкой на четность составляет $d_0 = 2$. Следовательно, для канала со стираниями гарантируется исправление $d_0 - 1 = 1$ стирания.

LDPC – коды позволяют построить мощные коды с хорошими корректирующими свойствами при простом итеративном декодировании. Для построения LDPC – кода рассмотрим диагональную блочную матрицу $\mathbf{H}_b$ на главной диагонали, которой стоят b проверочных матриц кода с проверкой на четность (см. (1.1)):

$$\mathbf{H}_b = \begin{pmatrix} \mathbf{H}_0 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_0 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \underbrace{\mathbf{0} \quad \mathbf{0} \quad \dots \quad \mathbf{H}_0}_{b} \end{pmatrix}. \qquad (1.1)$$

Обозначим теперь через $\pi(\mathbf{H}_b)$ случайную перестановку столбцов матрицы $\mathbf{H}_b$. Тогда проверочная матрица $\mathbf{H}$ LDPC – кода получается при использовании $\ell > 2$ таких перестановок в качестве слоев (см. (1.2)):

$$\mathbf{H} = \begin{pmatrix} \mathbf{H}_1 \\ \mathbf{H}_2 \\ \vdots \\ \mathbf{H}_\ell \end{pmatrix} = \begin{pmatrix} \pi_1(\mathbf{H}_b) \\ \pi_2(\mathbf{H}_b) \\ \vdots \\ \pi_\ell(\mathbf{H}_b) \end{pmatrix}. \qquad (1.2)$$

Таким образом, $\mathbf{H}$ представляет собой разреженную матрицу $\ell b \times n$, где $n = bn_0$ - длина LDPC – кода. Для заданного кода-компонента с проверочной матрицей $\mathbf{H}_0$, независимо выбирая случайные перестановки π_l, $l = 1, 2, ..., \ell$, которые все равновероятны, определим ансамбль LDPC – кодов, который обозначим как $\chi(n_0, \ell, b)$. Нижняя оценка на скорость R кода из $\chi(n_0, l, b)$ определяется следующей формулой (см. (1.3)):

$$R \geq 1 - \frac{\ell}{n_0} = 1 - \ell(1 - R_0). \qquad (1.3)$$

Равенство достигается только в случае, когда $\mathbf{H}$ имеет полный ранг.

Впервые корректирующая способность LDPC кодов для канала со стираниями была рассмотрена в [2], где было доказано, что существует LDPC код длины n с декодированием, исправляющем все стирания до кратности αn при сложности $O(n \ln n)$. Также был предложен метод оценки количества исправляемых стираний при независимом

итеративном декодировании и были представлены численные результаты для некоторых параметров кодов.

В данной статье метод, разработанный в [3] для оценки количества исправляемых ошибок LDPC кодом длины n при независимом итеративном декодировании со сложностью $O(n\ln n)$ для двоично-симметричного канала, применен для канала со стираниями. Также представлены численные результаты, полученные для некоторых параметров кодов и проведен сравнительный анализ с данными, полученными ранее в [2].

2. Алгоритм декодирования

Рассмотрим независимый итеративный алгоритм Λ декодирования LDPC кода, каждая $i-$àÿ итерация которого, $i=1,2,...,i_{\max}$, состоит из следующих шагов:

1. для декодируемой последовательности $\mathbf{r}^{(i)}$, где $\mathbf{r}^{(1)}$ - это принятая последовательность $\mathbf{r}$, выбираем коды-компоненты с ровно одним стиранием;
2. для каждого кода-компонента, выбранного на предыдущем шаге, определяем передаваемое значение на позиции стирания: она равна сумме по модулю два нестертых значений. Таким образом, получаем следующую декодируемую последовательность $\mathbf{r}^{(i+1)}$.

Пусть W - количество стираний в принятой последовательности, а a - количество кодов-компонентов с ровно одним стиранием (исправимым стиранием). Тогда в общем случае $a=\alpha W\ell$, где $\alpha\le 1$ - доля кодов-компонентов с ровно одним стиранием. В алгоритме Λ предполагается, что на каждой итерации, $i=1,2,...,i_{\max}$, найдется хотя бы один код с исправимым стиранием, т.е. $\alpha>0$. Следовательно, на первой итерации алгоритма Λ все исправимые стирания будут исправлены, а неисправимые приведут к отказу в декодировании. Таким образом, количество стираний в декодируемой последовательности уменьшиться. Ясно, что если на каждой итерации декодирования количество кодов-компонентов с ровно одним стиранием больше нуля, то количество стираний в декодируемой последовательности $\mathbf{r}^{(i)}$ будет уменьшаться с каждой итерацией и алгоритм восстановит переданное кодовое слово $\mathbf{v}$, т.е. $\mathbf{r}^{(i_{\max})}=\mathbf{v}$. Следовательно, можно утверждать:

Лемма 1: Для любого LDPC кода, если комбинация стираний такая, что на каждой итерации декодирования алгоритма Λ количество кодов-компонентов с исправимыми стираниями больше нуля, то алгоритм Λ восстанавливает переданное кодовое слово за $O(\log n)$ итераций, где n - длина кода.

Доказательство: Обозначим ε нижнюю границу доли исправимых стираний в каждой итерации, $0<\varepsilon<1$. Тогда, после x итераций, количество оставшихся стираний не превосходит $\omega n(1-\varepsilon)^x$. Последняя итерация декодирования $i_{\max}$ достигается, когда

$$\omega n(1-\varepsilon)^{i_{\max}}<1 \qquad (2.1)$$

Следовательно,

$$\log(\omega n)+i_{\max}\log(1-\varepsilon)<0 \qquad (2.2)$$

Откуда,

$$i_{\max}<\frac{1}{\log(\frac{1}{1-\varepsilon})}\log(\omega n) \qquad (2.3)$$

Таким образом, количество итераций декодирования составляет порядка логарифма от длины кода.

Сложность каждой итерации алгоритма Λ пропорциональна длине кода n. Следовательно, согласно *Лемме 1* суммарная сложность декодирования будет составлять $O(n\log n)$.

3. Основной результат

Как отмечалось ранее в [3] был предложен метод оценки количества исправляемых стираний ω_0 при независимом итеративном декодировании Λ, где ω_0 - корень уравнения:

$$H(\omega)+\ell\Psi(\omega,\mu)=0, \qquad (3.1)$$

где $\Psi(\omega,\mu)=\omega\ln\frac{\omega}{\mu}-\omega+\inf_{0<s}\{\mu\ln\left(e^s-s\right)-$ $-\omega\ln s\}$, а $\mu=\frac{1}{n_0}$ - избыточность кода с проверкой на четность.

В данной статье рассмотрен метод, предложенный в [3], применительно к каналу со стираниями. Был получен следующий результат:

Теорема 1: В ансамбле $\chi(n_0,\ell,b)$ существует код (с вероятностью p, которая $\lim_{n\to\infty}p=1$), который может исправить любую комбинацию стираний кратности до $\omega_\alpha n$ со сложностью

декодирования порядка $O(n\log n)$. *Значение* ω_α *равно наибольшему корню уравнения (3.2):*

$$h(\omega) - \ell F(\alpha, \omega, n_0) = 0, \qquad (3.2)$$

где $h(\omega)$ - *функция энтропии:* $h(\omega) = -\omega\log_2\omega - (1-\omega)\log_2(1-\omega)$ *;*

$$F(\alpha,\omega,n_0) \square\ h(\omega) - \frac{1}{n_0}h(\alpha\omega n_0) + \max\{\omega\log_2 s -$$

$$-\frac{1}{n_0}\log_2(g_0(s,n_0)) - \alpha\omega\log_2\left(\frac{g_1(s,n_0)}{g_0(s,n_0)}\right)\}, \qquad (3.3)$$

где $\alpha > 0$ - *доля кодов с исправимыми стираниями, и максимизация производиться по всем* s*, удовлетворяющим неравенству (3.4):*

$$\frac{\alpha\omega n_0}{1-\alpha\omega n_0} \le \frac{g_1(s,n_0)}{g_0(s,n_0)}, \qquad (3.4)$$

где $g_1(s,n_0)$ - *производящая функция исправимых комбинаций стираний для кода-компонента (см. (3.5)):*

$$g_1(s,n_0) = \binom{n_0}{1}s, \qquad (3.5)$$

$g_0(s,n_0)$ - *производящая функция неисправляемых комбинаций стираний для кода-компонента (см. (3.6)):*

$$g_1(s,n_0) = (1+s)^{n_0} - g_1(s,n_0). \qquad (3.6)$$

Как было показано в предыдущем разделе, алгоритм декодирования Λ исправляет любую комбинацию стираний кратности до W, если на каждой итерации $\alpha > 0$. *Теорема 1* позволяет утверждать, что в ансамбле $\chi(n_0,\ell,b)$ существует код, удовлетворяющий данным требованиям.

Доказательство: Введем понятие обобщенного синдрома $\mathbf{S} = (\mathbf{S}_1\,\mathbf{S}_2\ldots\mathbf{S}_\ell) =$ $= (\underbrace{S_{1,1}\,S_{1,2}\ldots S_{1,b}}_{\text{ñëîé 1}}\ \underbrace{S_{2,1}\,S_{2,2}\ldots S_{2,b}}_{\text{ñëîé 2}}\ldots\underbrace{S_{\ell,1}\,S_{\ell,2}\ldots S_{\ell,b}}_{\text{ñëîé }\ell})$, где $S_{l,i} = \begin{cases}1, \tau_{l,i} = 1\\ 0, \text{èíà÷å}\end{cases}$, где $\tau_{l,i}$ - количество стираний в i*-ом* коде l*-ого* слоя.

Покажем, что для фиксированной комбинации $W = \omega n$ стираний вероятность P_W того, что вес обобщенного синдрома $\mathbf{S}$ кода из ансамбля $\chi(n_0,\ell,b)$ не будет превосходить заданной величины $\alpha W\ell$, ограничена сверху:

$$P_W(|\mathbf{S}| \le \alpha W\ell) \le 2^{-n\ell F(\alpha,\omega,n_0)},$$

где $F(\alpha,\omega,n_0)$ - это функция (3.3).

Рассмотрим компоненту $\mathbf{S}_l$ обобщенного синдрома $\mathbf{S}$, соответствующую l*-ому* слою матрицы $\mathbf{H}$. Введем следующую производящую функцию вероятностей для l*-ого* слоя при кратности стираний W:

$$Q_W^l(t) = \sum_{j=0}^{b} P_W(|\mathbf{S}_l| = j)t^j, \qquad (3.7)$$

где $P_W(|\mathbf{S}_l| = j)$ - вероятность того, что для заданной кратности стираний W количество кодов-компонентов с ровно одним стиранием в l*-ом* слое будет в точности равно j.

Определенная выше вероятность равна:

$$P_W(|\mathbf{S}_l| = j) = \frac{N_j(W,n_0,b)\binom{b}{j}}{\binom{n}{W}}, \qquad (3.8)$$

где $N_j(W,n_0,b)$ - количество комбинаций стираний кратности W, при которых вес обобщенного синдрома $\mathbf{S}_l$ l-ого слоя матрицы $\mathbf{H}$ равен j, т.е. $|\mathbf{S}_l| = j$.

Введем далее следующую порождающую функцию для $N_j(W,n_0,b)$:

$$E_j = \sum_{i=0}^{n} N_j(i,n_0,b)s^i. \qquad (3.9)$$

Очевидно, что

$$E_j = \sum_{i=0}^{W} N_j(i,n_0,b)s^i + \sum_{i=W+1}^{n} N_j(i,n_0,b)s^i \ge$$

$$\ge N_j(W,n_0,b)s^W,$$

откуда следует

$$N_j(W,n_0,b) \le \frac{E_j(s)}{s^W}.$$

Полученная верхняя граница на $N_j(W,n_0,b)$ выполняется для всех s. Наиболее близкая граница определяется путем минимизации по s:

$$N_j(W,n_0,b) \le \inf_{s>0}\left\{\frac{E_j(s)}{s^W}\right\}. \qquad (3.10)$$

Каждый переданный символ проверяется ровно одним компонентным кодом Хемминга в каждом слое. Другими словами каждый из b кодов-компонентов одного слоя проверяет непересекающийся с другими набор, состоящий из n_o символов (n_0-набор). Проверка кода-компонента выполняется (т.е. соответствующий компонент $S_{l,i}$ обобщенного синдрома $\mathbf{S}_l$ равен нулю), если соответствующий n_0-набор передан без стираний, либо в нем присутствует более одного стирания. Иначе, компонент обобщенного синдрома равен единице. Таким образом, мы можем записать:

$$E_j(s) = (g_1(s,n_0))^j (g_0(s,n_0))^{b-j}, \qquad (3.11)$$

где $g_0(s,n_0)$ - производящая функция всех n_0-наборов, для которых соответствующий компонент $S_{l,i}$ обобщенного синдрома $\mathbf{S}_l$ равен нулю, а $g_1(s,n_0)$ - производящая функция n_0-наборов, для которых $S_{l,i}=1$.

В силу сказанного выше можем записать:

$$g_1(s,n_0) = \binom{n_0}{1} s, \qquad (3.12)$$

тогда

$$g_0(s,n_0) = (1+s)^{n_0} - g_1(s,n_0). \qquad (3.13)$$

Следовательно, подставляя (3.11) в (3.10), получаем

$$N_j(W,n_0,b) \le \inf_{s>0}\left\{\frac{(g_1(s,n_0))^j (g_0(s,n_0)^{b-j})}{s^W}\right\},$$

которое будучи подставленным в (3.8) дает

$$P_W(|\mathbf{S}_l| = j) \le \frac{\binom{b}{j}}{\binom{n}{W}} \inf_{s>0}\left\{\frac{(g_1(s,n_0))^i (g_0(s,n_0))^{b-j}}{s^W}\right\}$$

Затем, подставив полученное неравенство в (3.7), находим:

$$Q_W^l(t) \le \sum_{j=0}^{b}\left(\frac{\binom{b}{j}}{\binom{n}{W}} \inf_{s>0}\left\{\frac{(g_1(s,n_0))^i (g_0(s,n_0))^{b-i}}{s^W}\right\} t^j\right)$$

$$\le \binom{n}{W}^{-1} \inf_{s>0}\left\{s^{-W}\sum_{j=0}^{b}\binom{b}{j} t^j (g_1(s,n_0))^j (g_0(s,n_0))^{b-j}\right\}$$

$$= \binom{n}{W}^{-1} \inf_{s>0}\left\{s^{-W}(g_0(s,n_0)+tg_1(s,n_0))^b\right\}. \qquad (3.14)$$

Поскольку слои независимы, вес обобщенного синдрома $|\mathbf{S}|$ равен сумме ℓ независимых случайных величин $|\mathbf{S}_l|$, $l=1,2,\ldots,\ell$. Тогда производящая функция вероятностей веса обобщенного синдрома

$$Q_W(t) = \sum_{j=0}^{b\ell} P_W(|\mathbf{S}| = j) t^j \qquad (3.15)$$

равна произведению ℓ производящих функций $Q_W^l(t)$ (см. (3.7)), $l=1,2,\ldots,\ell$:

$$Q_W(t) = \prod_{l=1}^{\ell} Q_W^l(t) = (Q_W^1(t))^{\ell}. \qquad (3.16)$$

Из (3.14) и (3.16) мы получаем:

$$Q_W(t) \le$$
$$\le \binom{n}{W}^{-\ell} \inf_{s>0}\left\{s^{-W\ell}\left(g_0(s,n_0)+tg_1(s,n_0)\right)^{b\ell}\right\} \qquad (3.17)$$

Тогда вероятность того, что вес обобщенного синдрома будет не более $\alpha W\ell$ можно записать как

$$P(|\mathbf{S}| \le \alpha W\ell) = \sum_{j=0}^{\alpha W\ell} P_W(|\mathbf{S}| = j).$$

Из (3.15) мы можем вывести следующее неравенство, которое выполняется для всех $0 < t \le 1$:

$$Q_W(t) \ge \sum_{j=0}^{\alpha W\ell} P_W(|\mathbf{S}| = j) t^j$$
$$= t^{\alpha W\ell} \sum_{j=0}^{\alpha W\ell} P_W(|\mathbf{S}| = j) t^{j-\alpha W\ell}$$
$$\ge t^{\alpha W\ell} \sum_{j=0}^{\alpha W\ell} P_W(|\mathbf{S}| = j).$$

Первое неравенство выполняется для всех $t>0$, когда дополнительное условие $t \le 1$ необходимо для последнего неравенства, т.к. $t^{j-\alpha W\ell} \ge 1$ для $j \le \alpha W\ell$. Откуда следует верхняя граница

$$P(|\mathbf{S}| \le \alpha W \ell) = \sum_{j=0}^{\alpha W \ell} P_W(|\mathbf{S}| = j) \le \frac{Q_W(t)}{t^{\alpha W \ell}},$$

которая, после подстановки (3.17) и минимизации по t, дает

$$P(|\mathbf{S}| \le \alpha W \ell) \le$$

$$\le \binom{n}{W}^{-\ell} \inf_{0<t\le 1} \inf_{s>0} \left\{ (t^{\alpha} s)^{-W\ell} \left(g_0(s,n_0) + t g_1(s,n_0) \right)^{b\ell} \right\}$$

Полученное неравенство может быть переписано как

$$P(|\mathbf{S}| \le \alpha W \ell) \le$$

$$\binom{n}{W}^{-\ell} \inf_{0<t\le 1} \inf_{s>0} \left\{ s^{-W\ell} \left(g_0(s,n_0) \right)^{b\ell} \left(f(t,s) \right)^{b\ell} \right\}, \quad (3.18)$$

где зависимость от t локализована в

$$f(t,s) \triangleq \frac{1 + t\dfrac{g_1(s,n_0)}{g_0(s,n_0)}}{t^{\alpha\omega n_0}}.$$

Приравнивая $\dfrac{\partial f(t,s)}{\partial t} = 0$, находим значение t_0, на котором достигается минимум $f(t,s)$:

$$t_0 = \frac{\alpha\omega n_0}{1-\alpha\omega n_0} \left(\frac{g_1(s,n_0)}{g_0(s,n_0)} \right)^{-1} \quad (3.19)$$

где s такое, что $t_0 \le 1$.

Тогда

$$f(t_0,s) = \frac{1}{1-\alpha\omega n_0} \left(\frac{\alpha\omega n_0}{1-\alpha\omega n_0} \right)^{-\alpha\omega n_0} \left(\frac{g_1(s,n_0)}{g_0(s,n_0)} \right)^{\alpha\omega n_0} =$$

$$= 2^{h(\alpha\omega n_0)} \left(\frac{g_1(s,n_0)}{g_0(s,n_0)} \right)^{\alpha\omega n_0}, \quad (3.20)$$

где $h(\cdot)$ - двоичная функция энтропии, которая может быть записана как $h(p) = \log_2 \dfrac{p^{-p}}{(1-p)^{1-p}}$.

Подставляя (3.20) в (3.18) получаем

$$P(|\mathbf{S}| \le \alpha W \ell) \le \binom{n}{\omega n}^{-\ell} 2^{\frac{h(\alpha\omega n_0)\ell n}{n_0}} \times$$

$$\times \inf_{s>0} \left\{ s^{-\omega n \ell} \left(g_0(s,n_0) \right)^{\frac{\ell n}{n_0}} \left(\frac{g_1(s,n_0)}{g_0(s,n_0)} \right)^{\alpha\omega n \ell} \right\} \quad (3.21)$$

Используя неравенство

$$\binom{n}{\omega n} \le 2^{nh(\omega)}, \quad (3.22)$$

где асимптотическое равенство достигается при $n \to \infty$.

Тогда из (3.21) получаем

$$P(|\mathbf{S}| \le \alpha W \ell) \le 2^{-\ell n F(\alpha,\omega,n_0)}, \quad (3.23)$$

где $F \triangleq h(\omega) - \dfrac{1}{n_0} h(\alpha\omega n_0) + \max_{s>0:t_0\le 1} \{ \omega \log_2(s) -$

$- \dfrac{1}{n_0} \log_2 \left(g_0(s,n_0) \right) - \alpha\omega \log_2 \left(\dfrac{g_1(s,n_0)}{g_0(s,n_0)} \right) \}$, которая совпадает с (3.3), где минимизация производиться по всем позитивным s таким, что t_0, определенная в (3.19), меньше единицы.

Теперь рассмотрим вероятность того, что вес обобщенного синдрома не более $\alpha W \ell$ для любой комбинации стираний заданной кратности W. Если эта вероятность менее единицы, тогда существуют коды из ансамбля $\chi(n_0, \ell, b)$, для которых вес обобщенного синдрома достигает значения $\alpha W \ell$ для любой комбинации стираний кратности W. Таким образом, существование таких кодов гарантируется, если

$$\binom{n}{W} P(|\mathbf{S}| \le \alpha W \ell) < 1$$

Беря логарифм и используя неравенства (3.22) и (3.23), получим

$$h(\omega) - \ell F(\alpha, \omega, n_0) < 0 \quad (3.24)$$

Наибольшее значение ω, удовлетворяющее (3.24), обозначим как ω_α.

Наконец, из *Леммы 1* следует, для любого $\alpha > 0$ алгоритм Λ исправляет $\omega_\alpha n$ стираний со сложностью $O(n \log n)$, что завершает доказательство *Теоремы 1*.

Представленные выше формулы (3.1) и *(3.2)* позволяют получить численную оценку наибольшей кратности исправимой комбинации стираний для различных значений параметров кодов. Были вычислены значения ω_0 по (3.1) и ω_α по *(3.2)* при $\alpha = 10^{-4}$ в зависимость от скорости R LDPC – кода для количества слоев $\ell = 5$ и $\ell = 11$ (см. Рис.1). Как видно из графиков при малых значениях скорости R ω_α больше ω_0,

а при увеличении R становятся бесконечно близки, при этом ω_α не меньше ω_0. Так же стоит отметить, что при увеличении числа слоев, разница между ω_α и ω_0 уменьшается.

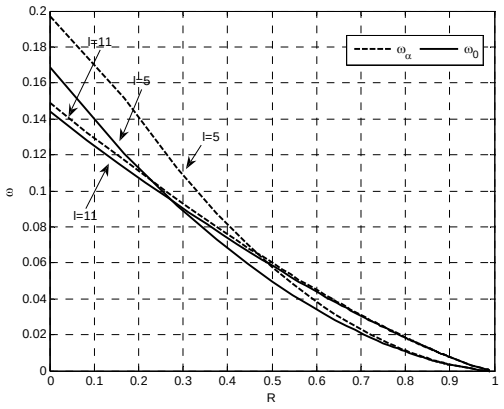


Рис.1. Кратность исправимой комбинации стираний, выраженная в долях от длины кода n, в зависимости от скорости кода R для $\ell = 5$ и $\ell = 11$

Таким образом, представленный в данной статье метод оценки кратности исправимой комбинации стираний улучшает ее при малых скоростях LDPC – кода.

4. Литература

[1] R. G. Gallager, “Low-Density Parity-Check Codes”, *MIT Press*, Cambridge, MA, USA, 1963.

[2] В.В. Зяблов, М.С. Пинскер, “Сложность декодирования низкоплотностных кодов при передаче по каналу со стираниями”, *Проблемы передачи информации*, т. X, вып. 1, стр. 15 – 28, янв. – март 1974г.

[3] V. V. Zyablov and M. S. Pinsker, “Estimation of the error-correction complexity for Gallager low-density codes,” *Problems of Inform. Transmission*, vol. 11, no. 1, pp. 23–26, Jan.–March 1975.

Адаптивная подстройка параметров кодирования речевого сигнала в системах IP-телефонии

Иманкулов Р.А.
Уральский государственный технический университет
roman@netangels.ru

Аннотация

Предложен способ оценки качества речевого сигнала в системах IP-телефонии, с помощью R-фактора. Представлены результаты эксперимента, определяющего коэффициенты для вычисления R-фактора для различных режимов кодека Speex. Приведено описание адаптивного поискового алгоритма, выполняющего в сетях IP-телефонии подстройку параметров кодирующего устройства в зависимости от характеристик канала связи. Представлены результаты испытаний алгоритма на каналах с ограниченной полосой пропускания.

1. Введение

IP-телефония – технология передачи речевого трафика по сетям IP, получает все большее распространение, постепенно вытесняя традиционную телефонию.

Однако необходимо отметить, что сети IP изначально не планировались для передачи данных в режиме реального времени, и потому не могут гарантировать требуемое оконечным устройством качество канала связи. Функционирование таких сетей в большинстве реальных каналах связи осуществляется по принципу *best efforts*. В соответствии с этим принципом, сетевая инфраструктура не может гарантировать обеспечение каких-либо наперед заданных значений качества, хотя и гарантирует, что будут приложены все усилия для того, чтобы передать данные с наименьшими задержками и потерями. В то же время для приложений IP-телефонии задержки и потери в сети являются параметрами, критичными для обеспечения качественной связи. В таких условиях возможным вариантом обеспечения приемлемого качества передачи речи представляется адаптивное управление параметрами кодирования и передачи речевого сигнала.

Возможность адаптивного управления качеством голоса в сетях IP реализуется динамическим изменением параметров кодирования речи оконечными устройствами в зависимости от текущих характеристик канала связи. Для этого необходим способ сбора статистической информации о канале связи, алгоритм принятия решение об изменении параметров кодирования и способ изменения этих параметров одновременно на передающей и приемной стороне.

В простейшем случае характеристики канала не зависят от параметров кодирования, и задача подстройки сводится к установке параметров кодирования, обеспечивающих максимальное качество кодирования речи в условиях идеального канала. Более сложным и интересным для практического применения является вариант, при котором характеристики канала связи изменяются под воздействием нагрузки. Типичным вариантом такого канала являются каналы с ограниченной полосой пропускания. Именно работе с такими каналами посвящена настоящая работа. Алгоритм адаптации реализован для кодека Speex [1], свободного от патентных ограничений и имеющего широкие возможности для адаптации. Полученные результаты могут быть применены и для других типов каналов и кодеков.

2. Оценка качества восстановленного сигнала

Качество восстановленного речевого сигнала в режиме реального времени может оцениваться только по косвенным признакам — потерям и задержкам пакетов, возникающим в процессе передачи информации. Основной способ оценки качества восстановленного речевого сигнала, подвергнутого искажениям, заключается в использовании метрик MOS (Mean Opinion Score, средних экспертных оценок), описанный в рекомендации [2]. Однако данный метод не подходит для оценки качества в режиме

реального времени. Вместо этого для таких оценок может быть использована E-модель, описанная в рекомендации [3]. E-модель основывается на принципе независимого влияния различных ухудшающих факторов (параметров канала, качества кодирования) на суммарное ухудшение речевого сигнала.

Системы IP-телефонии значительно отличаются по характеристикам от традиционных систем передачи речи с коммутацией каналов, что требует особого подхода при подсчете качества с использованием E-модели. Учитывая этот факт, были предложены исправленные и упрощенные оценки качества с использованием E-модели, предназначенные для систем IP-телефонии [4]. Проведенные различными авторами исследования показали, что упрощенная формула для подсчета фактора рейтинга (R-фактора) может быть представлена в следующем виде:

$$Q = 93{,}4 - Id(Ta) - Ie(codec,loss), \quad (1)$$

где:

Ta – абсолютная односторонняя задержка

Id – ухудшение качества, вызванное задержкой

Ie – потери качества при кодировании, а также из-за потерь пакетов в канале связи

Параметр *Id* при отсутствии эффекта эха полностью определяется односторонней задержкой распространения сигнала *Ta* и его вычисление не представляет принципиальной сложности.

Параметр *Ie* требует экспериментального определения. Для ряда наиболее популярных кодеков такие измерения проведены [5], в то же время для Speex такая работа ранее выполнена не была. Необходимо отметить, что на параметр *Ie* оказывает влияние как ухудшение качества, связанное с кодированием Speex, так и ухудшение качества за счет потерь в канале связи.

План эксперимента для оценки параметра Ie заключался в выполнении следующей последовательности действий:

- для заданного набора режимов кодирования Speex и заданного набора уровня потерь выполняется кодирование, а затем декодирование набора речевых файлов (всего было протестировано 20 звуковых файлов длиной от 15 до 120 секунд);
- полученные в результате декодирования звуковые фрагменты сравниваются с оригинальными по алгоритму PESQ, выставляющему оценку качества в пределах от 0 до 5.
- оценка MOS переводится в R-фактор с использованием формулы, приведенной в [3];
- принимая значение *Id* равным нулю, вычисляется *Ie* для каждого набора параметров кодера и канала.

Результаты измерения параметра Ie приведены в таблице 1.

Таблица 1. Значение *Ie* для различных характеристик кодера и канала

Режим кодера	Уровень потерь (%)						
	0	1	2	3	4	8	16
1	53, 5	53, 9	54, 2	54, 0	54, 2	55, 0	56, 5
8	44, 0	44, 4	44, 7	45, 1	45, 5	47, 1	50, 3
2	33, 0	34, 5	35, 6	36, 8	38, 1	41, 5	46, 6
3	27, 9	30, 3	31, 9	33, 6	35, 4	39, 7	45, 6
4	23, 1	26, 4	28, 5	30, 9	33, 0	37, 7	43, 9
5	21, 2	24, 8	27, 2	29, 8	32, 0	36, 8	43, 0
6	17, 9	22, 7	25, 3	28, 0	30, 6	35, 9	42, 9
7	15, 4	20, 9	24, 2	27, 3	29, 9	35, 5	42, 4

3. Управляемые параметры кодера

Основная задача адаптивного алгоритма состоит в том, чтобы найти режим кодирования, обеспечивающий наилучшее качество речевого сигнала на приемной стороне при данных условиях канала. Алгоритм адаптации имеет возможность управлять следующими параметрами кодера:

- коэффициент пакетизации *P* (положительное целое число);
- режим повторов *R* (целое число, принимающее значение из множества {0; 1});
- режим кодирования *S* кодека Speex (целое число от 1 до 8 включительно).

Коэффициент пакетизации определяет количество фреймов кодера в одном пакете RTP. Особенность сетей IP в применении к передаче речевого трафика заключается в высоких накладных расходах, возникающих при передаче. Эти расходы формируются за счет добавления к передаваемому речевому фрагменту заголовков канального уровня, IP, UDP и RTP. В связи с этим оказывается целесообразным передавать несколько речевых кадров в одном пакете.

Увеличение значения параметра *P* снижает суммарную нагрузку на сеть, и тем самым способно уменьшить потери и задержки, возникающие при передаче. С другой стороны, это увеличивает общую задержку, что приводит к противоположному эффекту — ухудшению качества речи на стороне слушателя.

Включение режима повторов приводит к тому, что каждый речевой фрагмент передается по каналу связи дважды: в первый раз это происходит в момент формирования, и во второй раз — с небольшой задержкой. На некоторых каналах связи такой подход, который можно назвать «продавливанием данных», позволяет снизить среднее количество потерь, и уменьшить корреляцию между потерями. Однако повышенная нагрузка на канал и в этом случае может негативно сказаться на его характеристиках.

Параметр S задает режим работы кодека Speex. Как и большинство современных кодеков, Speex использует алгоритм сжатия информации с потерями. Режим кодирования определяет степень сжатия, а также потери, которые возникнут в восстановленном речевом сигнале. Большее значение соответствует меньшему сжатию и лучшему качеству восстановленного сигнала. Однако, как и в предыдущих случаях, на узкополосных каналах большой битрейт может спровоцировать дополнительные потери и задержки.

Получается, что изменение любого из параметров может привести как к улучшению качества сигнала, так и к его ухудшению. Эта зависимость определяется свойствами самого канала, и потому невозможно заранее выбрать режим кодирования, оптимальный для всех случаев.

4. Описание алгоритма адаптации

Алгоритм имеет два режима работы — *режим поиска* и *режим наблюдения*.

Работая в *режиме поиска*, алгоритм исследует канал и подбирает оптимальные значения кодирования. В основу адаптации был положен алгоритм покоординатного спуска. Определение качества речевого сигнала производится с использованием R-фактора. Из начального состояния кодека, характеризуемого тройкой параметров (P_0, R_0, S_0), производится поиск максимума R-фактора вдоль направления оси P и определяется точка (P_1, R_0, S_0), в которой касательная к линии уровня параллельна оси P. Затем производится поиск максимума из этой точки в направлении оси R и определяется следующее состояние (P_1, R_1, S_0). Аналогично определяется максимум по координате S. Найденный таким образом режим считается оптимальным для данного канала, оптимальное значение качества Q_{opt} фиксируется, и алгоритм переходит в следующий режим.

В *режиме наблюдения* производится мониторинг качества Q и в каждый момент времени вычисляется $\Delta Q=|Q-Q_{opt}|$. При превышении значением ΔQ определенного порога выполняется переход в режим поиска, на этот раз подстройка режима кодирования выполняется из ранее найденной точки оптимума (P_1, R_1, S_1).

Такой подход позволяет алгоритму показывать удовлетворительные результаты на каналах, характеристики которых изменяются в пределах длительности телефонного разговора.

5. Тестирования алгоритма на каналах с ограниченной полосой

Для оценки возможностей реализованного алгоритма на тестовом стенде были поставлены эксперименты, цель которых заключалась в том, чтобы оценить возможное улучшение качества речевого сигнала за счет использования адаптации.

Сравнение выполнялось на стационарных каналах с заранее установленной полосой пропускания и долей случайных потерь пакетов. Выходное качество сигнала определялось алгоритмом PESQ [5]. Оценка предельных возможностей адаптации производилась многократной передачей фиксированного звукового файла, закодированного с использованием всех возможных комбинаций параметров кодера и способа передачи сигнала. При передаче сигнала варьировались число фреймов в RTP пакете, дублирование пакетов и режим кодирования Speex.

Изучалось поведение системы в диапазоне полос от 1000 до 7200 байт/с и потерь от 0 до 20%. Для иллюстрации поведения системы адаптации ниже приведены данные для 6% потерь.

Из приведенных графиков видно, что использование адаптивного алгоритма при 6% потерь обеспечивает приемлемое качество речевого сигнала в полосе от 2000 байт/с (красная линия на графике). В то же время использование кодирования с параметрами по умолчанию обеспечивает в данных условиях приемлемое качество, начиная с полосы 3800 байт/с и выше (зеленая линия на графике).

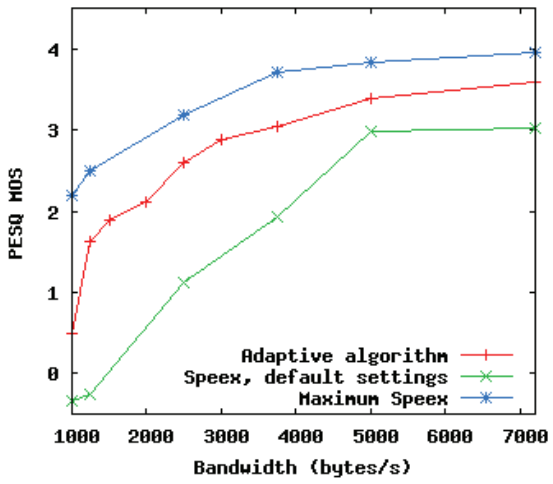


Рисунок 1. Результаты экспериментальной оценки качества речевого сигнала

Несмотря на то, что реализованный адаптивный алгоритм позволяет весьма существенно повысить качество восстановленного речевого сигнала, его сравнение с гипотетическим кодированием в наилучшем режиме (синяя полоса) позволяет говорить о том, что существуют дополнительные возможности улучшения качества адаптации.

[1] Valin, J. M. “The Speex Codec Manual”, *Speex documentation*, 2007, http://www.speex.org/docs/.

[2] P.800. ITU-T Recommendation *P.800: Methods for objective and subjective assessment of quality*, ITU-T Geneva, 1996.

[3] G.107. ITU-T Recommendation *G.107 E-модель - вычислительная модель, используемая для планирования передачи* ITU-T Geneva, 2005.

[4] L. C. G. Lustosa, L. S. G. Carvalho, P. H. A. Rodrigues and E. S. Mota, “*E-Model Utilization For Speech Quality Evaluation Over VoIP-Based Communication Systems”*, 22nd SBRC, 2004, pp. 933-938

[5] G.108. ITU-T Recommendation G.108: *Application of the E-model: A planning guide*, ITU-T Geneva, 1999.

Оценивание характеристик канала в системе множественного доступа, основанной на перестроении ортогональных частот.

Д.С.Осипов
Институт проблем передачи информации им. А.А. Харкевича РАН
d_osipov@iitp.ru

Аннотация.

Рассматривается система множественного доступа, основанная на перестроении ортогональных частот, отличающаяся от систем такого типа рассматривавшихся ранее. Задачей настоящей работы является разработка процедуры оценивания характеристик "восходящего" канала в вышеупомянутой системе и исследование ее свойств. Предлагаемая многоступенчатая процедура характеризуется относительно низкой сложностью и не требует введения специального режима передачи данных для целей оценивания, что позволяет использовать ее как в режиме "вхождение в связь" при накоплении первичных оценок, так и в режиме "передача данных" для улучшения оценок после декодирования. Эффективность предложенной процедуры оценивания подтверждается результатами моделирования.

Введение.

В настоящее время вследствие активного развития телекоммуникаций и ограниченности доступных пользователям ресурсов одним из наиболее активно развивающихся направлений исследований в области связи являются методы разделения пользователей в системах множественного доступа. Другой важнейшей задачей часто возникающей в современных системах беспроводной связи является защита информации и обеспечение надежной передачи данных т.е. защита от намеренного прослушивания и враждебного шумоподавления.

Кодовые методы разделения пользователей в системах множественного доступа позволяют эффективно решать обе вышеупомянутые задачи. В настоящей работе будет рассмотрена система с кодовым разделением множественного доступа, а именно система, основанная на перестроении ортогональных частот, отличающаяся от систем такого типа рассматривавшихся ранее (например от системы описанной в [1]).

Целью настоящей работы является разработка процедуры оценивания для такой системы и исследование показателей качества оценивания, которое обеспечивает использование предлагаемой процедуры. Так как предлагаемая процедура оценивания тесно связана с методом передачи и приема используемым в данной системе эта последняя будет описана ниже чуть более подробно.

Система множественного доступа, основанная на перестроении ортогональных частот: описание системы.

Рассматривается система множественного доступа без управления, в которой K активных пользователей одновременно осуществляют передачу на базовую станцию по каналу разбитому при помощи технологии OFDM на Q подканалов. Последнее, как уже было сказано, означает, что передача является асинхронной и некоординированой (т.е. ни один из пользователей не располагает информацией о других пользователях. Мы будем предполагать, что все пользователи ведут передачу, руководствуясь одними и теми же принципами. Кроме того, при описании функционирования каждого из пользователей в рассматриваемой системе мы будем различать два режима: “вхождение в связь” и “передача данных”.

Каждый из активных пользователей оснащен генератором номеров подканалов, который псевдослучайно выбирает из всего множества Q подканалов q подканалов. Базовая станция также оснащена K_0 аналогичными генераторами (по числу активных пользователей, которые могут одновременно работать в описываемой

системе). На этапе “вхождение в связь” происходит синхронизация генераторов. Последнее означает, что генератору каждого из активных пользователей соответствует один из генераторов на базовой станции, который при приеме каждого символа от соответствующего пользователя вырабатывает те же номера подканалов, которые вырабатывались генератором номеров этого активного пользователя при передаче соответствующего символа.

Предметом нашего рассмотрения будет режим передачи данных. Предполагается, что все активные пользователи передают кортежи (т.е. двоичные вектора) длины c ($c = c_c + c_s$). В настоящей работе нами будет без ограничения общности рассмотрена система с $c_c = 2$ и $c_s = 2$. Информационные биты делятся на две последовательности. Каждая из двух последовательностей кодируется (независимо) внешним плетеным турбо кодом в соответствующее кодовое слово. Биты первого кодового слова отображаются в номер подканала. Биты второго кодового слова отображаются в соответствии с выбранным типом модуляции в соответствующие амплитудные уровни. В настоящей работе мы используем вложенную амплитудную модуляцию. Ниже приведена схема используемого отображения:

00 ® 2
01 ® 1
11 ® - 1
10 ® - 2

Для того, чтобы добиться равномерного распределения ошибок по длине кодового слова биты, кодирующие амплитудный уровень, необходимо дополнительно переставлять таким образом, чтобы распределение ошибок по кодовым словам компонентных кодов было примерно равномерным. На приемном конце необходимо выполнить обратную перестановку.

Заметим, что, так как все пользователи выбирают наборы подканалов случайным образом и независимо друг от друга возможна ситуация, в которой в некоторый момент времени один и тот же подканал будет использоваться более чем одним пользователем. Такую ситуацию мы будем называть “коллизией”.

Как уже было сказано выше, базовой станции известны наборы подканалов, выбранные генератором активного пользователя. При этом, неизвестно какой именно из подканалов использовался для передачи. Для того, чтобы разрешить эту неопределенность используется многоступенчатая процедура принятия решения о битах предаваемой последовательности. Так как настоящая работа не посвящена описанию собственно процедуры принятия решения здесь будет приведено лишь краткое описание этой процедуры. Используя измерения сигнала на выходе подканала базовая станция рассчитывает проекции сигналов по каждому из подканалов на направление, задаваемое фазой переданного сигнала и вычисляет отношения правдоподобия для каждого из битов подкортежей, кодирующих номер подканала. Вычисленные значения поступают на вход декодера первого кода. Затем, используя новые оценки отношений правдоподобия, полученные после заранее заданного числа итераций декодирования, базовая станция принимает решение о номерах подканалов, которые использовались для передачи и (используя измерения на выходе этих подканалов) вычисляет отношения правдоподобия для каждого из битов подкортежей, кодирующих амплитудные уровни. Вычисленные значения поступают на вход декодера второго кода. Затем, используя новые оценки отношений правдоподобия, полученные после заранее заданного числа итераций декодирования, базовая станция принимает решение о битах вторых подкортежей. Принятие решения о битах вторых подкортежей завершает процедуру принятия решений.

Как видно из вышесказанного, принятый в описываемой системе метод приема требует точного знания частотных коэффициентов. Особенно важным является верное определение фазы т к. ошибка по фазе предопределяет точность вычисления проекции. Ниже будет приведено описание основных шагов предлагаемой процедуры оценивания.

Процедура оценивания: вычисление первичных оценок и выбор пилотных последовательностей.

В настоящей работе для оценивания характеристик канала мы процедуру, предполагающую использование известных (заранее заданных или полученных в результате декодирования) приемнику последовательностей.

Использование известных приемнику последовательностей позволяет применить для оценивания канала такие методы, как метод наименьших квадратов и метод максимума правдоподобия. Так как вычислительная

сложность этих методов велика (в основном, из-за необходимость решать задачу многомерной оптимизации или вычислять обратные матрицы больших размеров), большая часть исследований алгоритмов оценивания посвящены снижению сложности процедуры за счет дополнительных предположений и использованию связанных с этими предположениями аппроксимаций. Так, например, в работе [2] рассматривается техника оценивания на базе субоптимальной версии метода максимума правдоподобия, основанной на приближении обратной матрицы первым членом ее разложения в ряд (этот член может быть вычислен заранее) к которому она асимптотически стремится с увеличением числа ортогональных частот. В работе [3] для той же цели (разработки алгоритма оценивания канала со сложностью меньшей, чем у классического ML) предлагается использовать метод меняющихся направлений.

В настоящей работе будет использован более простой метод. Оценки компонент частотного отклика будем находить по формуле:

$$\hat{H}_{i,k} = \frac{Y_k}{v_{i,k}} = H_{i,k} + \underbrace{\frac{\sum_{m=1}^{J} H_{m,k} \cdot v_{m,k} \cdot \tau_m}{v_{i,k}}}_{\breve{H}_k} + \underbrace{\frac{\zeta_k}{v_{i,k}}}_{\breve{\breve{H}}_k} \quad (2)$$

здесь $\hat{H}_{i,k}$ -оценка коэффициента передачи k-ого подканала восходящего канала i-ого пользователя, Y_k - выход k-ого подканала, $v_{i,k}$ - символ пилотной последовательности переданный i-ым пользователем по k-ому подканалу, $H_{i,k}$ - коэффициент передачи k-ого подканала восходящего канала i-ого пользователя, τ_m - относительная задержка m-ого пользователя, $\breve{H}_k$ - искажение оценки за счет влияния "мешающих" пользователей, $\breve{\breve{H}}_k$ - искажение оценки за счет влияния фонового шума.

Уравнение (2) может быть использовано для оценивания канала для описания восходящего канала в системе без управления (например, в рассматриваемой нами системе множественного доступа, основанной на перестроении ортогональных частот).В рассматриваемой системе мы будем использовать его двояко: в режиме "установление связи" для предварительного оценивания частотных характеристик (в этом случае на базовую станцию передаются заранее известные "пилотные" последовательности), а в режиме "передача данных" для последующей коррекции частотных характеристик (в этом случае, в качестве известных последовательностей используются декодированные кодовые слова). Следует отметить, что применение любого механизма оценивания с использованием известных последовательностей возможно лишь в том случае, если предполагается, что частотная характеристика не меняется в течение всего времени, за которое передается последовательность. Для описанной выше системы это означает, что частотная характеристика не меняется в течение времени, равного продолжительности стольких OFDM блоков, сколько символов в пилотной последовательности.

В режиме "установление связи" на базовую станцию передаются пилотные последовательности. Используя имеющуюся у нее информацию о переданной некоторым пользователем тестовой последовательности и уравнение (2), базовая станция оценивает частотную характеристику канала. Следует отметить, что в системе множественного доступа использующей канал мобильной связи, вследствие особенностей распространения сигнала, частотные характеристики всех восходящих и нисходящих каналов различны. Это обстоятельство, а также различие моделей передачи от базовой станции к клиентам и от клиента к базовой станции приводят к существенным различиям в оценивании восходящего и нисходящего каналов. Передача от базовой станции, как уже было сказано выше, представляет собой широковещательную рассылку, организованную по принципу "от одного ко многим", поэтому для оценивания всех каналов "вниз" каждым из пользователей может использоваться одна и та же пилотная последовательность. Напротив, базовая станция должна оценивать столько различных восходящих каналов, сколько активных пользователей в системе, причем, если несколько пользователей передают тестовые последовательности, базовая станция не может использовать для восстановления частотной характеристики одного из них сигналы, пришедшие от прочих.

Процедура выбора пилотных последовательностей предопределяет вид распределения последней компоненты в выражении (2). Если все символы алфавита модуляции имеют равную амплитуду (например, если используется фазовая модуляция), оценки компонент частотного отклика, вычисленные в соответствии с (2), будут гауссовскими с одинаковыми дисперсиями. В случае если

алфавит модуляции состоит из символов с различной амплитудой (например, если используется 4-ASK, как в описанной выше системе), оценивание желательно намеренно выбирать в качестве символов тестовой последовательности символы с равными амплитудами (в противном случае дисперсии оценок частотного отклика окажутся различными).

Для того, чтобы избежать введения специализированного протокола передачи пилотных последовательностей, в описанной нами системе множественного доступа для передачи пилотных последовательностей мы используем адаптированную версию метода используемого для передачи данных. А именно, будем передавать последовательности той же длины, что и кодовые последовательности, используемые в режиме передачи данных, состоящие из символов с максимальной амплитудой (иными словами символы могут иметь значение 2 и -2) и передаваемых по подканалам, номера которых выбираются псевдослучайно (для этого достаточно дополнить случайные последовательности, вырабатываемые генераторами номеров подканалов, последовательностями, элементы которых выбираются из списка {1:q} и предопределяют выбор одного из q элементов вектора выданного генератором номеров подканалов). Базовая станция принимает сигнал, соответствующий некоторой пилотной последовательности, и, пользуясь соотношением (2), вычисляет оценки соответствующих компонент частотного отклика. Так как данный метод оценивания использует случайный выбор подканалов, по некоторым подканалам за время передачи пилотной последовательности может быть передано более одного символа. В этом случае оценка для данного подканала усредняется по всем переданным в нем символам. Ниже будут описаны методы коррекции оценок, позволяющие заметно повысить точность оценивания.

Коррекция полученных оценок.

Как видно из уравнения (2), оценки частотных характеристик, полученные в режиме “установление связи”, искажены аддитивными помехами, как от фонового шума, так и от взаимодействия с “мешающими” пользователями.

Для того, чтобы устранить выбросы, появление которых объясняется влиянием “мешающих” пользователей и заметно ухудшает точность оценивания, мы воспользуемся сравнительно простым методом, который использует информацию о характере распределения отсчетов аддитивного шума. Так как аддитивный шум распределен по гауссову закону, можно утверждать, что его отсчеты подчиняются известному “правилу трех сигма”, согласно которому вероятность того, что абсолютная величина отклонения окажется больше 3σ (где σ -среднеквадратическое отклонение) есть $p_d = 0,0027$. Поэтому можно утверждать, что в случае, если отсчеты, отклонение которых от среднего превышает по абсолютной величине утроенное среднеквадратическое, будут отброшены (исключены из набора измерений) вероятность того, что будут отброшены результаты корректных измерений есть p_d (т.е. весьма мала), а вероятность того, что отброшенные измерения не подчиняются нормальному распределению с заданными параметрами и следовательно искажены влиянием прочих пользователей, есть $1- p_d$ (т.е. напротив велика). Для того, чтобы воспользоваться этим правилом, необходимо выделить составляющую, соответствующую истинной частотной характеристике канала, иными словами, найти оценку среднего значения для каждого из измерений. Простейшим способом вычислить такую оценку является усреднение в скользящем окне (точность оценки полученной таким способом невелика, однако этот метод характеризуется низкой сложностью). Этот метод хорошо работает в том случае, если значения характеристики в пределах окна отличаются незначительно, кроме того, рассматриваемый числовой ряд содержит значения коэффициентов передачи лишь для части подканалов, поэтому для длины окна следует выбирать малые значения. Выбросы могут заметно исказить оценку, полученную при усреднении. Поэтому вместо среднего значения мы будем усеченное среднее, так как оно устойчиво к выбросам. Полученная в результате такой фильтрации оценка истинной частотной характеристики канала позволяет вычислить значения соответствующие аддитивной помехи. Этот числовой ряд, как и исходные измерения, содержит выбросы, которые искажают величину среднеквадратического отклонения. Поэтому для того, чтобы измерить его более точно необходимо вычислять межквартильный размах, т.е. разницу между 75 и 25 процентилем. Межквартильный размах обладает значительно большей устойчивостью к выбросам, чем среднеквадратическое отклонение, величина которого связана с межквартильным размахом соотношением $\sigma = 0.7431 \cdot IQR$. Заметим, что предлагаемая процедура устранения выбросов

производится отдельно для действительной и мнимой компонент частотного отклика.

Для того, чтобы уменьшить искажение воспользуемся процедурой коррекции по импульсному отклику. Для этого выполним обратное преобразование Фурье, и переведем вычисленную нами оценку частотного отклика в оценку импульсного отклика. Истинный импульсный отклик имеет конечную длину. Поэтому можно утверждать, что, начиная с некоторого отсчета, номер которого определяется максимальной длиной импульсного отклика, все ненулевые отсчеты обратного преобразования Фурье полученной нами оценки частотного отклика порождаются влиянием искажений различной природы (недостаток точности процедуры оценивания, фоновый шум, влияние других пользователей) и могут быть отброшены. Такое отбрасывание лишь улучшает качество оценивания, делая оценку импульсного отклика более адекватной истинному импульсному отклику, поэтому мы будем называть эту процедуру “коррекцией импульсного отклика”.

Следует отметить, что того, что номера подканалов в пилотных последовательностях назначаются случайно, коэффициенты передачи некоторых подканалов остаются неизвестными. Для того, чтобы воспользоваться описанной выше процедурой коррекции импульсного отклика необходимо использовать процедуру экстраполяции. В настоящей работе мы будем использовать линейную экстраполяцию, так как она характеризуется низкой сложностью.

Таким образом, процедура оценивания в самом общем виде выглядит следующим образом: базовая станция вычисляет оценки тех компонент частотного отклика, которым соответствуют символы принятой пилотной последовательности, при помощи процедуры устранения выбросов отбрасывает те оценки, которые с высокой степенью вероятности можно считать недостоверными, при помощи процедуры экстраполяции восстанавливает остальные компоненты частотного отклика, при помощи обратного преобразования Фурье преобразует оценку частотного отклика в оценку импульсного отклика и, отбрасывает отсчеты, с номером большим, чем максимальная длина импульсного отклика.

Моделирование.

Так как при приеме в вышеописанной системе множественного доступа используются коэффициенты передачи подканалов, т.е. компоненты частотного отклика, для того чтобы охарактеризовать качество предлагаемой процедуры оценивания, переведем полученные оценки в частотную область и будем измерять ошибку оценивания как по фазе, так и по амплитуде.

Так как выбор подканала является случайным уместно говорить о величине ошибки как по фазе так и по амплитуде как о случайной величине. Поэтому в дальнейшем мы будем говорить о вероятности того, что абсолютная величина ошибки по фазе или по амплитуде окажется больше чем некоторое пороговое значение (мы будем называть эту вероятность “вероятностью превышения”).

Ниже, в качестве примера, приведены соответствующие графики зависимостей, полученные для гористой местности и следующих параметров число активных пользователей K=50, общее число подканалов $Q = 4096$, отношение сигнал/шум 20 дБ, длина пилотной последовательности 300 символов.

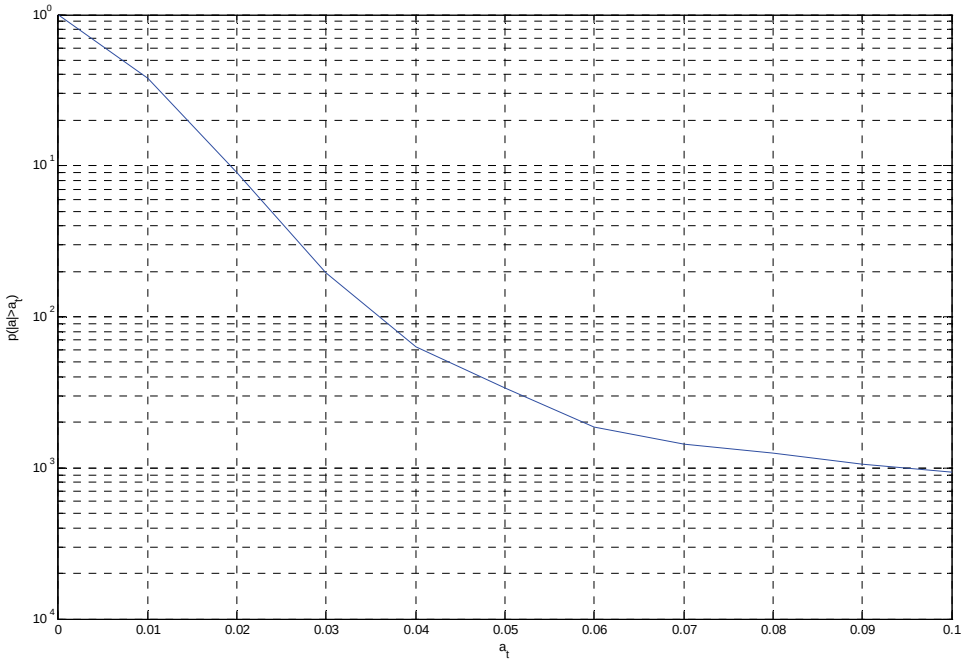


Рис. 1 Вероятность превышения порогового значения ошибки по амплитуде

Как видно из рис. 1 абсолютная ошибка оценивания по амплитуде относительно мала (заметим, что абсолютное значение истинной характеристики не превышает 2). Для нас однако, больший интерес представляет ошибка по фазе так как значительная ошибка по фазе приводит к ошибкам при том способе приема, который предлагается использовать в рассматриваемой системе.

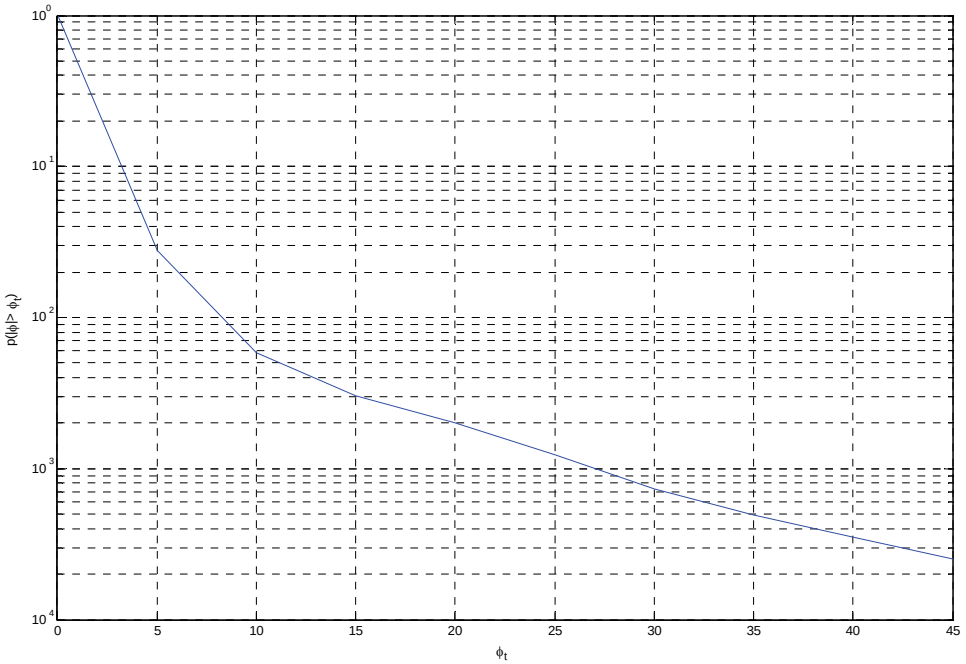


Рис. 2 Вероятность превышения порогового значения ошибки по фазе (в зависимости от величины порога в градусах)

Как видно из Рис. 2, вероятность того, что ошибка оценивания по фазе при использовании предложенной процедуры окажется большой весьма мала. В частности вероятность того, что абсолютное значение ошибки по фазе окажется больше, чем критическое для выбранного типа модуляции значение (т.е. 45 градусов) оказывается меньше, чем 10^{-3}, а распределение ошибки по фазе близко к нормальному с нулевым средним. При этом вероятность того, что абсолютное значение ошибки по фазе окажется больше, чем 10 градусов меньше чем 10^{-2} (заметим, что входные вероятности ошибки по подканалу оказываются в несколько раз больше). Это в частности позволяет сделать вывод о том что предлагаемая процедура может с успехом использоваться в системе описанного типа.

Заключение.

Как видно из результатов моделирования приведенных выше, использование предложенной выше процедуры оценивания в системе описанного типа при значениях отношения сигнал/шум близких к тем, которые достигаются в реальных системах, обеспечивает сравнительно малую величину как абсолютной ошибки по амплитуде, так и ошибки по фазе. Вероятность ошибочного приема вследствие погрешности оценивания также оказывается сравнительно малой, что свидетельствует о том, что предложенная процедура может с успехом использоваться для оценивания в системе множественного доступа, построенной в соответствии с описанными выше принципами.

Библиография.

[1] Nogueroles R. Performance Considerations of OFDMA Radio Communication Systems: Ph. D. Thesis Ulm: Universitat Ulm, 1998.

[2] S. Sezginer, P. Bianchi “Asymptotically Efficient Reduced Complexity Frequency Offset and Channel Estimators for Uplink MIMO-OFDMA Systems” to appear in IEEE Trans. on Signal Processing

[3] M. Pun, S.-H Tsai and C.-C. Jay Kuo "Joint Maximum Likelihood Estimation of Carrier Frequency Offset and Channel in Uplink OFDMA Systems" in proc. 47th annual IEEE Global Telecommunications Conference. Globecom 2004 p 3748-3752

Сравнительный анализ помехоустойчивости плетёных свёрточных (woven) и турбо кодов

Чеклин А.Б., Черкасов М.В., Адов А.А.
Институт проблем передачи информации им. А.А. Харкевича РАН, Москва
cheklin@iitp.ru

Аннотация

В докладе рассматриваются каскадные кодовые конструкции, описанные в литературе как турбо коды и плетеные (woven) коды. Обе эти конструкции строятся с использованием двоичных систематических сверточных кодов в качестве внешних и внутренних кодов каскадных схем кодирования. При построении этих кодов используются составляющие их сверточные коды со скоростями передачи от 1/4 до 8/9 и памятью от 2 до 6. Получающиеся результирующие турбо и плетеные коды имеют скорости передачи от 1/16 до 64/81. Декодирование каскадных конструкций основано на итеративном принципе с использованием оптимальных декодеров по максимуму апостериорной вероятности с «мягкими» входами и выходами для внешних и внутренних кодов. В качестве перемежителей (интерливеров) применяются различные перемежители европейского и американского стандартов.

В докладе представляются результаты компьютерного моделирования описанных выше каскадных схем в гауссовском и релеевском каналах. Для моделирования использовалась система компьютерного исследования конструкций плетеных сверточных и турбо кодов – «Code». Показывается, что широко применяемым сейчас турбо кодам во многих практически важных ситуациях хорошую альтернативу составляют плетеные сверточные коды.

1. Введение

Своевременность и качество доставки информации во многом обусловлены техническими характеристиками средств связи. Большой вклад в решении задачи улучшения их основных характеристик призвана внести теория помехоустойчивого кодирования. Использование методов помехоустойчивого кодирования в каналах связи позволяет увеличить скорость передачи информации и повысить верность передачи данных.

Одним из наиболее эффективных методов увеличения скорости передачи больших объемов данных при одновременном повышении качества их передачи по каналам связи является помехоустойчивое кодирование. В современных системах связи предпочтение отдается каскадным сверточным конструкциям. Наиболее широкое применение находят плетеные сверточные (woven) коды и турбо коды. Однако проектирование этих систем кодирования требует от разработчиков глубоких специальных знаний не только в области помехоустойчивого кодирования, но и в ряде специальных разделов математики. Избежать указанных трудностей при проектировании современных систем передачи информации позволяет программный комплекс "Code". Программный комплекс предназначен для разработки и моделирования каскадных систем помехоустойчивого кодирования на базе сверточных кодов. Он представляет собой современный инструмент проектирования систем связи с помехоустойчивым кодированием, который позволяет специалистам-разработчикам систем связи создавать нужные им корректирующие коды, а специалистам-исследователям при минимальных затратах разрабатывать и проводить сравнительный анализ различных перспективных систем кодирования.

2. Модель

При работе на программном комплексе можно реализовать три основных типа проектов: «Convolutional code», «Turbo code», «Woven turbo code». Предусмотрена работа с любым сверточным кодом, выбираемым из базы данных, содержащей сверточные коды со скоростями передачи от 1/3 до 8/9 и памятью 2, 3 и 4. В каждом из этих проектов пользователь может использовать любой сверточный код из этой базы данных. В проектах «Turbo code» и «Woven turbo code» эти сверточные коды используются в

качестве составляющих кодов каскадных конструкций. Комплекс позволяет получать каскадные коды со скоростью передачи от 1/6 до 4/5. Кроме того, для пользователя в программе предусмотрена база данных с большим числом интерливеров, среди которых имеются интерливеры как европейского, так и американского стандартов.

Двоичным плетеным сверточным (woven) кодом является комбинация нескольких сверточных кодов, причем результат комбинации тоже является сверточным кодом.

Плетеные сверточные (woven) турбо коды (WTC): кодер woven турбо кода состоит из l_0 «внешних» сверточных кодеров и одного внутреннего сверточного кодера (рисунок 1).

Информационная последовательность u разбивается на l_0 подпоследовательностей, которые подаются на вход l_0 «внешних» кодеров со скоростью $R^0 = b^0 / c^0$. Часть символов кодовых последовательностей ($v_l^{o,(1)}$), которые находятся в одних и тех же позициях для всех «внешних» кодеров, мультиплексируется и образует последовательность u^i. Последовательность u^i является входной для «внутреннего» кодера. Другая часть кодовых символов «внешних» кодеров ($v_l^{o,(2)}$ пунктирные линии на рисунке 1) не кодируется «внутренним» кодером. Последовательности $v_l^{o,(2)}$ составляют последовательность $v^{o,(2)}$, которая вместе с кодовой последовательностью «внутреннего» кода v^i составляют результирующую кодовую последовательность v. [1]

Интерливер: Обратим внимание на построчный блоковый интерливер (рисунок 2).

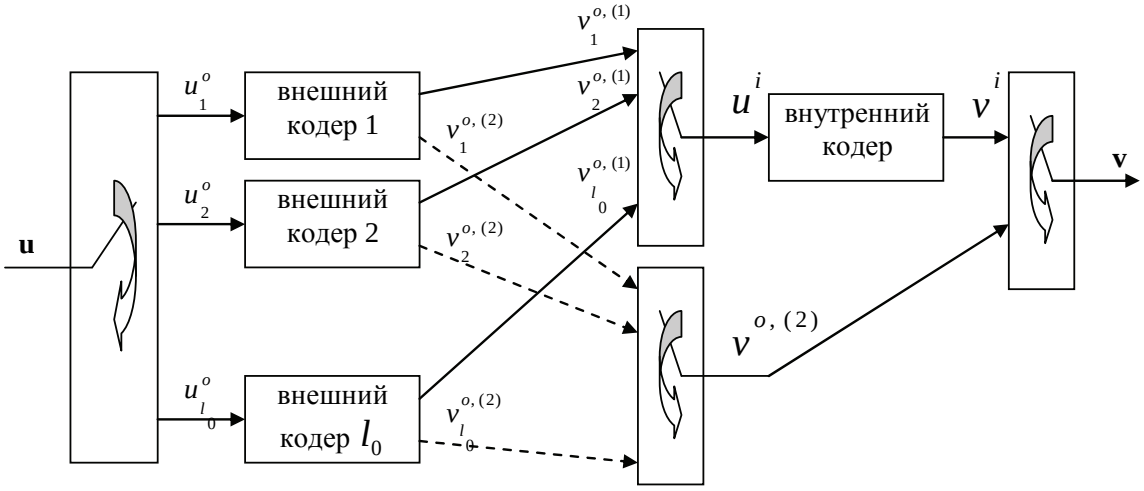


Рисунок 1.

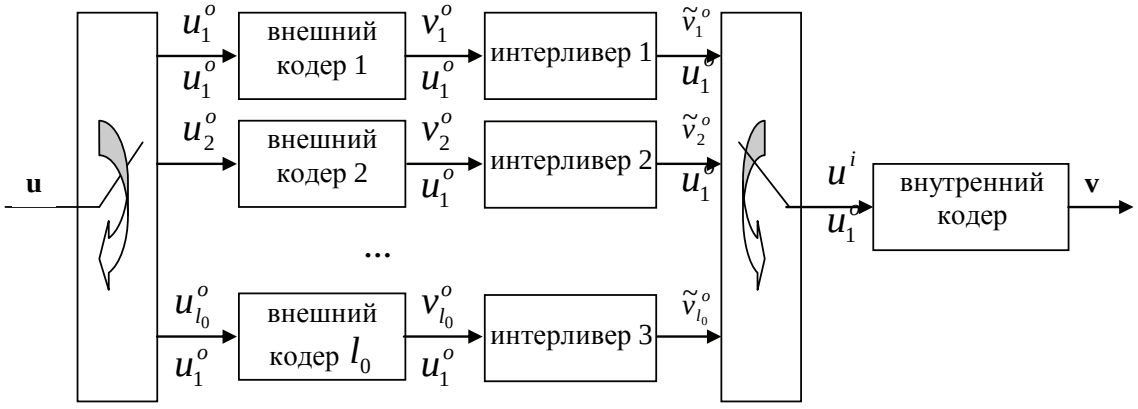


Рисунок 2.

Выходная кодовая последовательность каждого из «внешних» кодеров v_l^o переставляется произвольным независимым способом. Переставленные последовательности $\tilde{v}_l^o$ подаются на внутренний кодер. Блоковый интерливер может быть описан посредством функции перестановки $\pi(\cdot):\{1,...,N\}\to\{1,...,N\}$, которая является взаимно однозначной. N определяет длину перестановки. [2,3]

3. Результаты

Далее приведены графики зависимости сигнал/шум от BER для Гауссовского и Релеевского каналов. Как мы видим из графиков (рисунок 3 и 4) плетеные сверточные (woven) турбо коды по сравнению с Турбо кодами дают энергетический выигрыш порядка 0.4 – 0.5 ДБ (BER=10^{-6}).

Данные результаты получены впервые и нигде ранее не публиковались. Система компьютерного исследования конструкций плетеных сверточных и турбо кодов – «Code» является уникальной программой, не имеющей аналогов. В разработке данной системы я принимал активное участие под руководством научного руководителя О.Д. Скопинцева. Таким образом, новизна полученных результатов может быть подтверждена данным специалистом в области кодирования.

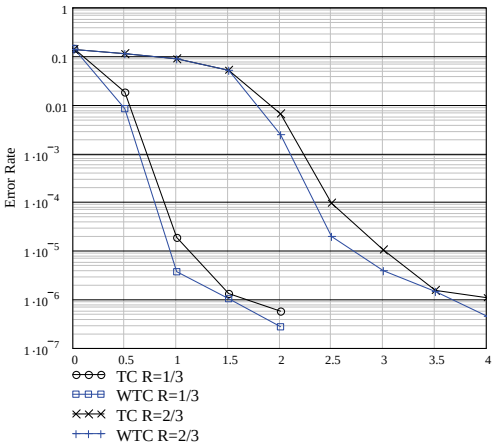


Рисунок 3. Гауссовский канал.

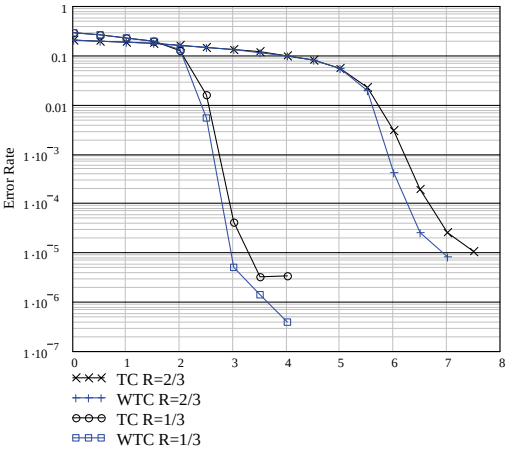


Рисунок 4. Релеевский канал.

Список литературы

[1] *Host S., Johanncsion R.., Zyablov V.V.* A First Encounter with Binary Woven Convolutional Codes // Proc. 4th Int.. Sympos. on Communication Theory and Applications. Lake District, UK. July, 13-18, 1997. P. 13-18.

[2] *Zyablov V., Shavgulidze S., Skopintsev 0., Host S., Jahannesson R.* On the Error Exponent, for Woven Convolutional Codes // Proc. Int. Sympos. on Information Theory. MIT, Cambridge, MA USA. August, 16-21, 1998. P. 160.

[3] *Johannesson R. and Zigangirov. K. Sh.* Fundamentals of Convolutional Coding. IEEE Press, Piscataway, N.J., 1999.

Программный комплекс для исследования систем передачи данных[1]

Ф.В. Грошев
ИППИ РАН
groshev@iitp.ru

Д.К. Зигангиров
ИППИ РАН
zig@iitp.ru

Аннотация

В статье рассматривается базовая версия программного комплекса (ПК) для исследования, анализа системы беспроводной передачи данных на основе OFDM, а также быстрой реализации отдельных блоков на процессорах цифровой обработки сигналов (ПЦОС). Базовая версия ПК реализует физический уровень системы передачи данных. Целью данной работы является разработка приемника и передатчика для исследования характеристик, оптимизации конструкции помехоустойчивого кодирования для дальнейшей реализации на ПЦОС или базовом матричном кристалле (БМК). Данный подход позволит исследовать свойства и структуру системы передачи данных, оптимизировать работу отдельных блоков при различных ограничениях на сложность реализации и различных каналах передачи данных. Кроме этого данный программный комплекс позволит специалистам-разработчикам систем передачи данных при минимальном объеме знаний в области помехоустойчивого кодирования создавать нужные корректирующие коды и сравнивать их с другими вариантами, а специалистам-исследователям при минимальных затратах разрабатывать и проводить сравнительный анализ различных перспективных систем кодирования.

1. Введение

Одной из лидирующих технологий по темпам роста в настоящее время является группа стандартов широкополосного беспроводного доступа с использованием OFDM. Технология OFDM (Orthogonal frequency division multiplexing) - хорошо забытое старое, основные постулаты которого были сформулированы еще в 1960-е годы. Но только теперь развитие микропроцессорной техники позволяет создавать недорогие приемопередатчики. Использование технология OFDM планируется и в перспективных системах беспроводной передачи данных 4 поколения. Увеличение скорости передачи, спектра сигналов и повышение мобильности пользователей требует разработки новых сигнально-кодовых конструкций, методов оценки канала передачи данных, различных способов удешевления конструкции приемопередатчиков и специализированных микросхем канального уровня.

Класс кодов помехоустойчивого кодирования, к которым сейчас проявляется наибольший интерес, носит название кодов с низкой плотностью проверок или низкоплотностных [1], разработаны в 60-е годы и также получили свое развитие в передачи данных с развитием микропроцессорной техники. Различные способы построения этих кодов, способам их декодирования определяет их конечные свойства и сложность их реализации на ПЦОС или БМК.

2. Постановка задачи

Цель данной работы - спроектировать программный комплекс (ПК) для разработки и моделирования систем передачи данных по беспроводным каналам передачи данных с использованием OFDM. ПК должен быть построен на основе следующих принципов: открытость комплекса, наглядность, простота использования, масштабируемость и обнавляемость.

3. Выбор среды проектирования

Разрабатываемый программный комплекс в большой степени использует параллельные вычисления в различных элементах и блоках, и для упрощения контроля и управления этими процессами, упрощения верификации программных модулей, упрощения переноса алгоритмов на ПЦОС или БМК было решено

[1] Эта работа поддержана государственным контрактом №02.514.11.4025 от 18.05.07 по теме: «Разработка и исследования сигнально кодовых конструкций для передачи и защиты меняющихся изображений»

использовать систему проектирования Matlab – Simulink.

Выбор среды Simulink для Matlab обусловлен следующими ее свойствами:

– Это открытая система, допускающая модификацию, дополнение и расширение функциональных блоков модели и системы в целом.
– Наглядность процесса моделирования и отображения результатов (графическое изображение блоков модели жестко определено Simulink).
– При работе с Simulink пользователь имеет возможность модернизировать библиотечные блоки,, создавать свои собственные, а также составлять новые библиотеки блоков.
– Приложение Simulink реализует принцип визуального программирования, поэтому сокращается время разработки систем моделирования и документации к ним. Разработчик может сконцентрироваться на научных исследованиях, и не тратить значительное время оформление работы. Этому способствует автоматический мастер генерации отчетов.

3. Состав программного комплекса

«Модель СПД-Simulink» состоит из блоков, соответствующих физическим блокам системы передачи данных (рис. 1):

– Источник,
– Передатчик,
– Приёмник,
– Канал,
– Блок оценки вероятности ошибки на бит (BER) в обратной связи,
– Генератор образующих матриц кода,
– Генератор импульсного отклика канала,
– Блок задания параметров работы модели.

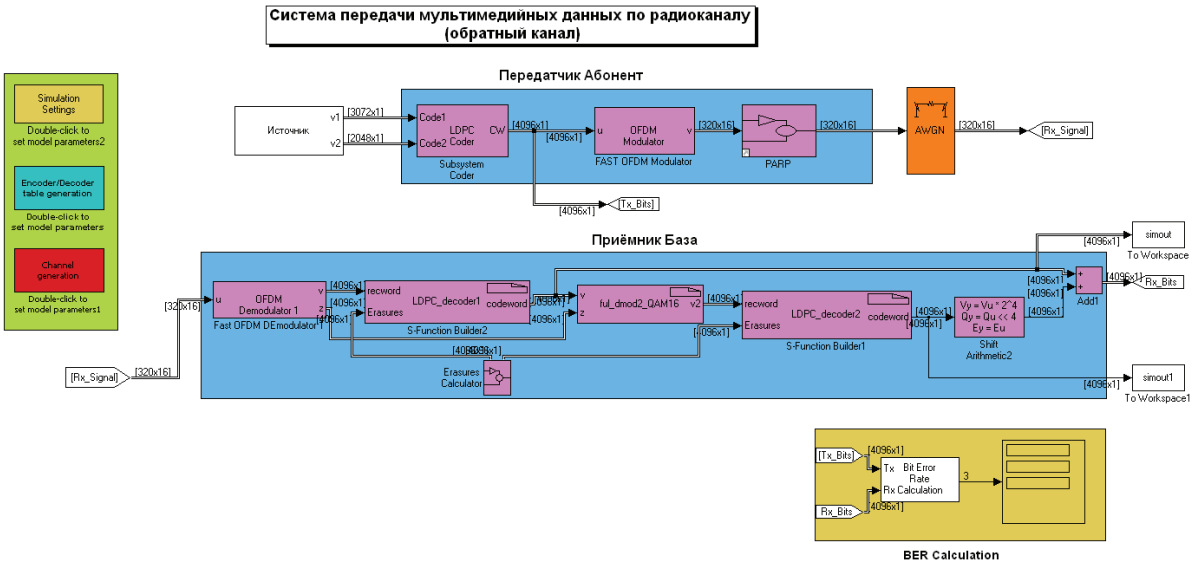


Рис. 1. Программный комплекс

4. Функциональное назначение программного комплекса

Представленная на рис.1 модель позволяет:

– строить низкоплотностные коды с заданными параметрами [2] (длиной, кодовым расстоянием, способом построения, различными компонентными кодами и скоростью);
– Генерировать и экспортировать в текстовом формате проверочные (или каркасные) матрицы кода;
– Импортировать проверочные (или каркасные) матрицы кода, созданные как вне данной системы, при условии соблюдения формата входных файлов, так и с ее помощью;
– Исследовать свойства полученных низкоплотностных кодов (оценка кодового расстояния, оценка скорости кода и сложности кодирования);
– Использовать стандартизованные (встроенные в Matlab) модели радиоканалов с меняющимися характеристиками или подключение специально разработанных моделей;
– Использование средств Simulink для анализа результатов моделирования, визуальное отображение полученных оценок вероятности

ошибки на бит (BER), расстояний и результатов моделирования;
- Моделировать процесс передачи данных;
- Исследовать вероятность ошибки на бит или символ при передаче данных;
- Исследовать модели канала передачи данных;
- Исследовать различные типы модуляций.

5. Результат моделирования

Для моделирования выбрана система передачи данных OFDM длиной 256 символов с вложенной модуляцией КАМ-256. На рис. 2, представлена сигнальная плоскость при отношении сигнал/шум в канале 20 дБ, а на рис. 3 сигнальная плоскость при 12,1дБ.

В качестве помехоустойчивого кода использовалась каскадная конструкция из двух низкоплотностных кода построенных на графах расширителях со скоростью ½ и ¾ и длиной 4096 полубайтовых символов.

При итеративном декодировании с «жестким» решением использовалось введение стираний наименее надежных символов по уровню 6дБ на первом уровне и 13 дБ на втором.

Для упрощения конструкции передатчика и борьбы с пик-фактором вводилось ограничение сигнала по уровню 7дБ.

Рис.2. Сигнальная плоскость 20 дБ

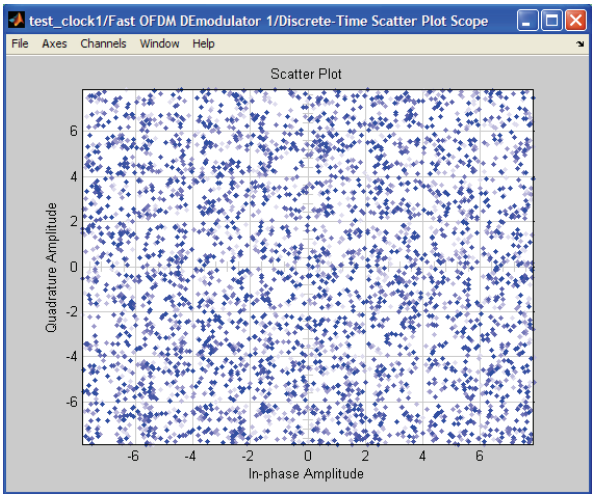


Рис.3. Сигнальная плоскость 12.1 дБ

На рис.4 представлены графики вероятности ошибки на бит (BER) в системе передачи данных с применением кодирования и без него. Можно видеть, что при соотношении сигнал/шум хуже 12дБ применение кодирования существенно уменьшает вероятность ошибки на бит. Например, на уровне BER = 10^{-4} энергетический выигрыш от применения кодирования составляет 6 дБ.

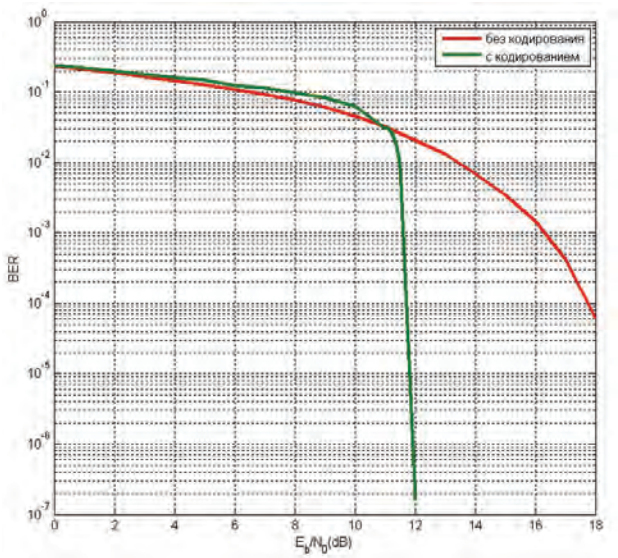


Рис.4. BER

6. Литература

[1] R. G. Gallager, Low-Density Parity-Check Codes, Ph.D. thesis, MIT Press, Cambridge, MA, USA, 1963

[2] S. Stiglmayr and V. V. Zyablov, “Asymptotically good low-density codes based on Reed Solomon codes,” in Proc. XI Int. Symp. Problems of Redundancy in Inform. and Control Systems, available online at http://www.k36.org/redundancy2007/proceedings.php, Saint Petersburg, Russia, July 2007, pp. 98–103.

[3] Simulink help

Применение вэйвлет-преобразования для анализа спектров EXAFS: исследование причины появления и анализ артефактов на спектрограммах

В.Н. Павлов, М.В. Чукалина
Институт проблем технологии микроэлектроники РАН

Аннотация

Исследование рентгеновских спектров поглощения (EXAFS) является важным инструментом в структурных исследованиях аморфных и жидких тел. Вейвлет-преобразование данных EXAFS существенно увеличивает чувствительность метода. Для модельной функции аналитически получен интерференционный член ее вэйвлет-преобразования Морле, что может упростить интерпретацию спектрограмм реальных сигналов.

1. Введение

Физико-химические свойства материалов во многом определяются их локальной структурой. Дифракционные методы, которые традиционно используются для анализа упорядоченных структур, неприменимы для исследования материалов, в которых отсутствует дальний порядок в расположении атомов, например, аморфных материалов, растворов и т.д. Начиная с 1970-х годов, успешно развивается группа методик, позволяющих получать структурную информацию из анализа формы спектров рентгеновского поглощения (зависимость коэффициента поглощения рентгеновского излучения в исследуемом образце от длины волны падающего излучения). Один из методов - EXAFS-спектроскопия (Extended X-ray Absorption Fine Structure). В работах [1-6] дано детальное описание метода. Цель данного метода - получить структурную информацию об окружении центрального атома, вблизи края поглощения которого измеряется спектр. Т.е. определить радиусы координационных сфер и тип рассеивателей, находящихся на этих координационных сферах.

2. Традиционный анализ EXAFS-спектров

Пример спектра поглощения в EXAFS области (дальняя тонкая структура) представлен на рис.1. Для того чтобы получить структурную информацию об образце, работают не со спектром поглощения, а с функцией $\chi(k)k^3$ (рис.2), полученной после нормировки спектра поглощения, см. например [5]. Здесь k - волновое число, определяемое

$$k = \sqrt{\frac{2m}{\hbar}(E - E_0)}, \qquad (1)$$

где m- масса покоя электрона, $\hbar$ - постоянная Планка, E - текущая энергия, E_0 - энергия края поглощения.

Для интерпретации EXAFS-измерений $\chi(k)$ аппроксимируется функцией

$$\chi(k) = S_0^2 \sum_{i=1}^{n} \frac{N_i}{R_i^2} \frac{F_i(k)}{k} e^{-\frac{2R_i}{\lambda}} e^{-2\sigma_i^2 k^2} \sin(2kR_i + \Psi), \qquad (2)$$

где N_i - число атомов в i-й координационной сфере, R_i - средний радиус координационной сферы, σ^2 - фактор Дебая-Уоллера, $F_i(k)$ - амплитуда обратного рассеяния.

Изначально для обработки $\chi(k)$ было предложено использовать преобразование Фурье (рис.3). Положение максимумов Фурье-преобразования соответствует радиусам координационных сфер. Идеи такого подхода впервые были предложены и детально рассмотрены в работах Штерна, Литтла и Сауерса [1-4]. Другой подход к решению уравнения (2), предложенный Бабановым, подробно изложен в работе [5]. Однако, если в одной или нескольких координационных сферах находятся атомы разных типов, то под одним пиком Фурье оказываются несколько разных по природе процессов рассеяния. В таких случаях необходима дополнительная информация, которая может быть получена дифракционными методами, но недоступна при анализе аморфных тел и жидкостей в экологических образцах.

3. Вейвлет-анализ EXAFS-спектров

В 2001 году было предложено использовать вэйвлет-преобразование для анализа функции $\chi(k)k^3$[7]. Амплитуда обратного рассеяния может быть в первом приближении аппроксимирована линейной комбинацией функций Гаусса. Положение центров функций Гаусса на оси k зависит от типа атома, по мере роста атомного номера Z положение максимума смещается в область больших k. Т.е. акт рассеяния, описываемый синусоидальной компонентой, промодулирован функцией амплитуды обратного рассеяния. Вэйвлет-преобразование является хорошим инструментом для анализа сигналов, "частотные" компоненты которых неустойчивы во "времени".

Функцию $\psi(t) \in L$ называют материнским вейвлетом, если она удовлетворяет условию допустимости

$$\int_{-\infty}^{+\infty} dt\psi(t) = 0, \qquad (3)$$

которое необходимо для обратимости вейвлет-преобразования, а именно конечности константы

$$C_\psi = \int_0^{+\infty} d\omega \frac{|\hat{\psi(\omega)}|^2}{\omega} \qquad (4)$$

Вейвлет-преобразование, порождаемое $\psi(t)$, сопоставляет каждой функции $f \in L^2$ число

$$W^\psi f(s,u) = \frac{1}{\sqrt{s}} \int dt f(t)\psi^*(\frac{t-u}{s}) \qquad (5)$$

Вейвлет-преобразование обратимо:

$$f(t) = \frac{1}{C_\psi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} W^\psi f(s,u) \frac{1}{s^2} \psi(\frac{t-u}{s}) ds du, \qquad (6)$$

если $C_\psi < +\infty$.

В качестве материнского вэйвлета было предложено использовать разные функции: Морле [1], Каши [9], FEFF-Morlet [10]. Иногда, например, выбор FEFF-Morlet был обусловлен требованиями задачи. Однако, чаще всего [11-12] для анализа спектра применяется материнский вэйвлет Морле.

Вейвлет Морле $\psi_{\sigma\eta}(t) = \frac{1}{\sqrt{2\pi\sigma}}(e^{i\eta t} - e^{-\frac{\eta^2\sigma^2}{2}})e^{-\frac{t^2}{2\sigma^2}}$ зависит от двух параметров - параметра масштабирования σ и частотного параметра η. Зачастую членом $e^{-\frac{\eta^2\sigma^2}{2}}$ пренебрегают, если в задаче существенны $|\eta\sigma| >> 1$. Положения атомов в (k,R)-пространстве вычисляются [8] как локальные максимумы спектрограммы $|W^\psi f(\eta,u)|^2$, где $W^\psi f(\eta,u)$ зависит от σ как от фиксированного параметра через $\psi_{\sigma\eta}$. При использовании этого вэйвлета самая простая интерпретация спектрограммы. Параметр масштабирования однозначно связан с радиусом координационной сферы, а положения вершин "гор" на спектрограммах по k-оси характеризуют число разнотипных рассеивателей на одной координационной сфере и тип рассеивателя [11] (рис.4)

Вместе с тем, при обработке некоторых спектров возникающая картина не столь однозначна. Мы предполагаем, что это связано с наличием интерференционных членов в $|W^\psi f(\eta,u)|^2$. Действительно, [8] исходит из того, что

$$|W^\psi f(\eta,u)|^2 \approx \hat{M}_1 [g_{\sigma_1}(\eta - R_1)]\hat{M}_2 [g_{\sigma_2}(u - k_1)] + \hat{M}_3 [g_{\sigma_3}(\eta - R_2)]\hat{M}_4 [g_{\sigma_4}(u - k_2)], \qquad (7)$$

где $\hat{M}_{1-4}$- линейные операторы, g_σ - гауссовская функция с дисперсией σ; интерференционные члены в (7) отброшены. В действительности, нужно включить в правую часть (7) слагаемое

$$\hat{M}_5 [g_{\sigma_5}(u - \frac{k_1+k_2}{2})]\hat{M}_6 [g_{\sigma_6}(\eta - \frac{R_1+R_2}{2})] \times \cos[(a\eta + bk_1 + ck_2)(d\eta + eR_1 + fR_2)], \qquad (8)$$

где a, b, c, d,e, f - постоянные, ($\hat{M}_5$ и $\hat{M}_6$ - линейные операторы) которое может приводить к: (i) появлению дополнительного локального максимума на спектрограмме (рис. 5), (ii) смещению максимумов в (k_1,R_1), (k_2,R_2) и (iii) даже исчезновению максимумов (k_1,R_1), (k_2,R_2).

4. Заключение

В работе впервые показано, что интерференционные эффекты в вейвлет-анализе EXAFS-данных играют важную роль. Получено качественное объяснение особенностей, присутствующих на некоторых экспериментальных спектрограммах, как максимумов интерференционных слагаемых. Намечен путь к улучшению разрешающей способности вейвлет-анализа EXAFS-данных - учет смещения реальных максимумов за счет их взаимного влияния. В дальнейшем планируется получить условие допустимости для идентификации особенностей спектрограмм, т.е. условие, невыполнение которого позволяет идентифицировать особенность спектрограммы как "артефакт". Также предполагается построить алгоритм для количественного анализа экспериментальных данных с учетом полученных здесь результатов.

Работа выполнена при частичной финансовой поддержке РФФИ 08-01-00285-а, 06-02-16117-а.

Литература

[1] D. E. Sayers and E. A. Stern, Phys. Rev. Lett, 27, 1204 (1971).

[2] E. A. Stern, Phys Rev B, 10, 3027 (1974).

[3] Stern, D. E. Sayers and F. W. Lytle, Phys Rev B, 10, 4836 (1975).

[4] W. Lytle, D. E. Sayers and E. A. Stern, Phys Rev B, 11, 4825 (1975).

[5] Д.И. Кочубей, Ю.А. Бабанов, К.И. Замараев, Р.В. Ведринский, В.Л. Крайзман, Г.Н. Кулипанов, Л.Н. Мазалов, А.Н. Скринский, В.К. Федоров, Б.Ю. Хельмер, А.Т. Шуваев. Рентгеновский метод изучения аморфных тел: EXAFS спектроскопия. Наука (1988). [1]

[6] Ведринский Р.В. EXAFS-спектроскопия - новый метод структурного анализа // Соросовский Образовательный Журнал. 1996. № 5. С. 79-84.

[7] H. Funke, M. Chukalina, WAVELET ANALYSIS OF EXAFS DATA - FIRST STUDIES , FZR Annual Reports 343 (2001) p.45.

[8] H. Funke and A.C. Scheinhost, M. Chukalina, Phys. Rev. B, vol.71, 094110 (2005).

[9] M.Munos, P. Argoul, F.Farges. Am. Mineral 88, 694 (2003).

[10] H. Funke, M. Chukalina and A. C. Scheinost J. Synchrotron Rad. 14, 426-432 (2007).

[11] M. Sahnoun and C. Daul O. Haas, J. Appl. Phys. 101, 014911 (2007).

[12] H Funke M Chukalina and A Rossberg Phys. Scripta. T115 232-234 (2005).

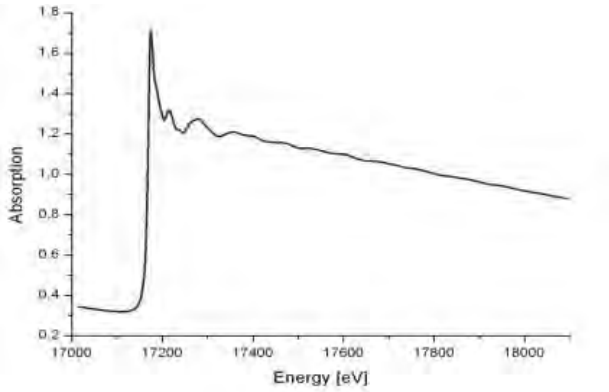


Рис. 1: Пример спектра поглощения. EXAFS-область составляет порядка 1000 еВ после края поглощения выбранного элемента. На кривой край поглощения находится при 17200 еВ.

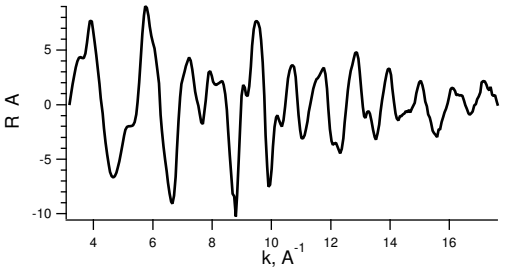


Рис. 2: Пример функции $\chi(k)k^3$

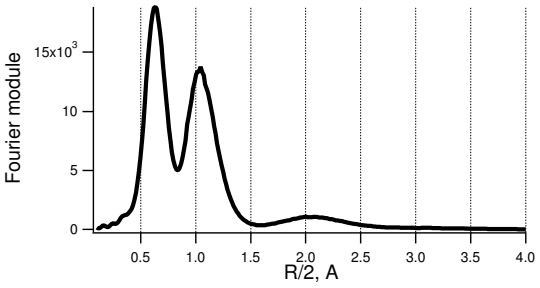


Рис. 3: Модуль Фурье-преобразования функции, представленной на рис.1

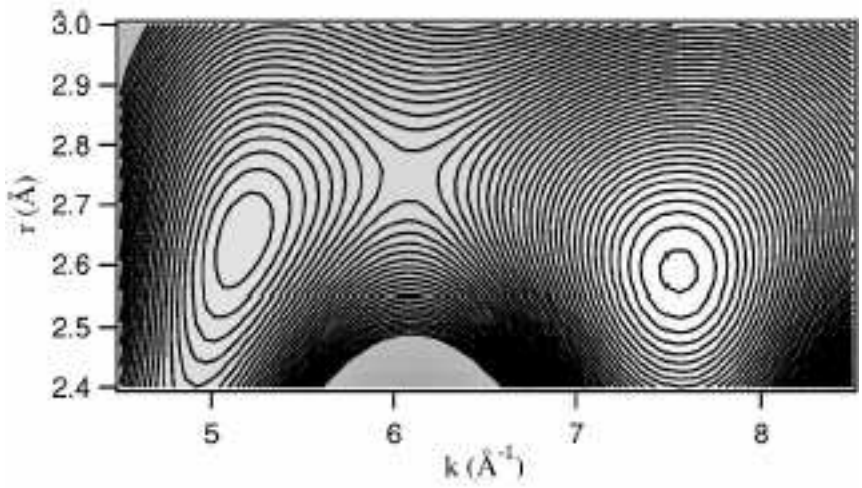


Рис. 4: Область вэйвлет-карты [11]. На одной координационной сфере радиуса $2.4A < R < 3.0A$ видны точно два максиума, отражающие наличие двух типов рассеивателей (как указано в работе, Al и Zn).

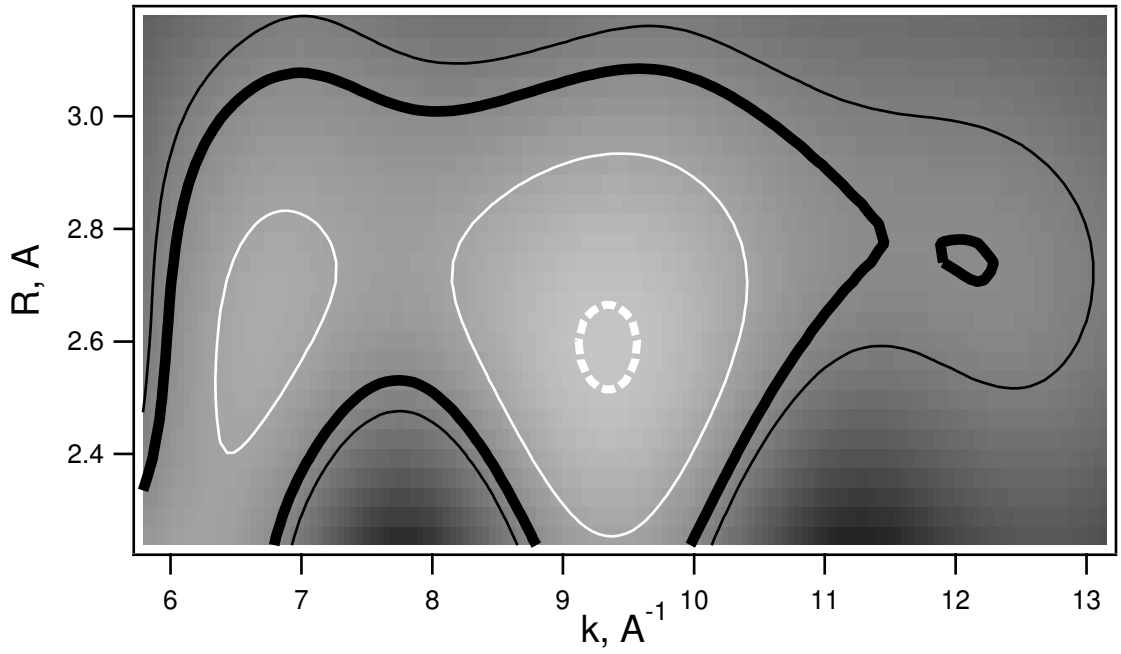
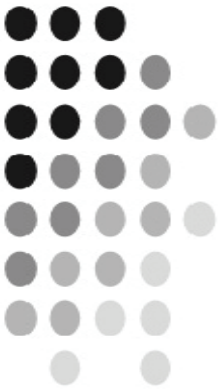


Рис. 5: Иллюстрация появление дополнительного максимума на спектрограмме. Дополнительный максимум выделен изолиниями черного цвета.

Раздел 4

Компьютерная лингвистика и моделирование естественного языка



Развитие системы построения семантического представления текста с использованием языка-посредника UNL на базе лингвистического процессора ЭТАП-3

Вячеслав Диконов
ИППИ РАН
dikonov@iitp.ru

Аннотация

В статье описывается текущее состояние работы по созданию на базе лингвистического процессора ЭТАП-3 системы получения семантического представления текста на английском языке в формате "Универсального Сетевого Языка" UNL. Предлагается обзор архитектуры системы, используемого словаря концептов, а также правил разрешения лексической неоднозначности.

1. Введение

Эта статья посвящена практической работе, которая выполняется в рамках длительного международного проекта. Статья содержательно связана с представленной на конференции ИТиС'07 публикацией [1], где содержится общий обзор семантического языка и проекта UNL. Ее основное содержание составляет описание хода работы и некоторых результатов проделанной в 2008 году работы по развитию системы получения семантического представления текста, использующей язык UNL. Задачей создаваемой системы является преобразование получаемой лингвистическим процессором ЭТАП синтаксической структуры в семантический граф. При этом происходит дальнейший анализ и обогащение построенного системой ЭТАП представления текста новой информацией. Источником дополнительной информации в этом процессе могут быть имеющиеся в системе словари и правила или человек, если программа работает в интерактивном режиме. Расширенное таким образом семантическое представление текста будет пригодно для различных задач автоматизированной обработки текстовой информации.

Рамки статьи не позволяют еще раз описывать все особенности UNL. Однако, прежде чем сосредоточиться на разрабатываемой системе, напомним используемую в рамках проекта UNL и далее в тексте терминологию.

UNL (Universal Networking Language) – язык для записи смысла текста, способный играть роль языка-посредника.

конвертер – программная система, способная преобразовывать текст на естественном языке в код на языке UNL.

деконвертер – система предназначенная для преобразования кода UNL в текст на естественном языке. В мире существует несколько деконвертеров, обеспечивающих поддержку разных языков.

концепт – произвольное понятие, выделяемое в толковом словаре естественного языка как отдельное значение слова или идиоматичного сочетания слов.

UNL-граф – способ представления текста на языке UNL. Граф состоит из концептов, соединенных направленными типизованными семантическими связями. Узлы графа могут содержать в себе не только выражающие концепты лексические единицы UNL, но и другие графы, тем самым образуя гиперграф.

UNL-код - запись графов UNL в машиночитаемом текстовом виде. Использует XML-подобный формат записи.

UW (Universal Word) **–** "универсальные слова" лексические единицы UNL. Каждое UW соответствует одному концепту. Предполагается, что все концепты всех языков должны получить по отдельному UW. Если в словаре UNL нет готового UW для нужного понятия, то его можно добавить и опубликовать.

общее правило – в ЭТАП правило применяемое к каждому слову предложения.

трафаретное правило – в ЭТАП правило, которое применяется к тем словам, словарные статьи которых содержат в себе ссылки на данное правило.

словарное правило – в ЭТАП правило которое применяется к одному определенному слову и находится в словарной статье этого слова.

2. Устройство конвертера

Система построения семантического представления текста на основе лингвистического процессора ЭТАП (далее конвертер) пользуется стандартными программными модулями ЭТАП и состоит из лингвистических правил и словарных

данных. Работа конвертера начинается с момента построения нормализованной синтаксической структуры. Она представляет собой дерево зависимостей, откуда уже были удалены служебные слова, такие как артикли, вспомогательные глаголы, некоторые частицы и предлоги. Следует отметить, что построенное лингвистическим процессором ЭТАП синтаксическое представление обеспечивает разрешение омонимии и может включать в себя ряд семантизированных отношений и дополнительных семантизированных признаков лексем.

Процесс конверсии разделен на три этапа:

1. Выработка некоторых атрибутов UNL, характеризующих предложение в целом. Например: *Could you, please, @polite look after the child? @interrogative*.
2. Замена английских лексем на UW с полным разрешением лексической неоднозначности, в том числе имеющей значение для перевода на другой язык. Например:
   - *you = to one person/to many,*
   - *child = young person/grown up person behaving like a child/member of an ethnic group ...*
3. Замена синтаксических отношений на семантические и перестройка структуры.

Конвертер способен работать в интерактивном режиме, позволяющем в процессе анализа предложения спрашивать пользователя о значении многозначных слов. Этот режим отличается от интерактивного режима ЭТАП при переводе с английского на русский и обратно тем, что выбор пользователя не может быть отменен правилами перевода. Кроме того, уже готовый UNL-код можно редактировать с помощью специализированного визуального редактора.

3. Текущее состояние работы

За прошедшее с момента прошлого обзора [1] время создаваемая система планомерно развивалась и сделала значительный шаг вперед. Проделанная работа обеспечивает качественный переход от автономно развиваемой модели, способной обрабатывать лишь отдельные тестовые фразы, к системе, которая может анализировать реальные тексты и обладает базовой совместимостью с параллельно создаваемыми деконвертерами наших партнеров, нацеленными на поддержку других языков. Вместе с тем, работа еще далеко не закончена. Совместимость с другими системами, а значит и возможность использовать UNL в качестве языка-посредника, требует наличия общего словаря, единого формата представления текста и единства интерпретации получаемых семантических структур. Последнее предполагает, что все порождаемые конвертером структуры всегда соответствуют спецификациям UNL, вне зависимости от используемого конвертером внутреннего формата предшествующего синтаксического представления. На данной стадии в достаточной мере реализованы два из трех перечисленных компонентов совместимости, и идет работа над более полным преобразованием различных синтаксических конструкций в отвечающие спецификациям UNL выражения. Так как система ЭТАП является базой не только конвертера, но и деконвертера UNL, параллельно устраняются найденные пробелы в ранее созданных правилах деконверсии.

Раньше преобразование древесной синтаксической структуры зависимостей ЭТАП в состоящий из семантических отношений UNL-граф было фрагментарным. Конвертер обслуживал только основные синтаксические связи и в реальных текстах оставлял множество непреобразованных фрагментов структуры. Теперь любой текст может быть автоматически преобразован в соответствующий формату UNL, но пока еще не всегда оптимальный код, который можно загрузить в созданный нашими партнерами редактор графов или преобразовать в текст на другом языке с помощью различных деконвертеров. Был значительно расширен набор правил третьего этапа конверсии, необходимый для достижения полной совместимости между разными реализациями UNL. В ходе этой работы выявляются отдельные недостатки и пробелы имеющихся спецификаций языка. Одним из результатов этого стал разработанный расширенный набор атрибутов UNL для выражения модальных значений.

Одновременно была проведена большая работа над используемыми в системе ЭТАП словарями английского языка и языка UNL. Ее целью было получение достаточно объемного словаря концептов, дающего возможность адекватного представления большого числа произвольных текстов, позволяющего фиксировать в UNL-коде разные значения большого числа слов, а также обеспечивающего лексическую совместимость документов на языке UNL с различными деконвертерами. Кроме того, для введения правил конверсии, которые отражают особенности отдельных лексических единиц, было необходимо создать соответствующие статьи словаря UNL. Благодаря импорту данных Wordnet и ручному редактированию словарь был расширен с 9000 до более чем 78000 концептов. Теперь конвертер впервые может различать и фиксировать значения свыше 45000 английских слов.

4. Расширение словаря

Лаборатория компьютерной лингвистики ИППИ давно занимается работой над общим для всех участников проекта UNL словарем, который должен обеспечивать возможность переноса получаемого UNL-кода между разными системами машинного перевода. Поскольку словарь UW должен пополняться разными участниками проекта UNL, необходимо было принять общие принципы формирования "универсальных слов" (UW). Та-

кой свод правил [4] был предложен И.М. Богуславским и принят партнерами в июле 2007 года в Гренобле. В дальнейшем лаборатория и испанский UNL-центр параллельно работали над созданием общего словаря UW. Сначала нами была вручную подготовлена часть словаря, охватывающая 600 наиболее частотных глаголов, а также предлоги и некоторые другие часто встречающиеся слова английского языка. В дальнейшем эта часть словаря расширялась и достигла объема 2726 английских слов и около 9000 UW. Одновременно с этим испанский центр UNL завершил проект по автоматизированному получению соответствующего согласованным общим правилам большого словаря UW на основе базы данных Wordnet. Вместе два словаря составляют основу общего словаря UNL, который будут использовать и совместно пополнять далее все партнеры нашего проекта.*

Полученные на основе Wordnet данные были импортированы в словари ЭТАП. Всего было добавлено 69703 новых UW и отсеяно более 121000. Столь значительное количество отсеянных UW объясняется особенностями Wordnet, где содержится множество терминологических словосочетаний, собственных имен и различных вариантов написания одних и тех же слов. Отсеянные UW можно разделить на две приблизительно равные группы:

а) составные UW, которые соответствуют идиоматическим выражениям или состоящим из нескольких слов терминам, например *morse_code(icl>code>thing,equ>morse)*;

б) UW, которые представляют значения отсутствующих в английском комбинаторном словаре ЭТАП редких или стилистически окрашенных слов, имен и названий, латинских терминов, причастий (которые рассматриваются в ЭТАП как формы глаголов) и т п., например *abronia(icl>caryophylloid_dicot_genus>thing)*, *profaned(icl>adj)* (от глагола «to profane») и др.

Группа составных UW (а) включает в себя 64450 концептов. Во время конверсии текста они должны с помощью словарных правил замещать собой определенные фрагменты синтаксической структуры, а при деконверсии вновь порождать соответствующие фрагменты структуры. В ближайшее время UW, которые представляют собой сочетания известных системе ЭТАП слов, будут включены в словарь конвертера.

Среди прочих отсеянных UW (б) имеется группа омонимов уже имеющихся в словаре ЭТАП слов. Их необходимо рассматривать как неизвестные слова даже при совпадении части речи. Система ЭТАП использует очень богатую словарную информацию при построении синтаксической структуры и сильно зависит от словарей. Если в тексте встречается неизвестное системе значение или омоним описанного в словаре слова, употребление которого не соответствует словарному описанию, то ЭТАП не сможет построить правильную синтаксическую структуру предложения. Например, фраза *REM sleep* может быть ошибочно переведена как *сон Рема* вместо *фаза быстрого сна* или *REM-сон*, так как в комбинаторном словаре имеется только статья омонимичного аббревиатуре REM имени с дескрипторами 'ЛИЦО', 'ЧЕЛОВЕК'. Для перевода UW *rem(icl>sleep>thing,equ>paradoxical_sleep)* она не подходит, так как эти омонимы имеют разные свойства.

Включение UW неизвестных слов в словарь ЭТАП требует трудоемкого расширения английского и русского комбинаторных и морфологических словарей. Альтернативой этому служит имеющийся в системе ЭТАП аварийный механизм обработки неизвестных слов путем приписывания их к специальной абстрактной статье словаря FICT-LEX или к шаблонам личных имен (FICT-GIVENNAME, FICT-SURNAME), географических названий (FICT-LOC), частей речи (FICT-A, FICT-S...) и т.п. Эти шаблоны позволяют приписать неизвестному слову минимальный набор свойств и включить его в структуру предложения. Отсутствующие в словаре UNL UW система приводит к заглавному слову, которое в дальнейшем обрабатывается как совпадающее по форме английское или неизвестное слово. Однако, как показано выше, при появлении омонимов известных системе слов возникают ложные интерпретации. Чтобы их исключить, все подобные омонимы необходимо ввести в словарь.

Вошедшие в словарь ЭТАП после объединения словарей UW также неоднородны. Есть три различные группы, которые отличаются по своему происхождению:

- Подготовленные вручную UW, удовлетворяющие требованиям общего словаря.
- Автоматически полученные UW из Wordnet, удовлетворяющие требованиям общего словаря.
- UW из различных специальных проектов, в частности проекта перевода Энциклопедии систем жизнеобеспечения (EOLSS), не всегда удовлетворяющие требованиям общего словаря. Среди них имеется значительное количество терминов.

Статус и связанные с ним возможности использования UW будут определяться тем, входит ли оно в текущую версию общего словаря. Не входящие в общий словарь UW должны рассматриваться как синонимы одного из UW общего словаря либо включаться в общий словарь. UW такого типа используются только при деконверсии. Они распознаются и немедленно заменяются на соответствующее UW общего словаря. Такие UW не должны использоваться при конверсии текста, чтобы не нарушать совместимость доку-

*В настоящее время французский центр занимается разработкой инфраструктуры хранения, совместного редактирования общего словаря и обмена данными между разными центрами UNL.

мента со стандартным общим словарем. Сейчас такой статус имеют включенные в словарь ЭТАП некоторые UW проекта EOLSS.

В будущем предполагается также дополнительное деление UW на исправленные, но еще не опубликованные, действующие UW и устаревшие, которые были исключены из общего словаря, но оставлены для совместимости со старыми документами. Для определения источника и статуса UW все статьи словаря UNL в системе ЭТАП содержат комментарии. Эти комментарии будут необходимы для отслеживания изменений состава UW и обеспечения дальнейшей синхронизации словаря ЭТАП с общим словарем UNL.

В настоящее время словарь покрывает наиболее часто используемую область лексики английского языка, включая отсутствующие в Wordnet фразовые глаголы и служебные слова. Он продолжает пополняться и редактироваться, если обнаруживаются лакуны или ошибки.

Каждая статья английского словаря, где есть зона перевода на UNL, содержит список UW, соответствующих различным значениям заглавного слова статьи. Этот список упорядочен по частоте встречаемости различных значений. В случае UW из Wordnet использовалась информация Wordnet. Если в зоне UNL нет правил для автоматического выбора UW для многозначных слов, то по умолчанию используется первое в списке, которое соответствует наиболее распространенному значению.

5. Правила разрешения лексической неоднозначности

Снятие лексической неоднозначности является важной частью процесса конверсии. Оно позволяет перейти от анализа структуры предложения и абстрактных грамматических значений к анализу смысла высказывания. Эта операция необходима для избежания двусмысленности во внутреннем представлении текста, способной привести к грубым ошибкам перевода. В лингвистическом процессоре ЭТАП эта задача частично решается при синтаксическом анализе (благодаря различиям синтаксических свойств омонимичных слов) и далее при переводе. Однако, до сих пор словари ЭТАП сознательно упрощались. Составители машинных комбинаторных словарей включали в них только минимально необходимые и часто употребляемые варианты перевода, исключая при этом более редкие и периферийные значения/омонимы. Таким образом достигалось уменьшение числа гипотез при анализе и переводе предложения. Это решение вполне обосновано на этапе синтаксического анализа, но делает ЭТАП неспособным хорошо переводить некоторые тексты на другой язык.

Задача получения качественного семантического представления на языке UNL требует более полно и детально различать и фиксировать значения слов. Если это не удается сделать автоматически, то использование UNL предполагает возможность ручной правки семантического представления. Во время конверсии текста конвертер может задавать вопросы пользователю о смысле отдельных слов, а готовые графы можно редактировать перед публикацией или деконверсией.

Определить значение отдельного слова в тексте можно только по контексту его употребления. Можно выделить узкий контекст, состоящий из слов синтаксически связанных с искомым напрямую или через служебные слова, контекст предложения и широкий контекст абзаца, главы, текста. Контекст понимания включает в себя также всю сумму общих знаний читателя о предметной области.

Разрабатываемый конвертер целиком опирается на возможности текущей версии системы ЭТАП. Это накладывает некоторые важные ограничения. ЭТАП не имеет интерфейса с какой-либо онтологией знаний, опирается на ввод лингвистической информации экспертами и не включает в себя модуля самообучения. Кроме того, текст анализируется исключительно по одному предложению, и программа не предоставляет возможности сдвигать границы предложений или накапливать какую-либо фоновую информацию в процессе обработки документа, чтобы применять их к следующим предложениям. Эти ограничения не позволяют использовать широкий контекст и, как следствие, ряд известных методов разрешения лексической неоднозначности на основе тезаурусов [3].

Единственным эффективным способом снятия лексической неоднозначности в таких условиях является анализ узкого контекста. Такой анализ может опираться на семантические ограничения, налагаемые на контекст отдельными значениями слова, и на устойчивые словосочетания. Первый метод известен давно и упоминается еще в 1973 г. [5]. Он позволяет получить практически значимый результат для развернутых предложений не осложненных эллипсисом и анафорами. Естественным недостатком является трудоемкость ручного описания большого числа контекстов для множества слов. Поэтому, опыты такого рода как правило ограничивались небольшим числом слов. Метод семантических ограничений удачно дополняется методом устойчивых словосочетаний [2], который обеспечивает высокую достоверность выбора значения, но имеет ограниченную сферу применения. Следует также отметить, что при создании UNL-конвертера ставится задача анализа широкого спектра реальных текстов без тематических ограничений, а это требует максимальной широты охвата лексики естественного языка, даже за счет полноты и точности различения значений отдельных слов в начале работы.

В настоящее время проводится эксперимент по созданию и тестированию механизма разрешения лексической неоднозначности в среде системы

ЭТАП на основе словаря концептов UNL, который позволял бы:

- реализовать оба упомянутых выше метода разрешения неоднозначности;
- постепенно улучшать качество определения значений каждого слова;
- использовать автоматические методы для сопоставления контекстов и значений, а также наполнения словаря необходимой информацией;
- применять тот же механизм не только для UNL-конверсии, но и для других целей.

Решением этой задачи стало введение набора универсальных трафаретных правил проверки свойств контекста, которые в отдельных сложных случаях дополняются словарными правилами. Эти правила можно назвать детекторами значений. Они должны отвечать нескольким взаимосвязанным требованиям:

- предельно широкая область применения;
- максимальная гибкость за счет изменения параметров обращения к правилам;
- в состав набора правил должны быть включены только реально необходимые критерии проверки узкого контекста;
- правила должны быть пригодны для автоматической генерации ссылок на них с помощью предполагаемого в будущем механизма самообучения системы.

Всего таких правил в настоящий момент 23, причем это количество определяется ограничениями языка записи правил в ЭТАП и может быть сокращено более чем вдвое. В рамках эксперимента они используются примерно в 250 статьях частотных слов английского языка. Каждое из введенных правил должно проверять различные свойства контекста слова, которые задаются аргументами правила. Если контекст отвечает заданному шаблону, то слово заменяется на соответствующее UW, которое также указывается при вызове правила. Исходя из требования минимизации набора детекторов был принят следующий алгоритм добавления правил:

- Определить признаки контекста, позволяющие истолковать слово одним из предлагаемых словарем UNL способов.
- Если нет правила-детектора, позволяющего проверить одновременно все найденные признаки, то ввести новое правило.
- Если подходящее правило есть, то вставить ссылку на него в зоне UNL английского комбинаторного словаря.
- Если имеется несколько возможных контекстов, или отдельные признаки могут меняться, то добавляется несколько ссылок на правила-детекторы с разными аргументами.

Правила-детекторы различаются по типу проверяемых свойств слов, которые составляют узкий контекст, (имя лексемы, принадлежность лексемы к определенному семантическому классу или наоборот отсутствие такой принадлежности), по направлению синтаксической связи (зависимое или хозяин) и расстоянию в структуре (прямая связь или опосредованная другой лексемой). Правила также подразделяются на группы в зависимости от того, какое количество элементов контекста (связанных узлов) они проверяют.

Внутри словарной статьи множество ссылок на правила-детекторы значений сортируются так, чтобы проверяющие большее число элементов контекста правила предшествовали остальным, а внутри образующихся групп первыми вызывались бы правила, которые проверяют слова на наличие сравнительно более редких свойств (сначала уникальное имя лексемы, затем небольшие по объему семантические классы). Таким образом, автоматически образуется множество проверок-фильтров, которые образуют дерево принятия решения о наиболее вероятном значении слова в данном контексте.

Эксперимент показал, что для определения значения слова в узком контексте нужно проверить свойства от одного до трех связанных с искомым валентными отношениями слов. Длина цепочки связей свыше 2 зависимых оказалась излишней, и таких правил в наборе детекторов нет. При ручном вводе в словарь правил-детекторов применялся принцип поиска минимально необходимого для приписывания слову значения критерия. Таблица 1 показывает, что число случаев, когда необходимо проверить свойства более чем одного соседа в контексте невелико, а случаи, когда недостаточно проверки двух соседей – исключительно редки.

Таблица1. Частотное соотношение в словаре ссылок на правила различных групп по числу проверяемых соседей

1 сосед	2 соседа	3 соседа
87%	12%	1%

Однако, при использовании более детальной семантической классификации и увеличении числа распознаваемых значений следует ожидать некоторого сглаживания этих пропорций, так как потребуется более детально описывать контексты.

На рисунке 1 показан фрагмент статьи английского глагола to address со ссылками на пра-

```
TRAF:UNL-CONV1.17    "величать – he addressed me with 'sir'"
T1:2-COMPL,LEN1:WITH,T2:'ЛИЦО',LUNL:ADDRESS{ICL>CALL>DO;COB>NAME;AGT>PERSON;OBJ>PERSON}
TRAF:UNL-CONV1.16    "задать вопрос"
T1:1-COMPL,LEN1:QUESTION1,LUNL:ADDRESS{ICL>ASK>DO;AGT>THING;OBJ>THING}
TRAF:UNL-CONV1.15    "обращать (речь)"/"обращаться"
T1:1-COMPL,T2:'ИНФОРМАЦИЯ',LUNL:ADDRESS{ICL>SEND>DO;AGT>THING;OBJ>THING;REC>THING}
T1:1-COMPL,T2:'ЛИЦО',LUNL:ADDRESS{ICL>SPEAK>DO;AGT>THING;REC>THING}
```

Рисунок 1. Ссылки на правила-детекторы значений

вила разрешения неоднозначности. При проверке контекста указывается тип синтаксических отношений, связывающих проверяемые слова (аргумент T1). В качестве идентификаторов семантического класса (здесь аргумент T2) могут выступать как семантические дескрипторы ЭТАП, так и некоторые синтаксические характеристики, которые являются семантически значимыми. Например, так можно использовать обозначения частей речи, классов местоимений (REL, INTERROG) и т п.. Кроме того, символ * позволяет отказаться от проверки на принадлежность к определенному классу и использовать то же правило для простого определения наличия зависимого по указанному отношению.

В таблице 2 приводится частотное соотношение различных детекторов по их употребимости на материале текущей версии комбинаторного словаря ЭТАП с указанием содержания проверки.

Таблица 2. Частотное соотношение ссылок на разные правила

Тип проверки		**Доля**
1 сосед		
1.15	Наличие непосредственного зависимого заданного класса	50,51%
1.16	Наличие и имя непосредственного зависимого	17,76%
1.13	Наличие и класс хозяина	9,07%
1.17	Наличие опосредованного зависимого заданного класса, связанного через указанное служебное слово	5,92%
1.14	Наличие непосредственного зависимого, отличного от указанного класса	1,58%
1.10	Наличие и имя хозяина	1,58%
1.18	Наличие и имя опосредованного зависимого, связанного через указанное служебное слово	0,79%
1.11	Отсутствие зависимого определенного класса	0,79%
2 соседа		
1.25	Наличие двух зависимых заданных классов	5,92%
1.24	Наличие хозяина заданного класса и зависимого заданного класса	1,18%
1.26	Наличие двух зависимых с указанными именами	0,79%
1.28	Наличие опосредованного зависимого, заданного класса и отсутствие другого зависимого заданного класса	0,79%
1.21	Наличие и принадлежность хозяина к двум разным классам	0,39%
1.27	Наличие опосредованного зависимого заданного класса, связанного через указанное служебное слово, и еще одного зависимого заданного класса	0,39%

Правила унифицированы и изначально предполагают возможность автоматического добавления ссылок на себя, хотя имеющиеся сейчас в составе системы ЭТАП инструменты не дают такой возможности. Тем не менее, это необходимо. Комбинаторный словарь английского языка в ЭТАП содержит примерно 76000 слов, и каждое из них может иметь от 1 до нескольких десятков выражаемых с помощью UW значений. Для определения каждого значения нужно столько ссылок на правила-детекторы, сколько существует типов контекста, где реализуется это значение. По приблизительной оценке, для разрешения неоднозначности всех слов текущего словаря нужно до полумиллиона обращений к трафаретным правилам. Ввести такое количество данных вручную практически невозможно, поэтому для решения данной задачи требуется дополнить систему ЭТАП модулем самообучения путем автоматической генерации ссылок на правила-детекторы по семантически размеченному корпусу и по результатам интерактивного ввода пользователя.

6. Заключение

Есть два необходимых направления развития системы. Во-первых, необходимо обеспечить поддержку всех синтаксических конструкций и порождение соответствующих графов. В условиях отсутствия зафиксированного в форме корпуса эталона UNL это требует еще и выбора наилучшего способа представления некоторых конструкций. Во-вторых, для обеспечения автоматической конверсии необходимо обеспечить разрешение лексической неоднозначности в объеме словаря UNL. Для решения этой задачи необходимо ввести механизмы более эффективного обучения системы, чем ручное программирование.

[1] В.Г. Диконов, “Обзор реализации семантического языка UNL в системе ЭТАП”, *Информационные технологии и системы*, ИТиС'07, Звенигород 2007, стр. 141-149.

[2] Кобрицов Б. П., Ляшевская О. Н., Шеманаева О. Ю. Снятие лексико-семантической омонимии в новостных и газетно-журнальных текстах: поверхностные фильтры и статистическая оценка , *Интернет-математика — 2005*. Москва, 2005.

[3] Лукашевич Н.В., Добров Б.В., Разрешение лексической многозначности на основе тезауруса предметной области, *Труды международной конференции "Диалог 2007"*, Бекасово 2007, М. Наука - 2007. стр. 400-406.

[4] I.M Boguslavsky “Guidelines for UW construction”, *manuscript*

[5] Weiss S. “Learning to disambiguate”, *Information Storage and Retrieval*, v.9, 1973

Использование конечных автоматов для морфологического анализа и синтеза на основе словарей системы ЭТАП

Казенников А.О.
Лаборатория компьютерной лингвистики ИППИ РАН
kzn@iitp.ru

Аннотация

В настоящей работе рассматривается применение методов, основанных на конечных автоматах для морфологического синтеза и анализа с использованием словарей системы ЭТАП. В настоящее время в системе ЭТАП морфологический анализ/синтез производится на основе таблиц. Такое устройство предполагает наличие раздельных таблиц для анализа и синтеза. Подход на основе конечных автоматов позволяет ускорить работу анализа и синтеза, а также существенно уменьшить объем дублирующейся информации. В работе рассматриваются вопросы конструирования и поддержки минимальности конечных автоматов и конечных преобразователей, а также возможность встраивания используемых методов в систему ЭТАП.

1. Введение

Первым шагом при работе системы ЭТАП является морфологический анализ. На этом шаге из словоформы получается лексема и ее характеристики. Кроме собственно морфологического анализа, происходит разбиение текста на отдельные составляющие — токены - и последующий их анализ. На вход морфологического модуля подается цепочка символов, а на выходе получаем шаблон фразы с разметкой (пунктуацией) и списком омонимов для каждого слова. Кроме того, морфологический модуль используется на самом последнем этапе — на этапе синтеза, когда лексема и ее список признаков преобразуется в законченную словоформу.

2. Описание и анализ существующего алгоритма морфологического анализа/синтеза

Задача морфологического анализа состоит в получении морфологической структуры входного текста. Морфологический анализ поочередно выделяет из входного текста словоформы и определяет для каждой из них список возможных лексем и морфологических характеристик.

Морфологический компонент состоит из двух частей :

- компиляция словарей в машинный вид:
- морфологический анализ/синтез на основе машинного вида словарей.

Статья словаря представляет собой декларативное морфологическое описание лексемы по частям слова — приставка, основа, тема, суффикс, окончания, частица[1]. Такое описание нельзя непосредственно использовать для анализа или синтеза, и поэтому необходимо некоторое оптимизированное представление статей словаря, которое строится на этапе компиляции.

При компиляции происходит последовательная обработка каждой статьи словаря. Она разворачивается в отсортированный список словоформ, каждая из которых разбита на две части – т.н. «голову» и «хвост». В состав «головы» входит приставка и корень, в состав «хвоста» — все остальные части слова. В текущей реализации словоформы с общей головой объединяются в новые списки. Хвосты словоформ такого списка составляют собой парадигму. Парадигмы являются общими в рамках одного языка и записываются в отдельную таблицу парадигм. Потом формируются структуры данных для записи на диск — две таблицы, первая, индексом которой является голова (обратная таблица), для синтеза, вторая — индекс которой лексема (прямая таблица), для анализа.

Анализ входного текста происходит итеративно. На каждом шаге анализа разбираемая словоформа представляется в виде двух частей: «головы» и «хвоста». После этого производится поиск «головы» словоформы по обратной таблице. Если такие записи найдены, то проверяется соответствие «хвоста» словоформы «хвостам» парадигмы. Если «хвост» совпал, то из записи извлекается лексема и признаки, и такой вариант разбора этой словоформы считается возможным. Процесс продолжается до исчерпания всех возможных вариантов.

Такой подход к анализу обладает следующими свойствами:

1. Простота архитектуры
2. Относительная простота реализации
3. Инкрементальность компиляции

Недостатки:

1. Необходимость хранить две таблицы с практически одинаковыми данными.
2. Поиск по таблице почти каждой буквы словоформы при анализе.

3. Использование конечных автоматов для анализа/синтеза

Современные развитие аппаратной части компьютеров позволяет использовать конечные автоматы для морфологического анализа и синтеза. Основное преимущество КА — высокая скорость поиска — время линейно зависит от длины искомого слова. Для описания морфологии обычно используется детерминированный конечный автомат (ДКА).

Формально ДКА можно представить в виде кортежа M = (Q, Σ, δ, q_0, F), где Q – конечное множество состояний автомата, $q_0 \subseteq Q$ – начальное состояние, F in Q – множество конечных состояний, S – алфавит(множество возможных входных символов), δ: Q × S → Q функция перехода. На основе δ вводится расширенная функция перехода:

$$\delta^*(q,\varepsilon) = q$$

$$\delta^*(q,ax) = \begin{cases} \delta^*(\delta(q,a),x), \text{ если } \exists\delta(q,a) \\ \varnothing \end{cases}$$

Для морфологического анализа/синтеза, используемого в ЭТАПе, можно использовать более специализированный ДКА, обладающий свойством ацикличности — при котором некоторое состояние при обходе любой строки символов может встречаться не более одного раза. Это следует из того, что число словоформ конечно, а словоформа имеет конечную длину. Кроме того, необходимо расширить понятие конечного состояния. В настоящей работе с конечным состоянием автомата связан не только факт принятия входной цепочки символов, но и возможные признаки принятой цепочки. Таким образом, при обходе автомата на выходе получается не только факт существования/отсутствия цепочки в автомате, но и то, какими признаками (характеристиками) она обладает.

Промежуточным представлением при переходе от результата трансляции статей в виде списка словоформ со списком признаков, к КА является префиксное дерево(trie). Хотя префиксное дерево не является автоматом, но легко преобразуется в автомат. КА на основе префиксного дерева обладает всеми необходимыми свойствами (простота реализации, скорость поиска, инкрементальность), но неэкономичен по памяти (Рис. 1).

4. Минимизация конечных автоматов

Для эффективного использования памяти применяется минимизация автомата. Минимизация связана с понятием эквивалентности состояний. Два состояния считаются эквивалентными, если эквивалентны множества $\{x \ni \Sigma^* \mid \delta^*(q,x) \in F\}$ этих состояний. Все состояния автомата можно разбить на классы эквивалентности. Основная идея минимизации состоит в том, что в автомате все состояния из одного класса эквивалентности заменяются каким-то одним представителем этого класса, и, таким образом каждый класс представлен только одним состоянием, и, следовательно, все состояния автомата являются уникальными и неэквивалентными.

Для практического использования проверка эквивалентности состояний является неэффективной, поскольку по времени она экспоненциальна. Поэтому минимизация автомата осуществляется при обратном порядке обхода состояний. Тогда для определения эквивалентности достаточно их равенства.

Для ацикличного ДКА используются следующие проверки для определения равенства двух состояний:

1. Если оба состояния обладают одинаковыми списками конечных признаков,
2. Если у них одинаковое число исходящих переходов,
3. Если метки переходов совпадают,
4. Если переходы ведут к одним и тем же состояниям.

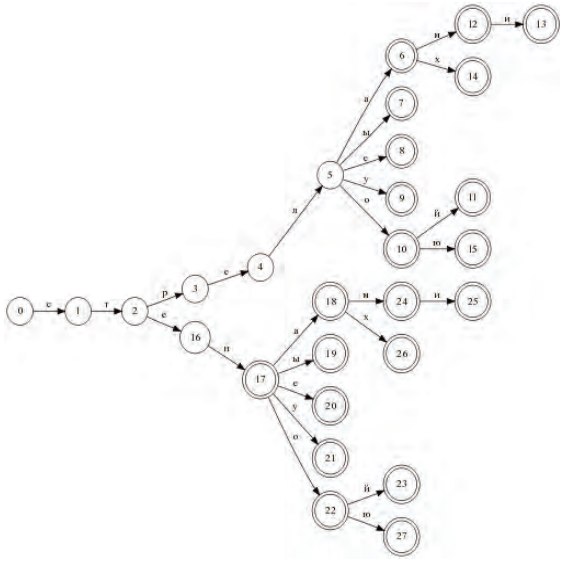


Рис 1. КА на основе префиксного дерева для статей словаря «стрела» и «стена»

5. Базовый алгоритм минимизации конечных автоматов

Самый простой алгоритм построения минимального автомата состоит из двух шагов:

1. построения автомата на основе префиксного дерева.
2. минимизация автомата при обратном обходе этого дерева.

Его основным достоинством является простота и скорость. Но такой алгоритм построения требует большой объем памяти для построения префиксного дерева, и не является инкрементальным — после того, как автомат был минимизирован, нет возможности добавлять к нему новые слова.

6. Инкрементальный алгоритм построения минимального ацикличного конечного автомата

В работе [2] предложен эффективный инкрементальный алгоритм минимизации ациклических ДКА. В алгоритме вводится понятие повторно используемого состояния — такого состояния, у которого более одного входящего перехода(состояние 5 из рис.2). Суть алгоритма состоит в том, что помимо минимального ДКА поддерживается реестр уникальных представителей классов эквивалентности и каждое новое состояние по возможности заменяется представителем существующего класса или добавляется как новый класс.

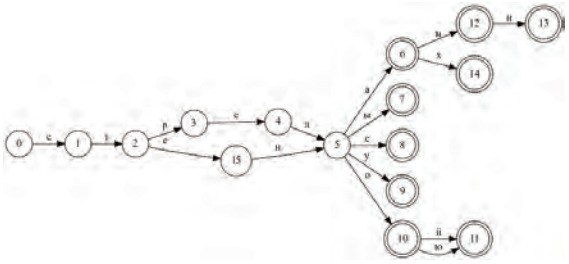


Рис 3. Минимальный КА для статей словаря «стрела» и «стена»

При добавлении новой строки она условно разбивается на две части: префикс и суффикс. Префикс соответствует максимальной общей подстроке, существующей в ДКА. Кроме того, запоминается последовательность состояний соответствующая префиксу. Если в этой последовательности есть хотя бы одно повторно используемое состояние, то оно клонируется, а так же клонируются все состояния, которые после него. После этого, к последнему состоянию последовательности(или его клону) добавляется цепочка состояний соответствующая суффиксу

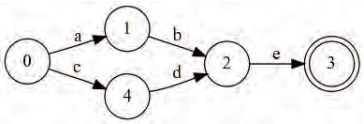


Рис 2. Минимальный КА со словами «abe» и «cde»

Затем происходит процесс минимизации цепочки состояний, соответствующей добавляемой строке. Эта цепочка обходится в обратном порядке, и если у текущего состояния есть эквивалентное, происходит замена состояний, или же текущее состояние добавляется в реестр состояний.

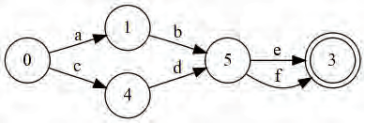


Рис 4. Некорректное добавление к КА «cdf»

Клонирование повторно используемых состояний в префиксе необходимо для предотвращения добавления лишних слов в автомат. Пусть есть автомат, распознающий слова «abe» и «cde»(Рис. 3). Необходимо добавить слово «cdf». Если просто добавить суффикс к состоянию 2, то в автомат так же добавиться слово «abf»(Рис. 4). Для корректного добавления необходимо клонировать состояние 2 и уже к нему добавить суффикс(Рис. 5).

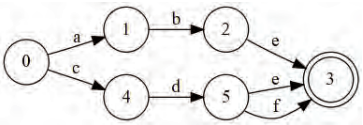


Рис 5. Корректное добавление к КА «cdf»

7. Алгоритм инкрементальной минимизации конечного автомата

На входе: fsa(минимизированный конечный автомат), word(добавляемое слово), features(характеристики слова, которые добавляются к конечному состоянию).

Предварительные действия:

1. Нахождение максимального общего префикса (common-prefix)
2. Поиск повторно используемых состояний в префиксе(confluence-index)
3. suffix-state — последнее состояние common-prefix. К нему будет добавляться суффикс слова.

Основная часть алгоритма, выполняется только если есть повторно используемые состояния. (confluence-index > 0):

1. клонирование состояний от confluence-index до последнего состояния common-prefix
2. Установка suffix-state равным последнему клону.

Заключительная часть:

1. Добавление к suffix-state суффикса слова
2. Добавление к последнему состоянию суффикса features.
3. Поиск и замена неуникальных состояний.

Функция поиска и замены неуникальных состояний обходит состояния слова в обратном порядке. Для каждого состояния:

1. Поиск состояния, эквивалентного данному.
2. Если такое найдено, то заменить текущее состояние найденным
3. Если нет, то добавить текущее состояние в реестр состояний.

Алгоритм, приведенный выше, отличается от того, что был в [2]. Представленный алгоритм является итеративным, и все его процедуры тоже. Это упрощает реализацию, а так же более эффективно, если компилятор языка реализации КА не поддерживает оптимизацию хвостовой рекурсии. Кроме того, алгоритм более общий — в нем все случаи, как граничные, так и нет обрабатываются единообразно.

Еще одно отличие связано с моментом, когда проводить минимизацию: в [2] предлагалось минимизировать состояния суффикса сразу после добавления. Такой подход более экономичен по памяти — автомат минимален даже во время добавления новой строки, но более затратен по времени — после минимизации необходимо убеждаться в отсутствии циклов.

8. Проблема полной инкрементальности

Этот алгоритм является инкрементальным: после каждого нового слова автомат остается минимальным. Но все равно сохраняется проблема удаления слова из автомата. Полную инкрементальность можно было бы реализовать следующим образом:

1. Для каждого конечного состояния хранить число словоформ, анализ которых приводит в это состояние.
2. Сохранять для каждого состояния все входящие переходы, а не только количество различных входящих переходов .

Недостаток такого решения в том, что резко увеличивается относительный объем данных для каждого состояния.

Возможна эмуляция свойства инкрементальности, когда ДКА сам по себе не является полностью инкрементальным, а некоторая структура, в которой он состоит — является. В настоящей работе использовался дополнительный автомат, который хранит удаленные лексемы и словоформы. Этот автомат не является минимальным, для того, чтобы в него самого можно было бы добавлять и удалять слова.При анализе слово сначала прогоняется через распознающий КА, а потом проверяется отсутствие слова в КА удаленных слов. Время проверки словоформы остается линейным, хотя и увеличивается. Для оптимизации скорости проверки необходима полная компиляция словаря, если число удаленных элементов больше некоторого.

9. Использование конечных преобразователей

В настоящей работе пока рассматривался только конечный автомат. С помощью него можно получить только признаки заданной словоформы. Более мощным средством является конечный преобразователь (КП). Преобразователь отличается от автомата наличием выходного алфавита и еще одним аргументов в функции перехода. В [3] показано, что КП эквивалентен КА, в котором вместо раздельных входного и выходного символа они рассматриваются как неделимая пара. Таким образом, алгоритмы, применяемые для минимизации КА применимы к КП.

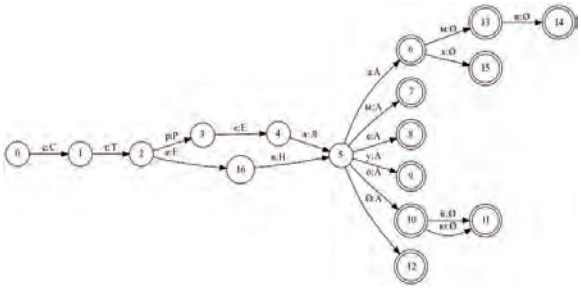


Рис 6. минимальный КП для статей словаря «стрела» и «стена»

При использовании КП изменяется сложность

обхода — при обходе только по одному символу (входному или выходному) нарушается свойство детерминированности автомата. Таким образом, в наихудшем случае время экспоненциально зависит от длины словоформы. (Вырожденный случай, когда все словоформы словаря принадлежат одной лексеме).

10. Обработка неизвестных слов

Кроме компактности, подход на конечных автоматах позволяет делать предположения о признаках слов, которых нет в словаре. Для этого строится конечных автомат на основе окончаний (хвостах парадигмы), взятых в обратном порядке. Получается достаточно компактный автомат(около 1000 состояний), который хорошо делает догадки о неизвестных словах.

11.Результаты

Все результаты получены на русском морфологическом словаре системы ЭТАП. Словарь содержит 123 тыс. статей, что составляет около 4.5 млн. словоформам.

Размер минимального конечного автомата: 109 тыс. состояний и 280 тыс. переходов, конечного преобразователя - 124 тыс. состояний и 300 тыс. переходов.

Минимизация значительным образом влияет на размер автомата. Минимальный автомат меньше автомата на основе префиксного дерева в 8-10 раз.

Построение минимальных КА/КП для морфологического анализа/синтеза является весьма компактным и эффективным представлением морфологических данных.

Время построения КА/КП составляет около 5 минут, что сравнимо с временем компиляции словарей по существующему алгоритму.

12. Благодарности

Автор выражает благодарность Н. В. Григорьеву за идею работы. А так же В. Г. Сизову за ценные замечания.

Литература

1.Ю.Д. Апресян, И.М. Богуславский, Л.Л. Иомдин, А.В. Лазурский, Н.В. Перцов, В.З. Санников, Л.Л. Цинман. Лингвистическое обеспечение системы ЭТАП-2. М. Наука 1989.
2.J. Daciuk; B. W. Watson; S. Mihov; R. E. Watson. «Incremental Construction of Minimal Acyclic Finite-State Automata».Computational Linguistics, Vol. 26, Number 1, 2000
3.E. Roche, Y. Schabes. «Finite State Language Processing». MIT Press.1997.

Извлечение информации о сочетаемости лексем из аннотированного корпуса текстов[1]

Дяченко Павел, Фролова Татьяна
Лаборатория компьютерной лингвистики ИППИ РАН
pavelvd@cl.iitp.ru, tfrolova@cl.iitp.ru

Аннотация

В настоящем докладе описывается извлечение из текстов информации о лексической сочетаемости слов, которая должна быть представлена в словарных статьях АОТ-систем для корректной обработки ими идиоматичных структур. Известно, что существует определенный набор синтаксических контекстов, в которых регулярно встречаются несвободные словосочетания. Таким образом, есть возможность получать информацию о несвободной лексической сочетаемости слов из синтаксически размеченного корпуса текстов. Для данного исследования используется морфологически и синтаксически размеченный корпус текстов «СинТагРус» объемом в 35 000 тысяч предложений, который был создан в Лаборатории компьютерной лингвистики ИППИ РАН. Из этого корпуса были автоматически выбраны определенные синтаксические контексты для пробного списка слов. Для пополнения и уточнения информации о сочетаемости в словарных статьях из результатов поиска были выбраны нестандартные словосочетания.

1. Лексическая сочетаемость слов и ее описание

В словарь системы машинного перевода, в частности, в словарь системы ЭТАП-3, должна быть помещена информация о способах анализа и перевода несвободных сочетаний слов. Например, в таких словосочетаниях как *бросать взгляд*, *накладывать ограничения*, *глубокий кризис* первое слово не может быть понято в своем основном значении и, возможно, должно переводиться на другие языки специальным образом. Поэтому в словарную статью одного из членов такого словосочетания вносится информация о том, как анализировать и переводить всё это словосочетание.

1.1. Лексические функции как способ описания несвободных словосочетаний

Было замечено, что существуют некоторые смыслы, которые выражаются идиоматично в разных языках и при разных словах. Для описания таких смыслов А.К.Жолковским и И.А.Мельчуком в [1,2] был предложен аппарат лексических функций. Например, прилагательное ГЛУБОКИЙ в словосочетании *глубокий кризис* в таком представлении является значением лексической функции MAGN. Лексическая функция MAGN ставит в соответствие слову X другое слово или словосочетание, обозначающее большую степень или интенсивность X-а (здесь и далее определения лексических функций приводятся по статье [3]). То же значение выражено нестандартным образом при других словах, ср. прилагательное БУРНЫЙ в словосочетании *бурная деятельность*, а также в других языках, ср. английское прилагательное SEVERE в словосочетании *severe crisis*. Было замечено также, что таких смыслов и, соответственно, таких лексических функций имеется лишь несколько десятков.

Значением лексической функции являются слова или словосочетания не только с определенным смыслом, но и выполняющие при данном слове X определенную синтаксическую функцию. Так, например, значением лексической функции MAGN для существительных являются прилагательные (ср. *глубокий* кризис) выполняющие при X-е функцию синтаксического определения.

[1] Работа частично поддержана средствами грантов РФФИ № 08-06-00344 и РФФИ № 07-06-00339.

1.2. Представление информации о лексической сочетаемости в комбинаторном словаре системы ЭТАП-3

В комбинаторном словаре системы ЭТАП-3 (более подробно о системе см. [4-6]) информация о способе анализа и перевода лексикофункциональных словосочетаний должна быть помещена в словарную статью слова X в специальной зоне лексических функций. Так, в словарной статье слова КРИЗИС в зоне лексических функций имеется запись «MAGN: *глубокий*». В комбинаторном словаре другого языка переводной эквивалент русского слова также снабжен информацией о способе выражения лексической функции MAGN, например, в словарной статье слова CRISIS английского комбинаторного словаря имеется запись «MAGN:*severe*». Это обеспечивает возможность адекватного перевода данной структуры.

Информация о нестандартной сочетаемости лексем, не поддающейся классификации в рамках системы лексических функций, описывается специальными правилами синтаксического анализа и перевода в словарной статье одного из членов такого словосочетания.

1.3. Источники сведений о лексической сочетаемости

В словарь системы ЭТАП-3 информация о лексической сочетаемости, в частности, информация о лексических функциях поступает либо непосредственно от составителя словаря (лингвиста и/или носителя данного языка) в результате интроспекции, либо из словарей данного языка: словарей сочетаемости, а также толковых и двуязычных словарей.

В последнее время, благодаря появлению корпусов текстов с разной глубиной разметки, стало возможным использовать свойство лексических функций выступать в стандартных синтаксических контекстах и получать информацию о нестандартной сочетаемости лексем из текстов. Так, например, для получения данных о лексической функции MAGN и других адъективных лексических функциях для данного слова можно осуществлять поиск слов и словосочетаний, зависящих от данного по определительному отношению в синтаксически размеченном корпусе.

2. Пробный поиск несвободных словосочетаний в корпусе текстов

В рамках данного исследования был осуществлен поиск данных о значениях адъективных лексических функций для ряда существительных и о тех нестандартных коллокатах этих существительных, которые имеют схожее с лексическими функциями синтаксическое оформление, на материале морфологически и синтаксически размеченного корпуса текстов «СинТагРус» объемом в 35 000 тысяч предложений, который был создан в Лаборатории компьютерной лингвистики ИППИ РАН. Использовалась полная версия корпуса по состоянию на июнь 2008 года.

Выбор контекстов с определенными синтаксическими условиями и сортировка этих контекстов по встречающимся в них прилагательным осуществлялся автоматически.

Алгоритм поиска представляет собой последовательный просмотр всех предложений размеченного корпуса. Для каждого предложения проверялось выполнение контекстного условия, которое на данном этапе исследования было сформулировано заранее и описано непосредственно в коде программы. В то же время, были выделены настраиваемые параметры поиска – например, объем просматриваемого корпуса, тип синтаксической связи между узлами контекста, а также морфологические характеристики интересующих нас коллокатов. Отобранные программой поиска контексты автоматически сортировались по частотности и сохранялись для последующей обработки.

Выбор нестандартных коллокатов, данные о которых должны быть помещены в словарь, осуществлялся вручную.

2.1. Условия поиска

Основные результаты для адъективных лексических функций и подобных им сочетаний получены в результате автоматического анализа определительных контекстов. То есть по заданному существительному находились прилагательные, зависящие от заданного существительного по определительному отношению, такие как прилагательное ГЛУБОКИЙ в словосочетании *глубокий кризис*. Ограничения на часть речи синтаксического определения не вводились, однако на пробном списке существительных были найдены и отобраны как нестандартные только словосочетания, в которых определительным отношением с заданными существительными были связаны прилагательные. Определения, относящиеся к другим частям речи, составили с заданными существительными свободные словосочетания, которые нет необходимости специальным образом описывать в словаре.

Для получения более полного результата выполнялся также поиск с более сложными синтаксическими условиями. Например, поиск прилагательных, которые связаны с синтаксическим определением заданного слова сочинительным отношением, таких как прилагательное ГЛУБОКИЙ в словосочетании *продолжительный и глубокий кризис*. А также прилагательных, выступающих в предикативной позиции в именных предложениях и предложениях с глаголом-связкой, таких как прилагательное ГЛУБОКИЙ в предложениях *кризис глубок* или *кризис был глубоким*.

Таким образом, всего для адъективных коллокатов было рассмотрено три типа контекстов. Такое расширение параметров поиска позволило получить очень небольшое количество дополнительных результатов (всего получено 222 нестандартных сочетания с прилагательными, из них лишь одно – в результате дополнительного поиска по контекстам с сочинительными и предикативными синтаксическими отношениями).

2.2. Интерпретация результатов

Из полученных 222 словосочетаний около четверти (53) могут быть описаны как сочетание существительного со значением лексической функции MAGN (определение см. выше), например, словосочетание *серьезное обвинение*. Двенадцать прилагательных являются значениями лексической функции BON (выражает стандартную положительную оценку X-а) от соответствующих существительных, например, словосочетание *теплый прием*. В одиннадцати случаях полученные прилагательные можно описать как значение лексической функции ANTIMAGN (обозначает небольшую степень X-а), например, в словосочетании *небольшое количество*. В трех случаях, в словосочетаниях *ошибочный подход*, *плохая работа*, *черное дело*, прилагательные являются значениями лексической функции ANTIBON (выражает стандартную отрицательную оценку X-а), в двух случаях, в словосочетаниях *прямое назначение*, *трезвый расчет* – значениями лексической функции VER (имеет значение свойства, которое нормально ожидается или требуется от X-а), в двух словосочетаниях *бесплодный поиск*, *неосуществленный проект*, – значениями лексической функции ANTIVER (имеет значение свойства, противоположного тому, которое нормально требуется или ожидается от X-а). Более половины словосочетаний (139) не могут быть описаны в терминах лексических функций и должны обслуживаться в словаре особыми правилами перевода, например, словосочетание *лесное хозяйство*.

3. Оценка результатов экспериментального поиска и их практическое применение

Полученные результаты были сопоставлены с теми сведениями о сочетаемости слов, которые приводятся в русском электронном словаре ABBYY Lingvo 12 и в Большом русско-английском словаре под ред. А. И. Смирницкого, представленном в электронном словаре МультиЛекс. В то же время, результаты эксперимента были применены для улучшения качества автоматического перевода в системе ЭТАП-3.

3.1. Сравнение с лексикографическими источниками

Из 222 несвободных словосочетаний, найденных в результате пробного поиска по корпусу, только 119 имеется хотя бы в одном из двух упомянутых выше словарей, а остальные 103 словосочетания не представлены ни в одном из них.

Таким образом, для описания лексической сочетаемости в словарных статьях текстовые данные представляют собой существенное дополнение к лексикографическим источникам.

3.2. Практическое применение результатов эксперимента в системе ЭТАП-3

Автоматический перевод выделенных словосочетаний на английский язык при помощи системы ЭТАП-3 дал следующие результаты: для 120 словосочетаний из выделенных 222 был получен удовлетворительный перевод, остальные 102 словосочетания были переведены неверно. После внесения информации об этих 102 словосочетаниях в русский комбинаторный словарь системы ЭТАП-3 был получен правильный перевод.

4. Перспективы метода

Поиск коллокатов слов на материале синтаксически размеченного корпуса предполагается развивать как путем расширения списка слов-аргументов лексических функций, так и путем добавления к условиям поиска других синтаксических контекстов, характерных для глагольных, субстантивных и адвербиальных

лексических функций. Ниже приведены примеры дельнейшего применения метода поиска по контекстам.

4.1. Поиск по значениям лексических функций

Поиск по текстам можно использовать и другим образом. Если для составления лексикографических портретов слов актуален поиск различных лексикофункциональных контекстов данного слова, то для систематического пополнения словаря поиск может быть «перевернут».

Несмотря на то, что выбор данного значения лексической функции для данного слова не может быть с точностью предсказан, имеется тенденция к семантической мотивированности этого выбора, как это было показано в [7-9]. То есть определенное значение лексической функции может быть характерным для целого класса слов-аргументов, обладающих общими семантическими признаками. Так, в 83 лексикофункциональных словосочетаниях, найденных в результате пробного поиска, описанного выше, встречается только 51 прилагательное.

Таким образом, поиск синтаксически схожих несвободных словосочетаний можно осуществлять не только по словам-аргументам, но и по значениям лексических функций, повернув стрелки в условиях поиска. Например, по прилагательному ОЖЕСТОЧЕННЫЙ, являющемуся значением лексической функции MAGN от существительного БОРЬБА, можно найти все существительные, от которых данное прилагательное зависит по определительному отношению. Такой поиск даёт ещё четыре лексикофункциональных словосочетания, информация о которых должна быть помещена в словарные статьи существительных СТОЛКНОВЕНИЕ, СОПРОТИВЛЕНИЕ, БОЙ и КОНКУРЕНЦИЯ.

При этом поиск по корпусу имеет некоторое преимущество по сравнению с извлечением информации из словарей. При поиске по корпусу словосочетания будут найдены независимо от того, заданы ли условия поиска по аргументам или по значениям лексических функций.

В то же время, например, при составлении лексикографического портрета слова по материалам словарей часть приведенных в словарях словосочетаний не будет найдена, поскольку эти словосочетания описываются в словарных статьях значений лексических функций, а не аргументов (ср. описание сложностей, возникающих в связи с таким свойством словарей в процессе обучения языку в [10]). Так, например, из 222 словосочетаний, представленных в словарях и одновременно являющихся результатами пробного поиска (ср. пункт 3 выше), только 36 находятся в словарных статьях соответствующих существительных, остальные 83 помещены в словарные статьи прилагательных, поэтому узнать об их наличии из словарной статьи существительного невозможно.

4.3. Поиск в морфологически размеченном корпусе

Дальнейшей перспективой развития метода является поиск в корпусе с морфологической разметкой. Поиск по Национальному корпусу должен дать значительно больше результатов, чем поиск по синтаксически размеченному корпусу, из-за существенно большего объема этого корпуса. В то же время, если задавать условия поиска только в терминах возможных морфологических признаков и возможного относительного расположения слов в предложении, уровень «шума» будет достаточно высок. Для исключения некоторого количества лишних контекстов из результатов поиска по Национальному корпусу кажется оправданным наложить дополнительные ограничения на относительное расположение и морфологические характеристики слов в словосочетаниях. Результаты поиска по синтаксически размеченному корпусу позволяют сформулировать предположения относительно содержания этих ограничений.

Так, например, выше было замечено, что, хотя значением лексической функции MAGN в принципе могут являться не только прилагательные, в итоговый список нестандартных определений-коллокатов для пробного списка существительных попали только они. С другой стороны, из 471 вхождения 222 несвободных словосочетаний только в 42 между прилагательным и существительным встречались другие слова, и только в 6 прилагательное следовало за существительным. Таким образом, ограничив поиск только прилагательными и только непосредственно предшествующими существительному, можно существенно снизить уровень «шума» в результатах поиска, лишь в достаточно небольшой степени пожертвовав полнотой материала.

[1] А.К. Жолковский, И.А. Мельчук. «О семантическом синтезе», *Проблемы кибернетики. Вып. 19.* М.: Наука, 1967, с. 177-238.

[2] И.А. Мельчук. *Опыт теории лингвистических моделей «Смысл – Текст»*. М., Наука, 1974.

[3] Ю.Д. Апресян, П.В. Дяченко, А.В. Лазурский, Л Л.Цинман. «О компьютерном учебнике лексики русского языка», *Русский язык в научном освещении*. №2(14), 2007, с. 48-112.

[4] Ю.Д. Апресян, И.М. Богуславский, Л Л. Иомдин, А.В. Лазурский, Н.В. Перцов, В.З. Санников, Л Л.Цинман. *Лингвистическое обеспечение системы ЭТАП-2*. М: Наука, 1989.

[5] Ю.Д. Апресян, И.М. Богуславский, Л Л. Иомдин, А.В. Лазурский, Л.Г. Митюшин, В.З. Санников, Л Л.Цинман. *Лингвистический процессор для сложных информационных систем*. М: Наука, 1992.

[6] Ju.D. Apresian, I.M. Boguslavsky, L.L. Iomdin, A.V.Lazursky, V.Z. Sannikov, V.G. Sizov, L.L. Tsinman. «ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the MTT», *MTT 2003, First International Conference on Meaning – Text Theory*. Paris: École Normale Supérieure, 2003, p. 279-288.

[7] Ю.Д. Апресян. «О семантической непустоте и мотивированности глагольных лексических функций», *Вопросы Языкознания № 4*, 2004, с. 3-18.

[8] Ю.Д. Апресян. «Лексические функции на службе компьютерной лингвистики» *Прикладна лінгвістика та лінгвістичні технології MEGALING-2006. Збірник наукових праць*. Київ, 2007, с. 7-20.

[9] Ю.Д. Апресян. «О семантической мотивированности лексических функций-коллокатов», *Вопросы Языкознания № 5*, 2008, (в печати).

[10] *Margarita Alonso Ramos*. «Glosas para las colocaciones en el Diccionario de Colocaciones del Español». *Diccionario y Fraseología. ed. by M. Alonso Ramos*. Coruña: Universidade da Coruña, 2006, p. 59-88.

Автоматическое уточнение правил в самозагрузочной системе извлечения информации из текстов на естественном языке

Мстислав Масленников
Московский Государственный Университет
steve@oit.cmc.msu.ru

Аннотация

По сравнению с полуструктурированными текстами, извлечение информации из текстов на естественном языке значительно сложнее, так как требуется аннотирование значительного количества тренировочных примеров. В текущей работе мы описываем новую самозагрузочную систему на основе правил, целью которой является интеграция эвристического знания от экспертов и обучение на основе правил. Стартуя с небольшого количества начальных эвристических правил, наша система последовательно создаёт новые правила, новые тренировочные примеры и повышает качество извлечения информации в рамках самозагрузочной модели обучения. Дополнительно, мы впервые вводим процедуру прямого и обратного уточнения правил для улучшения качества обучения.

1. Введение

Извлечение Информации, особенно Извлечение Собственных Имён, является критически важной технологией для поддержки вопросно-ответных систем, систем информационного поиска и понимания документов. Извлечение Собственных Имён (СИ) из полуструктурированных текстов значительно проще по сравнению с текстами на естественном языке. Это происходит из-за избыточности данных в полуструктурированных текстах, что значительно упрощает задачу по Извлечению Собственных Имён. Однако, наиболее полезная информация содержится в форме текстов на естественном языке, которая не соответствует любому строгому шаблону или структуре. Чтобы моделировать её корректно, требуется большое количество аннотированных данных. Существующие алгоритмы обработки текстов на естественном языке показывают, что это условие должно быть удовлетворено для получения удовлетворительного качества извлечения информации. Например, [28] сообщили о получении качества по метрике F_1 вида 94.1% на естественных текстах, используя весь доступный набор документов для обучения. В то же самое время, они утверждают что небольшое уменьшение в количестве тренировочных примеров ведёт к значительному ухудшению качества обучения. [2] заметили, что качественные алгоритмы обучения требуют по крайней мере двукратного увеличения тренировочных экземпляров, чтобы сократить количество ошибок наполовину.

Таким образом, задача получения большого количества аннотированных примеров зачастую является критически важной. Это обычно скучная и дорогая рутинная работа, которая обычно делается вручную. Последовательно, требуется автоматизированный метод для извлечения тренировочных примеров. В настоящий момент, самозагрузочный подход для автоматизированной аннотации и генерации правил был исследован в различных статьях ([1], [17]). Эти исследователи предложили, что неразмеченные данные могут быть полезными для сокращения необходимости для ручной аннотации набора текстов. Однако, большинство из них по-прежнему не достигают разумного результата без интенсивного участия человека [26], такого как через *активное обучение* [23]. Наилучший существующий метод [17][1] для текстов на естественном языке, использующий самозагрузку, достигает качества только 87.7% для извлечения людей, 82.3% для извлечения местоположений и 52.7% для организаций.

Другая мотивация нашего исследования состоит в том, что текущие наилучшие системы не используют достоинства эвристических и сгенерированных правил одновременно. Эвристические правила могут быть написаны экспертами по тематике для эффективного описания частей данных. На практике, их

[1] Использует 35 стартовых ключевых слов на корпусе MUC-7

получение экспертных правил обходится дешевле аннотирования большого набора данных. Однако, требуемое количество эвристических правил зачастую недостаточно для покрытия всех случаев Извлечения Информации. С другой стороны, сгенерированные правила имеют тенденцию быть слишком специфичными и требуют большого количества тренировочных примеров для обобщения. Существует также опасность чрезмерной тренировки или ошибок в тренировочных данных [5]. Таким образом, интеграция эвристических и сгенерированных правил является важной для практического Извлечения Информации.

В этой статье мы представляем новую самозагрузочную систему для получения правил для извлечения собственных имён из текстов на естественном языке. Наша система стартует только с небольшого количества первоначальных примеров и эвристических правил. Затем она постепенно генерирует новые правила, новые тренировочные примеры и улучшает качество Извлечения Информации в рамках самозагрузочной модели. Основной вклад нашего исследования состоит в разработке подхода по генерации правил, который в состоянии интегрировать эвристические правила и автоматически скорректировать ошибочные правила самостоятельно. Дополнительным вкладом является стратегия по уточнению полученных правил.

Оставшиеся части статьи организованы следующим образом. Раздел 2 описывает общую модель, стратегии генерации правил и автоматическое уточнение правил. Раздел 3 представляет экспериментальные результаты нашей системы. Раздел 4 обсуждает и сравнивает нашу парадигму с существующими похожими подходами. Наконец, раздел 5 завершает статью.

2. Самозагрузка для Извлечения Информации

2.1 Общая модель

Наша модель состоит из 3 частей: генерация правил, интеграция правил и уточнение правил. Мы стартуем весь самозагрузочный подход с использования небольшого множества эвристических правил (ЭП) и тренировочных примеров. Одно из ЭП, использованное для класса ЧЕЛОВЕК выглядит как:

HumCap -> Capwords & !список(nationality)

PersonPeriod -> HumCap+ Cap sym_. HumCap

ЧЕЛОВЕК -> список(personTitle) PersonPeriod

Для фрагмента “Mr. George W. Bush Jr” оно извлечёт ‘George’ как *HumCap*, ‘W’ как *Cap*, ‘.’ как *sym*, после этого ‘Bush’ как *HumCap*, и наконец вернёт ‘George W. Bush’ как *PersonPeriod* и после этого как ЧЕЛОВЕК (цель для возвращённого значения подчёркнута). Эти ЭП получены из более ранней эвристической системы извлечения Собственных Имён ([25]). Так как эти ЭП являются общими благодаря их общим типам шаблонов, они в состоянии извлечь достаточно большой набор новых примеров (положительных и отрицательных) из неразмеченного корпуса для генерации новых правил с помощью машины. Эти примеры, полученные во время первого цикла, могут быть в дальнейшем случайным образом выбраны и, в случае необходимости, верифицированы через взаимодействие с пользователем. Однако, наши результаты показывают, что это не является абсолютно необходимым. Единственное усилие, требуемое от человека, состоит в подготовке нескольких первоначальных правил и примеров для каждого класса СИ.

Архитектура системы представлена на Рисунке 1. Чтобы инициализировать систему, первоначальные ЭП наполняют *множество надёжных тренировочных примеров* и *текущий набор правил*. Множество надёжных тренировочных примеров используется для тренировки *валидатора* на основе метода Максимальной Энтропии (МЭ) для проверки и валидации на каждом цикле. Используя текущий набор правил, мы извлекаем новые примеры на множестве *неразмеченных данных* и начинаем итерацию. Новые примеры пропускаются через МЭ валидатор для проверки, и валидированные примеры добавляются в *набор нормальных тренировочных данных* для генерации правил. Генерация правил основана на получении новых правил с помощью *генератора правил* и *уточнителя правил* (см. следующие разделы). Полученное окончательное множество правил затем применяется снова к неразмеченным данным для получения дополнительных примеров для следующей итерации.

Не все новые примеры, извлечённые в каждом цикле, должны быть добавлены ко множеству надёжных данных в МЭ модели. Это потому-что такое добавление бы вызвало быстрое распространение ошибок и привело бы к плохому общему качеству [17]. Следовательно, мы лишь добавляем одно множество новых примеров ко множеству надёжных тренировочных данных после каждых N итераций, где N выбирается динамически на основе того, что новые извлечённые данные точны. Чтобы оценить эту точность для примеров каждого класса, мы измеряем качество правил, использованных для

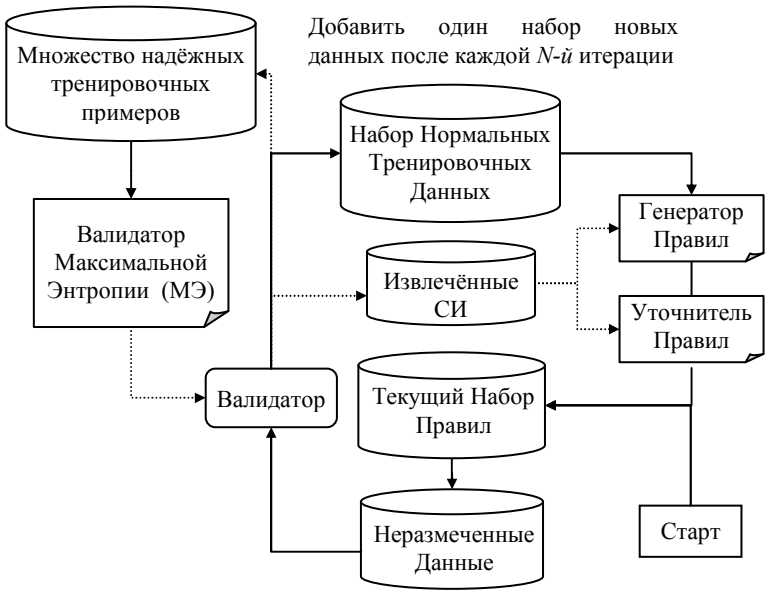


Рисунок 1. Самозагрузочная модель

их извлечения, базируясь на существующем множестве надёжных примеров, на основе:

$$GO(\{R_1,\cdots,R_n\}) = \frac{1}{n}\sum_{i=1}^{n} Точность(R_i) * Полнота(R_i) \quad (1)$$

Предполагая что правила упорядочены на основе точности, задача сводится к подбору наилучших m правил для каждого класса таких, что $GO \geq GO^0$, предопределённое пороговое значение. Мы затем добавляем новое множество неразмеченных примеров, извлечённых с помощью выбранных m правил для каждого класса, во множество надёжных тренировочных примеров нашей МЭ модели и продолжаем итерирование.

Вся самозагрузочная модель прерывается при выполнении трёх условий: (а) все неразмеченные данные для тренировки были использованы; (б) когда метрика F_1, базирующаяся на надёжных примеров для всех натренированных типов собственных имён не имеет дальнейшего улучшения в течение трёх последовательных циклов; и (в) она превышает фиксированный лимит итераций.

2.2 МЭ Валидатор

Модель Максимальной Энтропии (МЭ) используется для валидации извлечённых СЭ и новых тренировочных примеров в каждом цикле. Существует две причины для выбора МЭ модели по сравнению с известными статистическими моделями. Во-первых, они обеспечивают систематический способ тренировки статистической модели, который непротиворечив со всем имеющимся набором знаний. [16] сообщили, что совместное использование биграммных, фразовых и нетекстовых свойств не изменяет результаты работы МЭ значительно и проиллюстрировали что она может надёжно обрабатывать пересекающиеся свойства. Во-вторых, они не нуждаются в моделях с обратной обработкой (back-off) для компенсирования ненадёжной статистики, таких как Скрытые Модели Маркова (СММ). СММ также требует $O(|S|^2)$ переходных параметров (transition parameters) для оценивания [14], превращая разреженность данных в серьёзную проблему из-за того что количество состояний увеличивается.

Основная идея МЭ [19] базируется на понятии энтропии информации: распределение вероятности P для примеров со свойствами $\{x_1 \ldots x_m\}$, принадлежащими классу C_i, кодирует состояние информации с мерой максимальной неопределённости (значением энтропии) и находится в соответствии с существующей моделью информации $\lambda_1 \ldots \lambda_m$. Этот принцип позволяет нам оценить вероятность разметки экземпляра I, принадлежащего классу C_i, согласно формуле:

$$P(C_i \mid I) = \frac{1}{Z(I)}\exp[\lambda_1 f_1(I,C_i) + \ldots + \lambda_m f_m(I,C_i)] \quad (2)$$

где $f_k(I,C_i)$ является функцией с бинарным значением, которая отображает в '1' если размеченный пример I содержит свойство x_k, которое существует для класса C_i.

Информационная модель $\lambda_1 \ldots \lambda_m$ аппроксимируется нашими надёжными тренировочными примерами с использованием Обобщённого Итеративного Шкалирования (Generalized Iterative Scaling), описанного в [8]. $Z(I)$ является нормализующей константой, которая

получена в результате суммирования всех $P(C_i|I)$ для всех классов.

Валидированные примеры содержат размеченное СИ, которое мы используем для оценки сходства на более поздней стадии (см. 2.3.2). Набор свойств, которые мы используем для моделирования всех примеров в рамках МЭ являются неизменными, как и те использованные при генерации правил и приведённые в Таблице 1.

2.3 Генерация правил

Мы используем в целом 7 базовых лингвистических свойств для генерации правил. Они обобщены в Таблице 1 снизу.

Свойства	***Примеры***
Часть речи (POS)	NN, NP, PP
Лексическое представление	Car, Building
Именная группа	The President of US
Глагольная группа	Happily jogging
Главное существительное	[President]
Главный глагол	[jog]
Цифры	1,2,3

Таблица 1. Базовые лингвистические свойства

Дополнительно, наши сгенерированные правила состоят из упрощённых регулярных выражений, определённых пользователем типов шаблонов и булевских выражений, полученных из эвристических правил. Мы также используем списки известных объектов для поддержки генерации правил. Примером такого сгенерированного правила является:

ЧЕЛОВЕК -> список(personTitle) .? Capwords{1,3}

которое извлекает от 1 до 3 слов с заглавной буквы, предшествующими извлекаемому слову, такому как 'Mr.' из списка *personTitle*, как ЧЕЛОВЕК.

Использование эвристических шаблонов в наших правилах значительно улучшает качество полученного набора правил. По сравнению с другими подходами, которые не используют эвристических правил, например ([9], [3], [5]), это позволяет нашей системе иметь хороший старт с помощью использования шаблонов, которые люди используют для восприятия настоящих данных.

Процесс генерации правил стартует с измерения статистики свойств в тренировочных примерах. Свойства измеряются в соответствии с вероятностью, что они возникнут для данного СИ в определённом месте. Пусть f_i будет свойством i, f_i^k be свойством i, возникающем в контексте позиции k, $СИ_x$ будет произвольно выбранным Собственным Именем. Оценка свойства f_i для выбранного СИ определяется далее, где α является весом для вероятностей:

$$S(f_i) = \alpha \times P(f_i \mid СИ_x) + (1-\alpha) \times P(f_i^k \mid СИ_x) \quad (3)$$

При оценке свойства, мы имеем в виду появление свойства и контекст его местоположеения для выбранного СИ. Вероятность $P(f_i \mid СИ_x)$ может быть выражена как:

$$P(f_i \mid СИ_x) = \frac{\#\, f_i \text{ в } СИ_x}{\sum_{y=n}^{m} \#\, f_i \text{ в } СИ_y} \quad (4)$$

Похожим образом, $P(f_i^k \mid NE_x)$ может быть выражена как:

$$P(f_i^k \mid СИ_x) = \frac{\#\, f_i^k \text{ в } СИ_x}{\sum_{y=n}^{m} \#\, f_i^k \text{ в } СИ_y} \quad (5)$$

Для каждого размеченного тренировочного примера, и имеем набор контекстуальных свойств в форме:

$$<F_{-m}>\ldots<F_{-2}><F_{-1}><F_0><F_{+1}><F_{+2}>\ldots<F_{+m}>$$

где $<F_i>$ представляют набор свойств для слота i, и m является числом контекстуальных токенов имеющихся в виду. В особенности, $<F_0>$ является набором свойств для размеченного слота. Используя Равенство (3), мы измеряем статистику для каждого свойства среди наборов свойств в разных примерах и для разных типов СИ. С помощью этого метода, мы объединяем общую информацию во всех тренировочных примерах и выбираем наиболее обобщающие свойства для стартования генерации правил. Каждое сгенерированное правило R измеряется в соответствии с нашей модификацией ожидаемой ошибки Лапласа [22]:

$$L_k(R) = (n+1)/(n+k*p+1) \quad (6)$$

где n (или p) является количеством отрицательных (положительных) примеров, покрытых правилом, и $0 < k < 1$. k используется для повышения веса отрицательным примерам по сравнению с обычным Лапласианом. В наших экспериментах мы установили k = 0.5.

Для оптимального качества, нам нужно перебрать все возможные комбинации свойств и результирующую меру Лапласа их правил. Однако, это является вычислительно сложной задачей и временная сложность растёт экспоненциально при увеличении количества свойств. В качестве компромисса, мы лишь оцениваем меру Лапласа правил, сгенерированных с использованием наилучших n свойств (свойств с наибольшей статистической встречаемостью среди тренировочного множества). Окончательный результат представляется в виде правил с наибольшим

```
Цикл для каждого примера в стартовых-примерах
 Пока не покрыт(пример)
  Цикл для каждого правило в примере
   измерить(правило)
   если допустимое-правило(правило)
    тогда добавить(правило, множество-правил)
          пометить_добавление(правило,стартовые-примеры)
    иначе цикл по каждому непротестированному-свойству
          если протестировать-добавить(свойство,
                правило,множество-правил)
          тогда пометить_покрытие(правило,стартовые-примеры)
```

Рисунок 2. Базовый алгоритм генерации правил

покрытием, но маленькой мерой Лапласа. Базовый алгоритм для генерации правил похож на представленные в [5] и показан на Рисунке 2.

Проблема генерации правил возникает, когда существуют противоречия между классами. Например, для фрагмента "*Recently Ford said that…*", даже для людей трудно различить "*Ford*" как тип ОРГАНИЗАЦИЯ или ЧЕЛОВЕК. Последовательно, слово "*said*" не может быть использовано только лишь как хороший признак для различения этих двух типов СИ. Тем не менее, это свойство является хорошим индикатором, что предыдущее слово является кандидатом СИ.

2.4 Уточнение правил

Некоторые из сгенерированных правил являются слишком общими для использования в извлечении СИ. Источником таких ошибок является общность использования эвристических шаблонов. Примером такого сгенерированного правила является:

PERSON -> nonCapital Capwords{1,3} (R1)

которое извлекает от 1 до 3 слов с заглавной буквой после незаглавной буквы как класс ЧЕЛОВЕК.

Для данного предложения *"According to Mr Clinton"*, правило будет неправильно извлекать "Mr Clinton" в качестве ЧЕЛОВЕК, что неверно, так как 'Mr' не является частью СИ.Тем не менее, для другого предложения вроде *"According to Bill Clinton"*, правило работает корректно. Основная причина для генерации таких правил состоит в недостатке отрицательных примеров во время тренировочного процесса.

Для решения этой проблемы, правила с низкой точностью применяются для аккумулирования надёжных тренировочных примеров из предыдущих инераций для извлечения тренировочных примеров, используемых для уточнения. Извлечённые примеры, которые правильно промаркированы правилом, размечиваются как положительные. В то же самое время, любой пример, промаркированный правилом неправильно размечивается как отрицательный. Это может быть сделано, так как мы имеем набор аккумулированных надёжных примеров, которые имеют предопределённые классы.

После получения требуемых примеров, мы выболняем прямое и обратное обучение для уточнения правил. Необходимым ограничением для обратной тренировки является то, что мы не можем выбирать свойства из первоначальных правил. Это сделано, чтобы быть уверенным, что правила не начнут противоречить друг другу. Прямое обучение старается решить эту проблему добавлением дополнительных ограничений, тогда как обратное обучение добавляет новые исключения к правилу.

2.4.1 Прямое обучение. Целью прямого обучения является добавление ограничений без уменьшения полноты. Правила на этой стадии тренируются с использованием положительных и отрицательных примеров, извлечённых с помощью описанной выше процедуры. По сравнению с первоначальным обучением, мы имеем больше отрицательных примеров, которые возникают в отличных от СИ позициях. Таким образом, значение L_k (см. Равенство 6) моделируется точно на этой стадии. Новые свойства выбираются, базируясь на Равенстве (3) и добавляются к существующему правилу в коньюнкции с существующими свойствами. Используя один из предыдущих примеров (R1), частично уточнённое правило после процедуры прямого обучения превращается в:

PERSON -> nonCapital & PoS(предлог) Capwords{1,3} (R2)

Вместо использования любого слова с незаглавными (*nonCapital*) буквами для сопоставления с правилом, оно теперь ограничено словом с незаглавными буквами такого как "*to*". Количество добавляющихся свойств зависит от улучшения точности при поддержке текущего значения полноты. Это сделано с использованием

существующего базового алгоритма генерации правил на Рисунке 2. Все измерения точности и полноты базируются на аккумулированном наборе надёжных примеров.

2.4.2 Обратное обучение. Для обратного обучения, мы в первую очередь используем модификацию Равенства (3) для подсчёта наиболее перспективного свойства в отрицательных примерах из набора положительных примеров:

$$S(f_i) = \alpha \times P(f_i \mid not\ CИ_x) + (1-\alpha) \times P(f_i^k \mid not\ CИ_x) \quad (9)$$

Последовательно, мы используем эти перспективные свойства для улучшения правил с помощью добавления к ним отрицания или исключения первоначальных свойств. Эта процедура сделана через обращения первоначального обучения с перестановкой использования отрицательных примеров с положительными примерами. Свойства в правилах выбираются так, что они могут уменьшить значение ошибок L_k в Равенстве (6). Следуя тому же примеру (R2), мы находим что список *personTitle* обычно не является частью правила. Следовательно, превило модифицируется с *list(personTitle)* в качестве отрицательного свойства:

PERSON -> nonCapital & PoS(Предлог)
[Capwords{1,3} & !список(personTitle)] (R3)

С прямым и обратным обучением, мы в состоянии обрабатывать исключительные случаи когда правила с высокой полнотой не могут обрабатываться хорошо. При корректном применении, мы обнаружили что стадия уточнения правил может увеличить точность нашей системы.

3. Эксперименты

Мы самозагружаем процесс генерации правил на основе набора данных MUC-7 и статей из Penn Treebank Wall Street Journal (WSJ) [13]. Наши эксперимены нацелены на оценку стадий качества нашей системы: (а) количество первоначальных данных; и (б) уточнение правил.

Базируясь на эмпирических исследованиях, количество итераций зафиксировано на уровне 250. Мы заметили что качество начинает ухудшаться, когда количество итераций превышает это количество. Мы предполагаем, что это происходит из-за распространения ошибок в наборе правил. Некоторые неточные правила могут постоянно добавлять неправильные примеры в систему. Следовательно, качество множества примеров постепенно ухудшается после каждого цикла.

3.2 Отбор начальных данных

Важным моментом является выбор первоначальных эвристических данных. Мы используем два способа для отбора первоначальных данных из множества эвристических правил – базируясь на метриках F_1 или случайно. Результаты, показанные в Таблице 1, показывают что выбор на основе метрик F_1 приводит к лучшим результатам, чем просто случайный выбор.

Анализ результатов показывает, что есть две основные причины для этого. Во-первых, первоначальные правила могут извлечь пересекающиеся примеры. Во-вторых, они могут стать неоднозначными или конфликтовать с начальными данными из других классов. Поэтому система извлекает похожие тренировочные примеры между классами и приводит к плохо сгенерированным правилам в случае использования случайного выбора.

	Случайный выбор			***Выбор по F_1***		
Класс Собственных Имён	***P***	***R***	***F_1***	***P***	***R***	***F_1***
МЕСТОПОЛОЖЕНИЕ	0.70	0.63	0.66	0.76	0.96	0.84
ОРГАНИЗАЦИЯ	0.41	0.64	0.50	0.64	0.54	0.59
ЧЕЛОВЕК	0.81	0.79	0.80	0.81	0.89	0.85

Таблица 1. Результаты выбора первоначальных данных

	Без уточнения			*С уточнением*		
Класс Собственных Имён	***P***	***R***	***F_1***	***P***	***R***	***F_1***
МЕСТОПОЛОЖЕНИЕ	0.62	0.83	0.71	0.68	0.74	0.71
ОРГАНИЗАЦИЯ	0.47	0.62	0.53	0.53	0.58	0.55
ЧЕЛОВЕК	0.68	0.86	0.76	0.72	0.75	0.73

Таблица 2. Результаты уточнения правил

3.3 Уточнение правил

Здесь мы исследуем необходимость в уточнении правил. Правила применяются с и без уточняющей процедуры для проверки их точности и полноты. Результаты в Таблице 2 показывают, что уточнение может действительно помочь повысить точность за счёт полноты. Общая метрика F_1 улучшилась после уточнения.

4. Похожие работы

Несколько похожих исследований для самозагрузочного Извлечения Информации (ИИ) покрывают случай использования ключевых слов или извлекаемых шаблонов. [21] предложила использование эвристических понятийных шаблонов для создания словарей для ИИ используя заранее классифицированный тренировочный и неразмеченный корпуса текстов. Авторы [26] пытались найти шаблоны для извлечения без заранее размеченного набора текстов, стартуя с набора первоначальных шаблонов. Одно из основных различий между такими системами и нашей системой состоит в использовании эвристического знания. Вместо использования эвристических правил для выполнения основной задачи извлечения, мы используем эвристические правила в качестве начальных данных для самозагрузки и также как источник внешнего знания. Использование понятий в виде начальных данных в [17] является более близким к нашей работе. Однако, они не используют эвристические правила для поддержки тренировочного процесса. Другие исследования по самозагрузочным системам использовали неразмеченные данные с МЭ моделью [16], в которой несколько размеченных документов, вместо ключевых слов, используются для начала самозагрузочного процесса.

5. Заключение

В этой статье мы презентовали новую систему для получения правил для задачи извлечения собственных имён из текстов на естественном языке. Наш подход позволяет нам стартовать лишь с небольшого количества тренировочных примеров и эвристических правил, в то же время получая сравнимые результаты с существующими системами генерации правил. В отличие от других существующих систем, использующих лишь статистический подход, наша модель по генерации правил позволяет нашей системе автоматически корректировать ошибочные правила самостоятельно. Это сделано с помощью процедуры прямого и обратного уточнения на каждой итерации. Мы также использовали обучение общих правил и уменьшили разницу между сгенерированными и созданными вручную правилами, что не сделано в существующих методах.

6. Библиография

[1] Blum, A. and Mitchell, T. 1998. Combining labeled and unlabeled data with co-training. In *Proc of COLT'98*.

[2] Blum, A. and Langley, P. 1997. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1-2): 245–271.

[3] Borthwick, A. 1999. *A Maximum Entropy Approach to Named Entity Recognition*. PhD thesis, New York University.

[4] Califf, M.E. and Mooney, R. J. 1999. Relational Learning of Pattern-Match Rules for Information Extraction. In *Proc. of 16th NCAI*, 328-334.

[5] Ciravegna, F. 2001. An Adaptive Algorithm for Information Extraction from Web-related Texts. In *Proc. of IJCAI-2001 Workshop*.

[6] Cohen, W. W. and Sarawagi S. 2004. Exploiting Dictionaries in Named Entity Extraction: Combining Semi-Markov Extraction Processes and Data Integration Methods. In *Proc. of KDD 2004*, 89-98.

[7] Cohn, D.; Atlas, L.; and Ladner R. 1994. Improving generalization with active learning. *Machine Learning* 15.

[8] Darroch, J. N. and Ratcliff, D. 1972. Generalized Iterative Scaling for Log-Linear Models. *Annals of Mathematical Statistics*, 43(5): 1470-1480

[9] Freitag, D. and McCallum, A. 1999. Information Extraction with HMMs and Shrinkage. In Proc of *Workshop on Machine Learning for Information Extraction*, AAAI.

[10] Freitag, D. 1998. Multistrategy Learning for Information Extraction. In *Proc. of the 15th ICML*.

[11] Jones, R.; McCallum, A.; Nigam, K.; and Riloff, E. 1999. Bootstrapping for Text Learning Tasks. In *IJCAI-99 Workshop*, 52-63.

[12] Lin, D. K. 1997. Using Syntactic Dependency as Local Context to Resolve Word Sense Ambiguity. In *Proc. of ACL-97*.

[13] Marcus, M; Marcinkiewicz; and Santorini B. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2): 313-330.

[14] McCallum, A..; Freitag, D.; and Pereira, F. 2000. Maximum entropy Markov models for information extraction and segmentation. In *Proc. ICML 2000*, 591–598.

[15] Ng, V. and Cardie, C. 2003. Bootstrapping Coreference Classifiers with Multiple Machine Learning Algorithms. In *Proc. of EMNLP-2003*, ACL.

[16] Nigam, K.; Lafferty, J.; and McCallum, A. 1999. Using maximum entropy for text classification. In *Proc. of the IJCAI-99 Workshop*, 61–67.

[17] Niu, C.; Li, W.; Ding J.; and Srihari, R. K. 2003. A bootstrapping approach to named entity classification using successive learners. In *Proc of 40th Annual Meeting of ACL*.

[18] Pierce, D. and Cardie C. 2001. Limitations of Co-Training for Natural Language Learning from Large Datasets. In *Proc. of EMNLP-2001*, ACL.

[19] Pietra, S. D.; Pietra V. D.; and Laferty J. 1997. Inducing Features of Random Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4): 380-393.

[20] Peshkin L. and Pfeffer A. 2003. Bayesian Information Extraction Network. In *Proc. of 18th IJCAI*.

[21] Riloff E. 1996. Automatically Generating Extraction Patterns from Untagged Text. In *Proc. of 13th AAAI*, 1044-1049.

[22] Soderland S. 1999. Learning Information Extraction Rules for Semi-Structured and Free Text, *Machine Learning*, 34(1-44): 233-272.

[23] Thompson C.A.; Califf M.E.; Mooney J.R. 1999. Active Learning for Natural Language Parsing and Information Extraction. In *Proc. 16th IMLC*, 406-414

[24] Voorhees, E. M. 1993. Using WordNet to disambiguate word sense for text retrieval. In *Proc. ACM SIGIR*,171-180

[25] Yang, H., H. Cui, M.Y. Kan, M. Maslennikov, L. Qiu, and T.S. Chua. 2003. QUALIFIER in TREC-12 QA main task. In *Notebook of the Twelfth Text Retrieval Conference (TREC-2003)*. Maryland, USA, 480-488.

[26] Yangarber, R., Lin, W., and Grishman, R. 2002. Unsupervised learning of generalized names. In *Proceedings of the 19th international conference on Computational Linguistics (COLING 2002)*. Morristown, NJ, USA, 1-7.

[27] Yarowsky, D. 1995. Unsupervised word-sense disambiguation rivaling supervised methods. In *Proc of the Thirty Third Annual Meeting on Association for Computational Linguistics (ACL-1995)*. Cambridge, MA, 189-196.

[28] Zhou G.D and Su J. 2002. Named Entity recognition using an HMM-based chunk tagger. In *Proc. of 40th Annual Meeting of ACL*, 473-480.

Сравнение качества работы синтаксического анализатора системы ЭТАП и статистического синтаксического анализатора MaltParser на материале текстов из корпуса СинТагРус

В.Г. Сизов, Д.Р. Валеев, Л.Г. Крейдлин
Институт проблем передачи информации РАН,
sizov@iitp.ru, denis.valeev@gmail.com, lenya@iitp.ru

Аннотация

Статья представляет способ и результаты сравнения качества работы двух синтаксических анализаторов (парсеров) – основанного на правилах парсера, входящего в состав лингвистического процессора ЭТАП-3, и статистического парсера MaltParser. Оценка качества для каждой из рассматриваемых систем производилась путем сравнения по различным критериям построенных ими синтаксических структур с эталонными синтаксическими структурами. Ввиду того, что MaltParser использует входные данные с уже снятой частеречной и, в большой степени, лексической омонимией, аналогичное разрешение омонимии реализовано и для ЭТАП-3 – в специальном экспериментальном режиме. Источником эталонных структур и входных фраз, а также данных для обучения статистического парсера и снятия омонимии является глубоко аннотированный корпус русских текстов СинТагРус. Результаты сравнения двух парсеров показали, что после устранения частеречной и лексической омонимии парсер системы ЭТАП-3 существенно превосходит MaltParser практически по всем критериям, в частности, по критерию правильности установления синтаксического хозяина и правильности определения типов синтаксических отношений.

1. Вводные замечания

Синтаксический анализатор, разработанный группой ученых Института проблем передачи информации им. А.А. Харкевича (ИППИ РАН) для многоцелевого лингвистического процессора ЭТАП-3 (см. о нем, в частности, [5]), в значительной мере опирается на лингвистическую теорию «Смысл ⇔ Текст» И.А. Мельчука (см. [4]) и в первую очередь на синтаксический компонент этой теории.

Система MaltParser – это языково-независимый статистический парсер, созданный членами группы MALT (модели и алгоритмы для языковых технологий) факультета математики и системного проектирования университета Вэкшё (Швеция). MaltParser формирует статистическую модель синтаксического анализа на основе данных из корпуса размеченных текстов. Ниже описываются принципы работы MaltParser и исследуются результаты работы этого парсера, обученного на материале из корпуса СинТагРус. Подробнее о MaltParser можно прочитать в [7].

2. Основные компоненты парсера ЭТАП-3

2.1. Морфологический анализатор и предсинтаксический модуль

Парсер ЭТАП-3 начинает работу после того, как **морфологический анализатор**, который обрабатывает каждое отдельное слово предложения, закончит обработку всех слов предложения и выдаст его морфологическую структуру (МорфС). МорфС – это упорядоченная последовательность всех слов предложения, в которой каждый элемент содержит имя лексемы (лемму), атрибут части речи и набор словоизменительных морфологических характеристик. Если словоформа лексически и/или морфологически неоднозначна, она представлена в МорфС набором объектов, именуемых **омонимами**, каждый из которых состоит из леммы, атрибута части речи и набора морфологических характеристик.

Система ЭТАП-3 не располагает отдельным тэггером, который бы определял части речи слов, входящих в предложение. В какой-то степени его функции выполняет небольшой предсинтаксический модуль, частично разрешающий лексическую и морфологическую

неоднозначность, учитывая близкий линейный контекст. В среднем предсинтаксический модуль осуществляет чистку примерно 20% омонимов предложения.

2.2. Парсер

Синтаксический анализатор ЭТАП-3 получает на входе МорфС обрабатываемого предложения и строит для нее **дерево зависимостей**, используя набор синтаксических правил (синтагм). Каждая синтагма соответствует некоторой минимальной (бинарной) синтаксической конструкции и устанавливает одну гипотетическую именованную и направленную синтаксическую связь между двумя словоформами. В настоящее время в системе ЭТАП-3 для русского языка используется 78 различных типов синтаксических отношений (СинтО), именами которых и помечаются синтаксические связи в дереве зависимостей. Алгоритм синтаксического анализа использует синтагмы для того, чтобы построить все возможные гипотетические связи между словами предложения, а затем применяет к полученному набору связей систему различных фильтров, которые удаляют лишние связи таким образом, чтобы оставшиеся связи вместе с узлами формировали дерево зависимостей (размеченный ориентированный связный граф без циклов). Если с помощью фильтров не удается удалить все избыточные связи, алгоритм последовательно просматривает все узлы, в которые входит несколько синтаксических связей, определяет наиболее вероятную из них и удаляет все остальные. После каждой такой операции по удалению к набору связей вновь применяется система фильтров, до тех пор, пока не будет сформировано дерево. Более детально алгоритм описан в [3].

В процессе анализа парсер обращается к так называемому комбинаторному словарю русского языка, который в настоящее время насчитывает около 100 тысяч статей, содержащих обширную лингвистическую информацию о слове (синтаксические признаки, семантические признаки (дескрипторы), модель управления и пр.).

В целом работа парсера ЭТАП-3 может считаться робастной: в худшем случае, если для обрабатываемого предложения не удается получить адекватного дерева зависимостей, «безотказность» парсера достигается за счет того, что некоторые из слов предложения соединяются в аварийном порядке **фиктивным** СинтО. Словам, которые не были найдены в словаре, если их не удается надежно отнести к какой-либо конкретной части речи, приписывается атрибут части речи NID (неопознанное слово).

Как правило, в результирующем дереве зависимостей каждый узел соотносится со словом анализируемого предложения. Исключение составляют те случаи, когда слово образовано с помощью морфологического процесса словосложения и не имеет собственной словарной статьи в комбинаторном словаре (*восьмитомный, стодвадцатилетний* и т.д.). Для такого слова парсер генерирует два или несколько узлов в дереве зависимостей.

3. Основные компоненты парсера MaltParser

Статистический парсер MaltParser был выбран для сравнения и качественной оценки работы синтаксического анализатора ЭТАП-3, поскольку он является одним из немногих известных нам парсеров, которые строят структуры синтаксических зависимостей и работают с русским языком. Данная система не использует грамматических правил, а полностью опирается на метод индуктивного обучения: правила проведения синтаксических связей выводятся путем анализа уже имеющихся синтаксических структур (data-driven approach). Это дает возможность настраивать парсер на обработку текстов на любом языке, если для этого языка имеется синтаксически аннотированный корпус текстов. Сравнение MaltParser с другими анализаторами подобного типа на материале немецкого языка приводится в [8].

В структуре MaltParser можно выделить три основных компонента:

1. Собственно парсер, осуществляющий построение структур зависимостей на основе детерминистского алгоритма синтаксического анализа.
2. Основанную на истории статистическую модель (history-based model), выбирающую оптимальные действия при текущем состоянии парсера.
3. Обучающий (machine-learning) компонент, формирующий базу для статистической модели.

Для построения структуры фразы парсер использует алгоритм, энергично реагирующий на связи (arc-eager algorithm) в совокупности с восходящей обработкой данных для построения связей с левым хозяином и нисходящей

обработкой для построения связей с правым хозяином. Данный алгоритм позволяет строить дерево зависимостей за один проход, благодаря чему зависимость времени, затраченного на построение структуры фразы, от количества слов в этой фразе носит линейный характер.

На входе у MaltParser – последовательность словоформ, составляющих фразу. Для каждой словоформы должны быть указаны соответствующие ей имя лексемы и набор морфологических характеристик. Если словоформы могут интерпретироваться несколькими способами, неоднозначность должна быть снята. Поэтому для обработки текстов на естественном языке парсер должен применяться в сочетании с частеречным тэггером.

В процессе обработки списка входных узлов парсер формирует стек частично обработанных узлов. При проходе по списку входных узлов алгоритм пытается связать очередной входной узел с узлами из стека, либо как слугу, либо как хозяина. Если в какой-либо узел можно провести несколько связей с разными именами и/или хозяевами, то алгоритм выбирает наиболее вероятную из них. Входные узлы, у которых после окончания процедуры проведения связей не появилось синтаксического хозяина, заносятся в стек частично обработанных узлов. Узлы с установленными хозяевами далее не рассматриваются. Возможность проведения связей и выбор наиболее вероятной из нескольких возможных определяет механизм, называемый оракулом (oracle). Оракул, реализованный в MaltParser, для оптимального выбора использует основанную на истории статистическую модель.

4. Корпус «СинТагРус»

Первый аннотированный корпус русских текстов СинТагРус [1, 6] насчитывает свыше 35 тысяч фраз (около 500 тысяч слов). Корпус представляет собой массив документов в формате XML. Каждый документ содержит краткое описание (название, автор, дата создания и т.д.) и список элементов – предложений. Предложение, в свою очередь содержит список элементов – слов и знаков препинания. Как предложения, так и слова имеют обязательный атрибут – порядковый номер. У слов есть также другие обязательные атрибуты – лемма (имя лексемы), список морфологических характеристик, порядковый номер слова-хозяина (или признак вершины) и имя синтаксической связи, по которой слово зависит от хозяина.

Корпус строится в два этапа: сначала каждое предложение обрабатывается парсером ЭТАП-3, а затем полученные синтаксические структуры вручную редактируются лингвистами-экспертами. На стадии ручной правки в дерево зависимостей вносятся улучшения, которые не могли быть получены автоматически. В частности, могут вводиться дополнительные узлы (т. н. **фантомы**), делающие синтаксическую структуру эллиптических конструкций более прозрачной. Так, в предложении *Я приехал из Москвы, а он из Мадрида* возникнет еще один глагол *приехал* между словами *он* и *из*, и тем самым результирующее дерево зависимостей примет более изящный вид. Другой вид улучшений – указание лексических и грамматических характеристик для слова, отсутствующего в словаре системы ЭТАП-3. Если отсутствующее слово было опознано словообразовательной морфологией, производится также замена цепочки словообразовательного разбора на исходное слово (*гео+политика -> геополитика*)

В данном исследовании СинТагРус используется в качестве эталона (“золотого стандарта”) при оценке работы синтаксических анализаторов и как источник данных для обучения статистического парсера.

5. Оценка качества работы парсеров

5.1. Основные принципы оценки

В данном исследовании качественная оценка работы парсеров производилась путем сравнения эталонных синтаксических структур для фраз, взятых из корпуса СинТагРус, со структурами, полученными для тех же фраз каждой из рассматриваемых систем. При этом оценивалась степень расхождения между структурами по критериям, подробно описанным в [2]. В настоящей работе применялись только те из них, которые оказались релевантными и для MaltParser. Далее приводится краткое описание использованных нами типов оценки.

5.1.1. Синтаксические оценки слов

1. **Оценка правильности установления хозяев**: доля слов, для которых синтаксический хозяин (или его отсутствие) был установлен корректно
2. **Оценка правильности установления связей**: доля слов, для которых наличие корректной синтаксической связи или отсутствие хозяина было установлено верно,

однако связь может исходить из ошибочного хозяина.
3. **Оценка правильности установления хозяев и связей**: доля слов, для которых верно установлены как хозяин, так и синтаксическая связь.
4. **Статистика по разным типам связей**: для каждого типа связи вычисляются **точность, полнота** и **F-мера.**

5.1.2. Синтаксические оценки предложений

1. **Оценка вершины**: доля предложений, для которых вершина была установлена правильно.
2. **Оценка правильности скелета**: доля предложений, для которых все связи были установлены верно, не учитывая их тип.
3. **Оценка строгой правильности структуры**: доля предложений, для который все связи были установлены верно.

5.1.3. Нормализованная оценка синтаксических расхождений, основанная на штрафах

Данная оценка основывается на разного рода возможных отклонениях разбора (Р) от эталонной структуры (ЭС). Ниже перечисляются эти типы расхождений:

1. В узел структуры в Р идет фиктивная связь (в эталоне фиктивных связей, разумеется, быть не может).
2. Узел является вершиной в ЭС и не является вершиной в Р, или наоборот.
3. В ЭС некоторый узел X подчиняется узлу Z по отношению R, отсутствующему в списке отношений парсера. Такое может произойти, поскольку корпус СинТагРус содержит определенное количество специфических конструкций, не предусмотренных даже весьма полной грамматикой ЭТАПа-3.
4. В ЭС узел X подчиняется узлу Z по отношению R, а в Р он подчиняется узлу Z, но не по отношению R.
5. В ЭС узел X подчиняется узлу Z по отношению R, а в Р он подчиняется по отношению R, но не узлу Z.
6. В ЭС узел X подчиняется узлу Z по отношению R, а в Р он не подчиняется ни узлу Z, ни по отношению R.

Для каждого типа расхождений предусмотрен свой уровень штрафных баллов, на основании которых вычисляется нормализованная оценка для отдельных узлов, предложений и корпусов. Вычисления производятся следующим образом: сначала для каждого узла вычисляется суммарный штраф за все обнаруженные расхождения и делится на максимальный штраф, который этот узел может получить. Затем все полученные таким образом нормализованные штрафы узлов предложения суммируются, а сумма делится на общее количество узлов предложения; наконец, вычисляется сумма всех нормализованных штрафов предложений корпуса, которая делится на общее число предложений в корпусе.

Синтаксическая модель, на основе которой был размечен корпус СинТагРус, содержит близкие типы синтаксических отношений, которые в некоторых контекстах могут быть тождественными или трудноразличимыми. Например, для описания определительных конструкций используются два СинтО: определительное и описательно-определительное. Назначение за расхождение по именам этих отношений такого же штрафа, как за обычное расхождение по именам отношений, может исказить оценку качества парсера в сторону уменьшения. Поэтому в программу оценки включена опциональная возможность, позволяющая игнорировать некоторые типы синтаксических расхождений.

5.2. Проблемы, возникшие при оценке качества работы парсеров и способы их решения

5.2.1. Лексическая и грамматическая неоднозначность

Как уже отмечалось выше, входные данные сравниваемых систем существенно различаются: парсер ЭТАП-3 принимает на входе текст на естественном (русском) языке, тогда как MaltParser требует предварительной обработки текста частеречным тэггером. Применение такого тэггера могло бы исказить оценку качества работы MaltParser, так как его ошибки неизбежно наложились бы на ошибки самого парсера. Поэтому было решено формировать входные данные для MaltParser на основании информации, содержащейся в СинТагРус. При этом, однако, с точки зрения качественной оценки результатов MaltParser оказывался в более выгодной ситуации по сравнению с анализатором ЭТАП-3, так как вся частеречная и лексическая неоднозначность во входных данных для него оказывалась уже снятой в соответствии с эталоном.

Чтобы избежать занижения оценки парсера ЭТАП-3, был разработан специальный экспериментальный режим, при котором непосредственно перед началом синтаксического анализа устранялась вся частеречная и лексическая неоднозначность входного текста. Разрешение неоднозначности происходило на основании той же информации из корпуса СинТагРус, как и в случае с MaltParser, и потому можно говорить о том, что оценки качества работы двух парсеров являются сопоставимыми.

5.2.2. Внесение изменений в корпус

Потенциальное искажение оценки качества работы парсера ЭТАП-3 может возникнуть из-за того, что в процессе редактирования корпуса лингвистами-экспертами в него могут быть внесены изменения, которые невозможно получить автоматически в ЭТАП-3 (см. п. 4). В то же время во входных данных MaltParser эти изменения будут присутствовать, что приведет к занижению оценки ЭТАП-3 относительно MaltParser.

Для решения этой проблемы была разработана утилита, которая проверяла отсутствие расхождений в токенизации (т.е. в членении цепочки символов каждой фразы входного текста на отдельные единицы – слова) с эталоном и отсутствие неопознанных слов в результатах морфологического анализа. Массив входных данных для сравниваемых парсеров формировался из фраз, удовлетворяющих указанным критериям.

5.2.3. Смещение результатов оценки

В связи с тем, что оцениваемый парсер ЭТАП-3 использовался для первоначальной разметки корпуса СинТагРус, результаты сравнения могут быть искажены в пользу системы ЭТАП. Искажение, или смещение (bias), в значительной степени устраняется в ходе редактирования корпуса лингвистами-экспертами, но, по всей видимости, полностью не исчезает. В отредактированном корпусе можно выделить два типа смещения:

1. ошибочные конструкции, построенные парсером ЭТАП-3, и не исправленные лингвистами.
2. синтаксические конструкции, допускающие несколько равно возможных разборов (это может происходить в случаях реальной неоднозначности предложения (ср. конструкции типа *Весло задело платье*) или неединственности синтаксического описания (например, фраза *Реки не было видно* может анализироваться как трансформация фразы *Река не была видна,* и тогда *реки* – это подлежащее, либо как трансформация фразы *Реку не было видно*, и тогда *реки* – это дополнение

В случае (1) вполне вероятно, что парсер ЭТАП-3, вновь построив ошибочную конструкцию, получит за нее более высокую оценку, чем MaltParser, который провел правильную связь после обучения на корректных примерах. Для устранения смещения первого типа достаточно исправить вызвавшие его ошибки.

Во втором случае ЭТАП-3 скорее всего построит ту же конструкцию, что вошла в корпус, тогда как MaltParser с равной (а точнее – с неизвестной) вероятностью – любую из допустимых. Бороться со смещением такого типа сложнее. Отчасти его можно уменьшить, используя опцию игнорирования некоторых типов синтаксических расхождений, описанную в п. 5.1.3.

Искажения, вносимые смещением, можно оценить, проанализировав фразы, для которых парсер ЭТАП-3 дал лучший результат, чем MaltParser. Показателем смещения может служить получение синтаксическим анализатором ЭТАП-3 высокой оценки за неправильную конструкцию, или более высокой по сравнению с MaltParser оценки, если оба парсера построили одинаково допустимые конструкции. Такой анализ невозможно полностью формализовать, он может быть проведен только лингвистом-экспертом.

5.3. Результаты оценки

В корпусе СинТагРус на момент проведения эксперимента содержалось примерно 32000 фраз. Для оценочного тестирования парсеров было отобрано около 3200 фраз (10%), остальные использовались для обучения MaltParser. Из них требованию отсутствия неопознанных слов и расхождений токенизации соответствовали 1853 фразы. Результаты оценочного тестирования приведены в таблице 1. Из приведенных данных видно, что парсер системы ЭТАП-3 показал в ходе эксперимента более высокие результаты по всем используемым оценочным критериям. Особенно заметна разница в оценке правильности вершин предложения и имен синтаксических связей, тогда как оценки правильности выбора синтаксических хозяев очень близки.

Таблица 1. "Результаты оценочного тестирования ЭТАП-3 и MaltParser"

	ЭТАП	MaltParser
Обработано фраз	1853	
Обработано слов	22461	
Нормализованная ошибка	0,110026	0,142381
Правильная вершина	0,928764 (1721)	0,865623 (1604)
Правильный хозяин слова	0,934464 (20989)	0,932861 (20953)
Правильное имя отношения	0,912649 (20499)	0,886247 (19906)
Правильный хозяин и имя отношения	0,893460 (20068)	0,865010 (19429)
Полное совпадение структуры	0,348084 (645)	0,280626 (520)
Структура не построена	0	0

6. Заключение

Эксперимент по оценке качества работы синтаксических анализаторов с использованием корпуса русских текстов СинТагРус показал, что парсер ЭТАП-3 существенно превосходит MaltParser практически по всем критериям. Однако следует заметить, что для достижения такого качества работы коллективу разработчиков ЭТАП-3 потребовалось много человеческих ресурсов, тогда как для адаптации MaltParser к текстам на новом для него (например, русском) языке необходим лишь синтаксически размеченный корпус. Поэтому исследования, направленные на создание на базе ЭТАП-3 гибридного парсера, включающего, наряду с правилами синтаксического анализа, элементы статистической модели и машинного обучения, выглядит весьма перспективными.

7. Литература

[1] Ю.Д. Апресян, И.М. Богуславский, Б Л. Иомдин, Л Л. Иомдин, А.В. Санников, В.З. Санников, В.Г. Сизов, Л.Л. Цинман. Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы // *Национальный корпус русского языка 2003–2005 г. (результаты и перспективы)*. М: «Индрик», 2005, с. 193–214

[2] И.М. Богуславский, Л Л. Иомдин, Д.Р. Валеев, В.Г. Сизов. Синтаксический анализатор системы ЭТАП и его оценка с помощью глубоко размеченного корпуса русских текстов // *Труды международной конференции «Корпусная лингвистика – 2008»*. СПб: Изд-во Санкт-Петербургского университета, 2008 (в печати)

[3] Л Л. Иомдин, В.Г. Сизов, Л.Л. Цинман, Использование эмпирических весов при синтаксическом анализе. // *Обработка текста и когнитивные технологии*, № 6. Казань, Отечество, 2001, с. 64–72

[4] Мельчук И.А. *Опыт теории лингвистических моделей класса «Смысл ⇔ Текст»*. Москва, Наука, 1974

[5] Jurij Apresian, Igor Boguslavsky, Leonid Iomdin, Alexander Lazursky, Vladimir Sannikov, Victor Sizov, Leonid Tsinman, ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the MTT. // *MTT 2003, First International Conference on Meaning – Text Theory*. École Normale Supérieure, Paris, pp. 279-288.

[6] Juri Apresjan, Igor Boguslavsky, Boris Iomdin, Leonid Iomdin, Andrei Sannikov, Victor Sizov, A Syntactically and Semantically Tagged Corpus of Russian: State of the Art and Prospects // Proceedings *of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*. Genoa, 2006, pp 1378-1381.

[7] Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., and Marsi, E., MaltParser: A language-independent system for data-driven dependency parsing. // *Natural Language Engineering*, 13(2), 2007, pp. 95-135

[8] Smirnova S., A Comparison of Dependency Parsers for German. // *Master's Thesis in Computational Linguistics*. Eberhard-Karls-Universität, Tübingen, Germany, 2006, pp. 14-28, http://www.sfs.uni-tuebingen.de/iscl/Theses/Theses2006/Smirnova.pdf

Лексические функции и возможности оптимизации поиска информации в интернете (на материале параметрических слов)

Тимошенко С.П.
Институт проблем передачи информации им. А.А. Харкевича
nyrestein@gmail.com

Аннотация

На базе лингвистического процессора ЭТАП-3 была разработана опция экспериментального перифразирования, позволяющая определенным образом модифицировать класс поисковых запросов, которые ориентированы на поиск информации о численных значениях параметров объектов. Теоретически замена двух- и трехсловных поисковых запросов полученными структурами должна была увеличить точность поиска. Эксперимент, эмулирующий работу поисковой машины с перефразированием, показал, что показатель точности поиска повышается в среднем на 24%. Кроме того, было установлено, что с помощью получаемых структур, так называемых перифраз, возможно более точно искать некоторые другие типы данных.

1. Вводные замечания

Задача повышения точности поиска в Интернете не всегда хорошо решается чисто математическими методами. Для лингвиста это предсказуемо, поскольку в таких ситуациях мы имеем дело с асимметрией между означающим и означаемым. Применительно к поиску можно сказать, что означаемым является искомый смысл, а означающим – вся совокупность выражающих этот смысл предложений. Предложения языка L, выражающие один и тот же смысл, могут очень сильно отличаться друг от друга. Задача поисковой машины в таком случае – распознать смысл, игнорируя формальные различия. Соблазнительно попробовать решить эту задачу с помощью лингвистических инструментов, ориентированных на описание разнообразных способов выражения смысла – с помощью средств лексической семантики. Одним из таких инструментов является аппарат лексических функций, разработанный И.А. Мельчуком и А.К, Жолковским и значительно дополненный и усовершенствованный Ю.Д. Апресяном. См. [Жолковский, Мельчук 1965; Жолковский, Мельчук 1967; Жолковский, Мельчук 1969; Мельчук 1974; Мельчук, Жолковский 1984; Mel'čuk 1996; Апресян 1974; Апресян, Цинман 2002].

2. Роль лексических функций в перифразировании, ориентированном на задачи поиска

В самом общем виде можно сказать, что лексические функции – это тривиальные смыслы, словесное выражение которых в тексте зависит от того, при каком конкретном слове этот смысл выражается. Для некоторых фрагментов лексической системы языка разработанные лексической семантикой правила вида: «При слове X смысл f_1 выражается словом X´, при слове Y смысл f_1 выражается словом Y´» обладают большой предсказательной силой. Одним из таких фрагментов является класс параметрических слов. Под параметрическими словами мы понимаем имена существительные со значением параметра, допускающего числовое значение, например: *высота, вместимость, объем; продолжительность,возраст; мощность, сила, масса, давление, магнитуда; рождаемость, смертность; цена, стоимость, зарплата, выручка; энтропия, уровень, коэффициент, индекс* и т.д.

Аппарат лексических функций лежит в основе опции перифразирования, реализованной в системе ЭТАП-3. Перифразирование позволяет, опираясь на записанные в словарных статьях слов значения лексических функций, построить некоторое количество предложений, синонимичных или квазисинонимичных заданному, то есть предложений, имеющих с ним одинаковую структуру на глубинном семантическом уровне. Эта опция была приспособлена Л.Л. Цинманом для решения

поисковых задач: из именного словосочетания, в вершине которого находится параметрическое слово, можно получить ряд неполных предложений, содержащих все элементы, кроме численного значения параметра.

Ввод: глубина Марианской впадины

Перифразы, генерируемые системой перифразирования ЭТАП-3:

Глубина Марианской впадины равна[1].
Глубина Марианской впадины составляет.
Глубина Марианской впадины достигает.
Глубина Марианской впадины равняется.
Марианская впадина имеет в глубину.
Марианская впадина достигает в глубину.
Марианская впадина имеет глубину.
Марианская впадина достигает глубины.

Почему в качестве объекта эксперимента был выбран класс параметрических слов? Во-первых, количественные значения представляют собой тип информации, востребованный пользователями всемирной сети. Во-вторых, как мы уже отмечали выше, относительно этого класса правила лексической семантики хорошо исполняются. Параметры представляют собой распространенный тип предикатов. Однако несмотря на предикативную природу этих значений, прототипическими представителями данного класса являются не глаголы, а существительные [Апресян, 2006: 74]. В русском языке очень мало глаголов с соответствующими значениями, например: *стоить*, *весить*, *длиться*, *вмещать*; даже для выражения такого тривиального параметра, как 'высота', в русском языке нет глагола (в отличие, например, от английского: the Pisa tower rises 56 meters). Поэтому естественным способом приписывания какому-либо объекту определенного параметра является конструкция из параметрического существительного и вспомогательного глагола.

Отношения между существительным и глаголом в огрубленом виде предстают как отношения названия ситуации и вспомогательного глагола, служащего для выраженя категориальных значений – вида, времени и т.д. При более внимательном рассмотрении оказывается, что несвободные сочетания вида "глагол-связка +существительное" делятся на группы, объединенные общим элементом смысла, который и есть значение лексической функции.

Три лексические функции, позволяющие описать ситуацию приписывания какому-либо объекту значения параметра - это OPER1, FUNC2, LABOR1-2. Функция OPER1 позволяет обозначает ситуацию так, что название ситуации выступает при соответствующем глаголе первым дополнением, например: *Гора Котопакси имеет* ***высоту*** *почти 6 км.* Функция FUNC2 позволяет обозначить ситуацию так, чтобы название ситуации было подлежащим при функциональном глаголе: ***Высота*** *горы Котопакси составляет почти 6 км.* Функция LABOR1-2 позволяет так обозначить ситуацию, что название ситуации занимает место второго дополнения при функциональном глаголе, а места подлежащего и сказуемого занимают первый и второй актанты соответственно: *Гора Котопакси достигает 5 870 м* ***в высоту****.*

3. Эксперимент, определяющий эффективность поиска по перифразам

На основе исходного списка из 100 параметрических слов было составлено около 120 осмысленных запросов. С помощью системы ЭТАП-3 каждый из них преобразовывался в блок перифраз. В ходе эксперимента перифразы вводились по одной в поисковые системы «Яндекс» и «Google». Поскольку система перифразирования выдает целостные структуры, не выходящие за пределы одного предложения и не предусматривающие разрывов и пропусков, для тестирования была выбрана форма точного запроса. Применение логического оператора «ИЛИ» внутри точного запроса языками поисковых запросов не поддерживается. Употребление дизъюнкции при неточном запросе приводит к тому, что находятся не сайты с целостной структурой, а сайты, просто содержащие заданные слова, возможно, не связанные синтаксически. Поэтому группа слов, выражающая численное значение и зависящая от глагола, может относиться к своершенно другому объекту или к другому параметру.

Употребление точной формы запроса привело к тому, что различия в работе двух упомянутых поисковых систем практически нивелировались, за исключением различий в составе баз документов (например, "Яндекс" индексирует больше личных блогов, чем Google). Эти различия оказались настолько незначительными, что мы сочли возможным ими пренебречь.

Цель эксперимента заключалась в том, чтобы определить, насколько использование системы экспериментального перифразирования повышает эффективноость поиска. Для этого был разработан следующий протокол оценки эффективности. Релевантным признается такой

[1] Не все перифразы в этой опции составляются с опорой на лексические функции. Конструкции с нулевой связкой и кратким прилагательным порождаются особым правилом.

результат, когда искомая численная информация содержится непосредственно в сниппете, предлагаемом поисковой машиной. Показателем эффективности поиска считается количество релевантных результатов на первой странице (т. е. количество релевантных результатов среди первых десяти), выраженное в процентах. Если количество найденных документов меньше 10, то оно и принимается за 100%. В качестве контрольного уровня эффективности поиска был принят уровень эффективности поиска по исходным запросам – именным группам, подаваемым на вход системы перифразирования ЭТАПа-3. Эти запросы также оформлялись как точные – это ограничение дает возможность оценить чистый эффект использования перифраз.

В результате этих операций получаем данные по каждому запросу:

Запрос	1	2	3	
"твердость алмаза"	2044	791	3/10	30%
"твердость алмаза равна"	42	18	10/10	100%
"твердость алмаза составляет"	18	6	3/6	50%
"твердость алмаза достигает"	0	0	0	
"твердость алмаза равняется"	0	0	0	
"алмаз имеет твердость"	79	35	9/10	90%
"алмаз достигает твердости"	0	0	0	

Первая строка в данных по каждому запросу будет представлять эффективность поиска по необработанным словосочетаниям, а все остальные – по обработанным перифразам. Чтобы оценить среднее изменение результата, можно взять среднее значение эффективности необработанных запросов и среднее значение по всем перифразам вместе. Однако такой подход кажется нам неверным. Представим себе, как могла бы работать поисковая машина, умеющая генерировать перифразы. Она получает на входе запрос, создает перифразы, а затем обрабатывает каждую из них как точный запрос. Какие-то перифразы приносят результат, какие-то нет. Очевидно, что на экран в таком случае выводится сумма найденных по всем перифразам документов, а не сработавшие перифразы, конечно, ничего не добавляют, но и не ухуждшают картины. Поэтому мы рассчитывали повышение эффективности поиска для каждого запроса отдельно, и уже для этих результатов затем рассчитывался средний показатель, характеризующий общее изменение эффективности. Так, для приведенного выше запроса средний показатель увеличения эффективности составит

((100%-30%)+(50%-30%)+(90%-30%))/3=50% или 0,5.

Бывают случаи, когда необработанный запрос приносит какое-то количество релевантных ответов, а перифразы не приносят ничего. Например, *"абсолютный минимум температуры на Земле"*. Это связано с тем, что информация по этому вопросу очень часто представлена в Рунете перепечаткой статьи из Большой советской энциклопедии, где использована конструкция с нулевой связкой вместо глагола. В этом и других похожих случаях мы признавали результат использования системы перифразирования отрицательным: если результаты поиска по необработанному запросу были релевантны в 10 случаях из 10, то эффективность поиска с перифразированием равна -100% или -1, если в 6 случаях из 6 — -60% или -0,6 и т.д.

Распределение показателей изменения эффективности поиска при использовании системы перифразирования представлено на гистограмме.

Изменение эффективности поиска при использовании запросов на основе перифраз
аппроксимация:65*0,2*normal(x;0,2395;0,3966)

На графике видно, что в среднем точность поиска повышается на 24%. Любопытно, что если разделить параметрические слова по тематическим группам, то этот средний показатель будет варьироваться. Если смотреть данные по "географическим" запросам (параметры географических объектов, как природных, так и культурных, например, *ширина реки Амазонки в среднем течении* или *численность населения Киева*), то повышение эффективности поиска составит в среднем 18,5%, а в группе данных по физическими запросам (примеры запросов: *молярная масса меди*, *сила тяжести на Марсе*) этот показатель будет равен 27%. Результаты поиска по геометрическим объектам (примеры запросов: *площадь поверхности сферы*, *объем конуса*) оказались неожиданными. Реализующий значение лексической функции FUNC2 глагол *равняться*, а также добавленная в поисковое перифразирование конструкция с кратким

прилагательным *равен,* отличаются от других средств тем, что регулярно присоединяют в качестве первого дополнения формулу или её словесное описание: *Объём конуса равен одной трети произведения основания на высоту; Объем конуса равняется одной трети объема цилиндра с теми же основанием и высотой.* Показатель эффективности использования перифраз для поиска информации такого рода составляет 13%. Однако ЭТАП-3 можно настроить на поиск подобной информации, добавив в систему перифразирования глаголы *характеризовать*, *измеряться* и, возможно, некоторые другие.

4. Неточные запросы

Данные эксперимента показывают, что использование перифраз повышает точность поиска за счет выбора только тех документов, которые содержат искомую информацию, наличие которой однозначно предсказывается глаголом, реализующим ту или иную лексическую функцию. Однако требование точного запроса приводит к тому, что отсеиваются и релевантные документы, в которых мысль выражена немного по-другому: не будет найден документ, содержащий предложение *Гора Котопакси, высота которой составляет почти 6 км…* Неточный поиск по перифразе приводит к тому, что конструкция разрывается и по запросу *продолжительность жизни в Голландии составляет,* находятся страницы новостей, где есть информация о продолжительности жизни в Китае и сообщение о политической ситуации в Голландии. Однако некоторые попытки такого поиска показывают, что и в этом случае точность поиска по перифразам выше точности поиска без них. Это происходит из-за того, что наличие глагола, пусть даже и оторванного от именного словосочетания, меняет общую направленность страницы. Например, по запросу *высота юкки* находятся страницы, содержащие объявления о продаже пальмы какой-либо высоты, а по запросу *высота юкки достигает* находятся статьи из разнообразных справочников по цветоводству, содержащих общую информацию о растении. Можно сказать, что в действие вступает стилистический фактор, так как глаголы, реализующие лексические функции OPER1, FUNC2, LABOR1-2 часто используются в научно-публицистических и научных текстах. Это открывает для системы перифразирования определенные перспективы, тем более что в системе ЭТАП-3 предусматривалось применение стилистических фильтров.

5. Выводы

Эксперимент показал, что точность результатов поиска по запросам, предполагающим численные ответы, может быть увеличена за счет использования системы экспериментального перифразирования. В случае точного запроса количество релевантных результатов увеличивается на 24%. Точный запрос означает, что к искомым документам предъявляются самые жесткие требования. В случае менее жестких требований, то есть при использовании нестрогого запроса, результаты непредсказуемы. Помимо варьирования форм запросов, зависящего от поисковой машины, пути улучшения работы экспериментального перифразирования лежат в расширении числа используемых лексических функций и более точной их настройки на разные типы информации. В частности, эксперимент показал, что изменяя участвующие в перифразировании глаголы, можно получать либо конкретные числовые значения, либо формулы и их словесные описания.

Литература

Апресян 1974 — Апресян Ю. Д., *Лексическая семантика. Синонимические средства языка.*, М., 1974. Второе издание: Апресян Ю. Д. *Избранные труды*. Т. 1. Лексическая семантика. Синонимические средства языка. 2-е издание, исправленное и дополненное. М., 1995.

Апресян 2006 — Апресян Ю. Д., «Основания системной лексикографии», *Языковая картина мира и системная лексикография*, «Языки славянских культур», М., 2006.

Апресян, Цинман 2002 — Апресян Ю. Д., Цинман Л. Л., «Формальная модель перифразирования предложений для систем переработки текстов на естественных языках», *Русский язык в научном освещении*, 2002, №4. с. 102 -146.

Жолковский, Мельчук 1965 — Жолковский А. К., Мельчук И. А. «О возможном методе и инструментах семантического синтеза», *Научно-техническая информация*, 1965, №6.

Жолковский, Мельчук 1967 — Жолковский А. К., Мельчук И. А. «О семантическом синтезе», *Проблемы кибернетики*, Вып.19. 1967. с. 177 – 238.

Жолковский, Мельчук 1969 — Жолковский А. К., Мельчук И.А. «К построению действующей модели языка «Смысл <=> Текст», *Машинный перевод и*

прикладная лингвистика, 1969, вып.11, стр. 5-35.

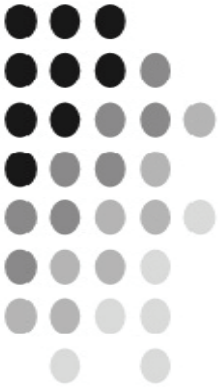
Мельчук 1974 — Мельчук И. А. *Опыт лингвистических моделей «Смысл <=> Текст». Семантика, синтаксис*. М., 1974.

Мельчук, Жолковский 1984 — Мельчук И.А., Жолковский А.К. *Толково-комбинаторный словарь современного русского языка. Опыт семантико-синтаксического описания русской лексики.* Wiener Slawistischer Almanach. Sonderband 14. Wien, 1984.

Mel'čuk 1996 — I. Mel'čuk, «Lexical Functions: A Tool for the Discription of Lexical Relations in a Lexicon». In: *Lexical Functions in Lexicography and Natural Language Processing.* Ed. By L. Wanner. Amsterdam (Philadelphia), 1996. P. 37 – 102.

Раздел 5

Информационные процессы в живых системах



Спектрально-временные характеристики акустических сигналов птиц

Гладышев А.М.
Белорусский государственный университет радиотехники и электроники
gl_peleng@rambler.ru

Аннотация

Приведены временные реализации записей акустических сигналов, издаваемых дятлом при стуке по дереву и звуки голоса дятла и кукушки. Вычислены оценки автокорреляционных функций и спектры для характерных фрагментов звуков, издаваемых дятлом и кукушкой, а также определены времена корреляции для этих фрагментов.

Результаты работы можно использовать при разработке оптимальных алгоритмов обнаружения слабого сигнала на фоне интенсивных помех в системах акустического мониторинга местности, при исследовании и обработке сигналов пения птиц и в природоохранной деятельности.

1. Введение

При проведении акустического мониторинга местности, как правило, прием полезного сигнала осуществляется на фоне сильных помех, источники которых в основном находятся вблизи от приемного устройства. К основным помехам естественного происхождения относятся: шумы ветра (в листве, ветвях деревьев, траве и т.д.), шумы гидродинамического происхождения (шум ручьев, рек, дождя и т.д.), звуки грома и звуки, издаваемые животными.

Акустические характеристики некоторых шумов естественного происхождения, находимые в открытых источниках, например в [1, 2], в основном содержат сведения о спектрах шумов. Временные реализации шумов реже представлены, а корреляционные свойства встречаются очень редко. Особенно это касается звуков, издаваемых животными. Поэтому возникла необходимость в получении данных характеристик, пригодных для разработки оптимальных алгоритмов обработки сигналов.

Среди звуков, издаваемых животными, наиболее существенными с точки зрения вероятности появления в месте приема и с точки зрения их интенсивности являются звуки птиц. В статье подробно рассмотрены спектрально-временные характеристики звуков, издаваемых широко распространенных в лесных ландшафтах птицами – кукушкой и дятлом.

2. Описание записей и методика их анализа

Акустические сигналы птиц записаны в аудио формате MPEG Layer 3 56 кбит/с, с частотой дискретизации 22050 Гц, 16 бит. На записях четко прослушиваются звуки кукушки и дятла, и слабо слышны посторонние звуки.

Анализ сигналов проводился в системе компьютерной математики MATLAB. Вычисление автокорреляционной функции (АКФ) осуществлялось по формуле:

$$B(m) = \sum_{i=1}^{N-|m|} X_i X_{i+m}, \qquad (1)$$

где $B(m)$ – значение АКФ для сдвига $m = -(N-1),\ldots,(N-1)$, N – число отсчетов X в выборке сигнала. Для нормирования АКФ полученная последовательность отсчетов делилась почленно на максимальное значение отсчета, которое получается при $m = 0$.

Амплитудно-частотный спектр (АЧС) сигнала вычислялся с помощью прямого дискретного преобразования Фурье (ДПФ). Для увеличения спектрального разрешения до 1 Гц сначала фрагмент сигнала дополнялся отсчетами, с нулевыми значениями так, чтобы общая длина фрагмента составила бы 1 с, а затем с помощью ДПФ и операции взятия модуля вычислялся АЧС [3, с.301]. Для более четкого выявления характера спектра полученный АЧС был возведен в квадрат. Затем была проведена процедура нормализации квадрата АЧС, аналогичная процедуре нормализации АКФ.

3. Результаты анализа сигналов

Для анализа были выбраны фрагменты записи звуков стука дятла по дереву и звуков голоса кукушки и дятла.

3.1. Анализ звуков кукушки

Полная временная реализация звуков, издаваемых кукушкой, показана на рисунке 1, а фрагмент реализации, на слух воспринимаемый как «Ку-ку» – на рисунке 2.

Звуки, издаваемые кукушкой, представляют собой последовательность пар акустических импульсов, имеющих каждый свою огибающую и частотное заполнение. Пара импульсов подробно показана на рисунке 2. Средний период повторения пар импульсов равен 1,45 с. Длительность первого акустического импульса по уровню 0,5 составляет порядка 76 мс, а второго – порядка 150 мс. Внутреннее заполнение первого и второго акустического импульса, показанное на рисунке 3 представляет собой сигнал близкий к синусоидальному, с периодом повторения около 1,67 мс для первого акустического импульса и 1,96 мс – для второго. Средний временной интервал между началами импульсов в паре составляет 0,4 с.

Оценка автокорреляционной функции проводилась отдельно для каждого акустического импульса в паре и показана на рисунках 4 и 5. Анализ АКФ первого и второго акустического импульса показывает наличие высокой корреляции сигнала в импульсе. Время корреляции, определяемое по уровню 0,5, составляет для первого импульса около 18,3 мс, для второго – 50 мс.

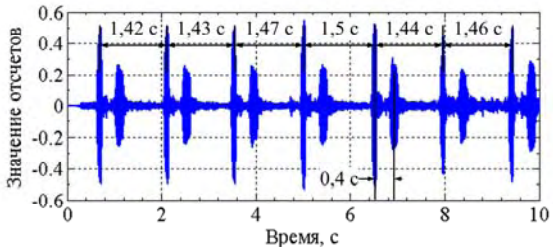


Рисунок 1. Временной график. Кукушка.

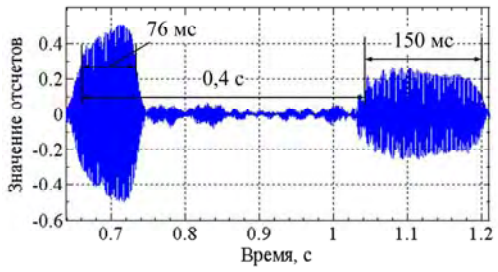


Рисунок 2. Временной график. Фрагмент. Кукушка.

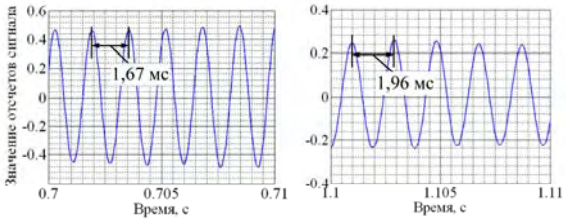


Рисунок 3. Фрагмент временной реализации.

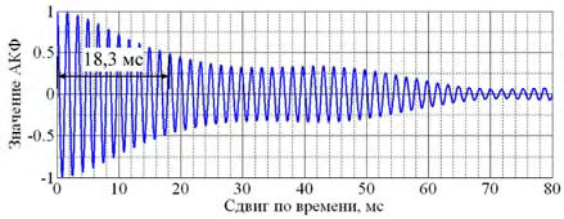


Рисунок 4. АКФ для фрагмента 0,63-0,93 с.

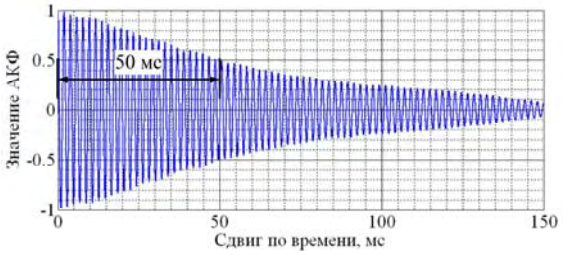


Рисунок 5. АКФ для фрагмента 0,93-1,23 с.

Огибающая АКФ для второго импульса близка к треугольной форме, что можно объяснить тем, что сам импульс имеет форму, близкую к прямоугольной.

Спектры звуков, издаваемых кукушкой, также вычислены отдельно для каждого акустического импульса.

Квадрат нормированного АЧС первого и второго акустических импульсов показан на рисунках 6 и 7. Анализ спектров показывает, что мы имеет дело с узкополосными сигналами, поскольку эффективная ширина спектра много меньше центральной частоты. Эффективная ширина спектра по уровню 0,5 для первого импульса составляет 14 Гц, а для второго – 7 Гц. Спектр первого импульса имеет два максимума – на частоте 584 Гц и на частоте 604 Гц.

Оба спектра, как для первого акустического импульса, так и для второго имеют асимметрию относительно максимальной частоты спектра.

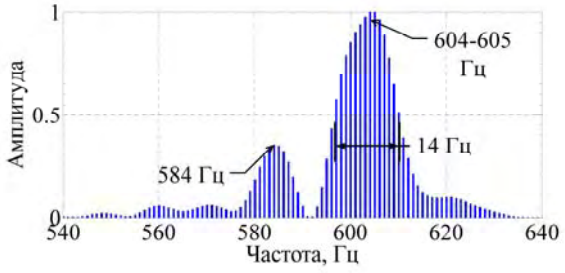


Рисунок 6. Квадрат АЧС фрагмента 0,6-0,8 с.

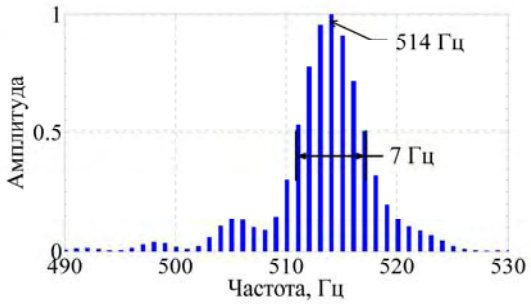


Рисунок 7. Квадрат АЧС фрагмента 1,0-1,2 с.

Асимметрия спектров может указывать на тот факт, что кроме амплитудной модуляции сигнала (огибающая акустического импульса) возможно, присутствует и угловая модуляция внутреннего заполнения.

Таким образом, на основании анализа звуков, издаваемых кукушкой можно сделать следующие выводы: *во-первых*, кукушка издает четко выраженную узкополосную последовательность звуков, состоящую из двух импульсов со средним временным интервалом 0,4 с между импульсами в паре и средним периодом повторения пар импульсов 1,45 с. *Во-вторых*, каждый импульс представляет собой сигнал синусоидальной формы с центральной частотой спектра около 604 и 514 Гц, и эффективной шириной полосы частот 14 и 7 Гц для первого и второго импульсов в паре соответственно. Спектры обоих импульсов ассиметричны. Второй импульс более коррелирован, чем первый и имеет время корреляции – 50 мс, а первый –18,3 мс.

3.2. Анализ звуков дятла

Дятлом издаются два характерных звука: первый – это стук клюва по дереву при поиске насекомых и выдалбливании дупла; второй – это непосредственно звук голоса дятла.

Полная временная реализация звуков, издаваемых дятлом, показана на рисунке 8, а фрагмент реализации, соответствующий стуку дятла по дереву – на рисунках 9-11.

3.2.1. Стук дятла. На рисунке 8 стук дятла представлен двумя пачками экспоненциальных импульсов. Средняя длительность пачки составляет 1,18 с, а количество импульсов в каждой пачке одинаково и равно 29.

Средний интервал между импульсами равен 60 мс, при этом вначале пачки импульсов интервал между первым и вторым импульсами составляет 107,6 мс для первой пачки и 91,7 мс для второй и имеет довольно большую разницу. Однако уже второй временной интервал между вторым и третьим импульсами составляет 77,9 мс для первой пачки и 78,52 мс для второй, и разница становится уже небольшой, какой и сохраняется до конца пачки. Временной интервал между импульсами в пачке постепенно уменьшается и к концу пачки достигает значения около 40 мс между 28 и 29 импульсами для обеих пачек.

Длительность одного импульса по уровню 0,5 составляет в среднем 3-5 мс для первой (рисунок 10) и 275 мкс для второй пачки импульсов (рисунок 11).

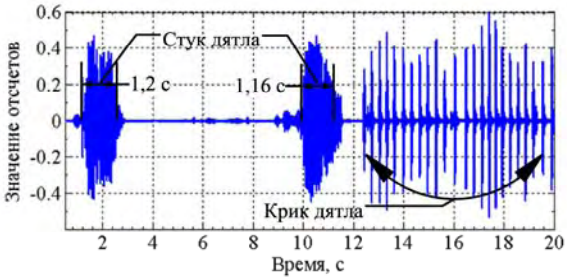


Рисунок 8. Временной график. Дятел.

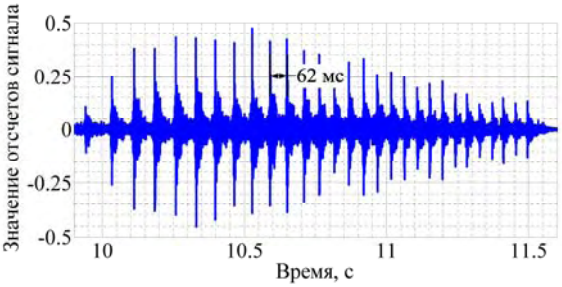


Рисунок 9. Стук. Вторая пачка импульсов.

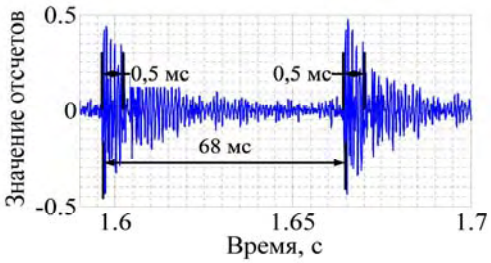


Рисунок 10. Стук. Импульсы первой пачки.

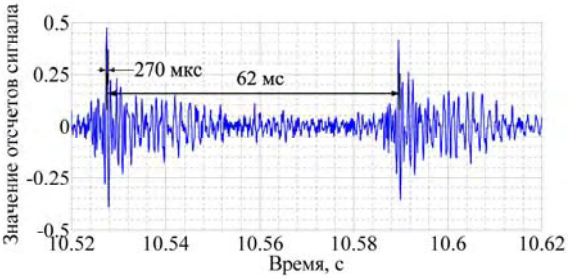


Рисунок 11. Стук. Импульсы второй пачки.

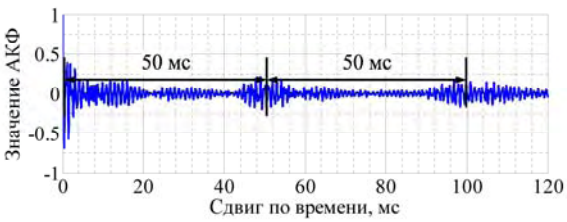


Рисунок 12. АКФ для фрагмента 2,0-2,5 с.

Оценка АКФ для фрагмента первой пачки импульсов приведена на рисунке 12. Форма АКФ для второй пачки схожа с формой АКФ для первой, поэтому график АКФ для второй пачки импульсов не приводится.

Анализ АКФ показывает слабую коррелированность сигналов стука дятла по дереву, причем как в самой пачке импульсов, так и самого экспоненциального импульса. Время корреляции импульсов первой пачки (полуширина основного зубца АКФ по уровню 0,5) составляет 124 мкс и практически одинаково для всей пачки. Небольшие всплески АКФ с периодом около 50 мс для фрагмента, показанного на графике, соответствуют среднему интервалу между импульсами в пачке. Сигнал второй пачки также практически не коррелирован. Время корреляции импульсов второй пачки несколько больше и составляет 139 мкс для промежутка времени 9,5-11 с и увеличивается к концу пачки до величины 147 мкс для промежутка времени 11-11,5 с.

Квадрат АЧС сигнала для фрагмента импульсов первой пачки показан на рисунке 13, а для второй пачки – на рисунке 14. Оба спектра построены для одинакового временного интервала анализа, равного 0,2 с. Форма АЧС в пределах пачки меняется незначительно.

Оба спектра линейчатые, поскольку время анализа спектра, равное 200 мс больше временного интервала между импульсами в пачке, составляющего около 70 мс для обоих временных интервалов (1,5-1,7 с и 10,0-10,2 с).

В отличие от звуков, издаваемых кукушкой стук дятла является широкополосным процессом, поскольку его ширина спектра сравнима с центральной частотой спектра.

Несмотря на то, что спектры для обеих пачек импульсов отличаются, в них можно выделить область частот 900-1200 Гц, в которой спектры похожи и которая является общей для обоих фрагментов акустического сигнала стука дятла.

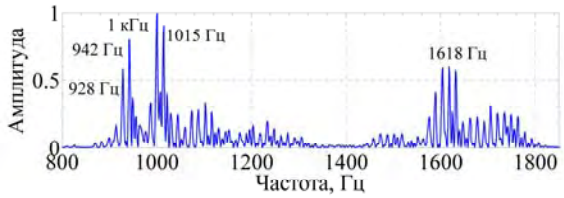


Рисунок 13. Квадрат АЧС фрагмента 1,5-1,7 с.

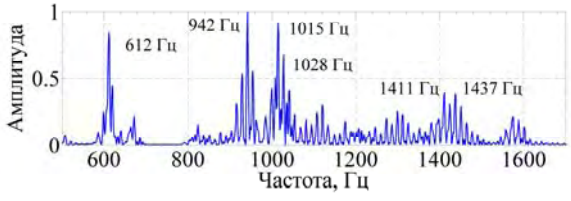


Рисунок 14. Квадрат АЧС фрагмента 10,0-10,2 с.

Можно отметить частоты 942 и 1015 Гц, как имеющие максимальное значение во всем спектре и присутствующие в спектрах обеих пачек импульсов. Кроме этого можно отметить область с частотами около 1,6 кГц, которая также присутствует в обоих спектрах.

Кроме этого следует отметить несимметричность спектра, что, как и для спектра звуков, издаваемых кукушкой, может свидетельствовать о присутствии угловой модуляции в сигнале.

3.2.2. Звук голоса дятла. Временная реализация звука голоса дятла, показанная на рисунке 8 представляет собой последовательность акустических импульсов с временным интервалом между импульсами от 0,26 до 0,45 с и средним значением временного интервала 0,32 с. Форма импульса показана на рисунке 15.

АКФ звука голоса дятла по форме похожа на АКФ стука дятла, показанную на рисунке 12. Акустический сигнал звука голоса дятла практически не коррелирован, время корреляции по уровню 0,5 составляет 0,3 с.

Квадрат АЧС звука голоса дятла показан на рисунке 16. Спектр звука голоса дятла – это спектр широкополосного процесса, максимум спектра которого находится на частоте около 4,5 кГц. Значения спектральных составляющих на частотах выше 4,6 кГц практически равны нулю. Основная энергия спектра содержится в интервале частот 2,9-4,6 кГц, а сам спектр несимметричен относительно своего максимума. В спектре можно выделить ряд частотных составляющих, имеющих максимальное значение: 2230, 3280, 3794, 4125, 4360 Гц. Спад спектра в сторону уменьшения частоты происходит по кривой, близкой к экспоненциальной. Спектральные составляющие ниже 2050 Гц практически равны нулю.

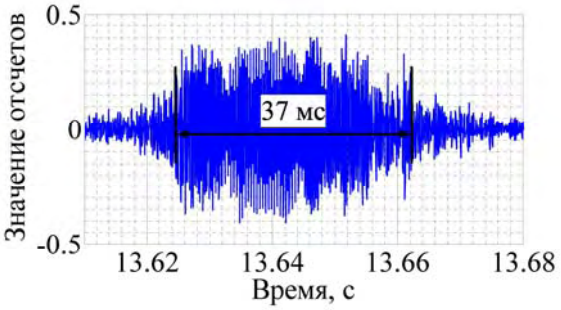


Рисунок 15. Форма импульса звука голоса дятла.

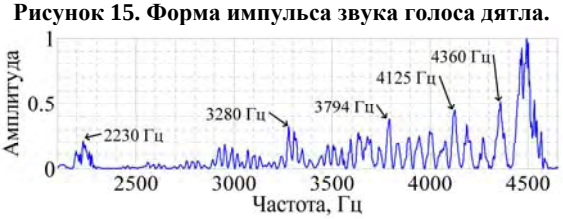


Рисунок 16. Квадрат АЧС фрагмента 13,95-14,15 с.

На основании анализа звуков, издаваемых дятлом можно сделать следующие выводы: *во-первых*, дятел издает четко выраженную последовательность звуковых импульсов, как при стуке клювом по дереву, так и своим голосом. Стук по дереву происходит в виде пачек экспоненциальных импульсов по 29 импульсов в каждой пачке. При этом ширина каждого импульса изменяется незначительно, а расстояние между импульсами в пачке уменьшается от значения, примерно равного 100 мс до 40 мс. Импульсы звука голоса дятла имеют среднюю ширину импульса по уровню 0,5, равную 37 мс. *Во-вторых*, как стук дятла, так и звук его голоса представляют собой широкополосные процессы. Основная энергия звука стука дятла находится в спектральном интервале 600-1800 Гц, а звука голоса дятла – в спектральном интервале 2900-4600 Гц. Спектры звуков стука клюва дятла по дереву и звука голоса дятла не симметричны относительно своего максимума.

[1] В.И. Арабаджи, К.И. Рудник, “О спектрах некоторых шумов естественного происхождения”, *Акустический журнал*, Т. 8, Вып. 4, 1962, с. 466-468.

[2] H.E Bass, “The propagation of thunder through the atmosphere”, *J. Acoust. Soc. Amer.*, Vol. 67(6), June 1980, pp. 1959-1966.

[3] Сергиенко А.Б., *Цифровая обработка сигналов: Учебник для вузов. 2-е изд.*, Питер, СПб., 2006.

Различные способы неспецифической активации шагательной активности в условиях разгрузки нижних конечностей у человека.

И. А. Солопова, В. А. Селионов
Институт Проблем Передачи Информации им. А.А.Харкевича РАН
solopova@iitp.ru

Аннотация

У многих позвоночных локомоторная активность связана со центральным генератором шагания (ЦГШ), активация которого зависит от его тонического состояния. Мы исследовали значение фактора тонических влияний на проявление локомоторных эффектов у человека, используя неспецифические, разномодальные афферентные входы для активации шагательной активности в условиях разгрузки нижних конечностей. С этой целью применяли вибростимуляцию мышц ног или электрическую стимуляцию дистальных нервов ног, прием Ендрассика и эффект Конштамма. Показано, что использование всех этих методов является существенным фактором, достаточным для активации локомоторного автоматизма у человека в условиях разгрузки. Наряду с этим исследовали влияние сигналов от разномодальных рецепторов стопы на характер вызванных движений созданием искусственной нагрузки на подошву вывешенной конечности. Дополнительный афферентный вход от рецепторов нагрузки изменял кинематический и электромиографический паттерн вызванных движений. Таким образом, неспецифические афферентные влияния могут повышать тоническую готовность ЦГШ и делать возможным активацию локомоторного автоматизма у человека. Создание афферентного входа, имитирующего нагрузку, является облегчающим фактором для инициации шагательных движений у человека.

1. Введение

Применение неспецифического раздражения афферентных проприоцептивных входов, дает возможность получить управляемый доступ к нейронным структурам, образующим центральные генераторы ритмики [1, 2]. Например, вибрация мышц горизонтально вывешенных ног вызывает циклические движения, сходные по своим кинематическим и электромиографическим характеристикам, с произвольными движениями, имитирующими шагательные движения, в тех же условиях. Предполагается существование сходных механизмов генерации этих движений. Результаты последних исследований показали, что спинной мозг человека обладает сетью, – центральным генератором шагания (ЦГШ), способной генерировать базовую ритмическую шагательную активность. Для запуска “шагания в воздухе” существенную роль играет повышение уровня неспецифической активации нейронных сетей спинного мозга, влияющих на тоническое состояние локомоторных структур. В качестве методов, способных вызвать неспецифическе изменение возбудимости спинного мозга может служить приём Ендрассика или эффект Конштамма. Приём Ендрассика преставляет собой произвольное изометрическое напряжение мышц сцепленных рук, приводящих к увеличению амплитуды сухожильных рефлексов и Н-рефлекса [3, 4], вызванных мышечных ответов при траскраниальной магнитной стимуляции [5]. Эффект Конштамма [6] служит средством проявления внутрицентральных механизмов генерации двигательных рефлексов и выражается в появлении непроизвольной тонической мышечной активности, сопровождающейся облегчением некоторых моторных реакций после сильного произвольного изометрического сокращения.

Короткая электрическая или механическая стимуляция кожи или кожных нервов, иннервирующих мышцы стопы, может вызывать различные ответы, зависящие от фазы цикла шага [7]. Безболезненная электрическая стимуляция, приложенная к дистальным нервам ноги может вызывать как облегченный, так и подавленный рефлекторный ответы [8, 9]. Такая стимуляция может имитировать реальные события, которые могут происходить при вертикальной локомоции

и наличии нагрузки на стопы, поэтому логично ожидать, что электрическая стимуляция нервов может возбуждать эти афферентные входы и приводить к инициации шагательных движений.

Афферентная обратная связь от рецепторов нагрузки играет ключевую роль в адаптации и модуляции работы ЦГШ в условиях реального окружения. Сенсорные входы могут оказывать существенное влияние на запуск, остановку или выбор двигательных паттернов [10]. Стопа является специализированной частью скелетно-мышечного аппарата, приспособленной для взаимодействия с опорой. Показана также потенциальная важность кожных входов от подошвы стопы в управлении ходьбой и позой [11]. Активность индивидуальных низко-пороговых механорецепторов в коже стопы вызывает мощную рефлекторную модуляцию выходящей ЭМГ-активности в мышцах, которые влияют на движения в стопе. При стимуляции нервов, иннервирующих стопу, ЭМГ ответы вызываются как на ипсилатеральной стороне стимуляции, так и контралатерально по отношению к ней. Несмотря на то, что ответы на контралатеральную стимуляцию, в среднем, меньше по сравнению с ответами, на ипсилатеральной стороне, наблюдается четкая модуляция ЭМГ-активности, которая обычно отличается от картин ипсилатеральной модуляции в тех же мышцах [12].

Механическая нагрузка на подошву стопы может активировать кожные афференты, имитируя контакт стопы с опорой, а кожная электростимуляция может быть использована для усиления мышечной активности в определенные периоды времени во время двигательной задачи. Непрерывное воздействие на рецепторы подошвы увеличивает возбудимость интернейронов в рефлекторных путях таким образом, что мотонейронный пул получает усиленный афферентный приток [13, 14].

Целью данной работы было: 1) исследование возможности вызова шагательных движений у человека в условиях разгрузки нижней конечности при использовании а) стимуляции разномодальных афферентных входов (вибростимуляции мышц, электрической стимуляции нервов, иннервирующих стопу), и б) приемов, облегчающих активацию ритмических движений (прием Ендрассика и феномен Конштамма). 2) оценка влияния афферентных входов от стопы на характер вызванных шагательных движений при приложении различной, искусственно созданной нагрузки на подошву вывешенной конечности.

2. Методика

В экспериментах приняли участие 25 здоровых испытуемых в возрасте 22-58 лет, не имеющих неврологических заболеваний и давших согласие на участие в эксперименте. Испытуемые лежали на правом боку, левая нога была вывешена в экзоскелетоне, дающем возможность горизонтальной ротации без трения в трех суставах: бедре, колене и стопе, и устранении влияния гравитации и внешнего сопротивления движению. Вывеска не ограничивала естественных движений ноги и позволяла совершать движения, имитирующие ходьбу как вперед, так и назад без ограничения амплитуды. Отсутствие опоры и горизонтальное положение испытуемого вносит сильные отличия в электромиографические и кинематические характеристики ходьбы: отсутствие необходимости в использовании большой группы мышц для совершения движений, изменения объемов движений в суставах движущейся конечности. Однако наличие в рисунке вызванных движений определенных фазовых соотношений [1, 2] между движениями в суставах вывешенной ноги, которые присущи нормальной наземной локомоции, позволяют интерпретировать полученные ритмические движения как шагательные.

Все испытуемые тестировались на возможность активации шагательной ритмики двумя способами: вибрационной стимуляцией или электрической стимуляцией. Непрерывную вибрационную стимуляцию наносили на четырехглавую мышцу бедра (m. Quadriceps femoris) вывешенной конечности. Вибратор представлял собой двигатель постоянного тока с прикрепленным эксцентриком, параметры вибрации: частота 20-60 Гц и амплитуда 0.8 мм. При электрической стимуляции непрерывная последовательность стимулов наносилась на поверхностные икроножный (n. Triceps surae) или малоберцовый (n. Peroneus fibularis) нервы невывешенной, неподвижной ноги. Применяли биполярную стимуляцию каждого из нервов, для нанесения стимулов использовали поверхностные электроды площадью 4 $см^2$ и специально сконструированный стимулятор постоянного тока с частотой следования стимулов 60 Гц и длительностью 0.3 мс. Использование частоты 60 Гц было оптимальным для облегчения ритмических рефлексов. Стимулы наносились с интервалом T=2c в течение промежутка времени d=1c. Располагая электроды вдоль стимулируемого нерва, добивались самой высокой плотности тока между электродами. Сила тока подбиралась таким образом, чтобы вызвать слабые сокращения мышц, не приводящие к движениям в суставах. Стимуляция была безболезненной и вовлекала, в основном,

низкопороговые афференты от механорецепторов стопы.

Для проверки влияния приема Ендрассика на генерацию циклической активности испытуемых просили растягивать сцепленные перед грудью руки, как на фоне подпороговой стимуляции, так и без внешних воздействий. Эффект Конштамма заключался в изометрическом напряжении группы гомонимных мышц бедра вывешенной ноги в течение 30-45 секунд против упоров, поставленных в одном или различных направлениях. После удаления упоров регистрировались возникающие движения.

В начале эксперимента регистрировали несколько проб произвольных шагательных движений. При активации шагательных движений при помощи электрической или вибрационной стимуляции испытуемых просили расслабиться и не вмешиваться в вызванные рефлекторные ответы. Для каждого вида электрической стимуляции, а также для вибрационной стимуляции эксперимент проводили в 5 условиях, которые чередовались в произвольном порядке: 1) вызванная локомоция без опоры под стопой, 2) вызванная локомоция с давлением дополнительной опоры на стопу 4Н, 3) вызванная локомоция с давлением дополнительной опоры на стопу 8Н, 4) шагательные движения с грузом 0.5 кГ, прикрепленным к стопе 5) шагательные движения с резиновым жгутом на переднем отделе стопы, создающим легкое давление на кожные рецепторы. В конце эксперимента тестировалось влияние приема Ендрассика на активацию ритмических рефлексов, а также проявление феномена Конштамма в условиях разгрузки ноги. В каждом из условий регистрировали по 2 пробы длительностью 60 с.

Для создания нагрузки, имитирующей контакт стопы с опорой (условия 2 и 3), на нее одевался «башмак», который с помощью резинового жгута имитировал равномерное давление на подошву с силой 4Н или 8Н. В условии 4 груз весом 0.5 кГ прикреплялся широкой полоской резины к передней части стопы со стороны подошвы. В условии 5 к стопе прикрепляли эту же резинку без груза.

В эксперименте регистрировали углы в тазобедренном, коленном и голеностопном суставах вывешенной ноги при помощи гониометров, расположенных латерально. Электромиограммы (ЭМГ) мышц голени и бедра той же ноги регистрировали при помощи поверхностных электродов и усилителя “ВАК” (США). Данные оцифровывались и подавались на компьютер с частотой оцифровки 800 Гц.

Вычисляли объем движений в каждом суставе, периоды шагательных движений, фазовые характеристики локомоции и электромиографическую активность каждой из регистрируемых мышц за цикл шага. Длительность цикла соответствовала усредненному периоду шагательных движений. Для выяснения значимости результатов использовали статистические программы ANOVA. Для парного сравнения использовали т-тест. Результаты статистического анализа считались достоверными, если вероятность ошибки была менее 0.05.

3. Результаты

3.1. Вибрационная и электрическая стимуляция как эффективные способы активации шагания «в воздухе»

Электрическая стимуляция нервов неподвижной ноги, как и вибростимуляция мышцы вывешенной ноги вызывали шагательные движения вывешенной конечности у 15 из 25 испытуемых. Латентный период вызванных движений у разных испытуемых составлял от нескольких секунд до нескольких десятков секунд, в ходе эксперимента происходило облегчение активации шагания с укорочением латентного периода начала движений. Для обоих видов стимуляции частота шагательных движений была устойчивой, слабо зависела от параметров воздействия и места его приложения. После выключения стимуляции движения могли продолжаться в течение нескольких циклов, нога могла совершать еще 3-5 и более циклов до полного затухания движения.

Кинематическая картина шагательных движений, вызванных как вибрацией, так и электрической стимуляцией, была сходной. Как правило, в циклические движения с периодом 2.5±0.4 с вовлекались голень и бедро, в голеностопном суставе движения отсутствовали, или же были незначительными. Объем движений вызванного вибрационной или электрической стимуляцией шагания для каждого из трех суставов ноги был сходен ($p>0.2$) и составил в среднем для тазобедренного сустава - 42.4±9.0º (в диапазоне 37º-53º), для коленного сустава – 60±13.5º (диапазон 60º-71º) и для голеностопного сустава 1.9±1.6º (в диапазоне 1.6-2.2º). Электрическая стимуляция вызывала шагательные движения с фазовыми соотношениями между тазобедренным и коленным суставами, характерными для ходьбы вперед (-0.24±0.09) или назад (0.24±0.3), причем ходьба назад наблюдалась лишь в 14% случаев, в то время как вибрационной стимуляции ходьба назад наблюдалась в половине случаев (49%).

Объем движений для каждого из трех суставов для ходьбы вперед или назад не отличался ($p>0.2$). При сравнении параметров вызванной локомоции и произвольного шагания в тех же условия вывески было обнаружено, что вызванные движения характеризовались более длительным периодом цикла шага ($p<0.03$). Кроме того было обнаружено, что объем вызванных движений в коленном суставе был значительно меньше ($p<0.03$), чем при произвольной ходьбе (79.7 ± 19.1^o).

Вызванные движения чаще всего сопровождались реципроктными вспышками электромиографической активности мышц-антагонистов бедра. Для мышц голени было характерно отсутствие ЭМГ активности или она была слабо выражена. Вызванная ходьба в условиях вывески ноги сопровождалась меньшей активностью мышц бедра и голени, чем произвольная ходьба.

3.2. Центральные тонические влияния

Прием Ендрасссика сам по себе не вызывал видимых движений в вывешенной ноге. Однако при выполнении приема после внешнего запуска 1-2 циклов движений движения продолжались в течение всего периода выполнения данного приема. Характеристики движений были сходными с аналогичными параметрами шагания, вызванными другими способами.

Феномен Конштамма также вызывал непроизвольные шагательные движения с соответствующими реципрокными вспышками в дистальных мышцах, после прекращения произвольного напряжения мышц вывешенной ноги. Частота движений вначале увеличивалась, достигая через 3-5 циклов максимальной величины, после чего прогрессивно затухала, сопровождаясь уменьшением амплитуды движений. Продолжительность колебаний составляла 5-60 с. Как во время выполнения приема Ендрассика, так и при феномене Конштамма движений в коленостопном суставе были незначительны ($<3°$).

3.3. Влияние дополнительных афферентных входов

Создание нагрузки на стопу подведением дополнительной опоры оказывало существенное влияние на дистальные отделы движущейся конечности. Голеностопный сустав активно вовлекался в локомоцию и объем движений в нем в среднем (по обоим видам стимуляции) составил 5.2 ± 3.3^o при нагрузке 4Н и 6.7 ± 3.8^o - при нагрузке 8Н. Статистический анализ показал, что возрастание объема движений по сравнению с шаганием без опоры было значимым как для условия с нагрузкой 4Н ($p<0.001$), так и для условия с нагрузкой 8Н ($p<0.001$). Амплитуда движений в коленном суставе также значимо возрастала при создании нагрузки по сравнению движениями в условии без нагрузки до 69.9 ± 23.7^o и 75.0 ± 20.1^o при нагрузке 4Н ($p<0.05$) и 8Н ($p<0.05$), соответственно. При возрастании нагрузки от 4Н до 8Н значимых изменений в амплитудах движений как в коленном, так и в голеностопном суставах обнаружено не было ($p=0.33$). В тазобедренном суставе независимо от степени давления на стопу объем движений не изменялся: 41.4 ± 16.7^o (4Н) и 43.1 ± 16.9^o (8Н) ($p=0.8$). Фазовые соотношения и периоды вызванных движений при подведении нагрузки были сходными с таковыми в условии без нагрузки.

Распределение электрической активности в движущейся конечности зависело от источника активации шагания. Наибольшее влияние искусственной нагрузки на стопу (условия 2 и 3) на электрическую активность мышц происходило при вибрационной стимуляции. При этом у всех испытуемых в какой-либо из дистальных мышц или в обеих сразу появлялась или возрастала электрическая активность (в среднем, на 100-150%), а у 75% испытуемых наблюдалось еще и возрастание ЭМГ-активности в какой-либо мышце бедра. При стимуляции малоберцового нерва подобные изменения в активности дистальных мышц наблюдались у 85% испытуемых, а в мышцах бедра ЭМГ активность практически не изменялась при приложении нагрузки, и лишь в нескольких случаях наблюдалось слабое (не более 40%) ее увеличение в одной из мышц. Для икроножного нерва в этих же условиях влияние дополнительной опоры на ЭМГ-активность дистальных мышц было еще меньшим и наблюдалось только у 40% испытуемых, а в мышцах бедра электрическая активность не изменялась.

Прикрепление дополнительного груза к стопе как при вибрационной, так и при электрической стимуляции не изменяло объем вызванных движений в тазобедренном и коленном суставах по сравнению с шаганием без дополнительных возмущений ($p=0.8$), в то время как движения в голеностопном суставе имели тенденцию к увеличению, однако это увеличение не было статистически значимым ($p=0.07$). Для всех способов активации шагания электрическими стимулами в условии 4 ЭМГ-активность мышц бедра и голени практически не изменялась по сравнению с активностью этих же мышц при невозмущенном шагани и лишь при применении

вибрации в 40% случаев инерционная нагрузка вызывала увеличение ЭМГ-активности мышц бедра и голени, но это увеличение было менее выражено, чем в условиях 2 и 3. Прикрепление резинового жгута к стопе не влияло на кинематические и электрофизиологические параметры вызванной локомоции.

4. Обсуждение

Вибрационная стимуляция или стимуляция малоберцового и икроножного нервов вызывает непроизвольные шагательные движения вывешенной конечности со сходными кинематическими и электрофизиологическими характеристиками. По-видимому, стимуляция запускает шагательный автоматизм, который, по крайней мере, частично имеет ту же природу, что и активация шагания при естественной локомоции. Например, в экспериментах на животных показано [15], что тонический афферентный приток, вызванный электрическими стимулами в низкопороговых афферентных волокнах, преобразуется в ритмическую активность мотонейронов. Вызванные шагательные движения могут быть связаны с тем, что поток импульсов, который формируется в результате раздражения сенсорных афферентов, изменяет тоническое состояние интернейронов, формирующих шагатательный генератор, который, в свою очередь, запускает шагательный автоматизм. Генерация ритмических движений происходит до тех пор, пока поддерживается постоянный афферентный тонический приток в сенсорных волокнах, проецирующихся в спинной мозг.

Отличительной чертой наших исследований является тот факт, что удалось активировать шагательные движения электрической стимуляцией, приложенной к контралатеральной ноге. Как показано в ряде работ [9, 12] ритмические движения, которые совершаются при сокращении мышц на одной стороне тела, дают фазово-связанные изменения возбудимости гомологичных моторных путей противоположной конечности. Установлено также, что поддерживаемое произвольное сокращение мышц на одной стороне тела ведет к увеличению возбудимости контралатеральных гомологичных моторных путей [16]. Наши результаты подтверждают предположение о перекрестных межгенераторных связях, дающих возможность прямого доступа к генератору контралатеральной конечности.

Настоящие результаты, а именно стабильность шагового цикла, схожесть параметров вызванного шагания с характеристиками произвольных шагательных движений, и, что особенно существенно, возможность запуска шагания с контралатеральной стороны, подтверждают ранее сделанные выводы [1, 17] о центральной природе шагательного автоматизма. Возможность активации шагательных движений как вперед, так и назад при неизменных условиях стимуляции также подтверждает это допущение. Однако ходьба назад в условиях активации шагательных движений электрической стимуляцией была выражена в гораздо меньшей степени и происходила лишь эпизодически по сравнению с условиями шагания при вибрации. По-видимому, активация шагания вибростимуляцией мышц может быть, в большей мере, связана с возбуждением мышечных афферентов, что приводит к более распределенной активации мотонейронов в спинном мозге. В то же время при запуске шагания электрической стимуляцией происходит более адресная посылка импульсации к мотонейронным пулам, запускающим ритмические движения.

Прием Ендрассика оказывает специфическое тоническое влияние на мышечную и рефлекторную активность [18]. Полученные в наших экспериментах ритмические движения вывешенной ноги, совершаемые на фоне приема Ендрассика, подтвердили его фасилитаторную роль во время активации ритмических рефлексов. В условиях разгрузки позный постэффект переходит в ритмические движения ног. Нами показано, что под влиянием повышенного мышечного тонуса (афферентация от напряженных мышц) в спинном мозге возникают периодические изменения возбудимости, формирующиеся собственными механизмами сегментарного уровня. По-видимому, эти периодические изменения возбудимости играют существенную роль в реализации двигательной программы, которая организуется спинальными нейронными сетями.

Существенный вклад в общую картину генерации шагательных движений вносят рецепторы нагрузки от подошвы. Дополнительные афферентные сигналы, обусловленные давлением на стопу, могут облегчать инициацию и поддержание индуцированного шагания “в воздухе”. В наших экспериментах локально-специфическая информация от механорецепторов стопы, генерируемая искусственно созданной нагрузкой на подошву во время вызванного шагания, существенно изменяла картину шагания: происходило увеличение объема движений в коленном и голеностопном суставах и часто наблюдалось усиление модуляции ЭМГ активности мышц бедра и голени. Подобные изменения паттерна вызванных шагательных

движений наблюдались в не зависимости от способа вызова шагания. Следует заметить, что при имитации нагрузки на стопу в 4Н и 8Н не наблюдалось значимых различий во влиянии на параметры движений. По-видимому, это обусловлено тем, что при создании определенного уровня давления импульсация механорецепторов достигает своей максимальной частоты, так что дальнейшее увеличение давления оказывается мало эффективным. Усиление ЭМГ-активности в дистальных мышцах подтверждает активное участие афферентного притока от дополнительной нагрузки в состоянии генератора ритмики, что сказывается в усилении моторного выхода.

Гравитационная составляющая, имитированная подведением опоры, оказывала большее влияние на вызванное шагание, чем инерционная добавка при прикреплении груза для обоих методов вызова шагания. Эти данные согласуются с работой [19], которые показали, что активные вертикальные генерированные силы против опоры во время бега изменялись в соответствии с изменениями веса тела, в то время как прикрепление дополнительного груза к голени не сказывалось на исследуемых движениях. Закрепление резинового жгута на стопе воздействовало только на тактильные кожные рецепторы, не затрагивая рецепторы других модальностей. Таким образом, для более эффективного влияния на общий рисунок вызванного шагания в условиях разгрузки необходимо использование наиболее сильного афферентного входа, а именно активация рецепторов давления вывешенной конечности. Дальнейшее изучение влияния различных афферентных раздражений дистальных отделов ноги, изменяющих рисунок естественного шагания, на движения, вызванные искусственной стимуляцией, может помочь прояснить механизмы возникновения и поддержания циклических движений и их связь с генератором шагания спинного мозга.

Работа выполнена при поддержке РФФИ (грант № 08-04-01200).

5. Список литературы

[1] В.С.Гурфинкель, Ю.С Левик, О.В.Казенников, В.А.Селионов. Существует ли генератор шагательных движений у человека? Физиология человека 1998. т.24. N 3. с.42-50.

[2] Gurfinkel V.S., Levik Yu.S., Kazennikov O.V., Selionov V.A. "Locomotor-like movements evoked by leg muscle vibration in humans" E. J. of Neuroscience. 1998. V.10. P. 1608-1612.

[3] Bussel B., Morin C., Pierrot-Deseilligny E. Mechanism of monosynaptic reflex reinforcement during Jendrassik maneuvre in man J.Neurosurg.Psychiatry. 1978. V.41. N1. P.40-44;

[4] Kameyama O., Hayes K.,C., Wolfe D. Methodological considerations contributing to variability of the quadriceps H-reflex Am.J.Phys.Med.Rehabil. 1989. V .68. N 6. P. 277-82.

[5] Pereon Y., Genet R., Guiheneuc P. Facilitation of motor evoked potentials: timing of Jendrassik maneuvre effects Muscle nerve, 1995. V.18. P.1427-1432.

[6] Mathis, J., Gurfinkel, V.S. & Struppler, A. Electroenceph. Clin. Neurophysiol. 1996. V.101, p.289-297.

[7] Wezel B.M.H., Engelen B.G.M., Gabreëls F.J.M., Gabreëls-Festen A.A.W.M., and Duysens J. A Fibers mediate cutaneous reflexes during human walking. J. Neurophysiol. 2000.V. 83. N. 5. p. 2980-2986.

[8] Duysens, J., Trippel, M., Horstmann, G. A., Dietz, V. Gating and reversal of reflexes in ankle muscles during human walking. Exp. Brain Res. 1990. V.82. p.351-358.

[9] Tax, A. A. M., Wezel B., Dietz V.. Bipedal reflex coordination to tactile stimulation of the sural nerve during human running. J. Neurophysiol. 1995. v.73. p.1947-64.

[10] Rossignol S., Dubuc R. and Gossard J. Dynamic Sensorimotor Interactions in Locomotion. Physiol. Rev. 2006. V.86. p.89-154.

[11] Fallon J. B., Bent L. R., McNulty P. A. and Macefield V. G. Evidence for strong synaptic coupling between single tactile afferents from the sole of the foot and motoneurons supplying leg muscles. J Neurophysiol 2005. V.94. p. 3795-3804.

[12] Carson R.G., Riek S, Mackey D. C., Meichenbaum D. P., Willms K., Forner M. and Byblow W. D. Excitability changes in human forearm corticospinal projections and spinal reflex pathways during rhythmic voluntary movement of the opposite limb. J Physiol. 2004. V. 560; N 3. p.929-940.

[13] Burke D., Dickson H.G. and Skuse N.F. Task-dependent changes in the responses to low-threshold cutaneous afferent volleys in the human lower limb. J. Physiol. 1991. V.432. p.445-458.

[14] Delwaide, P. J., Crenna, P. & Fleron, M. H. Cutaneous nerve stimulation and motoneuronal excitability. I. Soleus and tibialis anterior excitability after ipsilateral and contralateral sural nerve stimulation. J. Neurology, Neurosurgery and Psychiatry 1981. V.44. p. 699-707.

[15] Grillner, S. Neurobiological bases of rhythmic motor acts in vertebrates. Science 1985. V.228. p.143-149.

[16] Cernácek J (1961). Contralateral motor irradiation: Cerebral dominance – its changes in hemiparesis. Arch Neurol 4, p.165-174.

[17] Gurfinkel VS, Ivanenko YP, Levik YS, Kazennikov OV, and Selionov VA. The neural control of posture and locomotion: a lock with two keys. In: Motor Control, Today and Tomorrow. Sofia, Bulgaria: Academic Publishing House. 1999. p. 113-121.

[18] Delwaide PJ, Toulouse P Facilitation of monosynaptic reflexes by voluntary contraction of muscle in remote part of the body. Mechanisms involved in Jendrassik Manoeuvre. Brain, 1981, 104: 701-709.

[19] Chang Y-H., Huang H-W. C., Hamerski C. M. and Kram R. The independent effects of gravity and inertia on running mechanics. J. Exp. Biol. 2000. V.203. p.229-238.

Наследственные нарушения слуха: дифференцирование 60 случаев ранней глухоты и тугоухости

Диана Баталова
Институт проблем передачи информации имени А. А. Харкевича РАН
batalova@iitp.ru

Валентина Божкова
Институт проблем передачи информации имени А. А. Харкевича РАН
bgk@iitp.ru

Аннотация

Целью работы является создание простой и полезной для клиники системы предварительной диагностики генетических форм нарушений слуха на основе анализа фенотипических признаков. Система позволит установить наследственную природу нарушения и предсказать его наиболее вероятную генную форму. В основу системы положено пошаговое разделение типов с использованием набора ключевых признаков, выбранных на основе анализа электронной базы данных наследственных болезней человека (OMIM). Сведения о фенотипе получены из анкетных данных и анализа родословных. Иерархический процесс организован в виде графа вероятностного прогнозирования. Оценку эффективности системы предполагается провести по совпадению результата предварительной диагностики и данных ДНК-анализа, полученных независимым способом.

1. Введение

Нейросенсорные нарушения слуха – наиболее часто встречающееся сенсорное расстройство у людей. Оно обнаруживается у, более чем, 30 млн. детей. Как показали многие исследования, глухота – это гетерогенное заболевание, причем около половины случаев - это различные моногенные формы [1]. В связи с этим чрезвычайно актуальна ДНК-диагностика нарушений слуха, которая позволяет уточнить диагноз и оценить вероятность сопутствующих аномалий, предсказать характер развития дефекта, грамотно провести генетическое консультирование.

Последние несколько лет ознаменовались прорывом в области молекулярной генетики глухоты: было выявлено около 100 локусов на разных хромосомах, связанных с нарушениями слуха. В них идентифицировано несколько десятков генов с различными типами мутаций [2]. Молекулярные исследования подтвердили генетическую гетерогенность глухоты и одновременно подчеркнули трудность ее массовой диагностики. Она связана, в первую очередь, с тем, что скрининг для каждого пациента всех обнаруженных в настоящее время мутаций стандартными методами становится процедурой очень дорогостоящей. Из этой ситуации видно два выхода. Первый состоит в развитии недорогих методов, позволяющих одновременно выявлять сразу несколько мутаций в разных генах. Второй предполагает разработку вспомогательных диагностических приемов, позволяющих предсказать наиболее вероятную генную форму дефекта у данного пациента, чтобы затем направить его на конкретную ДНК-диагностику. И тот, и другой подход нуждается в априорных знаниях о частоте встречаемости различных форм нарушений слуха, которые могут быть получены только экспериментально.

2. Постановка задачи

Целью нашей работы является разработка ясной и удобной системы пошагового разделения различных форм нарушений слуха с использованием фенотипических признаков, которая в конечном счете приводила бы каждого пациента к наиболее вероятной для него форме заболевания. Такая система имеет явные преимущества перед системой классификации по совокупности признаков. Чтобы дифференцировать разные формы напрямую, нужно провести сопоставление множества наследственных типов. К примеру, база данных OMIM (электронная база данных по менделирующим наследственным болезням человека), поддерживаемая университетом Джона Хопкинса, в разделе глухота (Deafness) содержит материал по 719 наследственным формам [2]. Работа с ними представляет собой трудоемкое занятие даже для профессионала, в первую очередь потому, что не имеется ярких клинических признаков для отдельных

генетических форм, хотя они и различаются по возрасту проявления дефекта, динамике его прогрессирования, сопутствующим патологиям, типу наследования и др. признакам.

Эффективность системы предварительной диагностики зависит от выбора ключевых признаков и их приоритетности. После анализа литературы и базы данных OMIM в качестве ключевых признаков были выбраны следующие: время первоначального проявления дефекта, наличие семейной истории, тип наследования, сопутствующие патологии, динамика прогрессирования. Ряд признаков, такие как степень тугоухости были отнесены к дополнительным (не ключевым). Было предположено, что эти фенотипические признаки несут важную информацию о генетической природе заболевания. Сведения о фенотипе были получены методами анкетирования и анализа родословных. Дополнительные характеристики были получены из анализа аудиограмм. Конечной целью работы была проверка эффективности системы предварительной диагностики данными ДНК-тестирования. Кандидатное тестирование было проведено для *GJB2* гена, мутации в котором являются наиболее частой причиной ранней детской глухоты и тугоухости (тип DFNB1; связан с дефектом в белке коннексине 26). Нарушения слуха, вызванные дефектом *GJB2* гена, относятся к аутосомно-рецессивным, совпадают с началом функционирования улитки и не сопровождаются другими патологиями [3].

3. Результаты

Были проанализированы безвыборочно истории болезни и родословные 60 студентов сурдоотделения дефектологического факультета МПГУ, давших согласие на участие в исследовании. У всех из них имелся диагноз двустороннего нейросенсорного нарушения слуха (кроме одного случая отосклероза с кондуктивной тугоухостью), на уровне тугоухости II-IY степени или полной глухоты. Предлагаемая система предварительной диагностики нарушений слуха схематически представлена на рис. 1. На первом этапе разделение типов глухоты проводилось исходя из времени манифестации нарушения. Как показал анамнез, у большинства испытуемых (84%) нарушения слуха возникли в долингвальном периоде (I группа). Нарушения были выявлены или вскоре после рождения, или в течение первых трех лет жизни включительно. У остальных 16% (II группа) начало болезни приходилось на постлингвальный период, главным образом, на возраст 3-5 лет. В обеих группах встречались молодые взрослые как со средней степенью нарушений слуха, так и с полной глухотой, однако последние попадали преимущественно в I группу. Разделение на долингвальную и постлингвальную глухоту для наследственных случаев практически полностью совпадает с разделением на аутосомно-рецессивную и аутосомно-доминантную глухоту [2], что может быть использовано далее при формировании гипотезы о типе нарушения.

Для последующего дифференцирования была использована генеалогия. Родословные и анамнез подтвердили семейную природу заболевания в 53% случаев, ненаследственную (приобретенную) в 17% случаев, а 30% представляли собой спорадические случаи с неясной этиологией (таблица 1). В последней группе могли оказаться генетические нарушения, которые не были выявлены генеалогическим методом из-за недостатка данных или потому, что представляли собой спонтанные мутации, а также плохо диагностируемые приобретенные нарушения. Приблизительно то же соотношение наследственных, приобретенных и неясных спорадических случаев наблюдалось порознь в группах I и II.

Уже на этом этапе можно выделить группу испытуемых, которые с большой вероятностью имели дефект в *GJB2* гене - это молодые люди с глубокой глухотой и наличием семейной истории. Известно, что более половины наследственных случаев полной глухоты в США является следствием именно этого дефекта.

Далее были разделены синдромальные и несиндромальные (моносимптоматические) наследственные типы. В изученной группе вероятность синдромальных форм имелась в 8-ми семьях, т.е. вероятные синдромальные типы составили 14% от общего числа испытуемых, что соответствует мировым данным (таблица 1). Среди них оказались синдром Стиклера-Маршалла (тугоухость с миопией и нарушениями опорно-двигательного аппарата), синдром Джервелла и Ланге-Нильсена (глухота с нарушениями сердечного ритма), синдром Альпорта, а также несколько случаев, которые требуют специального анализа. Более половины из них имели долингвальное проявление дефекта.

В наследственной моносимптоматической подгруппе классический аутосомно-рецессивный тип (спорадическая глухота или тугоухость сразу у нескольких детей слышащих родителей) обнаружен только в 25 % случаев, что составляет 10 % от всех испытуемых. Все эти случаи – из группы I (тип IA, см. таблицу 1). Аутосомно-доминантный тип предположен в 4-х семьях (17% среди наследственных моносимптоматических случаев) (тип D) (часть из них – также из группы I). В остальных случаях тип наследования установить не удается. Это связано, в первую

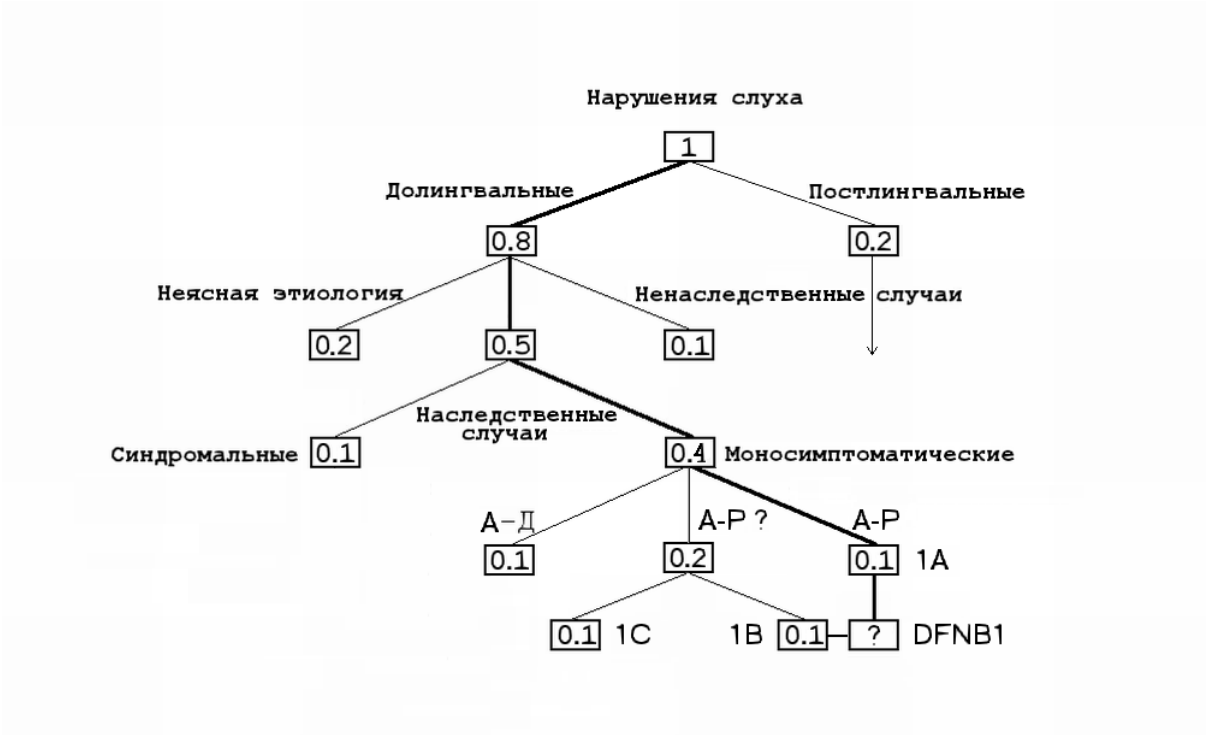


Рисунок 1. Схема предварительной диагностики нарушений слуха; А-Р – аутосомно-рецессивный тип, А-Д – аутосомно-доминантный тип, DFNB1 – один из видов детской глухоты

очередь, с высокой частотой ассортативных браков среди глухих и тугоухих. Так, у более чем половины детей с нарушениями слуха в нашей группе оба родителя были с глухотой или тугоухостью. По причине большой распространенности аутосомно-рецессивных мутаций в популяции глухих и, в частности, мутаций в гене *GJB2* оба родителя могут быть гомозиготными по гену *GJB2*. Кроме того, в таком браке один родитель может быть гомозиготен по аутосомно-рецессивному гену, а другой родитель, будучи гетерозиготным по аутосомно-доминантному гену, одновременно нести мутацию в том гене, по которому

Таблица 1. Генеалогия нарушений слуха в исследованной группе молодых взрослых.

Приобретенные случаи – 17 % (10 человек)		ДЦП, патология родов (2 случая) Резус-конфликт (1 случай) Менингит (2 случая) Другие инфекции (2 случая) Ототоксичность (3 случая)
Генные формы – 53 % (32 человек)	Из них: синдромальные – 26 % (8 человек)	Синдром Стиклера-Маршалла (3 случая) Синдром Джервелла и Ланге-Нильсена (1 случай) Синдром Альпорта (1 случай) Недиагностированные синдромы (3 случая)
	несиндромальные – 74 % (24 человека)	Аутосомно-рецессивный тип (классический) (IA) (6 случаев Аутосомно-доминантный тип (ID+IID) (4 случая) Аутосомно-рецессивный (с большой вероятностью) (IB) (9 случаев) Неясный тип наследования (IC) (5 случаев)
Спорадические случаи неясной этиологии – 30 % (18 человек)		

гомозиготен первый. Поэтому вероятность передачи дефекта аутосомно-рецессивного типа в семьях двух глухих родителей остается очень высокой. В связи с этим мы сочли возможным выделить из подгруппы испытуемых с неясной этиологией (тип IC) подгруппу, имеющую с высокой вероятностью аутосомно-рецессивный тип нарушения. В нее вошли дети, родители которых с большой вероятностью имели нарушения аутосомно-рецессивного типа, а также дети из таких семей, в которых имелись дальние родственники с нарушениями слуха (тип IB) (таблица 1).

Оценить эффективность предлагаемой схемы может показать кандидатное тестирование. В настоящее время тестирование проводится для *GJB2* гена, мутации в котором являются наиболее частой причиной ранней детской глухоты и тугоухости. Для ДНК-тестирования получены образцы крови 10 испытуемых: 4-х из подгруппы IA, 4-х из IB и 2-х из IC. ДНК-диагностика проводится в ДНК-лаборатории Медико-генетического научного центра РАМН.

Подгруппа IA с классическим аутосомно-рецессивным типом является кандидатом номер 1 дефектность *GJB2* гена, поскольку прямо подпадает под фенотип глухоты и тугоухости DFNB1 типа. Однако она не может представлять собой все случаи этого типа глухоты. По зарубежным источникам, при молекулярно-генетическом тестировании дефектность *GJB2* гена выявляется в более чем 20% всех случаев детской глухоты или в более чем 40% совокупных случаев наследственных и спорадических нарушений [3]. Вероятно, что дефект в *GJB2* гене имеет также большинство испытуемых из группы IB. При этом самая большая вероятность, по-видимому, у тех из них, которые имеют полную глухоту с рождения (см. выше).

4. Обсуждение

Предлагаемая схема предварительной диагностики нарушений слуха основана на том, что самые простые фенотипические признаки могут нести, на самом деле, важную информацию о генетической природе заболевания. Изучение мутаций генов, приводящих к нейросенсорной глухоте, показало, что значительная их часть нарушает морфогенез волосковых клеток, другая же вызывает деградацию уже сформированных рецепторов [4]. В первом случае дефект возникает в рецепторах в пренатальном периоде, во втором, они, как правило, деградируют в раннем детстве. И в том, и в другом случае результатом гибели волосковых клеток становится постепенная атрофия кортиева органа и потеря слуха. Насколько рано начинаются изменения, зависит в большой степени от типа мутации. Аутосомно-рецессивные мутации вызывают, как правило, врожденную глухоту и тугоухость, которая не прогрессирует. Аутосомно-доминантные мутации чаще вызывают поздние детские прогрессирующие нарушения, так как некоторое время система слуха еще может функционировать на минимальных количествах нормального белка. Однако понять все особенности этих механизмов можно будет только тогда, когда влияние мутаций на рецепцию звука будет изучено на клеточном и тканевом уровне.

5. Заключение

По зарубежным данным молекулярно-генетическое тестирование детской глухоты в 20% случаев выявляет дефект в *GJB2* гене. Остальные случаи представляют собой большое разнообразие приобретенных и наследственных форм, для большинства из которых молекулярной диагностики пока еще не разработано. Дифференцировать эти случаи на основе предварительного анализа клинической картины и родословных достаточно трудно. В настоящей работе мы предлагаем вариант простой и полезной для клиники системы пошагового дифференцирования нарушений слуха на основе ключевых фенотипических признаков. В ее основу положен анализ 60 случаев ранней детской глухоты разной этиологии. Эффективность системы проверяется ДНК-тестированием на наличие DFNB1 глухоты, вызванной мутациями в *GJB2* гене. По предварительным данным, такая система позволяет предсказать данный дефект с вероятностью 90%. Кроме того, она дает основания для предположения других наиболее часто встречающихся форм глухоты и тугоухости как синдромальной, так и несиндромальной природы.

6. Список литературы

[1] C. Morton, «Genetics, genomics and gene discovery in the auditory system», *Hum. Mol. Genet*, 2002, V. 11, pp. 1229-40.

[2] OMIM - база данных наследственных болезней человека http://www3.ncbi.nlm.nih.gov/omim.

[3] F. Denoyelle, S. Marlin, D. Weil, L. Moatti, P. Chauvin, E. Garabedian, C. Petit, «Clinical features of the prevalent form of childhood deafness, DFNB1, due to a connexin-26 gene defect: implications for genetic counseling», *Lancet*, 1999, V. 353, pp. 1298-30.

[4] В.П. Божкова, «Наследственные нарушения слуха: роль щелевых контактов», *Сенсорные системы*, 2008, Т. 22, N 1, с. 3-19.

Роль структурной организации ЭЭГ-информации при выявлении патогенетического механизма алалии.

Перемячкина С.Г., Попова С.М., Лукашевич И.П.
E-mail: luk@iitp.ru; peremyachkina-sv@mail.ru

Аннотация

Алалия – системное недоразвитие речи у детей, при котором нарушаются все компоненты речи. Причинами алалии как по нашим, так и по литературным данным являются перинатальные осложнения. Наши исследования показали, что в первую очередь к ним относятся факторы риска при внутриутробном развитии плода. Патогенетические механизмы этого речевого синдрома до сих пор не изучены.

Мы исследовали 24 электроэнцефалограммы (ЭЭГ) детей с алалиями в возрасте от 3 до 5,5 лет. Описание ЭЭГ проведено с помощью экспертной системы «ЭЭГ-ЭКСПЕРТ». В основу системы положена структурно-организованная схема представления ЭЭГ-информации. Схема отличается полнотой описания и универсальностью, а также включает возрастные критерии. Сравнение полученных результатов с нашим исследованием ЭЭГ-данных больных с сосудистой патологией зрительного бугра, позволило высказать предположение, что основным патогенетическим механизмом алалии является левостороннее поражение структур, связанных с генерацией альфа-ритма – задними отделами таламуса и коры.

1. Введение

Целью работы являлось выявление патогенетического механизма речевого синдрома алалии с помощью структурно-организованной ЭЭГ - информации.

Алалия – системное недоразвитие речи у детей, при котором нарушаются все компоненты речи и, в первую очередь, экспрессивная речь.

Метод структурной организации информации заключается в установлении основных систем и объектов в рассматриваемой области знаний; описании функций, выполняемых этими системами, и симптомов, которые выявляются при нарушении функций. При этом выделяется минимальное числа блоков (структурных единиц), знание которых достаточно для принятия решения. [1,2].

Анализ клинических данных показал, что наиболее информативными оказались данные родового анамнеза и показатели ЭЭГ.

Рассмотрим более подробно результаты этих исследований.

2.Структурная организация ЭЭГ-информации

Повышение уровня ЭЭГ- диагностики остается актуальной проблемой в области электрофизиологических исследований.

Решение этой проблемы можно искать в разных направлениях, например, создавая новые, более совершенные приборы; получая новые знания об ЭЭГ- феноменах, а также, используя метод структурной организации, имеющихся на сегодня знаний. Именно с последним из перечисленных способов и связано наше исследование.

Остановимся более подробно на характеристике и применении в практической и научной деятельности схемы представления ЭЭГ-информации.

Предлагаемый нами структурный анализа ЭЭГ [1] строится на распределении признаков, выделяемых при визуальной оценке электроэнцефалограммы, по блокам, характеризующим: состояние ритмогенных систем коры головного мозга; диффузные (общемозговые) отклонения электрической активности (ЭА); локальные отклонения ЭА, включая локальные индуцированные изменения; состояние подкорковых регуляторных структур разных уровней. Внутри каждого блока информация представляется в двух видах: в виде переменных: ЭЭГ- феноменов, их локализации и реакции на внешние воздействия; и функций, являющихся выводами из анализа переменных и требующих дополнительных знаний и аргументации.

Разработанная схема представления данных электроэнцефалографического обследования может быть использована для решения любых задач, связанных с визуальным анализом ЭА мозга, поскольку ее построение не зависело от

конкретных постановок. Эти задачи, помимо их значения для фундаментальных исследований мозга человека и практического значения для медицины, оказались удобными для отработки одного из ключевых вопросов информатики связанного с сокращением перебора - структурной организацией знаний и информации [3].

К достоинствам схемы следует отнести: полноту на современном уровне знаний с возможностью добавления новой информации без изменения структуры; компактность; использование традиционного языка описания и автоматическое формирование заключения, представляющего из себя совокупность функций по всем блокам. Структурированная таким способом ЭЭГ- информация, включающая определения и аргументацию для оценки функций, представляет собой базу знаний в области электроэнцефалографии и может быть использована для профессионального обучения.

Структурный подход был использован нами для создания автоматизированной системы “ЭЭГ-ЭКСПЕРТ”, которая успешно применяется на практике при ЭЭГ-исследованиях в ряде медицинских учреждений и в научно-исследовательских институтах Москвы и РФ.

Компьютерная автоматизированная система “ЭЭГ-ЭКСПЕРТ” предназначена для описания, хранения данных визуального анализа электроэнцефалограммы и формирования экспертного заключения о функциональном состоянии мозга и его отдельных структур.

Система включает в себя три параллельно функционирующие программы.

Программа "Диалог" представляет собой организованную в виде опросника схему описания данных визуального анализа ЭЭГ. Пользователю необходимо выбрать из перечня энцефалографических признаков, предъявляемых на экране монитора, те, которые соответствуют анализируемой кривой. Однородные по функциональному значению ЭЭГ-признаки объединены в блоки. Внутри блоков информация содержится в виде переменных, представляющих из себя описание ЭЭГ- феноменов, их топографии и реакции на функциональные пробы, а также функций - выводов, отражающих характер и выраженность изменений, описываемых в данном блоке. Вся записанная информация заносится в каталог, где хранится и при необходимости может быть выведена на экран.

Следующая программа "Справка" позволяет, находясь в любой позиции опросника, ознакомиться с описанием данного ЭЭГ-признака и получить рекомендации для определения локализации, характера и выраженности отмеченных изменений. Информация приведенная в “Справке” является результатом анализа, как собственных данных, так и материалов других исследователей, и является полной на современном уровне знаний. В «Справке» приведен полный атлас ЭЭГ-феноменов.

Третья программа "Вывод" автоматически формирует, выводит на экран, а затем на печатающее устройство описание ЭЭГ и заключение о функциональном состоянии мозга в соответствии с отмеченными в опроснике признаками. Заключение содержит информацию об отклонениях в электрической активности структур мозга, их характере и выраженности, соответствии уровня развития электрической активности возрасту (для детей и подростков), а также о снижении порога судорожной готовности при наличии соответствующего паттерна ЭЭГ. Алгоритм для формирования заключений был составлен с учетом опыта высококвалифицированных специалистов.
Приведем пример заключения:

ЦЕНТР ПАТОЛОГИИ РЕЧИ И НЕЙРОРЕАБИЛИТАЦИИ ДЗ г. МОСКВЫ

Ф.И.О.: Захаров Тимофей
Возраст: 4 года 9 мес.
Дата обследования: 14.01.08
Диагноз: алалия.

ЭЛЕКТРОЭНЦЕФАЛОГРАММА:

ХАРАКТЕРИСТИКИ АЛЬФА-РИТМА:

Альфа-ритм нерегулярный, дезорганизованный, частота 9 кол/с, амплитуда 40-70 мкв, доминирует в затылочных и теменных отведениях. Реакция следования на РФС отсутствует.

При первом предъявлении сенсорного сигнала возникает ориентировочная реакция в виде десинхронизации альфа-ритма. В течение первых 30 сек. нет выраженной реакции основного ритма на ГВ.

В левом полушарии наблюдается более дезорганизованный, сниженный по амплитуде альфа-ритм.

ДИФФУЗНЫЕ ИЗМЕНЕНИЯ ЭА:

На этом фоне наблюдаются диффузные полиморфные тета-колебания, диффузные бета-колебания, двухфазные пики без четкой локализации.

ЛОКАЛЬНЫЕ ИЗМЕНЕНИЯ ЭА:

В затылочной области левого полушария регистрируются полиморфные медленные колебания. ГВ не влияет на локальные изменения ЭА.

БИЛАТЕРАЛЬНО-СИНХРОННЫЕ ИЗМЕНЕНИЯ ЭА ГЛУБИННОГО ГЕНЕЗА:

В затылочных и теменных отделах регистрируются билатерально-синхронные отдельные острые волны тета-диапазона. ГВ генерализует изменения ЭА. Восстановление исходной фоновой активности происходит через 20 сек.

В лобных, центральных и передневисочных отделах регистрируются веретенообразные группы волн альфа-диапазона. ГВ не влияет на изменения ЭА.

ЗАКЛЮЧЕНИЕ:

Данные ЭЭГ указывают:

на умеренные изменения ЭА по общемозговому типу;

на негрубые резидуальные изменения ЭА в корковых отделах затылочной области левого полушария;

на выраженные функциональные изменения ЭА нижнестволового генеза; негрубые функциональные изменения ЭА лимбического происхождения.

3. Материал.

В работе были использованы данные обследования детей с диагнозом алалия в Детском отделении Центра патологии речи и нейрореабилитации ДЗ Москвы.

Были рассмотрены клинические данные 24 детей в возрасте от 3 до 5,5 лет, мальчики составили 75 % от общей группы.

Поскольку к основным причинам этого речевого синдрома в первую очередь относят перинатальные факторы риска, приведем отмеченные нами симптомы, характеризующие внутриутробные и родовые осложнения.

В анамнезе детей **с алалиями** отмечалась выраженная антенатальная патология. Были отмечены все три вида внутриутробной гипоксии (гемической, циркуляторной и тканевой): гестоз (отеки – 90% случаев, повышение давления – 80%), анемия (80%), генитальная инфекция (80%) с угрозой прерывания беременности (90%) а также клинически значимый токсикоз первой половины беременности (75% случаев). Такое массивное гипоксическое поражение ЦНС, могло привести к диффузной дефицитарности коры головного мозга плода, подкорковых и регуляторных стволовых структур. К тяжелому внутриутробному повреждению добавились травмы во время родов – 56,5% случаев (внутричерепные и внутрижелудочковые кровоизлияния) и инвазивные (в силу тяжести состояния новорожденного) реанимационные мероприятия у 70% детей, связанные в большинстве своем с антенатальной патологией. Следует отметить, что именно в этой группе практически у всех матерей (95% случаев) была проведена стимуляция в родах из-за слабости или дискоординации родовой деятельности.

Отметим, что среди осложнений преобладали антенатальные, преимущественно в первой половине беременности.

В итоге – ведущими перинатальными механизмами в структуре патогенеза алалии можно считать тяжелую сочетанную внутриутробную гипоксию плода (гемическую, циркуляторную и тканевую), приводящую к поражению всех отделов ЦНС: коры головного мозга, подкорковых и регуляторных структур.

Среди клинических данных наиболее информативными оказались ЭЭГ-показатели. ЭЭГ регистрировали от 12 отведений (O1O2; P3P4; C3C4; F3F4; T3T4; T5T6 по системе 10-20) с использованием биполярного и монополярного монтажей на 8-канальном электроэнцефалографе NEUROFAX (Nihon Kohen).

Для анализа были отобраны ЭЭГ - феномены, которые встречались более чем в 50% случаях. Эти показатели мы считали характерными для алалии. Приведем для сравнения распределение этих же данных в группе детей с задержками речевого развития (ЗРР), наиболее близких по характеру речевых нарушений к алалиям.

Таблица.

Распределение характерных для алалии ЭЭГ-показателей в группе детей с алалией и с ЗРР.

речевые синдромы ЭЭГ-показатели	Алалии в %	ЗРР в %
1. Локальные изменения ЭА в затылочных отделах левого полушария.	62	13
2. Снижение амплитуды альфа - ритма в левом полушарии.	62	27
3. п.1 + п.2 + (п.1 и п.2)	75	27
4. каудальные изменения ЭА (нижнестволовые)	75	10

Частота встречаемости ЭЭГ – признаков, характерных для алалии достоверно отличается от частоты тех же признаков в группе детей с ЗРР ($p<0,01$).

4. Обсуждение результатов.

Напомним, что основным ритмом ЭЭГ человека является альфа-ритм. Этот термин в случае визуального анализа ЭЭГ применяется только к колебаниям, доминирующим в затылочных областях и регистрируемым при закрытых глазах.

Характеристики альфа-ритма (частота, амплитуда и форма) и его реактивность меняются с возрастом. Комплексный анализ этих

показателей позволяет оценить соответствие развития электрической активности коры головного мозга возрастным нормативам, а также ее функциональное состояние.

Известно, что в генерации альфа-ритма участвуют специфические ядра таламуса и корковые нейронные цепи. Таламус представляет собой одно из основных подкорковых звеньев моторно-сенсорной системы и межцентральной регуляции электрических процессов мозга.

В нашей работе, проведенной в НИИ Нейрохирургии им. Н.Н. Бурденко было показано, что поражение в одном из полушарий головного мозга задних отделов таламуса приводит к угнетению альфа-ритма и наличию локальной патологической ЭА в затылочных отделах этого полушария [4].

Приведенные в таблице данные (п.1, 2 и 3) указывают, что у детей с алалиями можно предположить дисфункцию задних отделов таламуса в левом полушарии мозга.

Наши данные о внутриутробных осложнениях плода в первой половине беременности подтверждают это предположение. Из физиологии ЦНС развивающегося организма известно, что подкорковые структуры развиваются с 3 – 5 месяцев, кора головного мозга появляется лишь с 6-ого месяца.

Таким образом, можно считать, что в основе речевого синдрома алалии, в первую очередь, лежит поражение таламуса в левом полушарии.

Поскольку таламус является важным подкорковым отделом афферентных систем; кроме того, корковые отделы всех анализаторов имеют обратные связи с таламусом, то поражение таламуса в левом полушарии приводит к неполноценному развитию левого речевого полушария в целом и, как следствие, к недоразвитию всех компонентов речи.

Возникает вопрос, каковы последствия правополушарных поражений таламуса, которые безусловно, имели место наравне с левополушарными.

По нашим данным оказалось, что у 80% детей с ЗРР и очаговыми нарушениями ЭА в правом полушарии локальные изменения на ЭЭГ отмечены в задних отделах и у 55% детей имелось снижение альфа-ритма в правом полушарии [5].

Исследование высших психических функций в этой группе детей с ЗРР показало, что нарушение гностических функций, локализованных в задних отделах правого полушария и, в первую очередь, пространственной организации восприятия и движений приводит к задержке развития слухоречевой и номинативной памяти.

5. Выводы.

1. Структурная организация ЭЭГ - информации позволила провести сравнительный анализ результатов, полученных в группах детей с разными речевыми синдромами и в группе взрослых с поражением таламуса.
2. Поражение таламуса в левом полушарии мозга в онтогенезе можно считать одним из патогенетических механизмов развития алалии.
3. Аналогичное поражение в правом полушарии может приводить к ЗРР, в основном к нарушению слухоречевой и номинативной памяти.

6. Литература.

[1] И.П. Лукашевич, Р.И. Мачинская, Г.Е. Руденская, М.Н. Фишман, А.М. Эльнер, «Структурная организация медицинской информации для установления диагноза в детской неврологии», Медицинская техника, 1995, N 2, с.3-9.

[2] И.П. Лукашевич, Р.И. Мачинская, М.Н. Фишман, “Автоматизированная диагностическая система “ЭЭГ-ЭКСПЕРТ"
Медицинская техника, 1999, № 6, с.29-34.

[3] И.П. Лукашевич «Роль языка описания при комплексном анализе сложных систем» “Информационные процессы”, 2002 г., т. 2, №1, с. 37-44. http://www.jip.ru/

[4] И.П. Лукашевич, О.Б. Сазонова,
"Влияние поражения различных отделов зрительного бугра на характер биоэлектрической активности мозга человека".
Журнал высш. нервн. деят. им. ак. И.П.Павлова, 1996, т.46, N 5, с. 866-874.

[5] И.П. Лукашевич, Р.И. Мачинская, В.М. Шкловский,
«Роль структур правого и левого полушарий головного мозга в формировании речи и памяти у детей». Физиология человека. 2002 г., т.28, № 6, с.52-56.

Использование записей движений глаз для анализа механизмов зрительной оценки размеров на примере иллюзии Мюллера-Лайера

В.В. Огнивов, П.В. Максимов, Г.И. Рожкова
Институт проблем передачи информации РАН
127994 Москва, Б. Каретный пер., 19.
wwoo@yandex.ru

Аннотация

Цель работы – выяснение механизмов оценки длины в случае геометрических зрительных иллюзий. Метод - запись движений глаз в процессе сравнения длины отрезков в отсутствие дополнительных линий и в присутствии дистракторов, порождающих иллюзию Мюллера-Лайера. Регистрацию движений глаз производили при помощи системы, iView X™ Hi-Speed 1250 (SMI). В условиях бинокулярного наблюдения было протестировано 15 взрослых испытуемых от 18 до 60 лет. Зарегистрированные движения глаз трудно назвать «метрическими» или отражающими расположение информативных точек. Анализ записей позволяет отвергнуть любые гипотезы, связанные с непосредственным использованием глазодвигательной информации для оценки размеров в условиях наших экспериментов.

1. Введение

Систематическим изучением геометрических зрительных иллюзий исследователи занимаются более ста лет [1 - 12 и др.]. Долгое время методы исследования таких иллюзий базировались на использовании печатных образцов, что лимитировало возможности варьирования тестовых изображений. С появлением компьютеров появилась возможность генерировать стимулы на дисплеях и широко варьировать их параметры.

Одной из самых сильных геометрических иллюзий является иллюзия Мюллера-Лайера, которая исследуется с тех пор, как её придумал немецкий психиатр Франц Мюллер-Лайер в 1889 году и до самого последнего времени. Конфигурация, предложенная Мюллером-Лайером, представлена на **Рис. 1, а.** В этом изображении имеется два равных горизонтальных отрезка, но из-за наличия наклонных отрезков-дистракторов (называемых крыльями, усиками, стрелками и т. п.) наблюдателю кажется, что правый отрезок существенно меньше левого.

Наряду с такой конфигурацией, часто используется изображение, в котором два отрезка сведены вместе (Рис. 1, б). Такая версия была введена Брентано, в связи с чем соответствующую иллюзию часто называют иллюзией Брентано [13, 14]. Кроме того, в научных экспериментах используют также специальные модификации, различающиеся формой дистракторов (в, е, ж, з) или лишенные определенных деталей (г, д).

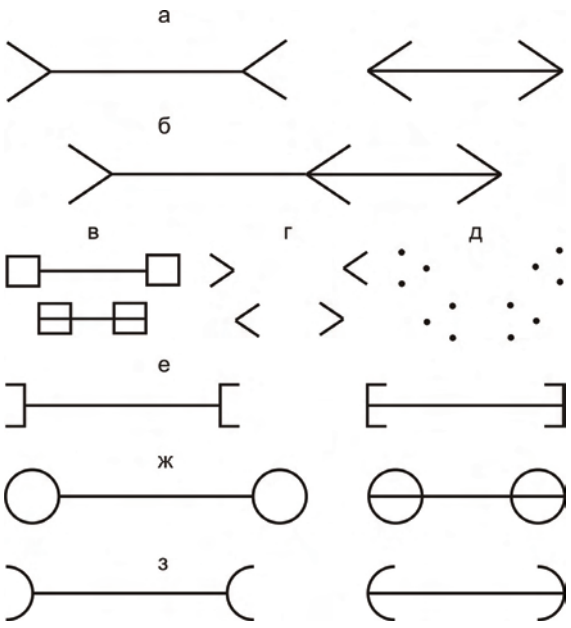


Рис. 1.
Исходная фигура Мюллера-Лайера (а), версия Брентано (б) и несколько модификаций, использованных в экспериментах разных авторов (в - з).

Для объяснения иллюзии Мюллера-Лайера предложено более десятка гипотез (теорий, моделей). Напомним некоторые из них.

- Влияние несовершенства оптики: при размывании изображений появляются локальные максимумы зачерненности, которые и выбираются в качестве точек отсчета. У острых углов эти максимумы находятся внутри, что приводит к кажущемуся укорочению/удлинению отрезков, примыкающих к вершине угла изнутри/снаружи.
- Взаимодействие контуров: тормозное влияние соседних линий характерным образом искажает представление фигур на нейронном уровне.
- Теория смещения оценок: на длину каждого отрезка влияет общий размер той части фигуры, в которую он входит.
- Теория неадекватного шкалирования глубины, ведущая начало от Тьерри и развиваемая в наше время Грегори: искажение длины отрезков происходит под влиянием механизмов константности пространственного восприятия.
- Влияние статистики зрительных сцен окружающего мира, т.е. физических источников соответствующих сетчаточных изображений: идентичные отрезки в тесте Мюллера-Лайера кажутся неравными, потому что различны распределения вероятности наблюдения их аналогов в реальном мире.
- Модель оценки размеров с использованием функции информативности: длины отрезков фактически измеряются по расстояниям между экстремумами функции информативности, которые локализуются внутри острых углов. (Для примера на **Рис. 2** показано положение минимумов и максимумов функции информативности на изучавшихся стимулах)
- Глазодвигательная теория: точками отсчета при оценке длины являются точки фиксации взора на концах отрезков, которые в реальности сдвинуты от концевых точек внутрь острых углов, поэтому амплитуды скачков глаз между концами двух горизонтальных отрезков в конфигурации Мюллера-Лайера оказываются различающимися **(Рис. 3).**

Каждая из предложенных теорий может объяснить существование иллюзии Мюллера-Лайера в принципе, но либо не согласуется с реальной величиной иллюзии, либо не выдерживает критики в определенных специальных условиях. Например, так называемая «глазодвигательная теория» в свое время была отвергнута на том основании, что иллюзия сохраняется при кратковременном предъявлении изображений, исключающем

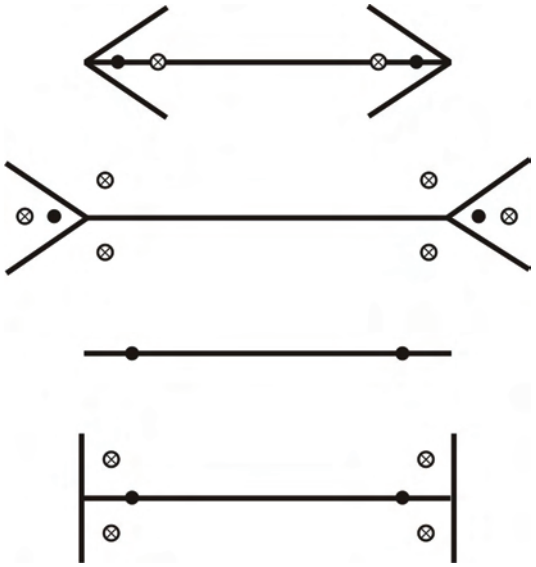


Рис. 2. Расположение минимумов (черные точки и максимумов (крестики) функции информативности на двух частях фигуры Мюллера-Лайера и на контрольных отрезках (по [15], с. 146])

возможность перевода взора, а также в условиях стабилизации сетчаточных изображений.

Наиболее глубоко проработанная теория оценки расстояний/размеров на основе экстремумов функции информативности [15] не согласуется с картами реального расположения точек фиксации внимания, полученными экспериментально.

Поскольку у каждой теории были достаточно веские основания, мы предположили, что иллюзия Мюллера-Лайера объясняется не одним механизмом, а несколькими, причем разные механизмы локализованы на разных уровнях обработки зрительной информации **(Рис. 3)** и

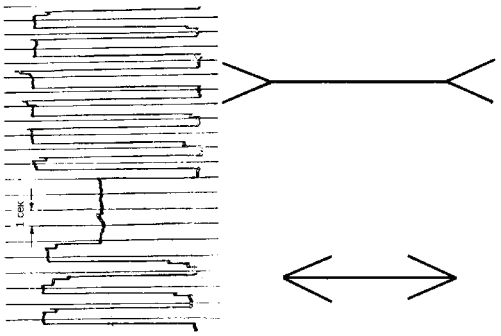


Рис. 3. Запись движения глаз при наблюдении фигуры Мюллера-Лайера с инструкцией переводить взгляд между концами отрезков (по [16])

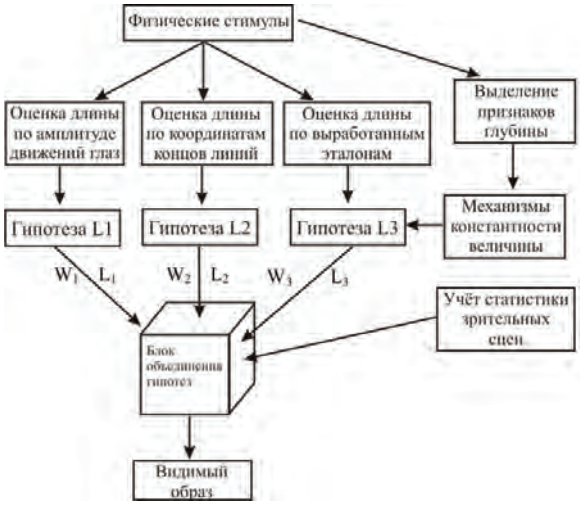


Рис. 4. Схема множественности механизмов иллюзии Мюллера-Лайера

не всегда функционируют одновременно. Обработка информации о размерах осуществляется разными способами в нескольких параллельных путях, а затем результаты частных оценок объединяются. Суммарный эффект, конечно, не равен простой сумме вкладов каждого механизма, а определяется правилами кооперации и конкуренции механизмов.

По результатам многих экспериментов и дискуссий, относительно движений глаз к настоящему времени сложилось мнение, что они не являются средством оценки расстояний, но отражают процесс перемещения внимания по рассматриваемому объекту. Понятие фиксации внимания было введено Вундтом – одним из основоположников теории зависимости восприятия от движений глаз. По-видимому, точки фиксации внимания не всегда совпадают с точками фиксации взора. Правда, в определенных ситуациях наблюдается высокая корреляция между картами этих точек, но есть и ситуации, когда корреляция близка к нулю. Например, очевидно, что при анализе стабилизированных изображений движения глаз не могут отражать движение взора по рассматриваемой картине, (поскольку сетчаточное изображение движется вместе с глазом, точка фиксации на нем неподвижна), но могут отражать перемещение внимания («внутреннего взора»).

В настоящей работе сделана попытка исследовать процессы сравнения длины двух отрезков без дистракторов (глазомер) и с дистракторами (иллюзия Мюллера-Лайера), анализируя движения глаз. Учитывая выводы наших прежних исследований, приведших к схеме **Рис. 4,** мы ожидали, что при сходных конфигурациях тест-объектов характер движений глаз будет сильно зависеть от сложности задачи (т.е. от различия в длине отрезков), так как задачи разного уровня сложности будут решаться разными механизмами.

2. Методика

2.1. Аппаратура для записи и анализа движений глаз

Запись движений глаз производили на установке iView X™ Hi-Speed 1250 (SensoMotoric Instruments (SMI) Ltd.). Это система видеослежения за темным зрачком, использующая инфракрасное освещение и компьютерную обработку изображений в реальном масштабе времени. В процессе обработки вычисляются координаты центра зрачка, которые после калибровки переводятся в данные о направлении взора. Кроме того, регистрирующая система отслеживает роговичный рефлекс, чтобы скомпенсировать изменения позиций видеокамеры по отношению к голове. На **рис. 5** показаны метки в центре зрачков (белые кресты) и на роговичных бликах (черные кресты), по которым определяется направление взора.

Пространственное разрешение установки в лучших условиях может доходить до одной угловой минуты, временное разрешение при бинокулярной регистрации соответствует частоте 500 Гц.

Общий вид экспериментальной установки с испытуемым показан на **рис. 6.** Голова испытуемого фиксировалась с помощью подбородника и лобного упора. Система видеослежения находилась над головой, и световые лучи от глаз попадали в объектив,

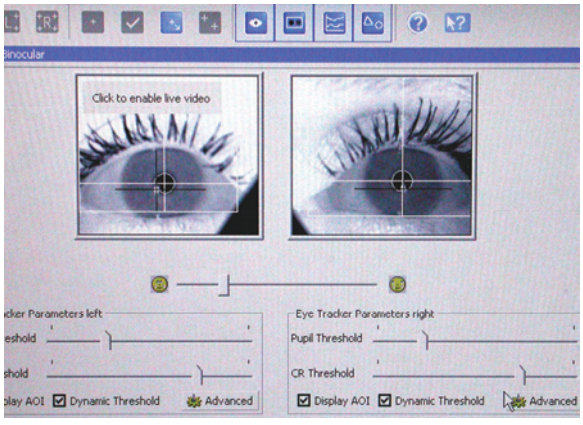


Рис. 5. Вид части дисплея, соединенного с компьютером, производящим обработку видеосигналов. На двух кадрах представлены изображения левого и правого глаз испытуемого в некоторый момент рассматривания стимула

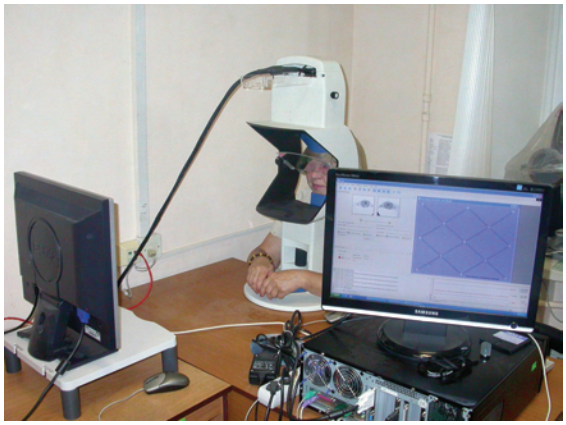


Рис. 6. Общий вид экспериментальной установки с испытуемым и двумя дисплеями – для стимуляции и контроля записи движений глаз.

отражались от наклонного полупрозрачного зеркала. На каждом испытуемом перед записью движений глаз производили калибровку прибора по 13 точкам в различных участках стимульного экрана.

При анализе записей движений глаз выделяли фиксации (фазы блуждания взора на малом участке изображения) и саккады (быстрые скачки с амплитудой более 0.2° и скоростью более 30°/сек от одной точки фиксации к другой).

2.2 Тестовый материал

Испытуемому последовательно предъявляли 10 изображений пары отрезков с различным отношением длин, причем в одной серии опытов отрезки были без дистракторов (для оценки глазомера), а в другой – с дистракторами (для оценки иллюзии). Различия в длине отрезков на 10 тестовых изображениях представлены в Таблице 1. Абсолютные значения длин отрезков варьировали в диапазоне 78-128 мм, а общая длина фигуры практически не менялась. В стимуле № 1 правый стимул был на 8% меньше левого, в стимуле № 2 отрезки были равны, а в стимуле № 10 длина правого отрезка более чем в полтора раза (на 64%) превосходила длину левого.

Таблица 1. Различия (в %) в длине двух горизонтальных отрезков для 10 использованных тестовых изображений

Номер образца	1	2	3	4	5	6	7	8	9	10
Разность %	-8	0	8	16	24	32	40	48	56	64

Стимулы предъявлялись на мониторе BenQ FP91G 19” с разрешением 1280 x 1024 пикселей, который располагался на расстоянии 1 м от глаз испытуемого. Угловые размеры отрезков составляли при этом 4,5 – 7,5^0.

2.3 Процедура

После инструктажа и калибровки испытуемому последовательно предъявляли серию тестовых изображений с парами отрезков. Испытуемый должен был определить, какой из отрезков длиннее и слегка приподнять указательный палец левой/правой руки, если длиннее кажется левый/правый отрезок, или оба пальца, если отрезки кажутся равными. Запись движений глаз включали в момент предъявления стимула и прекращали, как только испытуемый принимал решение. Стимулы предъявляли один за другим в псевдослучайном порядке. Процедуру калибровки прибора выполняли один раз в начале каждой тестовой серии.

2.4 Испытуемые

Было протестировано – 15 испытуемых - лет (из них – 4 мужского пола). Испытуемыми были сотрудники ИППИ в возрасте от 18 до 60 лет, как не имевшие опыта участия в подобных экспериментах, так и опытные. Перед началом эксперимента у всех испытуемых проверялась острота зрения.

3. Результаты и обсуждение

Типичные записи движений глаз при наблюдении тестовых изображений с дистракторами и без них, а также результаты обработки записей представлены на рис. 7-11.

Рис. 7 демонстрирует общий вид записей, получаемых для смещений взора по двум координатам (X и У).

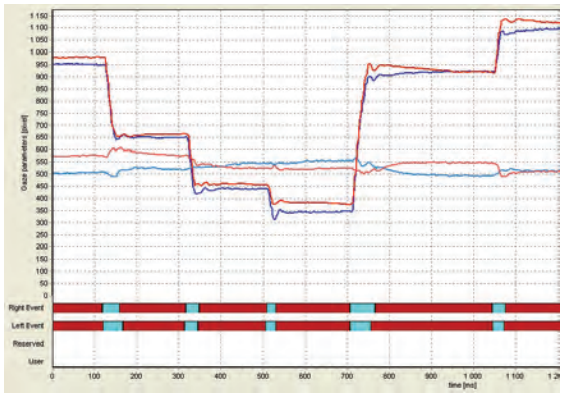


Рис.7: Общий вид записей (по оси абсцисс – время, по оси ординат – смещения левого и правого глаз по двум координатам

В данном представлении по оси абсцисс

откладывается время, а по оси ординат – позиции взора на экране монитора по горизонтали и вертикали. Движение правого глаза представлено красными линиями, а левого – синими линиями. Более насыщенные цвета соответствуют смещению по горизонтали. На приведенной записи одного испытания хорошо выделяются фазы фиксаций (практического постоянства координат), которые соответствуют горизонтальным участкам кривых, и скачки глаз – саккады, которые соответствуют переходам от одного горизонтального участка к другому. На графике четко видна серия синхронных для двух глаз горизонтальных скачков (ступени на записях X-координаты обоих глаз) и практическое отсутствие смещений глаз по вертикали (изменение У-координат представляют две линии, идущие по центру почти на одном уровне в течение всего измерения)

На **Рис. 8** приведены карты внимания для набора тестовых изображений с дистракторами и без них, полученные для одного конкретного испытуемого. Плотность точек на этих картах отражает время фиксации взора в соответствующем месте. Как видно из приведенных карт, в случае когнитивной задачи (сравнить длину отрезков), в отличие от случая глазодвигательной задачи (переводить взор между концами отрезков) не наблюдается концентрации точек фиксации у концов сравниваемых отрезков (сравни рис. 8 с рис. 3).

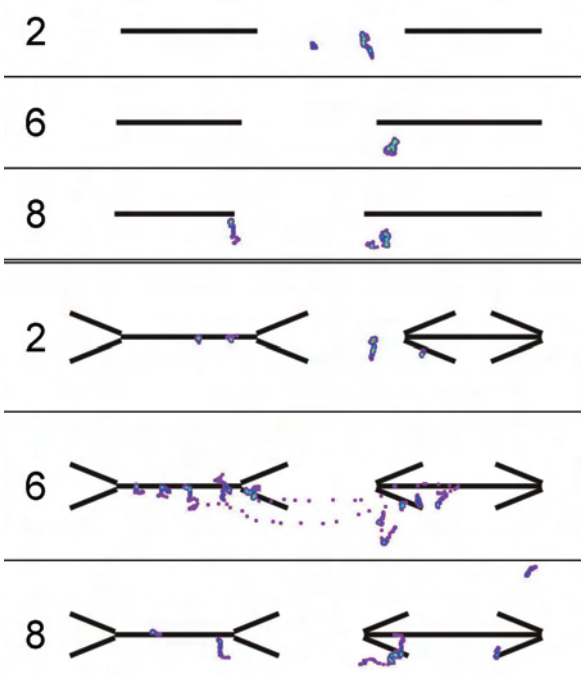


Рис.8: Карты внимания для стимулов 2, 5 и 8 в сериях «глазомер (вверху) и «иллюзия» (внизу). Точки фиксации наложены на соответствующие стимулы.

Заметим, что это имеет место как для отрезков с дистракторами, так и без дистракторов.

Следует специально подчеркнуть, что в простых случаях (при большой разнице в длине отрезков) для принятия решения хватает нескольких фиксаций, и не требуется никаких специфических «метрических» движений глаз (по терминологии Гиппенрейтер [17, 18]).

Можно думать, что в этих случаях процедура оценки длины ближе всего соответствует теории центрации Пиаже, предполагающей наличие у человека способности одномоментной оценки расстояний и размеров.

На **Рис. 9-11** приведены данные по зависимости числа и длительности фиксаций, а также общего времени принятия решения от типа стимула (с дистракторами, без дистракторов) и трудности задачи (различия в длине отрезков: номера стимула). Графики показывают, что для числа фиксаций и времени принятия решения имеются отчётливые максимумы, соответствующие субъективной близости сравниваемых отрезков по длине. Для отрезков без дистракторов (глазомер) – это стимулы 1, 2, 3, а для отрезков с дистракторами – стимулы 5, 6, 7.

Изменения паттернов на картах внимания Рис.8 и характер кривых Рис. 9-11 позволяют предположить, что по мере усложнения когнитивной задачи происходит смена тактики осмотра изображений и алгоритмов вычисления длины или смена доминирования параллельных механизмов в гипотетической схеме 3.

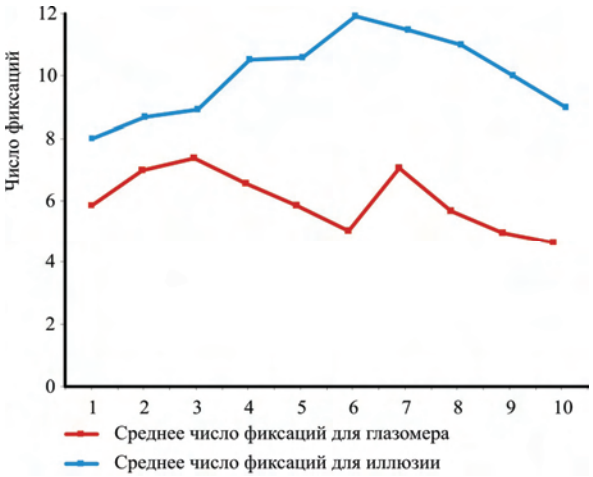


Рис.9: Среднее число фиксаций при сравнении отрезков в сериях «иллюзия» (15 испытуемых) и «глазомер» (10 испытуемых). Стимул 2 соответствует равным отрезкам, далее с увеличением номера различие в длине отрезков увеличивается шагами в 8%.

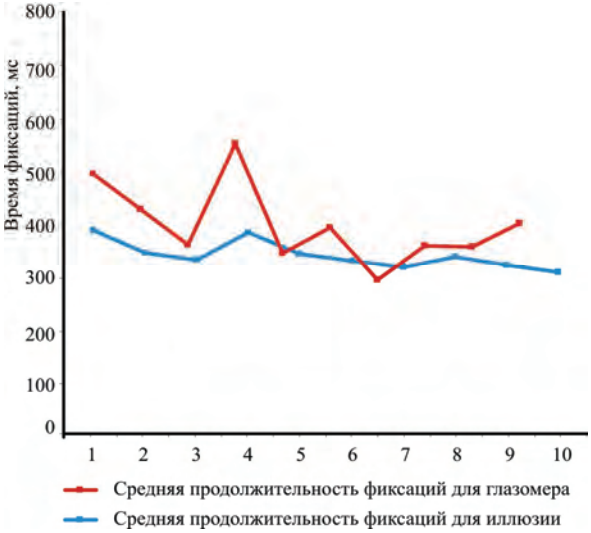


Рис.10: Средняя длительность фиксаций при сравнении отрезков в сериях «иллюзия» (15 испытуемых) и «глазомер» (10 испытуемых). Стимул 2 соответствует равным отрезкам, далее с увеличением номера различие в длине отрезков увеличивается шагами в 8%.

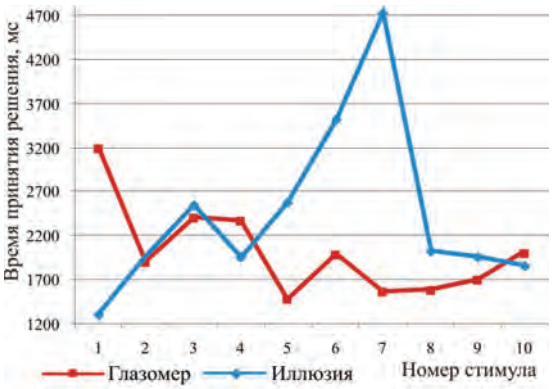


Рис.11. Среднее время принятия решения в сериях «иллюзия» (15 испытуемых) и «глазомер» (10 испытуемых). Стимул 2 соответствует равным отрезкам, далее с увеличением номера различие в длине отрезков увеличивается шагами в 8%.

5. Литература

[1] Грегори Р. Л. Иллюзии / Сб. Психология ощущений и восприятия / Под ред. Ю.Б. Гиппенрейтер. М.: ЧеРо. 2002. - С. 174-185.

[2] Dewar R.E. Stimulus determinants of the magnitude of the Mueller-Lyer illusion // *Percept. Mot. Skills.* 1967. - V.24. - P.708–710.

[3] Eysenck H. J., Slater, P. Effects of practice and rest on fluctuations in the Müller-Lyer illusion // *Br. J. Psychol.* 1958. - V. 49. - P. 246-255.

[4] Gregory R.L. The grammar of vision // Listener. 1971. - V.83. - P.242.

[5] Judd C.H. A Study of Geometrical Illusions. // *Psychl. Rev.* 1899. - V. 6. - P. 241.

[6] Judd C.H. The Müller-Lyer illusion // *Psychol. Rev. Monogr. Suppl.* 1905. - V.7. - P.55-81.

[7] Lewis E.O. Confluxion and contrast effects in the Müller-Lyer illusion. // *Br. J. Psychol.* 1909. - V.3. - P.21-41.

[8] Pollack, R. H., Chaplin, M. R. Effects of prolonged stimulation by components of the Müller-Lyer figure upon the magnitude of illusion // *Percept. Mot. Skills.* 1964. - V. 18. - P. 377-382.

[9] Predebon J. A comparison of length-matching and length-fractionation measures of Müller-Lyer distortions // *Percept. Psychophys.* 2005. - V.67. - Iss.2. - P.264-273.

[10] Predebon J. Decrement of the Brentano Müller-Lyer illusion as a function of inspection time // *Percept. Psychophys.* 1998. - V.27. - P.183-192.

[11] Predebon J. Length Illusions in Conventional and Single-Wing Müller-Lyer Stimuli // *Percept. Psychophys.* 2000. - V.62. - P.1086-1098.

[12] Predebon J. Selective attention and asymmetry in the Müller-Lyer illusion // *Psychonom. Bull. Review.* 2004. - V.11. - P.916-920.

[13] de Grave, D. D. J., Franz, V. H., Gegenfurtner, K. R. “The influence of the Brentano illusion on eye and hand movements” *Journal of Vision*, 2006. V. 6, pp. 727–738.

[14] de Grave, D. D. J., Smeets, J, B. J., Brenner, E. “Why are saccades influenced by the Brentano illusion?” *Exp. Brain Res.*, 2006. V. 175, pp. 177–182.

[15] Завалишин Н.В., Мучник И.Б. Модели зрительного восприятия и алгоритмы анализа изображений. М., Наука.1974г. 344с.

[16] Ярбус А Л. *Роль движения глаз в процессе зрения*. М.: Наука, 1965. 166 с.

[17] Гиппенрейтер Ю. Б. Опыт экспериментального исследования работы зрительной системы наблюдателя. / В сб.: Инженерная психология. М.: МГУ. 1964.

[18] Гиппенрейтер Ю.Б. *Движения человеческого глаза*. М.: Изд-во МГУ, 1978. 256 с.
chophysics, 1985. V. 37, pp. 335–344.

Движения глаз при восприятии неоднозначных изображений типа куба Неккера

М.Ю. Желтова, П.В. Максимов, В.А. Бастаков, А.К. Панютин
Институт проблем передачи информации РАН
127994 Москва, Б. Каретный пер., 19.
bastakov@iitp.ru

Аннотация

Цель работы – оценить корреляцию между движениями глаз и динамикой восприятия неоднозначных фигур. В качестве неоднозначной фигуры использован куб Неккера, при длительном наблюдении которого происходит чередование двух возможных интерпретаций контурного рисунка – как бы обращение куба. Регистрацию движений глаз производили при помощи установки для записи движения глаз, iView X™ Hi-Speed 1250 (SMI). В условиях бинокулярного наблюдения было протестировано 15 взрослых испытуемых от 18 до 60 лет. Построенные по результатам каждого эксперимента карты плотности фиксации взора в разных точках изображения демонстрируют большую индивидуальную вариабельность. Анализ этих карт и исходных треков позволяет отвергнуть любые гипотезы, связанные с непосредственным использованием характера движений глаз для детекции моментов обращения фигуры в условиях наших экспериментов..

1. Введение

Для понимания механизмов формирования видимых образов особый интерес представляют неоднозначные изображения, которые допускают две или более интерпретаций в смысле взаимного расположения по глубине вершин и граней (трансформирующиеся или обращающиеся фигуры) либо в смысле отнесения определенных фрагментов к объекту и фону.

Одной из самых простых и самых исследованных изображений такого рода является куб Неккера – контурное изображение, соответствующее параллельной проекции вершин и ребер куба на плоскость. Такое изображение и две возможные пространственные интерпретации приведены на **рис. 1.**

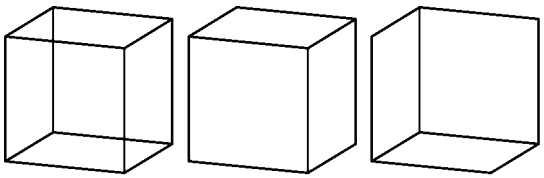


Рис. 1. Куб Неккера (а) и его возможные пространственные интерпретации (б, в).

Динамика трансформаций куба Неккера изучалась многими исследователями, причем во многих работах – в аспекте связи обращений фигуры с движениями глаз. Одни авторы придерживаются мнения, что обращениям куба способствует смена точек фиксации. Такая точка зрения высказывалась еще в работах начала века. В качестве альтернативы был выдвинут тезис, что обращение происходит в результате какого-то центрального процесса в сенсорной зрительной системе, а движения глаз являются следствием этого процесса и отражают его особенности (Гиппенрейтер, 1964; 1971).

Основной способ проверки этих двух гипотез состоял в подсчете числа случаев, когда обращение фигуры следовало за скачком и когда оно ему предшествовало. К началу 70-х гг, по итогам многих экспериментов, большинство исследователей склонилось к выводу, что обращения куба определяются центральными процессами. Затем группой отечественных психофизиологов были проведены более детальные эксперименты, которые показали, что результаты некоторых серий экспериментов могут лучше соответствовать первой гипотезе (Ярбус, 1965; Ефимов, Лебедев, 1976).

В экспериментах Гиппенрейтер с коллегами, кроме оценки корреляции обращений куба со скачками глаз, строили также карты плотности фиксаций взора, предполагая обнаружить максимумы плотности, соответствующие двум возможным интерпретациям изображения

(Гиппенрейтер, 1971). Такие максимумы как будто бы были получены, но число испытуемых было слишком мало, чтобы можно было верить этим выводам. Целью работы проверка предположений о положительной корреляции обращений куба со скачками глаз.

2. Методика

2.1. Аппаратура для записи и анализа движений глаз

Запись движений глаз производили на установке iView X™ Hi-Speed 1250 (SensoMotoric Instruments (SMI) Ltd.). Это система видеослежения за темным зрачком, использующая инфракрасное освещение и компьютерную обработку изображений в реальном масштабе времени. В процессе обработки вычисляются координаты центра зрачка, которые после калибровки переводятся в данные о направлении взора. Кроме того, регистрирующая система отслеживает роговичный рефлекс, чтобы скомпенсировать изменения позиций видеокамеры по отношению к голове. На **рис. 2** показаны метки в центре зрачка (белый крест) и на роговичном блике (черный крест), по которым определяется направление взора. Пространственное разрешение установки в лучших условиях может доходить до одной угловой минуты, временное разрешение при бинокулярной регистрации соответствует частоте 500 Гц.

Общий вид экспериментальной установки с испытуемым показан на **рис. 3.** Голова испытуемого опиралась на подбородник и фиксировалась с помощью лобного упора. Система видеослежения находилась над головой, а световые лучи от глаз попадали туда, отражаясь от наклонного полупрозрачного зеркала.

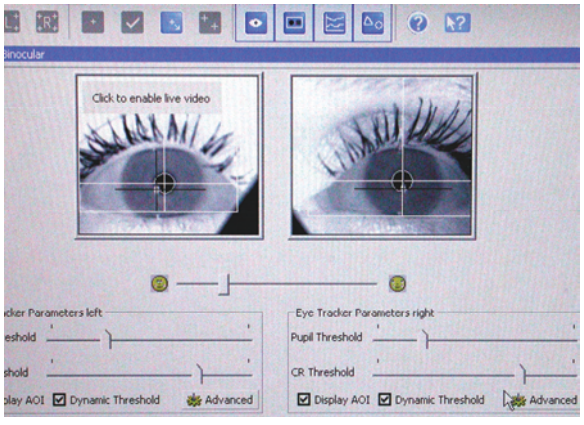


Рис. 2. Фрагмент интерфейса программы, работающей в составе регистрирующей системы

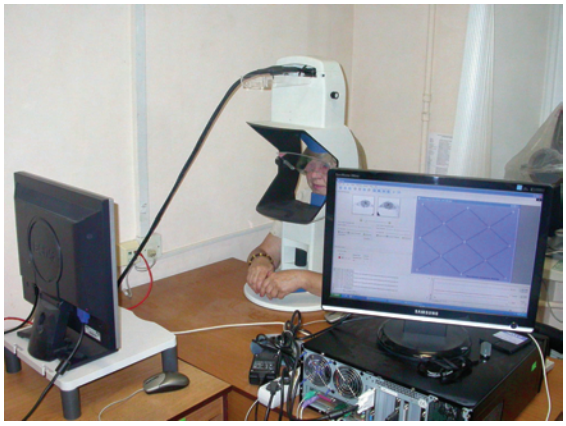


Рис. 3. Общий вид экспериментальной установки с испытуемым

На каждом испытуемом перед записью движений глаз производили калибровку прибора по 13 точкам в различных участках экрана.

При анализе записей движений глаз выделяли фиксации (фазы блуждания взора на малом участке изображения) и саккады (быстрые скачки с амплитудой более 0.2° и скоростью более 30°/сек от одной точки фиксации к другой).

2.2 Тестовый материал

Тестовое изображение представляло собой классический куб Неккера с длиной ребра 10 см. Общая протяженность изображения по горизонтали и вертикали составляла 15 см. Расстояние до экрана монитора равнялось 1 м, так что угловые размеры стимула укладывались в 30 градусов. Стимулы предъявляли на мониторе BenQ FP91G 19” с разрешением 1280×1024 пикселей.

2.3 Процедура

После инструктажа и калибровки испытуемому в течение 50 секунд предъявляли тестовое изображение. Как и в экспериментах Гиппенрейтер, результаты которых мы хотели воспроизвести, в двух сериях опытов использовали разные инструкции: (1) свободно рассматривать куб, отмечая легким движением указательных пальцев моменты обращений куба; (2) удерживать каждую интерпретацию как можно дольше с помощью ментальных усилий, стараясь минимизировать число обращений куба.

2.4 Испытуемые

Было протестировано – 15 испытуемых. Возраст испытуемых от 25 до 59 лет (из них 3 - мужского пола).

3. Результаты и обсуждение

Типичные карты фиксаций взора приведены на рис. 4-6. Плотность точек на этих картах отражает время фиксации взора в соответствующем месте.

Рис. 4 демонстрирует большую индивидуальную вариабельность карт, не позволяющую подтвердить вывод Гиппенрейтер с соавторами о наличии двух центров концентрации взора при рассматривании куба Неккера.

Рис. 5 убедительно показывает, что **характер инструкции** практически не влиял на карты внимания. Как видно из сравнения двух столбцов данных, относящихся к разным инструкциям, явных различий между ними не наблюдается.

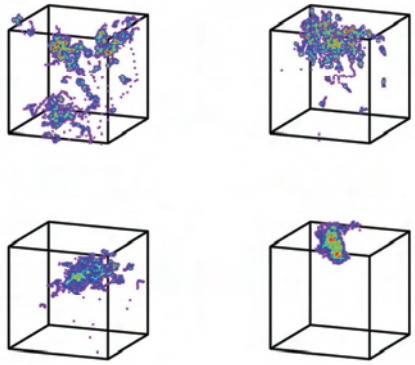


Рис. 4. Вариабельность карт внимания

В результате двух серий экспериментов убедительно показано, что любые гипотезы, связанные с непосредственным использованием характера движений глаз для детекции моментов обращения фигуры в условиях наших экспериментов не состоятельны. Инверсия изображения в случае неоднозначных фигур подчинено, по всей видимости, нескольким механизмам зрительного восприятия и связано с оценкой глубины зрительного пространства и формированием объемного изображения. Роль движений глаз, по-видимому, не является ведущей, а их вклад в процесс формирования объемного изображения требует дальнейших исследований.

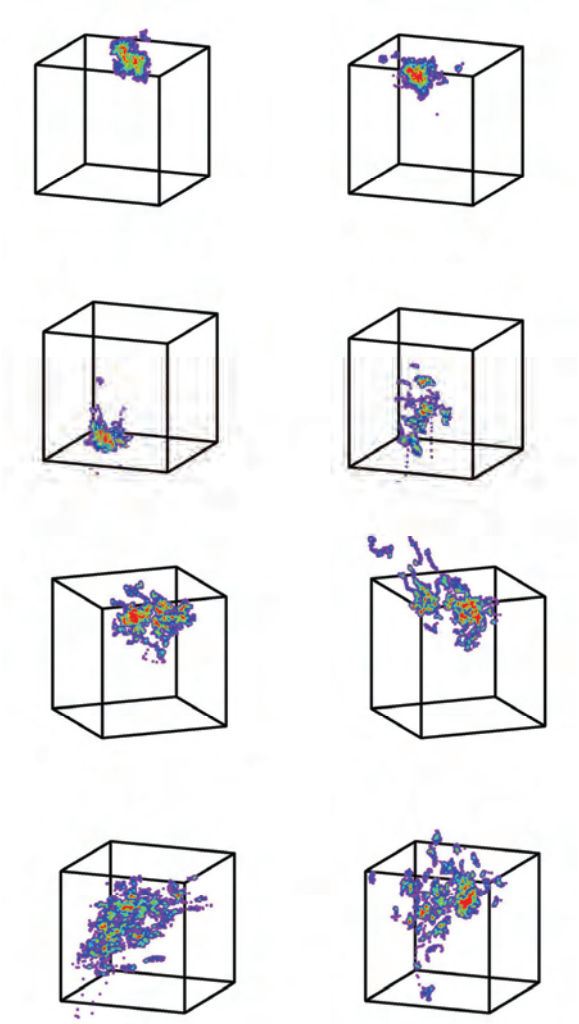


Рис. 5. Карты внимания при разных инструкциях, даваемых испытуемым, построенные для четырёх испытуемых. Левый столбец соответствует свободному рассматриванию, правый – удержанию интерпретации.

5. Литература

[1] Ярбус А.Л. *Роль движения глаз в процессе зрения*, М., 1965. С. 166.

[2] Ефимов С.К., Лебедев Д.Г. Фотокимограф для записи движений глаз // Вопр. психол. 1976. № 3. С. 152 - 153.

Компьютерная система для исследования лабильности зрительной системы

Грошев Ф.В.
ИППИ РАН
groshev@iitp.ru

Голубцов К.В.
ИППИ РАН
golub@iitp.ru

Трунов В.Г.
ИППИ РАН
trunov@iitp.ru

Айду Э.А.И.
ИППИ РАН
aidu@iitp.ru

Аннотация

В статье рассматриваются методы исследования функционального состояния зрительной системы человека. Описывается микропроцессорная система, позволяющая реализовать эти методы.

1. Введение

Психофизические методы исследования патологических процессов в зрительной системе, наряду с электрофизиологическими методами составляют основу современной офтальмологической диагностики. Основные психофизические методы включают исследование поля зрения (периметрия), исследование критической частоты слияния мельканий (КЧСМ), и исследование электрического фосфена (ЭФ). Разработана компьютерная система для исследования как КЧСМ, так и ЭФ. Объединение исследований, которые с разных сторон характеризуют функциональную подвижность или лабильность зрительной системы, сокращает время и повышает диагностическую эффективность обследования пациента. Разработанная система позволяет проводить стимулирующее терапевтическое воздействие на зрительную систему для лечения ряда заболеваний на ранних стадиях развития патологических процессов.

2. Описание методов диагностики

Под КЧСМ понимается минимальная частота прерывистого светового излучения, при которой глаз начинает воспринимать излучение как непрерывное. При малой частоте мельканий пациент видит серию вспышек, а с увеличением частоты вспышек появляется ощущение мерцания, вначале грубое, затем тонкое и, наконец, ощущение полной слитности предъявляемых стимулов. Средние показатели КЧСМ у здоровых взрослых и детей составляют 43-45 Гц на красный стимул, 45-47 Гц на зеленый стимул и 40-42 Гц на синий стимул. При таком заболевании, как глаукома, КЧСМ снижается до 30 Гц и ниже, в зависимости от степени заболевания. При исследовании ЭФ на глаз человека воздействуют импульсным током от 35 до 450 микроампер, в зависимости от патологии. Здоровый человек воспринимает ЭФ в виде очень слабых бесцветных или голубоватых вспышек. Минимальная сила тока, при которой человек воспринимает вспышку, определяется как порог электрической чувствительности сетчатки (ПЭЧ). Согласно экспериментальным и клиническим данным, ПЭЧ характеризует функциональное состояние внутренних слоев сетчатки, т.е. слоя ее ганглиозных клеток, и коррелирует с общей площадью патологических скотом. Чем больше площадь дефектов поля зрения, тем выше требуется ток, чтобы вызвать электрофосфен. Пороговая сила тока, при которой возникают едва заметные световые мелькания, зависит от частоты подаваемых импульсов тока. При плавном увеличении частоты стимуляции наступает момент, когда человек перестает ощущать ЭФ. Этот момент обозначается как критическая частота исчезновения мельканий ЭФ и является показателем функционального состояния “центрального” пучка зрительного нерва.

Эти два метода различаются по восприятию света и цвета. Так, если при КЧСМ выше 55 Гц человек видит равномерное свечение при любой фиксации взора, то при стимуляции ЭФ на частотах выше 50 Гц, восприятие света прекращается. Другое отличие заключается в том, что при исследовании КЧСМ человек видит свет того спектрального состава, который предъявляется, а при исследовании ПЭЧ только беловато–голубоватый.

Следует отметить, что у здоровых людей КЧСМ регистрируется выше 110 Гц на электроэнцефалограмме, а на высшем уровне зрительной системы имеется “фильтр”, который блокирует все частоты выше 60 Гц для

восприятия. Восприятие равномерного свечения при КЧСМ объясняется естественным тремором глаз, который, по данным А.Л. Ярбуса, полученным в опытах с присосками [5], составляет выше 90 Гц. В настоящее время исследованием КЧСМ и ЭФ уделяют большое внимание в клинике, особенно, при исследовании зрения у детей. Однако до сих пор эти исследования выполняют на разных приборах, иногда находящихся в разных кабинетах; многие из этих приборов без преувеличения можно назвать морально устаревшими. В разрабатываемой нами компьютерной системе обе методики могут быть реализованы в одном микропроцессорном устройстве, что сокращает время исследования и повышает его качество. Как отмечается в ряде работ, патологические изменения можно выявить при поражении более 20% зрительных волокон, а разработанная нами методика исследования позволяет выявлять патологические процессы на ранней стадии их развития.

3. Описание компьютерной системы

Разработанная компьютерная система позволяет исследовать КЧСМ, электрическую чувствительность и электрическую лабильность. Показатели КЧСМ и ПЭЧ отображаются на мониторе компьютера в наглядной форме и хранятся в базе данных.

Компьютерная система состоит из программной и аппаратной частей (см. рис.2). Программная часть работает под управлением персонального компьютера и реализует интерфейс управления аппаратной части и обработку и хранение в базе данных. Аппаратная часть осуществляет воздействие на зрительную систему человека с помощью световых и электрических стимулов. Может работать автономно.

Для исследования электрической чувствительности и лабильности зрительного нерва используется программно управляемый источник тока, подающий стимулирующие импульсы на электроды. Программа для управления источником тока создается или модифицируется на компьютере и загружается в память микроконтроллера.

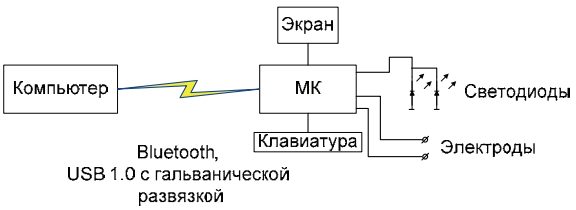


Рис. 1. Блок-схема компьютерной системы

Система имеет гальваническую развязку между компьютером и блоком стимуляции согласно ГОСТу на медицинские изделия. Питание аппаратной части осуществляется от аккумуляторов с напряжением 5В, что отвечает требованиям безопасности медицинских приборов.

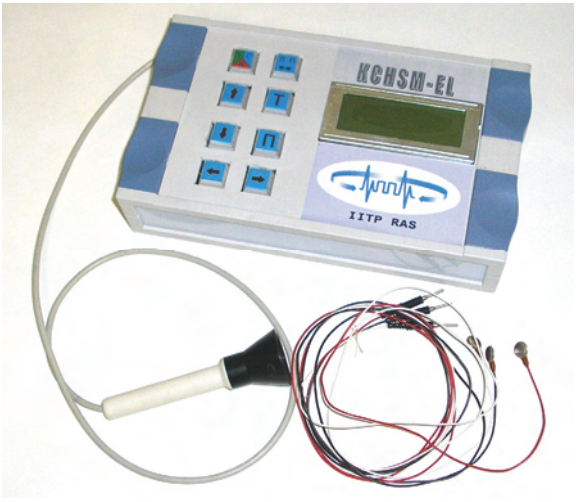


Рис.2. Внешний вид аппаратной части компьютерной системы

4. Литература

[1] Голубцов К.В., Орлов О.Ю., Софронов, "Новые средства диагностики в офтальмологии", П.Д. /От современной фундаментальной биологии к новым наукоемким технологиям / Российский фонд фундаментальных исследований, Труды конф. г. Пущино. 2002, с. 62.

[2] Голубцов К.В., Орлов О.Ю., Айду Э.А.-И., Трунов В.Г. и др. Компьютерная система для диагностики зрения. / Информационные процессы. 2002. Том 2. № 2, с. 275-278.

[3] Шигина Н.А., Куман И.Г., Голубцов К.В., Крутов С.В. Применение электрического тока в диагностике и лечении патологии зрительного нерва и сетчатки. / Клиническая офтальмология. Тема номера: – ГЛАУКОМА. 2001. № 2, с. 66-69.

[4] Егорова Т.С., Голубцов К.В. КЧСМ в определении зрительной работоспособности слабовидящих школьников, / Информационные процессы. 2002. Том 2, № 1, с. 106-110.

[5] А Л. Ярбус. Роль движения глаз в процессе зрения. Изд. Наука 1965.

[6] Ф.В. Грошев, К.В. Голубцов, В.Г. Трунов, "Микропроцессорная система для исследования дефектов зрительной системы человека", 30-я конференция молодых ученых и специалистов ИППИ РАН, 2007.

[7] Голубцов К.В. Куман И.Г. Хейло Т.С., Шигина Н.А., Трунов В.Г. и др. Мелькающий свет в диагностике и лечении патологических процессов зрительной системы человека. Информационные технологии. Том 1 №2 2003г. стр. 114-122.

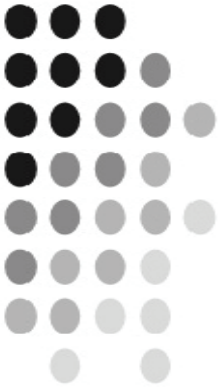
[8] Голубцов К.В. Компьютерная система для диагностики и лечения зрения. Сб. док. Новые информационные технологии в России. Белград (Сербия) 2004. стр.13-15.

[9] Golubtsov R.V., Aidu E.A.И. , Trunov V.G., Barinova N.E. Computer based devices for diagnosis and treatment of visual system. Exhibition – seminar "Advanced Information Technologies". Budapest - 2005.

[10] Иомдина Е.Н., Егорова Е.С., Голубцов К.В., Егорова И.В. Связь состояния аккомодации с временной контрастной чувствительностью глаза. “Биомеханика глаза 2005”. Московский научно-иследовательский институт глазных болезней им. Гельмгольца. Москва-2005. Стр. 48-51.

Раздел 6

Биоинформатика



Positive selection and alternative splicing of human genes

Vasily Ramensky[1], Ramil Nurtdinov[2], Alexei Neverov[3,4], Andrei Mironov[2,3], Mikhail Gelfand[2,5]

1. Engelhardt Institute of Molecular Biology of Russian Academy of Sciences, Moscow, 119991, Russia

2. Moscow State University, Moscow, 119992, Russia

3. State Scientific Center GosNIIGenetika, Moscow, 117545, Russia

4. Central Research Institute of Epidemiology, Moscow, 111123, Russia

5. Institute for Information Transmission Problems (Kharkevich Institute) of Russian Academy of Sciences, Moscow, 127994, Russia

E-mail: ramensky@imb.ac.ru

Abstract

Single nucleotide polymorphisms and human-chimpanzee divergence in the exons of 6,672 alternatively spliced human genes were studied with the aim to understand the forces driving the evolution of alternatively spliced sequences. The alternatively spliced regions from minor isoforms experience lower selective pressure at the amino acid level accompanied by selection against synonymous sequence variation. The results of the McDonald-Kreitman test suggest that, unlike the constitutive regions, they are also subject to the positive selection, with up to 27% of amino acids fixed by positive selection.

1. Introduction

Alternative splicing (AS) of genes is the processing of an RNA transcript into various mRNA molecules (and hence proteins) by including some exons and excluding others. AS is recognized as one of the major sources of proteome complexity [1,2] with the estimates of fraction of alternatively spliced genes in the human genome up to 70-80% [3,4]. Alternative splicing is also an important mechanism of accelerated genome evolution, associated with relaxation of selection pressure against certain types of evolutionary events, in particular, against amino acid substitutions [2,5], with evolutionary young exons evolving at a higher rate [6].

We show that the alternatively spliced exons experience lower selective pressure at the amino acid level accompanied by selection against synonymous sequence variation. The results of the McDonald-Kreitman test (see below) [7] suggest that, unlike the regions constitutively included in the mRNA, the alternatively spliced ones also experience the positive selection, with up to 27% of amino acids fixed by positive selection.

2. Materials and Methods

The splicing alternatives were derived from EDAS [8], a database containing genomic, protein, mRNA and EST sequences, and then processed by IsoformCounter [9], a conservative algorithm for construction of a representative and non-redundant set of protein isoforms compatible with AS data for a gene. This generated 52,151 constitutive and 14,196 alternative gene regions in 6,671 human genes. We define the exon inclusion level A, $0<A\leq1$, as the total number of all protein, mRNA and EST sequences that cover the corresponding exon and include the region, divided by the total number of sequences covering the exon. All regions were then divided into three major classes: minor, major and constitutive. Further, 6,465 SNPs from build 125 of the dbSNP database [10] and 50,649 human-chimpanzee substitutions derived from the whole genome alignments [11] were mapped to the genes.

3. Results

A common measure of the degree of evolutionary constraint on a sequence is the ratio *K*a/*K*s of nonsynonymous substitutions per nonsynonymous site (*K*a) to synonymous substitutions per synonymous site (*K*s) with higher values of the ratio indicating weaker negative selection acting on a sequence.

The ratios *K*a/*K*s for SNPs and divergence are very close for constitutive and major alternative regions but differ more than two-fold from those for functional minor alternative regions, confirming that lower negative selection against amino acid substitutions is especially characteristic of these fragments (Table 1). Table 1 also contains the results of the McDonald-Kreitman test [7], in which the numbers of synonymous polymorphisms (*P*s) and substitutions (*D*s) and the numbers of non-synonymous polymorphisms (*P*a) and substitutions (*D*a) are compared in a contingency table. A statistically significant excess of non-synonymous substitutions relative to polymorphisms (*D*a/*D*s > *P*a/*P*s) implies positive selection that provides fixation of advantageous mutations. The fraction α of fixed amino acid substitutions driven by positive selection is then estimated as $\alpha=1-(Pa/Ps)/(Da/Ds)$ [12]. This value equals 0.27 for the minor alternatives suggesting that they experience positive selection, unlike the major alternatives and constitutive exons with $\alpha<0$ indicating purifying selection (Table 1a).

We have repeated the analysis on SNPs from the basic set observed in two individual genomes. The values of α for constitutive, major and minor exons were –0.05, –0.02, 0.23, respectively, confirming the basic result albeit with a lower statistical significance. The test was repeated with SNPs for which the frequency of the derived (new) allele is known and not less than 5% (Table 1b) and 10%. With these presumably neutral SNPs, the fraction α of substitutions fixed by positive selection for minor alternatives is estimated then as 0.27 and 0.29, respectively, with Fisher's test value ~0.01. The Fisher test value F given in the last column of the table 1 reflects the probability that the difference between *D*a/*D*s and *P*a/*P*s is random.

The synonymous divergence and polymorphism rates are slightly lower in minor AS exons. This observation can be explained by RNA-level selection acting on synonymous sites of alternatively spliced exons in order to maintain the splicing regulation motifs, in particular, ESEs (exonic splicing enhancers) [13].

5. Discussion

Positive selection could not be observed in earlier studies relying on comparisons of *K*a/*K*s values in the alternative and constitutive regions of pairwise aligned mammalian genes, since the latter technique does not allow one to distinguish between positive selection and relaxed negative selection. The advantage of McDonald-Kreitman test is in that it detects the fraction of between-species amino acid diversity that cannot be explained by neutral variation. In our case, the increase of nonsynonymous diversity in minor isoforms is greater than that expected due to relaxation of negative selection estimated from the polymorphism levels and can naturally be interpreted as a trace of positive selection.

6. References

[1] B.R. Graveley, "Alternative splicing: increasing diversity in the proteomic world", *Trends Genet*, 2001, 17, 100-107.

[2] A.A. Mironov, J.W. Fickett, M.S. Gelfand, "Frequent alternative splicing of human genes", *Genome Res*, 1999, 9, 1288-1293.

[3] J.M. Johnson, J Castle., P. Garrett-Engele, Z. Kan, P.M. Loerch, C.D. Armour, R. Santos, E.E. Schadt, R. Stoughton, D.D. Shoemaker, "Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays", *Science*, 2003, 302, 2141-2144.

[4] D. Kampa, J. Cheng, P. Kapranov, M. Yamanaka, S. Brubaker, et al., "Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22", *Genome Res*, 2004, 14, 331-342.

[5] Y. Xing, C. Lee, "Alternative splicing and RNA selection pressure -- evolutionary consequences for eukaryotic genomes", *Nat Rev Genet*, 2006, 7, 499-509.

[6] X.H. Zhang, L.A. Chasin, "Comparison of multiple vertebrate genomes reveals the birth and evolution of human exons", *Proc Natl Acad Sci U S A*, 2006, 103, 13427-13432.

[7] J.H. McDonald, M. Kreitman, "Adaptive protein evolution at the Adh locus in Drosophila", *Nature*, 1991, 351, 652-654

[8] R.N. Nurtdinov, A.D. Neverov, D.B. Mal'ko, I.A. Kosmodem'ianskii, E.O. Ermakova, V.E. Ramenskii, A.A. Mironov, M.S. Gel'fand, "EDAS, databases of alternatively spliced human genes", *Biofizika*, 2006, 51, 589-592. Russian

[9] A.D. Neverov, I.I Artamonova, R.N. Nurtdinov, D. Frishman, M.S. Gelfand, A.A. Mironov, "Alternative splicing and protein function", *BMC Bioinformatics*, 2005, 7, 266.

[10] S.T. Sherry, M.H. Ward, M. Kholodov, J. Baker, L. Phan, E.M. Smigielski, K. Sirotkin, "dbSNP: the NCBI database of genetic variation", *Nucleic Acids Res*, 2001, 29, 308-311.

[11] W.J. Kent, C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, et al., "The human genome browser at UCSC", *Genome Res*, 2002, 12, 996-1006.

[12] N.G. Smith, A. Eyre-Walker, "Adaptive protein evolution in Drosophila", *Nature*, 2002, 415, 1022-4.

[13] T.I. Orban, E. Olah, "Purifying selection on silent sites - a constraint from splicing regulation?", *Trends Genet*, 2001, 5, 252-253.

Table 1. SNP and divergence density values and the McDonald-Kreitman test. The table contains the ratio values *K*a/*K*s of nonsynonymous substitutions per nonsynonymous site (*K*a) to synonymous substitutions per synonymous site (*K*s) for three types of gene regions. *K*a and *K*s denote both polymorphism densities and divergence, with discriminating labels "SNP" and "Div", respectively. The last column shows the significance computed by the Fisher test applied to the four numbers organized in a 2-by-2 contingency table. (a) All exons; (b) Genes with at least one SNP with known frequency, only neutral SNPs (both validated and non-validated) with the frequency of the derived (new) allele ≥5%.

Exon type	Number of exons	*Ka*/*Ks* (SNP)	*Ka*/*Ks* (Div)	α	The Fisher test significance
		(a) All exons			
Constitutive	52151	0.26	0.25	-0.04	0.118
Major	9799	0.26	0.26	-0.02	0.442
Minor	4397	0.55	0.76	0.27	0.001
		(b) All exons, SNPs with derived allele frequency ≥5%			
Constitutive	30060	0.25	0.25	0.04	0.149
Major	4801	0.23	0.29	0.19	0.015
Minor	2026	0.54	0.74	0.27	0.012

SDPfox - Web-сервер для анализа субстратной специфичности белков

П. Мазин[1*], О. Калинина[2,3], А. Рахманинова[1,2], А. Миронов[1,2], М. Гельфанд[2]
[1)]Московский государственный университет им. М.В. Ломоносова, факультет биоинженерии и биоинформатики, Москва, Россия; [2)]УНЦ «Биоинформатика» ИППИ им. А.А. Харкевича РАН, Москва, Россия; [3)]EMBL-Heidelberg, Germany
[*]iaa.aka@gmail.com

Аннотация

Предсказание функции и специфичности белка по аминокислотной последовательности или по выравниванию семейства последовательностей является одной из основных задач биоинформатики и приобрело в последние годы особую актуальность. Для выполнения белком своей биологической функции особенно важными являются остатки, которые можно условно разделить на два типа: одни отвечают за общую функцию белка и консервативны для всего семейства, другие - за специфическое распознавание взаимодействующей молекулы (Specificity Determining Positions, SDP) и, предположительно, консервативны в группе белков с одинаковой специфичностью (группа специфичности). В этой работе мы представляем Web-сервер SDPfox (http://storage.bioinf.fbb.msu.ru/SDPfoxWeb/main.jsp) и java-программу SDPproff, реализующие алгоритмы для поиска SDP и предсказания групп специфичности по аминокислотному выравниванию. Мы провели сравнение SDPproff с другими программами и серверами, разработанными для тех же целей, на реальных и генерированных данных. Тестирование показывает, что результаты SDPproff по крайней мере не хуже, чем результаты других методов, а в ряде случаев существенно превосходят их.

1. Введение

Предсказание функции и специфичности белка по аминокислотной последовательности или по выравниванию семейства последовательностей является одной из основных задач биоинформатики, и приобрело в последние годы особую актуальность. Для выполнения белком своей биологической функции особенно важными являются остатки, которые можно условно разделить на два типа: одни отвечают за общую функцию белка и консервативны для всего семейства, другие - за специфическое распознавание взаимодействующей молекулы и, предположительно, консервативны для белков с одинаковой специфичностью (Specificity Determining Positions, SDP)..

2. Результаты и обсуждение

Ранее нашей группой была разработана группа методов SDPclust для поиска SDP на основании выравнивания [1]. Этот метод позволяет анализировать семейства белков, для которых отсутствует информация о специфичности отдельных представителей. SDPclust включает в себя 4 взаимосвязанных метода: SDPlight – метод для быстрого поиска SDP [2], SDPprofile – метод, осуществляющий перегруппировку последовательностей в соответствии с наилучшим сходством по предсказанным SDP [3], SDPgroup – итеративная процедура для разбиения семейства на группы специфичности с использованием обучающей выборки (набора белков с известной специфичностью) [1], SDPtree – метод, строящий кластерное дерево предполагаемых специфичностей белков в случае, когда обучающей выборки нет, т.е. осуществляет деление семейства на группы специфичности.

В этой работе мы представляем Web-сервер SDPfox (http://storage.bioinf.fbb.msu.ru/SDPfoxWeb/main.jsp), реализующий методы SDPlight, SDPprofile и SDPgroup, и java-программу SDPproff реализующею все методы пакета SDPclust как в консольном, так и в визуальном режиме.

Мы провели сравнение SDPproff с другими программами и серверами осуществляющими предсказание детерминант специфичности и групп специфичности на реальных и

генерированных данных. По результатам тестирования SDPproff показывает результаты по крайней мере не хуже, чем другие методы, а в ряде случаев существенно превосходит их.

3. Благодарности

Авторы выражают благодарность к.б н. А.Б. Рахманиновой и проф. А.А. Миронову за ценное обсуждение.

4. Литература

[1] P.V. Mazin, A.B.Rakhmaninova, O.V. Kalinina, “SDPclust: a new tool for prediction protein specificity in MPA”. Moscow Conference on Computational Molecular Biology `07, 2007.

[2] П.В. Мазин и О.В. Калинина, “SDPlight – быстрый метод поиска позиций, определяющих специфичность в белковых выравниваниях” *Международная научная конференция студентов, аспирантов и молодых ученых “Ломоносов-2007”*, 2007

[3] O.V. Kalinina A.A. Mironov, M.S. Gelfand and A.B. Rakhmaninova, “Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families”, *Protein Science*, 2004, pp. 13: 443-456

Подтверждение участия давления отбора в поддержании стабильности CpG островов в геноме

Инна Перцовская[1,*], Нина Опарина[2], Дмитрий Виноградов[3], Андрей Миронов[1], Александр Фаворов[2,4]
[1]*Московский государственный университет им. М. В. Ломоносова,* [2]*ГосНИИГенетика,* [3]*Институт проблем передачи информации РАН,* [4]*Johns Hopkins University*
[*]*inna.perts@gmail.com*

Аннотация

Метилирование ДНК – модификация, играющая ключевую роль в регуляции экспрессии генов, дифференцировке клеток, защите генома и др. CpG острова – обычно защищенные от метилирования участки генома, в который количество CG динуклеотидов больше, чем в среднем по геному. Они часто обнаруживаются в промотерных областях генов и участвуют в регуляции экспрессии этих генов. CpG острова обнаружены только у хордовых, и, по-видимому, несколько независимых причин привели к их появлению и сохранению в эволюции. В данной работе на основе множественного выравнивания геномов были рассмотрены мутации цитозина с CpG парах внутри островов и в выборке 1000 наборах случайных участков генома у высших приматов и найдена значимая разница в вероятности мутаций вида CpG->notCpG в CpG островах и в среднем по геному. Таким образом, пользуясь методами биоинформатики и статистики, было показано влияние эволюционного отбора на сохранение повышенной концентрации CpG динулеотидов в CpG островах.

1. Введение

Полное секвенирование человеческого генома сделало ещё более актуальным изучение функциональных модификаций хроматина, что привело к появлению такого раздела биологии, как эпигенетика. Этот раздел объединяет в себе различные направления исследований модификаций ДНК и связанных с ней белков в ядре. Метилирование ДНК является единственной известной ковалентной модификацией ДНК, встречающейся у позвоночных. Обычно метилированию подвержен цитозин, находящийся в цепи ДНК перед гуанином (в CpG динуклеотиде). При этом специальный фермент – ДНК-метилтрансфераза – присоединяет метильную группу к 5 углероду цитозина C->5mC). Такая модификация наследуется дочерней цепью при репликации и играет важную роль в регуляции транскрипции и дифференцировки клеток. Показана определяющая роль метилирования ДНК в импринтинге и инактивации одной X хромосомы. В то же время, 5mC путем спонтанного дезаминирования может превращаться в Тимин, что часто приводит к мутациям C->T, которые не узнаются системами репарации ДНК и, как считается, приводят к вымыванию CpG динуклеотидов в процессе эволюции (ссылка). Таким образом, количество CpG пар в геноме значительно (для человека – примерно в 5 раз) меньше ожидаемого при биномиальном распределении.

Однако существуют особые участки ДНК, в которых количество CpG значительно больше, чем в среднем по геному и приближается к ожидаемому при нормальном распределении нуклеотидов. Такие участки получили название CpG островов. В настоящее время CpG острова аннотируются в геноме алгоритмически. В геноме человека порядка 28000 CpG островов со средней длиной 700 п н. Часто CpG острова обнаруживаются в промотерных областях генов (около половины генов имеют CpG остров в промотерной области). Большинство CpG островов в норме неметилированы, однако могут метилироваться при дифференцировки тканей, что приводит к инактивации транскрипции находящегося рядом гена.

Функции большинства CpG островов неизвестны. Были найдено большое количество CpG островов, находящихся в межгенных областях, а также несколько примеров в норме метилированных CpG островов, предположительно, блокирующих транскрипции мобильных элементов генома.

2. Гипотезы

Обедененность генома CpG динуклеотидами и феномен CpG островов, также как и метилирование цитозина в CpG парах, отчетливо наблюдается только у позвоночных. То есть это достаточно новые эволюционные изменения генома. Причины появления и сохранения CpG островов изучены плохо. Главенствующая гипотеза предполагает, что CpG острова – защищенные от метилирования участки генома. В результате этой защиты, природа которой неизвестна, цитозин не подвержен метил-зависимой мутации C->T. Однако такой механизм не объясняет сохранения группы в норме метилированных CpG островов.

Не отменяя важной роли защиты от метилирования в сохранение динуклеотидного состава CpG островов, мы предполагаем, что имеют место и другие причины. Первая возможная причина – размножение в геноме CpG богатого мобильного элемента или ретротранспозона на определенном этапе эволюции. Дальнейшая эволюция привела к постепенному дивергентному расхождение этих участков, однако количество CpG пар уменьшилось не так значительно, как в остальном геноме из-за того, что такие участки появились поздно. На возможность такого механизма появления некоторых островов указывает наблюдения идентичных островов на разных хромосомах генома человека.

Второй гипотетический механизм сохранения динуклеотидного состава CpG островов – не связанная с метилированием функциональная значимость именно такого состава. Например, возможно, есть определенные регуляторные белки, умеющие CpG богатые сайты связывания. В результате CpG острова находились под сильным давлением стабилизирующего отбора.

Поскольку вероятнее всего, что все эти механизмы и, возможно, некоторые другие, сыграли определенную роль в сохранение CpG островов в геноме позвоночных, определение степени влияния каждой отдельной причины в эволюции CpG островов. Мы разработали метод, позволяющий определить играет ли давление эволюционного отбора роль в появлении и сохранении CpG островов. Таким методом является подсчет вероятности рождения CpG пар в CpG островах по сравнению с средней скоростью рождения CpG в геноме. В то время как метил-зависимая мутабельность цитозина и рождение CpG островов в результате размножения ретроэлементов должны оказывать роль только на исчезновение CpG пар, давление отбора должно играть существенную роль именно в процессе рождения новых CpG пар. Таким образом мы можем детерминировать этот гипотетических механизм от других и проверить его независимо от других причин.

3. Материалы и методы

Для изучения эволюции CpG динуклеотидов в геномах разных групп позвоночных мы использовали множественное выравнивание с ftp-сервера Genome Browser. В зависимости от исследуемых групп животных, мы выбирали определенные тройные выравнивания из множественного (человек – мышь – крыса для грызунов (человек в данном случае выступает как отдаленный родственник) и человек – шимпанзе-макака для приматов (макака – удаленный родственник)). Используя такое тройное выравнивания, мы подсчитывали события замены цитозина на другие нуклеотида и обратные им события на линии приматов и грызунов соответственно при условии наличия гуанина во всех трех цепях после гуанина.

Следующей задачей было определение влияния отбора на сохранения CpG островов у приматов с помощью сравнения скорости обратных замен notCpG->CpG в CpG островах и в среднем по геному. Для этого мы использовали 1000 наборов случайных участков генома , полученных следующим образом: каждый раз мы брали разметку настоящих островов и рандомизировали начало участка внутри хромосомы, оставляя длину неизменной. Таким образом, мы получили набор «псевдоостровов» со сходной геометрией, что и реальные CpG острова.

Мы сгенерировали 1000 таких наборов и подсчитали в них обратные замены CpG так, как описано выше. По полученным данным мы построили гистограмму распределения вероятности замен островов (рис.???).

Мы создали базу данных MySQL, включающую таблицы с разметками CpG островов для разных геномов, разметки генов из Genome Browser, также таблицу с полученными нами данными по количеству событий перехода CpG в другие нуклеотиды и обратно и событий сохранения динуклеотидов для CpG островов человека.

4. Результаты

Полученные данные свидетельствуют о том, что метилирование определенно приводит к вымыванию CpG динуклеотидов и в островах, и в целом по геному как у приматов, так и у грызунов (таб. 1).

Участок ДНК	Направление мутаций	грызуны	приматы
геном	notCpG→CpG	**1,5**	**0.6**
	CpG→notCpG	**27,3**	**9.05**
CpG острова	notCpG→CpG	**3,1**	**0.63**
	CpG→notCpG	**17,2**	**1.89**

Таблица 1. Скорости мутаций цитозина в CpG динуклеотидах в геномах и CpG островах грызунов и приматов (в процентах)

При этом видно, что в целом мутационный процесс в CpG островах в любом направлении идет медленнее, что и натолкнуло нас на вторую гипотезу сохранения CpG островов. Для ее проверки мы сравнили скорость рождения CpG динуклеотидов, на которую метил-зависимые мутации CpG->TpG не оказывают влияние, со скоростью рождения CpG динуклеотидов в распределение, основанном на анализе скорости рождения в 1000 наборах случайных участков генома. Показано, что отличие скорости рождения в CpG островах и в среднем по геному достоверно с p-value<10^{-3}, что подтверждает гипотезу о функциональной природе сохранения CpG динуклеотидов в CpG островах. Особо значимый вклад в разницу скоростей рождения вносит фракция CpG островов, находящихся в межгенных участках. Для них P(notCpG->CpG)>0.009. Таким образом получается, что во внегенных CpG островах гораздо быстрее происходит процесс рождения новых CpG динуклеотидов.

5. Выводы

CpG острова имеют специфическое эволюционное поведение по сравнению с другими участками генома, особенно в отношении CpG динуклеотидов. Скорость рождения и смерти CpG динуклеотидов в CpG островах ниже, чем в среднем по геному, как на линии грызунов, так и на линии приматов. Исключение составляет скорость рождения CpG динуклеотидов во внегенных CpG островах, которая выше, чем средняя по геному, что указывает на верность гипотезы о сохранении CpG островов под давлением отбора.

Comparative Analysis of Transcription Regulation Strategies of Bacteriophages T4 and RB49

Pugach K.[1*], Severinov K.[1]
[1]*Institute of Molecular Genetics, Russian Academy of Sciences, Moscow, 123182*
[*]*xenik-alt@mail.ru*

Gelfand M.[2], Ravcheev D.[2]
[2]*Institute of Information Transmission Problems RAS, Moscow,127994*

Abstract

The genome of bacteriophage T4-related coliphage RB49 has recently been sequenced. Given the highly integrated strategy of transcription regulation during bacteriophage T4 development, it was reasonable to expect that RB49 would have a similar transcription regulation strategy. Contrary to this expectation, the RB49 genome has no recognizable homologs for essential T4 regulators such as motA, asiA, alc, and g42. On the other hand, homologs of T4 alt, denA, g33, g45, and g55 genes are readily detected. The absence (by whole-genome bioinformatic analysis) of sequences matching strong T4 early promoters in the RB49 genome can be regarded as evidence for a distinct mechanism of early genes transcription in RB49. The absence of asiA and motA (and T4 middle promoters), as well as the presence of putative strong σ^{70} RNAP promoters in front of many homologs of T4 middle genes, suggests that the middle transcription, if it exists in RB49, shall be also regulated differently than in T4. On the other hand, the late transcription of RB49 genome appears to proceed the same way as in T4. We will present the results of experimental study that i) identify temporal classes of RB49 transcription; ii) map RB49 promoters of different classes, and iii) identify RB49-encoded transcription factors that interact with host RNA polymerase.

1. Introduction

Escherichia coli bacteriophage T4 is a classical model object for studies of transcription regulation in prokaryotes. Transcription of about 40 *early* phage genes starts immediately after its DNA is injected into the host cell [1]. Early genes are transcribed by cellular σ70 RNA polymerase (RNAP) holoenzyme. Preferential transcription of these genes is due to high intrinsic strength of their promoters. Additional mechanism ensuring high activity of early promoters is ADP-ribosylation of the RNAP α subunit (by Alt and ModA), which prevents the recognition of host promoters UP-elements by the α subunit and releases large quantities of RNAP to transcribe from phage promoters. Another T4-encoded protein, Alc, terminates transcription of host DNA (contains cytosine) but has no effect on transcription of phage DNA (contains hydroxymethyl cytosine synthesized by phage-encoded hydroxymethyltransferase Gp42), leading to host transcription shut-off.

Two phage nucleases DenA and DenB degrade host DNA at the early steps of infection; phage DNA itself remains intact due to modified cytosine residues. After 2 to 3 min of infection (at 30 oC) early promoters are abruptly shut off and middle promoters are turned on [2]. Their activation requires two T4-encoded proteins, MotA and AsiA, the products of early genes. MotA binds to the "-30" region of these promoters (MotA box) and AsiA associates with the RNAP. Finally, specific T4-encoded sigma factor, the product of gene *55*, a middle gene, is required for T4 late promoter recognition.

Coliphage RB49 is a relative of T4 [3]. Given the apparently tight coherence of transcription regulation during T4 development it was reasonable to expect that RB49 will have the same transcription regulation strategy as T4. Contrary to the expectation, the RB49 genome appeared to have no homologs for essential T4 regulators such as *motA*, *asiA*, *alc*, and *g42*. On the other hand it has *alt*, *denA*, *g33*, *g45* and *g55* homologs.

The apparent presence (by bioinformatic analysis) of strong σ70 RNAP promoters in front of several putative early genes, as well as absence of sequences matching strong T4 early promoters in the RB49 genome can be regarded as evidence for a distinct mechanism of early genes transcription in RB49. The absence of *asiA* and *motA* (and MotA boxes) as well as the presence of strong σ70 RNAP promoter motifs in front of many homologs of T4 middle genes, suggests that the middle transcription, if it exists in RB49, shall be regulated differently. Lastly, the

existence of the *g55, g33, g45* homolog suggests that late transcription mechanisms in RB49 and T4 are similar. No evidence as to whether RB49 executes host transcription shut-off or host DNA degradation is available.

2. Results

It was shown that RB49 phage instead of three has only two temporal gene classes, called early and late. We identified promoter sequences of genes belonging to these two classes. Promoter sequences of all investigated early genes match the consensus sequence of *E. coli* σ^{70} – promoters. Promoter sequences of all but one late gene match the consensus sequence of T4 late promoters. On the contrary late *modB* gene promoter is almost equal to consensus sequence of *E. coli* σ^{70} – promoters. This result is of particular interest because it means that σ^{70}-containing RNAP should be active at the late stages of phage RB49 infection.
It was also shown that RB49 executes host transcription shut-off.

3. Conclusion

RB49 may have a simpler pattern of transcription than T4, with only two, rather than three, classes of temporally controlled promoters. These are early and late promoters, which have considerable similarity with *E. coli* σ^{70} and late promoters of T4 respectively.
RB49 infection causes host genes shut-off possibly by the same mechanism as in T4 (strong early promoters and ADP-ribosylation of RNAP).

4. References

[1] J.D., Karam, *et al. Molecular Biology of Bacteriophage T4*, American Society for Microbiology, Washington, DC (1994).

[2] K. Luke, *et al*. “Microarray analysis of gene expression during bacteriophage T4 infection”. *Virology*, ***299***, (2002), pp. 182-191.

[3] Desplats, C. *et al*. “Snapshot of the genome of the pseudo-T-even bacteriophage RB49”. *J.Bacteriol*., ***184***, (2002), pp. 2789-2804.

Комплексная регуляция дыхания бактерий: исследование методами сравнительной геномики

Д.А. Равчеев
Институт проблем передачи информации им. А.А. Харкевича РАН
ravcheyev@gmail.com

Аннотация

Многие прокариотические организмы, например Escherichia coli, могут использовать различные субстраты для дыхания, что позволяет им существовать в широком диапазоне условий. Подобная приспособленность обеспечивается за счет наличия множества различных дыхательных цепей а также сложной системы регуляции дыхания. Однако, несмотря на значимость данного вопроса, исследования регуляции дыхания к настоящему проводились лишь для некторых организмов.

В настоящей работе было проведено комплексное исследование регуляции дыхания в гамма-протеобактериях. Для этого были рассмотрены три глобальные регуляторы дыхания: белки Fnr и ArcA, отвечающие за переключение между аэробным и анаэробным метаболизмом и белок NarP, контролирующий нитрат-нитритное дыхание.

В ходе работы были изучены Fnr-, ArcA- и NarP-регулоны в 10 видах бактерий, относящихся к трем порядкам. Для каждого из трех регуляторов были найдены новые потенциально регулируемые гены. Также был показан ряд таксон-специфических перестроек регуляторных каскадов, и были обнаружены изменения в составе регулонов, коррелирующие с этими перестройками.

1. Введение

Гаммапротеобактерии, наиболее хорошо изученным представителем которых является энтеробактерия *Escherichia coli*, способны приспосабливаться к широкому диапозону условий существования. В том числе, такая приспособляемость достигается за счет наличия множества различных дыхательных цепей, что позволяет использовать для дыхания широкий спектр субстратов.

Несмотря на многообразие, все дыхательные цепи бактерий сходны по строение. Так, дыхательные цепи имеют так называемую модульную структуру, то есть состоят из заменяемых компонентов. Таковыми модулями являются дегидрогеназы доноров электронов, мембранорастворимые переносчики электронов – хиноны и оксидоредуктазы доноров электронов. Наличие в дыхательной цепи того или иного вида компонентов регулируется в соответствии с окружающими условиями и физиологическим состоянием клетки [1].

Понятно, что столь сложная система организации дыхания требует не менее сложной регуляции. Было показано, что в регуляции экспрессии генов дегидрогеназ и оксидоредуктаз сигналами являются концентрации акцепторов электронов, тогда как концентрации доноров электронов не имеют значения [2].

К настоящему моменту регуляция дыхания хорошо изучена в *E. coli*. Показано, что глобальная регуляция дыхания осуществляется целым рядом регуляторных систем. Первый уровень регуляции осуществляется белком Fnr, реагирующим непосредственно на концентрацию молекулярного кислорода. В анаэробных условиях данный белок димеризуется и становится способным к связыванию с ДНК. При этом Fnr активирует экспрессию генов анаэробного и репрессирует гены аэробного дыхания [3].

Следующий уровень регуляции осуществляется двукомпонентной системой ArcB-ArcA. Сенсорный белок ArcB реагирует на окислительно-восстановительный статус хинонов, в анаэробных условиях активируя ArcA, который, в свою очередь связывается с промотерными областями регулируемых генов. В целом, эффект данной системы похож на таковой для белка Fnr: в анаэробных условиях активируются гены анаэробного метаболизма и репрессируются гены аэробного [4, 5].

За следующий уровень регуляции ответственна дуплицированная двукомпонентная система, включающая сенсорные белки NarX и NarQ и факторы транскрипции NarL и NarP. Данная система аткивизируется в присутствии

ионов нитрата и нитрита. По всей видимости, удвоенная регуляторная система необходима для тонкой настройки системы нитрат-нитритного дыхания в зависимости от соотношений нитрата и нитрита. В активном состоянии белки NarL и NarP регулируют гены нитрат-нитритного дыхания и репрессируют гены других, менее эффективных систем нитрат-нитритного дыхания [6, 7].

Несмотря на то, что в *E. coli* регуляция дыхания исследуется уже достаточно давно, до сих пор многие вопросы остаются нерешенными. Так, до сих пор неясна структура сайтов связывания белка NarL. Кроме того, регуляция несколькими факторами транскрипции исследовалась лишь для единичных оперонов. Также практически не было исследований регуляции в родственных организмах. Поэтому в настоящей работе было проведено исследование комплексной регуляции дыхания в ряде геномов гамма-протеобактерий.

2. Данные и методы

2.1. Процедура определения генов, входящих в обобщенный регулон

Для предсказания структуры регулонов был использован сравнительно-геномный подход. [8] Так, ген считался принадлежащим к регулону, если в его предполагаемой регуляторной области были предсказаны потенциальные сайты связывания соответствующего фактора транскрипции, и такие сайты были найдены в нескольких геномах.

Для поиска потенциальных сайтов связывания факторов транскрипции был использован метод матриц (профилей) позиционных весов [8].

В случае, когда оперонная струкрура локуса не была исследована, гены считались принадлежащими к одному оперону, если они имели одинаковое направление считывания, расстояние между ними не превышало 200 п.н. и структура локуса сохранялась в нескольких родственных геномах.

2.2. Программное обеспечение

Для поиска ортологичных генов и потенциальных сайтов связывания был использован пакет программ GenomeExplorer [9]. Для построения матриц (профилей) позиционных весов была использована программа SignalX [9]. Поиск гомологов в базах данных проводился с помощью программы BLAST [10]. Для построения множественных выравниваний аминокислотных и нуклеотидных последовательностей, а также для построения филогенетических деревьев была использована программа ClustalX [11]. Для построения диаграмм Лого, отображающих структуру регуляторного сигнала, использовалась программа WebLogo [12].

2.3. Исследованные геномы

В настоящей работе были исследованы 13 полных геномных последовательностей гамма-протеобактерий, относящихся к трем порядкам: Enterobacteriales (*E. coli*, *Salmonella typhi*, *S. typhimurium*, *Yersinia pestis*, *Y. enterocolitica*, *Erwinia carotovora*), Pasteurellales (*Pasteurella multocida*, *Actinobacillus actinomycetemcomitans*, *Haemophilus ducreyi*, *H. influenzae*), Vibrionales (*Vibrio cholerae*, *V. fischeri*, *V. parahaemolyticus*, *V. vulnificus*). Все последовательности были взяты из базы данных GenBank [13].

3. Результаты и обсуждение

3.1. Эволюция регуляторных систем

Ортологи белка Fnr были обнаружены во всех исследованных геномах. Все найденные белки сохранили функциональные остатки цистеина, характерные для белка-регулятора из *E. coli*.

Также во всех изученных организмах были найдены ортологи ArcA. Однако, ортологи сенсорной киназы ArcB присутствовали не во всех геномах. Так, не удалось обнаружить ортологов в *H. ducreyi* и *H. Influenzae*. Тем не менее, присутствие в этих геномах гена регулятора ArcA указывает на возможность исследования в них регуляции данным белком.

Намного сложнее оказалась ситуация с системой регуляции нитрат-нитритного дыхания. Так, удвоенная система NarX-NarL и NarQ-NarP была найдена лишь в *E. coli*, *S. typhi*, *S. Typhimurium* и *E. carotovora*. Остальные же геномы содержали лишь одинарную систему NarQ-NarP. Поскольку вопрос о структуре сайта связывания NarL до сих пор остается спорным, было решено ограничить исследования только геномами, содержащими одиночную систему. Таким образом, в настоящей работе были исследованы десять геномов, содержащих Fnr, ArcA и NarP: *Y. pestis*, *Y. enterocolitica*, *P. multocida*, *A. actinomycetemcomitans*, *H. ducreyi*, *H. influenzae*, *V. cholerae*, *V. fischeri*, *V. parahaemolyticus*, *V. vulnificus*.

3.2. Предсказание регуляторных каскадов

Одним из наиболее значимых и в то же время неожиданных результатов работы стало

предсказание новых регуляторных каскадов с участием исследуемых факторов транскрипции.

В случае *E. coli* таковые каскады были показаны экспериментально. Известно, что в данном организме Fnr представляет собой главный регулятор дыхания и контролирует экспрессию генов других регуляторов. Так, Fnr репрессирует транскрипцию оперона *narXL* [14] и регулирует экспрессию гена *arcA* [15]. Кроме того, показана авторегуляция гена *fnr* [14]. Также была предсказана положительная ауторегуляция гена *arcA* [16]. Регуляторы же нитрат-нитритного дыхания формируют достаточно сложную сеть каскадов, где NarL активирует экспрессию оперонов *narXL*, *narP* и *narQ*, а NarP, в свою очередь, активирует транскрипцию оперона *narXL* [6].

Однако, ситуация в других исследованных геномах, как показали результаты данной работы, сильно отличается от наблюдаемой в *E. coli*. Так, в *Yersinia* spp., Fnr регулирует экспрессию *arcA* и оба гена *fnr* и *arcA* регулируют собственную экспрессию. При этом система регуляции нитрат-нитритного дыхания не принимает участия в регуляторных каскадах.

Для геномов группы Vibrionales характерна авторегуляция генов*fnr*, *arcA* и *narP*. Также белок ArcA регулирует ген *narP* gene, тогда как Fnr контролирует экспрессию *arcA* и *narP*. Таким образом, в данной группе организмов Fnr сохраняет позицию главного регулятора, но при этом значительно возрастает роль ArcA в регуляторных каскадах.

Наиболее существенные изменения в структуре регуляторных каскадов наблюдались в группе Pasteurellales. В большинстве организмов из этого порядка сайты связывания для всех трейх регуляторов, Fnr, ArcA и NarP, были найдены перед геном *fnr*,тогда как перед *arcA* и *narP* сайтов не обнаружено. Это позволяет полагать, что в данной группе роль Fnr в регуляторных каскадах уменьшается по сравнению с *E. coli*. Небольшие различия были зафиксированы в случае *H. ducreyi*, где перед геном *fnr* были обнаружены сайты только для Fnr и ArcA.

Таким образом, структура регуляторных каскадов достаточно сильно различается в разных порядках гамма-протеобактерий.

3.3. Новые члены регулонов

Как уже упоминалось, исследование комплексной регуляции дыхания проводились лишь для нескольких оперонов. В настоящей работе изучение регуляции множеством факторов транскрипции позволило выявить новые гены, входящие в Fnr-, ArcA- и NarP-регулоны. Ниже приводится описание двух наиболее интересных случаев обнаружения новых членов регулонов.

3.2.1. Гены АТФ-синтазы. АТФ-синтаза представляет собой основной компонент дыхания. Функция данного сложного белкового комплекса состоит в преобразовании энергии трансмембранного ионного градиента в энергию химической связи путем синтеза АТФ из АДФ. В *E. coli* все компоненты комплека АТФ-синтазы кодируются генами одного оперона, *atpIBEFHAGDC*. Ранее регуляция *atp*-оперона каким-либо из факторов транскрипции не была установлена. Однако, в настоящей работе была предсказана регуляция данного оперона. Так, в *Yersinia* spp. были обнаружены консервативные сайты связывания ArcA, а у представителей Vibrionales – консервативные Fnr-сайты.

3.2.2. Гены NADH-дегидрогеназы. В *E. coli* одним из основных дыхательных ферментов является протон-транспортирующая NADH-дегидрогеназа, компоненты которой кодируются опероном *nuoABCDEFGHIJKLMN*. Экспрессия данного оперона в *E. coli* контролируется белками NarL и Fnr [17]. Однако, ни в одном из исследованных геномов не было обнаружено ортологов для генов *nuo*-оперона. Тем не менне, во всех геномах был обнаружен оперон *nqrABCDEF*, кодирующий компоненты натрий-транспортирующую NADH-дегидрогеназу. При этом перед опероном *nqr* во всех исследованных геномах были найдены сайты связывания Fnr. Сайты связывания ArcA перед данным опероном были обнаружены в геномах представителей Vibrionales и Pasteurellales. Таким образом, в случае NADH-дегидрогеназ можно говорить о негомологичной замене с частичным изменением функции.

4. Благодарности

Данная работа выполнена совместно с М.С. Гельфандом и А.В. Герасимовой. Мы благодарим А.А. Миронова и А.В. Фаворова за любезно предоставленное программное обеспечение, а также О.В. Лайкову, Д.А. Родионова и А.Б. Рахманинову за ценные замечания.

Работа была частично поддержана Медицинским институтом Ховарда Хьюза, программой “Молекулярная и клеточная биология” Российской Академии Наук и INTAS.

5. Список литературы

[1] Neidhardt, F.C. (eds.), *Escherichia coli and Salmonella. Cellular and Molecular Biology*, ASM Press, Washington DC, U.S.A., pp. 1325-1333.

[2] S. Berry, "The chemical basis of membrane bioenergetics", *J. Mol. Evol.*, Springer-Verlag, Germany, 2002, pp. 595-613.

[3] P.J. Kiley and H. Beinert, "Oxygen sensing by the global regulator, FNR: the role of the iron-sulfur cluster", *FEMS Microbiol. Rev.*, Blackwell Pub., England, 1998, pp. 341-352.

[4] D. Georgellis, O. Kwon and E.C. Lin, "Quinones as the redox signal for the *arc* two-component system of bacteria", *Science*, American Assn. for the Advancement of Science, Unuted States, 2001, pp. 2314-2316.

[5] S. Iuchi and E.C. Lin, "*arcA* (*dye*), a global regulatory gene in *Escherichia coli* mediating repression of enzymes in aerobic pathways.", *Proc. Natl. Acad. Sci. U.S.A.*, National Academy of Sciences, Unuted States, 1988, pp. 1888-1892.

[6] A.J. Darwin and V. Stewar, "Expression of the *narX*, *narL*, *narP*, and *narQ* genes of *Escherichia coli* K-12: regulation of the regulators", *J. Bacteriol.*, American Society for Microbiology, Unuted States, 1995, pp. 3865-3869.

[7] A.J. Darwin, K.L. Tyson, S.J. Busby and V. Stewart, "Differential regulation by the homologous response regulators NarL and NarP of *Escherichia coli* K-12 depends on DNA binding site arrangement", *Mol. Microbiol.*, Blackwell Scientific Publications, England, 1997, pp. 583-595.

[8] A.A. Mironov, E.V. Koonin, M.A. Roytberg and M.S. Gelfand "Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes**"** *Nucleic Acids Res.*, Oxford University Press, England, 1999, pp. 2981-2989.

[9] А.А. Миронов, Н.П. Винокурова и М.С. Гельфанд, "Программное обеспечение анализа бактериальных геномов", *Мол. биология*, "Наука", Москва, Российская Федерация, стр. 253-262.

[10] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, M.W. Zhang and D.J. Lipman "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucleic Acid Res.*, Oxford University Press, England, 1997, pp. 3389-3402.

[11] J.D. Thompson, T.J. Gibson, F. Plewniak, F. Jeanmougin and D.G.Higgins, "The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools", *Nucleic Acid Res.*, Oxford University Press, England, 1997, pp. 4876-4882.

[12] T.D. Schneider, R.M. Stephens, "Sequence logos: a new way to display consensus sequences", *Nucleic Acid Res.*, Oxford University Press, England, 1990, pp. 6097-6100.

[13] D.A. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell and D.L. Wheeler, "GenBank", *Nucleic Acids Res.*, Oxford University Press, England, 2007, pp. 21-25.

[14] K. Takahashi, T. Hattori, T. Nakanishi, T. Nohno, N. Fujita, A. Ishihama and S. Taniguchi, "Repression of in vitro transcription of the *Escherichia coli fnr* and *narX* genes by FNR protein", *FEBS Lett.*, Elsevier Science B.V., Netherlands, 1994, pp. 59-64.

[15] I. Compan and D. Touati, "Anaerobic activation of arcA transcription in Escherichia coli: roles of Fnr and ArcA", *Mol. Microbiol.*, Blackwell Scientific Publications, England, 1994, pp. 955-964.

[16] А.В. Герасимова, М.С. Гельфанд, В.Ю. Макеев, А.А. Миронов и А.В. Фаворов, "Регулятор ArcA в гамма-протеобактериях. Определение сайтов связывания и описание регулона", *Биофизика*, "Наука", Москва, Российская Федерация, стр. 21-25.

[17] J. Bongaerts, S. Zoske, U. Weidner and G., "Transcriptional regulation of the proton translocating NADH dehydrogenase genes (*nuoA-N*) of *Escherichia coli* by electron acceptors, electron donors and gene regulator", *Mol. Microbiol.*, Blackwell Scientific Publications, England, 1995, pp. 521-534.

A Markov Chain Monte Carlo Technique for Identification of Combinations of Allelic Variants Underlying Complex Diseases

Alexander Favorov
GosNIIGenetika; Johns Hopkins University
favorov@sensi.org

Marina Sudomoina
RSMU
mol-biol@cardio.ru

Olga Favorova
RSMU
mol-biol@cardio.ru

Giovanni Parmigiani
Johns Hopkins University
gp@jhu.edu

Michael F. Ochs
Johns Hopkins University
mfo@jhu.edu

Abstract

We present an algorithm, APSampler, that is intended for identification of combinations of allelic variations that are positively or negatively associated with a disease in case-control studies. The algorithm sequentially proposes and tests potential combinations (patterns) of alleles at different loci. For each locus, a pattern identifies a genotype, or an allele carriership, or no preferred allele (i.e., any allele is equally acceptable). APSampler searches for sets of patterns for which carriers and non-carriers significantly differ by the disease level, using a variation of the Metropolis-Hastings algorithm. The a posteriori probability of each pattern is evaluated using a likelihood obtained from a battery of conditional rank sum statistics that ensures that each pattern is evaluated after removing the effect of other patterns from the same set. The software was applied to data from multiple sclerosis and hypertension studies, which included incomplete genotype data, and confirmed its effectiveness.

1. Introduction

Currently, the number of studies that are focused on the genetic basis of widespread disorders with a complex mode of inheritance, in which multiple genes of possibly small effect are involved, is continually increasing. A methodology to identify the cumulative contribution or genetic interaction of several polymorphous genes would accelerate the understanding of their importance in disease susceptibility and our ability to develop new treatments. A critical problem is the inability of standard statistical approaches, developed for relatively modest data sets, to be powerful in the face of the booming growth in our knowledge of genomics and in the volume of genetic data. Due to the human genome project and the development of high-throughput sequencing and genotyping technologies, there has been a rapid increase in the availability of genetic data for numerous polymorphous loci, including SNPs, repeat polymorphisms and insertions/deletions. This allows the collection of large sets of genetic data, which could be a base for the exploration of the genetic basis of complex diseases. The complexity increases because similar disease-prone phenotypes may be produced by different genes in the same pathways as well as by alternative sets of genes providing disease heterogeneity. The same “combinatorial burst” problem occurs often in other fields, and Bayesian statistical approaches together with Markov chain Monte Carlo (MCMC) [1] techniques have led to significant improvements in understanding.

In recent years, Bayesian approaches have become popular in bioinformatics and statistical genetics, as computational power has increased to a point where these methods can be fully utilized. In addition, there is a substantial body of information on gene locations, potential linkages, and SNPs, which is “vast and plain” and is often easily incorporated in an analysis by Bayesian approaches.

There are numerous recent examples of the application of Bayesian methods in genetics that include population studies, quantitative trait loci mapping, and family-based studies.

2. Algorithm description

We present here an algorithm, APSampler [2], for the exploration of potential combinations of allelic variations positively or negatively associated with a disease or with a phenotype. The algorithm relies on the rank comparison of phenotype for individuals with and without specific combinations of allelic variants (patterns) isolated in genetic backgrounds matched for the remaining significant patterns. It constructs a Markov chain to sample only potentially significant variants, minimizing the potential of large data sets to overwhelm the search.

The type of allelic patterns we seek are of interest in complex genetic diseases and include multiple alleles that are associated with a trait in combination rather than individually. We consider the general situation in which we have, for each individual, both a list of typed alleles at a fixed set of candidate loci and the phenotype of interest. Our method is based on ranks, so the phenotype can be measured as a continuous variable or as an ordinal categorical variable. While quantitative phenotypic measurements are powerful when available, it is useful in many applications to have a more general methodology that only requires comparing individuals to each other, as is the case with ranks.

Our approach is designed to search for correlations between complex genetic patterns and phenotype. These correlations are captured via differences in the distributions of phenotype across two subsets of the population, defined by whether a certain allelic pattern is present or not. We consider a broad range of possible genetic models by allowing every allele to potentially affect the phenotype irrespective of its counterpart on the other chromosome. For example, our approach covers dominant and recessive models, as well as their combinations. When looking for polygenic disease patterns, an important challenge arises from the fact that it is not sufficient to consider candidate patterns one by one, because one pattern may confound the measurement of association for another. Thus, we seek a set of patterns. While we do not consider explicitly the issue of removing the possible influence of environmental factors on the phenotype, such a generalization is possible by modifying the test statistic used to construct the likelihood. The algorithm is able to proceed genotypic data that is partially missing, i.e., data on distinct gene polymorphisms for some cases and controls were unknown.

3. Result verification

To compare allelic distributions in MS patients and controls and to verify the results of the APSampler, we used Fisher's exact test. To avoid noise that appears due to multiple hypotheses testing, each finding was tested by a permutation test. To execute the test, the initial information about disease levels was permuted in a balanced way and reassigned to the genotypes several time. Then, the APSampler was run on each of this permuted dataset and all the "findings" were tested by Fisher. The resulting Fisher significances were collected to a distribution that was used as the null-hypothesis distribution to obtain a p-value for each real (non-permuted) finding of significance.

4. Application

The program was applied to a set of real association studies data. The results of the application of the APSampler to a Russian population multiple sclerosis study is described in [3]. An arterial hypertension and ishemic stroke study on a Yakut population is represented in [4]. A study focused on hyperaldosteronism-associated arterial hypertension in Russian population is described in [5].

5. References

[1] C. Robert, G. Casella Monte Carlo Statistical Methods Springer Verlag, New York, 2003

[2] A.V. Favorov, T.V. Andreewski, M.A. Sudomoina, O.O. Favorova, G. Parmigiani, M.F. Ochs "A Markov chain Monte Carlo technique for identification of combinations of allelic variants underlying complex diseases in humans" *Genetics*, 2005, V. 171(4), pp. 2113-2121

[3] O.O. Favorova, A.V. Favorov, A.N Biko, T.V. Andreewski, M.A. Sudomoina, A.D. Alexeenkov, O.G. Kulakova, E.I. Gusev, G. Parmigiani, M.F. Ochs "Three allele combinations associated with multiple sclerosis" *BMC Med Genet*. 2006 Jul 26, pp. 7-63.

[4] M.A. Sudomoina, T.Y. Nikolaeva, M.G. Parfenov, A.D. Alekseenkov, A.V. Favorov, A.B. Gekht, E.I. Gusev, O.O. Favorova "Genetic risk factors of arterial hypertension: analysis of ischemic stroke patients from the Yakut ethnic group" *Dokl Biochem Biophys*. 2006 Sep-Oct, V. 410, pp. 324-346.

[5] N.M. Chikhladze, Kh.F. Samedova, M.A. Sudomoina, M. Thant, Z.M. Htut, G.N. Litonova, A.V. Favorov, I.E. Chazova, O.O. Favorova "Contribution of CYP11B2, REN and AGT genes in genetic predisposition to arterial hypertension associated with hyperaldosteronism" *Kardiologiia* 2008, V. 48(1), pp. 37-42. Russian.

Analyzing Drug-target Interaction from a Comparative Perspective

Olga V. Kalinina[1,2], Robert B. Russell[2]
[1] *Institute for Information Transmission Problems RAS, Bolshoi Karenty pereulok 19, Moscow, 127994, Russia*
[2] *EMBL-Heidelberg, Meyerhofstr., 1, 69126 Heidelberg, Germany*
kalinina@embl.de

Abstract

Understanding nature of interactions between proteins and low-molecular weight ligands is a key issue in chemoinformatics and drug design. Recent studies indicate that many popular drugs act not on a single but on multiple targets. In this study, we apply molecular modeling techniques to study the possible promiscuity both of small molecules and their target proteins. We compare proteins that bind the same small molecule and small molecules that are bound to each of them, and assess the potential new interactions. This suggests an effective way for prediction of novel interaction partners.

Understanding nature of interactions between proteins and low-molecular weight ligands is a key issue in chemoinformatics and drug design. The increasing amount of data allows to go beyond studying interactions of a single protein with a single small molecule and perform comparative studies in this field. It has been recently shown that drug-target interactions do not follow the simple lock-and-key model, where the keys (drugs) are specific and selective for the locks (target proteins): it is a frequent case when one protein binds different drugs, or one drugs fits many proteins [1]. Complementary, Keiser et al. demonstrated that low molecular weight ligands can be regarded as links between unrelated proteins and allow for prediction of new interactions [2].

In this study, we apply molecular modeling techniques to study the possible promiscuity both of small molecules and their target proteins and aim to predict new potential drugs on a large scale. We use similar small molecules as anchors to superimpose the proteins, to which they are bound, and fit other interaction partner into structures of proteins, which they were never observed to bind. Assessing geometry and chemical properties of the emerging interactions, we can conclude if the considered protein-small molecule pair is likely to interact.

We illustrate our approach by considering binding of phosphodiesterases (PDE) 4 and 5 with their specific inhibitors. These two enzyme families are specific for hydrolysis of cAMP and cGMP, respectively, and development of their specific inhibitors is an issue of great importance in drug industry. Sildenafil is a potent inhibitor of PDE5A, however, it binds PDE4B to a limited extent as well. Using a structure of human PDE4B bound to sildenafil, we fit other PDE5A-specific inhibitors into the pocket of PDE4B. We show that PDE4B is probably unable to bind tadalafil, another specific inhibitor of PDE5A, due to steric clashes with Tyr403, Arg404, Thr407 and Met411 of PDE4B. Experimental assays confirm virtual absence of binding of tadalafil to PDE4B [3].

We benchmark our technique by rediscovering existing interactions: if two divergent enough ligands bind two divergent enough proteins, we exclude one of the four interactions from the dataset and aim to reproduce it. The presented technique is widely applicable to discovery of new drug targets and in assessment of potential toxicity of drugs.

[1] Yildirim MA, Goh K-I, Cusick ME, Barabasi A-L, Vidal M. (2007) Drug-target network. *Nat Biotechnol.* 25(10), 1119-1126

[2] Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK. (2007) Relating protein pharmacology by ligand chemistry. *Nat Biothechnol.* 25(2), 197-206

[3] Card GL, England BP, Suzuki Y, Fong D, Powell B, Lee B, Luu C, Tabrizizad M, Gillette S, Ibrahim PN, Artis DR, Bollag G, Milburn MV, Kim SH, Schlessinger J, Zhang KY (2004) Structural basis for the activity of drugs that inhibit phosphodiesterases. *Structure* 12(12), 2233-2247

Изучение консервативности межгенных участков ДНК прокариот

М.А.Пятницкий
ГУ НИИ биомедицинской химии им В.Н.Ореховича РАМН
mpyat@mail.ru

Резюме

Хорошо известно, что даже некодирующие участки ДНК могут иметь высокие уровни консервативности Предположительно, такое сохранение в процессе эволюции свидетельствует об их функциональности. Для определения наличия действия естественного отбора на участок ДНК эукариот обычно выявляют различия в скорости замен нуклеотидов с другим участком, который считается эволюционирующим нейтрально (интроны, псевдогены). В случае прокариот такой подход невозможен. Целью настоящей работы было изучение характера эволюции и оценка доли консервативных позиций в некодирующих участках ДНК прокариот. С этой целью предложена простая модель эволюции некодирующих участков ДНК прокариот.

1. Введение

Известно, что некоторые участки некодирующей ДНК являются сильно консервативными. Предположительно, такое сохранение в процессе эволюции свидетельствует об их функциональности из-за наличия давления естественного отбора. В последнее время появился ряд работ в которых анализируются различные свойства некодирующих участков ДНК: повторы, периодичность, алгоритмы для поиска некодирующих РНК [1-5].

Для определения наличия действия естественного отбора на участок ДНК стандартным способом является сравнение скоростей замен нуклеотидов с другим участком, который *a priori* считается эволюционирующим нейтрально. Такие участки, в которых скорости замен значимо различаются, предполагаются находящимися под давлением естественного отбора. Основную трудность представляет выбор участка ДНК с нейтральной эволюцией. Для эукариот обычно используют интроны или псевдогены. В случае же прокариот такой подход невозможен.

Целью настоящей работы было изучение характера эволюции и оценка доли консервативных позиций в некодирующих участках ДНК прокариот. С этой целью нами была предложена простая модель эволюции некодирующих участков ДНК.

2. Материалы и методы

Опишем построение модели, используя в качестве примеров штаммы *Escherichia coli* и *Salmonella enterica*. Модель предполагает использование набора штаммов для двух близкородственных видов, т.е. имеющих общего предка. На рис.1 представлено условное филогенетическое дерево видов.

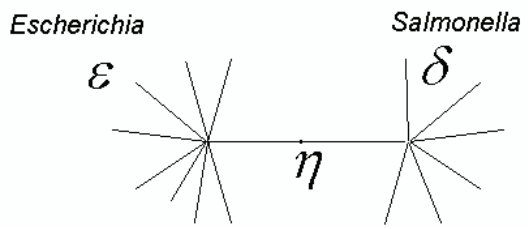


Рисунок 1 Предполагаемое филогенетическое дерево и параметры модели

На рис.2 показано выравнивание участка ДНК для ряда штаммов двух близкородственных видов. Каждую позицию выравнивания можно однозначным образом отнести в один из следующих классов: *CC*, позиция полностью консервативна во всех штаммах в обоих видах; *Cc*, позиция консервативна в каждом наборе штаммов в отдельности, но различается между видами; *CN*, позиция консервативна в штаммах первого вида, но не консервативна во втором виде; *NC*, позиция консервативна в штаммах второго вида, но не консервативна в штаммах первого вида; *NN*, позиция не является консервативной ни в одном из двух видов.

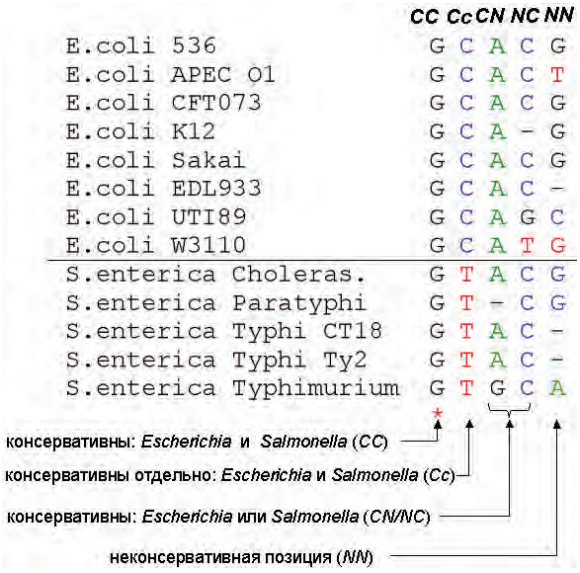


Рисунок 2 Типы позиций в выравнивании на примере *Salmonella* и *Escherichia*

Рассмотрим одну позицию выравнивания участка межгенной ДНК. Введем следующие обозначения (для каждой такой позиции): ω – вероятность для позиции быть нейтрально эволюционирующей, т.е. не находиться под действием отбора; η – вероятность мутирования позиции в общем предке двух видов; ε – вероятность мутирования позиции в первой ветке филогенетического дерева, например в *Escherichia*; δ – вероятность мутирования позиции во второй ветке дерева, например в *Salmonella*.

С учетом этого, общие доли позиций разных классов могут быть выражены через введенные параметры следующим образом:

$$\begin{cases} \frac{CC}{N} = (1-\omega) + \omega(1-\varepsilon)(1-\delta)(1-\eta) \\ \frac{CN}{N} = \omega(1-\varepsilon)\delta \\ \frac{NC}{N} = \omega(1-\delta)\varepsilon \\ \frac{Cc}{N} = \omega(1-\varepsilon)(1-\delta)\eta \\ \frac{NN}{N} = \omega\varepsilon\delta \end{cases} \quad (1)$$

где N – общее число позиций в выравнивании. Система не является переопределенной, поскольку $CC + Cc + NC + CN + NN = N$.

Полученная система уравнений (1) имеет следующее решение:

$$\omega = \frac{1}{N}\frac{(CN+NN)(NC+NN)}{NN}$$
$$\varepsilon = \frac{NN}{CN+NN}$$
$$\delta = \frac{NN}{NC+NN} \quad (2)$$
$$\eta = \frac{Cc \cdot NN}{CN \cdot NC}$$

Таким образом, величина 1 – ω дает искомую долю числа позиций, находящихся под действием естественного отбора.

Вкратце опишем процедуру получения межгенных участков ДНК которые анализировались в данной работе. Полные геномы микроорганизмов загружались с сайта NCBI. На первом этапе проводился поиск всех консервативных пар генов, т.е. пар соседних генов, которые имеют ортологов во всех анализируемых геномах. Для определения ортологичных генов нами был реализован наиболее распространенный алгоритм двунаправленного наилучшего совпадения. Результаты работы такого алгоритма можно представить в виде графа, где вершины (гены из всех изучаемых геномов) соединены ребром в том случае, если они являются наилучшим взаимным совпадением. Поиск клик в таком графе дает наборы взаимных ортологов, которые присутствуют во всех рассматриваемых геномах. В этом смысле полученные наборы генов аналогичны таковым из базы данных COG. Для каждой пары клик были отобраны только те гены, которые сохраняли одинаковую ориентацию во всех изучаемых геномах. Участки ДНК между отфильтрованными таким образом парами генов были выровнены с помощью программы множественного выравнивания MUSCLE.

3. Результаты

Полученные выравнивания межгенных участков ДНК объединялись и рассчитывались величины *CC*, *Cc* ,*CN*, *NC* и *NN*. Результаты расчета согласно (2) и исходные данные приведены в таблице 1. Показаны результаты для различных наборов штаммов и видов. Использовались данные по полным геномам близкородственных видов, для которых секвенировано не менее трех штаммов для каждого вида. Цифра в первой колонке таблицы после названия вида обозначает суммарное количество использованных штаммов.

Как видно из представленных данных, процент позиций, находящихся под действием

отбора достаточно широко варьирует. Для межгенной ДНК *E.coli* доля нуклеотидов находящихся под действием отбора составляет ~45%.

С целью валидации предложенного подхода, нами были рассчитаны те же показатели, но только для известных функциональных участков *E.coli K12* (база данных RegulonDB). Результаты приведены в таблице 2. Как и ожидалось, доля позиций предположительно находящихся под действием отбора больше в случае анализа известных функционально значимых сайтов существенно превосходит величины, рассчитанные для объединения межгенных участков и составляет ~60% для промоторов и сайтов связывания транскрипционных факторов. Для терминаторов эта доля мало отличается от средней величины и составляет ~48%.

Другим типом анализа явилось сравнение доли позиций находящихся под действием отбора для разных типов межгенных участков в зависимости от ориентации соседних генов (данные не приведены).

Тем не менее, предложенная модель имеет недостатки – при некоторых значениях исходных данных значения параметра η получаются больше единицы. В настоящий момент работа по устранению причин подобных оценок продолжается.

Как альтернатива предложенному подходу, в настоящее время проводится работа по разбиению позиций на несколько классов по скоростям эволюции с использованием метода максимального правдоподобия (используется пакет PAML).

Таблица 1 Результаты для всех ортологичных межгенных участков ДНК

Геномы	ω	ε	δ	η	CC	Cc	CN	NC	NN
Escherichia+salmonella 13	0.54	0.42	0.45	0.75	50117	12983	13956	12684	10211
Escherichia+salmonella 17	0.59	0.41	0.63	0.74	25034	5312	12303	5052	8689
Escherichia+salmonella 23	0.55	0.64	0.69	1.39	17186	3350	5396	4364	9793
Streptococcus pyogenes+agalactiae 15	0.38	0.60	0.06	1.49	15829	6189	268	6150	398
Streptococcus pyogenes+pneumomiae 15	0.61	0.34	0.14	0.66	5826	2635	664	2015	337
Clostridridia perfringens+botulinum 7	0.82	0.03	0.06	0.27	22131	9151	2130	1182	75
Chlamydia trachomatis+pneumoniae 8	0.92	0.03	0.31	0.37	14638	7224	8649	538	240
Burkholderia mallei+pseudomallei 8	0.40	0.64	0.52	0.16	99441	1622	11034	18330	19708

Таблица 2 *Результаты для функционально важных межгенных участков ДНК E.coli K12*

Геномы	ω	ε	δ	η	CC	Cc	CN	NC	NN
Промоторы:									
Escherichia+salmonella 13	0.40	0.36	0.55	0.99	5460	1037	1261	586	708
Escherichia+salmonella 17	0.45	0.33	0.66	0.92	2614	435	917	237	461
Escherichia+salmonella 23	0.40	0.61	0.64	*1.54*	2019	314	358	317	557
Сайты связывания:									
Escherichia+salmonella13	0.35	0.29	0.58	*1.25*	5189	1075	1182	353	484
Escherichia+salmonella 17	0.48	0.29	0.69	0.85	2232	376	991	180	404
Escherichia+salmonella 23	0.40	0.65	0.75	*2.49*	1327	215	253	162	474
Терминаторы:									
Escherichia+salmonella13	0.48	0.15	0.33	0.68	228	69	51	18	9
Escherichia+salmonella 17	0.57	0.20	0.65	0.54	116	20	70	9	17
Escherichia+salmonella 23	0.52	0.25	0.60	0.77	100	23	45	10	15

4. Реализация

Все программное обеспечение, которое использовалось в данной работе, было реализовано как набор платформенно-независимых скриптов на языках Perl и R. Дополнительно использовался набор пакетов расширения BioPerl. Также применялась система управления базами данных MySQL 5.0. Все исходные коды скриптов доступны по запросу к авторам.

5. Благодарности

Эта работа была проведена совместно с д.б н. М.С.Гельфандом. Мы благодарны к.б н. Г.А.Базыкину за плодотворные обсуждения.

6. Ссылки

[1] N.Rajewsky, N.D. Socci, M.Zapotocky, and E.D. Siqqia, “The Evolution of DNA Regulatory Regions for Proteo-Gamma Bacteria by Interspecies Comparisions”, *Genome Res*, 12, 2002, pp. 298-308.

[2] L.Wilson and P.Sharp, “Enterobacterial repetitive intergenic consensus (ERIC) sequences in Escherichia coli: Evolution and implications for ERIC-PCR”, *Mol Biol Evol.*, 23(6), 2006, pp. 1156-1168.

[3] S.Hosid, E.N.Trifonov and A.Bolshoy, “Sequence periodicity of E.coli is concentrated in intergenic regions”, *BMC Mol Biol.*, 5, 2004, pp. 1-14.

[4] A.Siepel, G.BejeranoJ.Pedersen et al., “Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes”, *Genome Res*, 15, 2005, pp. 1034-1050

[5] N.Molina, E. van Nimwegen, “Universal patterns of purifying selection at noncoding positions in bacteria.” *Genome Res*, 2008 Jan;18(1), pp 148-60

Эволюция под отбором: исследование паттерна замен в сайтах сплайсинга

Степан Денисов[1], Андрей Миронов[1,2], Михаил Гельфанд[2]
[1] *Факультет биоинженерии и биоинформатики МГУ им. М.В. Ломоносова,*
[2] *ИППИ РАН им. А.А. Харкевича*
stepan@bioinf.fbb.msu.ru

Аннотация

Целью работы было изучить события, происходившие в сайтах сплайсинга начиная от расхождения ветвей человека (H. sapiens), мыши (M. musculus) и собаки (C. familiaris) до настоящего времени отдельно для каждой ветви. Исследуемая выборка содержала ~30000 выравниваний ортологичных донорых сайтов из трех организмов и столько же выравниваний акцепторных сайтов. Для каждой позиции в выравнивании и для каждой ветви строилась матрица замен и исследовались ее свойства. Было показано, что вектора частот нуклеотидов во многих позициях сайтов, наблюдаемые в настоящий момент в геномах этих трех видов существенно отличаются от стационарных векторов соответствующих матриц замен. Мы обнаружили, что зачастую наблюдается уменьшение частоты "мажорных" букв (т.е. букв, частота которых наибольшая в данной позиции). В полипиримидиновом тракте акцепторного сайте происходит "размывание" буквы T, уменьшается частота C в позиции -3. A в донорном сайте наблюдается "размывание" A(+3), A(+4), G(+5) и G(-1). Мы также показали, что имеются различия в эволюции на разных ветвях и в различных группах сайтов, однако общие тенденции сохраняются.

1. Введение

Сплайсинг пре-мРНК является важным этапом экспрессии генов эукариотических организмов. Суть процесса заключается в том, что из первичного транскрипта вырезаются куски, называемые интронами, а оставшиеся фрагменты – экзоны – сшиваются. Реакция сплайсига осуществляется специальным РНК-белковым комплексом – сплайсосомой. На границах экзонов и интронов имеются специальные последовательности (сайты сплайсинга), которые распознаются сплайсосомой и в результате сплайсинг происходит по четко определенным позициям. Однако следует отметить, что на сплайсинг влияют также дополнительные последовательности, находящиеся в окрестности сайтов сплайсинга, поэтому наличие подходящих сайтов является необходимым, но не достаточным условием сплайсинга данного интрона.

Различают донорный (5') и акцепторный (3') сайты (см. рис. 1). Сайты сплайсинга сходны в пределах одного генома (а также между геномами), однако в некоторых позициях имеются существенные отличия. Каждый из сайтов содержит практически инвариантный динуклеотид (GU и AG у донора и акцептора, соответственно). Остальные позиции в большей или меньшей степени варьируют. На рис. 1 изображены консенсусные сайты сплайсинга, полученные на основе анализа 1683 интронов человека [1].

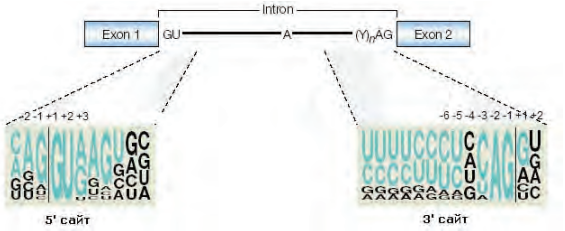


Рис.1. Сайты сплайсинга. Изображен фрагмент пре-мРНК: два экзона и интрон между ними. На границах экзонов и интрона расположены донорный (5') и акцепторный (3') сайты сплайсинга. Ниже изображены консенсусные последовательности сайтов сплайсинга. Высота букв в каждой позиции пропорциональна частоте их встречаемости. Также обозначены номера позиций в сайтах. Из [1] с изменениями.

В связи с тем, что сайты сплайсинга несут важную биологическую нагрузку, мутации в них

ограничены в связи с действием отбора. Целью данной работы было исследовать замены в сайтах сплайсинга начиная от расхождения ветвей человека (*H. sapiens*), мыши (*M. musculus*) и собаки (*C. familiaris*) до настоящего времени отдельно для каждой ветви. Нас также интересовал вопрос влияет ли тип альтернативного/конститутивного сплайсинга на эволюцию соответствующий сайтов.

2. Материалы и методы

Исследуемая выборка содержала 31881 выравнивание ортологичных донорых сайтов из трех организмов и столько же выравниваний акцепторных сайтов. Длина последовательностей в выравнивании составляла 30 нт (собственно сайты и фланкирующие последовательности: позиции -10..+20 для 5' сайта и -20..+10 для 3' сайта). Вся выборка была поделена на несколько подвыборок в зависимости от типа альтернативного/конститутивного сплайсинга, происходящего в данном месте генома у человека. А именно: (1) конститутивные экзоны, (2) кассетные экзоны, (3) внутренние альтернативные донорные сайты, (4) внешние альтернативные донорные сайты, (5) внутренние альтернативные акцепторные сайты, (6) внешние альтернативные акцепторные сайты, (7) удержанные интроны. Все описываемые ниже исследования производились как с общей выборкой, так и с каждой из подвыборок.

Для каждой позиции выравнивания восстановливалась буква, которая наиболее вероятно присутствовала у общего предка человека, мыши и собаки. В связи с тем, что филогенетическое дерево данных трех организмов является плохоразрешенным, наиболее целесообразной процедурой является “голосование”: если буквы в двух или трех организмах совпадают, то предку присваивается такая же буква. Если все буквы разные, то данная позиция игнорируется и не входит в итоговую статистику. Именно такая процедура и была применена. Таким образом, зная предковый нуклеотид, мы можем исследовать, какие замены происходили в сайтах начиная от расхождения ветвей человека, мыши и собаки до настоящего времени отдельно для каждой ветви.

Далее для каждой ветви отдельно и для каждой позиции в сайте сплайсинга строилась матрица замен размером 4x4. Столбцы – это предковые буквы, а строки – это буквы в потомке. Каждый элемент матрицы равняется количеству замен данного типа (например, A->C), произошедших в данной позиции сайта суммарно у всех сайтов данной выборки или подвыборки.

Далее мы исследовали свойства получившихся матриц

Однако было замечено, что доля выравниваний с различными буквами во всех трех организмах в отдельно взятой позиции заметна (~3-10%) и, возможно, игнорирование таких позиций может повлиять на результат. Скажем, если у человека, мыши и собаки в какой-то позиции выравнивания находятся буквы A, C и G, соответственно, то это есть свидетельство того, что в предке, скорее всего, нет буквы T. Эта информация не учитывается в описанной выше методике. Чтобы попытаться это учесть, мы применили следующую процедуру. Если две или три буквы совпадают, то предок восстанавливается прежним способом. Если же три буквы разные, то предковая буква выбирается случайным образом из этих трех. Эта процедура повторяется 1000 раз. В результате для каждого параметра, посчитанного на основе данной матрицы в данной позиции (вектор предковых частот $\boldsymbol{a}$, вектор частот в потомке $\boldsymbol{d}$, стационарный вектор $\boldsymbol{s}$ и др.– см. ниже) вычислялось среднее и выборочное среднеквадратичное отклонение. Ниже значения параметров будут приводиться в виде (среднее) ± (выборочное среднеквадратичное отклонение).

3. Результаты и обсуждение

Первое, что было замечено, что многие из полученных матриц существенно несимметричны относительно главной диагонали, что говорит о том, что число прямых и обратных замен между нуклеотидами значительно различается. Причем это наблюдается именно внутри сайтов сплайсинга, а не во фланкирующих последовательностях, что есть свидетельство направленного действия отбора. Однако, тот факт, что количеством прямых замен (например, A->C) больше, чем число обратных замен (C->A) не означает, что в данном примере частота буквы C будет увеличиваться, потому что существуют также другие замены. Поэтому необходимо разработать метод, который позволяет понять эволюционные тенденции с учетом всех замен.

Исходные матрицы замен были нормированы таким образом, чтобы сумма элементов по столбцам (столбцы – это предковые нуклеотиды) равнялась единице. Тогда полученная матрица (M) обладает следующим свойством:

$$M \times a = d \qquad (1)$$

где $\boldsymbol{a}$ – вектор-столбец частот нуклеотидов в данной позиции в предке (***a****ncestor*),

$\boldsymbol{d}$ - вектор-столбец частот нуклеотидов в данной позиции в потомке (***d****escendant*).

Разница между частотами нуклеотидов в предке и потомке в большинстве случаев

невелика. В большинстве позиций значения элементов вектора ***(d – a)*** по модулю обычно варьируют от 0,001 до 0,003 (для общей выборки). Однако, в отдельно взятых позициях эта разница более существенна, но не превышает по модулю 0,02. В таблице 1 приведены по 5 букв для общей выборки доноров и общей выборки акцепторов с наибольшими и наименьшими значениями разницы частот.

Нетрудно видеть, что убывание происходит внутри непосредственно донорного сайта, а наибольшее возрастание скорее во фланкирующих последовательностях. У акцепторных сайтов можно заметить уменьшение частоты нуклеотида T в полипиримидиновом тракте (позиции -12, -13, -14, -16).

Рассмотрим матрицу замен ***M***. Она описывает эволюцию одной позиции сайта на одной эволюционной ветке. Принимая во внимание тот факт, что вектора ***d*** и ***a*** отличаются не сильно,

Таблица 1. Буквы, частота которых наиболее существенно изменилась между общим предком и потомком (человек, мышь или собака).

	Организм	№ позиции	Буква	Разница частот $(d_i - a_i)$
Донорные сайты				
Увеличение частоты	Мышь	+11	C	0.0128±0.0007
	Мышь	+6	C	0.0121±0.0004
	Человек	+20	A	0.0119±0.0009
	Человек	+15	A	0.0117±0.0008
	Мышь	+16	C	0.0113±0.0008
Уменьшение частоты	Собака	+4	A	-0.0186±0.0004
	Мышь	-1	G	-0.0168±0.0002
	Собака	+5	G	-0.0164±0.0003
	Собака	-1	G	-0.0139±0.0002
	Мышь	+11	T	-0.0117±0.0007
Акцепторные сайты				
Увеличение частоты	Человек	-3	T	0.0202±0.0002
	Мышь	-3	T	0.0160±0.0002
	Мышь	-17	G	0.0144±0.0008
	Собака	-4	G	0.0137±0.0005
	Собака	-3	T	0.0135±0.0002
Уменьшение частоты	Собака	-14	T	-0.0228±0.0007
	Собака	-12	T	-0.0224±0.0005
	Мышь	-3	C	-0.0200±0.0006
	Собака	-13	T	-0.0194±0.0007
	Собака	-16	T	-0.0194±0.0007

представляется интересным для данной матрицы вычислить стационарный вектор частот ***s***, т.е. вектор, удовлетворяющий следующему соотношению:

$$M \times s = s \qquad (2)$$

Для каждой матрицы был получен стационарный вектор. Оказалось, что стационарные частоты букв практически в любой позиции существенно отличаются как от наблюдаемых частот (т.е. частот в потомке), так и от частот в общем предке (см. таблицу 2). Это довольно неожиданное наблюдение можно интерпретировать следующим образом. Во-первых, в настоящее время в сайтах сплайсинга человека, мыши и собаки не достигнуто состояние эволюционного равновесия, т.е. в сайтах сплайсинга и сейчас продолжает идти направленная эволюция под действием отбора. Во-вторых, сравнение векторов ***s*** и ***d*** дает возможность контрастно выявить эволюционные тенденции, которые наблюдались на ветвях человека, мыши и собаки и наблюдаются в настоящий момент. Стоит, однако, отметить, что не стоит интерпретировать стационарные частоты, как те частоты, которые будут достигнуты когда-то в будущем. Это было бы верно, если матрица ***M*** осталась неизменной на протяжении долгого периода эволюции. И, конечно, это может быть совсем не так. В таблице 2 приведены значения разницы между стационарными частотами и частотами в потомке (т.е. наблюдаемыми сейчас) для некоторых позиций (данные на общей выборке). Как можно видеть из таблицы зачастую наблюдается тенденция к уменьшению частоты "мажорных" букв (т.е. букв, частота которых наибольшая в данной позиции, см. рис.1). Это происходит с нуклеотидами A(+4), G(+5) в донорных сайтах, причем параллельно в трех ветках. Видно, что уменьшение частоты данных букв сопровождается более или менее равномерным возрастанием частоты остальных букв, т.е. имеет место “размывание” нуклеотидов. Однако не все тенденции являются одинаковыми для всех веток. Так у мыши и собаки очень ярко выражено уменьшение частоты G(-1), а у человека наоборот - частота G возрастает. В данном случае уменьшение частоты G сопровождается увеличением частот A и T, а уровень нуклеотида C остается прежним. На линии человека увеличение G в наибольшей степени скомпенсировано уменьшением T. Также легко видеть “размывание” A(+3), характерное для собаки и мыши и не свойственное человеку. Что касается акцепторного сайта, то там можно внаблюдать четкую тенденцию к размыванию буквы T на протяжении всего полипиримидинового тракта (в таблице 2 в качестве примера показаны позиции -10, -11 и -12). Сильно уменьшается доля буквы “мажорной” буквы C в позиции -3. Причем у человека и собаки параллельно существенно возрастает частота T.

Как видно из рассмотренных примеров, в эволюции сайтов сплайсинга имеются четкие

тенденции, которые зачастую параллельно проявляются в трех различных эволюционных ветках. Однако, у каждой ветви есть свои особенности. Эволюция сайтов отличается также и у различных групп сайтов в зависимости от типа сплайсинга. Общее направление эволюции во всех группах также сохраняется. Варьирует степень выраженности генеральных тенденций, а также имеется много мелких отличий. В целом, различия в эволюции между разными ветвями представляется более выраженными, чем между разными группами сайтов в пределах одной ветви.

Высокая степень консервативности донорных сайтов сплайсинга и наличие мажорных букв в консенсусе определяется в основном Уотсон-Криковскими и воббл-взаимодействиями между сайтом и U1 мяРНК – ключевым компонентом сплайсосомы [2, 3]. Кроме того, важную роль играет взаимодействие с U6 мяРНК [2]. Размывание “мажорных” букв говорит о том, что сила связывания с U1 ослабевает в течение эволюции млекопитающих. С одной стороны это может быть следствием многофакторного отбора: взаимодействие с U6 накладывает существенно другие ограничения, чем взаимодействие с U1, кроме того во взаимодействии принимают участие также белки сплайсосомы. Что качается экзонной части, то здесь также имеются ограничения, связанные с кодированием белковой последовательности, зависящие от рамки считывания [4]. С другой стороны, уменьшение силы взаимодействия с U1 может приводить к возрастанию роли дополнительных мотивов в окрестности сайтов сплайсинга [5]. И это, вероятно, делает сплайсинг более регулируемым, и более эволюционно лабильным, что может быть серьезным преимуществом. Стоит отметить, что было показано также, что наличие тех или иных нуклеотидов в различных позициях донорных сайтов скореллировано между собой [4]. Наличие подобных корреляций может также довольно неоднозначно влиять на результат. Аналогичные тенденции прослеживаются и в акцепторном сайте. Убывание частоты T в полипиримидиновом тракте наблюдается равномерно вдоль всего тракта, а не в каких-то отдельных позициях, что, видимо, связано с особенностями отбора в данном участке: важно наличие определенной доли пиримидинов в данном районе, а не конкретная их локализация. В отличие от донорного сайта, где особенности отбора определяются в основном взаимодействием с мяРНК, с полипиримидиновым трактом взаимодействует малая субъединица U2AF – белок.

Таблица 2. Разница между стационарными и наблюдаемыми частотами.

Позиция	Буква	*(s - d)* (человек)	*(s - d)* (мышь)	*(s - d)* (собака)
Донорные сайты				
-2	A	0,017 ± 0,006	-0,061 ± 0,002	**-0,159 ± 0,004**
-2	C	-0,003 ± 0,003	0,038 ± 0,002	0,065 ± 0,004
-2	G	-0,005 ± 0,003	0,023 ± 0,002	**0,079 ± 0,004**
-2	T	-0,008 ± 0,004	0,001 ± 0,002	0,016 ± 0,003
-1	A	-0,019 ± 0,002	**0,072 ± 0,003**	**0,112 ± 0,004**
-1	C	-0,012 ± 0,001	0,036 ± 0,002	0,051 ± 0,004
-1	G	**0,077 ± 0,003**	**-0,232 ± 0,004**	**-0,225 ± 0,004**
-1	T	**-0,046 ± 0,001**	**0,124 ± 0,004**	**0,062 ± 0,004**
+3	A	**0,027 ± 0,002**	**-0,027 ± 0,002**	**-0,076 ± 0,003**
+3	C	-0,016 ± 0,001	0,023 ± 0,001	0,025 ± 0,002
+3	G	0,005 ± 0,002	-0,021 ± 0,001	0,012 ± 0,002
+3	T	-0,016 ± 0,001	0,025 ± 0,001	0,039 ± 0,003
+4	A	**-0,034 ± 0,004**	**-0,072 ± 0,002**	**-0,169 ± 0,003**
+4	C	0,016 ± 0,002	0,033 ± 0,001	0,053 ± 0,002
+4	G	0,015 ± 0,003	0,008 ± 0,001	0,053 ± 0,002
+4	T	0,003 ± 0,002	0,031 ± 0,001	0,064 ± 0,003
+5	A	0,026 ± 0,003	0,005 ± 0,001	0,047 ± 0,002
+5	C	0,001 ± 0,001	0,025 ± 0,001	0,044 ± 0,002
+5	G	**-0,030 ± 0,003**	**-0,045 ± 0,002**	**-0,138 ± 0,002**
+5	T	0,002 ± 0,002	0,014 ± 0,001	0,047 ± 0,002
+6	A	0,029 ± 0,003	-0,015 ± 0,001	0,007 ± 0,002
+6	C	0,010 ± 0,002	0,020 ± 0,001	0,037 ± 0,002
+6	G	-0,001 ± 0,002	-0,014 ± 0,001	-0,003 ± 0,002
+6	T	-0,038 ± 0,004	0,009 ± 0,002	-0,041 ± 0,003
Акцепторные сайты				
-12	A	0,010 ± 0,002	0,007 ± 0,001	0,010 ± 0,001
-12	C	0,026 ± 0,002	0,001 ± 0,001	0,029 ± 0,002
-12	G	0,011 ± 0,002	0,016 ± 0,001	0,025 ± 0,002
-12	T	**-0,047 ± 0,002**	**-0,024 ± 0,001**	**-0,064 ± 0,001**
-11	A	0,020 ± 0,002	0,004 ± 0,001	0,016 ± 0,001
-11	C	0,016 ± 0,002	0,007 ± 0,001	0,021 ± 0,001
-11	G	0,021 ± 0,002	0,013 ± 0,001	0,022 ± 0,002
-11	T	**-0,057 ± 0,002**	**-0,025 ± 0,001**	**-0,059 ± 0,001**
-10	A	0,011 ± 0,002	0,002 ± 0,001	0,013 ± 0,001
-10	C	0,012 ± 0,002	-0,007 ± 0,001	0,030 ± 0,001
-10	G	0,004 ± 0,002	0,013 ± 0,001	0,020 ± 0,001
-10	T	**-0,027 ± 0,002**	-0,008 ± 0,001	**-0,062 ± 0,002**
-4	A	-0,003 ± 0,002	0,031 ± 0,001	0,009 ± 0,002
-4	C	-0,036 ± 0,002	-0,033 ± 0,001	-0,013 ± 0,002
-4	G	0,028 ± 0,003	0,026 ± 0,001	**0,046 ± 0,002**
-4	T	0,011± 0,003	-0,024 ± 0,001	-0,042 ± 0,002
-3	A	-0,019 ± 0,002	0,026 ± 0,001	0,030 ± 0,003
-3	C	**-0,103 ± 0,002**	**-0,063 ± 0,001**	**-0,090 ± 0,001**
-3	G	-0,005 ± 0,000	0,007 ± 0,001	0,007 ± 0,001
-3	T	**0,127 ± 0,002**	0,030 ± 0,001	**0,054 ± 0,002**
+1	A	**0,071 ± 0,006**	0,033 ± 0,003	0,008 ± 0,004
+1	C	-0,022 ± 0,004	0,014 ± 0,003	0,046 ± 0,005
+1	G	-0,040 ± 0,005	-0,066 ± 0,003	-0,081 ± 0,005
+1	T	-0,009 ± 0,004	0,019 ± 0,003	0,027 ± 0,004
+2	A	0,027 ± 0,006	0,003 ± 0,003	0,022 ± 0,005
+2	C	-0,006 ± 0,004	0,028 ± 0,002	0,019 ± 0,003
+2	G	-0,004 ± 0,004	-0,002 ± 0,002	-0,007 ± 0,003
+2	T	-0,017 ± 0,006	-0,029 ± 0,002	-0,034 ± 0,004

И это возможно влияет на вышеупомянутые особенности отбора в полипиримидиновом тракте.

Авторы выражают благодарность Р. Сутормину и Д. Первушину за полезные осуждения.

12. Литература

[1] L. Cartegni et al., “Listening to silence and understanding nonsense: exonic mutations that affect splicing” *Nature Rev*., Nature Pub. Group, UK, 2002, 3: 285-298.

[2] J. Rojers and R. Wall, ”A mechanism for RNA splicing” *Proc Natl Acad Sci USA*, National Academy of Sciences, USA, 1980, 77(4): 1877–1879.

[3] D.A. Wasserman and J.A. Steitz, “Interactions of small nuclear RNA's with precursor messenger RNA during in vitro splicing” *Science*, American Association for the Advancement of Science, USA, 1992, 257(5078):1918-25.

[4] I. Carmel , S. Tal, I. Vig, G. Ast, “Comparative analysis detects dependencies among the 5' splice-site positions” *RNA*, Cold Spring Harbor Laboratory Press, USA, 2004, 10(5):828-40.

[5] X. Roca, R. Sachidanandam, A.R. Krainer, “Determinants of the inherent strength of human 5' splice sites” *RNA*, Cold Spring Harbor Laboratory Press, USA, 2005, 11(5):683-98.

Фиксация точечных мутаций кодирующих участках генов плодовых мушек рода *Drosophila*

Е.О. Ермакова
ИППИ им. А.А. Харкевича РАН
ermakova@iitp.ru

Д.Б. Малько
ГосНИИгенетика
malko@genetika.ru

М.С. Гельфанд
ИППИ им. А.А. Харкевича РАН
gelfand@iitp.ru

Аннотация

У многоклеточных эукариот альтернативный сплайсинг — один из основных механизмов создания разнообразия белковых последовательностей. Сравнительный анализ гомологичных генов плодовой мушки Drosophila melanogaster и 10 других видов рода Drosophila, геномы которых находятся на различных эволюционных расстояниях от генома D. melanogaster (D. sechellia, D. yakuba, D. erecta, D. ananassae, D. pseudoobscura, D. persimilis, D. willistoni, D. mojavensis, D. virilis, D. grimshawi) показал, что в альтернативно сплайсируемых кодирующих участках плотность синонимичных замен ниже, а плотность замен, изменяющих аминокислоту в кодируемом белке, выше, чем в участках, всегда присутствующих в зрелом транскрипте. Скорость фиксации замен нуклеотидов в альтернативных участках различных типов сильно варьируется. Показано повышение интенсивности фиксации точечных мутаций в окрестностях вставок интронов.

1. Введение

Альтернативный сплайсинг у многоклеточных эукариот является одним из основных механизмов создания разнообразия белковых последовательностей. В отличие от дупликации целых генов, альтернативный сплайсинг способствует созданию белков, отличающихся только одним специальным участком и идентичных в остальной части аминокислотной последовательности. В данной работе исследована Фиксация точечных мутаций в альтернативных и постоянных участках генов, а также взаимодействие точечных мутаций с другим эволюционным фактором, вставкой и потерей интронов, на примере 11 полных геномов плодовых мушек рода *Drosophila (D. melanogaster, (D. sechellia, D. yakuba, D. erecta, D. ananassae, D. pseudoobscura, D. persimilis, D. willistoni, D. mojavensis, D. virilis, D. grimshawi).*

2. Материалы и методы

2.1 Поиск гомологичных генов и геномные выравнивания

Источники последовательностей полных геномов рассмотренных видов дрозофилы показаны в таблице 1. Мы использовали общепринятую топологию филогенетического дерева рассмотренных видов (рисунок 1), длины веток были оценены при помощи программы codeml из пакета PAML [1]. Сокращение названия каждого вида было построено из первой буквы названия рода и первых трёх букв названия вида. Далее Dnnn мы будем называть любой из рассмотренных видов, отличных от Dmel.

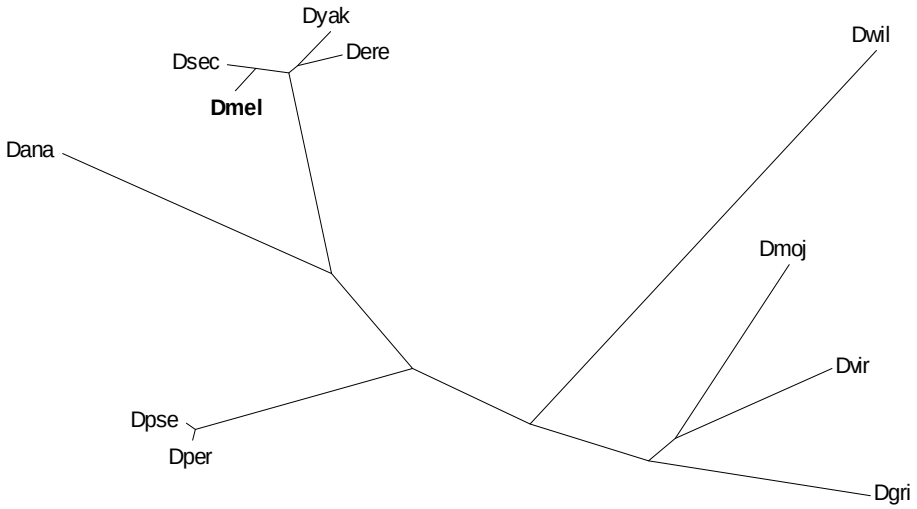
Таблица 1. Версии и источники полных геномов

Dana	droAna3, hgdownload.cse.ucsc.edu
Dere	droEre2, hgdownload.cse.ucsc.edu
Dgri	droGri2, hgdownload.cse.ucsc.edu
Dmel	5.2, ftp.flybase.net
Dmoj	droMoj3, hgdownload.cse.ucsc.edu
Dper	сборка 28 октября 2005, Институт Броуда
Dpse	2.0, ftp.flybase.net
Dsec	сборка 28 октября 2005, Институт Броуда
Dvir	droVir3, hgdownload.cse.ucsc.edu
Dwil	droWil1, hgdownload.cse.ucsc.edu
Dyak	droYak2, Вашингтонский университет

Рассматривались только гены Dmel, имеюшие гомологи во всех 10 геномах Dnnn. Поиск гомологов и выравнивание кодирующих участков

Рисунок 1. Филогенетическое дерево рассмотренных видов

Длины веток в формате Newick: (Dsec:0.05144, ((Dyak:0.08565, Dere:0.08320):0.02159, (Dana:0.52813, ((Dpse:0.01488, Dper:0.02113):0.40434, (Dwil:0.91236, ((Dmoj:0.37285, Dvir:0.30892):0.06326, Dgri:0.40364):0.22321):0.23411):0.22494):0.36614):0.06060, Dmel:0.05592);



геномных последовательностей были проведены при помощи программ BLAT [2] и Pro-Frame [3]. Сначала были найдены ближайшие гомологи генов Dmel из остальных геномов и построены попарные выравнивания кодирующих участков Dmel-Dnnn. Множественные выравнивание были построены формальным отождествлением последовательностей Dmel в попарных выравниваниях.

2.2 Аннотация альтернативного сплайсинга

Данные о сплайсинге в генах *D. melanogaster* были взяты из базы данных FlyBase [4]. Классификация кодирующих участков была проведена согласно этой аннотации.

Кроме деления кодирующих участков на однозначно сплайсируемые (постоянные экзоны и постоянные части кодирующих экзонов) и альтернативные, были выделены отдельные виды альтернативных участков: кассетные экзоны, гомологичные (предположительно, образовавшиеся за счёт дупликации) и негомологичные чередующиеся экзоны, удержанные интроны, удлинения экзона за счёт альтернативного донорного либо акцепторного сайта сплайсинга. Кодирующие участки одноэкзонных генов были выделены в отдельный тип.

2.3 Моделирование молекулярной эволюции

Для оценки плотности несинонимичных (т. е. меняющих аминокислоту) нуклеотидных замен в несинонимичных позициях (d_N) и синонимичных нуклеотидных замен в синонимичную позициях (d_S) был использован метод наибольшего правдоподобия [5], реализованный в программном пакете PAML [1] (использовалась версия 4). Множественные выравнивания кодирующих участков каждого из рассмотренных типов конкатенировались для получения средних оценок по геному. Количество нуклеотидных замен, произошедших на разных ветвях филогенетического дерева в отождествлённых участках, суммировалось.

3. Результаты и обсуждение

3.1 Фиксация нуклеотидных замен в альтернативных и постоянных кодирующих участках генов

Суммарная плотность синонимичных и несинонимичных замен и их отношение для различных типов кодирующих участках показаны в таблице 2.

Таблица 2. Фиксация нуклеотидных замен в кодирующих участках генов

тип кодирующих участков	d_N	d_S	d_N/d_S	длина конкатената в кодонах
одноэкзонные гены	0,434	5,369	0,081	187269
постоянные экзоны многоэкзонных генов	0,338	4,993	0,068	1964307
постоянные части альтернативных экзонов	0,368	4,929	0,075	31436
все альтернативные участки	0,498	3,489	0,143	16228
гомологичные чередующиеся экзоны	0,300	2,767	0,108	2371
негомологичные чередующиеся экзоны	0,308	1,883	0,164	573
кассетные экзоны	0,551	3,709	0,149	7519
удлинения за счёт донорного сайта сплайсинга	0,397	3,715	0,107	1068
удлинения за счёт акцепторного сайта сплайсинга	0,770	4,062	0,189	2124
удержанные интроны	0,519	5,236	0,099	1536

Таблица 3. Фиксация нуклеотидных замен в районах вставок и потерь интронов

тип кодирующих участков	d_N	d_S	d_N/d_S	длина конкатената в кодонах
кодирующие области генов, в которые вставляются интроны	0,383	5,120	0,075	426756
экзоны, в которые вставляются интроны	0,448	4,967	0,090	209185
кодирующие области генов, теряющих интроны	0,339	5,274	0,064	451042
экзоны, граничащие с 5' концом потерянного интрона	0,350	5,235	0,067	107376
экзоны, граничащие с 3' концом потерянного интрона	0,337	5,555	0,061	112626

В альтернативных участках плотность несинонимичных замен выше, чем в однозначно сплайсируемых (что, по-видимому, связано с более низкой интенсивностью экспрессии этих участков), а плотность синонимичных замен – ниже, чем в однозначно сплайсируемых (что, по-видимому, связано с большей плотностью *cis*-регуляторных элементов в альтернативных участках). Постоянные части альтернативных экзонов по количеству замен занимают промежуточное положение между постоянными экзонами и альтернативными участками. В кодирующих участках одноэкзонных генов фиксируется больше нуклеотидных замен, чем в постоянных участках многоэкзонных генов.

Более того, плотность нуклеотидных замен в альтернативных участках разных типов сильно отличается. Меньше всего нуклеотидных замен фиксируется в негомологичных чередующихся экзонах (даже меньше, чем в постоянных экзонах). В кассетных экзонах фиксируется вдвое больше несинонимичных замен, чем в чередующихся экзонах. Больше всего несинонимичных замен фиксируется в удлинениях акцепторного сайта сплайсинга, а больше всего синонимичных – в удержанных интронах. По-видимому, эти различия связаны с различиями в механизмах регуляции, точнее, с преимущественным расположением *cis*-регуляторных элементов относительно альтернативного кодирующего участка (вне или внутри него).

3.2 Фиксация нуклеотидных замен в районах вставки и потери интронов

Гены, потерявшие или приобретшие интроны в процессе эволюции рода *Drosophila*, рассматривались отдельно. Исследовались как гены в целом, так и отдельно экзоны, в которые вставился интрон, и экзоны, граничащие со вставленным интроном.

Окрестности потерянных интронов эволюционируют так же, как все гены в среднем, тогда как в экзонах-мишенях вставки интронов плотность нуклеотидных замен выше, чем в среднем по геному (таблица 3).

4. Благодарности

Мы благодарим И.И. Артамонову, Д.В. Виноградова, А.А. Миронова и Р.Н. Нуртдинова за плодотворные обсуждения и техническую поддержку.

5. Список литературы

[1] Z. Yang, “PAML: a program package for phylogenetic analysis by maximum likelihood,” *Comput Appl Biosci.*, Oxford University Press, England, 1997, pp. 555-556.

[2] W.J. Kent, “BLAT — the BLAST-like alignment tool,” *Genome Res.*, Cold Spring Harbor Laboratory Press, United States, 2002, pp. 656-664.

[3] A.A. Mironov, P.S. Novichkov, M.S. Gelfand, “Pro-Frame: similarity-based gene recognition in eukaryotic DNA sequences with errors,” Bioinformatics, 2001, pp. 13-15

[4] R.J. Wilson, J.L. Goodman, V.B. Strelets, and the FlyBase Consortium, “FlyBase: integration and improvements to query tools,” *Nucleic Acids Res.*, Oxford University Press, England, 2008 D588-D593; doi:10.1093/nar/gkm930

[5] N. Goldman, and Z. Yang, “A codon-based model of nucleotide substitution for protein-coding DNA sequences,” *Mol Biol Evol.*, Oxford University Press, United States, 1994, pp. 725-736.

Adaptation to the loss of leader protein by cardiovirus: analysis of compensatory mutations.

K. V. Fominykh[1,2], P. V. Lidsky[1]
[1] *M. P. Chumakov Inst. Of Poliomyelitis & Viral Encephalitides, Rus. Acad. Med. Sci.*
[2] *M. V. Lomonosov Moscow State University, Moscow, Russia*
foxenia@gmail.com

Abstract

Viruses of Picornaviridae family cause severe diseases of humans and animals. Recently, it was found that different picornaviruses may alter nucleo-cytoplasmic transport in the infected cells. Diversity of mechanisms of nucleo-cytoplasmic transport disruption appears to exist among different picornaviruses, so we suggest that such a host-pathogen interaction may confer to the viruses a selective advantage(s). In our investigation, we are attempting to ascertain the biological significance of the virus-induced changes in the nuclear envelope permeability. In our work, we made evolutionary experiment with mutant virus which is not able to disrupt nucleo-cytoplasmic traffic. After some sequential passages, mutant restored activity close to wild-type level without the re-acquisition of capacity to alter nucleo-cytoplasmic traffic. Here, we analyze nucleotide substitutions correlated with this phenotype change.

1. Introduction

Viruses of Picornaviridae family cause severe diseases of humans and animals. Investigation of interaction between picornaviruses and organisms is important for understanding pathogenesis and for development of effective therapeutic methods.

Recently, it was found that different picornaviruses may alter nucleo-cytoplasmic transport in the infected cells, in particular, by increasing bidirectional permeability of the nuclear envelope.

Diversity of mechanisms of nucleo-cytoplasmic transport disruption appears to exist among different picornaviruses, so we suggest that such a host-pathogen interaction may confer to the viruses a selective advantage(s). In our investigation, we are attempting to ascertain the biological role of the virus-induced changes in the nuclear envelope permeability.

2. Results and discussion

It was shown previously that it is the 67 aa viral leader (L) protein that is responsible for such changes in the case of cardioviruses [1, 2]. Deletion of L [3] resulted in the loss of mengovirus (*g. Cardiovirus*) capacity to injure the nuclear envelope, on the one hand, and in a cell-dependent decrease in viral fitness (as evidenced by smaller plaque size and a diminished yield of infectious virus in single-cycle growth experiments), on the other.

Sequential passages of the leader-less mutant in BHK-21 cells resulted in a significant restoration of viability, which was not accompanied with the re-acquisition of capacity to alter nucleo-cytoplasmic traffic. This discordant effect suggests that alteration in the nucleo-cytoplasmic traffic is not an obligatory prerequisite for efficient growth. We sequenced the genome of the newly isolated virus to find mutations correlated with its phenotype changes. Three mutations were detected: T99A in capsid protein VP2, V282A in nonstructural protein 2C and T47M in viral protease 3C. In this work, we analyze these nucleotide substitutions by bioinformatic methods and investigate their importance in the increase of viral fitness. Now we can suggest, for example, that mutation T99A in capsid protein VP2 might influence assembling of the capsid. The roles of other mutations are under investigation.

Our work shows that the virus can adapt to the deficient ability to disrupt normal nucleo-cytoplasmic transport by acquiring some compensatory mechanisms.

12. References

[1] S. Delhaye,V. van Pesch and T. Michiels, "The leader protein of Theiler's virus interferes with nucleocytoplasmic trafficking of cellular proteins", *J Virol.*, 2004 Apr, 78(8), pp. 4357-4362.

[2] P.V. Lidsky, S. Hato, M.V. Bardina, A.G. Aminev, A.C. Palmenberg, E.V. Sheval, V.Y. Polyakov, F.J. van Kuppeveld and V.I. Agol, “Nucleocytoplasmic traffic disorder induced by cardioviruses”, *J Virol.*, 2006 Mar, 80(6), pp. 2705-2717.

[3] J. Zoll, J.M. Galama, F.J. van Kuppeveld, W.J. Melchers, “Mengovirus leader is involved in the inhibition of host cell protein synthesis”, *J Virol.*, 1996 Aug, 70(8), pp. 4948-4952.

Изучение регуляции транскрипции в геномах рода *Shewanella* путем автоматического поиска регуляторных мотивов в рамках функциональных подсистем

Ставровская Е.Д.[1,3] , Сиприано М.[2], Родионов Д.А.[3,4], Дубчак И.Л.[2], Миронов А.А.[1,3], Гельфанд М.Г.[1,3]

[1]*Факультет Биоинженерии и Биоинформатики, Московский Государственный Университет имени М.В. Ломоносова, Москва, Россия*

[2]*Национальная лаборатория им. Лоуренса в Беркли, США*

[3]*Институт проблем передачи информации им. А. А. Харкевича РАН, Москва, Россия*

[4]*Институт медицинских исследований Бернхама, Ла Хойе, США*

stavrovskaya@gmail.com

Аннотация

В настоящее время изучение бактерий рода Shewanella приобрело важное практическое и стратегическое значение. Эти бактерии могут преобразовывать растворимые металлы, такие как уран, в нерастворимые формы. Они могут жить практически везде, как в аэробных так и в анаэробных условиях. Данные бактерии не вызывают заболеваний у человека, животных и других живых организмов.

Важным этапом при изучении геномов является определение участков связывания белков-регуляторов — регуляторных сайтов. Наличие общего регуляторного мотива перед генами, вовлеченными в один биологический процесс, говорит об их совместной регуляции. Кроме того, зная биологический процесс, в котором участвуют гены, и глядя на структуру найденного мотива, можно предсказать потенциальный белок, который регулирует эти гены. В данной работе мы исследовали гены из 61 функциональной подсистемы бактерий рода Shewanella с помощью автоматического поиска регуляторных мотивов. Результаты, полученные нашим автоматическим методом, коррелируют с результатами аккуратного ручного анализа.

Материалы и методы

Для автоматического поиска регуляторных мотивов перед генами в рамках функциональных подсистем мы использовали конвеер программ. Конвеер позволяет осуществлять поиск потенциальных регуляторных мотивов в рамках одной функциональной подсистемы для заданной группы геномов.

Функциональные подсистемы были взяты из базы данных SEED [1]. Подсистема состоит из функциональных ролей, которые составляют определенный биологический процесс или структурный комплекс. Функциональная роль – это функция, которую выполняет белок. Подсистему можно рассматривать как обобщенный метаболический путь и представить в виде обобщенной таблицы. Каждая колонка таблицы соответствует одной функциональной роли подсистемы, а строка – одному геному. Ячейка таблицы содержит гены, которые кодируют белки с выбранной функцией в рассматриваемом геноме.

Конвеер включает в себя процедуру, которая убирает слишком близкие геномы из исходной группы. Для определения расстояний между геномами мы использовали филогенетическое дерево организмов, которое построено по множественному выравниваю высококонсервативных белков из 74 кластеров ортологичных генов из базы данных COGs (http://www.microbesonline.org/treebrowseHelp.html#speciestree) [2]. Нами было выбрано пороговое значение 0,01. Геномы, расстояние между которыми меньше порога, считаются слишком близкими. Пороговое значение подобрано так, что близкими геномами считаются штаммы и некоторые близкие виды (например, *Neisseria gonorrhoeae* и *Neisseria meningitidis*). Поскольку функциональная подсистема может содержать гены не из всех запрашиваемых геномов, отсев близких геномов производится после проецирования группы геномов на исследуемую подсистему.

Конвеер реализует три стратегии поиска мотивов. Первая – искать мотивы отдельно для каждого генома (перед генами из каждой строки таблицы), а затем кластеризовать полученные мотивы. Вторая – искать мотивы в рамках каждой функциональной роли (перед генами из каждого столбца таблицы) и кластеризовать результаты. Третья – искать мотивы для всех генов подсистемы (перед всеми генами таблицы).

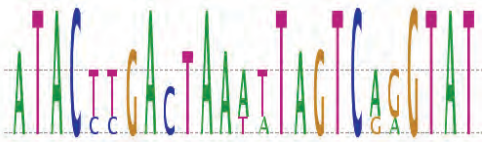
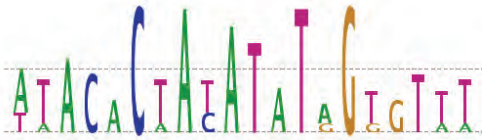
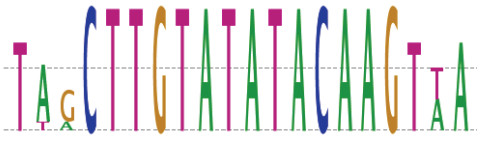
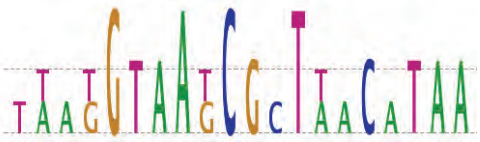
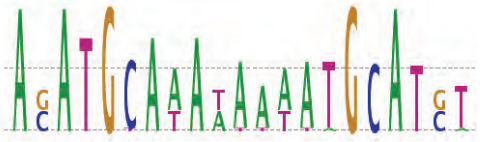
Для поиска мотива в наборе областей перед генами мы использовали алгоритм *SignalX* [3]. Для кластеризации полученных мотивов мы использовали алгоритм ClusterTree-RS [4]. Конвеер позволяет искать палиндромные безделеционные мотивы фиксированной длины. На вход подается описание функциональной подсистемы и группа геномов, в рамках которых осуществляется поиск, а также длина искомого мотива.

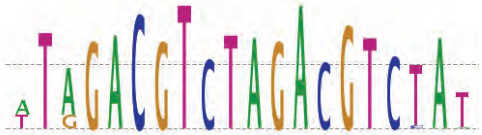
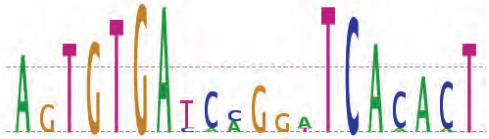
Результаты хранятся в базе данных. Для просмотра результатов с помощью технологии JSP (Java Servlet Pages) был реализован веб сервер (http://storage.bioinf.fbb.msu.ru:8180/SEED_motifs/).

Результаты и обсуждение

С помощью нашего метода мы проанализировали 61 функциональную подсистему в геномах рода *Shewanella*. Параллельно производился аккуратный ручной анализ данных функциональных подсистем. Результаты, полученные нашим методом, в большинстве случаев коррелируют с результатами ручного анализа. Некоторые результаты приведены в таблице 1.

Таблица 1. Некоторые результаты автоматического поиска регуляторных мотивов, коррелирующие с результатами ручного анализа.

функциональная подсистема	транскрипционный фактор	диаграмма лого
Fe-S_cluster_assembly	IscR	
Ribonucleotide_reduction	NrdR	
L-Arabinose utilization	AraR	
Histidine_Degradation	HutC	
Beta-Glucoside_Metabolism	BglR	
Nitrosative_stress	NsrR	

функциональная подсистема	транскрипционный фактор	диаграмма лого
	NorC	
Methionine Biosynthesis	MetJ	
	MetR	
TCA Cycle	? (найден мотив, похожий на crp)	

Обнаружено также большое количество потенциальных новых мотивов, которые нуждаются в дальнейшем уточнении.

Ссылки

[1] R. Overbeek, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crécy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank ED, Gerdes S, Glass EM, Goesmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L, Kubal M, Larsen N, Linke B, McHardy AC, Meyer F, Neuweger H, Olsen G, Olson R, Osterman A, Portnoy V, Pusch GD, Rodionov DA, Rückert C, Steiner J, Stevens R, Thiele I, Vassieva O, Ye Y, Zagnitko O, Vonstein V. ”The Subsystems Approach to Genome Annotation and its Use in the Project to Annotate 1000 Genomes”, *Nucleic Acids Research*, 33(17), 2005, pp.5691-5702.

[2] Tatusov RL, Koonin EV, Lipman DJ. «A genomic perspective on protein families.» *Science*. 278(5338),1997, pp. 631-637.

[3] Mironov A.A., Vinokurova N.P., Gel’falnd M.S. “Software for analyzing bacterial genomes.” *Mol Biol* 34(2) (Mosk)., 2000, pp. 253-262.

[4] Е.Д. Ставровская, ВЮ. Макеев, АА. Миронов “ClusterTree-RS: алгоритм кластеризации регуляторных синалов с помощью бинарного дерева.” Молекулярная биология 40(3), 2006, pp. 465-473

Non-5' CpG islands in the human genome: probable involvement in transcription regulation

Yulia A. Medvedeva,[1,*] Marina V. Fridman,[1] Nina J. Oparina,[2] Dmitri B. Malko,[1] Ekaterina O. Ermakova,[3] Ivan V. Kulakovskiy,[1,2] Vsevolod J. Makeev,[1,2]
[1]Institute of Genetics and Selection of Industrial Microorganisms, GosNIIgenetika, 1st Dorozhny proezd, 1, Moscow, 113545, Russia;
[2]Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Vavilova str., 32, Moscow, 199991, Russia;
[3]Institute for Information Transmission Problems (The Kharkevich Institute), Russian Academy of Sciences, Bolshoy Karetny per. 19, Moscow, 127994, Russia.
**Corresponding author: ju.medvedeva@gmail.com*

Abstract

Recent interest in epigenetic regulation has led to a focus on CpG islands (CGIs) due to their methylation status. While previous studies have addressed CGIs located 5' to genes, the functions of others have been neglected. We demonstrate that CGIs located far from the 5' regions of known genes can be functional. CGIs overlapping with coding exons reduce the synonymous substitution rate in CpG-containing codons independent of their distance from the gene start. Unmethylated regions are found not only in 5' CGIs, but also in intragenic and 3' CGIs. All types of CGIs contain frequent CAGE-tags. Moreover, intragenic, intergenic and 3' located CGIs hold an increased number of Sp1 binding sites. Although 5' CGIs typically encompass more unmethylated CpG pairs, more Sp1 sites, and more CAGE-tags than other types of the CGIs, these data suggest non 5' CGIs (especially 3' CGIs) are likely to participate in non-coding RNA transcription regulation.

.

1. Introduction

More than half of all human genes are associated with CpG islands (CGIs) proximal to their transcriptional start sites (TSS) or located in the 5' flanking sequences. Although the role of these CGIs is still disputed, growing evidence supports their participation in transcriptional regulation [1]. Significantly fewer genes have CGIs in their 3' flanking sequences or intragenically, and these CGIs have not previously been systematically studied at the genome-wide level. The intergenic CCIs are considered only to be involved in repression of mobile elements such as Alu-repeats in human genomes. As we here demonstrate 3' and to a lesser degree, intragenic and intergenic CGIs display similar properties as 5' CGIs. Our findings allow as proposing that all the types of CGIs probably perform the same biological function.

2. Results and discussion

Using the alignments of the protein coding regions of human and mouse genes that overlap and do not overlap with human CGIs, we show that codons containing CpG pairs are synonymously substituted in CGIs less frequently than similar codons in alignment segments not overlapping with CGIs. This phenomenon is equally characteristic for CGIs overlapping 5', intragenic and 3' exons but notably is almost never observed for nonsynonimous substitutions. In our opinion, this indicates that a CGI-dependent selection at the DNA (or RNA) level indeed exists, and reduces the synonymous substitution rate, but the corresponding selection pressure is smaller than the average selection for amino acid substitution in proteins.

We also compared the placement experimentally identified unmethylated regions (UMRs) and computationally identified CGIs in relation to defined genes. Approximately 50% of UMRs overlapped with CGIs. Although UMRs generally had a greater Obs/Exp rate and GC-content than the average for mammalian non-gene-associated genomic sequences, about half of UMRs displayed an Obs/Exp rate and GC content that was significantly lower than the threshold used in CGI definition. These data imply that the lack of methylation is not sufficient to explain the functionality of high Obs/Exp and G+C values of

CGIs, and some other explanation for CGI sequence properties is needed.

Out study shows the relative enrichment of CAGE tags not only in 5' CGIs but also in intragenic, 3'- and intergenic CGIs. This data suggests that all types of CGIs may be associated with transcription initiation. The suggested role of non-5' CGIs in transcription initiation agrees with the overrepresentation of high scoring Sp1 sites in many CGIs independent of their gene location. It is noteworthy that CGIs rich with CAGE tags contain more high-scoring Sp1 binding sites. Abundant Sp1 binding in gene 3' or intergenic regions was observed in genome-wide site location experiments [2], which reported that 36% of the clusters of Sp1, Myc and p53 binding sites lie within or immediately 3' to well-characterized genes. In these cases, non-coding RNA transcription may be initiated.

Recent studies of the mouse genome [3] demonstrated that a large number of ncRNA are initiated in the 3' regions of the genes, with specific enrichment at the 3' terminus of the final exon. A substantial number of these ncRNAs are complementary to known genes as anti-sense strands, which has led to the suggestion of an additional mechanism of gene silencing by natural antisense interference [4] (Katayama et al. 2005). The authors of this study also observed that sense-antisense pairing was almost universally associated with candidate imprinted loci. As genetic imprinting is frequently associated with an altered methylation status of CpG islands [5], CGIs located in 3' gene regions [4] or intragenic CGIs [6] may play an important role in this process by regulation of gene expression via inducing antisense-based gene interference.

3. References

[1] Rozenberg J.M., Shlyakhtenko A., Glass K., Rishi V., Myakishev M.V., FitzGerald P.C., Vinson C. “All and only CpG containing sequences are enriched in promoters abundantly bound by RNA polymerase II in multiple tissues”, *BMC Genomics*, 2008, 9, 67.

[2] Cawley S., Bekiranov S., Ng H.H., Kapranov P., Sekinger E.A., Kampa D., Piccolboni A., Sementchenko V., Cheng J., Williams A.J., Wheeler R., Wong. B., Drenkow J., Yamanaka M., Patel S., Brubaker S., Tammana H., Helt G., Struhl K., Gingeras T.R., “Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs”, *Cell*, 2004, 116, 499-509.

[3] Carninci P., Sandelin A., Lenhard B., Katayama S., Shimokawa K., Ponjavic J., Semple C.A., Taylor M.S., Engstrom P.G., Frith M.C., Forrest A.R., Alkema W.B., Tan S.L., Plessy C., Kodzius R., Ravasi T., Kasukawa T., Fukuda S., Kanamori-Katayama M., Kitazume Y., Kawaji H., Kai C., Nakamura M., Konno H., Nakano K., Mottagui-Tabar S., Arner P., Chesi A., Gustincich S., Persichetti F., Suzuki H., Grimmond S.M., Wells C.A., Orlando V., Wahlestedt C., Liu E.T., Harbers M., Kawai J., Bajic V.B., Hume D.A., Hayashizaki Y. “Genome-wide analysis of mammalian promoter architecture and evolution”, *Nature Genetics*, 2006, 38, 626-635.

[4] Katayama S.Y., Tomaru T., Kasukawa K., Waki M., Nakanishi M., Nakamura H., Nishida C.C., Yap M., Suzuki J. Kawai H., Suzuki P., Carninci Y., Hayashizaki C., Wells M., Frith T., Ravasi K.C., Pang J., Hallinan J., Mattick D.A., Hume L., Lipovich S., Batalov P.G., Engstrom Y., “Antisense transcription in the mammalian transcriptome”, *Science*, 2005, 309, 1564-1566.

[5] Teodoridis J.M., Strathdee G., Brown R., “Epigenetic silencing mediated by CpG island methylation: potential as a therapeutic target and as a biomarker”, *Drug Resist. Updat*., 2004, 7, 267-278.

[6] Ramser J., M.E. Ahearn, C. Lenski, K.O. Yariz, H. Hellebrand, M. von Rhein, R.D. Clark, R.K. Schmutzler, P. Lichtner, E.P. Hoffman, A. Meindl, and L. Baumbach-Reardon, “Rare missense and synonymous variants in UBE1 are associated with X-linked infantile spinal muscular atrophy”, *Am J Hum Genet.*, 2008, 82, 188-193.

Изучение системы деградации этаноламина методами сравнительной геномики

О.В. Цой, Д.А. Равчеев, М.С. Гельфанд, А. Мушегян
Московский Государственный Университет им. М.В. Ломоносова,
Институт проблем передачи информации им. А.А. Харкевича РАН,
Stowers Institute for Medical Research, Kansas City, MO
olgatsoj@list.ru

Аннотация

Одним из источников углерода, азота и энергии бактерий может служить этаноламин. Путь деградации этаноламина был экспериментально изучен у Salmonella typhimurium, в которой гены утилизации этаноламина образуют единый eut оперон, но ничего не известно об эволюции и этой системы. В данной работе была проанализирована эволюция и регуляция генов, ответственных за деградацию этаноламина. Была выдвинута гипотеза о происхождении ферментов деградации этаноламина и дальнейшей эволюции всей системы, включающей события горизонтального переноса генов от Proteobacteria к Actinobacteria и, предположительно, от предка Proteobacteria к предку Firmicutes. Также предсказан потенциальный мотив связывания регулятора EutR, контролирующего деградацию этаноломина. Также удалось установить связь пути деградации этаноламина с синтезом кофактора через систему регуляции.

1. Введение

Путь утилизации этаноламина является частным случаем реакций деградации диолов с образованием альдегидов. Все известные подобные пути требуют наличие фермента лиазы и его кофактора – кобаламина (витамина B12). Путь утилизации этаноламина был экспериментально изучен у *S. typhimurium*. Анализом различных штаммов *S. typhimurium*, мутантных по разным регионам *eut*-оперона, был определен состав оперона и роль продуктов отдельных генов в пути утилизации этаноламина *S. typhimurium* [1].

Основным ферментом пути является этаноламин лиаза – белковый комплекс, состоящий из двух субъединиц: большой EutB и малой, EutC. Этаноламин лиаза осуществляет главную реакцию деградации этаноламина – расщепление до ацетальдегида и аммиака, который используется как источник азота.

Дальнейшее превращение ацетальдегида осуществляется оксидоредуктазой EutE до ацетил-КоА или алкогольдегидрогеназой EutG до этанола. Ацетил-КоА в свою очередь может быть превращен в ацетат с получением одной молекулы АТФ с помощью ферментов ацетат киназы Ack и фосфоацетилтрансферазы EutD, либо отправлен в цикл Кребса. [1].

Кофактором этаноламин лиазы является аденозилкобаламин, который может быть получен извне или из цианокобаламина или гидроксикобаламина, синтезируемых в клетке. Превращение в аденозилкобаламин происходит с помощью аденозилкобаламинтрансферазы EutT [2]. Цианокобаламин и гидроксикобаламин являются ингибиторами этаноламин лиазы. От их воздействия ферментативный комплекс EutBC защищает белок EutA, обнаруженный у многих растущих на этаноламине бактерий [1].

У некоторых бактерий, способных к деградации диолов, обнаружен особый компартмент, так называемая метаболосома. У таких организмов превращения этаноламина до ацетил-коА происходят внутри этого компартмента [3].

В дополнение к перечисленным генам, в разных группах бактерий обнаружены транспортеры этаноламина EutH и Eat и регулятор транскрипции EutR.

Ранее предпринимались попытки экспериментального изучения регуляции *eut*-оперона. Для *S. typhimurium* была экспериментально показана роль белка EutR в регуляции экспрессии генов утилизации этаноламина. Штаммы *S. typhimurium*, мутантные по гену *eutR*, утрачивают способность расти на этаноламине. Эта способность появляется вновь при трансформации мутантов плазмидой, содержащей ген *eutR*. Также экспериментально показано, что EutR является активатором экспрессии деградации этаноламина [4].

У *S. typhimurium* гены, кодирующие многие вышеупомянутые ферменты, образуют единый оперон – *eutSPQTDMNEJGHABCLKR*. Белки EutT, EutD, EutE, EutG, EutA, EutB и EutC принимают участие в утилизации этаноламина, EutH осуществляет транспорт этаноламина в клетку, а EutS, EutM, EutN, EutL, EutK являются структурными белками метаболосомы. Роль белков EutP, EutQ и EutJ до настоящего момента остается неизученной.

Интересно, что путь утилизации этаноламина биохимически похож на путь утилизации пропандиола, регуляция которого экспериментально изучена в той же *S. thyphimurium*. Гены деградации пропандиола так же как и в случае деградации этаноламина образуют единый оперон – *pduABCDEGHJKLMNOPQSTUVWX*.
Установлено, что регуляция экспрессии генов утилизации пропандиола осуществляется PocR регулятором AraC семейства, а также прямо или опосредованно глобальными регуляторами Crp и ArcA [5]. PocR также контролирует синтез кобаламина [6]. Таким образом, для пути деградации пропандиола показана его связь с синтезом кофактора, что позволяет на основании сходства двух путей предположить такую же связь и для системы деградации этаноламина.

Наряду с предсказанием сайтов связывания EutR с промоторной областью *eut*-оперона, была поставлена задача нахождения связи между путями деградации этаноламина и синтеза кобаламина.

2. Данные и методы

2.1. Методы сравнительной геномики

Для поиска потенциальных сайтов связывания использовался метод филогенетического футпринтинга и метод матриц позиционных весов нуклеотидов.

Для определения принадлежности гена к регулону применяли метод проверки соответствия. Ген считался членом регулона, если в его регуляторной области обнаружен потенциальный сайт связывания, который сохраняется перед его ортологами в родственных геномах. В данной работе устанавливалась область поиска -400…+100 от старта трансляции.

В случае, когда оперонная структура неизвестна, полагали, что гены относятся к одному оперону если они имеют одинаковое направление считывания и расстояние между ними не превышает 200 пар нуклеотидов и оперонная структура сохраняется в нескольких родственных геномах.

2.2. Программное обеспечение

Для поиска сайтов и анализа ортологов использовался пакет программ Genome Explorer [7]. Построение матриц позиционных весов проводилось с помощью программы Signal X [7]. Для выравнивания аминокислотных и нуклеотидных последовательностей использованы программы ClustalX [8]. Для построения диаграмм Лого, отображающих структуру регуляторного сайта, использовали программу Weblogo [9]. Для построения филогенетических деревьев и оценки достоверности их топологии использовались пакеты программ PHYLIP [10] и MrBayes [11], а для реконструкции предкового состояния операторов программа MESQUITE [12]. Поиск ортологов проводился с помощью программы PSI-BLAST [13].

3.3 Исследованные геномы

Исследовано более 100 полных последовательностей геномов, относящихся к Proteobacteria, Firmicutes, Actinobacteria, Acidobacteria, Bacteroidetes и Chlorophlexi. Все последовательности получены из базы данных GenBank [14].

3. Результаты и обсуждение

3.1. Поиск ортологов генов утилизации этаноламина

Ортологи субъединиц основного фермента - EutB и EutC - обнаружены в более чем 100 бактериях из различных таксономических групп: Proteobacteria, Firmicutes, Actinobacteria, Acidobacteria, Bacteroidetes, Chlorophlexi, Flavobacteria, Lentisphaerae, Planctomycetacia и Fusobacteria.

3.2. Особенности оперонных структур генов утилизации этаноламина

Структура *eut* - оперона оказалась различной в разных группах бактерий. Можно выделить два главных типа подобных оперонов – «короткий» и «длинный»

Оперoны короткого типа состоят из генов основного фермента – *eutBC*, а также могут содержать гены пермеазы этаноламина *eat* и регулятора *eutR*. Подобная оперонная структура встречается почти во всех Proteobacteria и Actinobacteria. Для Enterobacteriales и Firmicutes характерен «длинный» тип оперона, содержащий

от 5 до около 20 генов (некоторые из этих генов могут и не относиться к метаболизму этаноламина).

Внутри Enterobacteriales наблюдается полиморфизм на уровне состава оперона. Так, если *S. typhimurium, C. koseri, K. pneumoniae* и *E. coli* содержат полный набор генов, то оперон в *S. sonnei* имеет вид *eutMNEJGHABCLKR*, а в *S. boydii* - *eutPQTMNEBCLKR* . В *S. dysenteriae* сохранились только гены *eutSCLKR*, а в *P. luminescens* - *eutBC*. Оперон *P. luminescens* при таком контексте все же является сокращенным вариантом «длинного» типа, а не «коротким» типом, что явно следует из анализа филогенетических деревьев, о котором подробнее будет сказано ниже.

Очень интересен случай сохранения двух типов оперона независимо внутри одного генома. Подобные случаи обнаружены в геномах *K. pneumoniae, M aquaeolei, P. fluorescens*.

Отдельные представители Fusobacteria, Chlorophlexi, Acidobacteria также содержат ортологи отдельных белков утилизации этаноламина. Fusobacteria обладает «длинным» типом, а Chlorobi и Acidobacteria – «коротким».

3.2. Эволюция системы утилизации этаноламина.

Подробное изучение строения оперонов в разных таксонах бактерий поставило вопрос об эволюции системы деградации этаноламина.

Мы обнаружили, что гены *eutB* и *eutC* из разных типов оперонов – «короткого» и «длинного» представляют сестринские ветви. Первая ветвь содержит белки из Proteobacteria и Actinobacteria, вторая – из *Enterobacteriales*, относящихся к Proteobacteria, и Firmicutes. В случае геномов, содержащих два *eut*-оперона (*K. pneumoniae, M aquaeolei, P. fluorescens*), *eutB* и *eutC* из оперонов разного типа лежат на разных ветвях.

Все эти случаи, когда расположение ветвей на филогенетическом дереве не соответствует таксономии, представляют особый интерес. Белки из бактерий группы Actinobacteria, сестринского по отношению к Proteobacteria и Firmicutes, оказываются внутри ветви, соответствующей белками из Proteobacteria. Actinobacteria содержат «короткий» тип оперона, также характерный для Proteobacteria. Такая топология является индикатором возможного горизонтального переноса от предка Alphaproteobacteria к бактерии порядка Actinomycetales с последующим вертикальным переносом.

Отдельные ветви, соответствующие белкам из представителей Fusobacteria, Chlorobi и Acidobacteria, которые являются таксонами самостоятельными, также находятся внутри Firmicutes и Proteobacteria. Возможно, в данном случае имел место горизонтальный перенос, о чем свидетельствует и то, что в остальных представителях данных групп ортологов генов утилизации этаноламина не найдено. Но сказать точно о горизонтальном переносе генов на данный момент невозможно. Положение же ветвей EutB из Chlorobi и Acidobacteria недалеко от корня допускает возможность недостоверного построения этой части дерева.

Другой яркий пример отличия топологии дерева EutB от дерева таксонов – перемещение ветви Gammaproteobacteria, относящихся к Proteobacteria, к Firmicutes.

Для детального исследования происхождения такой топологии мы реконструировали вероятное строение *eut-оперон*а у предка Firmicutes, предка Proteobacteria и предка Enterobacteriales в частности. На основании составов предковых оперонов можно выдвинуть следующую гипотезу об эволюции генов деградации этаноламина. Самый простой тип оперона предположительно был у предка Proteobacteria. Он включал только гены *eutBC*, а позднее – еще и ген транспортного белка. Далее этот основной набор ферментов был дополнен регулятором EutR. В таком виде оперон и сохранился у ныне живущих Proteobacteria.

В ветви Enterobacteriales событий было больше: здесь происходило накопление генов деградации этаноламина до современного вида и ускорение эволюции.

Что касается Firmicutes, то их *eut*-опероны могут быть результатом горизонтальных переносов от Enterobacteriales с заменой регулятора EutR на двух-компонентную систему. Также с помощью горизонтальных переносов внутри некоторых геномов возникло два независимых *eut*-оперона.

Таким образом, мы предполагаем, что система деградации этаноламина могла впервые возникнуть у Proteobacteria.

Дальнейшее детальное изучение эволюции генов деградации этаноламина требует новых геномных данных, охватывающих ветви, сестринские по отношению к Proteobacteria и Firmicutes.

3.4. Регуляция деградации этаноламина

EutR, регулятор AraC семейства, обнаружен у Enterobacteriales в составе *eut*-оперона и у Burkholderiales в непосредственной близости от *eut*-оперона. У Firmicutes и *F. nucleatum* (Fusobacteria) внутри *eut*-оперона находятся гены двух-компонентной регуляторной системы,

которые, предположительно, и контролируют утилизацию в этих организмах. В остальных группах бактерий не обнаружено ни ортологов EutR, ни ортологов генов белков двух-компонентной системы. Можно предположить, что в этих организмах либо присутствует некий еще неизвестный регулятор, либо экспрессия генов деградации этаноламина не регулируется.

У Enterobacteriales методом филогенетического футпринтинга в промоторной области *eut*-оперона найдены два консервативных мотива. Один из них при детальном рассмотрении обнаружил сходство с сайтом связывания CRP. Второй же мотив, как было предположено, является сайтом связывания EutR.

У представителей Burkholderiales в промоторной области *eut*-оперона обнаружен только один консервативный сайт. При сравнении сайтов из Enterobacteriales и Burkholderiales выявилось их сходство, что подтверждает наши предположения о том, что это сайт связывания именно регулятора EutR.

На основании полученных сайтов была построена матрица позиционных весов. Эту матрицу использовали для поиска генов, экспрессия которых может регулироваться EutR.

Профилем обнаружено порядка 60 генов, сгруппированных в 7 оперонов. Функции данных генов относятся к различным метаболическим процессам, таким как синтез клеточной стенки, метаболизм ксилозы, ДНК-связывающий белок и мембранные белки. Функциональная связь продуктов этих генов с системой утилизации этаноламина не ясна. Также мотив обнаружен перед геном *cbiA* – первым геном оперона синтеза кобаламина. Кобаламин является кофактором основного фермента деградации этаноламина, поэтому, подобно случаю с утилизацией пропандиола, регулятор EutR также может одновременно контролировать деградацию этаноламина и синтез необходимого для этого кофактора – кобаламина.

4. Список литературы

[1] D.M. Roof, J.R. Roth "Functions required for vitamin B12-dependent ethanolamine utilization in Salmonella typhimurium." *J Bacteriol.*, American Society for Microbiology, United States, 1989, 3316-3323.

[2] D.E. Sheppard, J.T. Penrod, T. Bobik, E. Kofoid, J.R.Roth "Evidence that a B12-adenosyl transferase is encoded within the ethanolamine operon of Salmonella enterica." *J Bacteriol.*, American Society for Microbiology, United States, 2004, 7635-7644.

[3] S.R. Brinsmade, T. Paldon, J.C. Escalante-Semerena "Minimal Functions and Physiological Conditions Required for Growth of Salmonella enterica on Ethanolamine in the Absence of the Metabolosome." *J Bacteriol.*, American Society for Microbiology, United States, 2005, 8039–8046.

[4] D.M. Roof, J.R. Roth "Autogenous regulation of ethanolamine utilization by a transcriptional activator of the eut operon in Salmonella typhimurium." *J Bacteriol.*, American Society for Microbiology, United States, 1992,6634-6643.

[5] M. Ailion, T.A. Bobik,J.R Roth. "Two global regulatory systems (Crp and Arc) control the cobalamin/propanediol regulon of Salmonella typhimurium." *J Bacteriol.*, American Society for Microbiology, United States, 1993, 7200-7208.

[6] T. A. Bobik, M. Ailion, J.R. Roth "A single regulatory gene integrates control of vitamin B12 synthesis and propanediol degradation" *J Bacteriol.*, American Society for Microbiology, United States, 1992, 2253-2266.

[7] А.А. Миронов, Н.П. Винокурова, М.С. Гельфанд, "Программное обеспечение анализа бактериальных геномов", *Мол. биология*, "Наука", Москва, Российская Федерация, 2000, 253-262.

[8] J.D. Thompson, T.J. Gibson, F. Plewniak, F. Jeanmougin, D.G. Higgins "The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools." *Nucleic Acid Res.*, Oxford University Press, England, 1997, 4876-4882.

[9] T.D. Schneider, R.M. Stephens, "Sequence logos: a new way to display consensus sequences", *Nucleic Acid Res.*, Oxford University Press, England, 1990, 6097-6100.

[10] J.Felsenstein "PHYLIP - Phylogeny Inference Package (Version 3.2)." *Cladistics*, Meckler Publishing, United States, 1989, 164-166.

[11] J.P. Huelsenbeck, F. Ronquist "MRBAYES: Bayesian inference of phylogenetic trees." *Bioinformatics*, Oxford University Press, England, 2001, 754-755.

[12] mesquiteproject.org

[13] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, M.W. Zhang, D.J. Lipman "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucleic Acid Res.*, Oxford University Press, England, 1997, 3389-3402.

[14] D.A. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, D.L. Wheeler, "GenBank", *Nucleic Acids Res.*, Oxford University Press, England, 2007, 21-25.

Alternative splicing tends to involve phosphorylation sites

Yerbol Z. Kurmangaliyev[1,2], Mikhail S. Gelfand[1,3]

[1]*Institute for Information Transmission Problems (the Kharkevich Institute) RAS, Bolshoi Karetny pereulok 19, Moscow, 127994, Russia*
[2]*National Center for Biotechnology of the Republic of Kazakhstan, Valihanova 43, Astana, 010000, Kazakhstan*
[3]*Faculty of Bioengineering and Bioinformatics, Moscow State University, Vorobievy Gory 1-73, Moscow 119992, Russia*

E-mail:kurmangali@mail.ru

Abstract

The analysis of high-throughput data on human phosphoproteome from the PHOSIDA database and the data about alternative splicing of human genes from the EDAS database demonstrated that phosphorylated serines and threonines, but not tyrosines tend to occur in alternatively spliced regions. This tendency is weak, but statistically significant and may be explained by structural reasons, as it is known that phosphosites more often occur in protein loops, which, like alternative regions, tend to the protein surface.

1. Introduction

Protein phosphorylation is one the most important and widely distributed types of posttranslational modifications. In eukaryotes, it plays a key role in the signal transduction and other processes. Recent advances in mass spectrometry allowed for large-scale identifications of phosphosites and demonstrated that at least one third of the human proteome is phophorylated [1,2].

A large set of *in vivo* phosphorylation sites identified in several recent largescale experiments has been collected in the PHOSIDA database [3]. The mapping of these data onto predicted protein secondary structures has shown that phosphoserines and phosphothreonines tend to be located in loops and hinges, whereas phosphotyrosines occur within regions of regular secondary structure. The phosphorylation sites have been shown to be more conserved than generic loops, and than non-phosphosites of the same proteins [3].

On the other hand, several recent studies demonstrated a profound influence of alternative splicing on protein structure and function (reviewed, e.g., in [4]). It has been demonstrated that alternative splicing tends to affect functional sites in proteins [5] and to involve domains that function in signal transduction and cellular communication [6,7].

This has prompted us to investigate the relation between alternative splicing and phosphorylation using newly available large-scale data.

2. Findings

Human phosphorylation sites were taken from the set of phosphopeptides from HeLa cells [8]. We used only phosphopeptides unique matches in the genome. This resulted in 1907 genes with 7584 phosphorylation sites. The control set consisted in non-phosphosites of the longest isoforms of these genes. The data about alternative splicing, including the inclusion levels of each codon, were taken from the EDAS database [9]. Briefly, the inclusion rate of a codon was defined as the fraction of EST alignments containing this codon in exon among all EST alignments containing this codon in exon or in intron.

About 30% of all residues, including phosphosites, had inclusion rate 99% and it was difficult to determine whether they this had resulted from bona fide alternative splicing or from splicing errors and experimental artefacts. All other residues were classified as constitutive (belonging to constitutive segments, inclusion rate 100%) or alternative (belonging to alternative segments, inclusion rate 99% or less).

The analysis demonstrated that phosphosites tend to alternatively spliced regions (Fig.1). In all cases this tendency is weak, but statistically significant. All phosphosites combined were alternative in 24% of cases and constitutive in 46% of cases, whereas for non-phosphosites (both combined and separate for different amino acids) these fractions were 21% and 51%, respectively (Chi-square test, P=0).

The same trend was observed for serine and threonine. Of phosphoserines, 24% were alternative, and 45% were constitutive (Chi-square test, $P=2\cdot10^{-8}$). Similarly, 25% of phosphothreonines were alternatively and 44% were constitutive (Chi-square test, $P=1\cdot10^{-4}$ respectively). No difference was observed for the relatively small subsample of phosphotyrosines.

One of the reasons for the observations made here may be the fact that alternative regions of proteins tend to the protein surface [10] and disordered regions of the structure [11], and thus the observed correlation is caused by a common factor rather than a direct functional link. A detailed analysis involving collecting or homology-modeling of structures of phosphorylated proteins is required to resolve this issue; this is beyond the scope of the present study.

3. References

[1] Cohen P: The role of protein phosphorylation in human health and disease. The Sir Hans Krebs Medal Lecture. Eur J Biochem 2001, 268:5001-5010.

[2] Ptacek J, Snyder M: Charging it up: global analysis of protein phosphorylation. Trends Genet 2006, 22:545-554.

[3] Gnad F, Ren S, Cox J, Olsen JV, Macek B, Oroshi M, Mann M: PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. Genome Biol 2007, 8:R250.

[4] Artamonova II, Gelfand MS: Comparative genomics and evolution of alternative splicing: the pessimists' science. Chem Rev 2007, 107:3407-3430.

[5] Kriventseva EV, Koch I, Apweiler R, Vingron M, Bork P, Gelfand MS, Sunyaev S: Increase of functional diversity by alternative splicing. Trends Genet 2003, 19:124-128.

[6] Liu S, Altman RB: Large scale study of protein domain distribution in the context of alternative splicing. Nucleic Acids Res 2003, 31:4828-4835.

[7] Yura K, Shionyu M, Hagino K, Hijikata A, Hirashima Y, Nakahara T, Eguchi T, Shinoda K, Yamaguchi A, Takahashi K, Itoh T, Imanishi T, Gojobori T, Go M: Alternative splicing in human transcriptome: functional and structural influence on proteins. Gene 2006, 380:63-71.

[8] Olsen JV, Blagoev B, Gnad F, Macek B, Kumar C, Mortensen P, Mann M: Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. Cell 2006, 127:635-648.

[9] Nurtdinov RN, Neverov AD, Mal'ko DB, Kosmodem'ianskii IA, Ermakova EO, Ramenskii VE, Mironov AA, Gel'fand MS: EDAS, databases of alternatively spliced human genes. Biofizika 2006, 51:589–592.

[10] Wang P, Yan B, Guo JT, Hicks C, Xu Y: Structural genomics analysis of alternative splicing and application to isoform structure modeling. ProcNatl Acad Sci USA 2005, 102:18920-18925.

[11] Romero PR, Zaidi S, Fang YY, Uversky VN, Radivojac P, Oldfield CJ, Cortese MS, Sickmeier M, LeGall T, Obradovic Z, Dunker AK: Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. Proc Natl Acad Sci USA 2006, 103:8390-8395.

Polymorphism Due to Multiple Amino Acid Substitutions at a Codon Site within *Ciona savignyi*

Nilgun Donmez[1]
Department of Computer Science, University of Toronto, Toronto ON M5S 3J4 Canada
nild@cs.toronto.edu

Michael Brudno
Banting & Best Department of Medical Research, University of Toronto, Toronto ON M5S 3J4 Canada
brudno@cs.toronto.edu

Georgii A. Bazykin[1]
Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute), Moscow 127994, Russia
gbazykin@iitp.ru

Alexey S. Kondrashov
Life Sciences Institute and Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109-2216, USA
kondrash@umich.edu

Abstract

We compared two haploid genotypes sequenced from the same Ciona savignyi *individual and identified codons at which these genotypes differ by two non-synonymous substitutions. Using the* Ciona intestinalis *genome as an outgroup, we showed that both non-synonymous substitutions tend to happen within the same lineage. Only in 64 (24%) of 267 codons, one substitution occurred in each of the two allelic lineages, although 133 (50%) of such codons are to be expected if substitutions were independent. Such clumping of non-synonymous substitutions indicates that a substantial fraction of non-synonymous differences between* C. savignyi *individuals are adaptive.*

1. Introduction

A wide variety of methods have been developed in order to detect natural selection at the level of DNA and protein sequences. In particular, detection of positive selection has attracted significant attention recently. One possible approach is to look for clumping of non-synonymous nucleotide substitutions that occur within a particular codon [1,2]. If such clumping is observed, and clumping of synonymous substitutions is weak or absent, positive selection for at least some amino acid replacements is a likely explanation.

So far, analysis of clumping has been performed for independently diverging lineages. Here, we take advantage of two recently sequenced genotypes of *Ciona savignyi* that belong to the same population and indeed to the same individual [3], and perform an analogous analysis for nucleotide substitutions that contribute to within-population variation.

2. Results and discussion

We aligned the two untranslated *C. savignyi* haplomes (i.e. haploid genotypes) against the translated *C. intestinalis* genes using the "protein2genome" functionality of the Exonerate software. Alignments with ambiguities in nucleotide sequences, internal stop codons, and internal gaps of length not a multiple of three were discarded. The remaining dataset consisted of 2,970 triplets of orthologous genes.

Among the 816,921 homologous codons in these 2,970 analyzed genes, 92.72% were identical in the two genotypes, and 6.95%, 0.28% and 0.05% differed at one, two and three nucleotide sites, respectively (no-, one-, two- and three-substitution codons). The mean evolutionary distances between genotypes are 0.129 at synonymous sites and 0.012 at non-synonymous sites, in agreement with [3].

At 48% of one-substitution codons, O (outgroup) coincides with either A (haplome A) or B (haplome B), and at 53% of one-substitution codons, O encodes the same amino acid as A and/or B. In such cases we assume that O reveals the Last Common Ancestor (LCA) codon or, at least, the amino acid it

[1] These two authors contributed equally to this work.

encodes. Otherwise, the LCA remains unknown. We assume that the nucleotide (amino acid) substitution occurred in the lineage of haplome A, if O coincides with (i.e. encodes the same amino acid as) B, and in the lineage of haplome B, if O coincides with A.

Let us now consider the 2,275 codons separated by two substitutions. If the substitutions were mutually independent, we would expect both substitutions to occur in the same lineage and in different lineages with the same probability of ~50% [1]. This is exactly what we observe if both substitutions are synonymous. In contrast, two non-synonymous substitutions occur in the same lineage 76% of the time and 24% of the time in different lineages. This clumping is not linked with correlated mutation events spanning two adjacent nucleotides, since clumping is strong in codons in which substitutions occurred in the first and the third nucleotide sites.

Clumping is even stronger in 429 three-substitution codons: while all substitutions are expected to occur in the same lineage in only 25% of cases, the observed value is 52%.

Our analysis of differences between the two genotypes from one *C. savignyi* individual closely follows that for the differences between mouse and rat genomes [1]. Surprisingly, the results of these analyses are also rather similar to each other. Our data reveal clumping of within-population non-synonymous polymorphisms at the same codon that is similar in magnitude to the previously observed clumping of non-synonymous substitutions that distinguish different genomes. Clumping of non-synonymous substitutions in divergent mammalian species [1] and HIV-1 strains [2] was interpreted as the signature of episodes of positive selection which had acted on corresponding sites at some time since divergence of the analyzed lineages. Although the exact fraction of adaptive amino acid substitutions is still debated, the consensus based on a variety of methods is that this fraction is substantial.

Conversely, the role of positive selection in maintaining polymorphism at the molecular level is believed to be small. Indeed, a transitive polymorphism that represents a phase of a selection-driven allele replacement must be short-lived and therefore unlikely to be encountered at any given instant. Still, clustering of amino acid replacements between genotypes belonging to the same population and separated, at a given codon site, by two or three non-synonymous substitutions seems to indicate that positive selection is involved.

The hypothesis that a substantial fraction of non-synonymous differences between individuals from the same population of *C. savignyi* are adaptive and are subject to positive selection is potentially feasible. *C. savignyi* population has a large effective size [3] and may be constantly accumulating new alleles; if a large fraction of replacements is due to positive selection, it is possible that we observe a large number of adaptive substitutions in the process of their fixation in the population. Because we observe 24%, instead of the expected 50%, of two-substitution codons with one non-synonymous substitution in each lineage, our hypothesis implies that ~50% of non-synonymous polymorphisms at two-substitution codons are temporary, transitive polymorphisms existing under directional selection. Assuming that this selection is generally weak, with selection coefficients of the order of 10^{-3}–10^{-4}, such a polymorphism exists for ~10^3–10^4 generations, and the per generation number of adaptive allele replacements that is needed to explain our data is not very high. Indeed, for the ~20,000 genes of *C. savignyi* genome [3] our estimate implies that ~1000 transitive polymorphisms with two non-synonymous substitutions at a codon and a high enough frequency of both alleles coexist within the population. Of course, there must also be many transitive one-substitution polymorphisms, so that the overall number of adaptive amino acid replacements per generation must by high.

Our hypothesis implies a testable prediction. Positively selected alleles reach high frequencies rapidly. Therefore, nucleotide diversity in the vicinity of a young, derived allele created by two or three advantageous substitutions must be reduced. This prediction can be tested once several more genotypes of *C. savignyi* are available, so that a derived allele can be encountered in more than one genotype.

3. References

[1] G.A. Bazykin, F.A. Kondrashov, A.Y. Ogurtsov, S. Sunyaev and A.S. Kondrashov, “Positive selection at sites of multiple amino acid replacements since rat–mouse divergence”, *Nature 429*, 2004, pp. 558-562.

[2] G.A. Bazykin, J. Dushoff, S.A. Levin, and A.S. Kondrashov, “Bursts of nonsynonymous substitutions in HIV-1 evolution reveal instances of positive selection at conservative protein sites”, *Proc Natl Acad Sci USA* 103:51, 2006, pp. 19396–19401.

[3] K.S. Small, M. Brudno, M.M. Hill, and A. Sidow, “Extreme genomic variation in a natural population”, *Proc Natl Acad Sci USA* 104:13, 2007, pp. 5698–5703.

Эволюция недавно дуплицировавшихся паралогов генома человека

Александр Панчин
Московский Госдарственный Университет им. М.В. Ломоносова Факультет Биоинженерии и Биоинформатики
alexpanchin@yahoo.com

Ирена И. Артамонова
Институт общей генетики им. Н.И. Вавилова РАН, Институт Проблем Передачи Информации им. А.А. Харкевича РАН
irenart@gmail.com

Аннотация

При изучении эволюции участков генов, кодирующих белки, часто рассматриваются две величины: количество синонимичных (Ks) и количество несинонимичных (Kn) замен на сайт. Предполагается, что на синонимичные сайты, не влияющие на последовательность кодируемого белка, отбор действует незначительно. Поэтому количество синонимичных замен является "молекулярными часами", по которым можно оценить время расхождения между двумя последовательностями. Соотношение Kn/Ks используется для определения силы и направления отбора.

В данной работе исследовалась эволюция недавно дуплицировавшихся паралогичных семейств генов человека. Большинство паралогичных семейств эволюционируют в условиях отрицательного отбора, который, по-видимому, ослабляется на короткий срок после дупликации. Примерно четверть всех генов таких семейств эволюционирует асимметрично: один из двух ближайших паралогов накапливает несинонимичные замены быстрее, чем другой. То есть, после дупликации одна копия гена может оставаться практически неизменной, в то время как другая претерпевает изменения. Так же встречаются пары ближайших паралогов, не одинаково накапливающие синонимичные замены, что свидетельствует о возможных неточностях измерения времени с помощью "молекулярных часов", использующих Ks.

1. Введение

Изучение эволюции генетических последовательностей играет важную роль в современной биологии. С момента полного прочтения генома человека [1] [2] работы в области биоинформатики стали особенно актуальными: изучая последовательности генов человека (и других организмов), можно узнать об их происхождении, биологической значимости и функции. Чаще всего эволюцию белок-кодирующих генов, характеризуют двумя величинами: количеством синонимичных (Ks) и количеством несинонимичных (Kn) замен, приходящихся на сайт. Время расхождения между двумя последовательностями можно оценить по значениям Ks. Синонимичные замены сравнитель слабо подвержены отбору, в следствии чего они накапливаются с примерно постоянной скоростью в разных последовательсностях.

Мутации в несинонимичных сайтах могут приводить к изменениям кодируемого белка, поэтому они в большей степени подвержены естественному отбору. Для определения направления и силы естественного отбора используется отношение Kn/Ks. Так Kn/Ks = 1, характеризует эволюционную нейтральность, Kn/Ks >> 1 положительный отбор (направленный на сохранение в популяции новых полезных мутаций), а Kn/Ks << 1 отрицательный отбор (направленный на сохранение важных консервативных позиций). Известно, что темпы эволюции гена зависят от того, является ли он уникальным в геноме, или имеет паралогов. Оказалось, что при сравнении генов близких видов эукариот, значения Kn/Ks больше у уникальных генов, чем у генов, представленных в геноме несколькими копиями [3]. Можно допустить, что дуплицируются главным образом гены, несущие важные функции, а также гены, эффект которых зависит от их копийности, и поэтому, с самого начала подверженные сильному отрицательному отбору.

Дупликация генов – явление, которое встречается очень часто. Примерно половина генов человека являются членами дуплицированных семейств [4]. Примерная оценка частоты возникновения дупликаций в геноме – 0.01 на ген на миллион лет. Два гена в

одном геноме, возникшие в результате дупликации, и, следовательно, имеющие похожие нуклеотидные последовательности, называются паралогами. Дупликация гена с эволюционной точки зрения - очень важное событие, так как, в частности, благодаря дупликациям образуются новые гены. Если в геноме существует какой-то ген, выполняющий важную функцию, то, при накоплении в нем случайных мутаций, он может утратить функцию, что, скорее всего, окажется неблагоприятным для организма. На одиночные полезные гены воздействует стабилизирующий отбор, поддерживающий важный ген в консервативном состоянии, что мешает модификации его функции. Если же ген дуплицировался, то, даже если одна копия утратит исходную функцию в результате накопившихся мутаций, эта функция сохранится за счет второй, неизмененной, копии гена. Чаще всего дупликация происходит в результате неравномерной рекомбинации между двумя гомологичными хромосомами, при этом на одной хромосоме оказывается две копии одного гена, а на другой - ни одной.

Целью нашей работы являлось изучение эволюции недавно разошедшихся семейств паралогичных генов человека, а именно:

- С помощью эволюционных тестов выявить наличие или отсутствие глобальных закономерностей между временем, прошедшим с момента дупликации разных представителей одного семейства, и силой действующего на них отбора.
- Выявить наличие или отсутствие асимметрии темпов эволюции близких паралогичных генов, находящихся в одном семействе. В случае наличия асимметрии, исследовать ее возможные причины и последствия.
- Выявить наличие или отсутствие взаимосвязи между возрастом изучаемых семейств и темпами их эволюции.

2. Методы и данные

Для получения семейств паралогичных генов в данной работе использовались белок-кодирующие последовательности мРНК человека из базы данных RefSeq. Повторяющие последовательности были маскированы с помощью программы RepeatMasker (http://www.repeatmasker.org/). Оставшиеся последовательности сравнивались между собой с помощью программы MegaBlast (http://www.ncbi.nlm.nih.gov/blast/megablast.shtml [5]). Похожие гены были объединены в кластеры по принципу “однократного сцепления” (single linkage clustering) – если последовательность похожа хотя бы с одной последовательностью из кластера, она добавляется в кластер.

Для каждого семейства мы построили множественное выравнивание открытых рамок считывания с помощью программы ClustalW. Выравнивания корректировались с учетом аминокислотных последовательностей кодируемых белков.

Для подсчета попарных значений Kn и Ks между представителями семейств паралогичных генов человека использовалась программа codeml из пакета программ PAML (http://abacus.gene.ucl.ac.uk/software/paml.html).

Мы использовали следующую процедуру для вычисления индивидуальных значений Kn и Ks для генов. Для каждого паралога находится ближайший сосед и второй ближайший сосед (аутгрупп). Условия определения гена - аутгруппа следующие: L[1-3] > L[1-2] и L[2-3] > L[1-2], где L[1-3] - расстояние от данного гена, до проверяемого аутгруппа, L[1-2] – расстояние от данного гена, до его ближайшего соседа, L[2-3] расстояние от ближайшего соседа гена, до проверяемого аутгруппа. Расстояние соответствует попарным значениям параметра Time программы codeml. Условия на аутгрупп отображают требования к расположению аутгруппа на эволюционном дереве (Рис. 1). Индивидуальные значения Kn и Ks подсчитываются следующим образом:

Kni = (Kn[1-3] + Kn[1-2] - Kn[2-3])/2

Ksi = (Ks[1-3] + Ks[1-2] - Ks[2-3])/2

Рис.1 Определение индивидуального расстояния до гена. Индивидуальное расстояние до гена 1 – это половина суммы расстояний от гена 1 до аутгруппа (ген 3) и расстояния от гена 1 до его соседа (ген 2) с вычетом расстояния между соседом и аутгруппом гена. В качестве расстояния можно использовать разные параметры, например, параметр Time программы codeml.

Чтобы проверить статистическую значимость гипотезы о асимметричности эволюции пары взаимно ближайших паралогов, мы использовали следующий подход:

Мы рассматривали тройки генов: два взаимно ближайших паралога и общий ближайший аутгрупп. Всего в нашей выборке генов таких троек было 101.

Сравнивались аминокислотные последовательности взаимно ближайших паралогов. Если последовательности таких

паралогов отличались в некоторой позиции, с помощью аутгруппа определялось в каком именно из двух паралогов произошла замена. Подсчитывалось число таких замен в каждом из двух взаимно ближайших паралогов. По наблюдаемой выборке числа замен мы проверили гипотезу о биномиальном, симметричном относительно пары генов (p=q=0.5), распределении замен в каждом случае. Для этого мы вычислили вероятность наблюдаемого или более экстремального распределения числа замен (p-value).

Для проверки гипотезы об асимметричной эволюции пар паралогов мы оценили вероятность наблюдать данное или большее число пар генов с бинарным свойством, частота одного из исходов которого 0.05 (порог значимости в предыдущем эксперименте).

3. Результаты и обсуждение

Наблюдаемые значения Kn почти всегда меньше, чем значение Ks, как для отдельных генов (Рис. 2), так и для генных семейств. Это свидетельствует о том, что выборка генов из семейства с тремя и большим числом паралогов подвержена преимущественно отрицательному отбору. Это согласуется с данными, описанными ранее в литературе. Во-вторых, мы наблюдаем уменьшение угла наклона кривой зависимости Kn от Ks, что свидетельствует об усилении отрицательного отбора по мере увеличения значений Ks как для отдельных генов, так и для семейств генов. Особенно хорошо это видно на Рис. 3. Объяснить данную общую тенденцию можно следующим образом: сразу после дупликации, по-видимому, снижается действие отрицательного отбора и гены начинают быстро эволюционировать, накапливая несинонимичные замены. Как только оба гена расходятся так сильно, что и тот, и другой начинают выполнять значимые для организма различные функции, и на них вновь усиливается давление отрицательного отбора и впоследствии происходят преимущественно синонимичные замены.

Для каждой пары взаимно ближайших паралогов с совпадающим ближайшим аутгруппом мы оценили статистическую значимость согласно алгоритму вычисления p-value, описанному в разделе “Методы”. 101 пара генов из нашей выборки удовлетворяла следующими критериям: они были взаимно ближайшими паралогами, и один и тот же ген являлся ближайшим аутгруппом для обоих генов из пары. 23 таких пары (22,8%) были статистически значимо асимметричны ($P < 0.05$). Пример такой асимметрии приведен на Рис. 4.

Статистическая значимость данной асимметрии (то есть вероятность обнаружить 23 или более значимо асимметричных гена, при пороге 0.05) была оценена при гипотезе биномиального распределении симметричных/асимметричных пар с параметром p=0.05. Она равна 2,8e-10. То есть найденная асимметрия статистически значима. Все статистически значимо асимметричные пары генов приведены в таблице 1. Мы исследовали с помощью программы Virtual Northern (http://cgap.nci.nih.gov/) значимо асимметричные пары паралогов с целью выявить отличия в их уровнях экспрессии и выяснить, влияет ли уровень экспрессии гена на действующий отбор. Уровень экспрессия гена может быть оценен по числу прочитанных последовательностей EST этого гена. (EST – Expression sequence tag - это подпоследовательность сплайсированного транскрипта длиной в несколько сот нуклеотидов, полученная путем секвенирования клонированных кДНК. Для получения кДНК используется обратная транскрипции мРНК из образца).

Одну из возможных гипотез возникновения асимметрии по Kn можно сформулировать так: один из генов в паре имеет низкий уровень экспрессии, поэтому меньше влияет на жизнедеятельность организма и, значит, подвержен более слабому давлению отрицательного отбора. Данные экспрессии представлены в таблице 1 (источник данных по экспрессии Cancer Genome Anatomy Project [6]) Мы не нашли какой либо корреляции между экспрессией генов и скоростью их эволюции. Интересно отметить, что многие из асимметричных генов ассоциированы с раком.

Аналогично асимметрии по числу несинонимичных замен, мы проанализировали число замен в синонимичных позициях для пар взаимно ближайших паралогов с общим аутгруппом. Пар, значимо асимметричных по синонимическим позициям, оказалось 9. Статистическая значимость такого наблюдения при условии истинности гипотезы о биномиальном распределении симметричных/асимметричных пар с параметром p=0.05 составляет 0.026. То есть, асимметрия между генами по числу синонимичных замен не так ярко выражена, как асимметрия по числу несинонимичных замен, но также статистически значима при пороге значимости 0.05.

Таким образом, возникновение асимметрии замен в несинонимичных позициях скорее всего связано с действием отбора, направленного на сохранение одной из копий дуплицировавшегося гена в исходном состоянии. В то же время, изменение аминокислотной последовательности

белка, кодируемого другой копией гена, происходит нейтрально или в условиях значительно более слабого давления отбора.

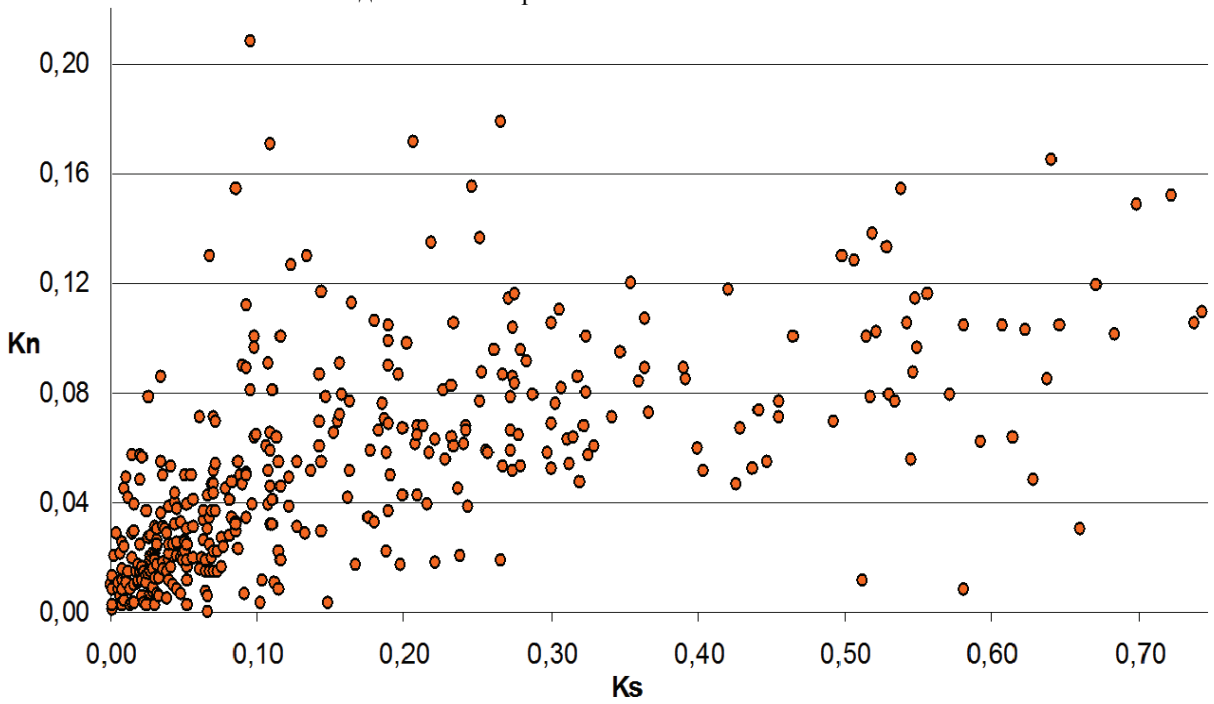


Рис. 2. Соотношения между индивидуальными значениями Kn и Ks для всех генов.

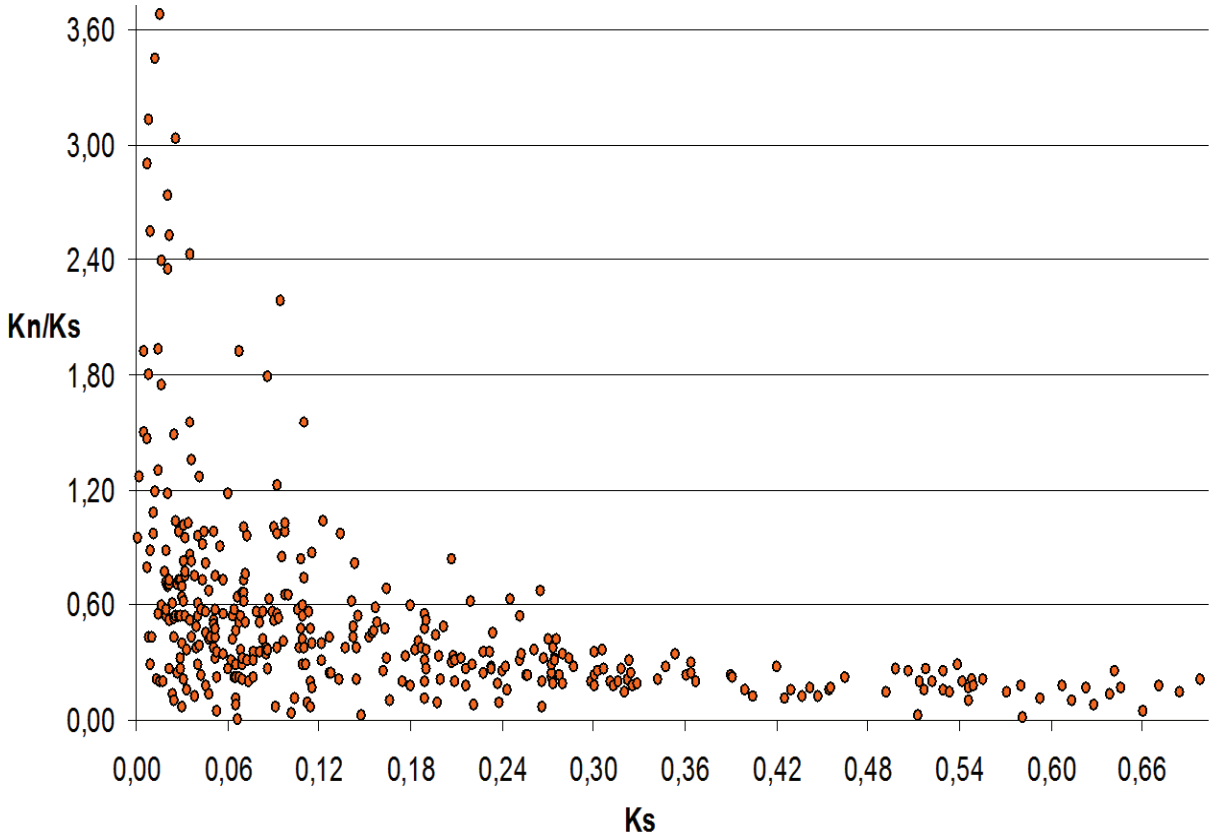


Рис. 3. Соотношения между значениями Kn/Ks, как меры силы действующего отбора, и значениями Ks.

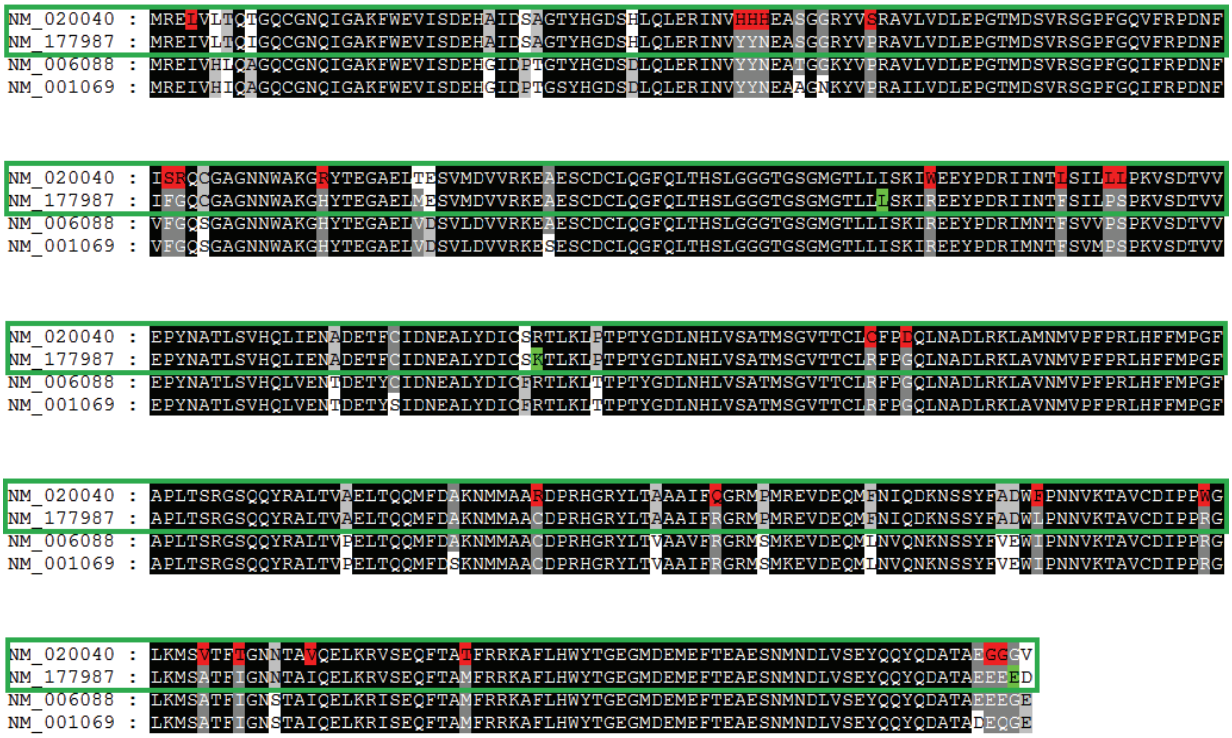


Рис. 4. Пример асимметричной пар взаимно ближайших паралогов (выделены зеленой линией). Красным отмечены аминокислотные замены в быстро эволюционирующем паралоге, зеленым в консервативном паралоге. Представлены 4 представителя бета полипептида тубулина.

Кластер	Ген 1	Ген 2	Замены в гене 1	Замены в гене 2	P value	EST гена 1	EST гена 1 в раковых тканях	EST гена2	EST гена 2 в раковых тканях	Функция
1054	NM_003661	NM_030882	34	17	2,41E-002	297	232	195	124	Apolipoprotein
1054	NM_145660	NM_145640	55	20	6,49E-005	43	30	323	127	Apolipoprotein
1143	NM_001004737	NM_001004736	8	1	3,91E-002	0	0	0	0	Olfactory receptor
11596	NM_002030	NM_001005738	51	15	1,01E-005	31	26	140	0	Formyl peptide receptor-like
1434	NM_001029891	NM_002629	9	1	2,15E-002	156	167	104	72	Butyrophilin
1584	NM_001005471	NM_030904	34	15	9,40E-003	0	0	0		Olfactory receptor
1862	NM_000896	NM_001082	36	15	4,60E-003	40	46	46	4	Cytochrome p450
2382	NM_005557	NM_000526	40	8	3,31E-006	57	382	256	1524	Keratin
2751	NM_001005479	NM_001005514	27	7	8,21E-004	0	0	0	0	Olfactory receptor
2nd1886	NM_020040	NM_177987	24	3	4,92E-005	0	1	1	2	Tubulin, beta polypeptide 4
2nd2341	NM_001097643	NM_176887	26	9	5,99E-003	2	0	N/A	N/A	Taste receptor
3533	NM_006891	NM_005210	26	12	3,36E-002	23	2	6	0	Crystallin, gamma B
3814	NM_152631	NM_001013736	64	36	6,64E-003	19	2	4	1	Unknown function
430	NM_001009562	NM_019844	77	34	5,49E-005	36	2	36	13	Solute carrier organic anion transporter
4666	NM_001006938	NM_001006933	10	1	1,17E-002	16	2	41	25	Transcription elongation factor A
4691	NM_130770	NM_182589	29	8	7,53E-004	1	0	0	0	5-hydroxytryptamine (serotonin) receptor
5920	NM_033179	NM_001004750	40	22	3,00E-002	0	0	0	0	Olfactory receptor
6336	NM_000149	NM_002034	19	6	1,46E-002	35	43	0	1	Fucosyltransferase 5
6559	NM_012373	NM_002551	17	5	1,69E-002	0	0	0	1	Olfactory receptor
8951	NM_001004692	NM_001004695	13	2	7,39E-003	0	0	0	0	Olfactory receptor
937	NM_001004482	NM_001004481	19	2	2,21E-004	0	0	0	0	Olfactory receptor
9751	NM_017534	NM_005963	64	27	1,32E-004	253	26	133	11	Myosin, heavy chain
PCDHGB	NM_032098	NM_032099	48	27	2,03E-002	434	337	N/A	N/A	Protocadherin gamma

Таблица 1. 23 пары асимметричных генов. Указаны функции кодируемых белков, число замен, статистическа значимости асимметрии и число EST по данным Cancer Genome Anatomy Project. Красным отмечены отрицательные ассоциации с раком, оранжевым положительные ($P < 0.03$)

4. Выводы

Гены изученных семейств в значительной степени подвержены отрицательному отбору, что скорее всего связано с важной функцией дуплицированных генов.

Сразу после дупликации, по-видимому, снижается действие отрицательного отбора и гены начинают быстро эволюционировать, накапливая несинонимичные замены

Примерно в четверти пар генов обнаружена статистически значимая асимметрия по числу несинонимичных замен. Эта асимметрия не связана с разницей в уровнях экспрессии двух паралогов. Такая асимметрия может объясняться

тем, что после дупликации одна копия гена сохраняется практически неизменной, а другая копия меняется, приобретая новую функцию.

Обнаружена статистически значимая асимметрия генов по числу синонимичных замен. Возможно, такие факторы, как, например, неодинаковая концентрация в клетке разных Т-РНК, уменьшают точность молекулярных часов, использующих Ks, как единицу измерения времени. Сила отбора, действующая на синонимичные сайты может оказаться большей, чем до сих пор представлялось. Однако это требует более тщательного исследования.

Эта работа поддержана грантом по программе "Молекулярная и клеточная биология" Российской академии наук.

5. Список литературы

[1] E.S. Lander et al., “Initial sequencing and analysis of the human genome”, *Nature*, 2001, 409, 860.

[2] “Finishing the euchromatic sequence of the human genome”, *Nature*, 2004, 431, 931.

[3] I.K. Jordan, E.V. Koonin, and Y.I., Wolf, “Duplicated genes evolve slower than singletons despite the initial rate increase”, *BMC Evolutionary Biology* 2004, 4, 22.

[4] G.M. Rubin et al., “Comparative genomics of the eukaryotes”. *Science*, 2000, 287, 2204.

[5] Z. Zhang, S. Schwartz, L. Wagner, and W. Miller, “A greedy algorithm for aligning DNA sequences”. *J Comput Biol*., 2000, 7, 203.

[6] R.L. Strausberg, K.H. Buetow, M.R. Emmert-Buck, and R.D. Klausner, “The cancer genome anatomy project: building an annotated gene index”. *Trends Genet.*,2000, 16, 103.

CRISPR-системы в метагеномах

Валерий А. Сорокин
Московский Государственный Университет им. М.В. Ломоносова
Факультет Биоинженерии и Биоинформатики
gtr-sorokin@mail.ru
Михаил С. Гельфанд
Институт Проблем Передачи Информации им. А.А. Харкевича РАН
gelfand@iitp.ru
Ирена И. Артамонова
Институт Общей Генетики им. Н.И. Вавилова РАН,
Институт Проблем Передачи Информации им. А.А. Харкевича РАН
irenart@gmail.com

Аннотация

Новый тип прокариотического иммунитета был недавно открыт и описан в общих чертах. Он получил название CRISPR-систем (от "Clusters of Regularly Interspaced Short Palindromic Repeats"). Такие системы состоят из кассет - специальных структур в ДНК, и белков, ассоциированных с этими кассетами. CRISPR-системы осуществляют защиту прокариотической (бактериальной или архейной) клетки от чужеродной ДНК, в частности, от атак фагов или плазмид.

Мы разработали схему поиска CRISPR-кассет в метагеномах. Анализ CRISPR-кассет самого большого на сегодняшний день метагенома Sorcerer II показал, что метагеномы обеднены такими кассетами по сравнению с отсеквенированными полными прокариотическими геномами в среднем. По паттерну сходства между двумя метагеномными контигами можно различить возможных агентов генетической агрессии и случаи горизонтального переноса всей кассеты между близкородственными и/или более далекими биологическими видами. Создана база данных, предоставляющая доступ к CRISPR-кассетам метагеномов.

1. Введение

Кассеты из коротких повторов, разделенных уникальными спейсерами, составляют основу так называемых CRISPR-систем (Clustered of Regularly Interspaced Short Palindromic Repeats). Эти системы служат прокариотическим иммунитетом нового, недавно открытого, типа, направленным против инородного генетического материала [1]. Среди всех прокариотических организмов, изученных на настоящий день, примерно половина всех бактерий и почти все археи обладают таким иммунитетом [2].

Типичная CRISPR-система состоит из кассеты, лидерной области и пула ассоциированных с системой белков (так называемые, cas- или CRISPR-associated proteins). Кассеты представляют собой серию последовательных высоко консервативных коротких (23-47 nt) прямых повторов, разделенных близкими по длине уникальными спейсерами, часто имеющими чужеродное (фаговое или плазмидное) происхождение [3]. Среди белков, ассоциированных с CRISPR-системой, выделяют до 45 функциональных семейств, часть из которых имеет, по-видимому, эндонуклеазную активность. В одном геноме встречается, как правило, от двух до дюжины cas-белков из различных семейств.

В данной работе мы исследовали распределение CRISPR-кассет в метагеномах (совокупностях геномов микроорганизмов, обитающих в одной экологической нише).

2. Результаты и обсуждение

2.1 Выявление CRIPSR-кассет в метагеномах

Основным недостатком всех трех существующих программ, разработанных для поиска CRISPR-кассет в геноме, является высокий уровень перепредсказания. Характерная структура метагенома создает дополнительные трудности для поиска таких кассет. Поэтому на данный момент не существует эффективного алгоритма, реализующего эту задачу. Мы

разработали схему для поиска CRISPR-кассет в метагеномах, основанную на комбинации всех трех существующих программ.

Так как каждая из программ построена на своем алгоритме, особенности всех программ и кассет, найденных каждой из них, различны. Поэтому мы сформировали первоначальную выборку из кассет, найденных всеми тремя программами. Повторы CRISPR-кассет имеют тенденцию к кластеризации по сходству последовательностей [4]. Поэтому мы добавили к этой выборке все кассеты с повторами, гомологичными повторам в первоначальной выборке. Кроме того, все кассеты, найденные хотя бы одной из программ, расположенные по близости от идентифицируемых cas-белков также участвовали в дальнейшем анализе.

2.2 Анализ CRISPR-кассет океанического метагенома Sorcerer II

В результате приложения описанной выше схемы к данным океанического метагенома, известного как Sorcerer II, было выявлено 192 CRISPR-кассет с 1939 уникальными спейсерами.

Была создана база данных, включающая все идентифицированные кассеты и предоставляющая быстрый доступ к ним. Все кассеты были разбиты на семейства по сходству последовательностей их повторов.

Анализ фланкирующих участков позволяет отличить кассеты, являющиеся результатом горизонтального переноса, от параллельной эволюции кассет родственных бактериальных штаммов. Но в большинстве случаев сходство последовательностей двух контигов, сосредоточенное внутри кассеты, найденной в одном из них, ограничивалось коротким участком, включающим единственный спейсер. По-видимому, такая модель соответствует сходству спейсера исследуемой кассеты и его прототипа в геноме агрессора. Выявление такого сходства может служить инструментом поиска фагового/плазмидного генетического материала среди всего нуклеотидного разнообразия метагенома.

Оба типа нуклеотидного сходства – сходство спейсера и его прототипа, а также сходство целых кассет, являющихся результатом горизонтального переноса – статистически значимо преобладают среди метагеномных контигов с одинаковой географической локализацией. Эти наблюдения подтверждают гипотезу об участии CRISPR-систем в защите прокариотических организмов, формирующих метагеном, от вторжения чужеродной ДНК, присутствующей в той же экологической нише.

Эта работа поддержана грантом по программе “Молекулярная и клеточная биология” Российской академии наук.

5. Список литературы

[1] Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, and Horvath P. “CRISPR provides acquired resistance against viruses in prokaryotes.” *Science*, 2007, 315, 1709-12.

[2] Grissa I, Vergnaud G, and Pourcel C. “CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats.” *Nucleic Acids Res.*, 2007, 35, W52-7.

[3] Grissa I, Vergnaud G, and Pourcel C. “The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats.” *BMC Bioinformatics*, 2007, 8, 172.

[4] Kunin V, Sorek R, and Hugenholtz P. “Evolutionary conservation of sequence and secondary structures in CRISPR repeats.” *Genome Biol.*, 2007, 8, R61.

ДНК коды, основанные на весовом стебельном сходстве Хэмминга

А.Г. Дьячков А.Н. Воронина
кафедра Теории Вероятностей, Механико-математический факультет, МГУ им. Ломоносова, Москва 119992, РОССИЯ
agd-msu@yandex.ru vorronina@gmail.com

Аннотация

В настоящей статье мы продолжаем исследование [1] функций сходства для q-ичных последовательностей, которые для случая $q=4$ могут быть использованы для моделирования энергии гибридизации ДНК-последовательностей. Коды, основанные на подобных функциях сходства, называются ДНК-кодами [1]. В настоящей статье исследуется аддитивная функция сходства, называемая весовым сходством Хэмминга. На множестве пар (блоков длины 2) элементов из $[0,...,q-1]$ задается весовая функция, обладающая свойством инвариант-ности относительно преобразования Ватсона-Крика [2]. Весовым стебельным сходством Хэмминга между последовательностями x и y называется сумма весов всех блоков длины 2, состоящих из последовательных символов x и y, содержащихся в наибольшей общей подпоследовательности Хэмминга между x и y. В работе получены верхняя граница Плоткина и нижняя граница Варшамова-Гильберта скорости ДНК-кодов, основанных на такой аддитивной функции сходства.

1. Введение

Одинарные цепочки ДНК могут рассматриваться как направленные четверичные последовательности с элементами из алфавита (*A,C,G,T*). При этом буква *A* является *комплементарной* к букве *T*, а *C* – к *G*.

Процесс формирования двойных спиралей ДНК путем образования водородных связей между комплементарными буквами из двух одинарных разнонаправленных цепочек ДНК называется *ДНК гибридизацией*. Согласно модели «Ближайшего соседа» *энергия* гибридизации может быть вычислена как сумма энергий (весов) всех образовавшихся *пар* (блоков длины 2) подряд идущих водородных связей.

Комплементарная к заданной цепочке ДНК может быть получена путем замены всех элементов цепочки на комплементарные и последующего изменения направления цепочки. Понятно, что максимальная энергия гибридизации может быть получена именно при соединении 2 комплементарных цепочек.

В биологических приложениях требуется формирование ансамблей, состоящих из пар взаимно-комплементарных цепочек ДНК, таких что при заданных условиях объединяться в двойные спирали смогут только комплементарные цепочки. Это накладывает естественные условия на максимальную энергию гибридизации между некомплементарными цепочками.

Мы определим специальную функцию сходства, позволяющую математически моделировать энергию гибридизации, благодаря чему описанную выше задачу можно будет естественным образом интерпретировать в терминах теории кодирования.

2. Определения, обозначения и примеры

Символ $\equiv$ обозначает равенство по определению. Множество целых чисел от 1 до *n* будем обозначать через [*n*]. Через $\lfloor u \rfloor$ **Ошибка! Закладка не определена.** $(\lceil u \rceil)$ будем обозначать целую часть снизу (сверху) вещественного числа u. Для $q=2,4,...$ – произвольного целого четного положительного числа – через $A \equiv \{0,1,...,q-1\}$ обозначим стандартный q-ичный алфавит размера $|A|=q$.

2.1. Общие определения

Для любого элемента $x \in A$ определим элемент $\bar{x} \equiv (q-1) - x \in A$, называемый *комплементарным* к элементу x. Для любой q-ичной последовательности длины n $x = (x_1, x_2, ..., x_n) \in A^n$ определим *комплементарную* к ней последовательность как $\tilde{\bar{x}} \equiv (\bar{x}_n, \bar{x}_{n-1}, ..., \bar{x}_2, \bar{x}_1) \in A^n$. Если $y \equiv \tilde{\bar{x}}$, то $x = \tilde{\bar{y}}$ для любого $x \in A^n$. Если $x = \tilde{\bar{x}}$, x называется *самосопряженной* последовательностью. Если $x \neq \tilde{\bar{x}}$, то пара $\left(x, \tilde{\bar{x}}\right)$ называется *парой взаимно комплементарных* последовательностей.

Пусть $w = w(a,b) \geq 0$, $a,b \in A$, является весовой функцией, определенной на A^2, и

$$w(a,b) = w(\bar{b}, \bar{a}), \qquad a,b \in A. \tag{1}$$

Условие (1) означает, что $w(a,b)$ инвариантна относительно преобразования Ватсона-Крика [2].

Рассмотрим две произвольных q-ичных последовательности длины n: $x = (x_1, x_2, ..., x_n)$, $y = (y_1, y_2, ..., y_n) \in A^n$.

Определение 1. Число

$$S_w(x,y) \equiv \sum_{i=1}^{n-1} s_i^w(x,y), \tag{2}$$

где $s_i^w(x,y) = w(a,b)$, если $x_i = y_i = a$ и $x_{i+1} = y_{i+1} = b$, и $s_i^w(x,y) = 0$ в противном случае, называется *весовым стебельным сходством Хэмминга* между x и y.

В виду (1)-(2), для любых $x, y \in A^n$ сходство $S_w(x,y) = S_w(y,x) \leq S_w(x,x)$ и $S_w(x, \tilde{\bar{y}}) = S_w(y, \tilde{\bar{x}})$.

Таким образом, функция $S_w(x, \tilde{\bar{y}})$ может быть использована для моделирования [1] *термодинамического сходства* (*энергии гибридизации*) между ДНК-последовательностями x, y и мы можем дать следующее биологически обоснованное [2], [4] определение расстояния между ДНК последовательностями.

Определение 2. Число

$$D_w(x,y) \equiv S_w(x,x) - S_w(x,y), \quad x,y \in A^n \tag{3}$$

называется *весовым стебельным расстоянием Хэмминга между* x *и* y. Заметим, что в большинство случаев $D_w(x,y) \neq D_w(y,x)$, то есть функция (3) не является симметичной.

2.2. Пример весовой функции

Весовая функция может задана таблицей, полученной из данных эксперимента, такой как, например, Таблица 1 из работы [2] (см. Таблицу 1).

Эта весовая функция представляет существенный практический интерес и мы приведем ниже графики верхней и нижней границ скорости для ДНК-кодов с данными весами.

2.3. ДНК коды, основанные на весовом стебельном сходстве Хэмминга

Пусть $x(1), x(2), ..., x(N)$, где $x(j) \equiv (x_1(j), x_2(j), ..., x_n(j)) \in A^n$, $j \in [N]$, являются *кодовыми словами* q-ичного *кода* $X = \{x(1), x(2), ..., x(N)\}$ *длины* n и *размера* N, где N=2,4,... *четно*. Пусть D, $0 < D < n-1$ – произвольное целое число.

Определение 3. Код X называется *ДНК* (n,D,w)-*кодом*, основанным на весовом стебельном сходстве Хэмминга, если выполнены следующие 2 условия: (i). Для любого $j \in [N]$ найдется $j' \in [N]$, $j' \neq j$, такое что $x(j') = \tilde{\bar{x}}(j) \neq x(j)$. Другими словами, X является набором из $N/2$ пар взаимно комплементарных последовательностей длины n.

(ii). Для любых $j' \neq j$ расстояние

$$D_w\left(x(j), x(j')\right) \geq D. \tag{4}$$

Обозначим через $N_w(n,D)$ максимальный размер ДНК (n,D,w)-кода с расстоянием (3). При фиксированном $d > 0$

$$R_w(d) \equiv \overline{\lim_{n \to \infty}} \frac{\log_q N_w(n, dn)}{n} \tag{5}$$

называется *скоростью* ДНК кода, основанного на весовом стебельном сходстве Хэмминга с расстоянием (3). В данной работе мы получим верхнюю и нижнюю границы скорости $R_w(d)$.

3. Верхняя граница Плоткина для $R_w(d)$

Пусть $p = p(a,b)$, $a,b \in A$, – произвольное распределение вероятностей на множестве пар A^2, то есть $\sum_{a,b \in A} p(a,b) = 1$ и $p(a,b) \geq 0$ для любых $a,b \in A$.

Таблица 1. Пример весовой функции

w(a,b)	*b=A(=0)*	*b=C(=1)*	*b=G(=2)*	*b=T(=3)*
a=A(=0)	1,02	1,46	1,29	0,88
a=C(=1)	1,46	1,83	2,17	1,29
a=G(=2)	1,32	2,24	1,83	1,46
a=T(=3)	0,60	1,32	1,46	1,02

Для фиксированной весовой функции $w = w(a,b)$, $a,b \in A$, удовлетворяющей условию (1), введем величину

$$T_w \equiv \max_{(7),(8)} \left\{ \sum_{a,b \in A} \left(p(a,b) - p^2(a,b) \right) w(a,b) \right\}, \quad (6)$$

где максимум берется по всем распределениям вероятности p, для которых выполняется следующие 2 условия:

$$\sum_{a \in A} p(a,b) = \sum_{a \in A} p(b,a) \quad \forall b \in A; \quad (7)$$

$$p(a,b) = p(\bar{b},\bar{a}) \quad \forall a,b \in A. \quad (8)$$

Стандартными методами может быть получена верхняя граница скорости $R_w(d)$:

Теорема 1. (Граница Плоткина) Если $d \ge T_w$, то $R_w(d) = 0$; кроме того,

$$R_w(d) \le \overline{R}_w(d) \equiv 1 - \frac{d}{T_w}, \text{ если } 0 < d < T_w. \quad (9)$$

4. Нижняя граница Варшамова-Гильберта для $R_w(d)$

Зафиксируем некоторое распределение вероятностей $p = p(a,b)$, удовлетворяющее условиям (7)-(8). Пусть $x = (x_1, x_2, ..., x_n) \in A^n$ и $y = (y_1, y_2, ..., y_n) \in A^n$ являются *независимыми* случайными последовательностями, построенными как цепи Маркова со стационарным начальным распределением $p(a)$, $a \in A$, и стационарной переходной матрицей $p(b \mid a)$, $a,b \in A$, такими что $p(b \mid a)p(a) = p(a,b)$ для любых $a,b \in A$.

Введем случайные величины η_i^p, такие что $\eta_i^p = 0$, если $x_i = y_i$ и $x_{i+1} = y_{i+1}$, и $\eta_i^p = w(x_i, x_{i+1})$ иначе. Тогда их сумма

$$S_n^p \equiv \sum_{i=1}^{n-1} \eta_i^p = D_w(x,y). \quad (10)$$

Из определения η_i^p сразу следует, что распределение вероятностей η_i^p, $i = 1,2,...,n-1$ $p_\eta(k) = \Pr\{\eta_i^p = k\}$ имеет следующий вид:

$$p_\eta(0) = \sum_{a,b \in A} p^2(a) p^2(b), \quad (11)$$

$$p_\eta(w(a,b)) = p(a,b) \sum_{(c,d) \ne (a,b)} p(c,d). \quad (12)$$

Обозначим через $E\eta$ математическое ожидание η. Для η_i^p и S_n^p будем иметь

$$E\eta_i^p = \sum_{(c,d) \ne (a,b)} w(a,b) p(a,b) p(c,d) =$$

$$= \sum_{a,b,c,d \in A} w(a,b) p(a,b) p(c,d) -$$

$$- \sum_{(c,d) = (a,b)} w(a,b) p(a,b) p(c,d) =$$

$$= \sum_{a,b \in A} w(a,b) p(a,b) \sum_{c,d \in A} p(c,d) -$$

$$- \sum_{a,b \in A} w(a,b) p^2(a,b) =$$

$$= \sum_{a,b \in A} \left(p(a,b) - p^2(a,b) \right) w(a,b),$$

$$ES_n^p = E \sum_{i=1}^{n-1} \eta_i =$$

$$= (n-1) \sum_{a,b \in A} \left(p(a,b) - p^2(a,b) \right) w(a,b).$$

Зафиксируем d, $0 < d < T_w$. Для фиксированного p введем функцию

$$\underline{R}_w(d,p) \equiv \varliminf_{n \to \infty} \frac{-\log_q \Pr\{D_w(x,y) \le dn\}}{n} =$$

$$= \varliminf_{n \to \infty} \frac{-\log_q \Pr\{S_n^p \le dn\}}{n}. \quad (13)$$

С помощью метода случайного кодирования для ДНК кодов [1] получаем

Предложение 1. Для $0 < d < T_w$ скорость $R_w(d)$ ДНК кодов, основанных на весовом стебельном Хэмминговском сходстве удовлетворяет неравенству

$$R_w(d) \ge \underline{R}_w(d) \equiv \max_{(7),(8)} \{\underline{R}_w(d,p)\}, \quad (14)$$

где $\underline{R}_w(d,p)$ определена в (13), а максимум берется по всем распределениям вероятности $p = p(a,b)$, удовлетворяющим условию (7)-(8).

Функция $\underline{R}_w(d)$ называется *границей случайного кодирования* (или границей Варшамова-Гильберта) для скорости (5) ДНК кодов, удовлетворяющих условию (4).

4.1. Вычисление границы случайного кодирования

Для заданного $p = p(a,b)$ для суммы случайных величин $S_n^p = \sum_{i=1}^{n-1} \eta_i^p$ введем

производящую функцию моментов

$$G_n(u,p) \equiv \sum_{a=0}^{n-1} \Pr\{S_n^p = a\} q^{ua} = E q^{uS_n^p} \qquad (15)$$

и *производящую функцию семиинвариантов*

$$\mu_n(u,p) \equiv \log_q G_n(u,p), \; -\infty < u < \infty. \qquad (16)$$

Определим случайные векторы

$$\xi_i^p \equiv (x_i, x_{i+1}, y_i, y_{i+1}), \quad i = 1,...,n-1. \qquad (17)$$

Можно проверить, что случайная последовательность ξ_i^p образует стационарную марковскую цепь с переходными вероятностями $p_\xi(t_1 \mid t_2) = \Pr\{\xi_i^p = t_1 \mid \xi_{i-1}^p = t_2\}$, где $t_i = (a_i, b_i, c_i, d_i)$. Легко видеть, что $p_\xi(t_1 \mid t_2) = 0$, если $a_1 \neq b_2$ или $c_1 \neq d_2$, и $p_\xi(t_1 \mid t_2) = p(b_1 \mid a_1) p(d_1 \mid c_1)$, если $a_1 = b_2$ и $c_1 = d_2$.

Кроме того, η_i^p, $i = 1,...,n-1$, определенные выше, могут быть записаны в виде:

$$\eta_i^p = f(\xi_i^p) \equiv \delta(\xi_i^p) w(x_i, x_{i+1}), \qquad (18)$$

где $\delta(\xi_i^p) = 0$, если $x_i = y_i$ и $x_{i+1} = y_{i+1}$, и $\delta(\xi_i^p) = 1$ в противном случае.

Таким образом, последовательность η_i^p является *детерминированной* функцией от марковской цепи (17)[1].

Для фиксированного $p = p(a,b)$ введем матрицу размера $(q^4 \times q^4)$:

$$P(u,p) \equiv \left\| p_\xi(t_j \mid t_l) q^{uf(t_j)} \right\|_{j,l=1}^{q^4}, \; -\infty < u < \infty,$$

где $t_j = (t_j^1,...,t_j^4)$, $j = 1,...,q^4$, являются упорядоченными в лексикографическом порядке возможными значениями случайного вектора ξ_i^p.

Пусть $\lambda(u,p)$ – собственное значение Перрона-Фробениуса [7] матрицы $P(u,p)$, то есть такое действительное собственное значение, что любое другое собственное значение $\hat{\lambda}$ не больше его по модулю: $|\hat{\lambda}| \le \lambda(u,p)$. Такое собственное значение существует согласно теореме Перрона-Фробениуса [7].

Теперь, используя стандартные методы доказательства [6] (стр. 230-232), можно вычислить производящии функции (15)-(16) и получить следующую асимптотическую $(n \to \infty)$ формулу:

$$\mu_n(u,p) = \log_q G_n(u,p) = n\mu(u,p) + O(1),$$

$$\mu(u,p) = \log_q \lambda(u,p).$$

Наконец, применим Принцип Больших Уклонений [7] к S_n и получим

Теорема 2. (Граница Варшамова-Гильберта). Для любого $0 < d < T_w$ скорость

$$R_w(d) \ge \underline{R}_w(d) \equiv \max_{(7),(8)} L(d,p), \qquad (19)$$

где максимум берется по всем распределениям вероятности $p = p(a,b)$, удовлетворяющим условиям (7)-(8) и

$$L(d,p) \equiv \max_{u \le 0} \{ud - \mu(u,p)\}, \qquad (20)$$

где $\mu(u,p) = \log_q \lambda(u,p)$. Для любого $p = p(a,b)$ $L(d,p)$ удовлетворяет свойствам

$$L(E\eta_i^p, p) = 0, \qquad (21)$$

$$L(d,p) > 0, \qquad 0 < d < E\eta_i^p. \qquad (22)$$

Кроме того, нижняя граница $\underline{R}_w(d) > 0$, если $0 < d < T_w$. Таким образом, для любого d, $0 < d < T_w$, скорость $R_w(d) > 0$.

Замечание 3. Если для построения Марковских цепей x и y мы возьмем распределение $p_{\max} = p_{\max}(a,b)$, для которого $\sum_{a,b \in A} \left(p_{\max}(a,b) - p_{\max}^2(a,b) \right) w(a,b) = T_w$, то $L(d, p_{\max}) > 0$ для $0 < d < E\eta_i = T_w$. С другой стороны, если для построения x и y мы возьмем p_{iid}: $p_{iid}(a,b) = \frac{1}{q^2}$ для любых $a,b \in A$, то $L(1, p_{iid}) = 1$. Таким образом, верхняя граница Плоткина и нижняя граница Варшамова-Гильберта имеют одинаковые

[1] Следует отметить, что сами η_i^p, $i = 1,...,n-1$, не являются марковской цепью. Например, для любого i, $3 \le i \le n-1$, для любого $k = w(a,b)$, условная вероятность $\Pr\{\eta_i^p = 0 \mid \eta_{i-1}^p = k, \eta_{i-2}^p = 0\} = 0$, но $\Pr\{\eta_i^p = 0 \mid \eta_{i-1}^p = k\} \neq 0$.

критические точки: $\underline{R}_w(0) = \overline{R}_w(0) = 1$ и $\underline{R}_w(T_w) = \overline{R}_w(T_w) = 0$. На Графике 1 приведены графики функций $L(d, p_{\max})$ и $L(d, p_{iid})$.

Замечание 4. Для приведенной в Таблице 1 весовой функции мы вычислили следующее максимизирующее распределение: см. Таблицу 2

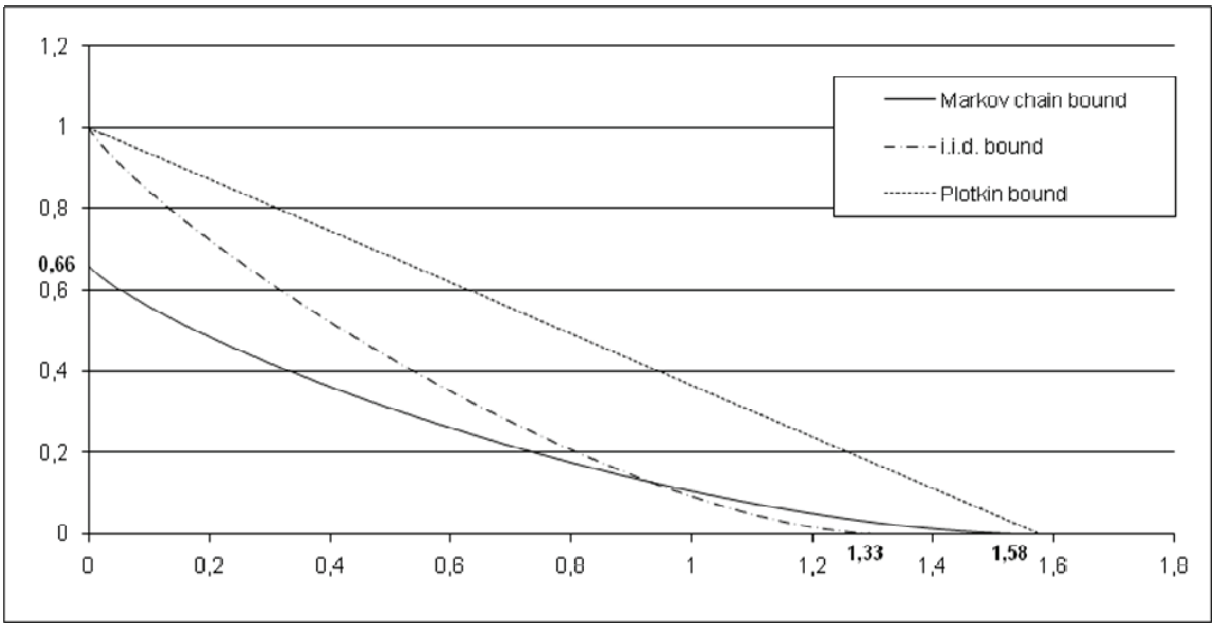


График 1. Границы скорости ДНК кода для q=4 для весовой функции из Таблицы 1

Таблица 2. Совместное распределение

p(a,b)	*b=A(=0)*	*b=C(=1)*	*b=G(=2)*	*b=T(=3)*
a=A(=0)	0,00	0,06	0,01	0,00
a=C(=1)	0,06	0,15	0,21	0,01
a=G(=2)	0,01	0,21	0,15	0,06
a=T(=3)	0,00	0,01	0,06	0,00

Таблица 3. Условное распределение

p(b|a)	*b=A(=0)*	*b=C(=1)*	*b=G(=2)*	*b=T(=3)*
a=A(=0)	0,00	0,87	0,13	0,00
a=C(=1)	0,14	0,35	0,49	0,02
a=G(=2)	0,02	0,49	0,35	0,14
a=T(=3)	0,00	0,12	0,88	0,00

и Таблицу 3. При указанном распределении $T_w = 1{,}58$. Это означает в частности, что при больших длинах n ДНК последовательностей экспоненциальный рост числа кодовых слов можно ожидать только для кодов с расстоянием $D \le 1.58n$.

4. Библиография

[1] А.Г. Дьячков, П.А. Виленкин, И.К. Исмагилов, Р.С. Сарбаев, А. Макула, Д. Торни, С. Уайт, "О ДНК кодах", *Проблемы Передачи Информации*, 2005, т.41, н.4, стр. 57-77.

[2] A. D'yachkov, A. Macula, T. Renz and V. Rykov, "Random Coding Bounds for DNA Codes Based on Fibonacci Ensembles of DNA Sequences", *Proc. of 2008 IEEE International Symposium on Information Theory, Toronto, Canada*, 2008, *in print*.

[3] A.G. D'yachkov, A.N. Voronina, "DNA Codes Based on Stem Hamming Similarity", *Proc. of Eleventh International Workshop on Algebraic and Combinatorial Coding Theory, Pamporovo, Bulgaria*, 2008.

[4] M.A. Bishop, A.G. D'yachkov, A.J. Macula, T.E. Renz, V.V. Rykov, "Free Energy Gap and Statistical Thermodynamic Fidelity of DNA Codes", *Journal of Computational Biology*, 2007, V. 14, N. 8, P. 1088-1104.

[5] F.J. MacWilliams, N.J.A. Sloane, *The Theory of Error-correcting Codes*, North Holland, Amsterdam, The Netherlands, 1977.

[6] В.Н. Тутубалин, *Теория вероятностей и случайных процессов*, Издательство МГУ, Москва, 1992.

[7] A. Dembo, O. Zeitouni, *Large Deviations Techniques and Applications*, Jones and Bartlett, Boston, MA, 1993.

[8] A.G. D'yachkov, D.C. Torney, "On Similarity Codes", *IEEE Trans. Inform. Theory*. 2000, V. 46. N. 4. P. 1558-1564.

Поиск вторичных структур в интронах представителей рода Drosophila

Храмеева Е. Е., Миронов А. А., Гельфанд М. С., Первушин Д. Д.
Московский государственный университет им. М.В. Ломоносова, Москва, Россия
khrameeva@yandex.ru

Аннотация

У многих организмов вторичная структура РНК оказывает значительное влияние на процессы транскрипции и трансляции, однако влияние таких структур на процесс сплайсинга до сих пор систематически не изучено. В данной работе был осуществлен сравнительный анализ интронов восьми представителей рода Drosophila на предмет наличия в них консервативных комплементарных мотивов, способных формировать стабильные РНК-структуры. Такие мотивы длиной 9 нуклеотидов были найдены в 52 интронах. Для большинства интронов данные участки можно продлить, причем исходный 9-нуклеотидный участок служит «затравкой», от которой расходятся в различной степени протяженные участки вторичной структуры. Для полученного набора интронов был осуществлен поиск альтернативных вторичных структур, т.е. таких, в которых кроме уже найденных комплементарных участков (A и B) имеется третий участок (C), также комплементарный A. Наличие альтернатив было показано для 35 из 52 интронов. Мы предполагаем, что в найденных генах регуляция сплайсинга осуществляется при помощи альтернативных вторичных структур.

1. Введение

Со времен открытия сплайсинга исследователей интересовал вопрос: какое влияние оказывает вторичная структура на процесс сплайсинга? Помимо цис-элементов, непосредственно определяющих границы экзонов (донорный и акцепторный сайты, “branch point” и полипиримидиновый тракт), на продукт сплайсинга значительное влияние оказывают энхансеры и сайленсеры. Энхансеры взаимодействуют со вспомогательными транс-факторами (как правило, серин-аргинин богатыми белками), которые воздействуют на остальные части сплайсеосомы, изменяя ход сплайсинга. Случаи взаимодействия цис-элементов друг с другом до сих пор считались редкими по сравнению с взаимодействиями между цис-элементами и транс-факторами. Результаты данной работы показывают, что механизмы сплайсинга, опосредованные вторичной структурой, могут быть более распространенными, чем предполагалось ранее.

В настоящее время экспериментально показано наличие порядка 15 случаев участия вторичной структуры в регуляции сплайсинга у различных организмов [1]: у вирусов (вирус гепатита B [2], аденовирус [3], вирус иммунодефицита человека типа 1 [4], вирус саркомы Рауса [5]), дрожжей [6,7], растений (*Nicotiana plumbaginifolia* [8]), насекомых (*Drosophila* [9]), а также у крыс и мышей [10]. Показано, что в организме человека вторичные структуры могут оказывать влияние на эффективность распознавания сайтов сплайсинга и таким образом участвовать в формировании изоформ гормона роста [11], гена *tau* [12], гена *Hprt* [13] и гена *hnRNPA1* [14]. Ошибки сплайсинга, обусловленные влиянием вторичных структур РНК, приводят к таким патологиям человека, как мышечная дистрофия [15], кистозный фиброз [16] и паркинсонизм [11].

Можно предположить, что вторичные структуры могут участвовать в регуляции сплайсинга по двум механизмам. (1) Наличие структурных элементов может ограничивать доступ участков, которые непосредственно определяют границы экзона (3′-сайты сплайсинга, 5′-сайты сплайсинга и “branch point”), для основных факторов сплайсинга [6]. (2) Если данные участки не входят в состав вторичных структур, то вторичные структуры могут регулировать эффективность сплайсинга, способствуя изменению расстояний между 3′-сайтами сплайсинга, 5′-сайтами сплайсинга и “branch point” [7].

В данной работе осуществлен сравнительный анализ интронов восьми представителей рода *Drosophila* на предмет наличия в них консервативных комплементарных мотивов,

способных формировать стабильные РНК-структуры. Исследованы окружения цис-элементов (а именно, донорных и акцепторных сайтов сплайсинга) с целью поиска элементов вторичной структуры РНК.

2. Материалы и методы

Объектом исследования в данной работе являются последовательности интронов представителей рода *Drosophila*. На данный момент сиквенированы геномы двенадцати представителей рода *Drosophila*. В данной работе мы исследовали только 8 наиболее близких видов (*D. melanogaster, D. sechellia, D. simulans, D. erecta, D. yakuba, D. ananassae, D. pseudoobscura, D. persimillis*). Для них была составлена база данных, содержащая порядка 50 тысяч интронов, принадлежащих 8 тысячам ортологичных генов.

В данной работе была поставлена задача поиска элементов вторичной структуры в окружении сайтов сплайсинга, а именно консервативных комплементарных мотивов, которые были способны формировать стабильные РНК-структуры. Для поиска таких мотивов рассматривались окружения донорных и акцепторных сайтов сплайсинга в интронах ортологичных генов восьми представителей рода *Drosophila*, при этом под окружением понимались 200 нуклеотидов интрона и 50 нуклеотидов экзона (рисунок 1). Если длина интрона была менее 400 нуклеотидов, его просто делили пополам.

По последовательностям окружений соответствующих сайтов сплайсинга из разных организмов были построены множественные выравнивания с помощью программы MUSCLE [17]. В каждом таком множественном выравнивании был осуществлен поиск консервативных участков длиной 9 нуклеотидов с одной разрешенной неконсервативной позицией. Далее осуществлялась проверка пар консервативных участков из донорного и акцепторного окружения на комплементарность, допуская одну GT пару на 9 нуклеотидов.

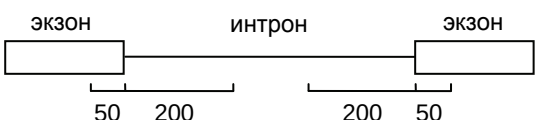


Рисунок 1. Область поиска консервативных комплементарных участков.

Кроме того, требовали присутствия хотя бы двух GC пар на 9 нуклеотидов, чтобы избежать нахождения AT-богатых последовательностей.

Процент ложных положительных предсказаний оценивали при помощи процедуры рандомизации. В результате каждой донорной части интрона оказалась сопоставленной акцептоная часть чужого интрона (рисунок 2). Далее для каждого нового «интрона» проводили обычную процедуру поиска консервативных комплементарных участков (см. выше). Процедуру рандомизации повторяли 100 раз, получая на каждом шаге число «интронов», в которых найдены консервативные комплементарные участки. По этим данным вычисляли среднее значение, стандартное отклонение и строили доверительный интервал.

Продление найденных консервативных комплементарных участков осуществлялось следующим образом. Для каждого интрона, в котором найдены консервативные комплементарные участки, были построены попарные выравнивания донор-акцептор для последовательностей из одного организма (по алгоритму Смита-Ватермана) [18]. При этом были найдены все возможные субоптимальные вторичные структуры для каждого интрона. Структуры с весом менее определенного порога не рассматривались (значение порога задается пользователем).

3. Результаты и обсуждение

3.1. Вторичные структуры найдены в 52 интронах

Консервативные комплементарные участки длиной 9 нуклеотидов в окружении донорных и акцепторных сайтов были найдены для 52 интронов. Процент ложных положительных предсказаний оценивался при помощи процедуры рандомизации (см. Материалы и методы) и составил 11±1% (на 95% уровне доверия).

На рисунке 3 приведен пример одного из 52-х интронов, в которых найдены консервативные комплементарные участки (интрон из гена CG11876).

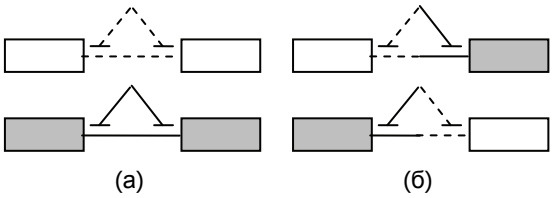


Рисунок 2. Контроль процедуры поиска консервативных комплементарных участков: (а) в опыте; (б) в контроле.

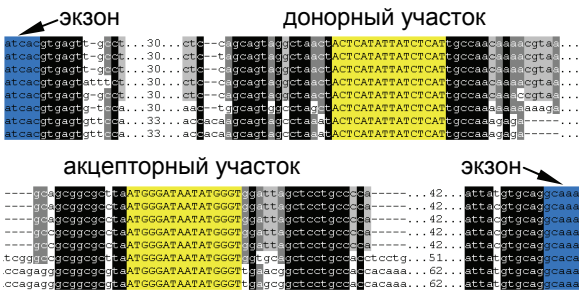


Рисунок 3. Консервативные комплементарные участки в интроне гена CG11876.

Здесь можно наблюдать типичную для всего набора интронов картину – консервативные комплементарные участки погружены в консервативную область, которая постепенно «размывается». При этом необходимо отметить, что длина консервативного участка как правило больше 9 нуклеотидов. Комплементарность также сохраняется на большей протяженности, чем 9 нуклеотидов.

Таким образом, для большинства интронов найденные участки вторичной структуры можно продлить. Все 52 интрона, в которых консервативные комплементарные участки, были подвергнуты процедуре продления (см. Материалы и методы). Также был сделан контроль процедуры продления. Для этого взяли выборку интронов, в которых не нашлись консервативные комплементарные участки, и далее действовали согласно схеме обычной процедуры продления (см. Материалы и методы). Оказалось, что субоптимальных вторичных структур находится значительно меньше, они короче и менее консервативны (данные не приводятся).

3.2. Визуализация вторичных структур

В результате процедуры продления для каждого интрона было получено много субоптимальных структур, отличающихся у разных организмов. Для удобства анализа таких структур был создан интерфейс визуализации вторичных структур. Программа написана на языке Java (код доступен по адресу http://kodomo.cmm.msu.ru/~Khrameeva/course/visual_code.rar). Она получает на вход выравнивания 52-х интронов и строит продленные вторичные структуры (рисунок 4). Интерфейс позволяет смотреть, как устроены вторичные структуры. На схеме слева представлена матрица комплементарности, которая является двумерным представлением всех возможных спиралей в данной молекуле РНК [19].

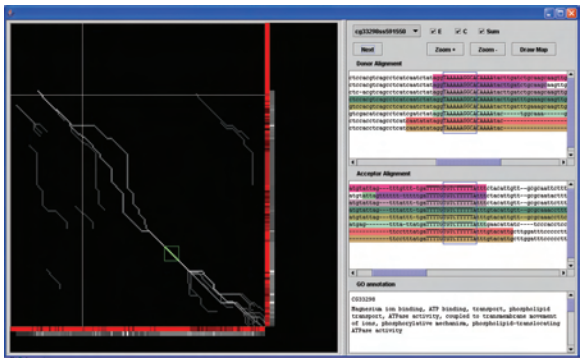


Рисунок 4. Интерфейс визуализации вторичных структур.

Белым цветом показаны все субоптимальные вторичные структуры (тем ярче, чем больше субоптимальных структур найдено на данном участке). По горизонтали – выравнивание донорной части интрона; по вертикали – выравнивание акцепторной части. По периметру схемы красным цветом показана энтропия колонки выравнивания (степень вырожденности позиции); белым – количество субоптимальных вторичных структур, найденных для данной колонки. В правой части интерфейс выводит выравнивания. В результате клика по схеме вторичных структур в правой части отображается соответствующий участок выравнивания (10 нуклеотидов выделены рамкой как на выравнивании, так и на схеме вторичных структур). На выравнивании каждая субоптимальная структура выделена своим цветом. Кроме того, выводится аннотация гена, которому принадлежит данный интрон.

При помощи данной программы было установлено, что для всех интронов с найденными консервативными комплементарными участками наблюдается схожая картина: исходный 9-нуклеотидный участок служит «затравкой», от которой расходятся в различной степени протяженные участки вторичной структуры (причем у разных организмов расходятся по-разному).

3.3. Анализ функций генов

Далее была поставлена задача выяснить, не являются ли найденные консервативные комплементарные участки особенностью только каких-то определенных групп генов. Для этого требовалось проверить, не перепредставлены ли какие-либо группы генов в выборке из 52 интронов с консервативными комплементарными участками.

Метод GOstat [20] позволяет решить эту задачу. С помощью GOstat было произведено сравнение двух выборок – нашей выборки из 46 генов, в интронах которых найдены

консервативные комплементарные участки, и выборки, состоящей из всех генов, аннотированных в базе данных Gene Ontology (GO). При этом перепредставленными назывались такие гены, процент которых в нашей выборке был значимо выше, чем в выборке из всех генов. Производилась коррекция на множественное тестирование.

По результатам GOstat, в нашей выборке перепредставлены гены дифференцировки (по онтологии биологических процессов) и гены ионных каналов (по онтологии молекулярных функций). Таким образом, наличие консервативных комплементарных участков характерно прежде всего для генов клеточной дифференцировки и ионных каналов.

3.4. Регуляция альтернативного сплайсинга

У дрозофилы на базе одного и того же гена *Dscam* может быть синтезировано более 38 тысяч разных сплайс-изоформ. При этом некоторые экзоны оказываются взаимно исключающими: один и только один экзон из данного набора входит в зрелую мРНК. Было показано, что взаимно исключающий характер сплайсинга обязан своим происхождением вторичной структуре РНК [21]. Перед всеми экзонами из взаимно исключающего набора располагается так называемый селекторный сайт, который комплементарен адаптерным участкам, расположенным около соответствующих экзонов. Таким образом, один и только один адаптер может спариваться с селекторным сайтом. Этот замечательный механизм регуляции альтернативного сплайсинга встречается не только у насекомых, но и у человека.

В связи с этим интересной представляется задача найти *in silico* альтернативные вторичные структуры в наборе из 52 интронов, для которых найдены консервативные комплементарные участки. Под интронами с альтернативными вторичными структурами понимались такие, в которых кроме уже найденных комплементарных участков (A и B) имеется третий участок (C), также комплементарный A (см. рисунок 5а). При этом требовалось, чтобы альтернатива C имела длину не менее 7 нуклеотидов и полностью перекрывалась с участком A по комплементарности (с одной разрешенной GT парой). Вместо консервативности было наложено требование, чтобы альтернативы присутствовали как минимум в 5 организмах.

Рассматривалось четыре варианта взаимного расположения комплементарных участков A и B и альтернативного участка C (рисунок 5а-г).

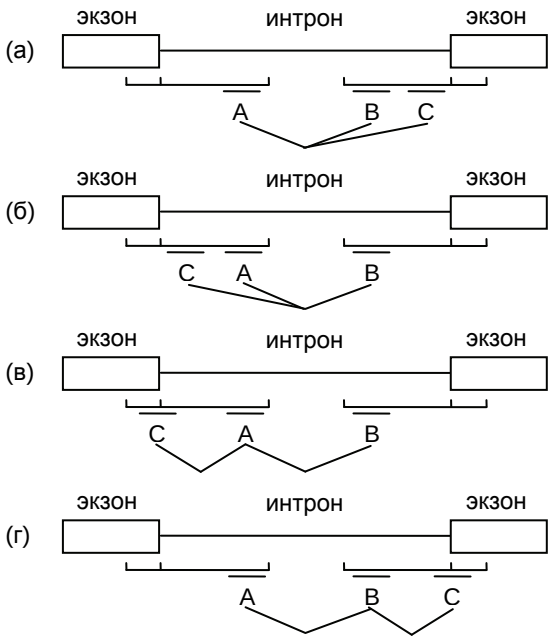


Рисунок 5. Варианты расположения альтернативы C (дуги указывают на комплементарность участков).

С перечисленными выше ограничениями в случаях (а)-(г) наличие альтернатив было показано для 12, 5, 5 и 13 интронов из 52, соответственно.

Был проведен такой же контроль, как для процедуры поиска консервативных комплементарных участков (см. Материалы и методы), где донорным частям сопоставлялись акцепторные части чужих интронов. Для каждого нового «интрона» находили по паре консервативных 9-ти нуклеотидных участков (конечно, не требуя комплементарности) и для них искали альтернативы по обычной схеме.

В контроле для четырех случаев, приведенных на рисунке 5а-г, получили 3, 3, 4 и 4 интрона, соответственно. Оцененные при помощи теста Фишера p-value для этих значений оказались значимыми на 5% уровне в случаях (а) и (г) (0.02 и 0.03, соответственно), что указывает на перепредставленность альтернативных структур вблизи акцепторных сайтов сплайсинга.

4. Заключение

В результате данной работы нами был разработан метод поиска консервативных вторичных структур РНК, ассоциированных со сплайсингом. Метод снабжен интерфейсом визуализации. С его помощью в 52 интронах найдены консервативные вторичные структуры. Показано, что найденные структуры являются статистически значимыми (процент ложных положительных предсказаний составляет 11±1%). Среди них перепредставлены гены клеточной

дифференцировки и ионных каналов. Также получены данные о перепредставленности альтернативных вторичных структур вблизи акцепторных сайтов сплайсинга. Тем не менее, вопрос об участии вторичных структур в регуляции альтернативного сплайсинга остается открытым и будет предметом наших дальнейших исследований.

5. Литература

[1] E. Buratti and F.E. Baralle, “Influence of RNA Secondary Structure on the Pre-mRNA Splicing Process”, *Molecular and Cellular Biology*, 2004, 24, pp. 10505-10514.

[2] D.D. Loeb, A.A. Mack, and R. Tian, “A secondary structure that contains the 5’ and 3’ splice sites suppresses splicing of duck hepatitis B virus pregenomic RNA”, *J. Virol.*, 2002, 76, pp. 10195-10202.

[3] S.H. Munroe, “Secondary structure of splice sites in adenovirus mRNA precursors”, *Nucleic Acids Res.*, 1984, 12, pp. 8437-8456.

[4] C.K. Damgaard, T.O. Tange, and J. Kjems, “hnRNP A1 controls HIV-1 mRNA splicing through cooperative binding to intron and exon splicing silencers in the context of a conserved secondary structure”, *RNA*, 2002, 8, pp. 1401-1415.

[5] J. Cabello-Villegas, K.E. Giles, A.M. Soto, P. Yu, A. Mougin, K.L. Beemon, and Y.X. Wang, “Solution structure of the pseudo-5’ splice site of a retroviral splicing suppressor”, *RNA*, 2004, 10, pp. 1388-1398.

[6] V. Goguel, Y. Wang, and M. Rosbash, “Short artificial hairpins sequester splicing signals and inhibit yeast pre-mRNA splicing”, Mol. Cell. Biol., 1993, 13, pp. 6841-6848.

[7] J.O. Deshler and J.J. Rossi, “Unexpected point mutations activate cryptic 3’ splice sites by perturbing a natural secondary structure within a yeast intron”, *Genes Dev.*, 1991, 5, pp. 1252-1263.

[8] H.X. Liu, G.J. Goodall, R. Kole, and W. Filipowicz, “Effects of secondary structure on pre-mRNA splicing: hairpins sequestering the 5’ but not the 3’ splice site inhibit intron processing in Nicotiana plumbaginifolia”, *EMBO J.*, 1995, 14, 377-388.

[9] Y. Chen and W. Stephan, “Compensatory evolution of a precursor messenger RNA secondary structure in the Drosophila melanogaster Adh gene”, *Proc. Natl. Acad. Sci. USA*, 2003, 100, pp. 11499-11504.

[10] T.P. Coleman and J.R. Roesser, “RNA secondary structure: an important cis-element in rat calcitonin/CGRP pre-messenger RNA splicing”, *Biochemistry*, 1998, 37, pp. 15941-15950.

[11] P.A. Estes, N.E. Cooke, and S.A. Liebhaber, “A native RNA secondary structure controls alternative splice-site selection and generates two human growth hormone isoforms”, *J. Biol. Chem.*, 1992, 267, pp. 14902-14908.

[12] A. Grover, H. Houlden, M. Baker, J. Adamson, J. Lewis, G. Prihar, S. Pickering-Brown, K. Duff, and M. Hutton, “5’ splice site mutations in tau associated with the inherited dementia FTDP-17 affect a stem-loop structure that regulates alternative splicing of exon 10”, *J. Biol. Chem.*, 1999, 274, pp. 15134-15143.

[13] E.E. Hennig, A.H. Conney, and S.J. Wei, “Characterization of hprt splicing mutations induced by the ultimate carcinogenic metabolite of benzo[a]pyrene in Chinese hamster V-79 cells”, *Cancer Res.*, 1995, 55, pp. 1550-1558.

[14] M. Blanchette and B. Chabot, “A highly stable duplex structure sequesters the 5’ splice site region of hnRNP A1 alternative exon 7B”, *RNA*, 1997, 3, pp. 405-419.

[15] N.N. Singh, R.N. Singh, and E.J. Androphy, “Modulating role of RNA structure in alternative splicing of a critical exon in the spinal muscular atrophy genes”, *Nucleic Acids Research*, 2006, 35(2), pp. 371-389.

[16] T.W. Hefferon, J.D. Groman, C.E. Yurk, and G.R. Cutting, “A variable dinucleotide repeat in the CFTR gene contributes to phenotype diversity by forming RNA secondary structures that alter splicing”, *Proc. Natl. Acad. Sci. USA*, 2004, 101, pp. 3504-3509.

[17] R.C. Edgar, “Local homology recognition and distance measures in linear time using compressed amino acid alphabets”, *Nucleic Acids Research*, 2004, 32(5), pp. 380-385.

[18] T.F. Smith and M.S. Waterman, “Identification of Common Molecular Subsequences”, *Journal of Molecular Biology*, 1981, 147, pp. 195-197.

[19] I.Jr. Tinoco, O.C. Uhlenbeck, and M.D. Levine, “Estimation of secondary structure in ribonucleic acids”, *Nature*, 1971, 230, pp. 363-367.

[20] T. Beissbarth and T.P. Speed, “GOstat: find statistically overrepresented Gene Ontologies within a group of genes”, *Bioinformatics*, 2004, 20(9), pp. 1464-1465.

[21] B.R. Graveley, “Mutually exclusive splicing of the insect Dscam pre-mRNA directed by competing intronic RNA secondary structures”, *Cell*, 2005, 123, pp. 65-73.

Эволюционный анализ генов семейства SSX

Елена Шустрова
Московский Государственный Университет им. М. В. Ломоносова, факультет Биоинженерии и Биоинформатики
lenashustrova@yandex.ru

Ирена Артамонова
Институт общей генетики РАН, Институт проблем передачи информации РАН
irenart@gmail.com

Аннотация

С/Т-антигены – это эндогенные белки организма, которые экспрессируются только в клетках зародышевого пути и плаценте, а также при развитии некоторых раковых опухолей. Семейство SSX является одним из представителей семейств таких антигенов. В данной работе мы исследовали эволюцию семейства SSX биоинформатическими методами. Локус семейства в геноме человека состоит из 11 генов, расположенных на X-хромосоме двумя отдельными регионами на расстоянии 4 Mb. Сравнение локусов семейства SSX в геномах человека, шимпанзе и мыши позволило выявить точные ортологические соотношения между отдельными генами, описать общую филогению семейства и восстановить эволюционную историю всего локуса в терминах элементарных геномных перестроек.

1. Введение

С/Т-антигены (Cancer/Testis-antigens) – группа антигенов, которые вызывают иммунный ответ в злокачественных опухолях, являясь при этом эндогенными белками организма. В норме соответствующие гены экспрессируются только в клетках зародышевого пути и плаценте. Одним из представителей семейств С/Т-антигенов является семейство SSX (synovial sarcoma X breakpoint) – семейство антигенов, экспрессирующихся при синовиальной саркоме. До сих пор неизвестна точная роль этих генов при развитии данной опухоли. Этим объясняется исследовательский интерес к данному семейству. Кроме того, продукты этих генов, как и другие С/Т-антигены, могут быть использованы в качестве маркеров противоопухолевых терапий.

Первым представителем семейства SSX стал ген SSX2. Он был определён как один их двух генов, участвующих в хромосомной перестройке t(X;18)(p11.2;q11.2), наблюдаемой в 70% случаев развития синовиальной саркомы [1]. Его партнёром на хромосоме 18 служит ген SYT, который кодирует активатор транскрипции. Вместе они образуют гибридный белок SYT-SSX2. Дальнейший анализ хромосомных транслокаций выявил ген SSX1, который вместе с тем же геном с 18 хромосомы образует гибридный белок SYT-SSX1 [2]. Остальные представители семейства – гены SSX3, SSX4, SSX5, SSX6, SSX7, SSX8 и SSX9 – были определены в результате анализа библиотек генов, а также с помощью поиска по базам данных [3, 4, 5, 6]. Как оказалось, ген SSX4 также вовлечён в хромосомную перестройку t(X;18)(p11.2;q11.2) [7]. Помимо этих 9 генов, были найдены 10 псевдогенов семейства. Все они, за исключением псевдогена ΨSSX10, расположены на хромосоме X [5].

Согласно экспрессионному анализу генов семейства SSX, гены SSX1, SSX2, SSX3, SSX4, SSX5 и SSX7 в норме экспрессируются только в тканях семенников. В тканях же почек, лёгких, кишечника, печени, матки, мозга, поджелудочной железы экспрессии обнаружено не было. Экспрессия в клеточных линиях меланомы была показана для генов SSX1, SSX2, SSX4, SSX5, SSX6 и SSX7. Экспрессия SSX8 и SSX9 не была замечена ни в одной из тканей или клеточных линий [4, 5, 8, 9].

Наиболее вероятная функция белковых продуктов генов семейства SSX – репрессия транскрипции [9], она связана с наличием в этих генах двух функциональных доменов – KRAB и SSXRD. Оба эти домена обладают активностью репрессора транскрипции [10, 11, 12, 13].

Мы провели эволюционный анализ семейства SSX, проанализировав экзон-интронную структуру генов человека и сравнив локусы семейства SSX в геномах человека, шимпанзе и

мыши. В результате, мы описали общую филогению семейства и восстановили эволюционную историю каждого гена в отдельности и всего хромосомного локуса в терминах элементарных геномных перестроек.

2. Методы и данные

Нуклеотидные и аминокислотные последовательности генов и их продуктов, последовательности соответствующих мРНК и EST (Expressed Sequence Tags) анализировались на основе сервиса UCSC Genome Browser, доступного на сайте http://genome.ucsc.edu [14]. Поиск по базе данных, доступных посредством этого сервиса, а также первичные выравнивание найденных последовательностей производились с помощью программы BLAT на том же сервере [15]. Для поиска далёких белков-гомологов использовалась поисковая программа BLAST (www.ncbi.nlm.nih.gov/BLAST/) [16]. Множественные выравнивания последовательностей строились с помощью программ CLUSTAL W [17] и T-COFFEE [18]. Построение филогенетических деревьев осуществлялось программой MEGA4 [19].

3. Результаты

3.1. Экзон-интронная структура генов семейства SSX в геноме человека

Семейство генов SSX в геноме человека состоит из 11 генов, расположенных на X-хромосоме. Все гены семейства имеют сходную экзон-интронную структуру, представленую на рисунке 1. В их структурах имеются лишь небольшие расхождения. Так, варьируют стартовые экзоны. Кроме того, для генов SSX2, SSX3, SSX4 и SSX5 существуют изоформы,

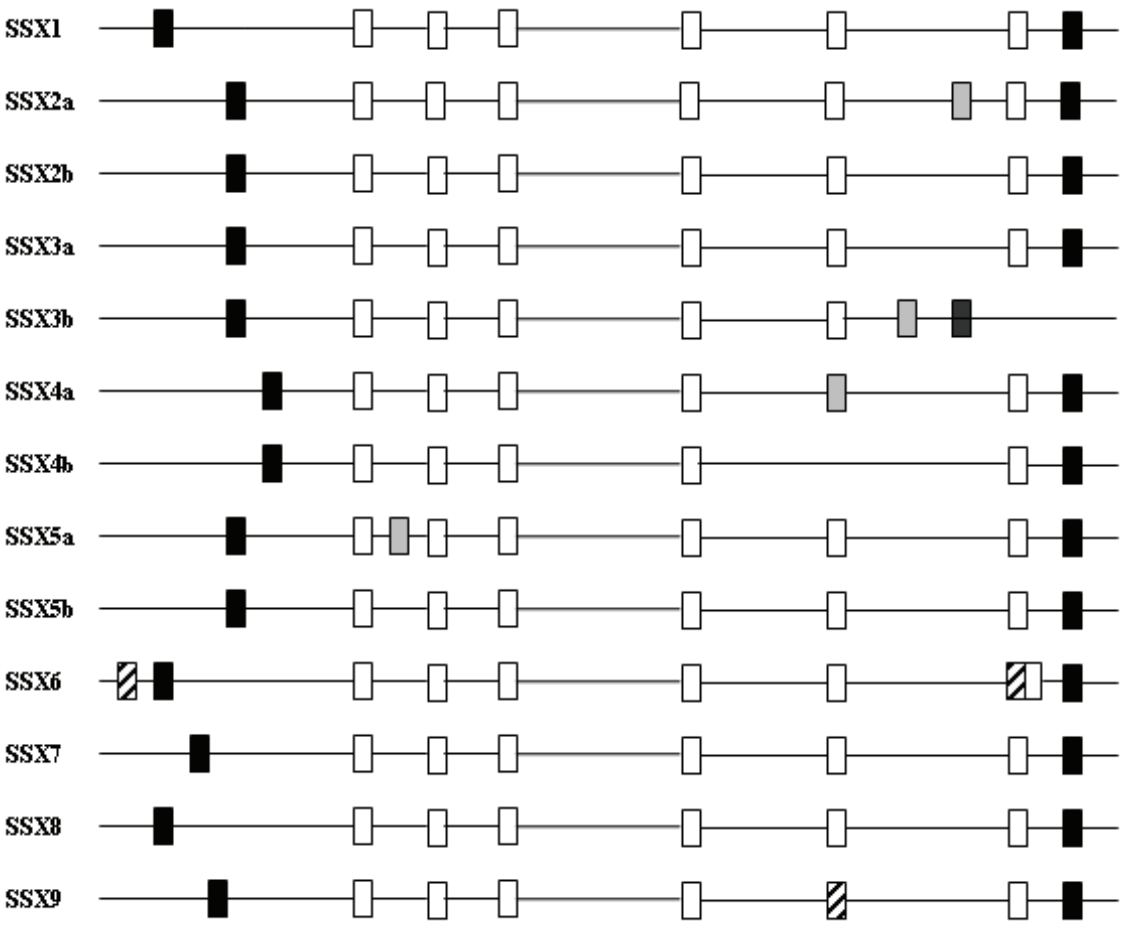


Рисунок 1. Схема расположения экзонов в генах семейства SSX. Экзоны, гомологичные друг другу, показаны белым цветом и расположены друг под другом. 5'- и 3'-экзоны показаны чёрным цветом, кассетные экзоны – серым. Заштрихованные экзоны содержат неканонические сайты сплайсинга.

имеющие в своём составе кассетные экзоны. Также в гене SSX3 имеется изоформа с альтернативным терминальным экзоном. Для гена SSX6 были выявлены неканонические донорные сайты сплайсинга для 1го и 7го экзонов. Анализ этого гена показал, что, видимо, функциональной является изоформа, использующая альтернативный стартовый экзон и альтернативный донорный сайт сплайсинга в 7м экзоне. Также встал вопрос о функциональной изоформе гена SSX9, поскольку его 6 экзон имеет неканонический донорный сайт сплайсинга. Вопрос остался неразрешённым, так как нет дополнительных данных, позволяющих сделать какой-либо вывод.

Кодирующая последовательность во всех генах начинается во 2м экзоне и заканчивается в предпоследнем.

3.2. Структура локуса в геномах человека и шимпанзе

Локус в геноме человека, содержащий гены семейства SSX, охватывает почти 5Mb X-хромосомы, он лежит на участке с координатами 47,852,032 – 52,807,342. Имеется две совсем недавние дупликации генов SSX2 и SSX4 – соответствующие копии генов практически неразличимы. В геноме шимпанзе семейство SSX состоит из 6 генов, расположенных на X-хромосоме на интервале с координатами 48,431,413 – 53,122,221. Порядок генов и их принадлежность к прямой или обратной цепи в обоих локусах проиллюстрированы на рисунке 2.

Экспансия семейства SSX происходила одновременно с разделением геномов шимпанзе и человека. Основные представители семейства уже присутствуют в геноме шимпанзе как отдельные гены, в то время как не удается идентифицировать копии гена SSX6, а также плохо различимы гены SSX7 и SSX9, SSX2 и SSX3. Совпадает принадлежность всех генов к прямой или обратной цепи, а также порядок, в котором они располагаются на хромосоме (за исключением генов SSX1 и SSX5 – они меняются местами в геноме шимпанзе). Помимо этого, в обоих локусах сохраняется интервал длиной 4Mb.

В геноме мыши семейство SSX также представлено несколькими генами. Филогенетическое дерево всех SSX генов приматов и мыши распадается строго на 2 таксон-специфичные клады (данные не представлены). Из этого мы сделали вывод, что данное семейство получило независимое распространение в отрядах грызунов и приматов. Видимо, у последнего общего предка приматов и грызунов семейство SSX было представлено только одним геном.

3.3. Эволюционная история локуса семейства SSX в терминах элементарных перестроек

Мы смоделировали эволюционную историю локуса сесейства SSX в терминах элементарных хромосомных перестроек, исходя из структуры локуса в геномах человека и шимпанзе, а также филогенетического дерева, построенного по нуклеотидным последовательностям генов человека, изображённого на рисунке 3. В этом анализе мы полагали, что параллельные события в эволюции значительно менее вероятны, чем хромосомные дупликации и потери. Кроме того, мы исходили из того, что тандемные дупликации более вероятны, чем дупликации в соседний локус. Более естественно восстанавливать события в направлении от настоящего к прошлому, сохраним этот стиль и в изложении.

Копии генов SSX2 и SSX4 настолько близки, что невозможно различить данные по транскрипции для этих генов. Очевидно, что эти копии - результат совсем недавних дупликаций. Различия между парами этих генов заключаются лишь в количестве повторов в поли-AC и поли-A треках, поэтому неясно, какой ген дуплицировался раньше, а какой – позже. Обе эти пары копий – результат тандемных инвертированных дупликаций. Эти копии, во-первых, слишком близки, а, во-вторых, отсутствуют в геноме шимпанзе. Поэтому, время происхождения этих дупликаций

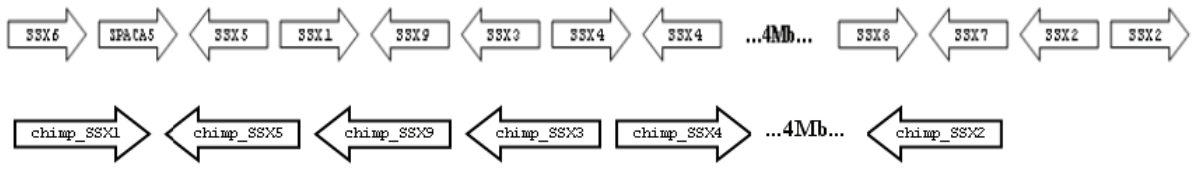


Рисунок 2. Карта локуса семейства генов SSX в геноме человека и в геноме шимпанзе.

заведомо меньше времени расхождения видов *Homo sapiens* и *Pan Troglodytes*, то есть 5-6 млн. лет.

Ближайшими попарными родственниками среди различных генов являются пары генов SSX1 и SSX8, SSX9 и SSX7, SSX3 и SSX2 (см. рисунок 3). Анализ их расположения в геноме и примерно одинаковое время, прошедшее после их расхождения приводит к выводу, что эти пары – результат дупликации целого сегмента, включающего 3 гена. Эта дупликация привела к встраиванию нового сегмента на расстоянии ~4 Mb от основного локуса семейства. Свидетельства этой дупликации есть и в геноме шимпанзе, причем относительное расстояние между локусами сохраняется. Вновь встроенный сегмент здесь представлен только геном SSX2, что может быть результатом последующей потери подсегмента, содержащего гены SSX8 и SSX7. Старая же копия сегмента представлена генами SSX3 и SSX9 – ген SSX1, видимо, также был потерян. Ген, аннотированный как SSX1 может являться ортологом гена SSX6. Кроме последующих потерь, отсутствие соответствующих генов в геноме шимпанзе может объясняться гэпами секвенирования, так как описываемый регион отличается многочисленными гэпами разного размера.

Данный сегмент также дуплицирован в геноме орангутана с сохранением относительного расстояния (данные не приведены), но свидетельства этой дупликации отсутствуют в геноме макаки. Таким образом, возраст данной дупликации ограничен интервалом от 13 до25 млн. лет.

До дупликации сегмента все события происходили внутри небольшого хромосомного локуса. Ген – предшественник генов SSX1, SSX6 и SSX8 дуплицировался в прямом направлении со встраиванием нового сегмента через 1-2 гена от старой копии, образовав современный ген SSX6 и ген-предшественник SSX1 и SSX8. Предыдущим событием, по-видимому, являлась прямая тандемная дупликация гена-предшественника SSX2, SSX3, SSX7 и SSX9, который, в свою очередь, появился в результате предыдущей прямой тандемной дупликации с образованием гена SSX5.

Предыдущая дупликация была инвертированной, разделив гены-предсшественники генов SSX1, SSX8 и SSX6 и генов SSX2, SSX3, SSX5, SSX7 и SSX9. Была ли эта дупликация тандемной, зависит от возраста гена SPACA5, разделяющего сегодня

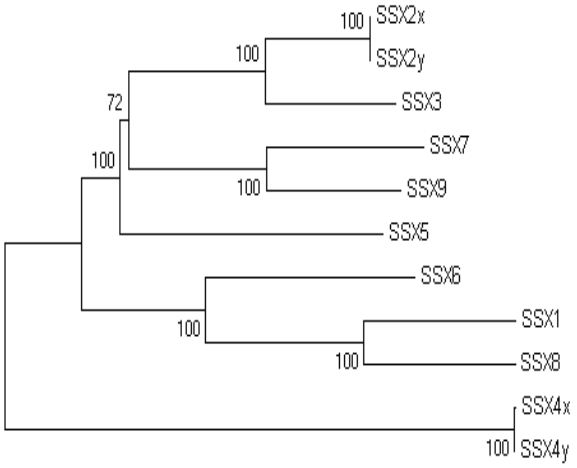


Рисунок 3. Дерево, построенное по нуклеотидным последовательностям генов SSX человека.

гены SSX6 и SSX5. Самым первым событием, положившим начало распространению SSX-семейства в геноме приматов, по современным свидетельствам, была тандемная дупликация общего гена-предшественника с отделением гена SSX4. Была ли эта дупликация прямой или обратной, сейчас невозможно установить, так как неизвестно направление гена-предшественника всех генов семейства, кроме SSX4. Так как мы полагаем, что у последнего общего предка приматов и грызунов семейство было представлено 1 геном, можно утверждать, что эта дупликация случилась менее 70 млн. лет назад. Общая схема описанных выше событий представлена на рисунке 4.

4. Заключение

Мы восстановили эволюционную историю хромосомного локуса семейства SSX на основе анализа представителей этого семейства в различных геномах современных приматов и мыши. Это позволило представить картину формирования мультигенного семейства из единственного гена в течение менее, чем 70 млн. лет. Исследования подобного рода помогут пролить свет на фундаментальные механизмы появления и фиксации геномных дупликаций в процессе эволюции.

Эта работа поддержана грантом по программе "Молекулярная и клеточная биология" Российской академии наук.

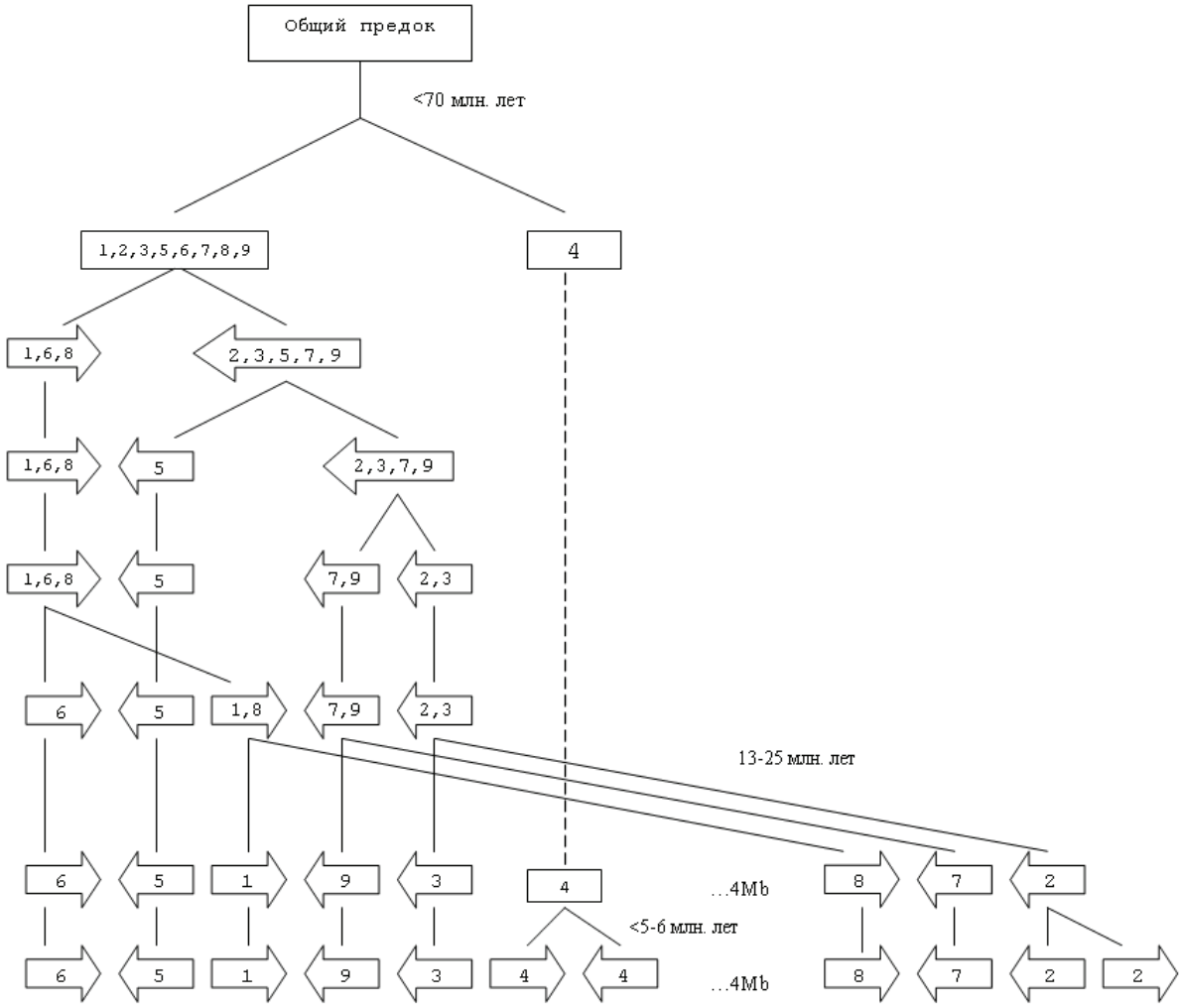


Рисунок 4. Схема эволюционных событий, происходивших в локусе семейства SSX.

5. Список литературы

[1] Clark J, Rocques PJ, Crew AJ, Gill S, Shipley J, Chan AM, Gusterson BA, and Cooper CS, “Identification of novel genes, SYT and SSX, involved in the t(X;18)(p11.2;q11.2) translocation found in human synovial sarcoma”, *Nat Genet*, 1994, 7, pp. 502-508.

[2] Crew AJ, Clark J, Fisher C, Gill S, Grimer R, Chand A, Shipley J, Gusterson BA, and Cooper CS, “Fusion of SYT to two genes, SSX1 and SSX2, encoding proteins with homology to the Kruppel-associated box in human synovial sarcoma”, *EMBO J*, 1995, 14, pp. 2333-2340.

[3] de Leeuw B, Balemans M, and Geurts van Kessel A, “A novel Krüppel-associated box containing the SSX gene (SSX3) on the human X chromosome is not implicated in t(X;18)-positive synovial sarcomas”, *Cytogenet Cell Gene*, 1996, 73, pp. 179-183.

[4] Gure AO, Türeci O, Sahin U, Tsang S, Scanlan MJ, Jäger E, Knuth A, Pfreundschuh M, Old LJ, and Chen YT, “SSX: a multigene family with several members transcribed in normal testis and human cancer”, *Int J Cancer*, 1997, 72, pp. 965-971.

[5] Güre AO, Wei IJ, Old LJ, and Chen YT, “The SSX gene family: characterization of 9 complete genes”, *Int J Cancer*, 2002, 101, pp. 448-453.

[6] dos Santos NR, de Bruijn DR, van Kessel AG, “Molecular mechanisms underlying human synovial sarcoma development”, *Genes Chromosomes Cancer*, 2001, 30, pp. 1-14.

[7] Skytting B, Nilsson G, Brodin B, Xie Y, Lundeberg J, Uhlén M, and Larsson O, “A novel fusion gene, SYT-SSX4, in synovial sarcoma”, *J Natl Cancer Inst*, 1999, 91, pp. 974-975.

[8] Türeci O, Chen YT, Sahin U, Güre AO, Zwick C, Villena C, Tsang S, Seitz G, Old LJ, and Pfreundschuh

M, “Expression of SSX genes in human tumors”, *Int J Cancer*, 1998, 77, pp. 19-23.

[9] Naka N, Araki N, Nakanishi H, Itoh K, Mano M, Ishiguro S, de Bruijn DR, Myoui A, Ueda T, and Yoshikawa H, “Expression of SSX genes in human osteosarcomas”, *Int J Cancer*, 2002, 98, pp. 640-642.

[10] Lim FL, Soulez M, Koczan D, Thiesen HJ, and Knight JC, “A KRAB-related domain and a novel transcription repression domain in proteins encoded by SSX genes that are disrupted in human sarcomas”, *Oncogene*, 1998, 17, pp. 2013-2018.

[11] Margolin JF, Friedman JR, Meyer WK, Vissing H, Thiesen HJ, and Rauscher FJ 3rd, “Krüppel-associated boxes are potent transcriptional repression domains”, *Proc Natl Acad Sci U S A*, 1994, 91, pp. 4509-4513.

[12] Pengue G, Calabrò V, Bartoli PC, Pagliuca A, and Lania L, “Repression of transcriptional activity at a distance by the evolutionarily conserved KRAB domain present in a subfamily of zinc finger proteins”, *Nucleic Acids Res*, 1994, 22, pp. 2908-2914.

[13] Witzgall R, O'Leary E, Leaf A, Onaldi D, and Bonventre JV, “The Krüppel-associated box-A (KRAB-A) domain of zinc finger proteins mediates transcriptional repression”, *Proc Natl Acad Sci U S A*, 1994, 91, pp. 4514-4518.

[14] Karolchik D, Baertsch R., Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, Weber RJ, Haussler D, and Kent WJ, “The UCSC Genome Browser Database”, *Nucl. Acids Res*, 2003, 31, pp. 51-54.

[15] Kent WJ, “BLAT - The BLAST-Like Alignment Tool”, *Genome Res*, 2002, 12, pp. 656-664.

[16] Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucleic Acids Res*, 1997, 25, pp. 3389-3402.

[17] Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, and Higgins DG, “Clustal W and Clustal X version 2.0.”, *Bioinformatics*, 2007, 23, pp. 2947-2948.

[18] Notredame C, Higgins DG, and Heringa J, “T-Coffee: A novel method for fast and accurate multiple sequence alignment”, *Journal of Molecular Biology*, 2000, 302, pp. 205-217.

[19] Tamura K, Dudley J, Nei M, and Kumar S, “MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0.”, *Molecular Biology and Evolution*, 2007, 24, pp. 1596-1599.

.

Регуляция дыхания бактерий: транскриптомика и сравнительная геномика

Марина Цыганова[1,*], Дмитрий Равчеев[2]
[1]*Московский государственный университет им. М.В. Ломоносова, факультет биоинженерии и биоинформатики, Москва, 119992*
[2] *Институт проблем передачи информации Российской академии наук, Москва, 127994*
**E-mail: marusya_tula@mail.ru*

Аннотация

В настоящее время для изучения экспрессии генов часто применяется метод биочипов, позволяющий одновременно получить и обработать информацию о работе большого числа генов. Несмотря на очевидные преимущества, этот метод имеет также и ряд существенных недостатков, связанных в основном со сложностью технологии эксперимента, высоким уровнем шума и ошибок эксперимента. Метод биочипов предоставляет данные о составе модулона – группы генов, экспрессия которых изменяется при мутации в гене белка-регулятора или при изменении условий среды.
В настоящей работе было произведено исследование регуляции дыхания энтеробактерий двумя факторами транскрипции – белками Fnr и ArcA. Полученные методом сравнительной геномики результаты были сопоставлены с данными биочипов для соответствующих регуляторов, благодаря чему выявлено соотношение регулонов и модулонов, а также показано расширение модулона по сравнению с регулоном за счет регуляторных каскадов

1. Введение

В настоящее время для изучения экспрессии генов часто применяется метод биочипов, позволяющий одновременно получить и обработать информацию о работе большого числа генов. Несмотря на очевидные преимущества, этот метод имеет также и ряд существенных недостатков, связанных в основном со сложностью технологии эксперимента, высоким уровнем шума и ошибок эксперимента. Метод биочипов предоставляет данные о составе модулона. Модулоном называется группа генов, экспрессия которых изменяется при мутации в гене белка-регулятора или при изменении условий среды. В случае регуляции в состав модулона входят гены, экспрессия которых прямо или опосредованно регулируется некоторой регуляторной системой.

Помимо экспериментальных технологий для изучения регуляции транскрипции в прокариотических организмах в настоящее время широко используются методы биоинформатических исследований. В случаях, касающихся регуляции транскрипции, методы биоинформатики состоят в поиске потенциальных регуляторных сайтов и выделении регулонов: групп генов, экспрессия которых непосредственно регулируется определенной регуляторной системой. Сравнительно-геномный метод, заключающийся в проверке консервативности потенциальных сайтов связывания регуляторов, значительно повышает качество предсказаний. Такой подход предполагает, что ген прямо регулируется изучаемым регулятором, если перед ортологичными ему генами в родственных геномах сохраняются потенциальные сайты [1].

Интуитивно понятно, что в состав модулонов входит больше генов, чем в состав соответствующих регулонов. Так как экспрессия большинства генов контролируется более чем одним регулятором, которые формируют сложные регуляторные каскады, то, по-видимому, одна из причин расширения модулона по отношению к регулону – влияние этих каскадов.

Для оценки вклада регуляторных каскадов в расширение модулона в данной работе были выбраны факторы транскрипции Fnr и ArcA контролирующие дыхание энтеробактерии *Escherichia coli* K12. В системе регуляции дыхания *E. coli* белок ArcA занимает подчиненное положение относительно Fnr: известно, что транскрипция гена *arcA*

регулируется последним [2], то есть, имеет место регуляторный каскад.

2. Материалы и методы

2.1. Исследованные геномы

Для выявления Fnr- и ArcA-регулонов с помощью сравнительно-геномных методов использовали следующие геномы бактерий порядка Enterobacteriales: *Escherichia coli* K12 *Salmonella typhi* Ty2, *Salmonella typhimurium* LT2, *Yersinia pestis* KIM, *Erwinia carotovorum* SCRI1043, *Yersinia pestis* KIM, *Yersinia enterocolitica* и *Photorhabdus luminescens* TTO1.Все последовательности были взяты из базы данных GenBank [3].

2.2. Определение состава регулона

В работе применяется метод выделения членов обобщенного регулона, основанный на попарном сравнении геномов организмов, относящихся к одной таксономической группе. Ген считали входящим в обобщенный регулон, если сайт связывания перед ним сохранялся более чем в трех родственных геномах. Однако если сайт был обнаружен только перед генами *E. coli* , *S. typhimurium* и *S. typhi*, то ген не включался в регулон, поскольку степень родства между данными организмами очень высока.

Поиск потенциальных сайтов связывания проводили с помощью метода матриц позиционных весов нуклеотидов [4]. Для этой цельи использовали описанные ранее матрицы для поиска сайтов связывания белков Fnr [5] и ArcA [6]. Поиск сайтов проводили в потенциальных регуляторных областях генов – в пределах -400 ... +100 п.н. относительно старта трансляции. Сайты считались значимыми, если их вес превышал пороговое значение: 3,50 для Fnr-сайтов и 4,25 для ArcA-сайтов. В случае, когда оперонная структура данного фрагмента последовательности неизвестна, полагали, что гены принадлежат к одному оперону, если они имеют одинаковое направление считывания, а раастояние между ними не превышает 100 п.н.

2.3. Использованные данные по регуляции экспрессии

Fnr- и ArcA-модулоны были выделены на основе экспериментальных данных метода биочипов из работ [7] и [8]. Для анализа потенциального регулона Fnr были использованы данные ChIP-chip исследования из работы [9].

2.4. Программное обеспечение

Для поиска ортологичных генов и потенциальных сайтов использовался пакет программ “Genome Explorer” [10].

3. Результаты и обсуждение

3.1. Выделение регулонов

В настоящей работе регулоны выделение регулонов производилось с помощью методов сравнительной геномики. Так, к потенциальному Fnr-регулону был отнесен 131 оперон, а к ArcA-регулону – 42 оперона.

3.2. Опероны функционального ядра

В ходе исследования нами была выделена группа из тринадцати оперонов, которые входят одновременно в Fnr-регулон и ArcA-регулон: *adhE*, *argR*, *focA-pflB*, *gapA*, *ndh*, *ompW*, *pdhR-aceEF-lpd*, *ychE*, *acnA*, *fadD*, *icd*, *mdh*, *gltA*. Эта группа была названа ядром регулонов. Опероны, составляющие эту группу, характеризуются рядом особенностей. Сайты перед этими оперонами имеют наибольшие из возможных веса. Кроме того, большое количество экспериментальных данных указывает на существование или возможность существования предсказанных нами сайтов.

Рассматриваемые опероны участвуют в ключевых этапах процесса дыхания (*adhE*, *focA-pflB*, *gapA*, *ndh*, *pdhR-aceEF-lpd*, *acnA*, *icd*, *mdh*, *gltA*) и других центральных процессах: синтез аминокислот (*argR*), бета-окисление жиров (*fadD*), мембранный транспорт (*ompW*).

3.3. Анализ потенциальных регулонов

Для анализа потенциальных Fnr и ArcA-регулонов, предсказанных методом сравнительной геномики, мы сравнили эти множества с множествами оперонов, регуляция которых была описана ранее в литературе. Кроме того, в случае Fnr для сравнения использовались данные об оперонах – мишенях FNR, идентифицированных с помощью ChIP-chip анализа [9].

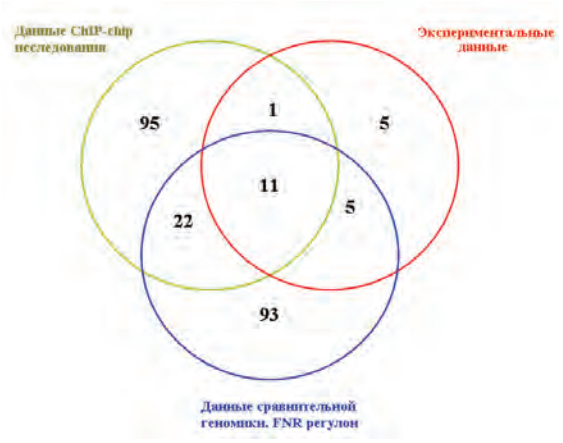


Рисунок 1. Диаграмма Венна для данных по FNR.

Как видно из диаграммы на Рис.1, Fnr-регулон практически полностью покрывает множестово оперонов, описанных как регулируемые Fnr в статьях, за исключением шести оперонов. Мы приняли множество оперонов, для которых регуляция Fnr была описана в литературе, за “золотой стандарт”. Исходя из этого, в случае пяти из них множество ChIp-chip их также не включает, следовательно, это ложноотрицательные результаты для обоих методов – ChIP-chip и сравнительной геномики. Можно объяснить причину того, что данные опероны не вошли в Fnr-регулон, как результат видоспецифичной регуляции.

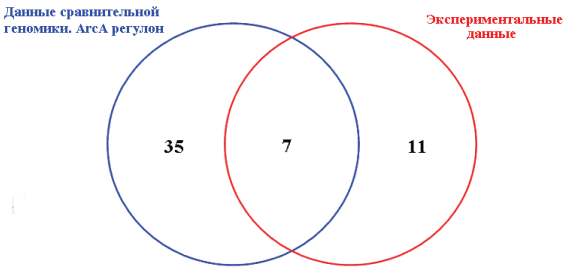


Рисунок 2. Диаграмма Венна для данных по ArcA.

Потенциальный ArcA-регулон включает только 7 из 18 оперонов, описанных как регулируемые ArcA в статьях (Рис.2). Из 11 оперонов, не вошедших в ArcA-регулон, три оперона, *lldPRD*, *aldA* и *glcDEFGBA*, характеризуются видоспецифичной для *E.coli* регуляцией, а восемь оперонов являются ложноотрицательным результатом.

3.4. Выделение предполагаемых Fnr- и ArcA-модулонов

Для выделения модулонов в данной работе были использованы данные биочипов для Fnr [7] и ArcA [8].

На настоящий момент времени не существует формальных критериев для выделения модулонов на основе эспресионных данных биочипов. В данной работе для выделения потенциальных модулонов рассматривались различные пороги по p-value. Все гены, изменяющие свою экспрессию при мутации регулятора с p-vaue ниже порогового значения, относятся к потенциальному модулону данного регулятора.

В качестве кандидатов рассматривались три множества экспрессионных данных, соответствующих трем пороговым значениям p-value, для каждого регулятора:

- гены, изменяющие свою экспрессию под влиянием фактора в 1.5 раза и более с p-value менее 0.05;
- гены, изменяющие свою экспрессию под влиянием фактора в 1.5 раза и более с p-value менее 0.01;
- гены, изменяющие свою экспрессию под влиянием фактора в 1.5 раза и более с p-value менее 0.001.

3.5. Выбор порога по p-value для Fnr-модулона и ArcA-модулона

Поскольку для выделения Fnr и ArcA-модулонов мы пользовались исходными данными, которые были получены в сходных условиях в результате аналогичных экспериментов, то при выборе порогового значения по p-value, мы можем ориентироваться на один их регуляторов, предполагая, что для другого пороговое значение будет тем же.

Для выбора порогового значения p-value описанные выше множества Fnr-модулона на трех уровнях значимости были пересечены с множествами, рассмотренными выше:

- опероны – мишени FNR, идентифицированные с помощью ChIP-chip анализа (множество “Данные ChIP-chip анализа”);
- опероны, регулируемые Fnr исходя из литературных данных (множество “Экспериментальные данные”);
- опероны – члены потенциального Fnr-регулона, предсказанного нами методом сравнительной геномики (множество “Данные сравнительной геномики. Fnr-регулон”)

Результат пересечения оформлен в виде диаграммы Венна (Рис. 3), где для каждого подмножества указан процент генов этого подмножества, относящихся к каждой группе значений p-value.

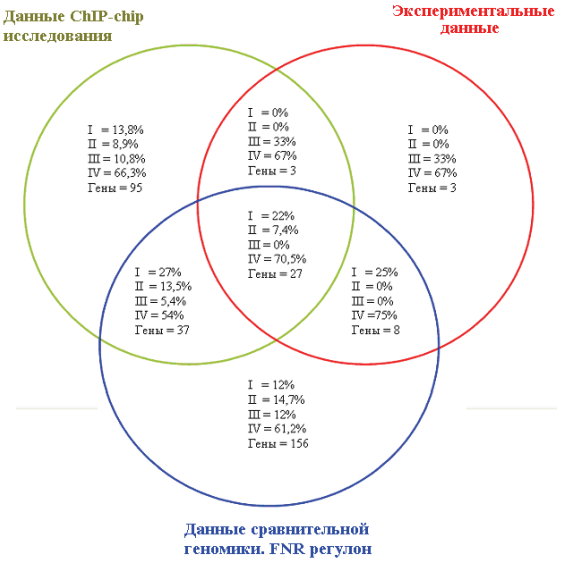


Рис.3 Диаграмма Венна для FNR с учетом результатов анализа данных биочипов. I - имеют уровень p-value ниже 0.001; II - имеют уровень p-value ниже 0.01, но выше 0.001; III -имеют уровень p-value ниже 0.05, но выше 0.01; IV - имеют уровень p-value выше 0.05.

Наибольший уровень достоверности демонстрировали множества оперонов, Fnr-зависимая регуляция которых подтверждена литературными данными или одновременно сравнительной геномикой и ChIP-chip исследованием. Оба эти двух множества оказались обогащены генами с p-value ниже 0.01.

Поскольку эти два подмножества являются частью множества оперонов, Fnr-регуляция которых имеет наибольшую степень достоверности, то за пороговое значение p-value следует принять значение, ниже которого лежит большинство значений p-value членов этих двух подмножеств, а именно 0.01.

При использовании данного значения p-value в качестве порогового к Fnr-модулону было отнесено 650 оперонов, а к ArcA-модулону – 722 оперона.

3.6. Соотношение сопряженных модулонов и регулонов

Сопряженными называются модулон и регулон одного регулятора, например, говорят о сопряженных модулонах и регулонах регулятора Fnr. Интуитивно понятно, что регулон должен являться частью сопряженного модулона.

Сопряженные Fnr-регулон и Fnr-модулон имеют 49 общих оперонов. Сопряженные ArcA-регулон и ArcA-модулон имеют 18 общих оперонов. Однако 82 оперона Fnr-регулона не вошли в Fnr-модулон, а 85 оперонов ArcA-регулона не вошли в ArcA-модулон. Наличие таких оперонов противоречит утверждению, что регулон является частью сопряженного модулона. Это может быть объяснено либо высоким уровнем перепредсказаний для Fnr-регулона, либо низким качеством экспрессионных данных.

3.7. Вклад регуляторных каскадов в расширение модулонов

Fnr и ArcA относятся к глобальным регуляторам дыхания в E.coli. Эти два регулятора формируют регуляторный каскад, в системе регуляции дыхания E. coli белок ArcA занимает подчиненное положение относительно Fnr: известно, что транскрипция гена *arcA* регулируется последним [2].

Для того чтобы проследить соотношения Fnr- и ArcA-модулонов, мы пересекли эти два множества. ArcA-модулон в значительной степени, но не входит целиком в состав Fnr-модулона. Эти два множества имеют 511 общих оперонов. В то же время, 211 оперонов ArcA-модулона не вошли в Fnr-модулон. Наличие оперонов ArcA-модулона, которые не входят в состав Fnr-модулона, скорее всего, это обусловлено высоким уровнем «шума» в результатах, полученных с помощью метода биочипов, или же связано с влиянием других регуляторов, экспрессия генов которых контролируется Fnr или ArcA. В целом же пересечение Fnr- и AcrA-модулонов подтверждает принцип иерархической организации регуляторных каскадов, так как значительная часть ArcA входит в состав Fnr-модулона. Из пересечения Fnr- и AcrA-модулонов следует также, что 139 оперонов Fnr-модулона не имеют общих членов с множеством ArcA-модулона. Эти опероны либо непосредственно регулируются Fnr, либо регулируются другими регуляторами второго уровня.

Для определения роли регуляторного каскада Fnr-ArcA в расширении Fnr-модулона по сравнению с Fnr-регулоном мы сопоставили полученные нами данные о составе регулона с данными, полученными с помощью метода биочипов и составили схему, визуализирующую соотношение Fnr-регулона, Fnr-модулона, и ArcA-модулона (Рис. 4). Как уже сказано, к Fnr-регулону нами было отнесено 131 оперон. Большая часть, а именно 601 оперонов Fnr-модулона, не находится под прямой Fnr-регуляцией. В тоже время, 40 из этих оперонов входят в ArcA-модулон. Эти соотношения иллюстрируют идею расширения модулона регулятора более высокого уровня за счет модуляции регулятором более низкого уровня.

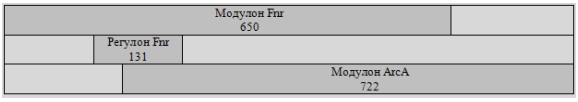


Рис. 4 Соотношение Fnr-регулона и Fnr- и ArcA-модулонов. Ниже указано количество оперонов, входящих в состав Fnr-модулона, Fnr-регулона и ArcA-модулона соответственно.

4. Благодарности

Данная работа выполнена совместно с М.С. Гельфандом. Мы благодарим А.А. Миронова и за любезно предоставленное программное обеспечение, а также Д.А. Родионова за ценные замечания.

Работа была частично поддержана Медицинским институтом Ховарда Хьюза, программой “Молекулярная и клеточная биология” Российской Академии Наук и INTAS.

5. Список литературы

[1] M.S. Gelfand, E.V. Koonin, A.A. Mironov, “Prediction of transcription regulatory sites in Archaea by a comparative genomic approach”, *Nucleic Acids Res*, Oxford University Press, England, 2000, pp. 695-705.

[2] I. Compan, D. Touati, “Anaerobic activation of arcA transcription in Escherichia coli: roles of Fnr and ArcA”, *Mol. Microbiol*., Blackwell Scientific Publications, England, 1994, pp. 955-964.

[3] D.A. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell and D.L. Wheeler, “GenBank”, *Nucleic Acids Res.*, Oxford University Press, England, 2007, pp. 21-25.

[4] М.С. Гельфанд, А.А. Миронов, “Компьютерный анализ регуляторных сигналов в полных бактериальных геномах” *Молекуляр. биология.*, “Наука”, Москва, Российская Федерация, 1999, стр. 772-778.

[5] А.В. Герасимова, Д.А. Родионов, А.А. Миронов, М.С. Гелфанд, “Компьютерный анализ регуляторных сигналов в бактериальных геномах. Участки связывания Fnr”, *Молекуляр. Биология*, “Наука”, Москва, Российская Федерация, 2001, стр. 1001-1009.

[6] A.V. Favorov, M.S. Gelfand, A.V. Gerasimova, D.A. Ravcheev, A.A. Mironov, V.J. Makeev, “A Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length”, *Bioinformatics*, Oxford University Press, England, 2005, pp. 2240-2245.

[7] K. Salmon, S. P. Hung, K. Mekjian, P. Baldi, and G. W. Hatfield, “Global gene expression profiling in *Escherichia coli* K12. The effects of oxygen availability and Fnr”, *J. Biol. Chem*, American Society for Biochemistry and Molecular Biology, United States, 2003, 278, pp. 29837-29855.

[8] K. A. Salmon, S. P. Hung, N. R. Steffan, R. Krupp, P. Baldi, G. W. Hatfield, and R. P. Gunsalus, “Global gene expression profiling in Escherichia coli K12: The effects of oxygen availability and ArcA”, *J. Biol. Chem*., American Society for Biochemistry and Molecular Biology, United States, 2005, pp. 15084-15096.

[9] D.C. Grainger, H. Aiba, D. Hurd, D.F. Browning, S.J. Busby, “Transcription factor distribution in Escherichia coli: studies with FNR protein”, *Nucleic Acids Res*., Oxford University Press, England, 2007, pp. 269-278.

[10] А.А. Миронов, Н.П. Винокурова, М.С. Гельфанд, “Програмное обеспечение анализа бактериальных геномов”, *Молекуляр. Биология*, “Наука”, Москва, Российская Федерация, 2000, 34, стр. 253-262.

Использование сравнительной геномики для улучшения достоверности аннотации сигнальных пептидов

Быкова Н.А.*, Сутормин Р.А.
Факультет Биоинженерии и Биоинформатики Московского Государственного Университета им. М.В. Ломоносова, 119992, Россия, Москва, Воробьевы горы 1-73
**nadelle4@mail.ru*

Аннотация

Экспорт белков из цитоплазмы определяется наличием N-концевого сигнального пептида. Данные о сигнальных пептидах широко используются в биотехнологии, а также могут быть использованы при разработке новых лекарств. Задача предсказания сигнальных пептидов пока не имеет удовлетворительного решения из-за отсутствия консервативных позиций в аминокислотной последовательности сигнального пептида и небольшого количества экспериментальных данных. В работе предложен алгоритм корректировки результатов программ предсказания сигнальных пептидов, основанный на гипотезе о консервативности наличия сигнального пептида внутри кластеров белковых ортологов. Предложенный алгоритм позволяет улучшить качество предсказаний как минимум в 1,3 раза.

1. Введение

Синтез белка в клетке, как известно, начинается в цитоплазме трансляцией мРНК на свободных рибосомах. Для того чтобы белки, выполняющие свою функцию в других компартментах клетки или за ее пределами достигли места локализации, они содержат в себе сигнальные последовательности. В этой работе рассматриваются N-концевые сигнальные пептиды бактерий, направляющие белки на экспорт из цитоплазмы. Белковые комплексы, осуществляющие перенос через плазматическую мембрану бактерий, называются транслоказами [1]. Компоненты транслоказ распознают сигнальный пептид и переносят белок на другую сторону мембраны, после чего сигнальный пептид отрезается с помощью сигнальных пептидаз.

Сигнальный пептид имеет длину 15-30 аминокислот и состоит из трех доменов: N, H и C. Основным для осуществления функции сигнального пептида является H-домен. Он представляет из себя тяж из 15-20 гидрофобных аминокислот, которые способны формировать гидрофобную альфа-спираль, вступающую во взаимодействие с компонентами транслоказ. C-домен сигнального пептида содержит в себе протеолитический сайт сигнальной пептидазы, что выражается в том, что в позициях -3 и -1 относительно места разрезания должны находиться аминокислоты с маленькой боковой цепью. N-домен положительно заряжен – по всей видимости, он взаимодействует с отрицательно заряженными фосфолипидами мембраны [2].

Задача предсказания сигнальных пептидов является важной, т к. помогает определить локализацию белка. Эти знания широко используются в биотехнологии, при разработке новых лекарств и выявлении механизмов генетических заболеваний [3]. При предсказании используются стандартные биоинформатические методы: применение позиционно-весовых матриц (программа PrediSi[4]), нейронные сети (SignalP3.0-NN[5]), скрытые марковские модели HMM (SigPfam[6]). Основными проблемами, с которыми приходится сталкиваться при предсказании сигнальных пептидов, являются: 1) отсутствие консервативности в позициях сигнального пептида (за исключением ограничения на размер боковой цепи в протеолитическом сайте) и вариация в длине 2) наличие гидрофобного H-домена обуславливает сходство сигнальных пептидов с сегментами трансмембранных белков 3) малое количество экспериментальных данных для обучения и оценки качества программ предсказания. В связи с обозначенными проблемами, программы предсказания имеют недостаточную точность. Так, при проверке результатов на независимых экспериментальных данных по анализу протеома одной бактериии, было показано, что точность программ не превышает 88% по чувствительности и 73% в специфичности [7].

Задачей данной работы является улучшение точности предсказания сигнальных пептидов. Для этого были использованы методы сравнительной геномики. Идея привлечения информации о гомологах была ранее реализована в программе PolyPhobius [8]. Согласно алгоритму, сначала осуществлялся поиск гомологов для каждой последовательности, потом строилось выравнивание для отобранных гомологов, и предсказание совершалось целиком для выравнивания. Наш подход позволяет избежать этапа выравнивания, которое с большой вероятностью может быть неправильным ввиду отсутствия консервативности в позициях сигнального пептида. Мы предлагаем алгоритм корректировки результатов программ предсказания, основанной на гипотезе о консервативности наличия сигнального пептида внутри кластеров ортологичных белков.

2. Материалы и методы

2.1. Гипотеза

В основе данной работы лежит гипотеза, которая заключается в том, что наличие сигнального пептида хорошо сохраняется в процессе эволюции, а значит консервативно внутри кластеров белковых ортологов. Это можно обосновать тем, что белки, являющиеся ортологами, как правило, имеют общую функцию и локализацию. Гипотеза подтверждается реальными данными – внутри белковых кластеров, используемых в работе (см. ниже), не наблюдается данных о наличии и одновременно отсутствии пептида, т.е. противоречий, несмотря на то, что существует достаточное количество кластеров с несколькими белками, в которых присутствие (или отсутствие) сигнального пептида проверено экспериментально.

Предполагаемые исключения могут наблюдаться при включении в кластеры паралогов, имеющих похожие аминокислотные последовательности.

2.2. Выборка экспериментальных данных

Выборка белков, имеющих экспериментальные данные о наличии или отсутствии сигнальных пептидов, содержит 8963 бактериальных белка из 197 геномов, распределенных по 385 кластерам ортологичных генов. Из них 5183 наблюдений (259 кластеров) описывают отсутствие сигнального пептида в белке (негативная выборка), остальные 3780 – присутствие. Выборка составлена путем объединения базы данных spDB [9] и обучающих выборок, используемых в работе программ: PrediSi, SignalP, SigPfam. Из них в тестовой выборке задействовано 8654 белка, т.к. для остальных не была определена принадлежность генома к грам-отрицательным или грам-положительным бактериям, которую необходимо знать для запуска программ предсказания. Выборка доступна по адресу http://www.bioinf.fbb.msu.ru/~nadya/sp_data.zip.

2.3. Алгоритм исправления ошибок.

Алгоритм исправления ошибок для одного метода предсказания, основанный на гипотезе о консервативности наличия сигнального пептида в ортологических кластерах, состоит в следующем:

1. Предсказания для белков группируются по кластерам. Если доля положительно предсказанных белков больше некоторого верхнего порога, то кластер целиком считается предсказанным положительно. Если доля положительно предсказанных белков меньше некоторого нижнего порога, то кластер считается предсказанным целиком отрицательно.
2. В таких “однозначных” кластерах отклонения считаются ошибочными и исправляются для каждого белка на значение предсказания для всего кластера.
3. В “неоднозначных” кластерах, где доля положительно предсказанных белков находится межу пороговыми значениями, исправления для белков не вносятся.

2.4. Алгоритм с голосованием

Метод голосования заключается в том, что решение о наличие сигнального пептида принимается, если две и более из трех программ предсказали его наличие. Такой способ можно рассматривать как отдельный способ предсказания. Применение алгоритма, описанного в предыдущем пункте, к этому способу ничем не отличается от обычного - метод дополняется нулевым пунктом проведения голосования среди трех методов.

2.5. База данных

Данные по геномам, белкам и таксономии были загружены с FTP сайта ресурса NCBI. БД содержит 1,872,850 белков из 566 не бактериальных геномов. 1,398,533 (538 геномов) из них оказались отнесенными к какому-либо кластеру (всего 237,419 кластера, см. ниже). Описываемый в работе алгоритм может быть использован только для этих белков. В связи с тем, что не для всех геномов удалось определить грам-положительность или грам-отрицательность (необходимый параметр программ предсказания сигнальных пептидов), предсказание проводилось для 1,290,702 белков из 472 геномов.

2.6. Программы предсказания

Для предсказания сигнальных пептидов использовалось 3 программы. Программа PrediSi использовалась со стандартным порогом – при выходном весе > 0.5 последовательность считается имеющей сигнальный пептид. Также на вход программе подавались сведения о грам-положительности и грам-отрицательности организма. Из программы SignalP бралась только часть, предсказанная с помощью нейронных сетей. В этом случае также использовались стандартные параметры и разные модели для грам-положительных и грам-отрицательных бактерий. Для программы SigPfam параметры были подобраны, исходя из того, чтобы при сравнении с экспериментальными данными количество перепредсказанных белков равнялось количеству недопредсказанных. Таким образом, пороговое значение для веса, выдаваемого SigPfam, было принято равным -6.

2.7. Ортологические кластеры

В работе использовалась база белковых кластеров NCBI [10], в которой содержится 237,419 бактериальных кластера, покрывающих 1,398,533 белков. Средний размер кластера - 6 белков, что значительно меньше, чем для кластеров базы данных COG. Такое более мелкое, чем в кластерах COG, разбиение представляется более подходящим для нашей задачи, т.к. с меньшей вероятностью содержит нежелательные паралоги. С другой стороны, в базе содержится 92,124 кластера размером 2 белка. В отношении таких кластеров могут возникнуть сомнения в применяемости предлагаемого алгоритма.

2.8. Подбор оптимальных порогов

Применение алгоритма требует наличия верхнего и нижнего порогов. В каждом случае мы рассчитали значения качества для каждой пары порогов, получив двумерный массив. Для каждого метода было выбрано оптимальное значение порогов, при котором доля ошибок минимальна.

3. Результаты

Вначале, используя три известных программы предсказания сигнальных пептидов (PrediSi, SignalP, SigPfam), мы предсказали наличие сигнальных пептидов для всех белков, имеющихся в геномной базе данных, созданной в рамках данной работы. Далее на выборке, содержащей экспериментальные данные о наличии сигнального пептида в белках, было проведено сравнение работы программ предсказания и предлагаемого алгоритма.

Показателем качества методов являлась доля ошибочных предсказаний, т.е. отношение количества ошибочных предсказаний к общему числу экспериментально проверенных наблюдений. Качество исходных методов составило 0,025, 0,011 и 0,071 соответственно для алгоритмов PrediSi, SignalP и SigPfam. После применения алгоритма исправления ошибок с оптимальными порогами, доля ошибок уменьшилась для каждого из методов, для метода SignalP она составила 0,009. Таким образом, после применения алгоритма, предсказания лучше соответствуют экспериментальным данным, что говорит в пользу гипотезы о консервативности наличия сигнальных пептидов внутри кластеров ортологичных белков.

Наилучшим оказалось применение алгоритма с использованием метода голосования. Доля ошибок составила в этом случае 0,0087, что в 1,3 раза меньше лучшего метода. Его мы и использовали для внесения исправлений в геномную базу. Всего было внесено 71501 исправлений, что составляет примерно 6% от размера базы данных. Из них 13264 отменяют сигнальный пептид, а 58238 – добавляют.

4. Обсуждение

К одному из недостатков проделанной работы можно отнести использующуюся экспериментальную выборку, которая содержит в себе тренировочные данные для используемых программ. Это привело к тому, что программы итак предсказывали довольно точно, и большую часть ошибок составляли артефактные, «не исправляемые» ошибки. Вероятно, применение такой выборки не дало возможность в полной мере продемонстрировать эффективность работы алгоритма и правильно оценить значения оптимальных порогов. С другой стороны, были выявлены случаи, при которых применение алгоритма может привести к усугублению ситуации. Тем не менее, можно сказать, что гипотеза, выдвинутая в работе, была подтверждена.

Рассмотрев различные случаи, мы подтвердили, что алгоритм позволяет делать то, что невозможно с помощью программ предсказания – находить потенциально новые виды сигнальных пептидов, т.к. он не основан на похожести по последовательности сигнального пептида. Полученные данные могут быть использованы для поиска и анализа возможных новых видов сигнальных пептидов. В отдельных случаях данный подход позволяет исправить аннотацию начала гена. К существенным недостаткам можно отнести то, что метод

применим только к белкам, относящимся к какому-либо кластеру.

5. Списиок литературы

[1] P. Natale, T. Brüser, A.J. Driessen, “Sec- and Tat-mediated protein secretion across the bacterial cytoplasmic membrane – Distinct translocases and mechanisms” Biochim Biophys Acta., 2007

[2] S.L. Rusch, D.A. Kendall, “Interactions that drive Sec-dependent bacterial protein transport”, Biochemistry, 2007, 46(34):9665-9673.

[3] K.C. Chou, “Prediction of protein signal sequences”, Protein Pept Sci., 2002, (6):615-622.

[4] K. Hiller et al., “PrediSi: prediction of signal peptides and their cleavage positions”, Nucleic Acids Res., 2004, 32(Web Server issue):W375-379.

[5] J.D. Bendtsen, H. Nielsen, G. von Heijne, S. Brunak, “Improved prediction of signal peptides: SignalP 3.0.”, J Mol Biol., 2004, 340(4):783-795.

[6] Z. Zhang, W.I. Wood, “A profile hidden Markov model for signal peptides generated by HMMER”, Bioinformatics, 2003, 19(2):307-308.

[7] N. Gupta et al., “Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation”, Genome Res., 2007, 17(9):1362-1377.

[8] L. Käll, A. Krogh, E.L. Sonnhammer, “An HMM posterior decoder for sequence feature prediction that includes homology information”, Bioinformatics, 2005, Suppl 1:i251-257.

[9] K.H. Choo, T.W. Tan, S. Ranganathan, “SPdb--a signal peptide database”, BMC Bioinformatics, 2005, 6:249.

[10] D.L. Wheeler et al., “Database resources of the National Center for Biotechnology Information”, Nucleic Acids Res., 2008, 36 (Database issue):D13-21.

Эволюционные изменения структуры новой некодирующей РНК человека

Артемий Бениаминов[12], Эрик Вестхоф[2], Ален Крол[2]
[1] *Институт молекулярной биологии, Вавилова 32, 119991 Москва*
[2]*Институт молекулярной и клеточной биологии, Рене Декарта 15, 67000 Страсбург*
abeniaminov@mail.ru

Многочисленные данные свидетельствуют об огромном разнообразии некодирующих РНК в эукариотах. Недавно обнаруженная новая некодирующая РНК HAR1 (Human Accelerated Region1) экспрессируется специфично в коре головного мозга человека. Участок генома, кодирующий РНК HAR1, высоко консервативен среди млекопитающих, эволюционировал наиболее быстро после расхождения общего предка человека и шимпанзе. Ферментативный и химический пробинги РНК HAR1 позволили нам установить вторичную структуру РНК, отличную от предложенной ранее. Кроме того мы обнаружили, что нуклеотидные различия обуславливают образование различных структур РНК HAR1 человека (клеверный лист) и шимпанзе (шпилька). Таким образом, эволюционные изменения проявились в значительной перестройке структуры РНК HAR1.

Исследование корреляций между аминокислотными последовательностями транскрипционных регуляторов и нуклеотидными последовательностями их сайтов связывания с ДНК

Ю. Д. Коростелев.
ФББ, МГУ
lan787@belozersky.msu.ru

О. Н. Лайкова
ГосНИИгенетика

А. Б. Рахманинова
ФББ, МГУ
abr@belozersky.msu.ru

Ранее в нашей группе был разработан метод поиска скоррелированных замен в последовательностях ДНК-связывающих белков и их сайтов связывания. При тестировании на семействе LACI-регуляторов были обнаружены статистически значимо скоррелированные пары позиций, хорошо соответствовавшие ДНК-белковым контактам в структуре комплекса PurR_Ecoli с его оператором purF (PDB: 1qpz) [1].

В настоящей работе исследована новая версия БД LACI (любезно предоставлена О. Лайковой), которая почти в 2 раза больше ранее изученной выборки. В программу введены процедуры взвешивания как регуляторов, так и сайтов. Полученные скоррелированные пары были сопоставлены с контактами в 3-х известных структурах (1qpz, 1efa, 1rzr) и с экспериментальными данными по точечному мутагенезу LACI_Ecoli и его оператора [2]. Исследованы перепредставленные пары слов (белковое слово − это а. о. в заданных позициях, нуклеотидное слово — нуклеотиды в заданных позициях).

Полученный топ-лист скоррелированных пар позиций включает практически все полярные контакты боковых групп а. о. с азотистыми основаниями, существующие хотя бы в одной из структур. Две позиции регуляторов, 16 и 20, являются абсолютными лидерами по скоррелированности. У этой пары а. о. лежат на одной стороне распознающей а-спирали. Полученные таблицы сопряженности частот а. о. в позициях 16, 20 и нуклеотидов в позициях 5, 7 идеально согласуются с экспериментальными данными [2]. (нумерация а.о по PurR_Ecoli, операторы – палиндромы, 20 п.н., нумерация с единицы).

Детальный анализ полученных предпочтительных пар позволил выявить несколько ярких взаимозависимостей: T^{16} уменьшает вероятность пары R^{20}-G_5, что, по-видимому, обусловлено стерическими взаимодействиями, а комбинация $S^{16}R^{20}$ предпочитает G_5C_6, что можно интерпретировать как возникновение «сложного» [3] контакта аргинина с 2-мя последовательными нуклеотидами. Существование этих и других взаимозависимостей побудило нас исследовать корреляции между парами слов (белковое слово - а. о. в позициях 16,20, нуклеотидное — нуклеотиды в наиболее скоррелированных позициях 5,6,7). Из полученных таблиц перепредставленных пар слов следует, например, что половина регуляторов с белковым словом $S^{16}R^{20}$ связывает сайты со словом $G_5C_6A_7$. И наоборот - сайт со словом $G_5C_6A_7$ связывает в 80% случаев белок с $S^{16}R^{20}$. А слово $G_5T_6A_7$ в 90% случаев встречается в паре со словом $A^{16}R^{20}$. Перепредставленные пары слов не являются филогенетическим следом, т.к. встречаются на разных ветвях дерева регуляторов.

Исследовано также семейство регуляторов NrtR. В предположительном ДНК-связывающем домене были обнаружены две сильно скоррелированых с ДНК позиции, расположенные в одной спирали через один виток, непосредственно взаимодействующие с основаниями ДНК. Что удивительно, помимо них, скоррелироваными с ДНК оказались 5 гидрофобных аминокислотных остатко, образующих гидрофобный кластер над распознающей спиралью. Возможно они задают

расположение спирали в большой бороздке ДНК. Корреляции также не являются следствием филогении: перепредставленные пары слов встречаются на разных ветвях дерева регуляторов, более того, был замечен случай скоррелированных замен в белке и операторе у 2-х очень близких регуляторов.

Таким образом предложенный метод позволяет не только предсказывать специфические контакты с ДНК, но и исследовать более тонкие взаимодействия между а. о. и нуклеотидами, влияющие на ДНК-белковое распознавание. Полученные таблицы сопряженности можно использовать для планирования генно-инженерных экспериментов.

Литература.

1) Коростелев Ю. Д. Лайкова О. Н. Рахманинова А. Б. Позиционно-Специфические Корреляции Между Последовательностями ДНК-Связывающих Доменов LACI-Семейства и их Операторов. *ИТИС 07*

2) Sartorius J, Lehming N, Kisters B, von Wilcken-Bergmann B, Müller-Hill B. lac repressor mutants with double or triple exchanges in the recognition helix bind specifically to lac operator variants with multiple exchanges. EMBO J. 1989 April; 8(4): 1265–1270.

3) Nicolas M. Luscombe, Roman A. Laskowski and Janet M. Thoronton Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. Nucleic Acids Research, 2001 Vol. 29, No 13 (2860-2874).

Deleterious effect of mitochondrial nucleotide repeats in *Chordata*

Popadin Konstantin[1], Mamirova Leila[1], Kuptsov Alex[2]

1. *Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute), Bolshoi Karetny pereulok 19, Moscow 127994, Russia.*
2. *Institute of Ecology and Evolution of the Russian Academy of Sciences, Leninskij prosp. 33, Moscow 119071, Russia.*

KonstantinPopadin@gmail.com

Abstract

*Repeats are hotspots for deletions in mitochondrial genome (1). The short mitochondrial genomes with deleted fragments possess higher replication rate and thus outnumber the normal ones during intracellular competition (2). Accumulation of such mitochondrial genomes with large deletions leads to a number of neuromuscular disorders (3,4). It seems likely, that repeats in mtDNA are **d**eleterious **i**n **l**ate **l**ive (DILL) mutations, which therefore might determine aging (5). In our work we test this suggestion. Firstly, we hypothesize that abundance of repeats in mitochondrial genome correlates with generation time across Chordata phylum. Secondly, we hypothesize that high temperature as well as basal metabolic rate determine increased mutability of mtDNA leading to higher deleterious effect of repeats. The theory of aging and influence of basal metabolic rate on mutability of mitochondrial DNA are discussed in light of the obtained results.*

1. Methods

We used 897 full mitochondrial genomes of vertebrate species, maximal lifespan is known for 282 analyzed genomes (database of Max Planck Society - Longevity Records: Life Spans of Mammals, Birds, Amphibians, Reptiles, and Fish). We estimated abundance of repeats in each genome by program (Orlov, unpublished). All statistical analyses were done in R language.

2. Results

Firstly, we analyzed effect of total number of repeats (direct, inverted, symmetrical and complemented) on maximal lifespan of chordate in the sample with known maximal lifespan (N = 282) (see figure 1). Warm-blooded sample (Eutheria, Metatheria, Monotremata & Aves, it is marked as squares in figure 1) demonstrates negative rank correlation with maximal lifespan (Kendall's rank correlation tau = -0.1647725, P = 0.002370), while cold-blooded sample (blue circles in figure 1) does not show any trend (Kendall's rank correlation tau = -0.05042379, P = 0.2358).

Secondly, we compared total number of repeats among warm-blooded (N = 308) and cold-blooded (N = 589) species in the initial sample (N = 897), irrespectively of their maximal lifespan. We observed that warm-blooded species contain fewer repeats in average as compared with cold-blooded ones: 907.6 versus 910.2 (P < 0.001, Mann-Whitney U test). Also total number of repeats in Eutheria (N = 202) is lower as compared to Metatheria (N = 29): 896 versus 1040.8 (P = 0.002, Mann-Whitney U test).

3. Discussion

Our results agree with our expectation that mtDNA of warm-blooded species are more constrained in terms of number of repeats than cold-blooded species. Eutheria and Metatheria demonstrate the same trend, bearing in mind that Metatheria species possess lower basal metabolic rate and body temperature than Eutheria. We suggest that high basal metabolic rate accelerates mutability of nucleotide repeats, either through increased total mtDNA mutagenic score (free oxygen radicals) or

increased rate of mtDNA replication. Future analyses are needed to decisively corroborate our hypothesis.

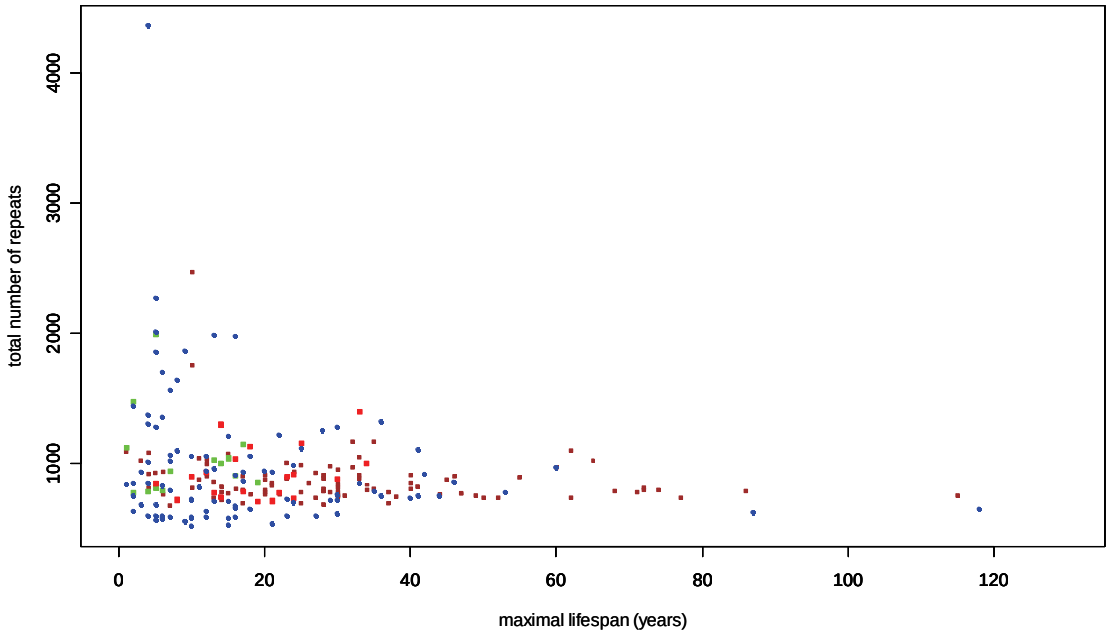


Figure 1. The relationship between total number of repeats in full mitochondrial genome and maximal lifespan of warm-blooded (Eutheria – brown squares, Metathera & Monotremata – green squares, Aves – red squares) and cold-blooded species (blue circles).

4. Acknowledgements

This research was supported by the Molecular and Cellular Biology Program of the Russian Academy of Sciences, by Russian Foundation for Basic Research Grants 08-04-09680, and by the Russian Science Support Foundation Grants to K.P.

5. References

[1] Samuels D.C. 2004. Mitochondrial DNA repeats constrain the life span of mammals. *Trends in Genetics* 20: 226-229.

[2] Taylor D.R., Zeyl C. and Cooke E. 2002. Conflicting levels of selection in the accumulation of mitochondrial defects in Saccharomyces cerevisiae. *PNAS* 99: 3690–3694.

[3] Kraytsberg *et al.* 2006. Mitochondrial DNA deletions are abundant and cause functional impairment in aged human substantia nigra neurons. *Nature genetics* 38: 518-520.

[4] Bender *et al.* 2006. High levels of mitochondrial DNA deletions in substantia nigra neurons in aging and Parkinson disease. *Nature genetics* 38: 515-517.

[5] Cortopassi G.A. 2002. A neutral theory predicts multigenic aging and increased concentrations of deleterious mutations on the mitochondrial and y chromosomes. *Free radical biology & medicine* 33: 605–610.

Реконструкция эволюции систем регуляции генов метаболизма гистидина в бактериях.

Алексей Витрещак
Институт Проблем Передачи Информации им. А.А.Харкевича, РАН (Большой каретный пер. 19)
**l_veter@mail.ru*

Аннотация

Бактерии используют различные механизмы для регуляции экспрессии генов. Т-бокс антитерминация является основным механизмом регуляции генов, вовлеченных в метаболизм аминокислот, в Грам-положительных бактериях. Однако, в некоторых грам-положительных бактериях наблюдаются лишь единичные случаи регуляции аминокислотных генов аттенуацией транскрипции или Т-бокс сайтами. Мы предположили, что подобные единичные случаи регуляции эволюционно возникли в тех геномах, где предположительно пропал некоторый неизвестный транскрипционный фактор. Используя методы сравнительной геномики, мы предсказали новый регуляторный сайт и транскрипцирнный фактор, связывающийся с ним, вовлеченные в регуляцию генов биосинтеза, транспорта и катаболизма гистидина в грам-положительных бактериях. Предсказан регулон для нового регуляторного сигнала. Построена эволюционная модель регуляции генов гистидина тремя различными регуляторными механизмами в Грам-положительных бактериях.

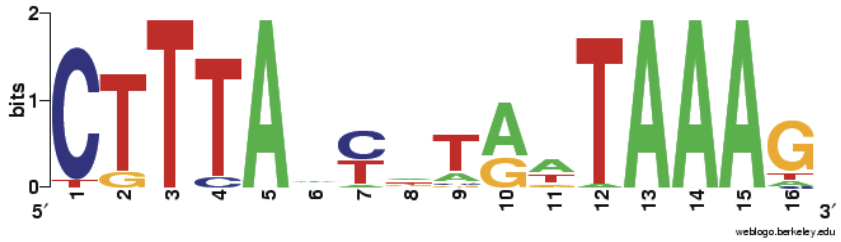
Введение.

Бактерии используют различные механизмы для регуляции экспрессии генов. В то время как наиболее общим регуляторным механизмом является регуляция транскрипции ДНК-связывающимися белками, имеются другие важные регуляторные механизмы, в частности, регуляция, осуществляемая альтернативными вторичными структурами мРНК в лидерной области генов: преждевременная терминация транскрипции и ингибирование инициации трансляции [1]. Т-бокс антитерминация является основным механизмом регуляции генов, вовлеченных в метаболизм аминокислот, в Грам-положительных бактериях. Однако, в некоторых грам-положительных бактериях наблюдаются лишь единичные случаи регуляции аминокислотных генов аттенуацией транскрипции или Т-бокс сайтами. Например, гистидиновый биосинтетический оперон *his* регулируется Т-боксами и аттенуацией транскрипции только в некоторых лактобациллах/стрептококках и в нескольких фирмикутах (*B. cereus, C. bartlettii, C. ramosum, C. difficile*), соответственно [3]. Мы предположили, что подобные единичные случаи регуляции эволюционно возникли в тех геномах, где предположительно пропал некоторый неизвестный транскрипционный фактор. Например, в *C. difficile,* исчез регулятор AhrC, что привело к недавней экспансии Т-боксов (расширению Т-бокс регулона): гены биосинтеза и транспорта аргинина в этой бактерии стали регулироваться аргинин-специфичными Т-боксами. Мы предположили подобный эволюционный эффект, названный «регуляторным взрывом» [3] и в случае с гистидиновыми генами. Кроме того, ранее были найдены единичные случаи регуляции гистидиновых генов аттенуатораим транскрипции в Грам-положительных бактериях [4]. Используя методы сравнительной геномики, мы предсказали новый регулятор YerC и его регуляторный сайт, вовлеченные в регуляцию генов биосинтеза, транспорта и катаболизма гистидина в грам-положительных бактериях. Построена эволюционная модель регуляции гистидиновых генов в Грам-положительных бактериях, обобщающая три вида регуляции: регуляцию Т-боксами, аттенуацию транскрипции и регуляцию новым транскрпционным фактором YerC.

Результаты.

Используя методы сравнительной геномики, мы предсказали новый регуляторный сайт, вовлеченный в регуляцию генов биосинтеза, транспорта и катаболизма гистидина в грам-

положительных бактериях. Это сайт представляет собой четный палиндром длины 16 (Рис 1а). В Bacillales левое плечо ослаблено, хотя нуклеотиды в ослабленных позициях достаточно консервативны. Это может быть объяснено эволюционным выбором более слабого сайта для регуляции экспрессии гистидинового оперона в Bacillales (Рис 1б).



а) Clostridiales, Thermoanaerobacteriales, Halanaerobiales, Bacillales.

б) Bacillales (*his* оперон).

Рис 1. **YerC регуляторный сайт .**

Методами сравнительной геномики и позиционного анализа найден белок регулятор YerC (из семейства TprR), в димерной форме связывающийся с найденным полиндромным сайтом. В некоторых Clostridiales YerC авторегулируется и позиционно связан с генами биосинтеза гистидина. Кроме того, как показал геномных анализ, YerC отсутствует в тех геномах, где зафиксированы единичные случаи регуляции Т-боксами или аттенуаторами транскрипции. Таким образом, наше предположение подтвердилось - в процессе эволюции YerC в этих геномах был потерян, что привело к экспансии Т-бокс сайтов и «рождению»/горизонтальному переносу новых аттенуаторов транскрипции.

YerC регулон. В грам положиетльных YerC регулирует гистидиновый биосинтетический оперон. Кроме того, кроме предсказанного гистидинового ABC транспортеар *hisXYZ*[3] YerC регулирует два новых транспортера, *yuiF* (Bacillales) и Cphy_3090 (Clostridiales, Thermoanaerobacteriales, Halanaerobiales) из семейства транспрортеров NahC и SSS, соответственно. Кроме того в некоторых Bacillales YerC регулирует катаболичекий оперон *hut*. В этом случае возможно сложное взаимодействие известного механизма hut антитерминации и посадки транскрипционного фактора YerC на предсказанный сайт.

Эволюционная модель регуляции гистидиновых генов. В общем предке фирмикут (корень дерева, рис 2) гистидинил-тРНК синтетаза и гены биосинтеза и транспорта гистидина регулируются Т-боксом и транскрипционным фактором YerC соответственно. Далее YerC пропадает в предке лактобацилл (рис 2) и происходит экспансия Т-боксов: гистидиновый биосинтетический оперон регулируется гистидин-специфичными Т-боксами. Далее недавняя потеря транскрипцирнного фактора YerC в трех клостридиях *C. difficile, C. bartletti, C. ramosum* (рис. 2, крайняя левая ветка) привела к появлению аттенуаторов транскрипции перед гистидиновым биосинтетическим опероном и Т-бокса перед гистидиновым ABC транспортером

(*C. difficile*). В остальных Clostridia гены биосинтеза и транспорта гистидина регулируются YerC. Аналогично независимая недавняя потеря транскрипцирнного фактора YerC в двух бациллах *(B. cereus, B. weihenstephanensis)* привела к зарождению или горизонтальному переносу гистидиновых аттенуаторов транскрипции. Как и в предыдущем случае в остальных бациллах гистидиновые гены биосинтеза и транспорта по-прежнему регулируются транскрипционным фактором YerC. Таким образом, в построенной эволюционной модели есть три независимых факта потери транскрипционного фактора сильно разнесенные по времени с последующей, по всей видимости, очень быстрой заменой механизма регуляции, от дубликации регуляторного сайта (в случае Т-бокса) до горизонтальнгого переноса или зарождения регуляторного сайта (в случае аттенуации транскрипции).

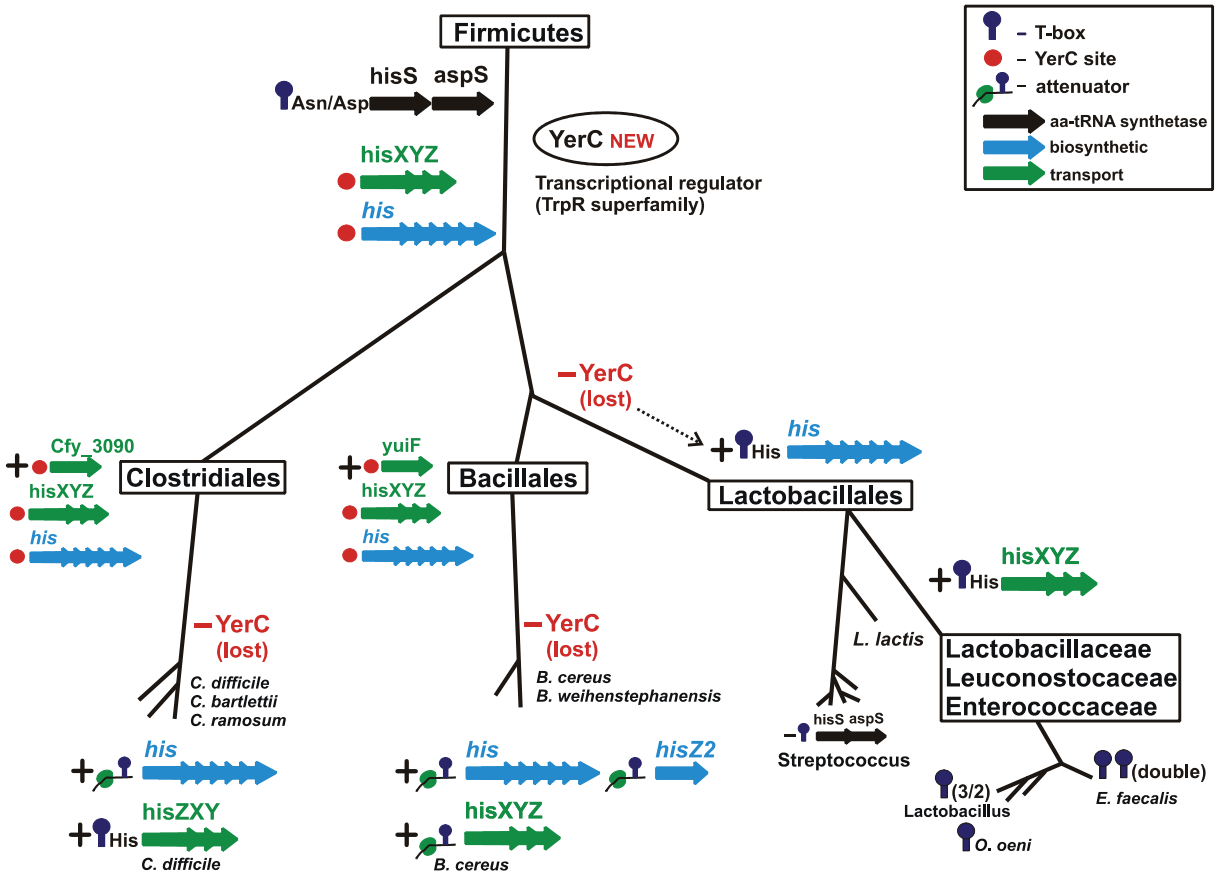


Рис 2. Эволюция регуляции гистидиновых генов в Грам-положительных бактериях.

Это совместная работа с М.С. Гельфандом.

Список литературы.

[1] M.S. Gelfand. “acterial cis-Regulatory RNA Structures” *Molecular Biology*, 2006, 40(4): pp. 541-550.

[2] F.J. Grundy, T.M. Henkin. “The T box and S box transcription termination control systems”. *Front Biosci.*, 2003, Jan 1, 8: pp. 20-31. Review.

[3] A.G. Vitreschak, A.A Mironov, V.A. Lyubetsky, M.S. Gelfand. **“**Comparative genomic analysis of T-box regulatory systems in bacteria”. *RNA*, 2008, Apr 14, 4:pp.717-735.

[4] A.G. Vitreschak, E.V. Lyubetskaya, M. Shirshin, M.S. Gelfand, V.A Lyubetsky. “Attenuation regulation of amino acid biosynthetic operons in proteobacteria: comparative genomics analysis”. *FEMS Microbiology letters*, 2004, May 15, 234(2): pp. 357-70.

Предсказание сайтов регуляции рибосомальных генов в бактериях.

Светлана Петрова[1], Алексей Витрещак[2], Михаил Гельфанд[1,2]*
[1]МГУ им. М. Ломоносова ,ФББ(Москва, Ленинские горы, стр. 73)
[2]Институт Проблем Передачи Информации им. А.А.Харкевича, РАН (Большой каретный пер. 19)
**plazmoliz[at]yandex[dot]ru*

Аннотация

Один из важных механизмов регуляции генов связан с образованием вторичной структуры мРНК в регуляторной (или лидерной) области гена. Такой механизм регуляции имеется как у бактерий, так и у эукариот (Nomura, 1999). Рибосомальный белок связывается с регуляторным сайтом мРНК, имитирующим сайт связывания рибосомального белка с рибосомой. При связывании запирается сайт инициации трансляции и рибосомальные гены не экспрессируются. Мы предсказали оперонные структуры и регуляторные сигналы для 6 оперонов: rpsO (S15), alpha ,spc (S8), S10 (L4), L11 (L1),Beta (L10) для 16 гамма-протобактерий.. Также изучили эволюцию сигналов, регулирующих рибосомальные гены на далеких эволюционных расстояниях. В частности, для оперона L11 (rplK-rplA) мы нашли подобную консервативную структуру у альфа-, бета-, гамма-протобактерий, грам-положительных бактерий м некоторых других. Причем в фирмикутах сайты найдены либо перед обоими генами оперона, либо перед одним геном rplK или rplA. Данный регуляторный сайт не был найден у цианобактерий, актиномицетов, дельта- и эплисон-протобактерий, Bacteroidetes. Вполне возможно, что в результате структурных перестроек оперона этот сигнал локализуется в другом месте.

Введение.

Экспериментально для некоторых рибосомных оперонов *E.coli* было показано значительное сходство между регуляторной структурой в лидерной области оперона и структурой рРНК.
Целью данной работы является выявление структур, схожих с теми, что регулируют трансляцию оперонов *E.coli*, у других бактерий,а так же выявление изменений в сайтах регуляции у бактерий, по мере роста эволюционного расстояния между ними и E.coli.

Мы предсказали оперонные структуры и регуляторные сигналы для 6 оперонов: rpsO (S15), alpha ,spc (S8), S10 (L4), L11 (L1),*Beta (L10)* для 16 гамма-протобактерий при помощи прогамм GenomeExplorer (Mironov, 2000), Operon (Витрещак, unpublished), ClustalW (Thompson, 1997), mfold(Zuker, 2003), и Multal(Mironov, unpublished). Поиск такой же регулирующей структуры, как у L11 проведен для бактерий с большим эволюционным расстоянием. В результате чего она была найдена у альфа-протобактерий, бета- протобактерий, гамма протобактерий, грамположительных , клостридий и микоплазм**.** И не обнаружена у цианобактерий, актиномицетов,дельта-протобактерий, Bacteroidetes, эплисон-протобактерий, а также архей.

Результаты.

Нами были рассмотрены лидерные области 6 оперонов: *rpsO (S15), alpha ,spc (S8), S10 (L4), L11 (L1),Beta (L10).* Для 16 других гамма протобактерий. При помощи программ GenomeExplorer (Mironov, 2000), Operon (Витрещак, unpublished)
были найдены ортологи для всех генов оперона. Далее были взяты участки последовательности, включающие в себя по 250 нуклеотид от гена перед которым лежит сигнал в опероне, а также 30 нуклеотид самого гена. Эти участки были выровнены программой Clustal(Thompson, 1997) (при этом, для получения лучшего результата часть последовательностей отсеивалась, подробнее в таблице 1 ниже), а так же программой Multal(Mironov, unpublished), которя так же давала информацию о расположении шпилек в исследуемой зоне. Для уточнения всех пространственных структур использовался алгоритм программы Mfold. В результате сравнения этих участков были получены возможные регуляторные структуры для некоторых бактерий, которые позволили выявить

наиболее консервативные участки в структуре регуляторов. Ниже описаны результаты для каждой из исследуемых структур.

κpsO (S15)
Для E.coli показано образование трех шпилек в лидерной области оперона, пред геном *rpsO*. Сравнительным анализом было показано, что первая (считая от 5' конца) шпилька является вариабельной, а 2 другие имеют в своем составе консервативные участки.
Для данной лидерной области возможно существование альтернативной вторичной структуры - псевдоузла. Соответственно белок репрессор S15 будет связываться либо с парой шпилек ,либо с псевдоузлом. При этом в случае двух шпилек область Шайна-Дальгарно входит в черенок третьей шпильки, а старт-кодон (ATG) в петлю этой же шпильки. В случае псевдоузла Шайна-Дальгарно входит в состав петли псевдоузла, а старт-кодон располагается на границе шпильки и петли псевдоузла.

alpha:
Регуляторный сайт Alpha оперона E.coli имеет структуру псевдоузла, вторичная структура которого представляют собой шпильку имеющую не только верхнюю петлю, но и боковые выпячивания. При этом 5 нуклеотидов петли псевдоузла взаимодействуют с 5 нуклеотидами, располагающимися на верху шпильки. И для всех выбранных последовательностей именно эти 5 нуклеотидов являются абсолютно консервативными. Шайна-Дальгарно и старт-кодон входят в петлю псевдоузла. Сама шпилька так же довольно консервативна, хотя в случае вибрионов и альтеромонад положение ее черешков было не вполне ясно. Но в результате множественного выравнивания вибрионов и альтеромонад, искомый псевдоузел был обнаружен, с той лишь разницей, что в петлю на вершине шпильки входит более, чем 5 нуклеотид.

Spc(S8):
Было известно, что регуляция данного оперона у E.coli может происходить путем образования в лидерной области одной из двух альтернативных структур, каждая из которых представляет по своему вторичному строению шпильку, но разной длины.
Однако сравнительным анализом подтверждается существование только ,более короткой, шпильки. При этом Шайна-Дальгарно и старт кодон входят в черешок шпильки. Следует также отметить особенность данной структуры, заключающуюся в том, что она заходит на кодирующую область 3-го гена (rplE) в опероне.

S10(L4):
Для *E. coli* показано образование 5 шпилек в лидерной облати этого оперона. Путем сравнительного анализа установлено, что только 3 из них достаточно консервативны. При этом Шайна-Дальгарно входит в черешок шпильки а, старт кодон (гена rpsJ) лежит в петле этой же шпильки ,т.е структура заходит на кодирующую область гена.

L11(L1):
Регуляторный участок данного оперона состоит из двух шпилек, одна из которых имеет 2 боковые петли. При этом, консервативными участками, найденными у всех рассмотренных бактерий являются 1 черешок и часть бокового выпячивания, следующая за ним в шпильке. Именно этот участок является схожим с участком на рРНК. Учитывая небольшой участок консервативности однозначно говорить входят ли область Шайна-Дальгарно и старт кодон гена rplK в состав шпильки не приходиться.

beta (L10):
Известно, что в регуляции данного оперона E.coli участвуют 5 шпилек. Для рассмотренных бактерий в их лидерных областях так же удалось построить все 5 шпилек, хотя консервативными являются только 3 из них – первая, вторая и третья, считая от 5' конца. Область Шайна-Дальгарно и старт кодон шпильку не связываются.

Таким образом, на основании этих данных можно предположить, что регуляция у других бактерий идет по схожему с *E.coli* принципу, при этом сами структуры регуляторных участков не являются уникальными для каждой бактерии, а имеют консервативные участки. При этом структуры имитируют участок связывания с рРНК и/или включают в свой состав область Шайна-Дальгарно и старт-кодон гена, в результате чего его экспрессия становиться невозможна.

Таблица 1. Таблица соответствия между известными регуляторными структурами в *E. coli* и их реализации в других гамма протобактериях. (Знаком «+» отмечено, что у данной бактерии обнаружена сходная структура. Знак «-» обозначает либо наличие регулирующей структуры , но имеющее более значительные отличие от E.coli либо отсутствие регуляции данного оперона.)

Bacteria	*Inf3-L35-L20*	*rpso(S15)*	*alpha*	*spc (S8)*	*S10(L4)*	*L11(L1)*	*beta(L10)*
Actinobacillus pleuropneumoniae	+	-	+	+	+	+	+
Buchnera aphidicola str.aps	-	-	-	-	-	-	-
Colwellia psychrerythraea	-	-	-	+	-	-	+
Escherichia coli	+	+	+	+	+	+	-
Haemophilus influenzae	-	+	+	+	+	+	+
Mannheimia succiniciproducens	+	+	+	+	+	+	+
Pasteurella multocida	-	+	+	+	+	+	-
Photorhabdus luminescens	+	+	+	+	+	+	+
Pseudoalteromonas haloplanktis	-	+	+	+	-	+	+
Pseudomonas aeruginosa	+	-	-	-	-	-	+
Shewanella oneidensis	+	+	+	+		+	+
Sodalis glossinidius	+	+	+	+	+	+	+
Vibro cholerae	+	+	+	+	+	+	+
Vibro fischeri	+	+	+	+	+	+	+
Wigglesworthia glossinidia	-	-	-	-	-	-	-
Xanthomonas campestris	-	-	-	-	-	-	-
Yersinia pestis	+	+	+	+	+	+	+

Поиск такой же регулирующей структуры, как у *L11* проведен для бактерий с большим эволюционным расстоянием. В результате чего она была
Найдена у альфа- протобактерий (Rhizobiales, Rhodobacterales), бета- протобактерий (Burkholderiales), у гамма протобактерий в тех группах где не была обнаружена вышеописанным методом (Pseudomonadales, Xanthomonadales), у грамположительных (Bacillales, Lactobacillales), у клостридий и микоплазм. Причем в фирмикутах сайты найдены либо перед обоими генами оперона, либо перед одним геном *rplK* или *rplA*. Не найдена регуляция у цианобактерий, дельта-протобактерий, Bacteroidetes, эплисон-протобактерий, а также архей. Стоит подчеркнуть, что регуляции не обнаружено в лидерной области перед геном (его ортологами) *rplK*, вполне возможно что в результате структурных перестроек оперона этот сингал локализуется в другом месте. Так, например, для архей известно, что гены *rplK* и *rplA* составляющие *L11* оперон у *E.coli*, в археях расположены на значительном расстоянии друг от друга и искомый сигнал лежит в начале гена *rplA*.

Список литературы.

[1] М. С. Гельфанд. *Молекулярная биология*. Цис-регуляторные структуры РНК бактерий ,том 40,№ 4, с. 609-619, 2006 год.

[2] Merianos H.J., Wang J., Moore P.B. 2004. “The structure of a ribosomal protein *S8/spc* operon mRNA complex”.*RNA*. , 954–964.

[3]
Guillier M., Allemand F., Raibaud S., Dardel F., Springer M., Chiaruttini C. 2002.”Translational feedback regulation of the gene for L35 in *Escherichia coli* requires binding of ribosomal protein L20 to two sites in its leader mRNA: A possible case of ribosomal”

[4] Mathy N., Pellegrini O., Serganov A., Patel D.J., Ehresmann C., Portier C. 2004.” Specific recognition of rpsO mRNA and 16S rRNA by *Escherichia coli* ribosomal protein S15 relies on both mimicry and site differentiation”.*Mol. Microbiol*., 661–675.

[5] Hanner M., Mayer C., Kohrer C., Golderer G., Grobner P.,Piendl W. 1994. “Autogenous translational regulation of the ribosomal MvaL1 operon in the archaebacterium *Methanococcus vannielii. J. Bacteriol”*. 409–418.

Построение HMM, основанной на правилах взаиморасположения сайтов связывания транскрипционных факторов, и ее применение для поиска корегулируемых генов в геномах рода Drosophila

А.А. Никулова*, Р.А. Сутормин, А.А. Миронов
Факультет Биоинженерии и Биоинформатики Московского Государственного Университета им. М.В. Ломоносова, 119992, Россия, Москва, Воробьевы горы 1-73
* nikanka@gmail.com

А.В. Фаворов
Государственный научно-исследовательский институт генетики и селекции промышленных микроорганизмов, 117545, Россия, Москва, 1-ый Дорожный пр-д, 1

Аннотация

Для понимания регуляции транскрипции генов необходимо знать закономерности, лежащие в основе группировки сайтов связывания факторов транскрипции и формирования ткане-специфичных регуляторных элементов. В данной работе мы попытались выявить правила взаиморасположения сайтов, сохраняющиеся в регуляторных областях гомологичных генов развития в геномах рода Drosophila . На основе выявленных правил была построена HMM, которая была применена для поиска схожих регуляторных модулей в предгенных областях других генов и выявления корегулируемых генов.

1. Введение

Понимание процесса экспрессии генов и его регуляции на разных этапах развития и в ответ на воздействия окружающей среды – одна из важнейших задач современной биологии. Регуляция экспрессии генов осуществляется на многих уровнях синтеза белка, как пред- так и посттрансляционно. На транскрипционном уровне контроль осуществляется белками (транскрипционными факторами, ТФ), связывающимися с определенными последовательностями ДНК в регуляторных областях генов, что приводит к активации или репрессии транскрипции.

В последнее время, становится все более очевидно, что у эукариот транскрипционная регуляция часто зависит от взаимодействия между множеством белков, и что клеточный контекст определяет, какой из регуляторов будет доступен для связывания ДНК и влияния на уровень транскрипции [1].

Такой способ регуляции, основанный на узнавании белками-регуляторами специфических последовательностей, должен был бы позволить легко искать ССТФ (сайты связывания факторов транскрипции) методами биоинформатики. Однако, малая длина регуляторных сайтов эукариот (от 5 до 15 нуклеотидов) и их вырожденность приводят к тому, что большое число потенциальных сайтов в геноме находится по чисто случайным причинам. Большинство из них не участвует в функциональной регуляции, будучи закрытыми хроматином, находясь вдали от других необходимых ССТФ или по каким-то другим причинам [2]. Поэтому для повышения эффективности предсказания регуляторных элементов компьютерными методами необходимо использовать другие критерии дополнительно к информации о последовательностях сайтов.

Известно, что ССТФ часто организованы в группы (кластеры, регуляторные модули), покрывающие участки ДНК протяженностью несколько сотен пар оснований. Эти модули координируют белок-белковые взаимодействия, тем самым регулируя уровень транскрипции генов [3]. Таким образом, поиск кластеров ССТФ, а не отдельных сайтов увеличивает точность предсказания. Этот подход достаточно широко используется для поиска регуляторных участков в геномах эукариот [4-5]. Однако большинство исследователей обращают внимание именно на близкое расположение сайтов, тогда как было показано, что во многих случаях важным фактором является порядок расположения ССТФ и расстояния между ними [6-7].

Вместе с тем, сравнительная геномика предоставляет мощный инструмент для определения функциональных регуляторных участков, так как биологически значимые элементы обычно сохраняются в процессе

эволюции. При этом набор регуляторных элементов перед гомологичными генами обычно сохраняется, а изменяется (если изменяется) только их локализация и иногда порядок расположения. В ситуациях, когда уровень консервативности между последовательностями низок, более разумным кажется подход, опирающийся не столько на консервативность последовательности, сколько на сохранение структуры регуляторных модулей (набор сайтов, их взаиморасположение) [7, 8].

2. Материалы и методы

В данной работе мы учитывали тенденцию сайтов образовывать кластеры и сохранение структуры регуляторных модулей между видами для поиска регуляторных участков в геномах рода *Drosophila*. Основная идея заключалась в том, что ТФ'ы взаимодействуют между собой, связываясь с соответствующими ССТФ в регуляторных областях. Соответственно, для того, чтобы такое взаимодействие было возможно, расстояния между сайтами должно сохраняться или подчиняться неким правилам (например, расстояние должно быть кратно длине витка спирали ДНК). При этом мы учитывали композиции только пар сайтов и считали, что в «гомологичных» кластерах может меняться расположение таких пар при сохранении расстояний в парах. Для реализации данной идеи была создана НММ, учитывающая расстояния между сайтами определенных типов.

На первом этапе мы разработали список правил, которым могут подчиняться расстояния между сайтами. Также с использованием известных позиционных весовых матриц были найдены все сайты во всех 12 геномах *Drosophila*, имеющие достаточно большой вес. Затем для каждого известного, экспериментально определенного регуляторного модуля из *D.melanogaster* была проведена оценка правил, которые наблюдаются в данном модуле и при этом достаточно редки в геометрически размеченных кластерах (скоплениях сайтов с весом не меньше заданного). Таким образом, каждому правилу приписывался свой вес.

Затем была сконструирована НММ, формирующая кластеры из предварительно найденных сайтов. Данная НММ выбирает наилучший путь среди путей, проходящих через имеющиеся сайты. Таким образом, эмиссиями в данном случае являются события включения/невключения сайтов в кластер. Исходные переходные вероятности основаны на весах правил, оцененных для конкретного кластера. Затем проводилось обучение НММ на предгенных областях гомологичных генов модифицированным методом Баума-Велча.

3. Результаты

Алгоритм, описанный в методах, был применен для формирования скрытых Марковских моделей, описывающих экспериментально показанные кластеры сайтов связывания транскрипционных факторов, связанных с развитием. Далее обученные НММ были применены для поиска похожих регуляторных областей в предгенных областях других генов. При этом предполагалось, что гены, имеющие схожие регуляторные модули корегулируются. Найденные таким образом гены могут рассматриваться в качестве кандидатов, корегулируемых совместно с генами, по которым строились НММ.

Полученные результаты для генов развития геномов рода *Drosophila* можно получить по запросу к авторам.

4. Список цитируемой литературы

[1] Levine, M., and R. Tjian, "Transcription regulation and animal diversity", *Nature*, 424, 2003. pp. 147-151.

[2] Wasserman, W.W. and Sandelin, A., "Applied bioinformatics for the identification of regulatory elements", *Nat. Rev. Genet.*, 5, 2004, pp. 276-287.

[3] Halfon, M.S., Y. Grad, G.M. Church and A.M. Michelson, "Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model", *Genome Res.*, 12, 2002, pp. 1019-1028.

[4] Berman, B.P., Nibu, Y., Pfeiffer, B.D., Tomancak, P., Celniker, S.E., Levine, M., Rubin, G.M., and Eisen, M.B., "Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome", *Proc. Natl. Acad. Sci.*, 99, 2002, pp. 757-762.

[5] Levy S., Hannenhalli S., "Identification of transcription factor binding sites in the human genome sequence", *Mamm. Genome.*, 13(9), 2002, pp. 510-514.

[6] Makeev V.J., Lifanov A.P., Nazina A.G., Papatsenko D.A., "Distance preferences in the arrangement of binding motifs and hierarchical levels in organization of transcription regulatory information", *Nucleic Acids Res.*, 31(20), 2003, pp. 6016-6026.

[7] O. Hallikas, K. Palin, N. Sinjushina, R. Rautiainen, J. Partanen, E. Ukkonen and J. Taipale, "Genome-wide Prediction of Mammalian Enhancers Based on Analysis of Transcription-Factor Binding Affinity", *Cell*, 124, 2006, pp. 47-59.

[8] Erives A. and Levine M. "Coordinate enhancers share common organizational features in the Drosophila genome", *Proc. Natl. Acad. Sci. USA*, 101, 2004, pp. 3851-3856.

Функциональная классификация генов полных бактериальных геномов на основе базы данных Clusters of Orthologous Groups (COG)

М. Казанов
Институт проблем передачи информации им. А. А. Харкевича
mkazanov@mail.ru

Аннотация

Гены полных бактериальных геномов кластеризованы в группы ортологов согласно классификации базы данных COG при помощи специально разработанного программного обеспечения. Классификация выполнялась на основе анализа результатов поиска гомологичных белковых последовательностей относительно базы данных COG при помощи алгоритма BLASTP. Всего было классифицировано более двух миллионов генов из 623-х полных бактериальных геномов. Проведен сравнительный анализ распределений генов по функциональным категориям. Полученная база данных предоставляется по запросу.

1. Введение

База данных ортологических групп (COG) была создана в 1997 году в Национальном центре по биотехнологической информации [1]. Первоначальная версия базы данных содержала 720 кластеров ортологических генов (6814 генов), полученных в результате анализа семи полных геномов, принадлежащих пяти филогинетически удаленным таксонам. Обновленная версия базы данных, представленная в 2003 году, содержала уже около пяти тысяч кластеров (138458 генов), являющихся результатом обработки 66-ти полных геномов одноклеточных организмов, а также включала группы ортологических генов из семи эукариотических геномов.

Создание базы данных COG внесло значительный вклад в исследования в области сравнительной и эволюционной геномики. Так, на момент июня 2008 года в базе данных публикаций PubMed [2] можно было найти более пятиста публикаций, содержащих ссылки на статьи, представляющие научному сообществу базу данных COG [1,3]. Кластеры ортологических генов явились основой для функциональной аннотации последовательностей, путем сравнения неаннотированных последовательностей с белковыми последовательностями, входящими в кластеры базы данных COG. При создании базы данных COG каждому кластеру ортологических генов была присвоена одна уникальная функция, так как ортологические гены, как правило, имеют одинаковую функциональность. Поэтому факт наилучшей схожести рассматриваемой последовательности с последовательностями из определенного кластера позволяет с большой степенью уверенности предсказать функцию неаннотированной последовательности, перенеся функцию из соответствующего кластера. Кроме этого, создание базы данных COG имело большое значение для исследований в области эволюционной геномики, так как каждый кластер ортологических генов, теоретически, представляет собой совокупность генов, возникших в процессе эволюции одного предкового гена.

Количество секвенированных полных бактериальных геномов продолжает расти и на момент начала 2008 года составляло более 600 геномов, что почти в 10 раз больше количества геномов, использованных при построеннии обновленной версии БД COG вышедшей в 2003 году. Используя новые данные о полных бактериальных геномах возможно значительно расширить ортологические кластеры последней версии базы данных COG, что еще более увеличит возможности функциональной аннотации и эволюционных исследований, выполняемых на ее основе. В данной работе было выполнено автоматическое расширение кластеров ортологических генов БД COG, на основе существующих данных о секвенированных полных бактериальных геномов, с помощью специально разработанного программного обеспечения. Полученная база данных результатов расширения ортологических кластеров предоставляется по запросу.

2. Материалы и методы

Расширение ортологических кластеров БД COG выполнялось путем сравнения транслированных последовательностей генов секвенированных полных бактериальных геномов с белками, входящими в состав БД COG, с помощью алгоритма BLASTP [4] и последующего анализа результатов выдачи данного алгоритма. Обновленная версия БД COG, как указано в [3],

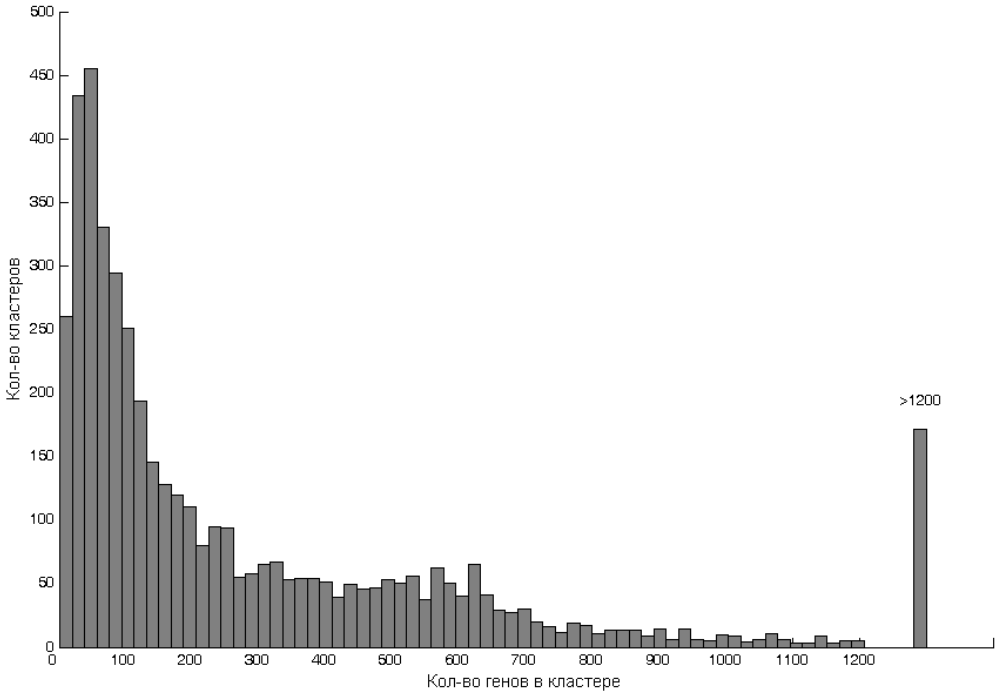


Рисунок 1. Распределение кластеров ортологических групп по числу генов в кластере.

создавалась аналогичным образом, однако полное описание метода обработки выдачи алгоритма BLASTP не было опубликовано, в связи с чем, был разработан собственный метод и реализующее его программное обеспечение.

Для каждого гена из 557 секвенированных полных бактериальных геномов, не входящих в БД COG, на вход разработанного алгоритма подавалась его транслированная последовательность. На выходе данному белку назначалось принадлежность либо к одному кластеру ортологических генов, либо, в случае мультидоменных белков, к нескольким кластерам.

Первоначально для каждого гена в стандартной выдаче алгоритма BLASTP, полученной в результате сравнения рассматриваемой белковой последовательности со всеми белками БД COG, были отфильтрованы потенциальные гомологи не удовлетворяющие следующим условиям:

1) Уровень Identity > 20%.
2) E-value < 10-5
3) Длина выравнивания превышает 2/3 длины белка, входящего в БД COG.

Первые два условия позволили отфильтровать кандидатов, имеющих низкий уровень схожести с рассматриваемой белковой последовательностью. Третье условие позволило отобрать белки БД COG, большая часть последовательности которых сходна со всей рассматриваемой последовательностью или с ее отдельной областью. В том случае, когда после выполнения фильтрации список кандидатов оставался пуст, рассматриваемая белковая последовательность не была отнесена ни к одному кластеру.

Если по окончании выполнения операции фильтрации среди списка кандидатов оставались кандидаты, относящиеся только к одному кластеру, то рассматриваемому гену назначался данный кластер. Если же среди хитов оставались хиты, принадлежащие различным кластерам, что, как правило, происходило в случае мультидоменных белков, то, первоначально, все кандидаты разбивались на группы, относящиеся к различным доменам белка. Далее, для каждой группы, выбирался наилучший по значению Identity кандидат, а кластер, к которому относился данный кандидат, назначался рассматриваемому гену.

Тестирование разработанного алгоритма выполнялось при помощи метода кросс-валидации [5]. Все белки БД COG разбивались случайным образом на 10 подгрупп. Далее, алгоритм применялся поочередно к каждой из подгрупп, и, полученная точность воспроизведения классификации БД COG фиксировалась. При назначении кластеров белковым последовательностям из каждой подгруппы база данных белков COG составлялась из белков остальных девяти подгрупп. Средняя точность воспроизведения по рассматриваемым десяти подгруппам составила около 97,5%. Разработанный метод был реализован на языке PL/SQL в рамках СУБД Oracle 10g Express Edition [6].

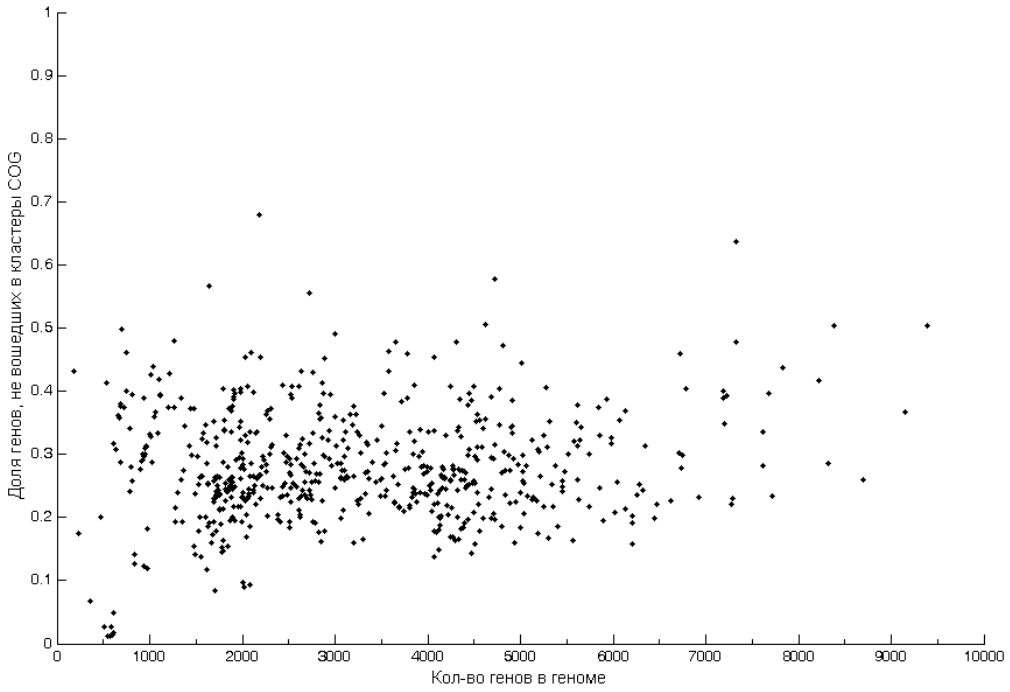


Рисунок 2. Доля генов, не вошедших в кластеры ортологических групп, в зависимости от размера геномов.

3. Результаты

Используя разработанный метод, мы обработали более двух миллионов белковых последовательностей. Средний размер кластеров в созданной базе данных оставил примерно 320 белков, по сравнению с 30 белками обновленной версии БД COG. Однако максимум распределения размеров кластеров приходится на гораздо меньшее число (рис. 1), а довольно большой средний размер кластера объясняется наличием значительного количества кластеров большого размера (>1000 генов).

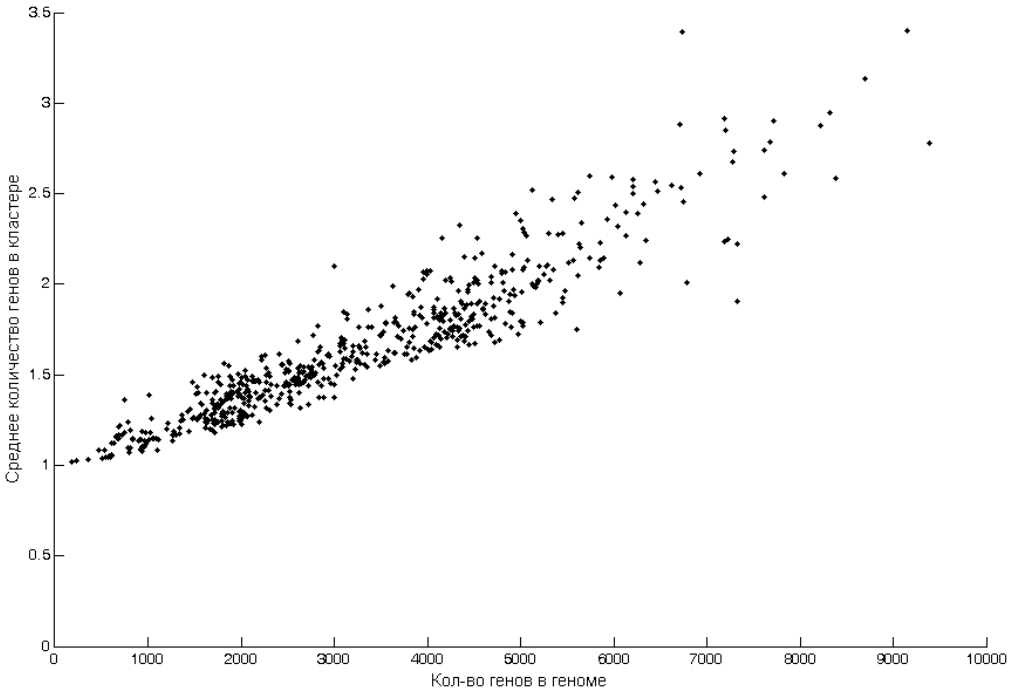


Рисунок 3. Зависимость среднего размера кластера ортологических групп в геноме от размера генома.

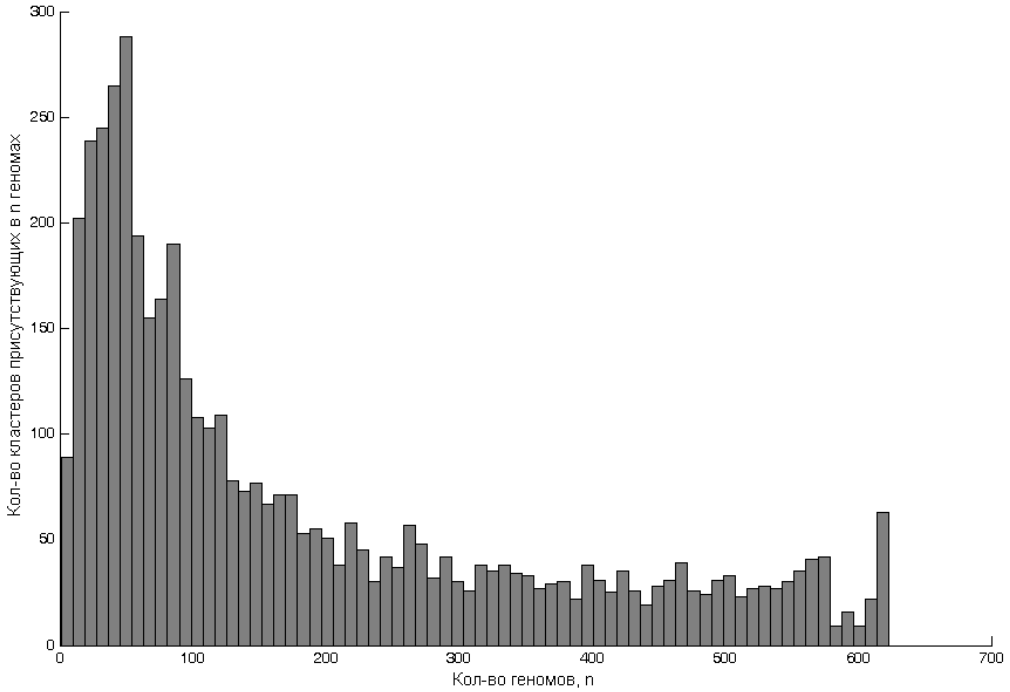


Рисунок 4. Распределение кластеров ортологических групп по присутствию в определенном числе геномов.

Наряду с приблизительно полутора тысячами классифицированными на принадлежность к кластерам генов (~70%), около шестисот тысяч генов не были причислены ни к одному кластеру. Доля генов, не причисленных к кластерам, для большинства геномов составляет примерно 30% и мало зависит от размера генома (рис.2). При этом среднее количество генов в кластерах COG в одном отдельно взятом геноме с увеличением размера генома также увеличивается (рис.3).

На рис. 4 представлено распределение кластеров ортологических генов по числу геномов, в которых присутствует каждый рассматриваемый кластер. На данном распределении хорошо видно, что количество кластеров, присутствующих одновременно в большинстве рассматриваемых геномов, незначительно. Этот факт, по-видимому, подтверждает известное наблюдение [7], что в процессе эволюции у филогенетически удаленных видов часто независимо формировались гены, выполняющие одну и ту же важную функцию.

На рис. 5 изображено распределение генов, отнесенных к кластерам БД COG, по функциональным категориям, в сравнении с распределением генов, входящих в последнюю версию БД COG. Хорошо видно, что оба распределения очень хорошо согласуются, что с одной стороны, подтверждает правильность работы примененного алгоритма, а с другой стороны, показывает, что вид распределения генов по функциональным категориям практически не изменился при почти десятикратном увеличении количества секвенированных полных бактериальных геномов.

4. Заключение

Примененная процедура расширения БД COG на основе доступных полных бактериальных геномов позволила существенно увеличить размеры кластеров ортологических генов. Большие размеры образованных кластеров и увеличение таксономического разнообразия рассмотренных бактериальных геномов дают возможность повышения мощности и детализации исследований в областях сравнительной и эволюционной геномики, основанных на использовании БД COG. Число генов, не присвоенных к кластерам, и средний размер кластера также позволяет предположить наличие еще около двух тысяч кластеров ортологических генов, не вошедших в БД COG.

5. Благодарности

Данная работа была выполнена совместно с М. С. Гельфандом. Автор выражает благодарность компаниям «ФОРС-Центр разработки» и «Атлантис коммуникэйшнс», на серверных площадках которых были произведены вычислительные работы в рамках данного исследования.

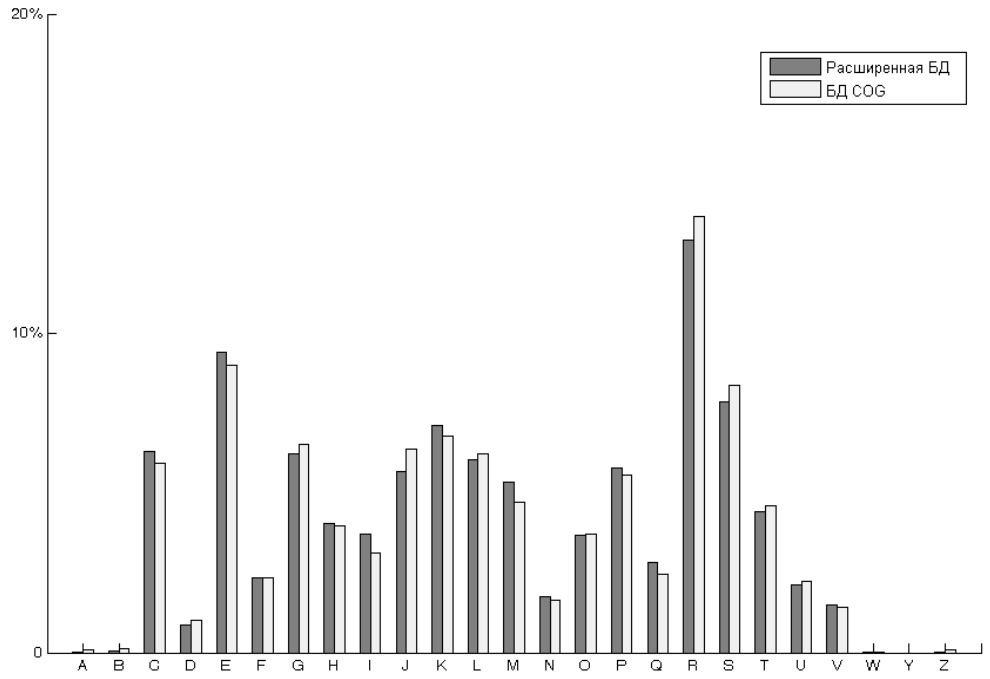


Рисунок 5. Функциональная классификация кластеров ортологических групп построенной расширенной базы данных в сравнении с БД COG. Обозначения функциональных категорий следующие: A – процессинг и модификация РНК; B – динамика и структура хроматина; C – генерация и накопление энергии; D - регуляция цикла клетки и митоз; E – метаболизм и транспорт аминокислот; F – метаболизм и транспорт нуклеотидов; G – метаболизм и транспорт углеводов; H – метаболизм коэнзимов; I – метаболизм липидов; J – трансляция; K – транскрипция; L – репликация и репарация; M – биогенез клеточных мембран; N – двигательная активность клетки; O – пост-трансляционные модификации; P – транспорт и метаболизм неорганических ионов; Q – биосинтез, транспорт и катаболизм вторичных метаболитов; T – трансдукция сигнала; U – внутриклеточный транспорт и секреция; Y – структура ядра; Z – цитоскелет; R – предсказана только биологическая активность; S – неизвестная функция.

[1] R.L. Tatusov, E.V. Koonin, D.J. Lipman, “A genomic perspective on protein families”, *Science*, 278, 1997, pp. 631-637.

[2] http://www.pubmed.gov

[3] R.L. Tatusov, N.D. Fedorova, J.D. Jackson, A.R. Jacobs, B. Kiryutin, E.V. Koonin, D.M. Krylov, R. Mazumder, S.L. Mekhedov, A.N. Nikolskaya, B.S. Rao, S. Smirnov, A.V. Sverdlov, S. Vasudevan, Y.I. Wolf, J.J. Yin, D.A. Natale, “The COG database: an updated version includes eukaryotes”, *BMC Bioinformatics* 11, 2003, pp. 4-41.

[4] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, “Basic local alignment search tool”, J. Mol. Biol. 215, 1990, pp. 403-410.

[5] P. A. Devijver, J. Kittler, *Pattern Recognition: A Statistical Approach*, Prentice-Hall, London, 1982.

[6] http://www.oracle.com/technology/products/database/xe/index html

[7] A. R. Mushegian, E. V. Koonin, “A minimal gene set for cellular life derived by comparison of complete bacterial genomes”, *PNAS*, 93(19), 1996, pp. 10268-10273.

Корреляции между аминокислотными последовательностями транскрипционных регуляторов суперсемейства CRP-FNR и нуклеотидными последовательностями их операторов.

Д.С.Митева
Московский Государственный Университет им. М.В.Ломоносова, Факультет Биоинженерии и Биоинформатики
desislava@rambler.ru

В.В.Степанова
Московский Государственный Университет им. М.В.Ломоносова, Факультет Биоинженерии и Биоинформатики
vita.stepanova@gmail.com

Аннотация

Определены статистически значимо скоррелированные позиции выравнивания ДНК-связывающих белков суперсемейства CRP-FNR и выравнивания их сайтов связывания. Наиболее скоррелированными являются позиции 180=>7/-7, 181=>5/-5(нумерация а.о. ведется по Crp_Ecoli; нумерация нуклеотидов относительно центра палиндрома). Эти пары соответствуют описанным в литературе распознающим контактам.
Кроме них еще скоррелированы пары 190=>7/-7, 183=>3 и 176=>3 функциональная роль которых не ясна, но можно сделать предположение о том, что а.о. в этих позициях определяют общее положение распознающей спирали относительно ДНК.

1. Введение

Специфическое ДНК-белковое распознавание лежит в основе всех основных генетических процессов, механизмы такого распознавания активно изучаются, однако до сих пор остается много непонятного. В нашей группе создан алгоритм, реализованный как программа DPCORR, для поиска статистически значимо скоррелированных позиций в выравнивании ДНК-связывающих белков и в выравнивании сайтов их связывания. На примере семейства LACI регуляторов получены первые интересные результаты. Задача настоящей работы – исследование таких корреляций у суперсемейства бактериальных транскрипционных регуляторов CRP-FNR.

По данным наших коллег была подготовлена выборка из 55 транскрипционных регуляторов семейств CooA, HcpR, DNR, NnrR, Crp и Fnr (среднее значение ID = 26%), а также выборка соответствующих им 551 их сайтов (среднее значение ID = 42%). Сайты связывания представляют собой более или менее идеальные четные палиндромы от 14 до 22н. С помощью инструментов БД SMART и Pfam были выделены ДНК-связывающие домены длины 49-52 а.о., и построено их множественное выравнивание с помощью программы MUSCLE. Нуклеотидное выравнивание было построено путем совмещения центров палиндромов. Принцип нумерации позиций нуклеотидов:

T	G	T	G	A	t	t	T	T	G	g	T	C	A	C	A	
-8	-7	-6	-5	-4	-3	-2	-1	1	2	3	4	5	6	7	8	

Нумерация позиций в белках - по аминокислотной последовательности CRP_ECOLI(UniProt AC P0ACJ8).

2. Результаты

Для каждой пары позиций 2-х выравниваний с помощью программы DPCORR были вычислено значение полной взаимной информации и статистическая значимость этого значения (Z-score). Пороговое значение Z-score было определено с помощью B-cutoff[1] и по алгоритму FDR[2]. Показано, что значимо скоррелированы 8 пар позиций (B-cutoff) или 7 пар (FDR). В порядке убывания Z-score - это пары 180=>7/-7, 181=>5/-5, -5, 190=>7/-7, 183/3 и 176/3.

Полученные пары были наложены на структуру 1CGP (PDB ID), ДНК-белковые контакты определены с помощью WHATIF (http://swift.cmbi.ru.nl/servers/html/index.html).
Оказалось, что первые четыре скоррелированные пары соответствуют плотным специфическим контактам Arg180=>Gua25 и Glu180=>Cyt5. Всего контактов боковых групп а.о. с азотистыми основаниями — 13 (в этих контактах участвуют 5

а.о. и 9 нуклеотидов), из них 2 пары, Arg180=>Gua25 и Glu181=>Cyt5, скоррелированы. Другими словами, далеко не все пары, соответствующие реальным контактам боковых групп а.о. с азотистыми основаниями, скоррелированы, но наиболее скоррелированные пары соответствуют реальным контактам, описанным в литературе как распознающие, см., например, таблицу особенностей HTH_CRP-домена в БД CDD (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml).

Систематический поиск по БД PubMed каких-либо экспериментальных сведений о рассмотренных нами белках показал, что данные по точечному мутагенезу ДНК-связывающего домена есть только для белка CRP_Ecoli. Показано, что замена Arg180 на Lys или Leu снижает эффективность связывания с ДНК[3], а мутирование Glu181 приводит к тому. что предпочтительный сайт связывания меняется с AA-TGTGA--T---TCA-ATW на AA-TGTAA—T---TCA-ATW[4].

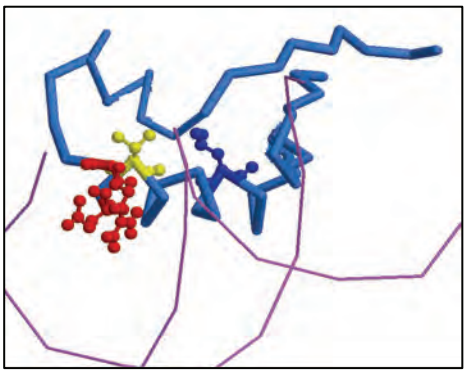


Рис.1. *Расположение* в структуре 1CGP аминокислотных остатков в наиболее скоррелированных позициях.
Показана только часть белка. В центре – распознающая спираль. В шарнирной модели красным цветом отмечены наиболее скоррелированные контактирующие с ДНК а.о., жёлтым цветом – Val183, синим цветом – Leu190, находящиеся на спине распознающей спирали, фиолетовым – ДНК.

3. Заключение

Таким образом, полученные нами результаты хорошо согласуются со структурными и экспериментальными данными и подтверждают сделанное ранее нашим коллегой Дмитрием Родионовым наблюдение о том, что в ряду подсемейств CRP-FNR суперсемейства Arg180 и Glu181 встречаются в белковых последовательностях, если в операторе есть гуанин в третьей позиции и аденин в шестой позиции соответственно[5].

С другой стороны, мы показали, что пары 190/ 7, 183/3и 176/3 хотя и слабее, но тоже скоррелированы. Соответствующие аминокислотные остатки не контактируют с ДНК, зато вовлечены во внутрибелковые взаимодействия. Так, в структуре 1CGP валин в позиции 183 находится на «спине» распознающей спирали 180-191 и образует гидрофобные контакты сразу с 3-мя остатками альфа-спирали 168-176, лейцин в позиции 190 также находится на «спине» распознающей спирали и образует гидрофобные контакты с остатками бета-тяжа 195-199.В этом же гидрофобном ядре расположен и валин 176. Можно предположить, что эти взаимодействия наиболее важны для правильной ориентации распознающей спирали в большой бороздке ДНК, представляется интересным экспериментально проверить это предположение.

4. Литература

[1] Kalinina OV, Novichkov PS, Mironov AA, Gelfand MS, Rakhmaninova AB.
SDPpred: a tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins.
Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W424-8.

[2] Benjamini, Y., Hochberg, Y. (1995). " Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing ", *Journal of the Royal Statistical Society B*, 57 289-300.

[3] Gent ME, Gronenborn AM, Davies RW, Clore GM. Probing the sequence-specific interaction of the cyclic AMP receptor protein with DNA by site-directed mutagenesis. *Biochem J*. 1987 Mar 15;242(3):645-53

[4] Ebright RH, Cossart P, Gicquel-Sanzey B, Beckwith J. Mutations that alter the DNA sequence specificity of the catabolite gene activator protein of E. coli.
Nature 311, 232 - 235 (20 September 1984)

[5] Dmitry A. Rodionov, Inna L. Dubchak, Adam P. Arkin, Eric J. Alm, Mikhail S. Gelfand. Dissimilatory Metabolism of Nitrogen Oxides in Bacteria: Comparative Reconstruction of Transcriptional Networks.
PLoS Comput Biol. 2005 Oct;1(5):e55. Epub 2005 Oct 28.

Эволюция альтернативного сплайсинга генов млекопитающих

Нуртдинов Рамиль Наилевич
Факультет биоинженерии и биоинформатики Московского Государственного Университета им. М.В. Ломоносова
n_ramil@mail.ru

Аннотация

В процессе эволюции организмы могут приобретать новые свойства либо путем добавления новых генов, либо путем добавления нуклеотидной информации в уже существующие гены, например созданием нового экзона. Альтернативный сплайсинг является идеальным механизмом тестирования и отбора таких экзонов. Первоначально новый экзон имеет небольшую частоту включения и отсутствует в большей части мРНК. Если кодируемый альтернативным вариантом мРНК белок окажется полезным, то возможность его создания может быть закреплена в популяции. Согласно современным оценкам доля альтернативно сплайсируемых генов человека или мыши составляет около 60%, причем большая часть альтернативных вариантов мРНК не консервативна и имеет крайне низкую частоту использования. Анализ консервативности альтернативного сплайсинга дает возможность выделить экзоны, появившиеся в ходе эволюции разных видов и таксонов. Был проведен анализ консервативности элементарных событий альтернативного сплайсинга генов мыши в геномах человека, собаки и крысы. Максимальная консервативность альтернативно сплайсируемых участков наблюдается в генах с наименьшей скоростью молекулярной эволюции белковой последовательности, в то время как в более быстро эволюционирующих генах возрастает доля неконсервативных альтернатив.

1. Введение

Характерной особенностью эукариотических генов является существование механизма вырезания из первичного транскрипта пре-мРНК протяженных участков, называемых интронами. Оставшиеся участки, экзоны, сшиваются, и получаемая мРНК впоследствии используется как матрица для синтеза белка. Процесс вырезания интрона и сшивки экзонов называется сплайсинг. Каждое событие сплайсинга вырезает один интрон и, как правило, спласинг интрона происходит независимо от сплайсинга остальных интронов. Сплайсинг осуществляется сплайсосомой, комплексом, состоящим из нескольких малых ядерных РНК и большого числа белков, непосредственно участвующих в процессе вырезания интрона. Согласно современным оценкам доля альтернативно сплайсируемых генов человека и мыши варьируется в интервале 40-75% [1,2,3]. Сравнение ДНК полных геномов мыши и человека показало, что большинство генов мыши имеют ортологичный ген в ДНК человека (ортологи или ортологичные гены - гены, имеющие общее эволюционное происхождение). Экзон-интронная структура ортологичных генов человека и мыши практически полностью консервативна, совпадет количество интронов в гене и длины экзонов [4]. Заметная доля альтернативно сплайсируемых участков является специфичной для определенных видов и/или для таксономических групп, при сравнении геномов человека и мыши [5,6,7,8], мыши и крысы [9], и сразу нескольких видов млекопитающих [10,11]. Однако парное сравнение не позволяет понять эволюционное происхождение неконсервативных участков, эти участки могут быть как и эволюционными новоприобретениями, так и эволюционными потерями.

В результате анализа консервативности событий альтернативного сплайсинга генов человека в геномах мыши и собаки [12] было показано, что консервативность кассетных экзонов и альтернативных сайтов сплайсинга зависит от EST-покрытия альтернативного варианта мРНК и его способности сохранять рамку считывания белковой последовательности. Большинство из неконсервативных кассетных экзонов и альтернативных сайтов сплайсинга обладают минимальным EST-покрытием в одну последовательность и могут являться ошибками

сплайсинга или ошибками выделения EST. Недостаточное EST-покрытие альтернативного варианта мРНК может быть обусловлено малым количеством последовательностей для гена целиком, например проведенный нами анализ данных EntrezGene [13] и UniGene [13] показал, что из приблизительно 30 тысяч генов человека 20% не имеют данных EST и 20% генов имеют менее 20 последовательностей EST в UniGene.

Сравнение данных человека, собаки и мыши не позволяет различить неконсервативные кассетные экзоны и альтернативные сайты сплайсинга от возможных ошибок. Необходимо провести анализ консервативности наблюденных неконсервативных вариантов в более близком геноме, например в геноме крысы для неконсервативных альтернатив мыши, или в геноме макаки для неконсервативных альтернатив человека. Консервативность в геноме крысы, с учетом поправки на сходство последовательностей ДНК мыши и крысы, является достаточным условием для признания альтернативного варианта функциональным, а отсутствие консервативности в геномах человека и собаки позволяет трактовать его как новоприобретение в ходе эволюции грызунов.

В ходе эволюции гены накапливают мутации, многие из которых приводят к аминокислотным заменам в последовательности белка. Однако скорость накопления мутаций различается для разных генов. Например, существует несколько сотен генов, для которых аминокислотные последовательности белков человека, мыши и собаки полностью совпадают, с другой стороны, множество генов поменяли до половины своих аминокислот в процессе эволюции. Многие из высоко консервативных генов играют ключевую роль в клеточных процессах, большинство из них можно обнаружить у бактерий и одноклеточных эукариот, поэтому не удивительно, что в этих генах наблюдается так мало меняющих аминокислотную последовательность мутаций. Быстро эволюционирующие гены необходимы для процессов видообразования и расхождения видов [14,15,16].

В ходе работы был проведен анализ консервативности элементарных событий альтернативного сплайсинга мыши в геномах человека, собаки и крысы. Для генов с разными скоростями молекулярной эволюции белковой последовательности была оценена доля неконсервативных и специфичных для грызунов кассетных экзонов.

2. Результаты и обсуждение

2.1 Создание базы данных альтернативно сплайсируемых генов EDAS

При создании базы данных EDAS использовалась версия 34.3 последовательности генома человека, опубликованная 2 августа 2004 года. Аннотация генома содержала информацию о 28957 генах. Последовательности EST и мРНК для выявленных генов человека были получены из 175.human версии UniGene, опубликованной 16 августа 2004 года. Было получено 4294590 EST и 114624 мРНК человека, кодирующая область была аннотирована для 91844 мРНК. Для мыши использовались версия 37.1 генома, 21 июня 2007 года, и версия UniGene37.1 от 7 сентября 2007 года. Аннотация генома мыши содержала информацию о 31811 генах. UniGene содержал данные о 3817531 EST и 215212 мРНК, кодирующая область была аннотирована для 126797 мРНК.

Последовательности мРНК и EST были выравнены с геномной последовательностью при помощи программы ProEST, белковые последовательности были выравнены при помощи программы ProFrame. В базе данных сохранялись все наблюденные выравнивания, однако в случае плохого выравнивания одного или нескольких экзонов (уровень сходства менее 80%), последовательность, порождающая данное выравнивание, игнорировалась. Для каждого гена на основе набора всех наблюденных экзонов и интронов был построен граф сплайсинга. Вершинами этого графа являются сайты сплайсинга, а ребрами – экзоны и интроны. Исходя из требуемого уровня достоверности, из данного графа можно удалять ребра с достоверностью ниже заданного порога. Анализ графа сплайсинга позволяет выделить альтернативный сплайсинг. Альтернативой называется наиболее близкая друг к другу, в смысле расстояния по графу, пара вершин, между которыми возможно два или более пути.

Многие альтернативы могут быть порождены в результате ошибок сплайсинга, поэтому была создана модель ошибочного сплайсинга, позволяющая с большой достоверностью выделять случаи “истинного” альтернативного сплайсинга. В рамках этой модели вероятность порождения ошибочного варианта мРНК описывается биноминальным распределением с параметром P_0=0.01 [17,18]. Для каждой наблюденной альтернативы в рамках данной модели оценивалась интегральная вероятность порождения ошибочной альтернативы, имеющей сходное EST-покрытие (K – EST-покрытие

альтернативного варианта и N – EST-покрытие всей альтернативы):

$$P = \sum_{i=K-1}^{N} \frac{N!}{i!\times(N-i)!}(0{,}01)^{i} \times (1-0{,}01)^{N-i}$$

Исследуемая альтернатива считалась ошибочной, в случае превышения данной вероятности значения 0.05.

2.2 Определение скоростей молекулярной эволюции ортогогичных генов

В качестве выборки ортологичных генов был использован материал из базы данных Homologene [13]. Для 11014 генов мыши из EDAS были получены ортологичные гены человека, собаки и крысы.

Для определения скоростей молекулярной эволюции генов необходимо в каждом гене выделить ортологичные экзоны и оценить их сходство. Использовались только последовательности белков человека, собаки и мыши, а скорость молекулярной эволюции генов крысы принималась равной скорости эволюции генов мыши. Набор экзонов был получен выравниванием белковых последовательностей мыши, человека и собаки с соответствующей последовательностью генома. Проведено попарное выравнивание кодируемых каждым экзоном аминокислотных последовательностей между разными видами. Ортологичные экзоны выявлялись при помощи метода BBH (best bidirectional hit или лучшее двустороннее сопоставление), распространенного на тройки выравниваемых объектов. Аминокислотные последовательности ортологичных экзонов были объединены в рамках соответствующих видов и выравнены друг с другом с использованием матрицы сходства Blosum62. Полученное распределение уровня сходства белковых последовательностей показано на Рисунке 1.

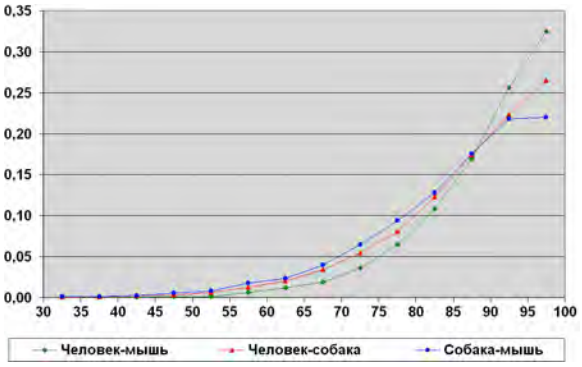


Рисунок 1. Распределение уровня сходства белков человека, мыши и собаки.

2.3 Определение консервативности экзонов

Первоначально последовательности ДНК генов были выровнены программой Blat [19]. Результатом этого выравнивания является цепочка консервативных на аминокислотном уровне участков. Программа Blat с высокой точностью находит участки сильной белковой консервативности, что с одной стороны позволяет идентифицировать экзоны, а с другой стороны отфильтровывает участки не экзонной консервативности, которые могут быть консервативными регуляторными элементами, повторами или просто случайными артефактами. Результатом такого первоначального выравнивания является набор пар сегментов ДНК, границами которых служат хорошо выравнивающиеся белковые экзоны. Для каждого такого сегмента исследовалась возможность реализации всех содержащихся внутри него интронов мыши в исследуемом гене человека, собаки или крысы. Результатом работы программы ProGene является набор пар ортологичных акцепторных и донорных сайтов. Внутренний кодирующий белок экзон считался консервативным, если его определены ортологи его акцепторного и донорного сайтов сплайсинга. Экзон, содержащий внутри себя терминирующий кодон, считался консервативным, если найден ортолог его акцепторного сайта.

Расхождение видов мыши и крысы произошло относительно недавно, поэтому консервативность в геноме крысы элементарных альтернатив мыши может быть объяснена сходством ДНК последовательностей, не накопивших достаточного количества мутаций, а не функциональной значимостью альтернативы, защищающей ее от случайных мутаций. При исследовании кассетных экзонов для учета такой возможности мы провели анализ консервативности случайно отобранных участков интронов мыши, имеющих сходное с кассетными экзонами выборки распределение длин: 16317 участков без терминирующих кодонов внутри, и 19111 участков, содержащих терминирующие кодоны. Анализ консервативности таких псевдоэкзонов показал, что консервативны 11.5% и 7.2% участков содержащих и не содержащих внутри терминирующий кодон, соответственно. Эти данные показывают, что достаточно большая доля кассетных экзонов мыши может быть консервативна в крысе по случайным причинам. Для отображения таких альтернатив была выделена отдельная группа – случайно консервативные кассетные экзоны.

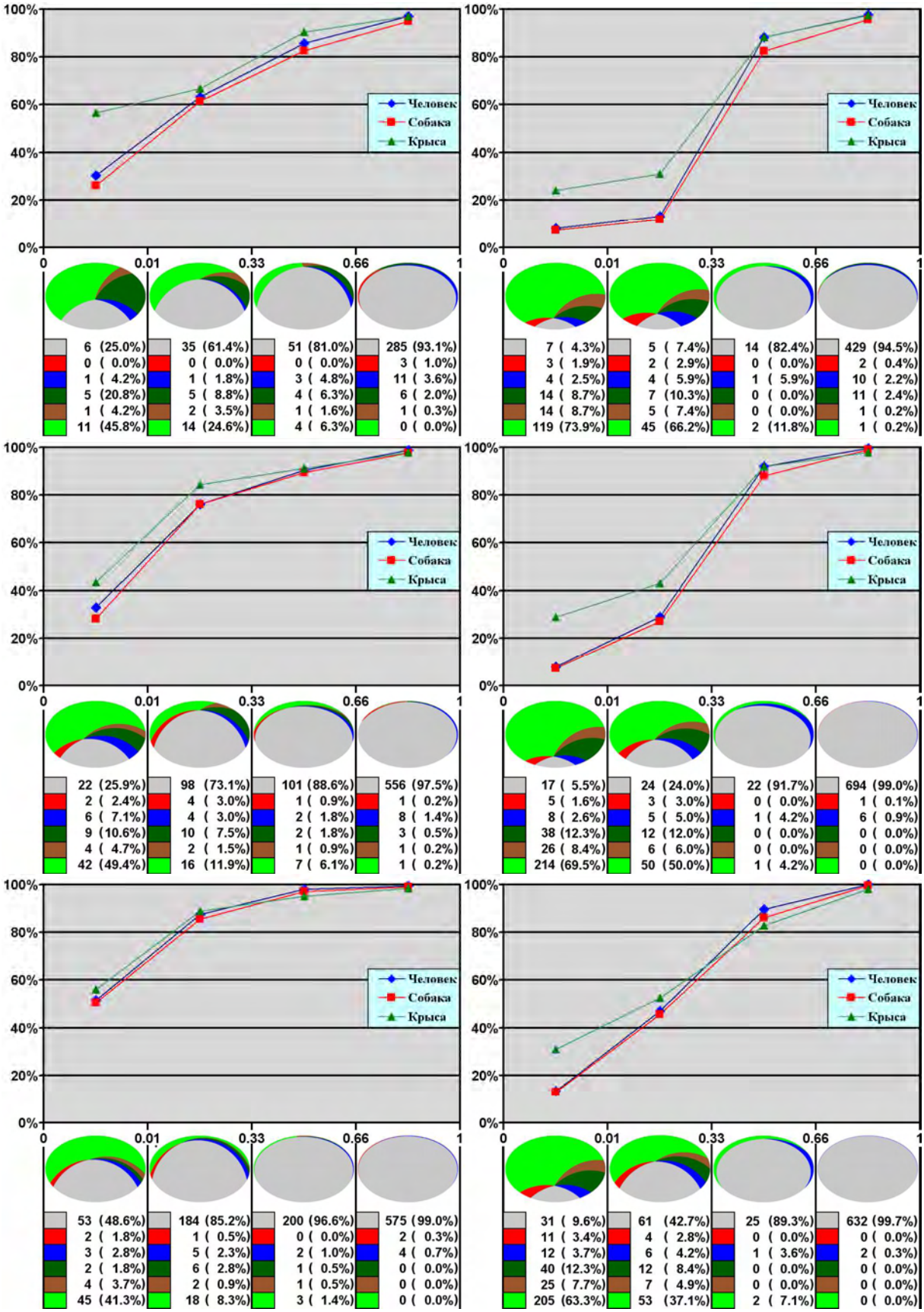


Рисунок 2. Консервативность сохраняющих (слева) и нарушающих (справа) рамку считывания кассетных экзонов мыши в генах человека, собаки и крысы для разных интервалов частоты включения кассетного экзона (продолжение на следующей странице).

Консервативность подсчитана отдельно для трех групп генов: вверху 2717 быстро эволюционирующих генов, имеющих уровень сходства менее 80%, 3939 гена с уровнем сходства от 80% до 92% посередине, 3447 крайне консервативных гена внизу, уровень сходства более 92%. На диаграммах серым цветом показана доля полностью консервативных альтернатив. Красным цветом показана доля альтернатив, консервативных только в генах собаки, а синим цветом - только в генах человека. Светло-зеленым цветом показаны неконсервативные кассетные экзоны. Экзоны мыши, консервативные только в генах крысы, разделены на две подгруппы, случайно консервативные экзоны - коричневые, а экзоны, появившиеся в ходе эволюции грызунов ("молодые экзоны") - темно-зеленые (определение см. в тексте).

Консервативность кассетных экзонов мыши в генах человека, собаки и крысы для разных интервалов частоты включения кассетного экзона в генах с разной скорость молекулярной эволюции белковой последовательности показана на Рисунке 2. Консервативность кассетных экзонов повышается с ростом частоты включения экзона, консервативность нарушающих рамку считывания кассетных экзонов в целом меньше чем консервативность сохраняющих рамку считывания. Однако разница в консервативности между сохраняющими и нарушающими рамку считывания пропусками экзонов крайне мала.

Доля "молодых экзонов", кассетных экзонов мыши, консервативных только в геноме крысы, падает с ростом частоты использования кассетного экзона. Это подтверждает гипотезу о том, что новые экзоны первоначально появляются в результате "экзонизации" части интрона и имеют крайне низкую частоту включения в мРНК, и постепенно наращивают частоту включения в ходе эволюции.

Существует положительная корреляция между скоростью молекулярной эволюции генов (скоростью накопления и фиксации в геномной последовательности мутаций) и скоростью эволюции альтернативного сплайсинга (консервативностью экзонов). Кассетные экзоны в быстро эволюционирующих генах существенно менее консервативны, чем кассетные экзоны в медленно эволюционирующих генах. Кроме того, в быстро эволюционирующих генах существенно выше доля "молодых экзонов", появившихся в ходе эволюции грызунов.

3. Литература

[1] Mironov AA, Fickett JW, Gelfand MS: "Frequent alternative splicing of human genes". *Genome Res*. 1999, 9: 1288-1293.

[2] Brett D, Pospisil H, Valcarcel J, Reich J, Bork P: "Alternative splicing and genome complexity". *Nat Genet*. 2002, 30: 29-30.

[3] Zavolan M, van Nimwegen E, Gaasterland T: "Splice variation in mouse full-length cDNAs identified by mapping to the mouse genome". *Genome Res*. 2002, 12: 1377-1385.

[4] Duret L, Mouchiroud D, Gouy M: "HOVERGEN: a database of homologous vertebrate genes". *Nucleic Acids Res.* 1994, 22(12): 2360-2365

[5] Nurtdinov RN, Artamonova II, Mironov AA, Gelfand MS: "Low conservation of alternative splicing patterns in the human and mouse genomes". *Hum Mol Genet*. 2003, 12: 1313-1320.

[6] Modrek B, Lee CJ: "Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss". *Nat Genet.* 2003, 34: 177-180.

[7] Thanaraj TA, Clark F, Muilu J: "Conservation of human alternative splice events in mouse". *Nucleic Acids Res.* 2003, 31: 2544-2552.

[8] Pan Q et al: "Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform". *Mol Cell* 2004, 16: 929-941.

[9] Wang W et al: "Origin and evolution of new exons in rodents". *Genome Res.* 2005, 9:1258-1264.

[10] Zhang XH, Chasin LA: "Comparison of multiple vertebrate genomes reveals the birth and evolution of human exons". *Proc Natl Acad Sci U S A* 2006, 36: 13427-1332.

[11] Alekseyenko AV, Kim N, Lee CJ: "Global analysis of exon creation versus loss and the role of alternative splicing in 17 vertebrate genomes". *RNA* 2007, 5: 661-670.

[12] Nurtdinov RN, Neverov AD, Favorov AV, Mironov AA, Gelfand MS: "Conserved and species-specific alternative splicing in mammalian genomes". *BMC Evol Biol*., 2007, 7: 249.

[13] Wheeler DL at al: "Database resources of the National Center for Biotechnology". *Nucleic Acids Res*., 2003, 31(1): D28-33.

[14] Jagadeeshan S, Singh RS: "Rapidly evolving genes of Drosophila: differing levels of selective pressure in testis, ovary, and head tissues between sibling species". *Mol Biol Evol*. 2005, 9: 1793-801.

[15] Haerty W at al: "Evolution in the fast lane: rapidly evolving sex-related genes in Drosophila". *Genetics*, 2007, 3:1321-1335.

[16] Torgerson DG, Kulathinal RJ, Singh RS: "Mammalian sperm proteins are rapidly evolving: evidence of positive selection in functionally diverse genes". *Mol Biol Evol*., 2002, 11: 1973-1980.

[17] Kan Z, States D, Gish W: "Selecting for functional alternative splices in ESTs". *Genome Res*., 2002, 12(12): 1837-1845.

[18] Neverov AD, Artamonova II, Nurtdinov RN, Frishman D, Gelfand MS and Mironov AA: "Alternative splicing and protein function". *BMC Bioinformatics*. 2005, 6: 266.

[19] Kent WJ: BLAT--the BLAST-like alignment tool. *Genome Res.* 2002, 12(4): 656-64.

Genome-wide assessment of codon usage correlation

Vinogradov D.V., Mironov A.A.
IITP RAS, Moscow, Russia; FBB MSU, Moscow, Russia
dimavin@bioinf.fbb.msu.ru

Abstract

Unequal usage of synonymous codons has been proved to impact the protein folding or the gene product expression rate significantly in some cases. Under the basic assumptions, functionally important properties should be conserved in orthologous groups. The goal of this research is to check whether the codon usage is conserved during the evolutionary process.

Average correlation coefficient was calculated for all protein families from the PFAM database and the Monte-Carlo method was used to evaluate the background distribution. The results suggest that the obvserved CC value can not be explained by stochastic reasons only, hence the codon usage should be under evolutionary pressure.

1. Introduction

Unequal usage of synonymous codons has been proved to impact the protein folding or the gene product expression rate significantly in some cases (e.g. [1], [2]). The common belief is that such cases are widespread, though no genome-wide analysis of codon usage conservation was published. Under the basic assumptions, functionally important properties should be conserved in orthologous groups. The goal of this research is to check whether the codon usage is conserved during the evolutionary process.

2. Material and methods

A set of manually curated multiple protein alignments was taken from the PFAM database (PFAM-A), then corresponding DNA alignments were built using the UniprotDB and EMBLCDS databases. Relative frequencies of codons for all species were calculated using CUTG database, and each line of every alignment was converted to a series of these frequencies. The averaged correlation coefficient for these series was used as a conservation measure. The probability to observe such correlation coefficient was calculated by a Monte-Carlo simulation of a codon selection among possible synonyms for a given amino acid.

3. Results

We found that the CC distribution of our data set is shifted significantly compared to the generated data set. This can be explained by the presence of evolutionary pressure towards conservation of codon frequency.

4. References

[1] C. Kimchi-Safraty et al. “A silent polymorphism in the MDR1 gene changes substrate specificity”, *Science*, 2007, **315**: 525-528

[2] Y.M. Zalucki, M.P. Jennings, “Experimental confirmation of a key role for non-optimal codons in protein export”, *BBRC*, 2007, **355**: 143-148

Кислород-зависимая регуляция азотфиксации: исследование методами сравнительной геномики

Е.Ю. Климова, Д.А. Равчеев
Московский Государственный Университет им. М.В. Ломоносова, Институт проблем передачи информации им. А.А. Харкевича РАН
ravcheyev@iitp.ru

Аннотация

Азотфиксация представляет собой процесс превращения молекулярного азота в ионы аммония, которые впоследствии включаются в органические соединения клети. Данный процесс достаточно широко таксономически распространен среди бактерий и архей, однако до сих пор его изучение было ограничено лишь несколькими модельными организмами. Наиболее хорошо азотфиксация исследована для бактерий порядка Rhizobiales. Центральный фермент азотфиксации, нитрогеназа, крайне чувствителен к присутствию молекулярного кислорода. По этой причине существует система регуляции азотфиксации, сигналом для которой является наличие молекулярного кислорода. В частности, ответ на присутствие кислорода на уровне экспрессии генов осуществляется гомологичными факторами транскрипции FnrN и FixK. В настоящей работе FnrN- и FixK-зависимая регуляция генов азотфиксации в Rhizobiales была исследована методами сравнительной геномики.

1. Введение

Азотфиксация представляет собой процесс превращения молекулярного азота в ионы аммония, которые впоследствии включаются в органические соединения клети. Как показали исследования, к азотфиксации способны различные группы бактерий и архей, причем таксономическое распространение организмов-азотфиксаторов крайне мозаично. Наиболее хорошо процесс азотфиксации изучен для бактерий порядка Rhizobiales из класса Alphaproteobacteria, являющихся симбионтами растений. К сожалению, экспериментальные исследования данного процесса проводились лишь для небольшого числа модельных организмов.

Центральным компонентом системы азотфиксации является фермент нитрогеназа, обеспечивающий собственно процесс превращения молекулярного кислорода в ионы аммония. Кроме того, известно большое число вспомогательных генов, требующихся для нормального протекания работы нитрогеназы. Исторически сложилось, что все гены азотфиксации делят на три группы: *nif*-, *nod*- и *fix*-гены [1].

Азотфиксация – сложный процесс, зависящий от множества факторов, таких как концентрация молекулярного азота, концентрация молекулярного кислорода, соотношение окисленных и восстановленных форм восстановительных эквивалентов и наличие энергетических эквивалентов. Понятно, что ответ на такое большое количество факторов требует достаточно сложной регуляторной системы. Так, в бактериях порядка Rhizobiales ответ на концентрацию азота осуществляется альтернативным сигма-фактором RpoN совместно с рядом вспомогательных белков, таких как NifA, NtrC и NtrX. Реакция же на опасный для нитрогеназы кислород осуществляется за счет факторов транскрипции FnrN и FixK и двукомпонентной системы FixL-FixJ [2].

В настоящей работе представлены результаты исследования FixK- и FnrN-зависимой регуляции в бактериях порядка Rhizobiales.

2. Данные и методы

2.1. Процедура определения генов, входящих в обобщенный регулон

Для предсказания структуры регулонов был использован сравнительно-геномный подход. [3] Так, ген считался принадлежащим к регулону, если в его предполагаемой регуляторной области были предсказаны потенциальные сайты связывания соответствующего фактора

транскрипции, и такие сайты были найдены в нескольких геномах.

Для поиска потенциальных сайтов связывания факторов транскрипции был использован метод матриц (профилей) позиционных весов [3].

В случае, когда оперонная струкрура локуса не была исследована, гены считались принадлежащими к одному оперону, если они имели одинаковое направление считывания, расстояние между ними не превышало 200 п.н. и структура локуса сохранялась в нескольких родственных геномах.

2.2. Программное обеспечение

Для поиска ортологичных генов и потенциальных сайтов связывания был использован пакет программ GenomeExplorer [4]. Для построения матриц (профилей) позиционных весов была использована программа SignalX [4]. Поиск гомологов в базах данных проводился с помощью программы BLAST [5]. Для построения множественных выравниваний аминокислотных и нуклеотидных последовательностей, а также для построения филогенетических деревьев была использована программа ClustalX [6]. Для построения диаграмм Лого, отображающих структуру регуляторного сигнала, использовалась программа WebLogo [7].

2.3. Исследованные геномы

В настоящей работе были исследованы полные геномные последовательности девяти бактерий из порядка Rhizobiales: *Bradyrhizobium japonicum*, *Bradyrhizobium* sp. BTAi1, *Sinorhizobium meliloti*, *Nitrobacter winogradskyi*, *N. hamburgensis*, *Rhodopseudomonas palustris*, *Rhizobium etli*, *R. leguminosarum* и *Agrobacterium tumefaciens*. Все последовательности были взяты из базы данных GenBank [8].

3. Результаты и обсуждение

3.1. Эволюция регуляторных белков

Поиск гомологов показал, что во всех исследованных геномах присутствует как минимум одна копия генов, кодирующих FnrN или FixK. В данном случае сложно говорить об ортологичности, поскольку полученные результаты свидетельствуют о том, что для генов *fnrN* и *fixK* многокопийность является скорее правилом, чем исключением. В сложившейся ситуации, когда в клетке имеется несколько регуляторов с высокой степенью схожести, мы не можем приписать отдельные типы сайтов к отдельным регуляторам и вынуждены искать сайты связывания для всех регуляторов, поскольку, по всей видимости, несколько регуляторов будут связываться с одинаковым сайтом. Также белки FnrN и FixK обнаруживают высокую степень сходства и, судя по всему, связываются с одинаковыми сайтами [9].

3.2. Эволюция регуляторных белков

На основании экспериментально известных сайтов, последовательности которых доступны по литературным данным, нами была построена матрица позиционных весов для поиска сайтов связывания белков FnrN и FixK. Общий сайт связвания данных белков представляет собой палиндром длиной 14 пар нуклеотидов и имеет консенсус TTGATC-2-GATCAA.

3.3. Исследование FnrN/FixK-регулонов

Путем сравнительно-геномного анализа во всех исследованных геномах была прослежена регуляция известных генов, участвующих в азотфиксации.

Несмотря на необычайный полиморфизм в плане копийности генов, FnrN/FixK-зависимая регуляция представляется крайне устойчивой. Так, в тех случаях, когда в геноме присутствует лишь одна копия оперона, перед ним всегда удается обнаружить достоверный потенциальный сайт. В случаях, когда наблюдаются дупликации оперонов, как минимум одна из копий является регулируемой. В качестве примеров можно привести устойчивую регуляцию оперонов *fixNOQP* и *fixGHIS*, для которых характерна многокопийность. Данные опероны кодируют белки, необходимые для образования кислород-чувствительных железо-серных кластеров.

4. Благодарности

Данная работа выполнена совместно с М.С. Гельфандом. Авторы благодарны А.А. Миронову за любезно предоставленное программное обеспечение.

Работа была частично поддержана Медицинским институтом Ховарда Хьюза, программой “Молекулярная и клеточная биология” Российской Академии Наук и INTAS.

5. Список литературы

[1] H.M. Fischer, “Genetic regulation of nitrogen fixation in rhizobia”, *Microbiol. Rev.*, American Society For Microbiology, United States, 1994, pp. 352-386.

[2] R. Dixon and D. Kahn, “Genetic regulation of biological nitrogen fixation”, *Nat. Rev. Microbial.*, Nature Publishing Group, England, 2004, pp. 621-632.

[3] A.A. Mironov, E.V. Koonin, M.A. Roytberg and M.S. Gelfand “Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes” *Nucleic Acids Res*., Oxford University Press, England, 1999, pp. 2981-2989.

[4] А.А. Миронов, Н.П. Винокурова и М.С. Гельфанд, “Программное обеспечение анализа бактериальных геномов”, *Мол. биология*, “Наука”, Москва, Российская Федерация, стр. 253-262.

[5] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, M.W. Zhang and D.J. Lipman “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs”, *Nucleic Acid Res.*, Oxford University Press, England, 1997, pp. 3389-3402.

[6] J.D. Thompson, T.J. Gibson, F. Plewniak, F. Jeanmougin and D.G.Higgins, “The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools”, *Nucleic Acid Res.,* Oxford University Press, England, 1997, pp. 4876-4882.

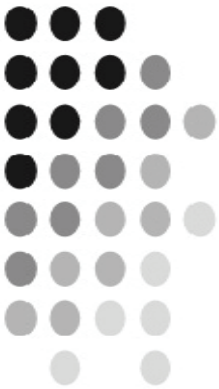
[7] T.D. Schneider and R.M. Stephens, “Sequence logos: a new way to display consensus sequences”, *Nucleic Acid Res.*, Oxford University Press, England, 1990, pp. 6097-6100.

[8] D.A. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell and D.L. Wheeler, “GenBank”, *Nucleic Acids Res.*, Oxford University Press, England, 2007, pp. 21-25.

[9] H. Korner, H.Z. Sofia and W.G. Zumft, (2003) Phylogeny of the bacterial superfamily of Crp-Fnr transcription regulators: exploiting the metabolic spectrum by controlling alternative gene programs. *FEMS Microbiol. Rev.*, Blackwell Pub., England, 2003, pp. 559-592.

Раздел 7

Математическая теория информации и управления, многокомпонентные случайные системы



Моделирование вычислительных систем непрерывными уравнениями

Захаров Владислав
Московский физико-технический институт (государственный университет)
vzaxarov@gmail.com

Тормасов Александр
Московский физико-технический институт (государственный университет)
tor@crec.mipt.ru

Аннотация

В работе формулируется и обосновывается математическая модель потребления ресурсов операционной системы, используя систему из дифференциальных и алгебраических уравнений. Модель может быть использована для исследования как процесса микро планирования на уровне планировщика ОС, так и на уровне макро планирования общего прохождения потока процессов через компьютерную систему.

1. Введение

Существующие в настоящее время модели вычислительных систем, в основном, предназначены для моделирования «физического» компьютера как совокупности процессора, памяти, шины обмена данными и так далее. Другой класс моделей используется для реализации разнообразных планировщиков ресурсов ОС. Но моделей, описывающих переход ресурсов операционной системы от одного потребителя к другому практически нет.

Модели такого типа дадут возможность описать достаточно сложные вычислительные системы различной степени распределённости.

2. Ресурсы операционной системы

Операционная система, работающая на компьютере, обладает определенным набором ресурсов, необходимых для запуска и работы процессов и нитей [1, 2].

К таким ресурсам можно отнести всевозможные экземпляры структур данных ядра ОС, например, открытые файлы, семафоры и т п. Особенностью этих ресурсов является то, что каждый процесс или нить запрашивает создание ресурса путем обращения к ядру ОС.

Другим классом ресурсов ОС являются «распределяемые» ресурсы, те, которые «всегда есть». Примерами таких ресурсов являются полосы пропускания сети или дисковой подсистемы, или доля CPU. В любой момент времени, когда работает планирование таких ресурсов, их не надо «отбирать» у текущего обладателя, надо просто перераспределить между потребителями [3].

3. Уравнения потребления «распределяемых» ресурсов

Рассмотрим вычислительную систему с одним потоком исполнения. У этого потока есть программа, состоящая из списка машинных комманд. В каждый момент времени поток может исполнять какую-то одну команду. Номер исполняемой команды в программе определяет внутреннее состояние потока. Будем моделировать это состояние непрерывной переменной x. Пусть переменная x изменяется от 0 до некоторого значения X. Количество ресурса определённого типа, полученное потоком в момент времени t в состоянии x, определяется функцией плотности ресурса $\rho(x,t)$. Скорость переноса ресурса будем обозначать как $u(x,t)$. Опишем закон, по которому будет изменяться количество ресурса, полученное одним потоком, в элементарном объёме от x до $x+dx$.

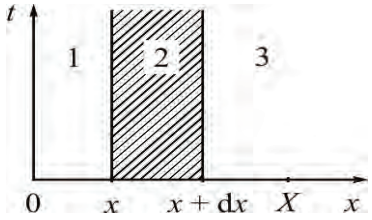


Рис. 1. Изображён элементарный объём в области определения $[0;X]\times[0;\infty)$ функции плотности ресурса

За малый промежуток времени dt из области 1 (см рис. 1) в область 2 втекает количество ресурса

$$\rho(x,t)u(x,t)dt\,, \qquad (1)$$

а в область 3 вытекает

$$\rho(x+dx,t)u(x+dx,t)dt\,. \qquad (2)$$

Количество ресурса, выделяемое потоку, контролируется планировщиком. Будем считать, что планировщик способен изменять количество ресурса, выделяемое одному потоку, со скоростью $f(x,t)$. Количество внесённого планировщиком ресурса за время dt будет равно

$$f(x,t)dtdx. \quad (3)$$

В итоге в области 2 скапливается количество ресурса, равное

$$\frac{\partial \rho}{\partial t}dtdx. \quad (4)$$

Теперь можно записать закон сохранения количества ресурса:

$$\rho(x,t)u(x,t)dt + f(x,t)dtdx = \\ = \frac{\partial \rho(x,t)}{\partial t}dtdx + \rho(x+dx,t)u(x+dx,t)dt. \quad (5)$$

Поделим обе части уравнения (5) на $dtdx$ и перенесём слагаемое $f(x,t)$ в правую часть, а остальные в левую:

$$\frac{\partial \rho(x,t)}{\partial t} + \frac{u(x+dx,t)\cdot\rho(x+dx,t) - u(x,t)\cdot\rho(x,t)}{dx} = \\ = f(x,t) \quad (6)$$

Или в традиционной форме записи:

$$\frac{\partial \rho(x,t)}{\partial t} + \frac{\partial (u\cdot\rho(x,t))}{\partial x} = f(x,t). \quad (7)$$

Система уравнений для совокупности распределяемых ресурсов, потребляемых одним потоком, будет иметь вид

$$\frac{\partial \rho_i(x,t)}{\partial t} + \frac{\partial (u\cdot\rho_i(x,t))}{\partial x} = f_i(x,t),\ i\in G. \quad (8)$$

Эти уравнения характеризуют то, как поток стремится потреблять ресурсы.

Если в вычислительной системе исполняется *n* потоков, то для $j^{го}$ потока можно написать систему уравнений

$$\frac{\partial \rho_i^j(x,t)}{\partial t} + \frac{\partial (u^j\cdot\rho_i^j(x,t))}{\partial x^j} = f_i^j(x,t), \\ i\in G,\ j=1...n, \quad (9)$$

индекс j здесь обозначает номер потока в системе. Конкретный вид правой части уравнения (9) приводится ниже. Область определения функций плотностей ресурсов $\rho_i^j(x,t)$ – это множество $D = [0;1]\times[0,\infty)$.

В частном случае, когда поток потребляет только процессорное время, скорость изменения внутреннего состояния пропорциональна плотности ресурса

$$u = k\rho,$$

и уравнение (9) превращается в уравнение Хопфа [5]:

$$\frac{\partial \rho(x,t)}{\partial t} + 2k\rho\frac{\partial \rho(x,t)}{\partial x} = f(x,t). \quad (10)$$

3. Моделирование планировщика ресурсов

Уравнения (9) являются уравнениями переноса. Решение таких уравнений постоянно вдоль характеристик. Ресурс, выделенный потоку в точке (x,t) переносится вдоль этих характеристик. Будем моделировать функцию $f(x,t)$ в виде

$$f(x,t) = \lambda(\rho^T(x,t) - \rho(x,t)) \quad (11)$$

Здесь λ – числовой параметр модели, этот параметр выбирается из условий устойчивости численного алгоритма, используемого для решения уравнений (8), $\rho(x,t)$ – текущее, а $\rho^T(x,t)$ – планируемое значение плотности ресурса. Если плотность ресурса равна планируемому значению, то она остаётся постоянной вдоль характеристики. В противном случае функция (11) линейно компенсирует возникающее расхождение ожидаемого и реального значения плотности ресурса.

В первом приближении величина планируемой плотности ресурса $\rho^T(x,t)$ – это функция внутренней потребности потока в ресурсе. Эта функция показывает, какое количество ресурса необходимо потоку, когда он достиг определённого внутреннего состояния при условии, что поток монопольно исполняется в вычислительной системе.

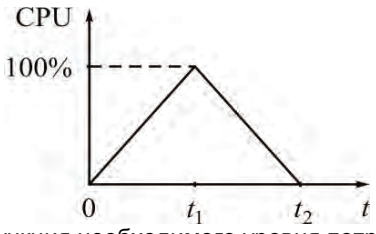


Рис. 2. Функция необходимого уровня потребления

Будем обозначать эту функцию как $S(x)$. Эта функция изменяется от 0 до 1, и в случае монопольного исполнения потока в системе она является ожидаемой долей ресурса. И уравнение (11) примет вид

$$f(x,t) = \lambda(\rho^{\max}S(x) - \rho(x,t)). \quad (12)$$

Здесь $\rho^{\max}$ – размерный коэфициент, характеризующий максимальное значение ресурса.

Теперь предположим, что в вычислительной системе исполняется несколько потоков, потребляющих общий ресурс. Планировщик

ресурсов вычислительной системы распределяет общий ресурс между потребителями. Будем полагать, что это распределение происходит на основе приоритетов потоков. Как это происходит для операционных систем семейства Windows, можно узнать в [6], а для UNIX систем информацию можно найти в [7]. Рассмотрим возможную процедуру определения ожидаемой доли ресурса. Допустим, планирование осуществляется на период времени t_0. Каждому потоку присвоен приоритет W_i. Согласно этому приоритету каждому потоку выделяется квант времени, в течение которого поток получает источник ресурса в монопольное владение. Он равен:

$$\tau_i = t_0 \frac{W_i}{\sum_{j \in l(t)} W_j}. \qquad (13)$$

Здесь $l(t)$ – индексы потоков, участвующих в планировании ресурса. Функция $l(t)$ зависит от времени, так как количество активных потоков может меняться.

Согласно функции внутренней потребности в ресурсе каждый поток использует ресурс лишь часть времени, отведённого ему (см рис. 3).

Рис. 3. Изображена диаграмма распределения процессорного времени между двумя потоками потребления

$$t_i = \tau_i S_i(x). \qquad (14)$$

Таким образом, максимальная ожидаемая доля ресурса $T_r^{\max}(W_i, S_i(x), t)$ может быть найдена как:

$$T_r^{\max}(W_i, S_i(x), t) = \frac{W_i S_i(x)}{\sum_{j \in l(t)} W_j S_j(x)}. \qquad (15)$$

Теперь рассмотрим ситуацию, в которой для функционирования потока необходим некоторый набор ресурсов *R*. Пусть для каждого ресурса *r* задана функция внутренней потребности в ресурсе $S_r(x)$.

Реальные функции потребности для всех ресурсов $S_r(x)$, очевидно, претерпят некоторые изменения, когда использование какого-либо ресурса r ограничено планировщиком. Рассмотрим поток, активно обрабатывающий информацию с жёсткого диска. В случае нехватки дисковой полосы пропускания этому потоку нечем будет «занять» процессорное время.

Положим на момент времени (*x*,*t*) планировщик ресурсов способен выделить потоку долю ресурса типа *r*, равную *M* единиц, эта доля определяется какой-либо политикой планирования. Если эта доля больше или равна $S_r(x)$, то ограничение никак не скажется. В другой ситуации, когда $M(t)$ меньше значения $S_r(x)$, доступный уровень ресурса изменится и будет равен *M*. Это отразится и на уровне потребления других ресурсов. Найдём эту зависимость.

За малый интервал времени поток ожидает получить объём ресурсов

$V_r(x,t) = Q_r \cdot \Delta T$, где

$$Q_r(x,t) = R_{\max} \cdot S_r(x,t) \cdot u(x,t). \qquad (16)$$

Относительный темп потребления ресурсов должен сохраняться, ибо поток может обрабатывать определённую долю, например, сетевой полосы пропускания, обладая только вполне определённой долей процессорного времени;

$$\frac{Q_r(x,t) \cdot \Delta T}{Q_r^*(t) \cdot \Delta T} = \frac{Q_n(x,t) \cdot \Delta T}{Q_n^*(t) \cdot \Delta T}. \qquad (17)$$

В (12) Q_r – требуемый потоком уровень потребления при монопольном выполнении в вычислительной системе, а Q_r^* – уровень при совместном исполнении нескольких потоков.

Формула (17) после подстановки (16) и преобразований

$$\frac{S_r(x,t)}{T_r^{\max}} = \frac{S_n(x,t)}{S_n^*(t)} \qquad (18)$$

Изменение доли одного ресурса, определяемой планировщиком, влияет на требуемый внутренний уровень всех других распределяемых ресурсов. Если теперь в планировании участвуют все виды ресурсов, то при определении необходимой доли для конкретного ресурса вида n следует учесть в уравнении (18) лишь тот ресурс *r* который даёт минимальное значение нового уровня потребления. Пусть S_n – требуемая доля при монопольном исполнении одного потока, а S_n^* – доля с учётом других потоков. Тогда условие минимизации относительных уровней потреблений имеет вид

$$\frac{S^*{}_i(x,t)}{S_i(x,t)} = \min_{r \in R} \frac{T_r^{\max}}{S_r(x,t)}. \qquad (19)$$

При $S_i = 0$ ресурс типа *i* для данного потока не принимает участия в планировании и не входит в набор *R*. То есть формула (14) всегда определена.

В итоге уравнения для перераспределяемых ресурсов можно записать так:

$$\frac{\partial \rho_i^j(x,t)}{\partial t} + \frac{\partial (u^j \cdot \rho_i^j(x,t))}{\partial x^j} =$$

$$= \lambda \cdot (\rho^{j\max}_i \cdot S^j{}_i(x,t) \cdot \min_{r \in R} \frac{T^{j\max}_r}{S^j{}_r(x,t)} - \rho_i^j(x,t)), \quad (20)$$

$$i = 1..m, j = 1..n.,$$

$$T^{j\max}_r(W_i, S^i(x), t) = \frac{W_i S^i(x)}{\sum_{k \in l(t)} W_k S^j(x)}. \quad (21)$$

Уравнения (20) и (21) записаны для n потоков в вычислительной системе. Эти потоки потребляют m типов ресурсов, i – индекс текущего ресурса.

4. Уравнения планирования процессорного времени

Вторая группа уравнений описывает ограничения, накладываемые планировщиком CPU. Эти уравнения можно записать в виде

$$\frac{du^j(x,t)}{dt} = \mu \cdot (\rho^{\max}_{cpu} \cdot T^j{}_{\text{cpu}}(x,t) - \text{u}^j) \cdot \quad (22)$$

Здесь $\rho^{\max}_{cpu}$ максимальная производительность процессора, $\text{T}^j_{\text{cpu}}(x,t)$ – ожидаемый уровень потребления процессорного времени для $j^{го}$ потока, μ – числовой параметр модели, который выбирается из условий устойчивости численного алгоритма, используемого для решения уравнений (22). В уравнении (22):

$\text{T}^j{}_{\text{cpu}}(x,t) = \text{S}^{*j}{}_{cpu}(x,t)$, где $\text{S}^{*j}{}_{cpu}(x,t)$ даётся соотношением (19).

Смысл уравнения (22) состоит в том, что компенсация завышенной или заниженной скорости распространения ресурса осуществляется пропорционально разности желаемого и полученного ресурса. На данном этапе модель учитывает только один ресурс, влияющий на скорость распространения ресурса. Это доля процессорного времени.

5. Модельная система уравнений

В итоге для модели вычислительной системы с n потоками и m типами ресурсов получаем следующую систему уравнений:

$$\begin{cases} \dfrac{\partial \rho_i^j(x,t)}{\partial t} + \dfrac{\partial (u^j \cdot \rho_i^j(x,t))}{\partial x^j} = \\ = \lambda \cdot (\rho^{j\max}_i \cdot S^j{}_i(x) \cdot \min\limits_{r \in R} \dfrac{T^{j\max}_r}{S^j{}_r(x)} - \rho_i^j(x,t)), \\ \dfrac{du^j(x,t)}{dt} = \\ = \mu \cdot (\rho^{j\max}_{\text{cpu}} \cdot S^j{}_{cpu}(x) \cdot \min\limits_{r \in R} \dfrac{T^{j\max}_r}{S^j{}_r(x)} - \text{u}^j(x,\text{t})), \\ i = 1...m, j = 1...n. \end{cases} \quad (23)$$

В системе $T^{j\max}_r(W_i, S^i(x), t) = \dfrac{W_i S^i(x)}{\sum_{k \in l(t)} W_k S^j(x)}$.

$l(t)$ – индексы активных потоков. Так как поток стартует, не обладая ресурсом, логично дополнить систему (23) условиями

$$\rho_i^j(0, \text{t}) = \rho_i^j(x,0) = 0,$$
$$\text{u}^j(x,0) = 1, \; i = 1...m, \; j = 1...n.$$

6. Численное решение модельной системы уравнений

Часть системы уравнений (23) является гиперболической, и для их решения предложено использовать сеточно-характеристический метод [4, 8-10].

Модель была использована для моделирования нагрузки на ftp-сервер (freeFtp) и Apache http-сервер. Для ftp-сервера были учтены такие ресурсы как: процессорное время, доля сетевой и дисковой полосы пропускания. Для http-сервера только процессорное время и доля сетевой полосы пропускания. Полученные численные результаты достаточно точно совпадают с практическими измерениями (см рис. 4-7).

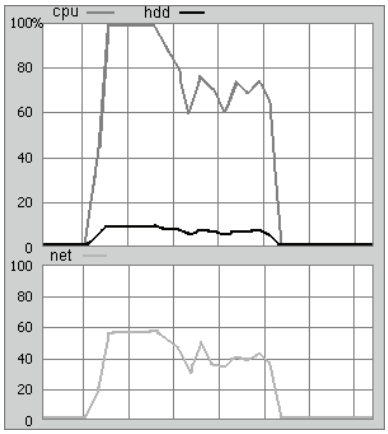


Рис 4 Типичная нагрузка на ftp-сервер при запросе от Core FTP LE – Windows

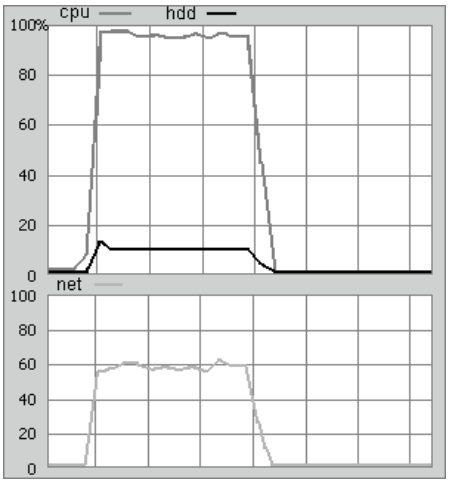


Рис 5 Типичная нагрузка на ftp-сервер при запросе от Konqueror – Linux

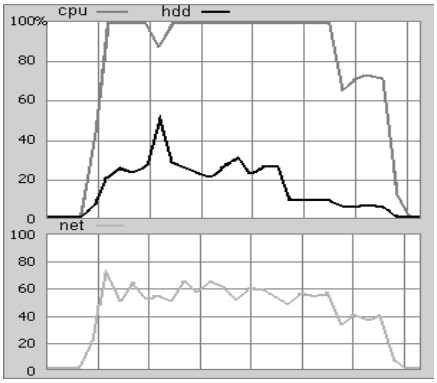


Рис. 6. Типичная нагрузка на ftp-сервер при запросе от Konqueror и CORE FTP LE, полученная экспериментально

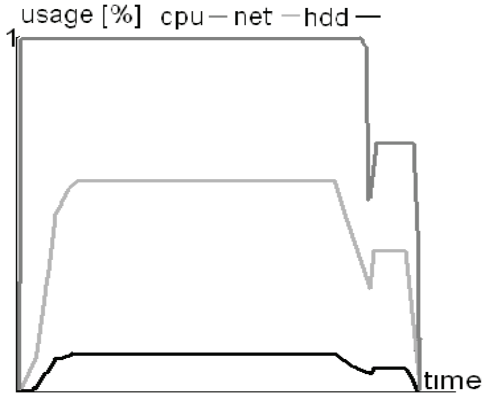


Рис. 6. Расчёт нагрузки на ftp-сервер при запросе от Konqueror и CORE FTP LE

7. Заключение

В ходе работы был реализован новый подход к моделированию реальных вычислительных систем. Он позволяет построить полную замкнутую модель вычислительной системы, для «распределяемых» типов ресурсов. Модель даёт возможность описать достаточно сложные вычислительные системы различной степени распределённости.

Для численного решения задач моделирования можно использовать сеточно-характеристический метод для гиперболических систем уравнений.

[1] Э. Таненбаум *Современные операционные системы* 2-е изд., Питер, Санкт-Петербург, 2002.

[2] Х.М.Дейтел, П. Дж. Дейтел, Д. Р. Чорнес, *Операционные системы. Основы и принципы*, Бином-пресс, Москва, 2006.

[3] А.Г. Тормасов, *"Модель потребления ресурсов вычислительной системой"*, Вестник НГУ. Сер. Информационные технологии, Т. 4, вып. 1, 2006.

[4] К.М. Магомедов, А.С. Холодов, *Сеточно-характеристические численные методы*, Наука, Моосква, 1988.

[5] А.А. Самарский, П.Н. Вабищевич, Г.П. Матус, *Разностные схемы с операторными множителями*. Минск, 1998.

[6] Д. Соломон, М. Руссинович, *Внутреннее устройство Microsoft Windows 2000*. Мастер-класс / Пер.с англ, Питер, Санкт-Петербург

[7] А.М. Робачевский, *Операционная система UNIX*. %, БХВ-Петербург, Санкт-Петербург , 2002.

[8] A.S. Kholodov, Y.A. Kholodov. *Computational models on graphs for nonlinear hyperbolic and parabolic system of equations*, 2-M.I.T. Conference on Computational Fluid and Solid Mechanics, June, 2003, Cambridge, MA, U.S.A.

[9] A. S. Kholodov, Y.A. Kholodov, Computational models on graphs for the nonlinear hyperbolic system of equations. ASME 2004 PVP, Vol. 476, PVP 2004-2580, pages 161-167.

[10] А.С.Холодов, Я.А.Холодов, Н.В.Ковшов. Сетевые вычислительные модели для нелинейных систем уравнений гиперболического типа. 2006, опубликовано на сайте http://swsoft.mipt.ru/nkovshov/SVM.doc

Моделирование стационарных входных потоков данных для стохастических моделей дискретных систем

Анна Шляева
МГТУ им. Н.Э. Баумана
shlyaeva@gmail.com

Аннотация

В работе рассмотрена задача автоматизации моделирования случайных входных данных для стохастических моделей дискретных систем. В общем случае входное воздействие представляет собой реализацию случайного процесса. Выполнен анализ существующих моделей стационарного случайного процесса: авторегрессионные процессы и процессы скользящего среднего, гамма-процессы, TES (Transform-Expand Sample)-процессы, ARTA (AutoRegressive-To-Anything)-процессы. Для формализации случайного входного процесса выбрана модель ARTA-процесса. Описаны методы оценки параметров и генерации реализации стационарного ARTA-процесса, методы оценки адекватности построенной модели входных данных и их реализация.

1. Введение

При создании модели сложной системы возникает задача формализации случайных входных данных, важность которой зачастую недооценивается, и основное внимание уделяется созданию модели функционирования сложной системы, при этом в ней учитываются особенности системы, практически не влияющие на интересующие пользователя выходные данные. Однако, если случайные величины, влияющие на поведение системы, заданы неверно, то даже при создании идеальной модели поведения системы получают в корне неверные результаты моделирования, что приводит к неверному принятию решений. Одной из актуальных задач в области имитационного моделирования сложных систем является автоматизация процесса моделирования случайных входных воздействий.

В общем виде входное случайное воздействие представляет собой реализацию случайного, дискретного во времени, процесса. Случайный процесс Y_t можно представить как сумму детерминированной последовательности (систематической составляющей) $f(t)$ и случайной последовательности U_t, подчиняющейся некоторому вероятностному закону [1]:

$$Y_t = f(t) + U_t.$$

Выделяют две возможные составляющие $f(t)$: тренд, или систематическое движение (медленно меняющиеся функции времени, примером которых могут быть полиномы низкой степени) $g(t)$, и сезонные изменения $h(t)$ (циклические последовательности, например, отрезки ряда Фурье) [1, 3].

Одним из частных случаев случайного процесса является стационарный случайный процесс – процесс, вероятностные характеристики которого не меняются со временем [2]. Стационарным процессом в широком смысле называют случайный процесс, математическое ожидание которого постоянно, а функция корреляции зависит только от сдвига между аргументами [2].

2. Модели стационарного случайного процесса

Рассмотрим методы моделирования входных данных в случае, когда в имеющейся выборке входных данных имеется зависимость (автокорреляция) между соседними наблюдениями:

- Игнорирование зависимости между входными данными;
- Использование линейных моделей;
- Использование случайных процессов в предположении известности безусловного распределения вероятностей;
- TES(Transform-Expand-Sample) процессы;
- ARTA (Autoregressive-To-Anything) процессы.

Игнорирование зависимости в реализации случайного процесса приводит к недостоверным результатам моделирования. Так, в работе Livny [8] показано, что пренебрежение автокоррелированностью входных данных при моделировании

систем массового обслуживания может привести к катастрофической ошибке в ходе моделирования (недооценка средней длины очереди в 3 раза и более). В то же время зачастую при моделировании входные данные являются автокоррелированными [10, 13] (в частности, потоки заявок при моделировании компьютерных сетей [13]).

Большинство разработанных моделей стационарного случайного процесса – линейные модели, такие как авторегрессионные процессы и процессы скользящего среднего. В работе [9] показано, что линейность этих моделей подразумевает Гауссовы безусловные распределения, однако реальное безусловное распределение рассматриваемого случайного процесса может быть не Гауссовым. В связи с этим были разработаны модели случайных процессов для других безусловных распределений: экспоненциального, гамма, геометрического и других дискретных распределений [7].

Альтернативный подход к моделированию стационарного случайного процесса – конструирование требуемого процесса как некоторого преобразования Гауссова линейного процесса. Примером таких моделей являются ARTA- и TES-процессы [5, 10].

TES-процессы (Transform-Expand Sample, процессы преобразования и расширения выборки) [4, 11, 12] позволяют согласовать безусловное распределение подобранного процесса с эмпирическим безусловным распределением наблюдений, а также приблизительно подобрать структуру автокорреляции (гарантируется соответствие автокорреляции с шагом 1). Последовательность независимых и одинаково распределенных величин $U(0,1)$ преобразуется в процесс, где безусловным распределением остается равномерное на отрезке $[0;1]$, но проявляется автокорреляция. Далее полученный процесс преобразуется в другой процесс с искомым безусловным распределением. Для определения TES-процессов с помощью программного обеспечения TEStool [4] требуется выполнять вручную некоторые интерактивные настройки до тех пор, пока представление данных не будет признано адекватным.

В работе Cario, Nelson [6] было предложено использовать для построения модели автокоррелированной случайной величины ARTA-процесс, который представляет собой стационарный случайный процесс с произвольным безусловным распределением вероятности F_X и структурой автокорреляции с конечным шагом p.

При моделировании ARTA-процесса сначала конструируется базовый процесс – нормированный Гауссов случайный процесс

$$Z_t = \sum_{h=1}^{p} \alpha_h Z_{t-h} + Y_t, t = 1,2,...,n \quad (1),$$

где $\alpha_h, h = 1,2,..., p$ – коэффициенты авторегрессии, Y_t – белый шум, представляющий случайную часть Z_t, не зависящую линейно от предыдущих наблюдений, причем $E[Y_t] = 0$ и $Var[Y_t] = \sigma_Y^2$,

из которого преобразованием $U_t = \Phi(Z_t)$, где Φ – функция нормального распределения, получают последовательность $\{U_t : t = 1,2,...,n\}$ автокоррелированных случайных величин, в основе которых лежит равномерное распределение.

Затем к полученной последовательности применяют обратную функцию безусловного распределения

$$X_t = F_X^{-1}(U_t) = F_X^{-1}(\Phi[Z_t]) \quad (2)$$

и получают реализацию требуемого случайного процесса $\{X_t : t = 1,2,...,n\}$.

Данный подход работает для всех безусловных распределений, однако для некоторых распределений обратная функция распределения F_X^{-1} не может быть получена аналитически и должна быть рассчитана приближенным численным методом.

Из рассмотренных методов моделирования стационарного случайного процесса наиболее универсальными являются ARTA- и TES-процессы, однако применение TES-процессов требует ручной настройки, в то время как моделирование ARTA-процессов может быть полностью автоматизировано. В то же время TES-процессы гарантируют соответствие структуры автокорреляции только с шагом, равным единице. Поэтому в результате анализа моделей стационарного случайного процесса для формализации выбраны ARTA-процессы (AutoRegressive-To-Anything) с безусловным распределением из семейства распределений Джонсона.

3. Оценка параметров стационарного ARTA-процесса

Для оценки параметров ARTA-процесса используется алгоритм, предложенный Biller и Nelson.

Алгоритм заключается в минимизации целевой функции

$$S_D(\psi)=\frac{1}{(n-p)^2}\sum_{t=p+1}^{n}\frac{(n-p+1)^2(n-p+2)}{(t-p)(n+1-t)}\times$$

$$\times\left(\Phi\left[V_{(t)}(\psi)\right]-\frac{t-p}{n-p+1}\right)^2 \quad (3)$$

где $V_{(t)}(\psi)$ – порядковые статистики

$$V_t(\psi)=\frac{\gamma+\delta f\left[\frac{X_t-\xi}{\lambda}\right]}{g(p,\alpha)}-$$

$$-\frac{\sum_{h=1}^{p}\alpha_h\left(\gamma+\delta f\left[\frac{X_{t-h}-\xi}{\lambda}\right]\right)}{g(p,\alpha)} \quad (4),$$

$$g(p,\vec{\alpha})\equiv\sigma_Y^2.$$

Минимизация проводится в области $\Psi=\left(\gamma,\delta,\lambda,\xi,\alpha_1,\alpha_2,...,\alpha_p\right)'$, где

$$\begin{cases}\delta>0,\ f\in\{S_U,S_B,S_L,S_N\},\\ \delta<\infty,\ f\in\{S_U,S_B\},\end{cases} \quad (5)$$

$$\begin{cases}\lambda>0,\ f=S_{U,}\\ \lambda>X_{(n)}-\xi,\ f=S_B,\\ \lambda=1,\ f\in\{S_L,S_N\},\end{cases} \quad (6)$$

$$\begin{cases}\xi<X_{(1)},\ f\in\{S_L,S_B\},\\ \xi=0,\ f=S_N,\end{cases} \quad (7)$$

$$\left|RootOf\left(1-\sum_{h=1}^{p}\alpha_hB^h=0,B\right)\right|>1, \quad (8)$$

Задачу оптимизации можно декомпозировать на две: определение параметров распределения Джонсона $\gamma,\delta,\lambda,\xi$ и параметров базового AR(p)-процесса $\alpha_1,\alpha_2,...,\alpha_p$. Таким образом, последовательно улучшаются оценки параметров $(\gamma,\delta,\lambda,\xi)$ и $(\alpha_1,\alpha_2,...,\alpha_p)$.

Пусть $C:\Psi\to\Psi$ и $D:\Psi\to\Psi$ – отображения, задаваемые следующими отношениями:

$C(\psi)\equiv\arg\min_{\gamma,\delta,\lambda,\xi}S_D(\psi|x)$ по $\psi\in\Psi_C$ и

$D(\psi)\equiv\arg\min_{\alpha_1,\alpha_2,...,\alpha_p}S_D(\psi|x)$ по $\psi\in\Psi_D$,

где Ψ_C – ограничения, обеспечивающие соответствие значений параметров распределения Джонсона семейству распределений ((5)-(7)), а Ψ_D – ограничение, обеспечивающее стационарность базового авторегрессионного процесса (8) и $\Psi=\Psi_C\cup\Psi_D$.

Данный алгоритм сходится при соблюдении следующих условий:

1. Если на границе области минимизации нет локальных минимумов целевой функции, то любое возможное решение ψ задачи оптимизации $\min_\psi S_D(\psi|x)$ лежит внутри области $cl\Psi$.

2. Функция $R_{(t)}(\psi)=\Phi\left[V_{(t)}(\psi)\right]$ трижды непрерывно дифференцируема для всех $\psi\in\Psi$.

Предложенная модель случайных входных воздействий позволяет оценивать параметры и генерировать случайные входные потоки с внутренней зависимостью и сезонной изменчивостью.

В общем виде алгоритм оценки параметров выглядит следующим образом (см. рисунок 1).

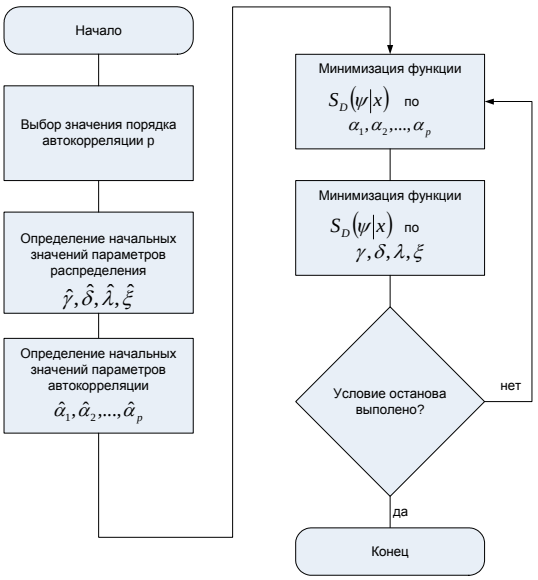


Рисунок 1 - Алгоритм оценки параметров ARTA-процесса

Параметры ARTA-процесса оцениваются для всех типов распределения Джонсона, и по результатам проверки по критериям согласия осуществляется выбор наиболее подходящего распределения.

Для выбора порядка автокорреляции базового процесса используется критерий Шварца. Начальные оценки параметров распределения выполняются методом наименьших квадратов. Оценки начальных значений параметров автокорреляции являются решением уравнений Юла-Уолкера:

$$\sum_{h=1}^{p}\alpha_h\rho_Z(h-k)=\rho_Z(k),\ k=1,2,...,p \quad (9),$$

где $\rho_Z(k)$ – коэффициент автокорреляции k-го порядка.

Для вычисления целевой функции на каждом шаге:

- Решаются уравнения Юла-Уолкера относительно коэффициентов корреляции (шаг от 1 до p);
- Вычисляется значение

$$\sigma_Y^2 = 1 - \sum_{h=1}^{p} \alpha_h \rho_Z(h);$$

- Вычисляются значения $V_t(\psi), t = 1,2,...,n$, используя (4);
- Значения порядковых статистик $V_{(p+1)}(\psi), V_{(p+2)}(\psi), ..., V_{(n)}(\psi)$ получают, упорядочив $V_t(\psi), t = 1,2,...,n$;
- Вычисляется значение целевой функции (3).

Для минимизации целевой функции используется алгоритм Левенберга-Маркардта, так как строго доказано, что этот метод обеспечивает сходимость алгоритма оценки параметров ARTA-процесса.

В качестве условия останова алгоритма выбрано следующее условие:

$$\left| S_D(\psi_k | x) - S_D(\psi_{k-1} | x) \right| \le S_D(\psi_{k-1} | x) \times \Delta,$$

где Δ – априори заданная относительная ошибка.

5. Генерация реализации ARTA-процесса

Алгоритм генерации реализации ARTA-процесса с заданными параметрами в необходимом объеме выглядит следующим образом в соответствии с определением ARTA-процесса (2):

1. Генерация $n + p$ нормально распределенных случайных величин Y с $E[Y] = 0$ и $Var[Y] = \sigma_Y^2$.

2. Определение n значений базового авторегрессионного процесса Z: $Z_t = \sum_{h=1}^{p} \alpha_h Z_{t-h} + Y_t$.

3. Определение n значений искомого процесса X: $X_t = \xi + \lambda f^{-1}\left[\frac{Z_t - \gamma}{\delta}\right]$.

Для генерации нормально распределенных величин используется преобразование методом полярных координат Бокса и Мюллера. Равномерно распределенные на отрезке $[0;1]$ случайные величины генерируются по алгоритму Mersenne Twister.

Значения авторегрессионного процесса вычисляются в соответствии с определением процесса (1). При этом необходимо определить первые p значений AR(p)-процесса. Для этого рассчитывается матрица ковариаций AR-процесса K и ее разложение Холецкого $K = VV^T$, далее для сгенерированного вектора нормально распределенных независимых величин Y размерности p начальные значения авторегрессионного процесса находят следующим образом: $Z = V \cdot Y$.

Вышеописанные алгоритмы оценки параметров и генерации реализации ARTA-процесса позволяют моделировать случайные входные воздействия, являющиеся стационарными процессами.

6. Оценка адекватности модели

После выбора модели входных данных необходимо оценить ее адекватность. Для стационарного случайного процесса требуется оценить адекватность выбора двух его ключевых особенностей: безусловного распределения вероятностей и структуры корреляции. Для оценки адекватности выбранного безусловного распределения вероятностей решено использовать такие статистические критерии согласия, как критерии Колмогорова-Смирнова и Андерсона-Дарлинга. В отличие от распространенного критерия хи-квадрат, они позволяют сравнивать распределения двух выборок без ограничения на независимость выборок. Для наглядности используется график квантиль-квантиль, по которому легко можно видеть, насколько хорошо построенная модель описывает исходную выборку. Для оценки адекватности подобранной структуры корреляции используются спектральные версии критериев Колмогорова-Смирнова и Андерсона-Дарлинга, а также сравнение коррелограмм и диаграмм разброса наблюдений.

7. Результаты

Для исследования адекватности модели ARTA-процесса было произведено моделирование стационарного процесса.

График исходной реализации представлен на рисунке 2а.

Процедура оценки параметров определила следующие параметры ARTA-процесса:

$$f = SU,\ \gamma = 1.915,\ \delta = 2.873,\ \lambda = 0.402,$$
$$\xi = 1.203,\quad p = 2,\ \alpha = \{1.061, -0.392\}$$

Была сгенерирована реализация ARTA-процесса с указанными параметрами того же объема (500 элементов), что и исходная. График реализации представлен на рисунке 2б.

Для сравнения использовалась модель случайной величины, полученная при рассмотрении исходной реализации как выборки независимых и одинаково распределенных данных (распределение подбиралось из семейства распределений Джонсона), и модель AR-процесса. На рисунке 2в приведена реализация сгенерированного процесса по модели случайной величины, на рисунке 2г – реализации по модели AR-процесса.

По характеру поведения графиков 2а-2г видно, что моделирование ARTA-процесса с автоматической оценкой параметров дает лучший результат. Рассмотрим графики квантиль-квантиль и диаграммы разброса наблюдений для разных моделей.

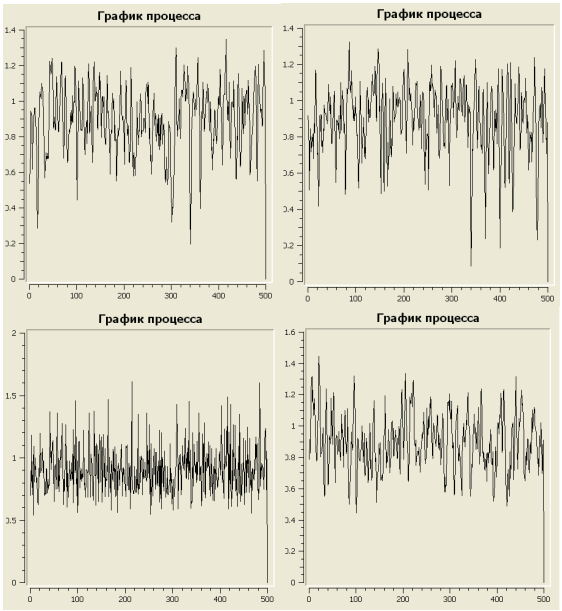


Рисунок 2 - Графики процесса (а - исходная реализация, б - ARTA-процесс, в – случайная величина, г – AR-процесс)

Для оценки адекватности выбранного безусловного распределения построены графики квантиль-квантиль (см. рисунки 3а-в). Из графиков видно, что автоматизированный подбор безусловного распределения дает хорошее приближение к истинному безусловному распределению моделируемого процесса, и лучшее распределение, чем оценка распределения методами, предназначенными для независимых данных (таким образом, методы моделирования случайной величины не только не подбирают структуру корреляции, но и неверно оценивают безусловное распределение, лежащее в основе данных).

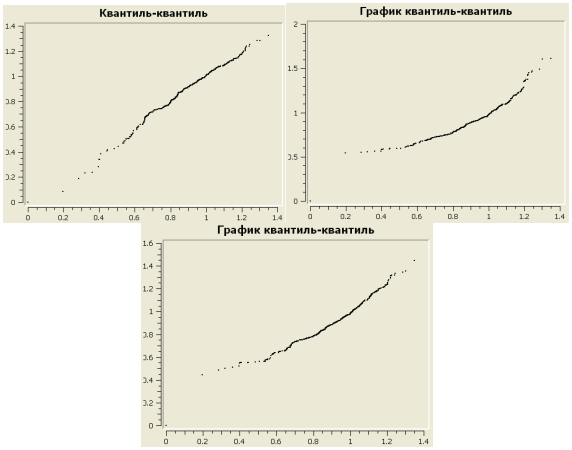


Рисунок 3 – График квантиль-квантиль (а - исходная реализация –ARTA-процесс, б - исходная реализация – случайная величина, в - исходная реализация - AR-процесс)

Для оценки адекватности выбранной структуры корреляции (то есть базового процесса) были построены диаграммы разброса для единичного шага исходной и сгенерированных реализаций (см. рисунки 4а-г).

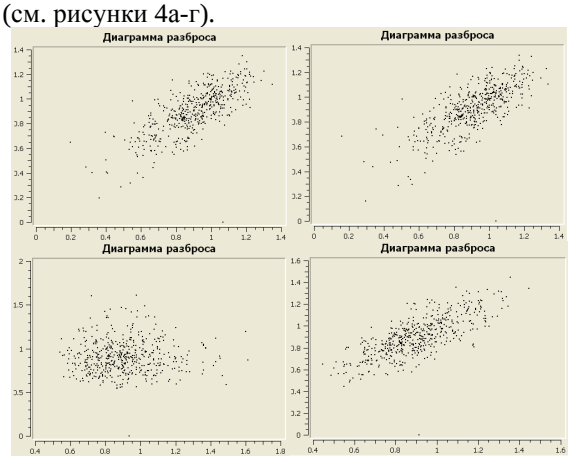


Рисунок 4 - Диаграмма разброса (а - исходная реализация, б - ARTA-процесс, в - случайная величина, г - AR-процесс)

По диаграммам разброса видно, что структура корреляции для ARTA-процесса подбирается лучше, чем в случае применения модели авторегрессионного процесса.

Таким образом, автоматизированная процедура моделирования ARTA-процесса дает приемлемые результаты, лучшие по сравнению с AR-процессами.

12. References

[1] Андерсон, Т. *Статистический анализ временных рядов*, Мир, Москва, 1976.

[2] Вентцель, Е.С., Л.А. Овчаров, *Теория случайных процессов и ее инженерные приложения*, Издательский центр «Академия», Москва, 2003.

[3] Кендалл, М., А. Стьюарт, *Многомерный статистический анализ и временные ряды*, Наука, Москва, 1976.

[4] Шляева, А.В., "Использование ARTA- и TES-процессов для моделирования входных данных стохастических имитационных моделей", *Сборник трудов научно-практического семинара "Новые информационные технологии в автоматизированных системах-9"*, 2006, 178-180с.

[5] Biller, B., B.L. Nelson, "Parameter estimation for ARTA processes", *Proceedings of the 2002 Winter Simulation Conference*, 2002, p.255-262.

[6] Cario, M.C., B.L. Nelson, **"**Autoregressive to anything: Time-series input processes for simulation", *Operations Research Letters*, 1996, p.51-58.

[7] Gaver, D.P., P.A.W. Lewis, "First-order autoregressive gamma sequences and point processes", *Advances in Applied Probability*, 1980, p.727-745.

[8] Livny, M., B. Melamed, A.K. Tsiolis, "The impact of autocorrelation on queueing systems", *Management Science*, 1993, p.322-339.

[9] Mallows, C.L., "Linear processes are nearly Gaussian", *Journal of Applied Probability*, 1967, p.313-329.

[10] Melamed, B., J.R. Hill, D. Goldsman, "The TES methodology: Modeling empirical stationary time series", *Proceedings of the 1992 Winter Simulation Conference*, 1992, p.135-144.

[11] Nelson, B.L., M. Yamnitsky, "Input modeling tools for complex problems", *Proceedings of the 1998 Winter Simulation Conference*, 1998, p.105-112.

[12] Schmeiser, B., "Advanced input modeling for simulation experimentation", *Proceedings of the 1999 Winter Simulation Conference*, 1999, p.110-115.

[13] Ware, P.P., T.W. Page, B.L. Nelson, "Automatic modeling of file system workloads using two-level arrival processes", *ACM TOMACS*, 1998, p.305-330.

Алгоритм прогнозирования функции распределения нестационарного временного ряда

Осминин Константин Павлович
OsminiK@yandex.ru

Аннотация

Предложен алгоритм, позволяющий прогнозировать выборочную функцию распределения временного ряда, в том числе нестационарного, на заданный временной горизонт. Алгоритм основан на эмпирическом уравнении Лиувилля.

1. Введение

В практической деятельности часто возникает задача прогнозирования нестационарных временных рядов, т.е. рядов, моменты или выборочные функции распределения (ВФР) которых меняются во времени. В то же время основные результаты теории вероятности и математической статистики и их приложений относятся лишь к стационарным рядам [1]. Чтобы применить строгие результаты теории к нестационарным рядам, чаще всего используется следующий прием [2]: выбирается такой отрезок ряда, на котором характеристики ряда, например, его функция распределения или несколько первых выборочных моментов, меняются несильно с точки зрения практических нужд исследователя. Затем проводится анализ такой выборки стационарными методами, для оценки точности которых используются классические доверительные интервалы тестов на стационарность [3]. После этого выборка «скользит» по оси времени, порождая соответствующие скользящие средние величины. Методологическим недостатком такого подхода является то обстоятельство, что доверительные интервалы представляют собой асимптотику по объему выборки статистики отклонений сумм случайных величин от их теоретических значений, но для нестационарных процессов нет теорем об асимптотической сходимости (ни по вероятности, ни в ином смысле). Поэтому оценки, получаемые с помощью асимптотических методов, могут приводить для нестационарных рядов к существенно неточным результатам. Кроме того, объем выборки, наиболее оптимальный поставленной задаче, остается в таком подходе неопределенным.

Таким образом, для повышения точности анализа случайных данных и их прогноза требуется ввести меру нестационарности временного ряда на отрезке, более адекватную задачам практического прогнозирования. В совместных с Ю.Н. Орловым работах автора [4 - 6] вводится мера нестационарности ряда равномерно ограниченных случайных величин на основе функционала расстояния на пространстве неотрицательных суммируемых функций, нормированных на единицу.

Пусть требуется составить прогноз ВФР на заданный промежуток времени вперед, так называемый горизонт, с точностью не ниже заданной. Обоснования, того, в какой мере прогноз ВФР дополняет прогноз самого ряда, приведены в [4, 5]: прогноз ВФР показывает, с какой вероятностью ожидается то или иное значение ряда, поскольку доверительные интервалы в их классическом понимании для нестационарных процессов не определяются. В этом случае возникает задача определения объема выборки, оптимального для прогнозирования. Оптимальность объема выборки связана с тем, что имеются два типа ошибки в оценке статистик ряда. Первая ошибка связана с конечностью объема выборки. Она велика при малых объемах выборки и мала при больших. Вторая ошибка обусловлена нестационарностью процесса. Она, напротив, мала при малых объемах выборки и велика при больших. Например, характерная зависимость выборочной дисперсии нестационарного ряда от объема выборки имеет минимум в области квази-стабилизации.

В работе [6] был сформулирован алгоритм определения вышеуказанного оптимального объема выборки, кратко описываемый ниже. На основе этого метода в данной работе представлен алгоритм, прогнозирующий сам ряд и его ВФР на заданный горизонт с заданной точностью.

2. Основные определения

Пусть имеется временной ряд *{x(t)}*, *t=1..Tmax*. Из практических соображений удобно считать ряд равномерно ограниченным по времени, так что без ограничения общности полагаем, что область значений ряда – это интервал (-1;1). Разобьем его на N равных отрезков и введем плотностную ВФР ряда *f(y,t)*, построенную по выборке объема T на промежутке *(t-T,t]* целочисленных значений времени, и заканчивающейся в момент времени t:

$$f(i/N,t)=\frac{1}{T}\#\left\{x(t'):\frac{i}{N}\le x(t')<\frac{i+1}{N}\right\}. \qquad (1)$$

В этой формуле i меняется от –N до N-1, а *t'* от *t-T+1* до *t*.

Расстояние между двумя ВФР одного ряда, построенных по выборкам одинакового объема, но в разные моменты времени *t'* и *t''*, определяется как

$$\rho\big(f(y,t'),f(y,t'')\big)=\int_{-1}^{1}\big|f(y,t')-f(y,t'')\big|\,dy. \qquad (2)$$

В данном случае под интегралом следует понимать соответствующую сумму. Интеграл написан для сокращения записи и облегчения понимания.

Ошибка прогноза ВФР определяется как расстояние (2) между спрогнозированной и истинной функциями распределения.

Объем выборки T, по которому строится ВФР, определяется одним из следующих способов:

1. T определено из априорных соображений. Это могут быть специфические требования к ВФР, ограниченный объем статистики ряда и пр.
2. В качестве T берется оценка оптимального объема выборки, определяемая с помощью алгоритма CHS, описываемого в работе [7]. Имея изначально заданные величины горизонта прогноза τ и близости ВФР ε, этот подход позволяет гарантировать, что ошибка прогнозирования ВФР на τ шагов вперед будет иметь ошибку не выше ε, т к. самый простой прогноз, наивный, уже имеет эту ошибку. Грубо говоря, алгоритм делает следующее: строится функционал расстояния в смысле (2) между ВФР в момент времени t и «сдвинутой» ВФР построенной в момент времени t+ τ. При этом объем выборки T, по которому построены обе ВФР, одинаков и является параметром функционала. Далее по заданному ε для каждого t находится такое значение T', что при всех значениях T, больших либо равных T', значение функционала меньше ε. Далее квантили полученной статистики T', зависящей от t, используются в качестве оценки оптимального объема выборки.

3. Алгоритм прогнозирования ВФР

Задача, решаемая этим алгоритмом, формулируется так: определить прогноз величин *f(y,t)* в моменты времени *t=tMax+1,..,tMax+τ* по данным *x(t)* в моменты времени *t=1..tMax*.

Шаги алгоритма:

1. Введем конечно-разностную производную *p(t)=x(t+1)-x(t)*. Вычислим *p(t)* в моменты времени *t=1..tMax-1*.
2. Установим *t=tMax*.
3. Рассмотрим ВФР по двум переменным x и p, обозначим ее через *F(y,p,t)*. Аналогично определению одномерной ВФР, построим массив

$$F\left(\frac{i}{N},\frac{2j}{N},t\right)=\frac{1}{T}\#\left\{\frac{i}{N}\le x(t')<\frac{i+1}{N},\ \frac{2j}{N}\le p(t')<\frac{2(j+1)}{N},t'\in(t-T,t]\right\} \qquad (3)$$

Здесь *i* и *j* меняются от *-N* до *N*-1. Таким образом, выполняется равенство

$$f(y,t)=\sum_{j=-N}^{N-1}F\left(y,\frac{2j}{N},t\right). \qquad (4)$$

4. Введем функцию «средней скорости»

$$u(y,t)=\frac{\sum_{j=-N}^{N-1}\frac{2j}{N}\cdot F\left(y,\frac{2j}{N},t\right)}{f(y,t)}. \qquad (5)$$

Если знаменатель равен нулю, то средняя скорость также приравнивается к нулю.

5. Определим прогноз ВФР в следующий момент времени через его значения в предыдущий и значения средней скорости согласно формуле

$$\tilde f\left(\frac{i}{N},t+1\right)=f\left(\frac{i}{N},t\right)\left(1-u\left(\frac{i}{N},t\right)\right)+u\left(\frac{i-1}{N},t\right)f\left(\frac{i-1}{N},t\right) \qquad (6)$$

6. Определим выборочную двумерную функцию распределения в момент времени t+1 по формулам

$$\tilde F\left(y,\frac{2j}{N},t+1,T\right)=F\left(y,\frac{2j}{N},t,T\right)+f(y,t+1,T)-f(y,t,T)\ \text{ïðè}\ \ j=(\lfloor(y-x(t))\cdot N\rfloor)/2, \qquad (7)$$

$$\tilde F\left(y,\frac{2j}{N},t+1,T\right)=F\left(y,\frac{2j}{N},t,T\right)\ \text{ïðè}\ \ \text{îñòàëüíûõ}\ \ j.$$

7. Определим

$$\tilde u(y,t+1)=u(y,t)+p\times\times\left(\tilde f(y,t+1)-f(y,t)\right) \qquad (8)$$

8. Установим *t=tMax+1*.
9. Определим прогноз ВФР по формуле

$$\tilde{f}\left(\frac{i}{N},t+1\right)=\tilde{f}\left(\frac{i}{N},t\right)\left(1-\tilde{u}\left(\frac{i}{N},t\right)\right)+ \\ +\tilde{u}\left(\frac{i-1}{N},t\right)\tilde{f}\left(\frac{i-1}{N},t\right) \quad (9)$$

10. Определим прогнозные значения выборочной двумерной функции распределения по формулам

$$\tilde{F}(y,p,t+1,T)=\tilde{F}(y,p,t,T)+ \\ \left(\tilde{f}(y,t+1,T)-\tilde{f}(y,t,T)\right)\cdot\tilde{f}(y-p,t,T) \quad (10)$$

при $p-1<y<p+1$ и

$$\tilde{F}(y,p,t+1,T)=\tilde{F}(y,p,t,T)$$

при остальных y и p.

11. Определим прогнозные значения средней скорости в момент времени t+1 по формуле (6).

12. Повторим действия пп.9-11 при $t=tMax+2,..,\ tMax+\tau$.

Алгоритм завершен.

4. Комментарии и обоснования

Обоснования алгоритма приведены в работах [5] и [8]. В них описаны свойства горизонтной статистики квазистационарного ряда и показано, что она остается стационарной с заданной точностью внутри горизонта прогнозирования.

Формулы (6) и (9), позволяющие определить прогнозное значение ВФР, есть разностная версия эмпирического уравнения Лиувилля

$$\frac{\partial f}{\partial t}+\frac{\partial(uf)}{\partial x}=0, \quad (11)$$

а формула (5) есть разностный аналог соотношения

$$u(y,t,T)\,f(y,t,T)=\int pF(y,p,t,T)dp. \quad (12)$$

Для описания эволюции двумерной выборочной функции распределения используются формулы (7) и (10) для сохранения согласованности с одномерной выборочной функцией распределения (4).

Прогнозный алгоритм, не заменяет собой различные методы обработки ряда на предварительном этапе, исключающие из него периодические или трендовые составляющие. Он применяется тем эффективнее, чем лучше пред-обработан ряд, т.е. чем ближе остаток ряда к белому шуму. Алгоритм предполагается к применению в тех случаях, когда временной ряд обработан всеми доступными методами, но амплитуда и/или дисперсия остатка слишком велика. В этом случае алгоритм, примененный к остатку, поможет улучшить точность прогноза ряда.

5. Применение алгоритма

Описанный алгоритм был применен для анализа и прогноза нестационарных временных рядов, возникающих в финансовой деятельности: это ряд биржевых цен на акции некоторых крупных компаний (Майкрософт, Дженерал Моторз, Дженерал Электрик и др.), а также почасовые цены на электроэнергию на российском оптовом рынке электроэнергии и мощности (ОРЭМ) в европейской ценовой зоне. Алгоритм был также протестирован на некоторых стационарных временных рядах, порождаемых хаотической динамикой в системах Эно и Лоренца. Было численно подтверждено, что алгоритм доставляет прогнозу ВФР декларируемую точность.

Пример прогнозирования ВФР с точностью 5% для ряда цен на ОРЭМ на сутки вперед (т.е. на горизонт в 24 шага) по оптимальной выборке в приведен на Рис 1.

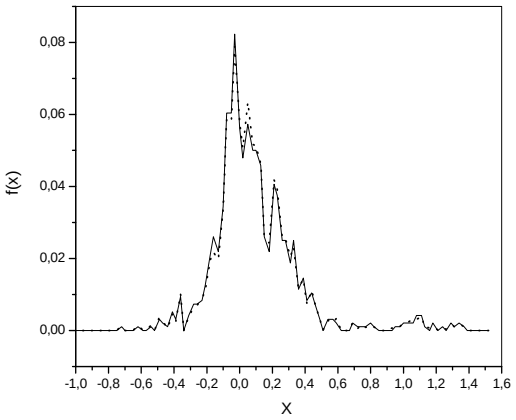


Рис. 1. Фактическая (сплошная линия) и прогнозная (пунктир) ВФР для ряда ОРЭМ.

6. Заключение

В работе представлен алгоритм прогнозирования выборочной функции распределения. Имея такой прогноз, прогноз самого временного ряда можно получить, используя алгоритм, описанный в 7 параграфе работы [7].

Наличие совместного прогноза ряда и его ВФР, доставляемое описанным выше алгоритмом, позволяет получать детальное представление об изучаемом процессе. К примеру, кроме значения прогноза самого ряда, оцениваются границы доверительности такого прогноза.

Данный подход может быть полезен и при аппроксимации данных, если требуется изучить

характеристики случайных величин, которые могут не стабилизироваться с течением времени.

Метод позволяет также отслеживать изменения в случайном процессе, т.е. определять появление новых тенденций в поведении ряда и предсказывать моменты их появления. К примеру, горизонтный ряд $h(t,\tau)$ может быть использован в качестве индикатора изменения закономерностей ряда $x(t)$. Резкое изменение горизонтного ряда по переменной t сигнализирует о возможном появлении нового фактора, влияющего на динамику исходного ряда, либо о существенном изменении имеющихся факторов. Профиль горизонтного ряда является показателем относительной неизменности совокупности определяющих ряд факторов. Данный индикатор может быть полезен, например, для анализа изменения структуры рынка ценных бумаг, изменения эластичности спроса, а период перехода горизонтного ряда к новой устойчивой структуре может характеризовать скорость протекания процессов саморегуляции в системе. Алгоритмы, реализующие такой подход, являются перспективным направлением дальнейших исследований.

8. Литература

[1] Уилкс С. Математическая статистика. (пер. с англ.) – М.: Наука, 1967. – 632 с.

[2] Кендалл М., Стюарт А. Статистические выводы и связи. (пер. с англ.) – М.: Наука, 1973. – 900 с.

[3] Кобзарь А.И. Прикладная математическая статистика. – М.: Физматлит, 2006. – 816 с.

[4] Орлов Ю.Н., Осминин К.П. Анализ нестационарных временных рядов. / Препринт ИПМ им. М.В. Келдыша РАН, № 36, 2007. – 24 с.

[5] Орлов Ю.Н., Осминин К.П. Методика определения оптимального объема выборки для прогнозирования нестационарного временного ряда. // Информационные технологии и вычислительные системы, 2008, № 3, с. 39-50.

[6] Орлов Ю.Н., Осминин К.П. Построение выборочной функции распределения для прогнозирования нестационарного временного ряда. // Математическое моделирование, 2008, № 7 (в печати)

[7] Осминин К.П. Алгоритмы построения статистик для анализа и прогнозирования нестационарных временных рядов // Информационные технологии и вычислительные системы, 2008, №4 (в печати)

[8] Орлов Ю.Н., Осминин К.П. О квазистационарности статистики «горизонтного» ряда для нестационарных временных рядов. / В сб. Современные проблемы фундаментальной и прикладной математики. М.: МФТИ, 2008. С. 123-132.

Многоэтапные процедуры проверки статистических гипотез и субоптимальность

Федор Цитович
Институт проблем передачи информации им. А.А. Харкевича РАН
ftsitovich@gmail.com

Аннотация

Многоэтапные последовательные процедуры проверки статистических гипотез часто имеют практическое преимущество над стандартными последовательными процедурами. В данной работе исследуется субоптимальность многоэтапных процедур. В статье показано, что, не обладая достаточной априорной информацией об истинном распределении, и как следствие, ограничиваясь субопимальными процедурами, количество этапов в многоэтапной процедуре может быть уменьшено до 2-х с сохранением свойств субоптимальности. Таким образом, строящиеся процедуры оказываются существенно более простыми, чем оптимальные многоэтапные процедуры, что позволяет уменьшить время обработки результатов наблюдений.

1. Введение

В [1] рассмотрена задача построения робастных субоптимальных правил проверки непараметрических гипотез, представляющих собой слабые возмущения заданных распределений, задающих соответствующие простые гипотезы. В работе было показано, что введение дополнительных поправок в стандартный последовательный тест Вальда позволяет получать гарантийные решающие правила и в случае неопределенности в законе распределения наблюдений, причем построенные правила являются близкими к асимптотически оптимальным правилам проверки этих гипотез.

Вместе с тем, в задачах последовательной проверки гипотез бывает важным знать момент окончания наблюдений, а решение продолжать наблюдения может быть существенно более дорогостоящим, чем стоимость одного наблюдения c. Это обусловлено различными причинами. Одна из них состоит в организации наблюдений, когда проще задать проведение серии наблюдений заранее заданной длины, чем после каждого наблюдения принимать такое решение. Вторая состоит в том, что, как это следует, например, из [1] или [2], на каждом шаге приходится решать некоторую оптимизационную задачу для принятия решения о необходимости продолжения наблюдений. Решение такой задачи требует значительных вычислительных затрат. Стоимость таких затрат включают в параметр M, $M >> c$. Для учета ограничения на моменты принятия решения вводится ограничение на решающее правило, состоящее в том, что остановка наблюдений возможна лишь в определенные моменты времени, которые задаются до начала соответствующей серии наблюдений. Таким образом возникают многоэтапные решающие правила, в которых стоимость этапа равна M и требуется минимизировать стоимостной функционал, представляющий собой среднее значение стоимости всех наблюдений и всех этапов. В [4] рассмотрена соответствующая задача для проверки простых гипотез, где получено асимптотически оптимальное правило проверки гипотез в такой постановке. В этой работе построена процедура проверки гипотез, увязывающая отношения между M/c и $-\ln\alpha$, где α — вероятность ошибки решающего правила. В этом случае, в зависимости от значений указанных выше параметров и полученных результатов наблюдений на момент завершения очередного этапа, выбирается продолжительность следующего этапа, причем количество этапов может стремиться к бесконечности при $\alpha \to 0$. Таким образом, построена достаточно сложная процедура, обладающая более строгими свойствами оптимальности, чем стандартные асимптотически оптимальные процедуры, однако требующая для своей реализации точное знание о законах распределения наблюдений при каждой из гипотез.

В том случае, когда распределения наблюдений точно не известны, это правило оказывается

излишне сложным, поскольку неопределенность в распределении наблюдений оказывает более существенное влияние на результат, чем введение дополнительных этапов, связанных с обеспечением заданных значений вероятности ошибок. Поэтому цель настоящей работы состоит в обобщении рассмотренной в [4] задачи на случай сложных гипотез, рассматриваемых в [1].

2. Постановка задачи

Пусть $(\Omega, \mathcal{F}, \mu)$ — вероятностное пространство, на котором определен класс $\mathcal{P} = \{\mathsf{P}\}$ вероятностных мер, абсолютно непрерывных относительно меры μ и взаимно абсолютно непрерывных. Мера P порождает распределение случайной величины x со значениями из множества $X \subset \mathbf{R}$, $\mathbf{R}$ — множество действительных чисел, функция распределения которой имеет плотность относительно меры μ, которая обозначается через $f = f(x)$.

Будем решать задачу проверки сложных гипотез

$$\mathcal{H}_1 : f \in \mathcal{G}_1, \ldots, \mathcal{H}_m : f \in \mathcal{G}_m, \tag{1}$$

где $\mathcal{G}_i := \left\{ g_{i,h} : g_{i,h_i}(x) = g_i(x)(1 + h_i(x)) \right\}$, $g_i(x)$ — плотности, задающие гипотезы $\mathcal{H}_i$, известны, а $h_i(x)$ не известны, причем таковы, что:

$$1) \sup_{-\infty < x < +\infty} |h_i(x)| \le \varepsilon < 1, \tag{2}$$

$$2) \int_{-\infty}^{+\infty} g_{i,h_i}(x)\, d\mu(x) = 1. \tag{3}$$

Множества $\mathcal{G}_i$ зависят, вообще говоря, от ε, но далее для простоты обозначений эту зависимость будем опускать.

Множество $\mathcal{G}_i$ можно трактовать как непараметрическую окрестность функции $g_i(x)$, состоящую из плотностей, относительное отклонение которых от $g_i(x)$ не превосходит некоторую малую величину ε. В качестве одного из примеров возникновения таких окрестностей в статистике может быть следующий. В серии проводимых наблюдений может присутствовать малая доля "выбросов", т.е. наблюдений, вызванных механизмом формирования, отличным от исследуемого $g_i(x)$ (например, задаваемым плотностью $g(x)$). Предполагается, что отделить такие наблюдения от "правильных" не представляется возможным. Тогда для этой ситуации можно применить окрестности, задаваемые соотношениями (2) и (3) в случае, когда $h_i(x) = a\left(\frac{g(x)}{g_i(x)} - 1\right)$, где a — доля выбросов в выборке. Заметим, что аналогичные процедуры можно строить и для других окрестностей плотностей $g_i(x)$, однако оценки для средней продолжительности последовательной процедуры проверки гипотез будут зависеть от вида этих окрестностей.

Пусть $x_1, x_2, \ldots$ — наблюдения, т.е. независимые одинаково распределенные случайные величины с плотностью $f(x)$, $\left\{\mathcal{F}_n\right\}$ — естественная фильтрация, порождаемая наблюдениями, т.е. $\mathcal{F}_n = \sigma(x_1, \ldots, x_n)$, $\mathcal{F}_0$ — тривиальная σ-алгебра.

Определение 1. Процедура d для задачи проверки гипотез (1) называется допустимой, если выполнены следующие условия:

1. задана последовательность продолжительности этапов наблюдений $N_1 > 0, N_2 > 0, \ldots$, а $\tau_0 = 0, \tau_i = \tau_{i-1} + N_i$, $i > 0, \ldots$ — соответствующая последовательность моментов остановки этапов наблюдений, при этом N_i является $\mathcal{F}_{\tau_{i-1}}$-измеримой целочисленной случайной величиной;

2. момент завершения наблюдений τ совпадает с одним из моментов τ_i, причем τ — марковский момент остановки относительно фильтрации $\left\{\mathcal{F}_n\right\}$, т.е. $\forall\, n \in \mathbf{N}\ \left\{\omega : \tau(\omega) \le n\right\} \in \mathcal{F}_n$ и $\mathsf{P}(\tau < \infty) = 1$;

3. решающее правило $\delta(\cdot)$ является $\mathcal{F}_\tau$ измеримым отображением в множество $\{1, 2, \ldots, m\}$, т.е. $\delta = \delta(x_1, \ldots, x_\tau)$, причем $\delta = i$ означает, что принимается гипотеза $\mathcal{H}_i$;

4. стратегия d обеспечивает заданный уровень вероятности ошибки принятия неправильного решения, т.е. при всех $i \ne j$ выполнено неравенство

$$\sup_{f \in \mathcal{G}_j} \int_{\delta = i} f(x)\, d\mu(x) \le \alpha. \tag{4}$$

Множество допустимых процедур обозначим через $\mathcal{D}(\alpha)$.

Определение 2. Функцией риска стратегии $d = \langle \tau, \delta \rangle$ для задачи проверки гипотез (1) назовем максимальную среднюю продолжительность процедуры с учетом стоимости этапов, т.е.

$$R_{\mathcal{H}_i}(d) = \sup_{f \in \mathcal{G}_i} \mathsf{E}_f(c\tau + Mi^*), \tag{5}$$

где E_f — математическое ожидание, когда наблюдения имеют плотность f, а i^* — число этапов процедуры, которое определяется из условия $\tau_{i^*} = \tau$.

Наличие $\sup_{f \in \mathcal{G}_i}$ в функции риска обусловлен тем обстоятельством, что "размеры" областей $\mathcal{G}_i$ предполагаются малыми (например ε — малое

положительное число), время, необходимое для построения состоятельной оценки наблюдаемого распределения $h(x)$, слишком велико и поэтому решающее правило должно работать одинаково хорошо для всех возможных распределений. С другой стороны, малость областей $\mathcal{G}_i$ приводит к тому, что замена $\sup_{f\in\mathcal{G}_i}$ на точное распределение не приведет к существенному изменению значения функции риска.

3. Описание субоптимальной процедуры d_0

Для $f \in \mathcal{G}_i, i = 1, \ldots, m$, обозначим через $A(f)$ альтернативное множество для плотности f, т.е.

$$A(f) := \bigcup_{\substack{i=1,\ldots,m,\\ i\neq i_0}}^{m} \mathcal{G}_i, \quad \text{если } f \in \mathcal{G}_{i_0}.$$

Будем использовать следующие стандартные обозначения:

$$z_{f,g}(x) := \ln\frac{f(x)}{g(x)},\ x \in X,$$
$$I(f,g) := \mathsf{E}_f z_{f,g}(x);$$
$$I_i := \min_{g_k\in A(g_i)} I(g_i, g_k);$$
$$I_+ := \max_{\substack{i,j=1,\ldots,m\\ i\neq j}} I(g_i, g_j);$$
$$I_- := \min_{i=1,\ldots,m} I_i;$$
$$l_f(g;n) := \sum_{i=1}^{n} z_{f,g}(x_i);$$
$$l_g(n) := \min_{g_k\in A(g_i)} l_{g_i}(g_k;n);$$
$$L_i(n) := \inf_{g\in A(g_i)} l_{g_i}(g;n) =$$
$$= l_i(n) - n\ln(1+\varepsilon),\ i = 1,\ldots,m. \qquad (6)$$

Последнее равенство получено в [1].

1 этап. Проводится N_1 испытаний, после чего проверяется условие:

$$\exists i_0 : L_{i_0}(\tau_1) \geq -\ln\beta, \qquad (7)$$

где N_1 определяется из условия:

$$N_1 := \left[\frac{-\ln\beta}{I_+}\right] + 1, \qquad (8)$$

$$\beta := \frac{\alpha}{4(m-1)}. \qquad (9)$$

Заметим, что число N_1 не является случайным, что соответствует требованию 1 определения 1.

Если условие (7) выполняется, то процедура завершается, принимается гипотеза $\mathcal{H}_{i_0}$, иначе выполняется следующий этап процедуры.

2 этап. На основе статистического материала 1-го этапа строится оценка максимального правдоподобия для плотностей $g_i(x), i = 1, \ldots, m$:

$$\hat{i} = \arg\max_{i=1,\ldots,m} \sum_{j=1}^{N_1} \ln g_i(x_j) \qquad (10)$$

Длительность второго этапа N_2 определяется как:

$$N_2 := \left[(1+\Delta)\frac{-\ln\beta}{I_{\hat{i}}}\right] + 1 - N_1, \qquad (11)$$

где $\Delta := \Delta(\varepsilon)$ — некоторое положительное число, рассматриваемое как параметр стратегии.

Проводится N_2 испытаний, после чего проверяется условие

$$L_{\hat{i}}(\tau_2) \geq -\ln\beta. \qquad (12)$$

Если условие (12) выполняется, то процедура завершается, принимается гипотеза $\mathcal{H}_{i_0}$, иначе выполняется следующий Этап процедуры.

3 этап. Проводится N_3 испытаний, где

$$N_3 = \left[2\frac{-\ln\beta}{I_-}\right] + 1, \qquad (13)$$

после чего проверяется условие

$$\exists i_0 : L_{i_0}(N_3) \geq -\ln\beta, \qquad (14)$$

при этом при вычислении L_{i_0} используются только данные, полученные на этом этапе. Если условие (14) выполняется, то процедура завершается, принимается гипотеза $\mathcal{H}_{i_0}$, иначе выполняется снова выполняется этап 3.

Как видно из сказанного, третий этап может выполняться независимо несколько раз, если во всех предыдущих случаях никакая гипотеза не была принята. Отметим, что процедура строится таким образом, что выход на третий этап происходит только в том случае, если на предыдущих этапах никакая гипотеза не была принята. За счет выбора параметра Δ вероятность такого события мала, а вероятность неудачного завершения третьего этапа так же мала, поэтому при задании параметров процедуры на третьем этапе использовались грубые оценки для необходимой продолжительности наблюдений: продолжительность этапа примерно в 2 раза больше необходимой при самом неблагоприятном стечении обсто-

ятельств и никакая информация, известная на начало этого этапа, не используется.

4. Утверждения

Теорема 1. *Если*

$$\Delta > \max_{i=1,\dots,m} \frac{\varepsilon I_i + \ln(1+\varepsilon)}{(1-\varepsilon)I_i - \ln(1+\varepsilon)}, \qquad (15)$$

то при достаточно малых α процедура d_0 является допустимой процедурой.

Теорема 2. *Пусть $\mathsf{E}_{g_i}|\ln\frac{g_1(x)}{g_2(x)}|^2 \le C_i < \infty$. Тогда существует число K_1 такое, что функция риска процедуры d_0 удовлетворяет условиям*

$$R_{\mathcal{H}_i}(d_0) \le M(2+\gamma(\alpha)) + \\ + c\left((1+\Delta)\frac{-\ln\alpha}{I_i} + K_1\right), \qquad (16)$$

причем $\gamma(\alpha) \to 0$ при $\alpha \to 0$.

Замечание. Условие $\mathsf{E}_{g_i}|\ln\frac{g_1(x)}{g_2(x)}|^2 \le C_i < \infty$ может быть ослаблено как в [1], но это усложнит вид остаточного члена асимптотического разложения функции риска при $\alpha \to 0$.

5. Доказательства

Доказательство теоремы 1. Очевидно, что процедура d_0 удовлетворяет условиям 1–3. Проверим выполнение условия 4. Пусть, для определенности, $f \in \mathcal{G}_{i_0}$. Введем следующие обозначения: A_1 — событие, состоящее в том, что процедура завершилась после первого этапа, A_2 — процедура завершилась после второго этапа, B_i — процедура завершилась после i-ой попытки третьего этапа, C_i — событие, состоящее в том, что процедура завершилась после первого этапа и принята гипотеза $\mathcal{H}_i$, D_i — процедура завершилась после второго этапа и принята гипотеза $\mathcal{H}_i$, E_i — процедура завершилась после третьего этапа и принята гипотеза $\mathcal{H}_i$. Для событий E_i не указывается номер попытки, поскольку согласно определению процедуры d_0 на третьем этапе все попытки независимые и одинаковые с точки зрения распределения анализируемых на них событий. Из определения процедуры d_0 следует, что

$$\mathsf{P}_f(\delta \ne i_0) \le \sum_{i\ne i_0}\mathsf{P}_f(C_i) + \sum_{i\ne i_0}\mathsf{P}_f(D_i) + \\ + (1-\mathsf{P}_f(B_1^c))^{-1}\sum_{i\ne i_0}\mathsf{P}_f(E_i), \qquad (17)$$

где B_1^c — противоположное B_1 событие. Последнее слагаемое получено на основании того, что все попытки прохождения третьего этапа независимы.

Оценки вероятностей $\mathsf{P}_f(C_i)$, $\mathsf{P}_f(D_i)$ и $\mathsf{P}_f(E_i)$ проводятся по аналогичной схеме, поэтому оценим только $\mathsf{P}_f(C_i)$.

$$\mathsf{P}_f(C_i) = \mathsf{P}\left(L_i(N_1) \ge -\ln\beta\right) = \\ = \mathsf{E}_f\left(\mathcal{I}\left(L_i(N_1) \ge -\ln\beta\right)\right) \le \\ \le \mathsf{E}_{g_i}\left(\exp\left(-L_i(N_1)\right)\mathcal{I}\left(L_i(N_2) \ge |\ln\beta|\right)\right) \le \\ \le \mathsf{E}_{g_i}\left(\beta\mathcal{I}\left(L_i(N_1) \ge |\ln\beta|\right)\right) \le \beta.$$

Здесь $\mathcal{I}(A)$ — индикаторная функция события A. При доказательстве использован переход от меры, порождаемой плотностью f, к мере, порождаемой плотностью g_i, и использован тот факт, что $f \in A(g_{i_0})$, $i \ne i_0$.

Подставляя аналогичные неравенства для $\mathsf{P}_f(D_i)$ и $\mathsf{P}_f(E_i)$ в (17), получаем

$$\mathsf{P}_f(\delta \ne i_0) \le (m-1)\beta + (m-1)\beta + \\ + (1-\mathsf{P}_f(B_1^c))^{-1}(m-1)\beta = \\ \left(2 + (1-\mathsf{P}_f(B_1^c))^{-1}\right)(m-1)\beta.$$

С учетом (9) получаем

$$\mathsf{P}_f(\delta \ne i_0) \le \frac{2+(1-\mathsf{P}_f(B_1^c))^{-1}}{4}\alpha.$$

Из доказательства теоремы 2 следует, что $\mathsf{P}_f(B_1^c) \to 0$ при $\alpha \to 0$, поэтому неравенство (15) справедливо при достаточно малых α, для которых $\mathsf{P}_f(B_1^c) \le \frac{1}{2}$.

Доказательство теоремы 2. Используя введенные при доказательстве теоремы 1 обозначения, получаем

$$\mathsf{E}_f(c\tau + Mi^*) = c[N_1 + \mathsf{P}_f(A_1^c)\mathsf{E}_f(N_2) + \\ + \mathsf{P}_f(A_2^c)(1-\mathsf{P}_f(B_1^c))^{-1}N_3] + \\ + M(1+\mathsf{P}_f(A_1^c) + \\ + \mathsf{P}_f(A_2^c)(1-\mathsf{P}_f(B_1^c))^{-1}), \qquad (18)$$

а на основании (11)

$$\mathsf{E}_f(N_2) \le 1 - N_1 + \mathsf{P}_f(\delta = i_0)\times \\ \times\left[(1+\Delta)\frac{-\ln\beta}{I_{i_0}}\right] + \\ + \mathsf{P}_f(\delta \ne i_0)\left[(1+\Delta)\frac{-\ln\beta}{I_-}\right]. \qquad (19)$$

Далее будем использовать очевидные оценки $\mathsf{P}_f(A_1^c) \le 1$, $\mathsf{P}_f(\delta = i_0) \le 1$. Для оценки вероятностей других вероятностей придется использовать более точные оценки.

Из свойств оценки максимального правдоподобия следует, что

$$\mathsf{P}_f(\delta \neq i_0) \leq k_1 \alpha^{\gamma_1}, \tag{20}$$

для некоторого $\gamma_1 > 0$, причем k_1 и γ_1 не зависят от α и распределения f из G_{i_0}.

Из вида статистики (6) следует, что задача проверки сложных гипотез сводится, фактически, к аналогичной задаче проверки простых гипотез, поэтому можно воспользоваться оценками из [3]. Выбор параметра Δ (15) обеспечивает выполнение условия, что $\frac{\mathsf{E}_f L_{i_0}(N_2)}{-\ln\beta} - 1 > k_2 > 0$, причем число k_2 не зависит от α и распределения f из G_{i_0}, поэтому

$$\mathsf{P}_f(A_2^c) \leq k_3 \alpha^{\gamma_3}, \tag{21}$$

для некоторого $\gamma_1 > 0$, причем k_3 и γ_3 не зависят от α и распределения f из G_{i_0}.

Аналогично

$$\mathsf{P}_f(B_1^c) \leq k_4 \alpha^{\gamma_4}, \tag{22}$$

для некоторого $\gamma_4 > 0$, причем k_4 и γ_4 не зависят от α и распределения f из G_{i_0}.

Подставляя (20) в (19) получаем

$$\mathsf{E}_f(N_2) \leq \left[(1+\Delta)\frac{-\ln\beta}{I_{i_0}}\right] + \\ + k_1\alpha^{\gamma_1}\left[(1+\Delta)\frac{-\ln\beta}{I_-}\right]. \tag{23}$$

Теперь из (14), (21), (22) и (23) получаем, что

$$\mathsf{E}_f(\tau) \leq \left[(1+\Delta)\frac{-\ln\beta}{I_{i_0}}\right] + \\ + k_1\alpha^{\gamma_1}\left[(1+\Delta)\frac{-\ln\beta}{I_-}\right] + \\ + \frac{k_3\alpha^{\gamma_3}}{1-k_4\alpha^{\gamma_4}}\left[2\frac{-\ln\beta}{I_-}\right] + 3 \leq \\ \leq \left[(1+\Delta)\frac{-\ln\beta}{I_{i_0}}\right] + 3 + \\ + k_5\alpha^{\gamma_5}\left[(1+\Delta)\frac{-\ln\beta}{I_-}\right], \tag{24}$$

причем $k_5 > 0$ не зависит от α и распределения f из G_{i_0}, $\gamma_5 = \min(\gamma_1, \gamma_3)$. Из определения i^* и (21), (22) получаем, что

$$\mathsf{E}_f(i^*) \leq 2 + \frac{k_3\alpha^{\gamma_3}}{1-k_4\alpha^{\gamma_4}}\left[2\frac{-\ln\beta}{I_-}\right] \leq \\ \leq 2 + k_6\alpha^{\gamma_2}, \tag{25}$$

причем k_6 не зависит от α и распределения f из G_{i_0}.

Подставляя (24) и (25) в (18) получаем утверждение (16).

6. Заключение

В работе [5] показано, что асимптотически оптимальная стратегия должна использовать как минимум два этапа. Поэтому сравнивая полученную в теореме 2 оценку с соответствующей оценкой из [1] видим, что процедура d_0 является субоптимальной.

Если $Mc^{-1} << \varepsilon \ln|\alpha|$, то введение ограничения на принятие решения, связанное с тем, что оно принимается только после завершения очередного этапа, оказывается несущественным, поскольку его вклад меньше того, который вносит погрешность в определении распределения наблюдений.

Если $Mc^{-1} >> \ln|\alpha|$, то стоимость, обусловленная введением этапов, оказывается главной и в этом случае следует ограничиться фактически статической процедурой, когда используется только третий этап описанной выше процедуры.

7. Литература

1. *Цитович Ф.И.* Некоторые субоптимальные последовательные правила проверки гитопез // Сборник трудов 30-й конференции молодых уччных и специалистов ИППИ РАН: Информационные технологии и системы ИТиС'07. М.: ИППИ, 2007. С. 110-115.

2. *Малютов М.Б., Цитович И.И.* Асимптотически оптимальная последовательная проверка гипотез // Проблемы передачи информации. Том 36. Вып.4. 2000. С 98–112.

3. *Malyutov, M.B., Tsitovich, I.I.*. Second order optimal sequential tests // Proc. Internat. Workshop Optimum Design 2000, Cardiff, UK, April 2000. Kluwer, Netherlands. 2000. P. 67–78.

4. *Bartroff J.* Asymptotically optimal multistage tests of simple hypotheses // Ann. Statist. 2007. V. 35. N 4. P. 2075-2105.

5. *Lorden G.* Asymptotic efficiency of three-stage hypothesis tests // Ann. Statist. 1983. V. 11. P. 129–140.

Игровые стратегии и адаптивное прогнозирование

В.В.Вьюгин, Е.Л.Румянцев, Ф.М.Цфасман
ИППИ им. А.А.Харкевича РАН, Б.Каретный пер. 19, Москва, 127994, Россия, vyugin@iitp.ru

Аннотация

Предложен вариант алгоритма универсального прогнозирования Фостера, Вохра и Какаде. Разработаны стратегии игры на фондовом рынке, использующие прогнозы этого алгоритма. Приведены результаты экспериментов на реальных данных.

1 Введение

Рассматривается следующая задача прогнозирования: предсказатель получает в режиме он-лайн некоторую числовую последовательность исходов $\omega_1, \omega_2, \ldots, \omega_{n-1}, \ldots$, при этом его задачей является вычисление оценки p_n будущего события ω_n по известным $n-1$ исходам. Число p_n может рассматриваться как прогноз среднего значения будущего исхода или как вероятность того, что $\omega_n = 1$, в случае если ω_i принимают значения 0 или 1.

Типичным примером является задача предсказания погоды на завтра, например, событие $\omega_n = 1$ может интерпретироваться как дождь в n-ый день, а число p_n - как его вероятность, вычисленная на основе наблюдений погоды $\omega_1, \omega_2, \ldots, \omega_{n-1}$ за предыдущие $n-1$ дней. Предсказатель погоды считается хорошо калибруемым, если дождь случается также часто, как он прогнозируется предсказателем. Например, если дождь случается в 80% всех дней, для которых предсказатель давал прогноз $p_n = 0.8$ и т.д. В этом примере $p_n \in [0, 1]$.

Если $\omega_n = S_{n+1} - S_n$ - это приращения цен некоторого финансового инструмента в некоторые последовательные моменты времени $n = 1, 2, \ldots$, то число p_n может рассматриваться как прогноз среднего значения этого приращения. В этом примере p_n может принимать отрицательные значения.

При вероятностном подходе мы предполагаем, что имеется некоторая вероятностная мера на $\omega_1, \omega_2, \ldots, \omega_n$, тогда $p_n = E(\omega_n | \omega_1, \omega_2, \ldots, \omega_{n-1})$, где E - условное математическое ожидание случайной величины ω_n относительно $\omega_1, \omega_2, \ldots, \omega_{n-1}$ по этой мере. Однако на практике мы часто имеем дело с единственной исторической последовательностью исходов $\omega_1, \omega_2, \ldots, \omega_{n-1}$, и меру, описывающую будущие значения величины ω_n довольно трудно определить. В данной работе предположение о наличии такой меры не используется. Тем не менее, можно указать метод прогнозирования произвольной последовательнсти $\omega_1, \omega_2, \ldots, \omega_{n-1}$, удовлетворяющий требованию *калибруемости*.

В данной работе мы рассматриваем последовательности чисел $\omega_n \in [-1, 1]$ для всех n. Пусть Ξ - множество всех конечных последовательностей действительных чисел из интервала $[-1, 1]$. Под детерминированной *предсказательной системой* понимается произвольная частичная функция $f : \Xi \to [-1, 1]$.

Приведем точное определение калибруемости, предложенное Дейвидом [1]. Предсказательная система f *калибруется* на бесконечной последовательности $\omega_1\omega_2\ldots$, если для каждой характеристической функции $I(p)$ подинтервала $[-1, 1]$ *калибровочная ошибка* стремится к нулю, т.е.

$$\frac{\sum_{i=1}^{n} I(p_i)(\omega_i - p_i)}{\sum_{i=1}^{n} I(p_i)} \longrightarrow 0, \qquad (1)$$

если знаменатель отношения (1) стремится к бесконечности при $n \to \infty$, где $p_i = f(\omega^{i-1})$. Характеристическая функция $I(p_i)$ определяет некоторое правило выбора, которое определяет те номера исходов i, для которых мы вычисляем отклонение прогноза p_i от соответствующего исхода ω_i.

Под *рандомизированной* предказательной системой понимается случайная величина $f(\alpha; x)$, определенная на некотором вероятностном пространстве Ω_x снабженным некоторым распределением вероятностей Pr_x, где $x \in \Xi$ - параметр. Далее, как обычно, мы опускаем переменную α. Для каждой бесконечной последовательности ω (которая в данном случае является парметром), вероятностные распределения $Pr_{\omega^{i-1}}$ порождают распределение вероятностей Pr на прямом

произведении вероятностных пространств $\Omega_{\omega^{i-1}}$, $i = 1, 2, \ldots$. В этом случае, при фиксированной последовательности ω можно рассматривать вероятность Pr события (1).

Фостер и Вохра [2], а также Какаде и Фостер [3], определили для каждого $\Delta > 0$ рандомизированную предсказательную систему f такую, что для произвольной бесконечной последовательности $\omega = \omega_1\omega_2\ldots$ и характеристической функции $I(p)$ произвольного подинтервала $[-1, 1]$ вероятность Pr события

$$\left|\frac{1}{n}\sum_{i=1}^{n} I(\tilde{p}_i)(\omega_i - \tilde{p}_i)\right| \leq \Delta$$

стремится к 1 при $n \to \infty$, где $\tilde{p}_i = f(\omega_1, \ldots, \omega_{n-1})$ - случайная величина.

2 Алгоритм вычисления калибруемых прогнозов

Приведем адаптированный вариант рандомизированного алгоритма Какаде и Фостера. Пусть $\omega_1\omega_1\ldots$ - произвольная последовательность действительных чисел из $[-1, 1]$, поступающая в режиме он-лайн. Построим алгоритм для вычисления случайной величины, выдающей прогноз $p_n \in [-1, 1]$ будущего значения $\omega_n \in [-1, 1]$ по начальному фрагменту $\omega_1 \ldots, \omega_{n-1}$. Основное требование к таким прогнозам - они должны с вероятностью 1 удовлетворять условию калибруемости. Соответствующее распределение вероятностей является внутренним по отношению к алгоритму и строится в процессе конструкции.

Предварительно разобъем интервал $[-1, 1]$ на равные части длины $\Delta = 2/K$ с помощью рациональных точек $v_i = i\Delta$, $i = 0, 1, \ldots, K$. Пусть V обозначает множество всех этих точек. Любое число $p \in [-1, 1]$ представляется в виде линейной комбинации соседних точек подинтервала разбиения, содержащего p,

$$p = \sum_{v\in V} w_v(p)v = w_{v_{i-1}}(p)v_{i-1} + w_{v_i}(p)v_i,$$

где $p \in [v_{i-1}, v_i]$, $i = \lfloor p/\Delta + 1\rfloor$, и

$$w_{v_{i-1}}(p) = 1 - \frac{p - v_{i-1}}{\Delta},\ w_{v_i}(p) = 1 - \frac{v_i - p}{\Delta}.$$

Полагаем $w_v(p) = 0$ для всех остальных значений $v \in V$.

В дальнейшем, детерминированный прогноз p, выдаваемый алгоритмом приведенным далее, будет округляться до v_{i-1} с вероятностью $w_{v_{i-1}}(p)$ и до v_i с вероятностью $w_{v_i}(p)$.

Построим детерминированный алгоритм, выдающий прогноз p. Пусть прогнозы $p_1, \ldots, p_{n-1}$ уже определены (полагаем $p_1 = 0$). Построим прогноз p_n.

Рассмотрим вспомогательную величину

$$\mu_{n-1}(v) = \sum_{i=1}^{n-1} w_v(p_i)(\omega_i - p_i).$$

Имеем

$$\begin{aligned}(\mu_n(v))^2 = (\mu_{n-1}(v))^2 + \\ 2w_v(p_n)\mu_{n-1}(v)(\omega_n - p_n) + \\ (w_v(p_n))^2(\omega_n - p_n)^2.\end{aligned} \quad (2)$$

Суммируем (2) по v

$$\begin{aligned}\sum_{v\in V}(\mu_n(v))^2 = \sum_{v\in V}(\mu_{n-1}(v))^2 + \\ 2(\omega_n - p_n)\sum_{v\in V} w_v(p_n)\mu_{n-1}(v) + \\ \sum_{v\in V}(w_v(p_n))^2(\omega_n - p_n)^2.\end{aligned} \quad (3)$$

Изменим порядок суммирования в сумме вспомогательных величин

$$\begin{aligned}\sum_{v\in V} w_v(p)\mu_{n-1}(v) = \\ \sum_{v\in V} w_v(p)\sum_{i=1}^{n-1} w_v(p_i)(\omega_i - p_i) = \\ \sum_{i=1}^{n-1}(\sum_{v\in V} w_v(p)w_v(p_i))(\omega_i - p_i) = \\ \sum_{i=1}^{n-1}(\bar{w}(p), \bar{w}(p_i))(\omega_i - p_i) = \sum_{i=1}^{n-1} K(p, p_i)(\omega_i - p_i),\end{aligned}$$

где $\bar{w}(p) = (w_1, \ldots, w_{v_K}) = (0, \ldots, w_{v_{i-1}}(p), w_{v_i}(p), \ldots, 0)$ - вектор вероятностей округления, $p \in [v_{i-1}, v_i]$, и скалярное произведение

$$K(p, p_i) = (\bar{w}(p), \bar{w}(p_i)) \quad (4)$$

соответствующих векторов (ядро). По определению - $K(p, p_i)$ - непрерывная функция (точнее, кусочно-линейная).

Второй член правой части суммы (3) при подходящем значении p_n всегда можно сделать меньшим или равным нулю. Действительно, в качестве p_n берем корень $p_n = p$ уравнения

$$\sum_{v\in V} w_v(p)\mu_{n-1}(v) = \sum_{i=1}^{n-1} K(p, p_i)(\omega_i - p_i) = 0, \quad (5)$$

если он существует. В противном случае, если левая часть уравнения (5) (которая является непрерывной по p функцией) больше нуля для всех значений p_n, то полагаем $p_n = 1$, если она меньше нуля, то полагаем $p_n = -1$. Таким образом определенное значение p_n выдаем в качестве детерминированного прогноза.

Третий член (3) ограничен числом 4. Действительно, так как $|\omega_i - p_i| \le 2$ для всех i, имеем

$$\sum_{v\in V}(w_v(p_n))^2(\omega_i - p_n)^2 \le 4\sum_{v\in V} w_v(p_n) = 4.$$

Отсюда и по (3), если последовательно выбирать прогнозы p_i согласно указанному правилу, получим

$$\sum_{v\in V}(\mu_n(v))^2 \le \sum_{i=1}^{n}\sum_{v\in V}(w_v(p_i))^2(\omega_i - p_i))^2 \le 4n.$$

Так как сумма квадратов не превосходит $4n$, каждый член этой суммы также не превосходит $4n$. Поэтому для каждого $v \in V$ имеет место неравенство

$$|\sum_{i=1}^{n} w_v(p_i)(\omega_i - p_i))| \le 2\sqrt{n}. \qquad (6)$$

Пусть теперь $\tilde{p}_i$ - случайная величина, принимающая значения $v \in V$ с вероятностями $w_v(p_i)$ (на самом деле, для каждого p ненулевыми являются только значения $w_v(p)$ для двух соседних границ подинтервала разбиения, содержащего детерминированный прогноз p_i). Пусть также $I(p)$ - характеристическая функция произвольного интервала. Для любого i математическое ожидание случайной величины $I(\tilde{p}_i)(\omega_i - \tilde{p}_i)$ равно

$$E(I(\tilde{p}_i)(\omega_i - \tilde{p}_i)) = \sum_{v\in V} w_v(p_i)I(v)(\omega_i - v). \qquad (7)$$

Согласно усиленному закону больших чисел, Pr-вероятность события

$$|\sum_{i=1}^{n} I(\tilde{p}_i)(\omega_i - \tilde{p}_i) - \sum_{i=1}^{n} E(I(\tilde{p}_i)(\omega_i - \tilde{p}_i))| \le \sqrt{n\log n} \qquad (8)$$

стремится к 1 при $n \to \infty$.

По определению детерминированного прогноза p_i и функции $w_v(p)$

$$|\sum_{v\in V} w_v(p_i)I(v)(\omega_i - v) - \sum_{v\in V} w_v(p_i)I(v)(\omega_i - p_i)| < \Delta \qquad (9)$$

для каждого i.

Так как (6) имеет место для каждого $v \in V$, суммируя (9) по $i = 1, \ldots, n$, получаем верхнюю оценку для абсолютной величины суммы математических ожиданий (7)

$$|\sum_{i=1}^{n} E(I(\tilde{p}_i)(\omega_i - \tilde{p}_i))| =$$

$$|\sum_{i=1}^{n}\sum_{v\in V} w_v(p_i)I(v)(\omega_i - v)| < \Delta n + 2|V|\sqrt{n} \qquad (10)$$

для всех n, где $|V| = 2/\Delta$ - число подинтервалов разбиения.

Из (10) и (8) получаем, что Pr-вероятность события

$$\left|\sum_{i=1}^{n} I(\tilde{p}_i)(\omega_i - \tilde{p}_i)\right| \le \Delta n + \sqrt{n\log^2 n} + 4\sqrt{n}/\Delta \qquad (11)$$

стремится к 1 при $n \to \infty$. В частности, Pr-вероятность события

$$\left|\frac{1}{n}\sum_{i=1}^{n} I(\tilde{p}_i)(\omega_i - \tilde{p}_i)\right| \le \Delta \qquad (12)$$

стремится к 1 при $n \to \infty$.

3 Игра на поиск тренда в подпоследовательности

Рассмотрим игру на повышение (понижение) на фондовом рынке. Пусть $S_1, S_2, \ldots$ - последовательность цен некоторого финансового инструмента в дискретные моменты времени, $\Delta S_i = S_{i+1} - S_i$ - их приращения. Предполагаем, что эти приращения нормированы так, что $\Delta S_i \in [-1, 1]$ при всех $i = 1, 2, \ldots$.

FOR $i = 1, 2 \ldots$

Участники рынка наблюдают цены финансового инструмента $S_1, \ldots, S_i$.

Предсказатель предлагает вероятностный прогноз приращения ΔS_i финансового инструмента S_i - случайную величину $\tilde{p}_i \in [-1, 1]$.

Инвестор определяет количество M_i приобретаемых или продаваемых единиц инструмента. Инвестор может пропустить данный шаг игры полагая $M_i = 0$.

Рынок объявляет цену S_{i+1}.

Инвестор подсчитывает изменение своего капитала

$$\Delta\mathcal{K}_i = M_i\Delta S_i.$$

ENDFOR

Здесь был рассмотрен случай $M_i \geq 0$, что соответствует длинной позиции при игре с акциями. Фондовый рынок также допускает случай $M_i < 0$, что соответсвует игре в короткой позиции (в этом случае, дальнейшие рассуждения аналогичны).

В момент времени n наблюдаем значения $\Delta S_1, \ldots, \Delta S_{n-1}$ и вычисляем вероятностный прогноз p_n для ΔS_n с использованием алгоритма универсального прогнозирования.

Предварительно рассмотрим игру с одной акцией. Пусть число $\epsilon > 0$ определяет условие входа в игру (правило выбора подпоследовательности для игры). На шаге n входим в игру, т.е. покупаем $M_n = 1$ акцию и продаем ее в конце этого периода, если $p_i > \epsilon$; в противном случае, пропускаем этот шаг.

В конце торгового периода получаем доход или убыток в размере ΔS_i на одну акцию. Суммарный доход на одну акцию за n шагов игры равен

$$\mathcal{K}_n = \sum_{i=1}^{n} I(\tilde{p}_i > \epsilon)\Delta S_i.$$

Будет рассматривать точность округления Δ как функцию от n, т.е., $\Delta = \Delta_n$.

Из (11) получаем, что вероятность события

$$\mathcal{K}_n = \sum_{\tilde{p}_i > \epsilon} \Delta S_i = \sum_{i=1}^{n} I(\tilde{p}_i > \epsilon)(\Delta S_i - \tilde{p}_i) + \sum_{i=1}^{n} I(\tilde{p}_i > \epsilon)p_i \geq \quad (13)$$
$$\epsilon \sum_{i=1}^{n} I(\tilde{p}_i > \epsilon) - (\Delta_n n + \sqrt{n \log n} + 2\sqrt{n}/\Delta_n)$$

стремится к 1, при $n \to \infty$. Здесь и далее $I(\tilde{p}_i > \epsilon)$ есть конкретное правило выбора, которое в предыдущем разделе обозначалось $I(\tilde{p}_i)$.

Отсюда, вероятность того, что средний доход от одной акции (на один вход на рынок) удовлетворяет неравенству

$$k_n = \frac{\mathcal{K}_n}{\sum_{i=1}^{n} I(\tilde{p}_i > \epsilon)} \geq \left(\epsilon - \frac{\Delta_n n + \sqrt{n \log n} + 2\sqrt{n}/\Delta_n}{\sum_{i=1}^{n} I(\tilde{p}_i > \epsilon)} \right) \quad (14)$$

стремится к 1, при $n \to \infty$.

В частности, для произвольного μ, $0 < \mu < 1$, вероятность того, что событие

$$\sum_{i=1}^{n} I(\tilde{p}_i > \epsilon) \geq \frac{\Delta n + \sqrt{n \log n} + 2\sqrt{n}/\Delta}{\mu\epsilon} \quad (15)$$

влечет событие $k_n \geq (1-\mu)\epsilon$ стремится к 1 при $n \to \infty$. Оптимальное значение $\Delta_n = \sqrt{2}n^{-\frac{1}{4}}$. Тогда условие (15) можно заменить на

$$\sum_{i=1}^{n} I(\tilde{p}_i > \epsilon) \geq \frac{cn^{\frac{3}{4}}}{\mu\epsilon}, \quad (16)$$

где c - некоторая константа. Оценка (16) не является оптимальной и требует улучшения. Число ϵ - определяет условие выбора подпоследовательности входов в игру. Значение ϵ выбирает инвестор.

Найдется ли достаточно длинная подпоследовательность входов, удовлетворяющая (15) или (16) зависит от величины отклонения исходной последовательности цен $S_1, S_2, \ldots$ от модели броуновского движения.

4 Игровая стратегия с ограниченным снизу риском

Стратегия игры с одной акцией может быть легко модифицирована в стратегию с ограниченным снизу риском, если выбирать число акций M_i зависящим от текущего капитала ([4]). Пусть начальный капитал инвестора составлял $\mathcal{K}_0$ единиц. Пусть также $0 < \delta < 1$. Определим $M_i = \delta\mathcal{K}_i$, где $\mathcal{K}_i$ - капитал инвестора на i-ом входе в игру. Тогда изменение капитала на i-входе в игру определяется соотношением

$$\mathcal{K}_{i+1} = \mathcal{K}_i + \delta\mathcal{K}_i\Delta S_i. \quad (17)$$

Так как $|\Delta S_i| \leq 1$ для всех i по основному предположению, получаем, что вероятность события

$$\ln \mathcal{K}_n = \ln \mathcal{K}_0 + \sum_{i=0}^{n-1} I(\tilde{p}_i > \epsilon) \ln(1 + \delta\Delta S_i) \geq$$
$$\ln \mathcal{K}_0 + \delta \sum_{i=0}^{n-1} I(p_i > \epsilon)\Delta S_i -$$
$$\delta^2 \sum_{i=0}^{n-1} I(p_i > \epsilon)(\Delta S_i)^2 \geq$$
$$\ln \mathcal{K}_0 + \delta(\epsilon \sum_{i=1}^{n} I(\tilde{p}_i > \epsilon) - cn^{\frac{3}{4}})$$
$$-\delta^2 \sum_{i=0}^{n-1} I(\tilde{p}_i > \epsilon) \quad (18)$$

стремится к 1 при $n \to \infty$. Последнее неравенство следует из и (14). Мы также использовали неравенство $\ln(1+x) \geq x - x^2$ при $|x| \leq 1$.

Отсюда получаем, что вероятность того, что суммарный доход при такой игре удовлетворяет неравенству

$$\mathcal{K}_n = \mathcal{K}_0 \prod_{i=0}^{n-1} \sum_{i=0}^{n-1} I(\tilde{p}_i > \epsilon)(1 + \delta \Delta S_i) \geq \mathcal{K}_0 e^{\delta \sum_{i=0}^{n-1} I(\tilde{p}_i > \epsilon) \left((\epsilon - \delta) - \frac{cn^{\frac{3}{4}}}{\sum_{i=0}^{n-1} I(\tilde{p}_i > \epsilon)} \right)} \quad (19)$$

стремится к 1. Из получаем, что для произвольного μ, $0 < \mu < 1$, вероятность того, что событие

$$\sum_{i=0}^{n-1} I(\tilde{p}_i > \epsilon) \geq \frac{cn^{\frac{3}{4}}}{\mu(\epsilon - \delta)}$$

влечет событие

$$\mathcal{K}_n \geq \mathcal{K}_0 e^{\delta(\epsilon-\delta)(1-\mu) \sum_{i=0}^{n-1} I(\tilde{p}_i > \epsilon)} \geq \mathcal{K}_0 e^{\delta(\frac{1}{\mu} - 1) cn^{\frac{3}{4}}} \quad (20)$$

стремится к 1 при $n \to \infty$. В этих оценках мы использовали условие $\epsilon > \delta$.

Правая часть (20) всегда положительна, поэтому в игре можно потерять только начальный капитал $\mathcal{K}_0$, т.е. $\mathcal{K}_n > 0$ для любой последовательности цен S_i такой, что $|\Delta S_i| \leq 1$ для всех i.

5 Результаты расчетов

На рис. 1 приведены прогнозы средних значений приращений цен акции ГАЗПРОМа. В качестве ядра использована функция $K(p, p_i) = e^{-c(p_i - p)^2}$, где $c > 0$.

В таблице 1 и на рис. 2-3 (ниже) приводятся результаты экспериментов игры на основе некоторого эвристического алгоритма, построенного на основе идей раздела 3. В качестве исходных данных использовались поминутные цены акций российских компаний ГАЗПРОМ, ЛУКОЙЛ, МТС, Роснефть, Сбербанк, РАО ЕС за период с июля по декабрь 2007 г., взятые из сайта *www.finam.ru*. В первой колонке приведен тикер-символ компании, во второй - период игры, в третьей шаг игры, в четвертой - количество акций, используемых в тире, в пятой - средняя цена акции за период игры, в пятой - полученный доход (с учетом выплаченной комиссии за участие в игре). Необычно большой доход в игре

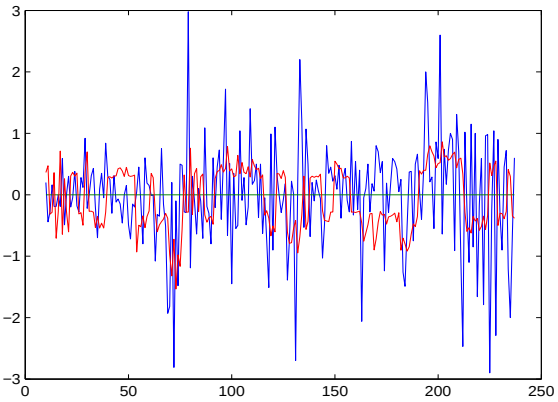


Рис. 1. Пример подпоследовательности приращений цен акции ГАЗПРОМа и прогнозов их средних значений. Было отобрано 237 значений из общего числа 1000. Отбирались те моменты i, для которых прогноз $p_i > \epsilon$ или $p_i < -\epsilon$, где $\epsilon > 0$.

с акциями Сбербанка связан с удачным входом на 14706 минуте игры, когда цена акции выросла с 46400 до 52200 руб. Горизонтальные отрезки на графиках соответствуют достаточно длинным интервалам времени, когда входов в игру не было и капитал не изменялся. Эксперименты проводились только для игры в длинной позиции, т.е когда $M_i \geq 0$ (см. раздел 3).

Отклонение ϵ выбиралось динамически, с учетом значения скользящего среднего квадратичного отклонения цены финансового инструмента. В качестве ядра использовалась функция $K(p, p_i) = \cos(\pi(p_i - p))$ (см. также [4]).

Список литературы

[1] Dawid, A.P.: Calibration-based empirical probability [with discussion]. Ann. Statist. 13, 1251-1285 (1985)

[2] Foster, D.P., Vohra, R.: Asymptotic calibration. Biometrika. 85, 379–390 (1998)

[3] Kakade, S.M., Foster, D.P.: Deterministic calibration and Nash equilibrium. LNCS 3120 (John Shawe Taylor and Yoram Singer. ed). 33–48 (2004)

[4] Vovk, Vladimir, Takemura, Akimichi, Shafer, Glenn: Defensive Forecasting. Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics. 365–372 (2005) (http://arxiv.org/abs/cs/0505083)

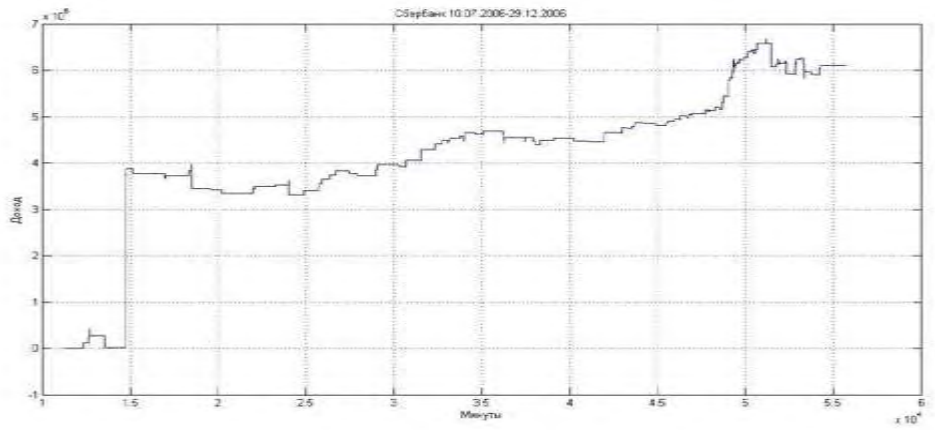


Рис. 2.

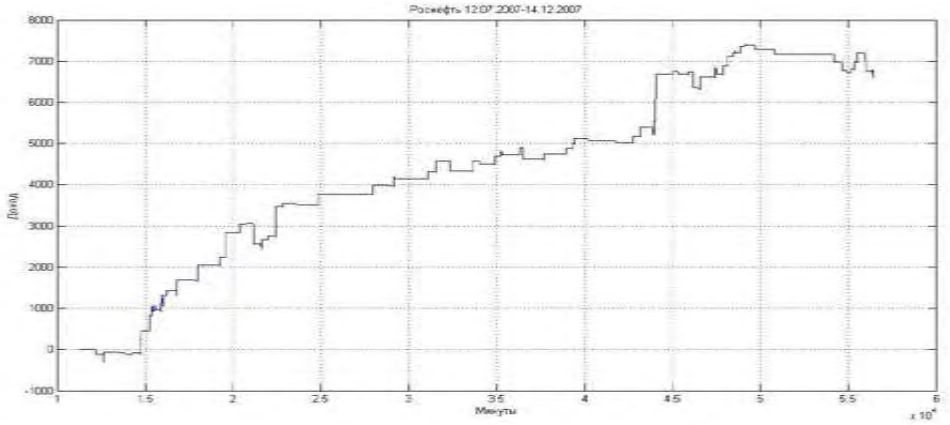


Рис. 3.

Таблица 1. Результаты игры на нескольких финансовых инструментах российского фондового рынка.

Акция	Период	Шаг	Кол-во акций	Средняя цена акции	Доход
ГАЗПРОМ ао	2007.07.09-2007.12.03	мин.	500	284	661
ЛУКОЙЛ	2007.07.10-2007.12.04	мин.	500	2063	13500
МТС ао	2007.07.27-2007.12.28	мин.	500	314	21247
Роснефть	2007.07.12-2007.12.14	мин.	500	215	6602
Сбербанк	2006.07.10-2006.12.29	мин.	500	60723	6103575
РАО ЕЭС	2007.07.09-2007.12.04	мин.	500	31.8	485

Уравнение Пенлеве VI и дополнительные особенности фуксовых уравнений

Р. Р. Гонцов
ИППИ РАН
rgontsov@inbox.ru

Аннотация

В работе рассматриваются проблема Римана—Гильберта для фуксовых систем линейных дифференциальных уравнений, ее аналог для скалярных фуксовых уравнений и их связь с уравнением Пенлеве VI.

1. Введение

В середине XIX века Б. Риман рассмотрел задачу о построении линейного дифференциального уравнения

$$\frac{d^p u}{dz^p} + b_1(z)\frac{d^{p-1}u}{dz^{p-1}} + \ldots + b_p(z)u = 0 \qquad (1)$$

с заданными регулярными особыми точками $a_1, \ldots, a_n \in \overline{\mathbb{C}}$ (полюсами коэффициентов) и заданной монодромией (см. [8]).

Напомним, что особая точка a_i уравнения (1) называется *регулярной*, если в ее окрестности любое решение уравнения имеет не более чем степенной рост. Существует простой критерий регулярности особой точки уравнения (1), полученный Л. Фуксом [9]: *особенность a_i уравнения (1) является регулярной тогда и только тогда, когда коэффициент $b_j(z)$ имеет в этой точке полюс порядка не более j ($j = 1, \ldots, p$).* Линейные дифференциальные уравнения с регулярными особыми точками также называются *фуксовыми*.

Монодромия линейного дифференциального уравнения характеризует ветвление его решений в особых точках и определяется следующим образом. В окрестности неособой точки z_0 рассмотрим фундаментальную систему решений $u_1, \ldots, u_p$ уравнения (1). Результатом аналитического продолжения этих функций вдоль произвольной петли γ, начинающейся в точке z_0 и лежащей в $\overline{\mathbb{C}} \setminus \{a_1, \ldots, a_n\}$, будет, вообще говоря, другая фундаментальная система решений $\tilde{u}_1, \ldots, \tilde{u}_p$. При этом наборы функций $(u_1, \ldots, u_p)$, $(\tilde{u}_1, \ldots, \tilde{u}_p)$ как два базиса в пространстве решений линейного дифференциального уравнения связаны с помощью невырожденной матрицы перехода G_γ, соответствующей петле γ: $(u_1, \ldots, u_p) = (\tilde{u}_1, \ldots, \tilde{u}_p)G_\gamma$. Отображение $[\gamma] \mapsto G_\gamma$ (которое зависит только от гомотопического класса $[\gamma]$ петли γ) задает представление $\chi : \pi_1(\overline{\mathbb{C}} \setminus \{a_1, \ldots, a_n\}, z_0) \longrightarrow \mathrm{G}$ фундаментальной группы пространства $\overline{\mathbb{C}} \setminus \{a_1, \ldots, a_n\}$ в пространство $\mathrm{G} = GL(p, \mathbb{C})$ невырожденных комплексных матриц порядка p. Это представление и называется *монодромией* уравнения (1).

Матрицей монодромии уравнения (1) в особой точке a_i (относительно фундаментальной системы решений $u_1, \ldots, u_p$) называется матрица G_i, соответствующая простой петле γ_i, обходящей точку a_i, т. е. $G_i = \chi([\gamma_i])$. Матрицы $G_1, \ldots, G_n$ являются образующими *группы монодромии* $\mathrm{Im}\,\chi$ и в силу условия $\gamma_1 \ldots \gamma_n = e$ в фундаментальной группе связаны соотношением $G_1 \ldots G_n = I$ (здесь и далее I обозначает единичную матрицу).

Если изначально вместо фундаментальной системы решений $(u_1, \ldots, u_p)$ рассмотреть другую фундаментальную систему решений $(u_1', \ldots, u_p') = (u_1, \ldots, u_p)C$, $C \in \mathrm{G}$, то соответствующие ей матрицы монодромии будут иметь вид $G_i' = C^{-1}G_iC$. Аналогичным образом устроена и зависимость матриц G_i от выбора начальной точки z_0. Таким образом, монодромия линейного дифференциального уравнения определена с точностью до сопряжения на постоянную невырожденную матрицу и, говоря точнее, является элементом пространства $\mathcal{M}_a = \mathrm{Hom}\left(\pi_1(\overline{\mathbb{C}} \setminus \{a_1, \ldots, a_n\}), \mathrm{G}\right)/\mathrm{G}$ классов сопряженности представлений группы $\pi_1(\overline{\mathbb{C}} \setminus \{a_1, \ldots, a_n\})$.

А. Пуанкаре [7] было показано, что число параметров, от которых зависит фуксово уравнение порядка p с n особыми точками, меньше размерности пространства $\mathcal{M}_a$ представлений монодромии (см. также [4], лекция 8). Поэтому в общем случае при построении фуксова уравнения с заданной монодромией необходимо возникают дополнительные (помимо $a_1, \ldots, a_n$) *ложные* особые точки, в которых коэффициенты уравнения имеют особенности, но решения являются однозначными мероморфными функциями и, следовательно, матрицы монодромии в этих точках единичны. (В дальнейшем под дополнительными особыми точками уравнения

будем понимать именно такие особенности.) Таким образом, в общем случае задача Римана имеет отрицательное решение.

Аналогичная задача для систем линейных дифференциальных уравнений называется проблемой Римана—Гильберта и состоит в построении *фуксовой системы*[1]

$$\frac{dy}{dz} = \left(\sum_{i=1}^{n} \frac{B_i}{z - a_i} \right) y, \tag{2}$$

где $y(z) \in \mathbb{C}^p$, $B_i \in \mathrm{Mat}(p, \mathbb{C})$, p линейных дифференциальных уравнений с заданными особыми точками $a_1, \ldots, a_n$ (если бесконечность не входит в число особенностей системы, то $\sum_{i=1}^{n} B_i = 0$) и монодромией

$$\chi : \pi_1(\overline{\mathbb{C}} \setminus \{a_1, \ldots, a_n\}, z_0) \longrightarrow \mathrm{G}. \tag{3}$$

Монодромия системы линейных дифференциальных уравнений определяется так же, как и для скалярного уравнения (1), нужно только вместо фундаментальной системы решений уравнения рассмотреть фундаментальную матрицу $Y(z)$ системы — матрицу, столбцы которой образуют базис в пространстве решений системы.

Контрпример к проблеме Римана—Гильберта был получен А. А. Болибрухом [2]. История решения этой проблемы намного сложнее, чем ее аналога для скалярных фуксовых уравнений (до А. А. Болибруха долгое время она ошибочно считалась решенной в положительном смысле; подробнее см. в [1]).

В данной работе мы расскажем об интересной взаимосвязи между, с одной стороны, проблемой Римана—Гильберта и ее аналогом для скалярных фуксовых уравнений, и известным нелинейным дифференциальным уравнением Пенлеве VI — с другой.

2. Метод решения проблемы Римана—Гильберта

При исследовании задач, связанных с проблемой Римана—Гильберта, весьма полезным инструментом служат линейные калибровочные преобразования вида

$$y' = \Gamma(z)\, y \tag{4}$$

неизвестной функции $y(z)$. Данное преобразование называется *голоморфно (мероморфно) обратимым* в некоторой точке z_0, если матрица $\Gamma(z)$ голоморфна (мероморфна) в этой точке и $\det \Gamma(z_0) \neq 0$ ($\det \Gamma(z) \not\equiv 0$). Преобразование (4) переводит систему (2) в систему с матрицей коэффициентов

$$B'(z) = \frac{d\Gamma}{dz}\Gamma^{-1} + \Gamma \left(\sum_{i=1}^{n} \frac{B_i}{z - a_i} \right) \Gamma^{-1}, \tag{5}$$

которая называется соответственно *голоморфно* или *мероморфно эквивалентной* исходной системе в окрестности точки z_0.

Важным свойством мероморфных калибровочных преобразований является то, что они сохраняют монодромию.

Локально, в окрестности каждой точки a_k, нетрудно предъявить систему, для которой a_k была бы фуксовой особенностью, а матрица монодромии в этой точке совпадала бы с соответствующей образующей $G_k = \chi([\gamma_k])$ представления (3). Это — система

$$\frac{dy}{dz} = \frac{E_k}{z - a_k}\, y, \qquad E_k = \frac{1}{2\pi i} \ln G_k, \tag{6}$$

с фундаментальной матрицей $(z - a_k)^{E_k}$. (По определению $(z - a_k)^{E_k} = e^{E_k \ln(z - a_k)}$, а ветвь логарифма матрицы G_k выбирается таким образом, чтобы собственные значения ρ_k^j матрицы E_k удовлетворяли условию $0 \leqslant \mathrm{Re}\, \rho_k^j < 1$.)

Пусть $\Lambda_k = \mathrm{diag}(\lambda_k^1, \ldots, \lambda_k^p)$ — диагональная целочисленная матрица, элементы λ_k^j которой образуют невозрастающую последовательность, а S_k — невырожденная матрица, приводящая матрицу E_k к верхнетреугольному виду $E_k' = S_k E_k S_k^{-1}$. Тогда, согласно (5), преобразование

$$y' = \Gamma(z)\, y, \quad \Gamma(z) = (z - a_k)^{\Lambda_k} S_k,$$

переводит систему (6) в систему с матрицей коэффициентов

$$\frac{\Lambda_k}{z - a_k} + (z - a_k)^{\Lambda_k} \frac{E_k'}{z - a_k} (z - a_k)^{-\Lambda_k}, \tag{7}$$

для которой точка a_k является также фуксовой особенностью[2], а матрица G_k — матрицей монодромии.

Согласно теореме Левеля [5] любая фуксова система в окрестности особой точки a_k *голоморфно* эквивалентна системе с матрицей коэффициентов вида (7). В то же время, класс *мероморфной* эквивалентности системы в окрестности регулярной (в частности, фуксовой) особой точки a_k однозначно определяется матрицей G_k ее локальной монодромии, поскольку такая система мероморфно эквивалентна системе вида (6).

[1] Заметим, что по теореме Соважа особые точки фуксовой системы являются регулярными особенностями (см., например, [4], теорема 4.1).

[2] Из вида матриц Λ_k и E_k' следует голоморфность матрицы $(z - a_k)^{\Lambda_k} E_k' (z - a_k)^{-\Lambda_k}$.

Будем называть набор $\{\Lambda_1, \ldots, \Lambda_n, S_1, \ldots, S_n\}$ матриц, обладающих описанными выше свойствами, набором *допустимых матриц*.

Проблема Римана—Гильберта решается положительно, если от локальных систем с матрицами коэффициентов (7) удается перейти к глобальной фуксовой системе, заданной на всей сфере Римана. При исследовании этого вопроса эффективным оказывается использование голоморфных векторных расслоений и логарифмических (фуксовых) связностей. Опишем вкратце данный подход (подробнее см. в [4]).

1. Сначала по представлению (3) над проколотой сферой Римана $B = \overline{\mathbb{C}} \setminus \{a_1, \ldots, a_n\}$ строится голоморфное векторное расслоение F ранга p с голоморфной связностью ∇, имеющей заданную монодромию (3). Расслоение F над B получается из голоморфно тривиального расслоения $\widetilde{B} \times \mathbb{C}^p$ над универсальным накрытием $\widetilde{B}$ проколотой сферы Римана после отождествлений вида $(\tilde{z}, y) \sim (\sigma\tilde{z}, \chi(\sigma)y)$, где $\tilde{z} \in \widetilde{B}$, $y \in \mathbb{C}^p$ и σ — элементы группы накрывающих преобразований пространства $\widetilde{B}$, отождествленной с фундаментальной группой $\pi_1(B)$. Таким образом, $F = \widetilde{B} \times \mathbb{C}^p / \sim$ и $\pi : F \longrightarrow B$ — естественная проекция. Нетрудно показать, что при некотором выборе покрытия $\{U_\alpha\}$ проколотой сферы Римана склеивающий коцикл $\{g_{\alpha\beta}\}$ расслоения F будет задаваться постоянными матрицами $g_{\alpha\beta}$.

Связность ∇ можно задать набором $\{\omega_\alpha\}$ матричных дифференциальных 1-форм $\omega_\alpha \equiv 0$, на пересечениях $U_\alpha \cap U_\beta \neq \varnothing$ очевидно удовлетворяющих условиям склейки

$$\omega_\alpha = (dg_{\alpha\beta})g_{\alpha\beta}^{-1} + g_{\alpha\beta}\omega_\beta g_{\alpha\beta}^{-1}. \qquad (8)$$

При этом из построения расслоения F следует, что монодромия связности ∇ совпадает с χ.

2. Далее пара (F, ∇) продолжается до расслоения F^0 с логарифмической связностью ∇^0 над всей сферой Римана. Для этого набор $\{U_\alpha\}$ нужно дополнить малыми окрестностями $O_1, \ldots, O_n$ точек $a_1, \ldots, a_n$ соответственно. Продолжение расслоения F в каждую точку a_i выглядит следующим образом. Для некоторого непустого пересечения $O_i \cap U_\alpha$ положим $g_{i\alpha}(z) = (z - a_i)^{E_i}$ в этом пересечении. Для любой другой окрестности U_β, имеющей непустое пересечение с O_i, определим $g_{i\beta}(z)$ как аналитическое продолжение матричной функции $g_{i\alpha}(z)$ в $O_i \cap U_\beta$ вдоль подходящего пути (так, чтобы набор $\{g_{\alpha\beta}, g_{i\alpha}(z)\}$ определял коцикл для покрытия $\{U_\alpha, O_i\}$ сферы Римана). Продолжение связности ∇ в каждую точку a_i задается матричной дифференциальной 1-формой $\omega_i = E_i dz/(z - a_i)$, имеющей простой полюс в этой точке. Тогда набор $\{\omega_\alpha, \omega_i\}$ определяет логарифмическую связность ∇^0 в расслоении F^0, поскольку наряду с условиями (8) для непустых пересечений $U_\alpha \cap U_\beta$ также выполняются условия склейки

$$(dg_{i\alpha})g_{i\alpha}^{-1} + g_{i\alpha}\omega_\alpha g_{i\alpha}^{-1} = \frac{E_i}{z - a_i}dz = \omega_i$$

для $O_i \cap U_\alpha \neq \varnothing$ (см. (6)). Пара (F^0, ∇^0) называется *каноническим продолжением* пары (F, ∇).

3. Аналогично построению пары (F^0, ∇^0) осуществляется построение семейства $\mathcal{F}$ расслоений F^Λ с логарифмическими связностями ∇^Λ, имеющими заданные особенности $a_1, \ldots, a_n$ и монодромию (3). Для этого нужно заменить матрицы $g_{i\alpha}(z)$ в построении пары (F^0, ∇^0) матрицами

$$g_{i\alpha}^\Lambda(z) = (z - a_i)^{\Lambda_i} S_i (z - a_i)^{E_i},$$

а формы ω_i — формами

$$\omega_i^\Lambda = \left(\Lambda_i + (z - a_i)^{\Lambda_i} E_i' (z - a_i)^{-\Lambda_i}\right) \frac{dz}{z - a_i},$$

где $\{\Lambda_1, \ldots, \Lambda_n, S_1, \ldots, S_n\}$ — всевозможные наборы допустимых матриц. При этом снова будут выполнены условия склейки

$$(dg_{i\alpha}^\Lambda)(g_{i\alpha}^\Lambda)^{-1} + g_{i\alpha}^\Lambda \omega_\alpha (g_{i\alpha}^\Lambda)^{-1} = \omega_i^\Lambda$$

на непустых пересечениях $O_i \cap U_\alpha$ (см. (7)).

Собственные значения $\beta_i^j = \lambda_i^j + \rho_i^j$ матрицы $\Lambda_i + E_i'$ называются *показателями* логарифмической связности ∇^Λ в точке $z = a_i$. (Из вида форм ω_i^Λ следует, что показатели в точке $z = a_i$ — это собственные значения матрицы-вычета $\operatorname{res}_{a_i} \omega_i^\Lambda$.)

По теореме Биркгофа—Гротендика любое голоморфное векторное расслоение E ранга p над сферой Римана эквивалентно прямой сумме

$$E \cong \mathcal{O}(k_1) \oplus \ldots \oplus \mathcal{O}(k_p)$$

одномерных расслоений, имеющей координатное описание вида

$$\left(U_1 = \mathbb{C},\ U_\infty = \overline{\mathbb{C}} \setminus \{a_1\},\ g_{1\infty} = (z - a_1)^K\right),$$

где $K = \operatorname{diag}(k_1, \ldots, k_p)$, $k_1 \geqslant \cdots \geqslant k_p$ — набор целых чисел, который называется *типом расщепления* расслоения E.

Расслоение E голоморфно тривиально тогда и только тогда, когда оно имеет нулевой тип расщепления.

Число $\deg E = \sum_{i=1}^p k_i$ называется *степенью* расслоения E. Для пары $(F^\Lambda, \nabla^\Lambda)$ степень расслоения F^Λ совпадает также с суммой $\sum_{i=1}^n \sum_{j=1}^p \beta_i^j$ показателей связности ∇^Λ (см. [4], предложение 10.3).

Если некоторое расслоение F^Λ семейства $\mathcal{F}$ окажется голоморфно тривиальным, то соответствующая логарифмическая связность ∇^Λ будет определять глобальную фуксову систему (2), решающую проблему Римана—Гильберта. С другой стороны, вследствие упомянутой выше теоремы Левеля, из существования фуксовой системы с заданными особыми точками $a_1, \ldots, a_n$ и монодромией (3) следует тривиальность некоторого расслоения семейства $\mathcal{F}$.

Таким образом, *проблема Римана—Гильберта разрешима тогда и только тогда, когда хотя бы одно из расслоений семейства $\mathcal{F}$ голоморфно тривиально.*

3. Проблема Римана—Гильберта для скалярных фуксовых уравнений и уравнение Пенлеве VI

Как говорилось ранее, задача о построении фуксова дифференциального уравнения (1) с заданными особенностями $a_1, \ldots, a_n$ и монодромией (3) в общем случае имеет отрицательное решение. При построении необходимо возникают дополнительные особые точки. В случае, когда представление (3) *неприводимо* (т. е. когда у образующих $G_1, \ldots, G_n$ нет общего инвариантного подпространства), А. А. Болибрухом [3] была получена формула для минимального числа таких особенностей, которую мы приводим ниже.

Рассмотрим семейство $\mathcal{F}$ голоморфных векторных расслоений F^Λ с логарифмическими связностями ∇^Λ, построенное по представлению (3). *Фуксовым весом* расслоения F^Λ называется величина

$$\gamma(F^\Lambda) = \sum_{i=1}^{p} (k_1 - k_i),$$

где $(k_1, \ldots, k_p)$ — тип расщепления расслоения F^Λ.

Если представление (3) *неприводимо*, то для типа расщепления расслоения F^Λ имеют место неравенства

$$k_i - k_{i+1} \leqslant n-2, \qquad i = 1, \ldots, p-1 \quad (9)$$

(см. [4], теорема 11.1). Поэтому для такого представления определена величина

$$\gamma_{\max}(\chi) = \max_{F^\Lambda \in \mathcal{F}} \gamma(F^\Lambda) \leqslant \frac{(n-2)p(p-1)}{2},$$

называемая *максимальным фуксовым весом* неприводимого представления χ.

Минимально возможное число m_0 дополнительных особых точек, возникающих при построении фуксова уравнения (1) по *неприводимому* представлению (3), выражается формулой

$$m_0 = \frac{(n-2)p(p-1)}{2} - \gamma_{\max}(\chi). \quad (10)$$

Уравнение Пенлеве VI ($\mathrm{P_{VI}}$) возникает при решении задачи о построении фуксова дифференциального уравнения *второго* порядка с *четырьмя* заданными особенностями и *неприводимой* монодромией. Далее мы расскажем о том, как это происходит.

Уравнение $\mathrm{P_{VI}}$ — это нелинейное дифференциальное уравнение второго порядка

$$\begin{aligned} \frac{d^2u}{dt^2} &= \frac{1}{2}\left(\frac{1}{u} + \frac{1}{u-1} + \frac{1}{u-t}\right)\left(\frac{du}{dt}\right)^2 - \\ &- \left(\frac{1}{t} + \frac{1}{t-1} + \frac{1}{u-t}\right)\frac{du}{dt} + \\ &+ \frac{u(u-1)(u-t)}{t^2(t-1)^2}\left(\alpha + \beta\,\frac{t}{u^2} + \right. \\ &+ \left. \gamma\,\frac{t-1}{(u-1)^2} + \delta\,\frac{t(t-1)}{(u-t)^2}\right) \end{aligned} \quad (11)$$

относительно неизвестной функции $u(t)$. *Подвижными особенностями* решений этого уравнения (положение которых зависит от начальных условий) могут быть только полюсы. В таком случае говорят, что уравнение удовлетворяет *свойству Пенлеве*. В общем виде (11) уравнение $\mathrm{P_{VI}}$ впервые записано Р. Фуксом [10] (сыном Л. Фукса). Им было предложено два способа получения этого уравнения. Первый, на котором мы здесь остановимся, связан с изомонодромными деформациями линейных дифференциальных уравнений. Второй, более геометрический подход, использует эллиптические интегралы.

Рассмотрим четыре точки $t, 0, 1, \infty$ ($t \in D(t^*)$, где $D(t^*) \subset \mathbb{C} \setminus \{0,1\}$ — диск малого радиуса с центром в точке t^*) и *неприводимое* $SL(2,\mathbb{C})$-представление

$$\chi^* : \pi_1(\mathbb{C} \setminus \{t, 0, 1\}) \longrightarrow SL(2,\mathbb{C}),$$

порожденное матрицами G_1, G_2, G_3, соответствующими точкам $t, 0, 1$.

В зависимости от расположения точки t возможны два случая.

1) Любое векторное расслоение F^Λ из семейства $\mathcal{F}$, построенного по данным четырем точкам и представлению χ^*, такое, что $\deg F^\Lambda = 0$, голоморфно тривиально[3].

2) Среди элементов семейства $\mathcal{F}$ существует голоморфно нетривиальное расслоение F^Λ степени нуль.

[3] Данный случай имеет место для почти всех значений $t \in D(t^*)$ (см. [4], упражнение 16.4).

Из неравенств (9) следует, что $\gamma_{\max}(\chi^*) \leqslant 2$, поэтому в первом случае типы расщепления голоморфно нетривиальных расслоений F^Λ (ненулевой степени) могут быть только $(k, k-1)$ или (k, k). Вариант $(k+1, k-1)$ невозможен, поскольку тогда расслоение, построенное по набору матриц $\Lambda_1 - kI, \Lambda_2, \Lambda_3, \Lambda_4$, имеет нулевую степень, т. е. голоморфно тривиально, но в то же время его тип расщепления равен $(1, -1)$. Следовательно, в первом случае $\gamma_{\max}(\chi^*) = 1$.

Во втором случае тип расщепления голоморфно нетривиального расслоения нулевой степени равен $(1, -1)$, и в этом случае $\gamma_{\max}(\chi^*) = 2$.

Итак, ввиду формулы (10), для почти всех значений $t \in D(t^*)$ набор точек $t, 0, 1, \infty$ и неприводимое двумерное представление χ^* реализуются фуксовым дифференциальным уравнением второго порядка с одной дополнительной особенностью. Обозначим ее через $u(t)$ (рассматривая как функцию параметра t). Оказывается, что функция $u(t)$ удовлетворяет уравнению (11) при некоторых значениях констант $\alpha, \beta, \gamma, \delta$ [4]. Этот интересный факт можно пояснить при помощи *изомонодромных деформаций фуксовых систем*.

Выберем значение t^0, при котором существует фуксова система

$$\frac{dy}{dz} = \left(\frac{B_1}{z - t^0} + \frac{B_2}{z} + \frac{B_3}{z-1}\right) y$$

с особыми точками $t^0, 0, 1, \infty$, представлением монодромии χ^* и такая, что $B_i \in sl(2, \mathbb{C})$, а матрица $B_\infty = -B_1 - B_2 - B_3$ диагональна.

Существование такой системы (даже семейства систем, зависящего от дискретного параметра $\mathbf{m} = (m_1, m_2, m_3, m_\infty) \in \mathbb{Z}_+^4$) показать несложно.

Действительно, если набор $\Lambda = \{\Lambda_1, \Lambda_2, \Lambda_3, \Lambda_\infty\}$ допустимых матриц удовлетворяет условиям $\operatorname{tr}(\Lambda_k + E_k) = 0$, $k = 1, 2, 3, \infty$, и собственные значения матрицы $\Lambda_\infty + E_\infty$ отличны от нуля, то для почти всех значений t^0 соответствующее расслоение F^Λ голоморфно тривиально и логарифмическая связность ∇^Λ определяет фуксову систему, обладающую нужными свойствами.

Вследствие соотношений $e^{2\pi i E_k} = G_k$, $\det G_k = 1$, сумма $\rho_k^1 + \rho_k^2$ собственных значений матрицы E_k является целым числом, причем эта сумма равна 0 или 1 ввиду условия $0 \leqslant \operatorname{Re} \rho_k^j < 1$. Зафиксируем порядок собственных значений ρ_k^1, ρ_k^2 и положим $\rho_k = \rho_k^1$.

[4] Уравнение $\mathrm{P_{VI}}$ получено Р. Фуксом именно как дифференциальное уравнение, которому удовлетворяет дополнительная (пятая) особенность $\lambda(t)$ некоторого фуксова уравнения второго порядка с особыми точками $0, 1, t, \infty$ и монодромией, не зависящей от параметра t.

1) Если $\rho_k^1 + \rho_k^2 = 0$, то в качестве матриц Λ_k можно взять любые матрицы вида $\Lambda_k = \operatorname{diag}(m_k, -m_k)$, $m_k \in \mathbb{Z}_+$ (но если $\rho_\infty = 0$, то $m_\infty \in \mathbb{N}$).

2) Если $\rho_k^1 + \rho_k^2 = 1$, то в качестве матриц Λ_k можно взять любые матрицы вида $\Lambda_k = \operatorname{diag}(m_k, -m_k - 1)$, $m_k \in \mathbb{Z}_+$.

Таким образом, для почти всех значений $t^0 \in D(t^*)$ представление χ^* реализуется фуксовыми системами

$$\frac{dy}{dz} = \left(\frac{B_1^{\mathbf{m}}}{z - t^0} + \frac{B_2^{\mathbf{m}}}{z} + \frac{B_3^{\mathbf{m}}}{z-1}\right) y \qquad (12)$$

($\mathbf{m} = (m_1, m_2, m_3, m_\infty) \in \mathbb{Z}_+^4$) с особыми точками $t^0, 0, 1, \infty$, при этом собственные значения матриц $B_k^{\mathbf{m}}$ суть $\pm(m_k + \rho_k)$, а матрицы $B_\infty^{\mathbf{m}} = -B_1^{\mathbf{m}} - B_2^{\mathbf{m}} - B_3^{\mathbf{m}}$ диагональны. (Напомним, что собственные значения матриц $B_k^{\mathbf{m}}$ совпадают с показателями системы (12) — показателями логарифмической связности в голоморфно тривиальном расслоении.)

Всякая система вида (12) может быть вложена в *изомонодромное семейство*[5] *Шлезингера*

$$\frac{dy}{dz} = \left(\frac{B_1^{\mathbf{m}}(t)}{z - t} + \frac{B_2^{\mathbf{m}}(t)}{z} + \frac{B_3^{\mathbf{m}}(t)}{z-1}\right) y \qquad (13)$$

($B_k^{\mathbf{m}}(t^0) = B_k^{\mathbf{m}}$) фуксовых систем с особенностями $t, 0, 1, \infty$, голоморфно зависящее от параметра $t \in D(t^0)$, где $D(t^0)$ — диск малого радиуса с центром в точке t^0. При этом $B_1^{\mathbf{m}}(t) + B_2^{\mathbf{m}}(t) + B_3^{\mathbf{m}}(t) = -B_\infty^{\mathbf{m}} = \operatorname{diag}(-m_\infty - \rho_\infty, m_\infty + \rho_\infty)$. Б. Мальгранжем [6] показано, что матричные функции $B_k^{\mathbf{m}}(t)$ продолжаются на универсальное накрытие T пространства $\mathbb{C} \setminus \{0, 1\}$ как мероморфные функции. Множество $\Theta_{\mathbf{m}} \subset T$ их полюсов называется Θ-*дивизором Мальгранжа*.

Обозначим через $B^{\mathbf{m}}(z, t) = (b_{ij}^{\mathbf{m}}(z, t))$ матрицу коэффициентов семейства (13). Поскольку верхний правый элемент матрицы $B_1^{\mathbf{m}}(t) + B_2^{\mathbf{m}}(t) + B_3^{\mathbf{m}}(t) = -B_\infty^{\mathbf{m}}$ равен нулю, то такой же элемент матрицы $z(z-1)(z-t)B^{\mathbf{m}}(z, t)$ при каждом фиксированном t является многочленом первой степени по z. Определим $\tilde{u}_{\mathbf{m}}(t)$ как единственный корень этого многочлена. Далее мы воспользуемся следующей теоремой (подробное доказательство см. в [4], теорема 18.1).

Теорема 1. *Функция $\tilde{u}_{\mathbf{m}}(t)$ удовлетворяет уравнению (11), где константы α, β, γ, δ связаны с параметром $\mathbf{m} = (m_1, m_2, m_3, m_\infty)$ сле-*

[5] Изомонодромность означает, что монодромия систем этого семейства не зависит от значения параметра t. Более того, собственные значения матриц $B_k^{\mathbf{m}}(t)$ также не зависят от t и совпадают с собственными значениями $\pm(m_k + \rho_k)$ матриц $B_k^{\mathbf{m}}(t^0) = B_k^{\mathbf{m}}$ (подробнее см. в [4], лекция 13).

дующими соотношениями:

$$2\alpha = (2m_\infty + 2\rho_\infty - 1)^2, \quad \beta = -2(m_2 + \rho_2)^2,$$

$$\gamma = 2(m_3 + \rho_3)^2, \quad \delta = \frac{1}{2} - 2(m_1 + \rho_1)^2.$$

□

Рассмотрим вектор-строки $q_0^{\mathbf{m}} = (1, 0)$,

$$q_1^{\mathbf{m}}(z,t) = \frac{dq_0^{\mathbf{m}}}{dz} + q_0^{\mathbf{m}} B^{\mathbf{m}}(z,t) = (b_{11}^{\mathbf{m}}, b_{12}^{\mathbf{m}})$$

и составленную из них матрицу

$$\Gamma_{\mathbf{m}}(z,t) = \begin{pmatrix} q_0^{\mathbf{m}} \\ q_1^{\mathbf{m}} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ b_{11}^{\mathbf{m}} & b_{12}^{\mathbf{m}} \end{pmatrix},$$

мероморфно обратимую на $\overline{\mathbb{C}} \times D(t^*)$, поскольку $\det \Gamma_{\mathbf{m}}(z,t) = b_{12}^{\mathbf{m}}(z,t) \not\equiv 0$ в силу неприводимости представления χ^*. Определим мероморфные на $\overline{\mathbb{C}} \times D(t^*)$ функции $a_{\mathbf{m}}(z,t)$, $b_{\mathbf{m}}(z,t)$ так, чтобы выполнялось соотношение

$$\begin{aligned} q_2^{\mathbf{m}}(z,t) &:= \frac{dq_1^{\mathbf{m}}}{dz} + q_1^{\mathbf{m}} B^{\mathbf{m}}(z,t) = \\ &= (-b_{\mathbf{m}}, -a_{\mathbf{m}}) \Gamma_{\mathbf{m}}(z,t). \end{aligned}$$

Непосредственно проверяется равенство

$$\frac{d\Gamma_{\mathbf{m}}}{dz}\Gamma_{\mathbf{m}}^{-1} + \Gamma_{\mathbf{m}} B^{\mathbf{m}}(z,t) \Gamma_{\mathbf{m}}^{-1} = \begin{pmatrix} 0 & 1 \\ -b_{\mathbf{m}} & -a_{\mathbf{m}} \end{pmatrix},$$

означающее, что при каждом фиксированном $t \in D(t^0)$ калибровочное преобразование $y' = \Gamma_{\mathbf{m}}(z,t) y$ переводит соответствующую систему семейства (13) в систему

$$\frac{dy'}{dz} = \begin{pmatrix} 0 & 1 \\ -b_{\mathbf{m}} & -a_{\mathbf{m}} \end{pmatrix} y',$$

первая координата решения которой является, как обычно, решением скалярного уравнения

$$\frac{d^2w}{dz^2} + a_{\mathbf{m}}(z,t)\frac{dw}{dz} + b_{\mathbf{m}}(z,t) w = 0. \qquad (14)$$

Данное (фуксово) уравнение имеет особые точки $t, 0, 1, \infty$ и монодромию χ^*, но есть у него и дополнительная особенность $u_{\mathbf{m}}(t)$ — это, как следует из построения функций $a_{\mathbf{m}}(z,t)$, $b_{\mathbf{m}}(z,t)$, нуль функции $\det \Gamma_{\mathbf{m}}(z,t) = b_{12}^{\mathbf{m}}(z,t)$. В силу теоремы 1 функция $u_{\mathbf{m}}(t)$ удовлетворяет уравнению $\mathrm{P_{VI}}$. (Заметим, что $u_{\mathbf{m}}(t) \neq t, 0, 1, \infty$, если $t \in D(t^*) \setminus \widetilde{\Theta}$, где $\widetilde{\Theta}$ — счетное множество, состоящее из таких значений параметра t, что соответствующее семейство $\mathcal{F}$, построенное по особенностям $t, 0, 1, \infty$ и представлению χ^*, содержит нетривиальное расслоение степени нуль.) Таким образом, можно сформулировать следующее утверждение.

Утверждение 1.

i) *Множество точек $t, 0, 1, \infty$ и неприводимое $SL(2,\mathbb{C})$-представление χ^* реализуются семейством скалярных фуксовых уравнений (14), зависящим от параметра $\mathbf{m} \in \mathbb{Z}_+^4$, с одной дополнительной особенностью.*

ii) *Дополнительная особенность $u_{\mathbf{m}}(t)$ каждого из уравнений этого семейства (как функция параметра $t \in D(t^*)$) удовлетворяет уравнению $\mathrm{P_{VI}}$ с параметрами α, β, γ, δ, выраженными в теореме 1.*

iii) *Особые точки функции $u_{\mathbf{m}}(t)$, продолженной на T, являются полюсами, а множество $\widetilde{\Theta} \supset \bigcup_{\mathbf{m}} \{t \in D(t^*) | u_{\mathbf{m}}(t) = t, 0, 1$ или $\infty\}$ является счетным множеством значений параметра, при которых рассматриваемая проблема Римана—Гильберта для скалярных фуксовых уравнений решается без дополнительных особенностей.* □

Список литературы

[1] Д. В. Аносов, В. П. Лексин, "Андрей Андреевич Болибрух в жизни и науке", *УМН*, 2004, т. 59(6), сс. 3–22.

[2] А. А. Болибрух, "Проблема Римана—Гильберта на комплексной проективной прямой", *Матем. заметки*, 1989, т. 46(3), сс. 118–120.

[3] A. A. Bolibruch, "Vector bundles associated with monodromies and asymptotics of Fuchsian systems", *J. Dynam. Control Systems*, 1995, v. 1(2), pp. 229–252.

[4] А. А. Болибрух, *Фуксовы дифференциальные уравнения и изомонодромные деформации*, М.: МЦНМО (в печати).

[5] A. Levelt, "Hypergeometric functions", *Proc. Konikl. Nederl. Acad. Wetensch., Ser. A*, 1961, v. 64, pp. 361–401.

[6] B. Malgrange, "Sur les déformations isomonodromiques. I. Singularités régulières", *Progress in Math.*, 1983, v. 37, pp. 401–426.

[7] A. Poincaré, "Sur les groupes des équations linéaires", *Acta Math.*, 1884, v. 4, pp. 201–311.

[8] Б. Риман, "Две теоремы общего характера, касающиеся линейных дифференциальных уравнений с алгебраическими коэффициентами", *Сочинения*, М.: Гостехтеоретиздат, 1948, сс. 176–186.

[9] L. Fuchs, "Zur Theorie der linearen Differentialgleichungen mit veränderlichen Koefficienten", *J. Reine Angew. Math.*, 1868, v. 68, pp. 354–385.

[10] R. Fuchs, "Sur quelques équations différentielles linéaires du second ordre", *C. R. Acad. Sci. Paris*, 1905, v. 141, pp. 555–558.

О минимаксной асимптотической оптимальности первого порядка в задаче наискорейшего обнаружения изменения интенсивности пуассоновского процесса

Бурнаев Евгений Владимирович
Институт проблем передачи информации им. А.А. Харкевича РАН
burnaev@iitp.ru

Аннотация

Рассматривается задача о разладке для пуассоновского процесса в минимаксной постановке. Для минимаксного риска получены оценки снизу и сверху. Показано, что метод обнаружения разладки на основе процесса Ширяева-Робертса является асимптотически оптимальным первого порядка (при больших значениях среднего времени до ложной тревоги), поскольку для этого метода асимптотики полученных оценок минимаксного риска совпадают.

1. Введение. Постановка задачи

Пусть на некотором вероятностном пространстве (Ω, F, P) наблюдается случайный процесс $X = (X_t)_{t\geq 0}$, имеющий вид $X_t = \int_0^t I(s<\theta)\, dN_s^{\lambda_0} + \int_0^t I(s\geq\theta)\, dN_{s-\theta}^{\lambda_1}$, $X_0 = 0$, где $N^{\lambda_0} = (N_t^{\lambda_0})_{t\geq 0}$ и $N^{\lambda_1} = (N_t^{\lambda_1})_{t\geq 0}$ – пуассоновские процессы с интенсивностями λ_0 и λ_1 соответственно, $I(A)$ – индикатор события A.

Неслучайный момент $\theta \in [0,\infty]$ интерпретируется как момент появления "разладки" (по терминологии [1]-[5]) в ходе наблюдения за процессом X: при $t<\theta$ наблюдаемый процесс $X_t = N_t^{\lambda_0}$, а при $t\geq\theta$ процесс $X_t = N_\theta^{\lambda_0} + N_{t-\theta}^{\lambda_1}$.

Будем через $\tau = \tau(\omega)$ обозначать момент подачи сигнала о появлении "разладки", т.е. пусть τ есть марковский момент со значениями в $[0,\infty]$ относительно потока (F_t^X), где $F_t^X = \sigma\{\omega : X_s, s\leq t\}$, $t\geq 0$.

В работах [1]-[6] рассматривались различные постановки задачи обнаружения разладки и были найдены *оптимальный байесовский метод* и *оптимальный метод* в предположении, что разладка появляется *"на фоне установившегося стационарного режима наблюдения"*.

В настоящей работе доказана асимптотическая оптимальность первого порядка статистики Ширяева-Робертса (введенной в [1]-[6]) для *минимаксного критерия* (рассмотренного, например, в [6], [8]) в задаче о разладке для пуассоновского процесса.

Пусть P_t – распределение вероятностей процесса X в предположении, что момент разладки $\theta = t$. Распределение P_∞ отвечает случаю отсутствия разладки ($\theta=\infty$), P_0 соответствует случаю наличия разладки в момент $\theta = 0$. Обозначим через $M_T = \{\tau : E_\infty\tau = T\}$ – класс моментов остановки τ, для которых среднее время $E_\infty\tau$ до подачи (ложной) тревоги равно T. Также будем рассматривать более широкий класс моментов остановки $\overline{M}_T = \{\tau : E_\infty\tau \geq T\}$.

Качество системы наблюдения, идентифицируемой с выбором момента $\tau\in M_T$ ($\tau\in\overline{M}_T$), будем оценивать величиной $C(T;\tau) = \sup\limits_{\theta\geq 0} E(\tau-\theta \mid \tau\geq\theta)$. Момент $\tau_T^*\in M_T$ ($\tau_T^*\in\overline{M}_T$) естественно называть *оптимальным в минимаксном смысле*, если $C(T) = C(T;\tau_T^*) = \inf\limits_{\tau\in M_T} C(T;\tau)$ ($\overline{C}(T) = C(T;\tau_T^*) = \inf\limits_{\tau\in\overline{M}_T} C(T;\tau)$).

Отметим, что минимаксная задача о разладке в непрерывном времени для пуассоновского процесса до сих пор не рассматривалась. Известны только результаты, касающиеся минимаксной задачи наискорейшего обнаружения изменения сноса у броуновского движения. Так в [6], [8], [12] получены асимптотики первого и второго порядков минимаксной функции риска в задаче о разладке

для броуновского движения (использовался момент остановки на основе процесса Ширяева-Робертса). В [7], [8] доказана минимаксная оптимальность *непрерывной* версии метода кумулятивных сумм (CUSUM), предложенного в [11] для критерия, введенного в [10].

В данной работе найден такой момент остановки τ, для которого в предельном случае (при больших значениях интенсивностей λ_0 и λ_1) доказана асимптотическая оптимальность первого порядка (при больших значениях среднего времени до ложной тревоги).

Задача о разладке возникает при решении многих познавательных и прикладных задач. С ней приходится сталкиваться, например, при изучении явлений живой природы, решении задач правления, технической и медицинской диагностики, обработке данных измерений, обнаружении неисправностей технических устройств, анализе финансовых данных и т.д. [13]-[16].

Полученные результаты позволяют теоретически обосновать правомерность практического применения процедуры обнаружения разладки на основе статистики Ширяева-Робертса к процессам, имеющим скачковую компоненту (примеры практического применения см. в [13]-[16]).

2. Основной результат

Обозначим через $\psi = (\psi_t)_{t\ge 0}$ – процесс Ширяева-Робертса, имеющий вид $\psi_t = \int_0^t \frac{L_t}{L_s} ds$, где $L_t = \frac{d(P_0 \mid F_t^X)}{d(P_\infty \mid F_t^X)} = \exp\left(\log\left(\frac{\lambda_1}{\lambda_0}\right) X_t - (\lambda_1 - \lambda_0)\cdot t\right)$ – производная Радона-Никодима сужения $P_0 \mid F_t^X$ меры P_0 относительно сужения $P_\infty \mid F_t^X$ меры P_∞ [5]. Процесс $\psi = (\psi_t)_{t\ge 0}$ удовлетворяет уравнению $d\psi_t = dt + \left(\frac{\lambda_1}{\lambda_0} - 1\right)\psi_{t-} d(X_t - \lambda_0 t)$, $\psi_0 = 0$. Пусть $\rho = \frac{1}{2}\left(\frac{\lambda_1 - \lambda_0}{\sqrt{\lambda_0}}\right)^2$, $E = 0.577\ldots$ – константа Эйлера [9]. Обозначим также через $A = A(T)$ решение уравнения $E_\infty \tau_{A(T)} = T$ для некоторого фиксированного $T > 0$, где $\tau_{A(T)} = \inf\{t \ge 0 : \psi_t \ge A(T)\}$.

Выполняется следующая

Лемма. При любых $T > 0$ уравнение $E_\infty \tau_{A(T)} = T$ имеет решение $A = A(T)$, непрерывно зависящее от T. При этом для $\lambda_1 < \lambda_0$ решение $A = A(T) = T$, а для $\lambda_1 > \lambda_0$ решение удовлетворяет неравенствам $\frac{\lambda_0}{\lambda_1} T < A(T) < T$. Для этого решения момент остановки $\tau_{A(T)} \in M_T \subset \overline{M}_T$.

Теорема. При любых $T > 0$ момент остановки $\tau_{A(T)}$ таков, что $C_1(T) \le \overline{C}(T) \le C(T) \le C_2(T)$, где $C_1(T) = \frac{1}{T} E_\infty \int_0^{\tau_{A(T)}} \psi_s ds$, $C_2(T) = E_0 \tau_{A(T)}$.

При этом $C_1(T) = V_1(0;T)/T$ и $C_2(T) = V_2(0;T)$, где $V_1(x;T)$ и $V_2(x;T)$ – решения дифференциально-разностных уравнений

$$(1 - (\lambda_1 - \lambda_0)x) V_1'(x;T) + \lambda_0\left(V_1\left(\frac{\lambda_1}{\lambda_0} x;T\right) - V_1(x;T)\right) = -x,$$

и

$$(1 - (\lambda_1 - \lambda_0)x) V_2'(x;T) + \lambda_1\left(V_2\left(\frac{\lambda_1}{\lambda_0} x;T\right) - V_2(x;T)\right) = -1.$$

Если $\lambda_0 \to \infty$ и $\lambda_1 \to \infty$ так, что для некоторой фиксированной константы $L \in R$ выполняется равенство $\lambda_1 = \lambda_0 + \sqrt{\lambda_0} L$, то $C_1(T) = C_1^0(T) + \varepsilon \cdot C_1^1(T) + \ldots$ и $C_2(T) = C_2^0(T) + \varepsilon \cdot C_2^1(T) + \ldots$, где малый параметр $\varepsilon = 1/\sqrt{\lambda_0}$. Для функций $C_1^0(T)$ и $C_1^1(T)$ имеют место следующие асимптотики:

$$C_1^0(T) = \begin{cases} \frac{1}{\rho}\left(\frac{\rho T}{2} + O\left((\rho T)^2\right)\right), T \to 0 \\ \frac{1}{\rho}\left(\log(\rho T) - (1 + E)\right) + \\ \qquad O\left(\frac{\log^2 \rho T}{\rho^{3/2} T}\right), T \to \infty, \end{cases}$$

при $\lambda_1 < \lambda_0$

$$C_1^1(T) = \begin{cases} -\frac{1}{3}\sqrt{\frac{2}{\rho}}\left(\frac{(\rho T)^3}{2} + O\left((\rho T)^4\right)\right), T \to 0 \\ -\frac{1}{3}\sqrt{\frac{2}{\rho}}\left(\log(\rho T) - \left(\frac{5}{2} + E\right)\right) + \\ \qquad O\left(\frac{\log^2 \rho T}{\rho^{3/2} T}\right), T \to \infty, \end{cases},$$

а при $\lambda_1 > \lambda_0$ $LC_1(T) < C_1^1(T) < RC_1(T)$, где

$$LC_1(T)=\begin{cases}-\sqrt{\frac{2}{\rho}}\left(\rho T+O\left((\rho T)^2\right)\right), T\to 0\\ -\frac{1}{3}\sqrt{\frac{2}{\rho}}\left(2\log(\rho T)-\left(2E-\frac{5}{2}\right)\right)+\\ \qquad O\left(\frac{\log^2\rho T}{\rho^{3/2}T}\right), T\to\infty,\end{cases}$$

$$RC_1(T)=\begin{cases}\sqrt{\frac{2}{\rho}}\left(\rho T+O\left((\rho T)^2\right)\right), T\to 0\\ \frac{1}{3}\sqrt{\frac{2}{\rho}}\left(4\log(\rho T)-\left(4E+\frac{5}{2}\right)\right)+\\ \qquad O\left(\frac{\log^2\rho T}{\rho^{3/2}T}\right), T\to\infty.\end{cases}$$

Для функций $C_2^0(T)$ и $C_2^1(T)$ имеют место следующие асимптотики:

$$C_2^0(T)=\begin{cases}\frac{1}{\rho}\left(\rho T+O\left((\rho T)^2\right)\right), T\to 0\\ \frac{1}{\rho}\left(\log(\rho T)-E+O\left(\frac{\log^2\rho T}{\rho T}\right)\right), T\to\infty,\end{cases}$$

при $\lambda_1<\lambda_0$

$$C_2^1(T)=\begin{cases}-\frac{1}{3}\sqrt{\frac{2}{\rho}}\left(2(\rho T)^2+O\left((\rho T)^4\right)\right), T\to 0\\ -\frac{1}{3}\sqrt{\frac{2}{\rho}}\left(\log(\rho T)-\left(E+\frac{3}{2}\right)\right)+\\ \qquad O\left(\frac{\log\rho T}{\rho^{3/2}T}\right), T\to\infty,\end{cases}$$

а при $\lambda_1>\lambda_0$ $LC_2(T)<C_2^1(T)<RC_2(T)$, где

$$LC_2(T)=\begin{cases}-\sqrt{\frac{2}{\rho}}\left(\rho T+O\left((\rho T)^2\right)\right), T\to 0\\ \frac{1}{3}\sqrt{\frac{2}{\rho}}\left(\log(\rho T)-\left(E+\frac{9}{2}\right)\right)+\\ \qquad O\left(\frac{\log^2\rho T}{\rho^{3/2}T}\right), T\to\infty,\end{cases}$$

$$RC_2(T)=\begin{cases}\sqrt{\frac{2}{\rho}}\left(\rho T+O\left((\rho T)^2\right)\right), T\to 0\\ \frac{1}{3}\sqrt{\frac{2}{\rho}}\left(\log(\rho T)-\left(E-\frac{3}{2}\right)\right)+\\ \qquad O\left(\frac{\log^2\rho T}{\rho^{3/2}T}\right), T\to\infty.\end{cases}$$

3. Список литературы

[1] А.Н. Ширяев, “Обнаружение спонтанно возникающих эффектов”, *ДАН СССР*, 1961, Т. 138, Вып. 4, С. 794-801.

[2] А.Н. Ширяев, “Задача наискорейшего обнаружения нарушения стационарного режима”, *ДАН СССР*, 1961, Т. 138, Вып. 5, С. 1039-1042.

[3] А.Н. Ширяев, “Об оптимальных методах в задачах скорейшего обнаружения”, *ТВП*, 1963, Т. 8, Вып. 1, С. 26-51.

[4] Ширяев, А.Н., *Статистический последовательный анализ*, Наука, М., 1976.

[5] Peshkir, G., Shiryaev, A.N., *Optimal stopping and free-boundary problems*, Birkhauser, 2006.

[6] E.A. Feinberg, A.N. Shiryaev, “Quickest detection of drift change for Brownian motion in generalized Bayesian and minimax settings”, *Statistics and Decisions*, 2006, Vol. 24, pp. 445-470.

[7] А.Н. Ширяев, “Минимаксная оптимальность метода кумулятивных сумм (CUSUM) в случае непрерывного времени”, *УМН*, 1996, Т. 51, Вып. 4, С. 173-174.

[8] A.N. Shiryaev, “Quickest detection problems in the technical analysis of the financial data”, *Mathematical Finance Bachelier Congress (Paris, 2000)*, 2002, Springer, Berlin, pp. 487-521.

[9] Градштейн, И.С., Рыжик, И.М., *Таблицы интегралов, сумм, рядов и произведений*, Физматгиз, М., 1963.

[10] G. Lorden, “Procedures for reacting to a change in distribution”, *Ann. Math. Statist.*, 1971, Vol. 42, pp. 1897-1908.

[11] E.S. Page, “Continuous inspection schemes”, *Biometrika*, 1954, Vol. 41, pp. 100-115.

[12] Е.В. Бурнаев, А.Н. Ширяев, Е.А. Файнберг, “Об асимптотической оптимальности второго порядка в минимаксной задаче скорейшего обнаружения момента изменения сноса у броуновского движения”, *ТВП*, 2008, Т. 53, Вып. 3.

[13] Basseville, M., Nikiforov, I.V., *Detection of abrupt changes: theory and application*, Prentice-Hall, 1993.

[14] Бассевиль, М., Банвенист, А., *Обнаружение изменения свойств сигналов и динамических систем*, Наука, М., 1989.

[15] Никифоров, И.В., *Последовательное обнаружение изменения свойств временных рядов*, Наука, М., 1983.

[16] Тартаковский, А.Г., *Последовательные методы в теории информационных систем*, Радио и связь, М., 1991.

Оптимальное нечеткое агрегирование информации в системах принятия решений

Алфимцев А.Н., Девятков В.В.
Московский Государственный Технический Университет им. Н.Э. Баумана
{alfim, deviatkov}@mail.ru

Аннотация

В статье проводится сравнительный анализ двух нечетких операторов агрегирования Суджено и Шоке, по критериям надежности принимаемых решений и вычислительной сложности. Под надежностью принятия решения понимается процент правильно принятых решений из числа всех попыток. В большинстве существующих систем, использующих нечеткие операторы агрегирования, отдается предпочтение нечеткому оператору Шоке перед нечетким оператором Суджено, но в данной статье будет отмечено, что надежность принятия решений, при определенных условиях, с помощью обоих операторов одинаковая. И так как, оператор Суджено вычислительно-эффективней, чем оператор Шоке, то применение данного оператора в системах принятия решений оптимальней, чем применение оператора Шоке.

1. Введение

Агрегирование, слияние или фьюзинг информации - один из важнейших аспектов построения современной интеллектуальной системы. Системы принятия решений, компьютерного зрения, обработки изображений используют агрегирование данных из различных источников информации для правильного принятия решения. Стремление к объединению нескольких источников информации для получения результирующего решения, объясняется тем, что каждый источник по отдельности может иметь высокую неопределенность или неточность данных, которая уменьшается при их агрегировании [2, 5, 8, 9, 10].

В настоящее время наиболее популярными операторами агрегирования, позволяющими учитывать взаимозависимость источников информации, являются операторы (интегралы) Суджено [9] и Шоке [6]. С использованием нечеткого оператора, как оператора агрегирования информации, извлекаемой на разных уровнях абстракции из изображения, в работе [7] проводился динамический анализ транспортного потока. Система распознавала опасное поведение автомобиля на дороге и предсказывала аварийную ситуацию. В [2] представлена автоматическая система выбора пути для пешеходов. В системе для принятия решения, с помощью нечеткого оператора, учитывались как индивидуальные предпочтения клиентов системы о пути следования, так и объективные данные о самой дороге. В [9] с помощью агрегирования информации нечетким оператором, с высокой надежностью распознавались транспортные средства на поле боя.

Принятие решения на основе распознавания изображения лица человека было описано в [6] и [4]. В первом случае агрегированию подвергалась информация, полученная из вейвлетного разбиения изображения на части. Во втором, агрегировалась информация, полученная как от основного источника информация (видеокамеры), так и от дополнительного источника (инфракрасного сенсора).

В большинстве этих систем в качестве оператора агрегирования отдается предпочтение нечеткому оператору Шоке, перед нечетким оператором Суджено, но в данной работе будет показано, при каких условиях в системах принятия решений, которые базируются на информации, полученной из распознанного изображения лица человека, оптимальнее использовать оператор Суджено, а не Шоке.

В разделе 2 даются определения понятиям нечеткой меры и нечеткого оператора. В разделе 3 рассматривается процедура агрегирования информации для распознавания объекта с использованием операторов Суджено и Шоке. В разделе 4 производится сравнение операторов агрегирования. Раздел 5 является заключительным.

2. Нечеткие меры и нечеткие операторы

Нечеткой мерой называется функция $g: 2^R \to [0,1]$, где R множество каких-либо параметров, характеризующих некоторый объект [1]. Нечеткая мера $g(Q_i)$ характеризует совокупную значимость параметров, входящих в множество Q_i. Нечеткая мера удовлетворяет ряду условий: $g(\varnothing)=0,\ g(Y)=1;$ если $Q,P \in Y$ и $Q \subset P$, то $g(Q) \le g(P)$ [9].

Если μ_i это некоторая функция принадлежности, определяемая на универсуме допустимых отсчетов $y_i \in Y_i$, $i=1,...,m$, относящихся к одному моменту времени, а $Y=\{Y_1,...,\ Y_m\}$ множество алгоритмов распознавания, являющиеся источниками информации для некоторой системы принятия решений, то нечеткие операторы можно определить следующим образом.

Нечеткий оператор Суджено:

$$A_k = A_k^C = \max_{i=1}^{i=m} [\min(\mu_i^k, g(Q_i)] \quad (1)$$

, где

$\mu_1^k(y_1) \ge \mu_2^k(y_2) \ge ...\mu_m^k(y_m),\ Q_i = \{Y_1,...,\ Y_i\},$ $i=1,...,m$.

Нечеткий оператор Шоке:

$$A_k = A_k^{III} = \sum_{i=1}^{i=m} [\mu_i^k(y_i) - \mu_{i+1}^k(y_{i+1})] g(Q_i) \quad (2)$$

,где

$\mu_1^k(y_1) \ge \mu_2^k(y_2) \ge ...\mu_m^k(y_m),\ Q_i = \{Y_1,...,\ Y_i\},$ $i=1,...,m,\ \mu_{m+1}^k(y_{m+1})=0$.

В работе [3] приведена интерпретация оператора Шоке и Суджено на примере операторов небольшой размерности, на основании чего сделано заключение, что оператор Шоке является обобщением понятия взвешенного среднего арифметического, а оператор Суджено - обобщением концепции взвешенной медианы (при использовании не менее трех алгоритмов распознавания).

Вследствие простоты наиболее распространены методы вычисления нечеткой меры, основанные на понятии g_λ-нечеткой меры, введенной Суджено. Нечеткая мера называется g_λ-нечеткой мерой, если для нее справедливо условие: для всех $Q,P \subset Y$ таких, что $Q \cap P = \varnothing$, имеет место

$g(Q \cup P) = g(Q) + g(P) + \lambda g(Q) g(P)$ для некоторого $\lambda > -1$.

Рассмотрим процедуру наиболее популярного метода вычисления g_λ-нечеткой меры [7] обозначая ее по-прежнему просто g.

Шаг 1. Для каждого Y_i, $i=1,...,m$ выбрать значение нечеткой меры $g(Y_i) \in [0,1]$ как степень важности алгоритма Y_i. Значения $g(Y_i)$ могут быть установлены экспертом, получены в результате наблюдений или каким-либо другим путем.

Шаг 3. Найти значение λ, используя уравнение (3)

$$\lambda + 1 = \prod_{i=1}^{m} (1 + \lambda g(Y_i)) . \quad (3)$$

Шаг 3. Для всех $Q_i = \{Y_1,...,\ Y_i\},\ i=1,...,m$ вычислить рекурсивно нечеткие меры $g(Q_i)$, используя следующие выражения:

$$g(Q_1) = g(Y_1) , \quad (4)$$

$$g(Q_i) = g(Y_i) + g(Q_{i-1}) + \lambda g(Y_i)\, g(Q_{i-1}) ,$$

$$i = 2,...,m.$$

3. Процедура агрегирования информации

Прежде, чем переходить к рассмотрению процедуры агрегирования информации для распознавания объекта с использованием нечетких операторов агрегирования, описанной в [9] и развитой в [5], формализуем процедуру предварительной обработки алгоритмов распознавания для формирования множества $\mu(y_{ij_i}), y_{ij_i} \in Y_i, i=0,...,m,\ j_i=0,...,n_i$.

Шаг 1. Задается совокупность пустых множеств $Y_i^k = \varnothing, k=1,...,K$.

Шаг 2. Для каждого эталонного объекта *k*, *k*=1,…, *K*, формируется своя эталонная модель распознавания $G_i^k, k=1,...,K$.

Шаг 3. Для распознаваемого объекта по тем же принципам формируется модель G .

Шаг 4. Модель G сравнивается с каждой моделью $G_i^k, k=1,...,K$, в результате чего вычисляется множество отсчетов $\{y_i^1, y_i^2,...,\ y_i^K\}$, характеризующих близость

модели G соответственно к моделям $G_i^k, k=1,...,K$.

Шаг 5. Формируются множества $Y_i^k \cup y_i^k, k=1,...,K$, которые принимаются за новые множества Y_i^k. Если множества Y_i^k перестают изменяться, то осуществляется переход к шагу 6 (могут использоваться и другие критерии перехода к шагу 6). В противном случае процедура начинается с шага 2.

Шаг 6. Множества Y_i^k объединяются, в результате чего получается множество $Y_i = \bigcup_{k=1}^{K} Y_i^k$, которое упорядочивается (если оно числовое, то упорядочивание осуществляется по возрастанию) и его элементы индексируются $i=1,...,m,\ j_i=0,...,n_i$, в результате чего получается множество $Y_i = \{ y_{ij_i} \in Y_i \,|\, i=1,...,m,\ j_i=0,...,n_i \}$. На множестве Y_i задается функция принадлежности $\mu(y_{ij}), y_{ij} \in Y_i,\ i=1,...,m,\ j=0,...,n_i$.

Объект θ считается распознанным, если существует эталонный объект $\Theta^k,\ k=1,2,...,K$, схожий с объектом θ. Критерием схожести или степенью близости распознаваемого объекта θ к эталонному объекту Θ^k является значение функции $A^k[\mu_1(y_1),\ \mu_2(y_2),...,\ \mu_m(y_m)]]$.

Таким образом, общая процедура агрегирования информации с помощью нечетких операторов Суджено или Шоке будет следующей.

Шаг 1. Выполнить процедуру предобработки и для каждого $Y_i, i=1,...,m$ выбрать значение $g(Y_i) \in [0,1]$, как степень важности Y_i.

Шаг 2. Найти значение λ, используя уравнение (3).

Шаг 3. Для распознаваемого объекта по каждому алгоритму $i=1,...,m$ и для каждого эталонного объекта $k=1,...,K$ вычислить множество функций принадлежности $\mu(y_i^k), y_i^k \in Y_i,\ i=1,...,m$

Шаг 4. Для каждого $k=1,...,K$ упорядочить множество функций $\mu(y_i^k)$ таким образом, чтобы $\mu(y_{j_1}^k) \ge \mu(y_{j_2}^k) \ge ... \ge \mu(y_{j_m}^k), j_n \in \{1,...,m\}$.

Шаг 5. Для каждого $k=1,...,K$ вычислить рекурсивно значения нечетких мер $g(Q_i^k)$, где $Q_i^k = \{Y_{j_1},...,\ Y_{j_i}\},\ i=1,...,m$, используя формулу (4).

Шаг 6. Вычислить для всех $k=1,...,K$ значения операторов $A_k = A_k^{\ C}$ (или $A_k = A_k^{\text{Ш}}$). Распознаваемый объект считается совпадающим с тем эталонным объектом, для которого значение A_k максимально.

4. Сравнение нечетких операторов агрегирования

Рассмотрим, как зависят значения нечетких операторов от вкладов алгоритмов распознавания, характеризуемых значениями функций принадлежности и значениями нечетких мер.

Случай 1. $\mu_i(y_i) \le g(Q_i)$ для всех $i=1,...,m$. Заметим, что если $\mu^k{}_1(y_1) \ge \mu^k{}_2(y_2) \ge ... \ge \mu^k{}_m(y_m)$, то $A_k^C = \mu_1^{\ k}(y_1)$. Поскольку в операторе Шоке каждая разность $\mu_i^k(y_i) - \mu_{i+1}^k(y_{i+1})$ умножается на $g(Q_i) \le 1$, то очевидно, что для этого случая $A_k^C \succ A_k^{Ш}$.

Случай 2. $\mu_i(y_i) \ge g(Q_i)$ для всех $i=1,...,m$.

В этом случае значение $A_k^C = \max_{i=1}^{i=m} [g(Q_i)]$. Если $g(Q_i)=1$ хотя бы для одного $i=1,...,m$, то оператор Суджено $A_k^C = 1$, а оператор Шоке $A_k^{Ш} \le 1$ (единичного значения он достигает только в случае, если хотя бы одна разность $\mu_i^k(y_i) - \mu_{i+1}^k(y_{i+1}) = 1$). Таким образом, для этого случая также $A_k^C \succ A_k^{Ш}$.

При этих условиях значение оператора Суджено превышает значение оператора Шоке, но это не означает, что надежность распознавания, а значит и надежность принимаемого решения, при этих условиях будет выше. Под надежностью принятия решения понимается процент правильно принятых решений из числа всех попыток.

Поэтому для того, чтобы определить надежность принятия решения обоих операторов при условиях $\mu_i(y_i) \le g(Q_i)$ или $\mu_i(y_i) \ge g(Q_i)$, были проведены эксперименты с системой принимающей решение о контроле допуска человека. Система основывается на

распознавании человека по его изображению, полученному от видеокамеры. Для распознавания человека система использует три алгоритма распознавания: Скрытую Марковскую модель (СММ), Алгоритм определения цвета (АОЦ), Алгоритм нахождения соотношений (АНС).

В алгоритме 1 используется множество отсчетов $Y_1 = \{ y_{1j_1} = P_{1j_1} - Q_{1j_1} \,|\, j_1 = 0,...,n_1 \}$, которое получается с помощью сформулированной в разделе 3 процедуры, используя эталонные СММ $G_1^k, k = 1,..., K$. Этим СММ соответствуют вероятности $P_1^k, k = 1,..., K$ распознавания с помощью этих моделей эталонных лиц, на которых происходило обучение моделей. Для вновь распознаваемого лица, используя те же модели $G_1^k, k = 1,..., K$, вычисляются вероятности $Q_1^k, k = 1,..., K$ его распознавания с помощью моделей G_1^k. Множество отсчетов $y_1^k = P_1^k - Q_1^k,\ k = 1,..., K$ вычисляется, как разность вероятностей P_1^k и Q_1^k. На множестве $Y_1 = \{ y_{1j_1} = P_{1j_1} - Q_{1j_1} \,|\, j_1 = 0,...,n_1 \}$ задается функция принадлежности $\mu(y_{1j_1}), j_1 = 0,...,n_1$. Ее график показан на рисунке 1а, из которого видно, что $\mu(y_{1j_1}) = \{1/\, y_{10} = 0;\ 1/\, y_{11} = 1{,}5;\ 0/\, y_{12} = 2\}$.

В алгоритме 2 используется множество отсчетов

$$Y_2 = \{ y_{2j_2} = \sqrt{(r_{2j_2}^k - \tilde{r}_{2j_2})^2 + (g_{2j_2}^k - \tilde{g}_{2j_2})^2 + \overline{+(b_{2j_2}^k - \tilde{b}_{2j_2})^2} \,|\, j_2 = 0,...,n_2 \},$$

которое получается, используя эталонные цветовые модели $G_2^k, k = 1,..., K$ в цветовом пространстве *RGB*. Здесь $r_{2j_2}^k, g_{2j_2}^k, b_{2j_2}^k \,|\, j_2 = 0,...,n_2, k = 1,..., K$ - наборы эталонных значений цветов красного, зеленого и синего, на которых происходило обучение моделей, $\tilde{r}_{2j_2}, \tilde{g}_{2j_2}, \tilde{b}_{2j_2} \,|\, j_2 = 0,...,n_2$ - аналогичные наборы. для вновь распознаваемого лица. На множестве Y_2 задается функция принадлежности $\mu(y_{2j_2}), j_2 = 0,...,n_2$. Ее график показан на рисунке 1б, из которого видно, что $\mu(y_{2j_2}) = \{1/\, y_{20} = 0;\ 0{,}5/\, y_{21} = 20;\ 0/\, y_{22} = 40\}$.

В алгоритме 3 используется множество отсчетов $Y_3 = \{ y_{3j_3} = \sqrt{(A_{3j_3}^k - \tilde{A}_{3j_3})^2 + (B_{3j_3}^k - \tilde{B}_{3j_3})^2 + \overline{+(C_{3j_3}^k - \tilde{C}_{3j_3}^3)^2 + (D_{3j_3}^k - \tilde{D}_{3j_3})^2 + (E_{3j_3}^k - \tilde{E}_{3j_3})^2} \,|\, j_3 = 0,...,n_3 \}$, которое получается, используя эталонные модели отношений $G_3^k, k = 1,..., K$. Здесь $A_{3j_3}^k, B_{3j_3}^k, C_{3j_3}^k, D_{3j_3}^k, E_{3j_3}^k, j_3 = 0,...,n_3$ наборы значений соответственно расстояний между глазами, глазами и носом, носом и ртом, глазами и ртом, глазами и подбородком, на которых происходило обучение моделей, $\tilde{A}_{3j_3}, \tilde{B}_{3j_3}, \tilde{C}_{3j_3}, \tilde{D}_{3j_3}, \tilde{E}_{3j_3}, j_3 = 0,...,n_3$ - аналогичные наборы для вновь распознаваемого лица. На множестве Y_3 задается функция принадлежности $\mu(y_{3j_2}), j_3 = 0,...,n_3$. Ее график показан на рисунке 1в, из которого видно, что $\mu(y_{3j_3}) = \{1/\, y_{20} = 0;\ 0{,}5/\, y_{21} = 5;\ 0/\, y_{22} = 10\}$.

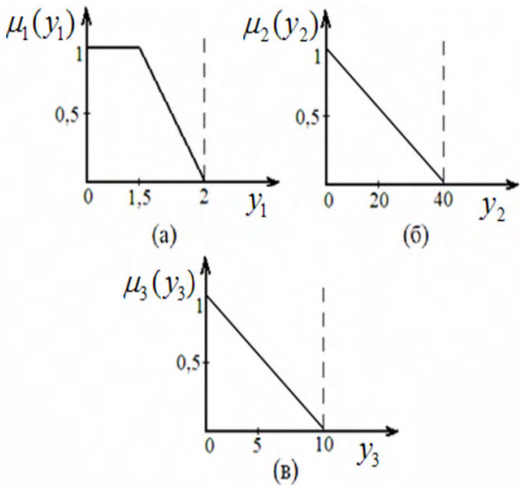


Рис. 1. Функции принадлежностей, соответствующие алгоритмам (а)-Y_1, (б)- Y_2, (в)- Y_3

Для проведения экспериментов, система была обучена и сформирована база данных, хранящая необходимую информацию для алгоритмов распознавания. Пример базы данных представлен на рисунке 2.

Было проведено шестьдесят экспериментов по контролю доступа, в первой половине из которых выполнялось условие $\mu_i(y_i) \le g(Q_i)$, в другой $\mu_i(y_i) \ge g(Q_i)$.

В результате проведенных экспериментов было установлено, что при заданных условиях надежность принятия решения системы, использующей нечеткий оператор агрегирования Суджено или Шоке, одинаковая. Кроме того, не было ни одного случая, когда система принимала

одного человека за другого человека, имеющегося в базе данных.

С другой стороны, для вычисления нечеткого оператора Суджено потребуется выполнить m операций нахождения минимума для двух n-разрядных двоичных чисел, а затем столько же операций нахождения максимума. Поэтому верхнюю оценку вычислительной сложности нечеткого оператора Суджено можно оценить как $O(mn)$. Для вычисления нечеткого оператора Шоке потребуется выполнить m операций сложения, вычитания и умножения n- разрядных двоичных чисел. Если для умножения использовать алгоритм Шенхаге – Штрассена, то верхняя оценка вычислительной сложности нечеткого оператора Шоке составит $O(mn\ln n\ln\ln n)$.

Таким образом, при данных условиях, часто возникающих в задачах распознавания, оптимальней с точки зрения вычислительной сложности использовать нечеткий оператор Суджено, вместо нечеткого оператора Шоке.

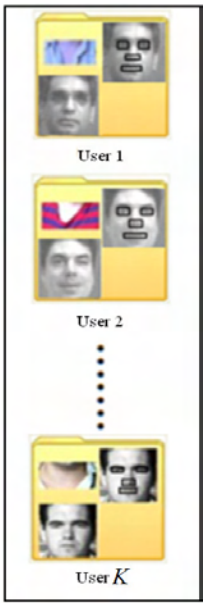


Рис. 2. База данных системы

5. Заключение

В статье рассмотрено использование нечетких операторов агрегирования Суджено и Шоке для объединения данных из нескольких источников информации.Приведена процедура агрегирования информации для задач распознавания и процедура предварительной обработки алгоритмов распознавания.

Проведенный сравнительный анализ двух нечетких операторов агрегирования Суджено и Шоке, по критериям надежности принимаемых решений и вычислительной сложности, показал, что при условии $\mu_i(y_i) \le g(Q_i)$ или $\mu_i(y_i) \ge g(Q_i)$, в качестве оператора агрегирования оптимальней использовать нечеткий оператор Суджено, вместо нечеткого оператора Шоке.

В дальнейшем планируется провести дополнительные эксперименты с большим количеством источников информации, подтверждающие данное утверждение и продолжить исследования возможных соотношений значений функций принадлежности и значений нечеткой меры нечетких операторов.

6. Литература

[1] *Нечеткие множества в моделях управления и искусственного интеллекта*/ А. Н. Аверкин, И. З. Батыршин, А. Ф. Блишун и др.- М.:Наука, 1986.-312 с.

[2] Akasaka Y., Onisawa T. *Individualized pedestrian navigation using fuzzy measures and integrals* // Proc of IEEE Intern. Conf. on syst., man and cybern.- Hawai, 2005.-Vol.2.- P.1461-1466.

[3] Bolanos M.J. *Numerical experimentation and comparison of fuzzy integrals* // Mathware & Soft Computing.-1996.-Vol. 3.-P. 309-319.

[4] Chen X., Jing Z., Xiao G. *Fuzzy Fusion for Face Recognition* // Fuzzy Systems and Knowledge Discovery.-2005.-Vol.361.-P.672-675.

[5] Grabisch M., Roubens M. *Application of the Choquet Integral in Multicriteria Decision Making* // Proc. of Fuzzy Measures and Integrals.- Paris, 2000.-P. 415-434.

[6] Kwak K., Pedrycz W. *Face recognition: A study in information fusion using fuzzy integral* // Patt. Recog. Lett.-2005.-Vol. 26.-P. 719-733.

[7] Liu Z. *Dynamic image sequence analysis using fuzzy measures* // IEEE trans. on sys., man, and cybern.-2001.-Vol. 31, №4.-P.557-572.

[8] Popescu M., Keller J.M., Mitchell J.A. *Fuzzy Measures on the Gene Ontology for Gene Product Similarity* // IEEE/ACM trans. on comput. biology and bioinformatics.-2006.-Vol. 3.-P. 263-274.

[9] Tahani H., Keller J. M. *Information fusion in computer vision using the Fuzzy integral* // IEEE transactions on systems, man, and cybernetics.-1990.- Vol. 20, №3.-P.733-741.

[10] Wang X., Chen J. *Multiple Neural Networks fusion model based on Choquet fuzzy integral* // Proc. of the Third Intern. Conf. on Mach. Learn. and Cybern.-Shanghai, 2004.-Vol.4.-P. 2024-2027.

Об одном достаточном условии PSPACE-разрешимости транзитивных модальных логик

Илья Шапировский
ИППИ РАН
shapir@iitp.ru

Аннотация

В работе описано достаточное условие PSPACE-разрешимости транзитивных модальных логик. Показано, что логики, полные относительно т. н. упорядоченных множеств шкал умеренного роста, являются PSPACE-разрешимыми. В явном виде построен алгоритм проверки выполнимости на таких множествах шкал.

1. Введение

Теория модальных логик — раздел математической логики с широким кругом приложений: модальные логики используются в задачах формальной верификации моделей программ, в задачах искусственного интеллекта, как средство описания пространственной и временной информации, в компьютерной лингвистике. При этом одним из наиболее актуальных вопросов оказывается вычислительная сложность модальных логик.

В работе описан метод доказательства PSPACE-разрешимости транзитивных модальных логик, основанный на семантическом условии, характеризующем реляционные структуры (шкалы Крипке) данной логики (этот метод обобщает подход, предложенный в [5]). Оказывается, многие транзитивные логики могут быть заданы как логики *упорядоченных множеств шкал умеренного роста*. Мы покажем, что такие логики являются PSPACE-разрешимыми, и что для них в явном виде возникает достаточно простая (в смысле описания) разрешающая процедура.

Работа организована следующим образом. В параграфе 2. содержатся некоторые предварительные сведения из теории модальных логик. В параграфе 3. рассматриваются сохраняющие истинность преобразования на упорядоченных множествах шкал. В параграфе 4. рассматриваются упорядоченные множества шкал умеренного роста и дается разрешающая PSPACE-процедура для логик таких шкал. В параграфе 5. мы иллюстрируем наш метод на примерах различных логик (в частности, мы дадим новые доказательства PSPACE-разрешимости для логик K4 и S4, а также рассмотрим менее известную *логику компактного включения* LM).

2. Основные понятия

Определение 1. Множество *модальных формул* строится из счетного множества *пропозициональных переменных* $PV = \{p_1, p_2, \dots\}$, пропозициональной константы $\bot$ (*ложь*), двуместной связки $\to$ (*импликация*) и одноместной связки $\Diamond$ (модальность *"возможно"*) следующим образом:

- $\bot, p_1, p_2, \dots$ — формулы;
- Если φ, ψ — формулы, то $(\varphi \to \psi)$ — формула;
- Если φ — формула, то $\Diamond\varphi$ — формула.

Формулы $\neg\varphi$, $\varphi\wedge\psi$, $\varphi\vee\psi$ и константа $\top$ (*истина*) рассматриваются как сокращения и определяются стандартно, в частности: $\neg\varphi := \varphi \to \bot$, $\top := \neg\bot$. Модальность "необходимо" вводится следующим образом: $\Box\varphi := \neg\Diamond\neg\varphi$.

Определение 2. Множество модальных формул L называется *(нормальной) модальной логикой*, если

- L содержит все булевы тавтологии,
- L содержит формулу $\neg\Diamond\bot$,
- L содержит формулу $\Diamond(p_1 \vee p_2) \to \Diamond p_1 \vee \Diamond p_2$,
- L замкнуто относительно *правила Modus Ponens*, *правила подстановки формул вместо пропозициональных переменных* и *правила монотонности*:

$$\text{если } \varphi \to \psi \in \mathrm{L}, \text{ то } \Diamond\varphi \to \Diamond\psi \in \mathrm{L}.$$

K обозначает наименьшую модальную логику. Для логики L и формулы φ пусть $\mathrm{L}+\varphi$ обозначает наименьшую логику, содержащую $\mathrm{L}\cup\{\varphi\}$.

Определение 3. *Шкалой Крипке* (или просто *шкалой*) F называется пара (W,R), где W — непустое множество и $R\subseteq W\times W$. *Оценкой* на шкале F называется отображение

$$\theta : PV \to \mathcal{P}(W),$$

где $\mathcal{P}(W)$ обозначает множество всех подмножеств W. *Моделью* (*Крипке*) M над шкалой F называется пара (F,θ), где θ — оценка на F. Отношение истинности $\mathsf{M},w\vDash\varphi$ (*"формула φ истинна в точке w модели M"*) определяется с помощью индукции по построению формулы:

$$\begin{array}{lcl}
\mathsf{M},w\nvDash\bot; & & \\
\mathsf{M},w\vDash p & \Leftrightarrow & w\in\theta(p);\\
\mathsf{M},w\vDash\varphi\to\psi & \Leftrightarrow & \mathsf{M},w\nvDash\varphi \text{ или } \mathsf{M},w\vDash\psi;\\
\mathsf{M},w\vDash\Diamond\varphi & \Leftrightarrow & \exists v(wRv\ \&\ \mathsf{M},v\vDash\varphi).
\end{array}$$

R называется *отношением достижимости* в шкале F (в модели M). Всюду в этой работе мы полагаем, что шкалы и модели *транзитивны*, т. е. R удовлетворяет условию

$$\forall x\forall y\forall z(xRy\&yRz\Rightarrow xRz).$$

Шкала (W,R) называется *конусом*, если $W=R(w)\cup\{w\}$ для некоторой точки w (мы полагаем $R(w)=\{v\mid wRv\}$); в этом случае w называется *корнем* шкалы (W,R).

Формула φ называется *истинной* (*выполнимой*) *в модели* M, если она истинна в любой (в некоторой) точки этой модели. Формула φ называется *общезначимой* (*выполнимой*) *в шкале* F, если она истинна (выполнима) в любой (в некоторой) модели над F. Формула φ называется *общезначимой* (*выполнимой*) *в классе шкал* $\mathcal{F}$, если φ общезначима (выполнима) в любой (в некоторой) шкале $\mathsf{F}\in\mathcal{F}$. Множество всех формул, общезначимых в классе $\mathcal{F}$, будем обозначать через $\mathbf{L}(\mathcal{F})$.

Логика L называется *полной относительно класса шкал* $\mathcal{F}$, если $\mathrm{L}=\mathbf{L}(\mathcal{F})$. Логика L называется *полной (по Крипке)*, если она полна относительно некоторого класса шкал. Логика L называется *финитно аппроксимируемой*, если она полна относительно некоторого класса конечных шкал.

Для множества формул Ψ запись $\mathsf{F}\vDash\Psi$ обозначает, что все формулы из Ψ общезначимы в F.

Логика называется *транзитивной*, если она содержит формулу $\Diamond\Diamond p\to\Diamond p$. Положим $\mathrm{K4}=\mathrm{K}+\Diamond\Diamond p\to\Diamond p$, $\mathrm{S4}=\mathrm{K4}+p\to\Diamond p$. Напомним, что для любой шкалы F

$$\mathsf{F}\vDash\mathrm{K4}\Leftrightarrow\mathsf{F}\text{ транзитивна},$$

$$\mathsf{F}\vDash\mathrm{S4}\Leftrightarrow\mathsf{F}\text{ транзитивна и рефлексивна}.$$

Более того, K4 полна относительно класса всех (конечных) транзитивных шкал, S4 полна относительно класса всех (конечных) транзитивных и рефлексивных шкал (см., например, [1]).

Шкала $\mathsf{F}'=(W',R')$ называется *слабой подшкалой* шкалы $\mathsf{F}=(W,R)$, если $W'\subseteq W$, $R'\subseteq R$; модель $\mathsf{M}'=(\mathsf{F}',\theta')$ называется *слабой подмоделью* $\mathsf{M}=(\mathsf{F},\theta)$, если еще $\theta(p)=\theta'(p)\cap W'$ для любого $p\in PV$. В случае, когда $R'=R\cap(W'\times W')$, шкала F' и модель M' называются *сужениями* на множество W' шкалы F и модели M соответственно, и обозначаются $\mathsf{F}|W'$ и $\mathsf{M}|W'$.

Для формулы φ через $Sub(\varphi)$ обозначим множество всех подформул φ.

Определение 4. Пусть Ψ — множество формул, замкнутое относительно взятия подформул (т. е. если $\varphi\in\Psi,\psi\in Sub(\varphi)$, то $\psi\in\Psi$). Пусть M' — слабая подмодель модели M. M' называется *селективной фильтрацией модели* M *по* Ψ, если для любого $w\in\mathsf{M}'$, для любой формулы ψ, имеем:

$$\Diamond\psi\in\Psi\ \&\ \mathsf{M},w\vDash\Diamond\psi\ \Rightarrow\ \exists u\in R'(x)\ \mathsf{M},u\vDash\psi,$$

где R' — отношение достижимости в M'.

Лемма 5 (см., например, [6]). *Если M' — селективная фильтрация модели M по Ψ, то для любого $w\in\mathsf{M}'$, для любой формулы $\psi\in\Psi$ имеем:*

$$\mathsf{M},w\vDash\psi\Leftrightarrow\mathsf{M}',w\vDash\psi.$$

3. Упорядоченные множества шкал

Рассмотрим конечный (строгий или нестрогий) частичный порядок $\mathsf{G}=(W,R)$.

Для точки $w\in W$ и шкалы $\mathsf{F}=(V,S)$ определим шкалу $\mathsf{G}[\mathsf{F}/w]=(W',R')$, полученную *заменой в* G *точки* w *на* F:

$$\begin{array}{l}
W'=(W-\{w\})\cup V',\ \text{где } V'=\{w\}\times V,\\
R'=(R|W-\{w\})\cup\{((w,u'),(w,u''))\mid u'Su''\}\cup\\
\quad\cup(V'\times(R(w)-\{w\}))\cup((R^{-1}(w)-\{w\})\times V').
\end{array}$$

Заметим, что $\mathsf{G}[\mathsf{F}/w]$ — транзитивная шкала (в силу транзитивности G и F).

Для класса шкал $\mathcal{F}$ положим (Рис. 1):

$$\mathsf{G}[\mathcal{F}]=\{\mathsf{G}[\mathsf{F}_1/w_1]\dots[\mathsf{F}_n/w_n]\mid\mathsf{F}_1,\dots,\mathsf{F}_n\in\mathcal{F}\},$$

где m — число элементов G, $W=\{w_1,\dots,w_m\}$. Иными словами, класс $\mathsf{G}[\mathcal{F}]$ получен всевозможными заменами всех точек шкалы G на шкалы из класса F.

Для класса конечных частичных порядков $\mathcal{G}$ положим $\mathcal{G}[\mathcal{F}]=\bigcup\{\mathsf{G}[\mathcal{F}]\mid\mathsf{G}\in\mathcal{G}\}$.

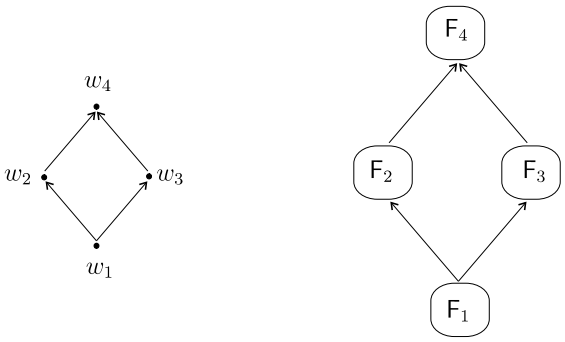


Рис. 1.

Пример 1. Пусть C_n обозначает n-элементную шкалу $\mathsf{C}_n = (C_n, C_n \times C_n)$, где $C_n = \{1, \ldots, n\}$ $(n \geq 1)$; такие шкалы будем называть *сгустками*. Пусть также $\mathsf{C}_0 = (\{0\}, \varnothing)$ (*вырожденный сгусток*).

Пусть $\mathcal{PO}$ обозначает класс всех конечных нестрогих частичных порядков.

Тогда если $\mathcal{F} = \{\mathsf{C}_0\}$, то $\mathcal{PO}[\mathcal{F}]$ — класс всех строгих частичных порядков (здесь и далее совпадение классов шкал подразумевается с точностью до изоморфизмов шкал); если $\mathcal{F}$ — класс всех конечных (невырожденных) сгустков, то $\mathcal{PO}[\mathcal{F}]$ — класс всех конечных транзитивных (и рефлексивных) шкал.

Пусть $\mathsf{F} \in \mathcal{PO}$. Через $Ht(\mathsf{F})$ обозначим высоту F, т. е. наибольшую из длин строго возрастающих цепей в F (подразумевая под длиной цепи число её элементов). Ветвление в точке w шкалы F — число непосредственно следующих за w точек. Через $Br(\mathsf{F})$ обозначим ветвление F, т. е. наибольшее из ветвлений в точках F.

Пусть $\mathcal{T}$ обозначает класс всех конечных транзитивных деревьев; для $h, b \geq 1$ положим $\mathcal{T}_{h,b} = \{\mathsf{T} \in \mathcal{T} \mid Ht(\mathsf{T}) \leq h,\ Br(\mathsf{T}) \leq b\}$ — класс всех конечных транзитивных деревьев высоты не более h и ветвления не более b.

Пусть $\langle\varphi\rangle$ обозначает число всех подформул формулы φ.

Лемма 6. *Для всякого класса шкал $\mathcal{F}$ и формулы φ, если φ $\mathcal{PO}[\mathcal{F}]$-выполнима, то φ $\mathcal{T}_{\langle\varphi\rangle,\langle\varphi\rangle}[\mathcal{F}]$-выполнима.*

Доказательство. Используя стандартную технику *развертки* (*unravelling* — см., например, [1]), можно показать, что φ выполнима в некоторой шкале $\mathsf{G} \in \mathsf{T}[\mathcal{F}]$, где $\mathsf{T} \in \mathcal{T}$. Тогда $\mathsf{M}, (w_0, v) \vDash \varphi$, где M — некоторая модель над G, $w_0 \in \mathsf{T}$, $(w_0, v) \in \mathsf{G}$.

Пусть $\mathsf{T} = (W, R)$. Для точки $w \in W$ положим $\Psi_w = \{\Diamond\psi \in Sub(\varphi) \mid \mathsf{M}, (w, v) \vDash \Diamond\psi$ для некоторой пары $(w, v) \in \mathsf{G}\}$.

Определим по индукции множества W_i. Положим $W_0 = \{w_0\}$, $\Psi_i = \bigcup\{\Psi_w \mid w \in W_i\}$.

Если $\Psi_i \neq \varnothing$, определяем W_{i+1}. Сначала для всякого $w \in W_i$ определим U_w: если $\Psi_w = \varnothing$, положим $U_w = \varnothing$; если же $\Psi_w = \{\Diamond\psi_1, \ldots, \Diamond\psi_l\}$, то полагаем $U_w = \{u_1, \ldots, u_l\}$, где u_i — R-максимальная точка в множестве
$\{u \mid u \in R(w),\ \mathsf{M}, (u, v) \vDash \psi_i$ для некоторых $(u, v) \in \mathsf{M}\}$.
Положим $W_{i+1} = \bigcup\{U_w \mid w \in W_i\}$.

Заметим, что $|\Psi_{i+1}| < |\Psi_i|$, следовательно при некотором $l < \langle\varphi\rangle$ получим $\Psi_l = \varnothing$. Полагаем $W' = W_0 \cup \cdots \cup W_l$, $V' = \{(w, v) \in \mathsf{G} \mid w \in W'\}$, $\mathsf{T}' = \mathsf{T}|W'$, $\mathsf{M}' = \mathsf{M}|V'$, $\mathsf{G}' = \mathsf{G}|V'$. По построению, $\mathsf{T}' \in \mathcal{T}_{\langle\varphi\rangle,\langle\varphi\rangle}$ и $\mathsf{G}' \in \mathcal{T}_{\langle\varphi\rangle,\langle\varphi\rangle}[\mathcal{F}]$. Кроме того, M' — селективная фильтрация M через $Sub(\varphi)$, следовательно φ $\mathcal{T}_{\langle\varphi\rangle,\langle\varphi\rangle}[\mathcal{F}]$-выполнима. ▲

Следствие 7. *Пусть $\mathcal{F}$ — некоторый класс конечных шкал и* $\mathrm{L} = \mathbf{L}(\mathcal{PO}[\mathcal{F}])$. *Тогда* L *разрешима.*

Доказательство. Разрешимость L следует из разрешимости проблемы выполнимости в любом из классов $\mathcal{T}_{n,n}[\mathcal{F}]$. ▲

Далее мы покажем, как для логик классов $\mathcal{PO}[\mathcal{F}]$ могут быть построены разрешающие процедуры, использующие лишь полиномиальную память.

4. Условная выполнимость

Для построения разрешающих PSPACE-алгоритмов введём понятие *условной истинности*, которое является незначительной модификацией стандартного понятия истинности в модели Крипке.

Определение 8. Рассмотрим модель M и множество формул Ψ. Для формулы φ и точки w модели M определим отношение истинности $(\mathsf{M}, w|\Psi) \vDash \varphi$:

$$
\begin{array}{lcl}
(\mathsf{M}, w|\Psi) \nvDash \bot & & \\
(\mathsf{M}, w|\Psi) \vDash p & \Leftrightarrow & \mathsf{M}, w \vDash p \\
(\mathsf{M}, w|\Psi) \vDash \varphi \to \psi & \Leftrightarrow & (\mathsf{M}, w|\Psi) \nvDash \varphi \text{ или } (\mathsf{M}, w|\Psi) \vDash \psi \\
(\mathsf{M}, w|\Psi) \vDash \Diamond\varphi & \Leftrightarrow & \varphi \in \Psi \text{ или } \Diamond\varphi \in \Psi \text{ или } \exists v \in R(w)\ \ (\mathsf{M}, v|\Psi) \vDash \varphi,
\end{array}
$$

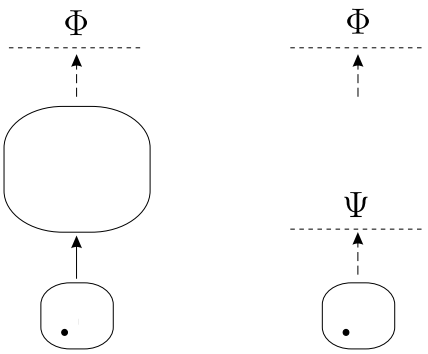
где R — отношение достижимости в модели M.

$(\mathsf{M}, w|\Psi) \vDash \varphi$ читается как "φ *истинна в точке w модели* M *при условии* Ψ".

Заметим, что $(\mathsf{M}, w|\varnothing) \vDash \varphi \Leftrightarrow \mathsf{M}, w \vDash \varphi$.

Для шкал без общих точек F_1 и F_2, пусть $\mathsf{F}_1 + \mathsf{F}_2$ и $\mathsf{F}_1 \sqcup \mathsf{F}_2$ обозначают соответственно их *порядковую* и *дизъюнктную* сумму. Напомним, что

$$\mathsf{F}_1 \sqcup \mathsf{F}_2 = (W_1 \cup W_2, R_1 \cup R_2),$$



$(\mathsf{M}_0 + \mathsf{M}, w|\Phi) \vDash \varphi$ $(\mathsf{M}_0, w|\Psi \cup \Phi) \vDash \varphi$

Рис. 2.

$$\mathsf{F}_1 + \mathsf{F}_2 = (W_1 \cup W_2, R_1 \cup (W_1 \times W_2) \cup R_2).$$

(Если $W_1 \cap W_2 \neq \varnothing$, то пусть

$$\mathsf{F}_1 + \mathsf{F}_2 = \mathsf{G}_+[\mathsf{F}_1/1][\mathsf{F}_2/2],$$

$$\mathsf{F}_1 \sqcup \mathsf{F}_2 = \mathsf{G}_\sqcup[\mathsf{F}_1/1][\mathsf{F}_2/2],$$

где $\mathsf{G}_+ = (\{1,2\}, \{(1,2)\})$, $\mathsf{G}_\sqcup = (\{1,2\}, \varnothing)$, т.е. в случае пересекающихся шкал мы рассматриваем суммы их изоморфных копий.) Если M — модель над шкалой $\mathsf{F}_1 + \mathsf{F}_2$ (над шкалой $\mathsf{F}_1 \sqcup \mathsf{F}_2$), M_1 и M_2 — модели над шкалами F_1 и F_2 соответственно, являющиеся подмоделями M, то M называется порядковой (дизъюнктной) суммой моделей M_1 и M_2.

Предложение 9. *Рассмотрим модели M_0, M, их порядковую сумму $\mathsf{M}_0 + \mathsf{M}$ и множество формул Φ. Пусть*

$$\Psi = \{\psi \in Sub(\varphi) \mid (\mathsf{M}, v|\Phi) \vDash \psi \text{ для некоторого } v\}.$$

Тогда для всяких φ, $w \in \mathsf{M}_0$ имеем

$$(\mathsf{M}_0 + \mathsf{M}, w|\Phi) \vDash \varphi \Leftrightarrow (\mathsf{M}_0, w|\Psi \cup \Phi) \vDash \varphi.$$

В частности, если

$$\Psi = \{\psi \in Sub(\varphi) \mid \mathsf{M}, v \vDash \psi \text{ для некоторого } v\},$$

то

$$\mathsf{M}_0 + \mathsf{M}, w \vDash \varphi \Leftrightarrow (\mathsf{M}_0, w|\Psi) \vDash \varphi.$$

Доказательство этого предложения проводится непосредственно, индукцией по построению формулы.

Скажем, что формула φ выполнима в конусе F при условии Ψ (в обозначениях: $\mathsf{F} \mid \Psi \Vdash \varphi$), если $(\mathsf{M}, w|\Phi) \vDash \varphi$, где M — некоторая модель над F и w — корень F. Если $\mathcal{F}$ — класс конусов, и $\mathsf{F} \mid \Psi \Vdash \varphi$ для некоторого $\mathsf{F} \in \mathcal{F}$, то скажем, что формула φ выполнима в классе $\mathcal{F}$ при условии Ψ (в обозначениях: $\mathcal{F} \mid \Psi \Vdash \varphi$).

Непосредственным следствием предложения 9 является

Предложение 10. *Рассмотрим конусы $\mathsf{F}, \mathsf{G}_1, \ldots, \mathsf{G}_n$. Пусть $\mathsf{H} = \mathsf{F} + (\mathsf{G}_1 \sqcup \cdots \sqcup \mathsf{G}_n)$. Тогда для любой формулы φ и условия Ψ следующие утверждения равносильны:*
(i) $\mathsf{H} \mid \Psi \Vdash \varphi$;
(ii) для некоторых $\Psi_1, \ldots, \Psi_n \subseteq Sub(\varphi)$ имеем: $\mathsf{F} \mid \Psi \cup \Psi_1 \cup \cdots \cup \Psi_n \Vdash \varphi$ и $\mathsf{G}_i \mid \Psi \Vdash \wedge\Psi_i$ для всех $i = 1, \ldots, n$.

Таким образом, (условная) выполнимость на шкалах вида $\mathsf{H} = \mathsf{F} + (\mathsf{G}_1 \sqcup \cdots \sqcup \mathsf{G}_n)$ сводится к проверке условной выполнимости на шкалах F и $\mathsf{G}_1, \ldots, \mathsf{G}_n$.

Определение 11. Последовательность $(\mathcal{F}_n)_{n\in\mathbb{N}}$ классов конусов называется *умеренной* (*d-умеренной*), если существует алгоритм, проверяющий для любых $\Psi, \Phi \subseteq Sub(\varphi)$ условие

$$\mathcal{F}_{\langle\varphi\rangle} \mid \Psi \Vdash \bigwedge \Phi$$

с использованием объема памяти $O(\langle\varphi\rangle^d)$, $d \in \mathbb{N}$.

Из определения условной выполнимости следует, что всякая последовательность множеств конусов, в которой каждое множество $\mathcal{F}_n$ может быть эффективно описано с использованием полиномиальной относительно n памяти, является умеренной. А именно, имеет место

Предложение 12. *Рассмотрим последовательность конечных множеств конечных конусов $(\mathcal{F}_n)_{n\in\mathbb{N}}$. Пусть существует полиномы P, Q, такие, что для всякого множества $\mathcal{F}_n$ и для всякой шкалы $(W_k^{(n)}, R_k^{(n)}) \in \mathcal{F}_n$ мощности $|\mathcal{F}_n|$ и $|W_k^{(n)}|$ ограничены $P(n)$ и вычисляются с ограничением на используемую память $Q(n)$, и пусть для любых точек $x_i, x_j \in W_k^{(n)}$ условие $x_i R_k^{(n)} x_j$ также может быть проверено с использованием $Q(n)$ памяти. Тогда $(\mathcal{F}_n)_{n\in\mathbb{N}}$ — умеренная последовательность.*

Следствие 13. *Если в последовательности множеств конусов $(\mathcal{F}_n)_{n\in\mathbb{N}}$ все множества $\mathcal{F}_n$ конечны, состоят из конечных шкал и совпадают, начиная с некоторого k, т. е. $\mathcal{F}_k = \mathcal{F}_{k+1} = \ldots$, то эта последовательность является умеренной.*

Пример 2. $(\mathcal{F}_n)_{n\in\mathbb{N}}$ является умеренной последовательностью, если:

- $\mathcal{F}_n$ состоит из всех (невырожденных) сгустков мощности не более n: для всех n $\mathcal{F}_n = \{\mathsf{C}_0, \ldots, \mathsf{C}_n\}$ или для всех n $\mathcal{F}_n = \{\mathsf{C}_1, \ldots, \mathsf{C}_n\}$;
- $\mathcal{F}_n$ состоит из единственной шкалы — одноэлементного сгустка: для всех n $\mathcal{F}_n = \{\mathsf{C}_0\}$ или для всех n $\mathcal{F}_n = \{\mathsf{C}_1\}$.

Для классов шкал $\mathcal{F}$ и $\mathcal{G}$ положим

$$\mathcal{F} + \mathcal{G} = \{\mathsf{F} + \mathsf{G} \mid \mathsf{F} \in \mathcal{F},\ \mathsf{G} \in \mathcal{G}\}.$$

Предложение 14. *Если $(\mathcal{F}_n)_{n\in\mathbb{N}}$ и $(\mathcal{G}_n)_{n\in\mathbb{N}}$ — умеренные последовательности, то $(\mathcal{F}_n + \mathcal{G}_n)_{n\in\mathbb{N}}$ — умеренная последовательность.*

Доказательство. По предложению 9. ▲

Пример 3. Если $\mathcal{F}_n = \{\mathsf{C}_0 + \mathsf{C}_1, \ldots, \mathsf{C}_0 + \mathsf{C}_n\}$, то $(\mathcal{F}_n)_{n\in\mathbb{N}}$ — умеренная последовательность.

Следующая ключевая лемма показывает, что умеренные последовательности не исчерпываются последовательностями, в которых мощности шкал растут полиномиально.

Лемма 15. *Если $(\mathcal{F}_n)_{n\in\mathbb{N}}$ — d-умеренная последовательность, то $(\mathcal{T}_{n,n}[\mathcal{F}_n])_{n\in\mathbb{N}}$ — $\max\{3, d\}$-умеренная последовательность.*

Доказательство. Пусть $h, b \geq 0$. Заметим, что для произвольного класса шкал $\mathcal{F}$ и шкалы G, если $\mathsf{G} \in \mathcal{T}_{h+1,b}[\mathcal{F}]$, то или шкала G изоморфна некоторой шкале из класса $\mathcal{F}$, или же G изоморфна шкале $\mathsf{F} + (\mathsf{G}_1 \sqcup \cdots \sqcup \mathsf{G}_{b'})$, где $1 \leq b' \leq b$, $\mathsf{F} \in \mathcal{F}$, $\mathsf{G}_1, \ldots \mathsf{G}_{b'} \in \mathcal{T}_{h,b}[\mathcal{F}]$.

Пусть процедура `SatModerate` проверяет для произвольных φ, $\Phi, \Psi \subseteq Sub(\varphi)$ условие $\mathcal{F} \mid \Psi \Vdash \bigwedge \Phi$. Тогда, в силу сделанного выше замечания и предложения 10, следующий алгоритм проверяет условие

$$\mathcal{T}_{h,b}[\mathcal{F}] \mid \Psi \Vdash \bigwedge \Phi$$

для произвольных φ, $\Phi, \Psi \subseteq Sub(\varphi)$, $h, b > 0$:

```
Function SatTree(φ; Φ, Ψ; h, b) returns boolean;
Begin
  if SatModerate(φ, Φ, Ψ) then return(true);
  if h > 1 then
   for every integer b' such that 1 ≤ b' ≤ b
    for every Φ_1, ..., Φ_b' ⊆ Sub(φ)
     if ⋀_{1≤j≤b'} SatTree(φ, Ψ_j, Ψ, h − 1, b) then
      if SatModerate(φ, Φ, Ψ ∪ Ψ_1 ··· ∪ Ψ_b') then
       return(true);
  return(false);
End.
```

Кроме того, если используемый алгоритмом `SatModerate` объём памяти ограничен $f(\langle\varphi\rangle)$, то `SatTree` использует объём памяти $O\left(f(\langle\varphi\rangle) + \langle\varphi\rangle bh\right)$. Действительно, глубина рекурсии h и на каждом уровне рекурсии `SatTree` использует объем памяти порядка $\langle\varphi\rangle b$.

Таким образом, если f — полином степени d, $n = b = h = \langle\varphi\rangle$, то для проверки условной выполнимости в классе $\mathcal{T}_{n,n}[\mathcal{F}_n]$ нам потребуется $O(n^3 + n^d)$. ▲

ТЕОРЕМА 16. *Пусть* L — *транзитивная логика, и пусть* $\mathrm{L} = \mathrm{L}(\mathcal{PO}[\mathcal{F}])$ *для некоторого класса шкал $\mathcal{F}$. Если существует умеренная последовательность $(\mathcal{F}_n)_{n\in\mathbb{N}}$, такая, что*
i) $\mathcal{F}_n \subseteq \mathcal{F}$ для всех $n \in \mathbb{N}$,
ii) всякая L*-выполнимая формула φ является $\mathcal{PO}[\mathcal{F}_{\langle\varphi\rangle}]$-выполнимой,*
то L *является PSPACE-разрешимой.*

Доказательство. Рассмотрим формулу φ, $n = \langle\varphi\rangle$. В силу леммы 6, выполнимость в классе $\mathcal{PO}[\mathcal{F}_n]$ равносильна выполнимости в классе $\mathcal{T}_{n,n}[\mathcal{F}_n]$. Поскольку (в силу леммы 15) последовательность $(\mathcal{T}_{n,n}[\mathcal{F}_n]_n)_{n\in\mathbb{N}}$ является умеренной, проверка L-выполнимости может быть осуществлена с использованием полиномиальной относительно $\langle\varphi\rangle$ памяти. Тем самым, L PSPACE-разрешима. ▲

Следствие 17. *Если* $\mathrm{L} = \mathrm{L}(\mathcal{PO}[\mathcal{F}])$ *для некоторого конечного класса конечных конусов $\mathcal{F}$, то* L *является PSPACE-разрешимой.*

Доказательство. Следует из теоремы 16 и следствия 13: достаточно положить $\mathcal{F}_n = \mathcal{F}$ для любого n. ▲

5. Примеры

Покажем на примере конкретных логик, как могут быть использованы результаты предыдущего параграфа.

Сначала рассмотрим известные логики — K4, S4, логику Гёделя-Лёба GL и логику Гжегорчика GRZ. PSPACE-разрешимость (более того, PSPACE-полнота) этих логик хорошо известна, см., например, [4, 8, 2]. Покажем, как этот результат следует из теоремы 16.

Логика GRZ (GL) является логикой класса всех частичных (строгих) порядков (см., например, [1]). Иначе этот факт можно записать так: $\mathrm{GRZ} = \mathrm{L}(\mathcal{PO}[\{C_1\}])$, $\mathrm{GL} = \mathrm{L}(\mathcal{PO}[\{C_0\}])$. В силу следствия 17, GL и GRZ PSPACE-разрешимы.

Теперь рассмотрим логики K4, S4. Пусть формула φ выполнима в некоторой транзитивной шкале F. Эта шкала изоморфна одной из шкал класса $\mathsf{G}[\mathcal{F}]$, где G — конечный частичный порядок, $\mathcal{F}$ — класс всех конечных сгустков. С помощью селективной фильтрации несложно показать, что из этого следует выполнимость формулы в классе $\mathsf{G}[\mathcal{F}_\varphi]$, где $\mathcal{F}_\varphi$ — класс всех сгустков, не превосходящих по мощности $\langle\varphi\rangle$ (доказательство этого факта можно найти, например, в [5]). Положим

$$\mathcal{F}_n^{\mathrm{K4}} = \{\mathsf{C}_0, \ldots, \mathsf{C}_n\}, \quad \mathcal{F}_n^{\mathrm{S4}} = \{\mathsf{C}_1, \ldots, \mathsf{C}_n\}.$$

Тогда для всякой формулы φ имеем:
φ K4-выполнима $\Leftrightarrow$ φ $\mathcal{PO}[\mathcal{F}^{\mathrm{K4}}_{\langle\varphi\rangle}]$-выполнима;

φ S4-выполнима $\Leftrightarrow$ φ $\mathcal{PO}[\mathcal{F}^{\mathrm{S4}}_{\langle\varphi\rangle}]$-выполнима.

Поскольку $(\mathcal{F}^{\mathrm{K4}}_n)_{n\in\mathbb{N}}$ и $(\mathcal{F}^{\mathrm{S4}}_n)_{n\in\mathbb{N}}$ — умеренные последовательности, теорема 16 даёт доказательство PSPACE-разрешимости логик K4 и S4.

Рассмотрим менее известный пример — логику *компактного включения*:

$$\mathrm{LM} = \mathrm{K4} + \Diamond\top + \Diamond p_1 \wedge \Diamond p_2 \to \Diamond(\Diamond p_1 \wedge \Diamond p_2).$$

Эта логика впервые была рассмотрена в работе [3] в связи с вопросами модальной аксиоматизации релятивистского времени, в дальнейшем было показано [6], что LM действительно является логикой отношения хронологического будущего на областях в пространствах Минковского. В работах [7, 9] было показано, что LM является логикой *компактного включения* на различных множествах регионов действительного пространства (множество A компактно содержит множество B, если замыкание B является подмножеством внутренней части A).

Покажем, как из теоремы 16 следует PSPACE-разрешимость логики LM. В [6] было показано, что LM полна относительно класса всех конечных транзитивных шкал, удовлетворяющих следующим условиям:

$$\forall x\forall y_1\forall y_2\exists z(xRy_1 \wedge xRy_2 \Rightarrow xRz \wedge zRy_1 \wedge zRy_2);$$

$$\forall x\exists y(xRy).$$

Это означает, что если F — конечна, то $\mathsf{F} \vDash \mathrm{LM}$ в том и только в том случае, если F изоморфна некоторой шкале из класса $\mathcal{PO}[\mathcal{F}]$, где

$$\mathcal{F} = \{\mathsf{C}_0 + \mathsf{C}_1, \mathsf{C}_0 + \mathsf{C}_2, \mathsf{C}_0 + \mathsf{C}_3 \dots\}.$$

Положим $\mathcal{F}^{\mathrm{LM}}_n = \{\mathsf{C}_0 + \mathsf{C}_1, \dots, \mathsf{C}_0 + \mathsf{C}_n\}$. Для всякой формулы φ имеем:

φ LM$-$выполнима $\Leftrightarrow$ φ $\mathcal{PO}[\mathcal{F}^{\mathrm{LM}}_{\langle\varphi\rangle}]-$выполнима.

Последовательность $(\mathcal{F}^{\mathrm{LM}}_n)_{n\in\mathbb{N}}$ является умеренной (предложение 14, пример 3), следовательно, LM PSPACE-разрешима.

Список литературы

[1] A. Chagrov, M. Zakharyaschev. Modal Logic. Oxford University Press, 1997.

[2] L. Farinas del Cerro, O. Gasquet. A general framework for pattern-driven modal tableaux. Logic Journal of the IGPL, 10(1):51-83, 2002.

[3] R. Goldblatt. Diodorean modality in Minkowski spacetime. Studia Logica. 1980. V. 39. № 2-3. P. 219-236.

[4] R. Ladner. The computational complexity of provability in systems of modal propositional logic. SIAM Journal of Computing, 6:467-480, 1977.

[5] I. Shapirovsky. On PSPACE-decidability in transitive modal logics. Advances in Modal Logic. 2005. V. 5. P. 269-287.

[6] I. Shapirovsky, V. Shehtman. Chronological future Modality in Minkowski spacetime. Advances in Modal Logic. 2003. V. 4. P. 437-459.

[7] I. Shapirovsky, V. Shehtman. Modal logics of regions and Minkowski spacetime. J. Logic and Comput. 2005. V.15. № 4. P. 559-574.

[8] E. Spaan. Complexity of modal logics. PhD thesis, University of Amsterdam, 1993.

[9] И. Шапировский. О модальных логиках некоторых геометрических структур. Проблемы передачи информации. 2007. Т. 43, № 3. С. 97–104.

Сложность алгоритмов проверки и поддержания степеней непротиворечивости алгебраических байесовских сетей.

Александр Владимирович Сироткин
СПИИРАН
avs@iias.spb.su

Аннотация

Алгебраические байесовские сети — одна из парадигм представления данных и знаний с вероятностной неопределенностью. В работе рассматриваются степени непротиворечивости алгебраических байесовских сетей (локальная, экстернальная, интернальная и глобальная). Приводятся алгоритмы проверки и поддержания для каждой из степеней непротиворечивости, а также рассматриваются оценки вычислительной сложности указанных алгоритмов: приводится оценка числа задач линейного программирования, которые необходимо решить, а также сложность каждой из задач, выраженная числом участвующих в ней переменных и ограничений.

1. Введение

Развитие современной вычислительной техники открывает перед человечеством широкие возможности. Но есть задачи, с которыми современные компьютеры справляются хуже человека. К таким проблемам можно отнести распознавание изображений и голоса, оценка ситуаций, игра в го и другие. Но все же, возможность обработки больших объемов данных позволяет, если не заставить компьютер самостоятельно решать различные проблемы, то, по крайней мере, сильно в этом помочь человеку. Именно такую задачу преследуют современные интеллектуальные системы поддержки принятия решений (ИСППР). Существуют разные математические модели и методы, на которых базируются те или иные ИСППР, в данной статье мы рассмотрим одну из них, а именно — алгебраические байесовские сети (АБС).

Парадигма АБС была предложена В. И. Городецким [1] в 1983 году и с тех пор поручила существенное развитие в работах А. Л. Тулупьева [3, 7]. Наиболее полно об этой парадигме можно прочитать в монографии [8], а так же в ряде других монографий [4, 5, 6, 7]. Базируясь на аппарате вероятностной логики, АБС позволяют без дополнительных доказательств и объяснений считать все результаты корректными, что не всегда можно сделать в ряде других парадигм.

В данной статье мы вкратце изложим подход Н. Нильссона к введению вероятностей над логическими формулами, опишем основной элемент АБС — фрагмент знаний, а так же покажем, как работать с совокупностью ФЗ и какое влияние они оказывают друг на друга.

Первые два раздела вводят понятийный аппарат и частично пересекаются с работой [9], в которой этот аппарат рассмотрен более подробно.

2. Основные понятия вероятностной логики

Зафиксируем алфавит $A = \{x_i\}_{i=0}^{n-1}$, состоящий из атомарных пропозициональных формул.

Определим идеал цепочек конъюнкций (идеал конъюнктов), как

$$\left\{ x_{i_1} x_{i_2} \ldots x_{i_k} \mid 0 \le i_1 < \ldots < i_k \le n-1, k \le n \right\},$$

где $x_{i_1} x_{i_2} \ldots x_{i_k}$ означает конъюнкцию соответствующих переменных; сам знак конъюнкции мы для удобства будем опускать. Сопоставим каждому конъюнкту вида $x_{i_1} x_{i_2} \ldots x_{i_k}$ число $2^{i_1} + 2^{i_2} + \ldots + 2^{i_k}$ — номер конъюнкта.

Литерал $\tilde{x}_i$ обозначает, что на его месте в формуле может стоять либо x_i, либо его отрицание $\overline{x}_i$. Тогда множество квантов над алфавитом $A = \{x_i\}_{i=0}^{n-1}$ — $Q = \{\tilde{x}_0 \tilde{x}_1 \ldots \tilde{x}_{n-1}\}$. Для нумерации квантов мы воспользуемся способом, аналогичным нумерации квантов. Выделим «положительную» часть конъюнкта (множество положительно означенных переменных) и рассмотрим ее как конъюнкт. Номер этого конъюнкта и будет номером рассматриваемого кванта. Таким образом, единице в двоичной записи номера соответствует положительное вхожде-

ние переменной, а нулю — отрицательное, при этом рассматриваются все n бит (то есть с учетом лидирующих нулей).

Пусть $F_0(A)$ — множество всех пропозициональных формул над заданным алфавитом $A=\{x_i\}_{i=0}^{n-1}$. Определим множество отличающихся формул как фактор-множество всех формул по условию эквивалентности: $F(A)=F_0(A)/\equiv$. Мощность множества $F(A)$ равна 2^{2^n}.

По теореме о совершенной нормальной дизъюнктивной форме любая пропозициональная формула эквивалентна дизюънкции некоторого набора квантов. Так как при любом означивании переменных истинен ровно один квант, то можно определить на квантах вероятность и распространить ее на все формулы, а именно вероятность произвольной формулы равна сумме вероятностей конъюнктов входящих в ее СДНФ представление.

Определение вероятности на элементах множества Q потребует следующих ограничений:

$$\forall q\in Q\ \ p(q)\ge 0;\sum_{q\in Q}p(q)=1. \qquad (1)$$

Введя обозначение $\mathbf{P_q}$ для вектора состоящего из вероятностей квантов, в порядке их нумерации получим:

$$\mathbf{P_q}\ge\mathbf{0};\ (\mathbf{1},\mathbf{P_q})=1. \qquad (2)$$

Обозначив через $\mathbf{P_c}$ вектор вероятностей конъюнктов в порядке их нумерации (напомним что нумерация начинается с нуля, нулевой конъюнкт соответствует тождественной истине, а значит нулевой элемент вектора будет единица), можно, согласно [8], определить матрицу $\mathbf{I}_n$, такую что:

$$\mathbf{P_q}=\mathbf{I}_n\times\mathbf{P_c}. \qquad (3)$$

При этом ограничения (1) принимают вид:

$$\mathbf{P_q}=\mathbf{I}_n\times\mathbf{P_c}. \qquad (4)$$

Матрица $\mathbf{I}_n$ имеет очень четкую структуру, которую удобнее всего описать рекуррентно [8].

$$\mathbf{I}_1=\begin{pmatrix}1 & -1\\ 0 & 1\end{pmatrix},\ \mathbf{I}_n=\mathbf{I}_1^{[n]}.$$

Обратная ей $\mathbf{J}_n$, удовлетворяющая условию:

$$\mathbf{P_c}=\mathbf{J}_n\times\mathbf{P_q}, \qquad (5)$$

определяется $\mathbf{J}_n=\begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix}^{[n]}$.

3. Непротиворечивость фрагмента знаний

Основным элементом АБС является фрагмент знаний (ФЗ); дадим определение его непротиворечивости:

Если оценка каждого конъюнкта задана скалярно (точечно), при этом совокупность всех оценок представлена в виде вектора оценок $\mathbf{P_c}$, тогда для проверки непротиворечивости достаточно проверить выполнение условий (4).

Если заданы два вектора — верхние и нижние оценки $\mathbf{P}^-$ и $\mathbf{P}^+$; мы говорим, что подобные оценки непротиворечивы [3, 8], если:

$$\forall i:1\le i\le 2^n-1\ \forall\varepsilon:\mathbf{P}^-[i]\le\varepsilon\le\mathbf{P}^+[i]$$
$$\exists\mathbf{P_c}:\left(\mathbf{P}^-\le\mathbf{P_c}\le\mathbf{P}^+\right)\&\qquad .$$
$$\&\left(\mathbf{P_c}[i]=\varepsilon\right)\&\left(\mathbf{I}_n\times\mathbf{P_c}\ge\mathbf{0}\right)$$

Здесь запись вида $\mathbf{P_c}[i]$ означает i-тый элемент вектора $\mathbf{P_c}$. Если выразить словами, то для любой оценки любого конъюнкта (кроме нулевого, оценка для которого всегда единица), лежащей в границах определенных векторами $\mathbf{P}^-$ и $\mathbf{P}^+$, существует совокупность оценок вероятностей всех остальных конъюнктов, лежащих в границах, определенных $\mathbf{P}^-$ и $\mathbf{P}^+$, и задающих непротиворечивый фрагмент знаний со скалярными оценками.

В случае с интервальными оценками вопрос не только в том, что непротиворечивы они или нет, но и если противоречивы, то можно ли их сузить до непротиворечивых (совместны ли они). Решения этих двух вопросов взаимосвязаны. Для проверки непротиворечивости необходимо решить серию задач линейного программирования (ЗЛП) [3, 8]. Переменными данных задач будут точечные вероятности $\mathbf{P_c}$, а ограничения будут двух типов: $\mathbf{P}^-\le\mathbf{P_c}\le\mathbf{P}^+$ и $\mathbf{I}_n\times\mathbf{P_c}\ge\mathbf{0}$. Осталось определить только целевые функции. Целевыми функциями будут максимизация и минимизация $\mathbf{P_c}[i]$ для каждого $i:1\le i\le 2^n-1$. Решение этой серии ЗЛП позволяет определить, непротиворечив ли ФЗ; в этом случае все ЗЛП будут разрешимы и соответствующие максимумы и минимумы совпадают с заданными границами. Если хотя бы одна из ЗЛП дала результат, отличный от заданных границ, то соответствующие максимумы и минимумы дадут наибольший по включению набор интервальных оценок, задающий непро-

тиворечивый ФЗ и лежащий в указанных границах. А если хоть одна из ЗЛП оказалась неразрешима, то значит такого сужения не существует и ФЗ — противоречив.

Для фрагмента знаний заданного над n переменными, нам потребуется решить $2\cdot(2^n-1)$ задачу линейного программирования, каждая из которых будет содержать 2^n-1 переменных и 2^n ограничений, вытекающих из вероятностной логики, и $2\cdot(2^n-1)$ граничных условия для вероятностей конъюнктов. Мы выделяем ограничения вида $x\le a$ или $x\ge a$ (будем называеть их граничными оценками) отдельно, от остальных, так как во многих реализациях симплекс-метода или других способах решения ЗЛП, учет подобных ограничений реализован эффективнее.

Несмотря на экспоненциальные размеры, следует заметить, что, во-первых, относительно объема заданной нам информации, эти размеры линейны, так как фрагмент знаний над n переменными содержит $2\cdot(2^n-1)$ оценки вероятностей конъюнктов. А, во-вторых, n предполагается не очень большим (порядка четырех), а, значит, сложность каждой из задач может быть оценена константой сверху.

4. Вторичная структура АБС

Перейдем от рассмотрения отдельных фрагментов знаний к их совокупности. Самое простое определение: алгебраическая байесовская сеть — это просто набор фрагментов знаний.

Но для удобства реализации, а так же для удобства структурирования нам потребуется эти фрагменты знаний организовать в некоторую структуру. Для этого определим понятия графа и дерева смежности [4].

Графом смежности — называется ненаправленный граф, в котором:

1. Между каждой парой узлов, веса которых содержат общие элементы, существует путь, такой, что в веса каждого из узлов этого пути входят все элементы, общие для начального и конечного узлов;
2. Вес одного узла не входит полностью в вес никакого другого узла.

Дерево смежности — это граф смежности, в котором каждая пара узлов соединена путем и такой путь единственный.

Для любого набора фрагментов знаний можно построить граф смежности, в узлах которого будут располагаться эти ФЗ [4, 8]. Построение подобных графов смежности из набора фрагментов знаний, в большинстве случаев происходит в процессе формирования соответствующего набора ФЗ. Мы будем предполагать, что наша АБС уже задана, как граф смежности с фрагментами знаний в узлах.

5. Степени непротиворечивости АБС

При переходе от фрагмента знаний к алгебраической байесовской сети возникает потребность определения непротиворечивости всей сети сразу. Мы выделяем четыре степени непротиворечивости [5, 8]:

- глобальная непротиворечивость;
- интернальная непротиворечивость;
- экстернальная непротиворечивость;
- локальная непротиворечивость.

Мы расположили данные степени непротиворечивости в порядке их «значимости». Но, к сожалению, значимость, в нашем случае, оказывается синонимом вычислительной сложности. Чем выше степень непротиворечивости в нашем списке, тем сложнее ее проверить. Перейдем к формальным определениям [5, 8].

АБС считается *глобально* непротиворечивой, если ее с имеющимися оценками можно погрузить в непротиворечивый объемлющий фрагмент знаний и при этом оценки на формулах из АБС не изменятся.

АБС считается *интернально* непротиворечивой, если каждый фрагмент знаний в сети непротиворечив, а также для каждой формулы из АБС для любого точечного значения из интервала оценки ее истинности можно выбрать согласованные (т.е. совпадающие на одинаковых формулах) точечные значения во всех фрагментах знаний, так, что все получившиеся ФЗ с точечными оценками будут непротиворечивы.

АБС считается *экстернально* непротиворечивой, если каждый фрагмент знаний в сети непротиворечив, а также оценки истинности каждой формулы, входящей одновременно в два фрагмента знаний или более, совпадают.

АБС считается *локально* непротиворечивой, если каждый фрагмент знаний в сети непротиворечив. Этот уровень непротиворечивости не предъявляет требований к согласованности оценок на формулах, общих для двух или нескольких фрагментов знаний.

В монографии [8] доказано, что из более сильной непротиворечивости всегда следует более слабая. В обратную же сторону утверждение не верно. В общем случае для любой пары существует пример алгебраической байесовской сети, на котором более слабая степень непротиворечивости есть, а более сильной нет.

6. Алгоритмы поддержания непротиворечивости и их сложность

Начнем с самой сильной непротиворечивости — глобальной. Как следует из ее определения, нам следует погрузить всю сеть в один фрагмент знаний и проверить его непротиворечивость.

Для этого нам потребуется решить $2k$ задач линейного программирования, каждая из которых будет содержать $2^n - 1$ переменную, 2^n линейных ограничений и $2k$ граничных оценок вероятностей. Здесь k — число разных конъюнктов, участвующих в АБС, а n — число атомарных переменных. Заметим, что это самый общий и самый сложный случай.

Если известно, что АБС — ациклическая, то можно процесс поддержания интернальной непротиворечивости, даст тот же результат что и процесс поддержания глобальной непротиворечивости [3, 5]. Если же в сети присутствую циклы, то глобально противоречивая сеть, обязательно будет интернально противоречиво, обратное в общем случае не верно.

Для поддержания интернальной непротиворечивости необходимо решить серию задач линейного программирования, но в отличие от предыдущего случая точные оценки будут зависеть от конкретной конфигурации сети. Оценим их сверху. Будем полакать, что общее число атомарных пропозиций — n, и над ними задано m фрагментов знаний; i-тый ФЗ задан над n_i переменными; а общее число различных конъюнктов в сети — k.

В таких условиях нам потребуется решить $2k$ задач линейного программирования, каждая из которых будет содержать k переменных, $2k$ граничных оценок вероятностей и $\sum_{i=1}^{i=m} 2^{n_i}$.

Если обозначить, через n_0 максимум по всем n_i, то последнюю сумму можно оценить как $m \cdot 2^{n_0}$.

Так как размеры ФЗ ограничены, то последняя оценка оказывается не экспоненциальной, а *линейной*. Число переменных, можно оценить следующим образом $m \le k \le m \cdot 2^{n_0}$.

В итоге если описанную выше точную оценку можно переформулировать как: требуется решить линейное число ЗЛП, с линейным числом ограничений и переменных, относительно количества ФЗ.

Перейдем теперь к экстернальной степени непротиворечивости. К сожалению, как показано в [2], наличие экстернальной непротиворечивости не гарантирует нам наличия интернальной непротиворечивости, но отсутствие экстернальной непротиворечивости ведет к гарантированному отсутствию интернальной и тем более глобальной непротиворечивости. Таким образом, мы сможем заметно «дешевле» выявить противоречивость сети. В случае если сеть окажется экстернально непротиворечивой, придется все таки проверять и интернальную непротиворечивость.

Опишем алгоритм проверки экстернальной непротиворечивости. Любая ациклическая алгебраическая байесовская сеть может быть представлена в виде дерева смежности с фрагментами знаний в узлах [5, 8]. В этом дереве все ребра ненаправленные, поэтому нет строгого понятия отношения родитель–сын, но если выбрать одну из вершин и зафиксировать ее как корень дерева, то в силу того, что родитель для каждой вершины в дереве единственный, а корень — родитель для всех своих соседей, то отношение родитель–сын распространяется на все дерево единственным образом. Предположим, что корень зафиксирован, и опишем алгоритм поддержания экстернальной непротиворечивости.

1. Выберем произвольный непомеченный узел дерева смежности, у которого нет, непомеченных детей. Если таких нет, то переходим к шагу 4.
2. Поддержим непротиворечивость в ФЗ соответствующем выбранному узлу, с учетом оценок всех узлов-детей.
3. Если процесс поддержания непротиворечивости успешен, то пометим данный узел и перейдем к шагу 1. В противном случае сообщаем, что сеть экстернально противоречива и выходим.
4. Снимаем все пометки и помечаем корень дерева.
5. Выберем произвольный непомеченный узел дерева смежности, у которого помечен родитель. Если таких нет, то переходим к шагу 8.
6. Поддержим непротиворечивость в ФЗ соответствующем выбранному узлу, с учетом оценок узла-родителя.
7. Если процесс поддержания непротиворечивости успешен, то пометим данный узел и перейдем к шагу 5. В противном случае сообщаем, что сеть экстернально противоречива и выходим.
8. Сообщаем, что сеть экстернально непротиворечива и выходим.

Оценим сложность данного алгоритма. Если сеть экстернально непротиворечива, нам потребуется провести процесс поддержания непроти-

воречивости ФЗ $2m-1$ раз. Будем считать, что количество атомов в i-том ФЗ — n_i, а ФЗ, соответствующий узлу-корню дерева, имеет номер один, тогда надо решить:

1. Для корня: $2(2^{n_1}-1)$ ЗЛП с $2^{n_1}-1$ переменными, $2(2^{n_1}-1)$ граничными условиями и 2^{n_1} ограничениями;
2. Для остальных вершин: $4(2^{n_i}-1)$ ЗЛП с $2^{n_i}-1$ переменными, $2(2^{n_i}-1)$ граничными условиями и 2^{n_i} ограничениями.

Но, как мы уже отмечали раньше, n_i сильно ограничены (на практике не больше четырех), таким образом, мы получаем задачи линейного программирования, вычислительная сложность которых может быть оценена константой.

По сравнению с процессом поддержания экстернальной непротиворечивости, потребуется решить в два раза больше задач линейного программирования, но каждая из задач, в соответствии с приведенными оценками будет вычислительно проще. Таким образом вычислительная сложность поддержания экстернальной непротиворечивости меньше чем сложность поддержания интернальной непротиворечивости.

Для полноты картины осталось рассмотреть только локальную степень непротиворечивости. Для ее вычисления необходимо поддержать непротиворечивость каждого ФЗ. То есть решить $4(2^{n_i}-1)$ ЗЛП с $2^{n_i}-1$ переменными, $2(2^{n_i}-1)$ граничными условиями и 2^{n_i} ограничениями.

Еще одним важным замечанием является то, что если сеть противоречива, то для случая интернальной и глобальной непротиворечивости, это выявится решении первой же ЗЛП и в таком случае, все остальные ЗЛП решать уже не потребуется. Но, несмотря на это, экстернальная непротиворечивость остается вычислительно более выгодной, так как ее сложность линейно возрастает с увеличением числа фрагментов знаний, а сложность каждой из ЗЛП в глобальной и экстернальной степенях непротиворечивости растет полиномиально от числа переменных (количество которых можно оценивать, как линейное от количества ФЗ), и этот полином заметно больше, чем линейная функция.

7. Заключение

В статье мы, находясь в рамках парадигмы алгебраических байесовских сетей, описали алгоритмы и оценили сложность, для различных видов непротиворечивости АБС. Подводя итог укажем не точные оценки сложности приведенные в статье, а порядковые. В качестве единиц в которых мы будем оценивать, возьмем число ФЗ в сети. Сводный результат приведен в табл. 1.

Табл. 1. Сводные оценки сложности алгоритмов проверки и поддержания степеней непротиворечивости

Задача	Число ЗЛП которое требуется решить	Число переменных в каждой ЗЛП	Общее число ограничений в каждой ЗЛП
Проверка глобальной непротиворечивости	1	экспоненциальное	экспоненциальное
Поддержание глобальной непротиворечивости	Экспоненциальное	Экспоненциальное	Экспоненциальное
Проверка интернальной непротиворечивости	1	Линейное	Линейное
Поддержание интернальной непротиворечивости	Линейное	Линейное	Линейное
Проверка экстернальной непротиворечивости	Линейное	Константа	Константа
Поддержание экстернальной непротиворечивости	Линейное	Константа	Константа
Проверка локальной непротиворечивости	Линейное	Константа	Константа
поддержание локальной непротиворечивости	Линейное	Константа	Константа

Таким образом подтверждается, что в случае ациклических АБС замена глобальной непротиворечивости на интернальную, дает существенный выигрыш в скорости [5, 8].

В случае, когда есть подозрения на противоречивость сети, следует вначале провести процесс поддержания экстернальной непротиворечивости, и только если он даст положительный результат переходить к более трудоемким степеням непротиворечивости.

8. Литература

[1] Городецкий В. И. «Алгебраические байесовские сети — новая парадигма экспертных систем» // *Юбилейный сборник трудов институтов Отделения информатики, вычислительной техники и автоматизации РАН*. Т. 2. М.: РАН, 1993. С. 120–141.

[2] Сироткин А.В. «Интернальная и экстернальная степени непротиворечивости ациклических алгебраических байесовских сетей» // X Санкт-Петербургская международная конференция «Региональная информатика - 2006 (РИ-2006)», Санкт-Петербург, 24–26 октября 2006 г.: Материалы конференции. СПб.: СПОИСУ, 2006. С. 52.

[3] Тулупьев А. Л. *Алгебраические байесовские сети. Логико-вероятностный подход к моделированию баз знаний с неопределенностью*. СПб.: СПИИРАН, 2000. 292 с.

[4] Тулупьев А. Л. *Алгебраические байесовские сети: локальный логико-вероятностный вывод*. СПб.: СПбГУ; Анатолия. 2007. 80 с. — (Элементы мягких вычислений.)

[5] Тулупьев А. Л. *Алгебраические байесовские сети: глобальный логико-вероятностный вывод*. СПб.: СПбГУ; Анатолия. 2007. 40 с. — (Элементы мягких вычислений.)

[6] Тулупьев А. Л. *Байесовские сети: логико-вероятностный вывод в циклах*. СПб.: Издательство С.-Петербургского университета, 2008. 140 с. — (Элементы мягких вычислений.)

[7] Тулупьев А. Л. *Алгебраические байесовские сети: теоретические основы и непротиворечивость*. СПб.: СПИИРАН, 1995. 76

[8] Тулупьев А. Л., Николенко С. И., Сироткин А. В. *Байесовские сети: логико-вероятностный подход*. СПб.: Наука, 2006. 607 с.

[9] Тулупьев А. Л., Сироткин А. В. Байесовские и марковские сети: логико-вероятностный вывод в базах фрагментов знаний с неопределенностью // Этот сборник.

Барьерно - мультипликативные алгоритмы для двойственных задач

Гасникова Евгения
Московский физико-технический институт
egasnikova@ya.ru

Аннотация

В работе предложен способ получения барьерно - мультипликативных алгоритмов для решения определенного класса задач выпуклого (вогнутого) программирования. В частности, для задач максимизации энтропии при аффинных ограничениях, которые возникают, например, при моделировании трафика в world wide web. Частными случаями семейства предложенных алгоритмов являются метод балансировки Шлейховского - Брэгмана (1934 г., 1967 г.), итерационные алгоритмы Дубова – Имельбаева - Попкова (1983 г.), Попкова (1998 г.).

1. Введение

В статье рассматривается определенный класс задач выпуклого (вогнутого) программирования. Пожалуй, основной характеристикой этого класса является возможность явно регулярно выразить решение задачи оптимизации через двойственные переменные. Отмеченная особенность полностью сводит решение исходной задачи к решению двойственной задачи. При этом в приложениях часто оказывается, что число ограничений в исходной задаче на порядок меньше размерности пространства в котором происходит оптимизация. Таким образом, двойственную задачу нужно решать в пространстве значительно меньшей размерности, чем в исходном случае. Решение двойственной задачи сводится к нахождению (для определенности) минимума выпуклой функции на неотрицательном ортанте. Для решения этой задачи используется барьерный метод (с помощью замены переменных все пространство отображается на неотрицательный ортант, и рассматривается новая задача уже безусловной минимизации). При этом барьеры выбираются мультипликативными. Далее, полученная задача безусловной минимизации решается с помощью метода Коши или метода Ньютона. Оказывается, что выбор барьеров мультипликативными обеспечивает, при выполнении некоторых условий регулярности, глобальную сходимость возникших алгоритмов.

При обосновании сходимости (глобальной сходимости) активно используется первый и второй метод Ляпунова.

Предложенные итерационные алгоритмы демонстрируются на примере известной задачи максимизации энтропии при аффинных ограничениях [1] - [3].

На примере задачи ранжирования web-страниц в поисковой системе (модель J.A. Tomlin'а [4]), продемонстрировано как могут возникать задачи максимизации энтропии при аффинных ограничениях.

2. Двойственная задача

Рассмотрим задачу выпуклого (вогнутого) программирования

$$H(\vec{x}) \to \sup_{\vec{x} \in A \cap \mathbb{R}^m_+}, \qquad (1)$$

где $\mathbb{R}^m_+ = \left\{\vec{x} \in \mathbb{R}^m : \vec{x} \geq \vec{0}_m\right\}$ (здесь, и везде в дальнейшем, неравенство $\vec{x} \geq \vec{0}_m$ - понимается покомпонентно), а выпуклое множество A определяется как:

$$\vec{\Lambda}(\vec{x}) = \vec{q} - T\vec{x} = \vec{0}_l,\ \vec{F}(\vec{x}) \geq \vec{0}_w,$$

где $l + w \leq m$ (причем $\mathrm{icr}\left(A \cap \mathbb{R}^m_+\right) \neq \varnothing$);

$$H(\vec{x}), F(\vec{x}) \in C^2\left(\mathbb{R}^m_{++}\right)\ \left(\mathbb{R}^m_{++} \overset{def}{=} \mathrm{int}\,\mathbb{R}^m_+\right)$$

- вогнутые функции, причем *матрица Гессе* функции $H(\vec{x})$ при $\vec{x} \in \mathbb{R}^m_{++}$

$$\left\| \frac{\partial^2 H(\vec{x})}{\partial x_i \partial x_j} \right\|_{\substack{i=1,\ldots,m \\ j=1,\ldots,m}} \overset{def}{=} G_H(\vec{x})$$

- (здесь, и везде в дальнейшем, i - номер строки, j - номер столбца) *отрицательно определенная*.

Кроме того, потребуем, что бы функционал $H(\vec{x})$ был *коэрцитивным* в следующем смысле

$$\forall\ c_0 \in \mathbb{R},\ \vec{c} \in \mathbb{R}^m\ \exists\ R_c > 0:$$
$$\left\{\vec{x} \in \mathbb{R}^m_+ :\ H(\vec{x}) + \langle \vec{c}, \vec{x} \rangle \geq c_0\right\} \subseteq$$
$$\subseteq \left\{\vec{x} \in \mathbb{R}^m :\ 0 \leq x_k \leq R_c,\ k = 1,..,m\right\}, \text{ где}$$

$\langle \vec{c}, \vec{x} \rangle = c_1 x_1 + ... + c_m x_m$ - скалярное произведение, порождающее евклидову норму $\|\vec{x}\| = \sqrt{\langle \vec{x}, \vec{x} \rangle}$, и

$$\left. \frac{\partial H(\vec{x})}{\partial x_k} \right|_{\vec{x} \in \mathbb{R}^m,\ x_k = 0+} = \infty,\ k = 1, ..., m.$$

Построим функцию Лагранжа для задачи (1)

$$L(\vec{x}, \vec{\lambda}, \vec{\mu}) = H(\vec{x}) + \langle \vec{\lambda}, \vec{\Lambda}(\vec{x}) \rangle + \langle \vec{\mu}, \vec{F}(\vec{x}) \rangle.$$

Предположим, что при $\vec{x} \in \mathbb{R}^m_{++}$

$$\left(\left\| \frac{\partial^2 L(\vec{x}, \vec{\lambda}, \vec{\mu})}{\partial x_i \partial \lambda_j} \right\|_{\substack{i=1,...,m \\ j=1,...,l}} ; \left\| \frac{\partial^2 L(\vec{x}, \vec{\lambda}, \vec{\mu})}{\partial x_i \partial \mu_j} \right\|_{\substack{i=1,...,m \\ j=1,...,w}} \right) =$$

$$= \left(-\mathrm{T}^T ; \left\| \frac{\partial F_j(\vec{x})}{\partial x_i} \right\|_{\substack{i=1,...,m \\ j=1,...,w}} \right) \stackrel{def}{=} M(\vec{x})$$

- матрица полного ранга $l + w$ (*условия сильной регулярности ограничений*).

Сделанные предположения позволяют утверждать (теорема Куна – Таккера [5] - [7]), что задача (1) имеет единственное решение $\vec{x}(\vec{\lambda}^*, \vec{\mu}^*)$, где $\vec{x}(\vec{\lambda}, \vec{\mu}) \in \mathbb{R}^m_{++}$ - определяется на множестве

$$(\vec{\lambda}, \vec{\mu}) \in \mathbb{R}^l \oplus \mathbb{R}^w_+$$

(притом единственным образом) из системы

$$\frac{\partial L(\vec{x}, \vec{\lambda}, \vec{\mu})}{\partial x_i} = 0,\ i = 1, ..., m,$$

причем матрица Якоби отображения $\vec{x}(\vec{\lambda}, \vec{\mu})$

$$\left\| \frac{\partial \vec{x}(\vec{\lambda}, \vec{\mu})}{\partial (\vec{\lambda}, \vec{\mu})} \right\| = \left(\left\| \frac{\partial x_i(\vec{\lambda}, \vec{\mu})}{\partial \lambda_j} \right\|_{\substack{i=1,...,m \\ j=1,...,l}} ; \left\| \frac{\partial x_i(\vec{\lambda}, \vec{\mu})}{\partial \mu_j} \right\|_{\substack{i=1,...,m \\ j=1,...,w}} \right) =$$

$$= -G_H(\vec{x}(\vec{\lambda}, \vec{\mu}))^{-1} M(\vec{x}(\vec{\lambda}, \vec{\mu})) \stackrel{def}{=} J(\vec{x}(\vec{\lambda}, \vec{\mu}))$$

- полного ранга $l + w$. Точка $(\vec{\lambda}^*, \vec{\mu}^*) \in \mathbb{R}^l \oplus \mathbb{R}^w_+$ определяется как решение двойственной к (1) задачи [5] - [7]

$$\sup_{\vec{x} \in \mathbb{R}^m} L(\vec{x}, \vec{\lambda}, \vec{\mu}) \stackrel{def}{=} \tilde{L}(\vec{\lambda}, \vec{\mu}) \to \min_{(\vec{\lambda}, \vec{\mu}) \in \mathbb{R}^l \otimes \mathbb{R}^w}. \quad (2)$$

Причем $\tilde{L}(\vec{\lambda}, \vec{\mu}) = L(\vec{x}(\vec{\lambda}, \vec{\mu}), \vec{\lambda}, \vec{\mu})$, и матрица Гессе функции $\tilde{L}(\vec{\lambda}, \vec{\mu})$ при $(\vec{\lambda}, \vec{\mu}) \in \mathbb{R}^l \oplus \mathbb{R}^w_+$

$$\begin{pmatrix} \left\| \frac{\partial^2 \tilde{L}(\vec{\lambda}, \vec{\mu})}{\partial \lambda_i \partial \lambda_j} \right\|_{\substack{i=1,...,l \\ j=1,...,l}} ; \left\| \frac{\partial^2 \tilde{L}(\vec{\lambda}, \vec{\mu})}{\partial \lambda_i \partial \mu_j} \right\|_{\substack{i=1,...,l \\ j=1,...,w}} \\ \left\| \frac{\partial^2 \tilde{L}(\vec{\lambda}, \vec{\mu})}{\partial \mu_i \partial \lambda_j} \right\|_{\substack{i=1,...,w \\ j=1,...,l}} ; \left\| \frac{\partial^2 \tilde{L}(\vec{\lambda}, \vec{\mu})}{\partial \mu_i \partial \mu_j} \right\|_{\substack{i=1,...,w \\ j=1,...,w}} \end{pmatrix} =$$

$$= J^T(\vec{x}(\vec{\lambda}, \vec{\mu})) G_H(\vec{x}(\vec{\lambda}, \vec{\mu})) J(\vec{x}(\vec{\lambda}, \vec{\mu})) -$$

$$+ M^T(\vec{x}(\vec{\lambda}, \vec{\mu})) J(\vec{x}(\vec{\lambda}, \vec{\mu})) +$$

$$+ J^T(\vec{x}(\vec{\lambda}, \vec{\mu})) M(\vec{x}(\vec{\lambda}, \vec{\mu})) =$$

$$= -M^T(\vec{x}(\vec{\lambda}, \vec{\mu})) G_H(\vec{x}(\vec{\lambda}, \vec{\mu}))^{-1} M(\vec{x}(\vec{\lambda}, \vec{\mu})) \stackrel{def}{=}$$

$$\stackrel{def}{=} G_{\tilde{L}}(\vec{\lambda}, \vec{\mu})$$ - положительно определенная.

Поэтому задача (2) имеет единственное решение, которое можно также рассматривать как единственное решение системы

$$\begin{cases} \partial \tilde{L}(\vec{\lambda}, \vec{\mu}) / \partial \lambda_k = \Lambda_k(\vec{x}(\vec{\lambda}, \vec{\mu})) = 0, \quad k = 1, ..., l; \\ \mu_k \partial \tilde{L}(\vec{\lambda}, \vec{\mu}) / \partial \mu_k = \mu_k F_k(\vec{x}(\vec{\lambda}, \vec{\mu})) = 0, \\ \partial \tilde{L}(\vec{\lambda}, \vec{\mu}) / \partial \mu_k \geq 0, \quad \mu_k \geq 0, \quad k = 1, ..., w. \end{cases} \quad (3)$$

Если $l + w \ll m$, то, естественно, вместо задачи (1) решать задачу (2) или (3). При этом в нашем распоряжении должен быть эффективный алгоритм расчета $\vec{x}(\vec{\lambda}, \vec{\mu})$ (например, явная аналитическая зависимость (см. п. 3.3)).

3. Барьерно - мультипликативные алгоритмы решения двойственной задачи

Для того чтобы строить глобально сходящиеся итерационные алгоритмы поиска решения задачи (2) или (3), воспользуемся первым и вторым методом Ляпунова (см., например, работы [3], [5], [8] - [13]).

3.1. Первый и второй метод Ляпунова

Пусть необходимо найти

$$\vec{y}^* = \arg \min_{\vec{y} \in \Upsilon \subseteq \mathbb{R}^N} \Phi(\vec{y}).$$

Будем считать, что $\vec{y}^* \in \Upsilon \subseteq \mathbb{R}^N$ является изолированным корнем уравнения

$$\vec{f}(\vec{y}) = \vec{0}_N, \text{ где } \vec{f}(\vec{y}) \in C^2(\mathbb{R}^N).$$

Построим итерационный процесс

$$\vec{y}^{n+1} = \vec{y}^n + \vec{\gamma} \otimes \vec{f}(\vec{y}^n); \quad (4)$$

$$\vec{y}^n \in \mathbb{R}^N,\ n \in \mathbb{N};\ \vec{\gamma} > \vec{0}_N$$

($\otimes$ - покомпонентное умножение векторов).

Для локальной (по начальным данным $\vec{y}^0 \in \mathbb{R}^N$) сходимости к $\vec{y}^*$ итерационного процесса (4) (со скоростью геометрической прогрессии) достаточно установить (согласно первому методу Ляпунова [5], [10], [12], [13]), что матрица Якоби преобразования $\vec{f}(\vec{y})$ в точке $\vec{y}^*$ - устойчива (гурвицева), т.е. действительные

части всех собственных значений этой матрицы отрицательны.

Для исследования области сходимости итерационного процесса (4) (при выполнении достаточного условия первого метода Ляпунова) эффективным средством является второй метод Ляпунова [3], [5], [8], [9], [11] - [13]. Наряду с итерационным процессом (4) рассматривается система обыкновенных дифференциальных уравнений (СОДУ), порождающая по схеме Эйлера итерационный процесс (4):

$$\frac{d\vec{y}}{dt} = \vec{f}\left(\vec{y}\right). \qquad (5)$$

Из обратных теорем теории устойчивости [14] следует, что в некоторой окрестности (открытое связное множество, не обязательно ограниченное) $\mathrm{O}\left(\vec{y}^*\right)$ точки $\vec{y}^*$ существует функция Ляпунова $V\left(\vec{y}\right)$ СОДУ (5). Точнее говоря, существует целый класс таких функций, причем всегда найдется такая метрика $\rho\left(\cdot,\cdot\right)$, эквивалентная обычной евклидовой, что $V\left(\vec{y}\right) = \rho\left(\vec{y}, \vec{y}^*\right)$ [15]. Напомним, что гладкая функция $V\left(\vec{y}\right)$ называется функцией Ляпунова СОДУ (5), если

$$\forall\ \vec{y} \in \mathrm{O}\left(\vec{y}^*\right) \setminus \vec{y}^* \rightarrow \dot{V}\left(\vec{y}\right) \stackrel{def}{=} \left\langle \mathrm{grad}\, V\left(\vec{y}\right),\ \vec{f}\left(\vec{y}\right)\right\rangle < 0$$

$$\dot{V}\left(\vec{y}^*\right) = 0,\ \forall\ \vec{y} \in \mathrm{O}\left(\vec{y}^*\right) \setminus \vec{y}^* \rightarrow V\left(\vec{y}\right) > V\left(\vec{y}^*\right).$$

При этом, согласно [14], от $V\left(\vec{y}\right)$ можно также требовать выполнение следующего условия:

$$\exists\ V\left(\vec{y}^*\right) < C^* \le \infty :\ \forall\ V\left(\vec{y}^*\right) < C < C^*$$

множество Лебега функции Ляпунова $V\left(\vec{y}\right)$

$$L_V\left(C\right) \stackrel{def}{=} \left\{\vec{y} \in \mathbb{R}^N :\ V\left(\vec{y}\right) \le C\right\} \text{ - ограничено.}$$

Положим

$$C^{**} = \sup\left\{C \in \left(V\left(\vec{y}^*\right), C^*\right):\ L_V\left(C\right) \subset \mathrm{O}\left(\vec{y}^*\right)\right\},$$

$$Y\left(t, C\right) = \bigcup_{\vec{y}_0 \in L_V\left(C\right)} \vec{y}\left(t; \vec{y}_0\right),\ Y_n\left(C\right) = \bigcup_{\vec{y}_0 \in L_V\left(C\right)} \vec{y}_n\left(\vec{y}_0\right),$$

где $\vec{y}\left(t; \vec{y}_0\right)$ - решение (5) с начальным условием $\vec{y}\left(0\right) = \vec{y}_0$, а $\vec{y}_n\left(\vec{y}_0\right)$ - решение (4) с тем же начальным условием. Тогда

$$\forall\ V\left(\vec{y}^*\right) \le C < C^* \rightarrow \lim_{t\rightarrow\infty} h\left(Y\left(t, C\right), \left\{\vec{y}^*\right\}\right) = 0$$

$$\forall\ V\left(\vec{y}^*\right) \le C < C^{**}\ \exists\ \vec{\gamma}_C > \vec{0}_N :\ \forall\ \vec{0}_N < \vec{\gamma} \le \vec{\gamma}_C$$

$$\lim_{t\rightarrow\infty} h\left(Y_n\left(C\right), \left\{\vec{y}^*\right\}\right) = 0, \text{ где} \qquad (6)$$

$$h\left(A, B\right) = \max\left\{\sup_{\vec{z}_1 \in A} \inf_{\vec{z}_2 \in B} \left\|\vec{z}_1 - \vec{z}_2\right\|, \sup_{\vec{z}_2 \in B} \inf_{\vec{z}_1 \in A} \left\|\vec{z}_1 - \vec{z}_2\right\|\right\}$$

- метрика Хаусдорфа.

Подчеркнем, что сходимость (6) процесса (4) к точке $\vec{y}^*$ при $\vec{y}_0 \in L_V\left(C\right)$ ($C^* > C \ge V\left(\vec{y}^*\right)$), была установлена исходя только из свойств СОДУ (5), порождающей по схеме Эйлера итерационный процесс (4). Как следствие, компоненты вектора $\vec{\gamma}_C > \vec{0}_N$ получаются при выводе (6) сильно заниженными. Для получения более точных оценок $\vec{\gamma}_C > \vec{0}_N$ часто помогают функции Ляпунова итерационного процесса (4) [5], [8], [9] (условие убывания $V\left(\vec{y}\right)$ вдоль траекторий СОДУ (5), заменяется условием убывания вдоль траекторий процесса (4)). Заметим, что при весьма общих предположениях (см., например, [5], [9]) функция Ляпунова СОДУ (5), будет функцией Ляпунова процесса (4). При этом искать функцию Ляпунова для СОДУ (5), как правило, проще чем для процесса (4).

Продемонстрируем сказанное в этом пункте на примерах [5], [12], [13].

Пример 1 (метод Коши). Положим

$$\vec{f}\left(\vec{y}\right) = -\mathrm{grad}\ \Phi\left(\vec{y}\right).$$

Будем считать, что матрица

$$G_\Phi\left(\vec{y}^*\right) = \left\|\frac{\partial^2 \Phi\left(\vec{y}^*\right)}{\partial y_i \partial y_j}\right\|_{\substack{i=1,\ldots,N \\ j=1,\ldots,N}}$$

- положительно определенная. Тогда

$$J\left(\vec{f}\left(\vec{y}^*\right)\right) = \left\|\frac{\partial f_i\left(\vec{y}^*\right)}{\partial y_j}\right\|_{\substack{i=1,\ldots,N \\ j=1,\ldots,N}} = -G_\Phi\left(\vec{y}^*\right)$$

- устойчивая матрица.

Приведем три различные функции Ляпунова

$$V_1\left(\vec{y}\right) = \Phi\left(\vec{y}\right),\ V_2\left(\vec{y}\right) = \frac{1}{2}\left\|\vec{f}\left(\vec{y}\right)\right\|^2,$$

$$V_3\left(\vec{y}\right) = \frac{1}{2}\left\|\vec{y} - \vec{y}^*\right\|^2.$$

Заметим, что методы, имеющие функцию Ляпунова $V_1\left(\vec{y}\right) = \Phi\left(\vec{y}\right)$, часто называются релаксационными методами [8] или методами спуска [7].

Пример 2 (метод Ньютона). Положим

$$\vec{f}\left(\vec{y}\right) = -G_\Phi\left(\vec{y}\right)^{-1} \mathrm{grad}\ \Phi\left(\vec{y}\right).$$

Будем считать, что матрица $G_\Phi\left(\vec{y}^*\right)$ - положительно определенная. Тогда

$$J\left(\vec{f}\left(\vec{y}^*\right)\right) = -\underbrace{\begin{Vmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 \end{Vmatrix}}_{N} \stackrel{def}{=} -I_N,$$

- устойчивая матрица.

Функцией Ляпунова будет $V_2\left(\vec{y}\right)$.

Пример 3 (барьерно - мультипликативный метод). Рассмотрим задачу (2). Положим

$$\vec{y} = \left(\vec{\lambda}, \vec{\mu}\right) \in \mathbb{R}^l \oplus \mathbb{R}^w_+ = \Upsilon,\ \Phi\left(\vec{y}\right) = \tilde{L}\left(\vec{\lambda}, \vec{\mu}\right).$$

Поскольку

$\vec{y} \in \Upsilon \Leftrightarrow \vec{y} = \left(\vec{u}^1, \vec{u}^2 - \vec{u}^2/2\right),\ \left(\vec{u}^1, \vec{u}^2\right) \in \Box^{l+w}$, то естественно перейти к новым переменным $\vec{u} = \left(\vec{u}^1, \vec{u}^2\right)$ и рассматривать задачу

$$\vec{u}^* = \arg\min_{\vec{u} \in \Box^N} \breve{\Phi}(\vec{u}),\ \text{где}\ \breve{\Phi}(\vec{u}) = \Phi\left(\vec{y}(\vec{u})\right).$$

Матрица

$$G_{\breve{\Phi}}\left(\vec{u}^*\right) = \begin{pmatrix} I_l & 0_{l\times w} \\ 0_{w\times l} & \operatorname{Diag}\left(\left\{\mu_k^*\right\}_{k=1}^w\right) \end{pmatrix} G_{\tilde{L}}\left(\vec{\lambda}^*, \vec{\mu}^*\right) +$$

$$+ \begin{pmatrix} 0_{l\times l} & 0_{l\times w} \\ 0_{w\times l} & \operatorname{Diag}\left(\left\{\dfrac{\partial \tilde{L}\left(\vec{\lambda}^*, \vec{\mu}^*\right)}{\partial \mu_k}\right\}_{k=1}^w\right) \end{pmatrix}$$

- положительно определенная, в том и только том случае, когда выполняются условия *строгой дополняющей нежесткости* (везде в дальнейшем предполагается, что эти условия выполняются)

$$\partial \tilde{L}\left(\vec{\lambda}^*, \vec{\mu}^*\right)\big/\partial \mu_k = 0 \Rightarrow \mu_k^* > 0, \quad k = 1,..., w.$$

Положим

$$\vec{\mathrm{f}}(\vec{u}) = -\operatorname{grad} \breve{\Phi}(\vec{u})$$

$$\left(\vec{\mathrm{f}}(\vec{u}) = -G_{\breve{\Phi}}(\vec{u})^{-1} \operatorname{grad} \breve{\Phi}(\vec{u})\right).$$

Рассуждая далее также как и в примере 1 (примере 2), получим, что $J\left(\vec{\mathrm{f}}\left(\vec{u}^*\right)\right)$ - устойчивая матрица и, например,

$$\breve{V}_1(\vec{u}) = \breve{\Phi}(\vec{u})\ \left(\breve{V}_2(\vec{u}) = \frac{1}{2}\left\|\vec{\mathrm{f}}(\vec{u})\right\|^2\right)$$

– функция Ляпунова.

Возвращаясь к исходным переменным, имеем

$$G_\Phi\left(\vec{y}^*\right) = G_\Phi\left(\vec{\lambda}^*, \vec{\mu}^*\right) = G_{\breve{\Phi}}\left(\vec{u}^*\right),\ \vec{f}(\vec{y}) = \vec{f}\left(\vec{\lambda}, \vec{\mu}\right) =$$

$$= -\left(\left\{\partial \tilde{L}\left(\vec{\lambda}, \vec{\mu}\right)\big/\partial \lambda_k\right\}_{k=1}^l ; \left\{\mu_k \partial \tilde{L}\left(\vec{\lambda}, \vec{\mu}\right)\big/\partial \mu_k\right\}_{k=1}^w\right)$$

$$\left(\vec{f}(\vec{y}) := G_\Phi(\vec{y})^{-1} \vec{f}(\vec{y}) = G_\Phi\left(\vec{\lambda}, \vec{\mu}\right)^{-1} \vec{f}\left(\vec{\lambda}, \vec{\mu}\right)\right),$$

$J\left(\vec{f}\left(\vec{y}^*\right)\right)$ - устойчивая матрица.

$$V_1(\vec{y}) = \Phi(\vec{y}) = \Phi\left(\vec{\lambda}, \vec{\mu}\right) = \tilde{L}\left(\vec{\lambda}, \vec{\mu}\right)$$

$$\left(V_2(\vec{y}) = V_2\left(\vec{\lambda}, \vec{\mu}\right) = \frac{1}{2}\left\|\vec{f}\left(\vec{\lambda}, \vec{\mu}\right)\right\|^2\right)$$

- функция Ляпунова.

Замечание 1. Пусть

$$\begin{cases} \dfrac{d\lambda_k}{dt} = g_k^1\left(\partial \tilde{L}\left(\vec{\lambda}, \vec{\mu}\right)\big/\partial \lambda_k\right), \quad k = 1,..., l; \\ \dfrac{d\mu_k}{dt} = g_k^2\left(\partial \tilde{L}\left(\vec{\lambda}, \vec{\mu}\right)\big/\partial \mu_k\right)\mu_k, \quad k = 1,..., w, \end{cases} \tag{7}$$

где $g_k^1(s)$, $k = 1,..., l$; $g_k^2(s)$, $k = 1,...w$ - гладкие функции при $s \in \Box$; и

$$\forall\ k = 1,..., l,\ s \in \Box \rightarrow dg_k^1(s)/ds < 0,\ g_k^1(0) = 0;$$

$$\forall\ k = 1,..., w,\ s \in \Box \rightarrow dg_k^2(s)/ds < 0,\ g_k^2(0) = 0.$$

Тогда итерационный процесс, порожденный СОДУ (7) по схеме Эйлера, локально сходится к решению задачи (2) (или, что то же самое, (3)). Для исследования области сходимости может использоваться функция Ляпунова

$$V_1\left(\vec{\lambda}, \vec{\mu}\right) = \tilde{L}\left(\vec{\lambda}, \vec{\mu}\right).$$

Необходимость в введении $\left\{g_k^1(\cdot)\right\}_{k=1}^l$, $\left\{g_k^2(\cdot)\right\}_{k=1}^w$ обусловлена, в частности, желанием сделать задачу лучше обусловленной [5].

Замечание 2. Если для решения задачи оптимизации (1) использовать двойственный метод модифицированной функции Лагранжа [12] с нулевым штрафом

$$M\left(\vec{x}, \vec{u}^1, \vec{u}^2\right) = L^\bullet\left(\vec{x}, \vec{u}^1, \vec{u}^2\right) = L\left(\vec{x}, \vec{u}^1, \vec{u}^2 - \vec{u}^2/2\right),$$

то придем к СОДУ примера 3

$$\frac{du_k^1}{dt} = -\partial \tilde{L}\left(\vec{u}^1, \vec{u}^2 - \vec{u}^2/2\right)\big/\partial u_k^1, \quad k = 1,..., l,$$

$$\frac{du_k^2}{dt} = -\partial \tilde{L}\left(\vec{u}^1, \vec{u}^2 - \vec{u}^2/2\right)\big/\partial u_k^2, \quad k = 1,..., w,$$

переходящей в (7) при $\vec{\lambda} = \vec{u}^1$, $\vec{\mu} = \vec{u}^2 - \vec{u}^2/2$; $g_k^1(s) = -s$, $k = 1,..., l$; $g_k^2(s) = -s$, $k = 1,...w$.

3.2. Глобальная сходимость барьерно – мультипликативного метода

Рассмотрим СОДУ (7). По построению (см. пример 3) траектории СОДУ (7), стартовавшие в нулевой момент времени из $\Upsilon = \Box^l \oplus \Box_+^w$, все время будут оставаться в Υ, т.е. $\left(\vec{\lambda}_0, \vec{\mu}_0\right) \in \Upsilon \Rightarrow$

$$\Rightarrow \forall\ t \geq 0 \rightarrow \left(\vec{\lambda}\left(t; \left(\vec{\lambda}_0, \vec{\mu}_0\right)\right), \vec{\mu}\left(t; \left(\vec{\lambda}_0, \vec{\mu}_0\right)\right)\right) \in \Upsilon. \tag{8}$$

Если в определении Υ заменить $\Box_+^w$ на $\Box_{++}^w$, то формула (8) по-прежнему останется верной.

На границе $\partial\Upsilon$ множества Υ (и только на ней) имеется конечное число положений равновесия СОДУ (7). Эти положения равновесия (и только они) являются одновременно нулями производной функции Ляпунова СОДУ (7) в силу системы (7)

$$\dot{V}_1\left(\vec{\lambda}, \vec{\mu}\right) = \sum_{k=1}^l g_k^1\left(\partial \tilde{L}\left(\vec{\lambda}, \vec{\mu}\right)\big/\partial \lambda_k\right)\partial \tilde{L}\left(\vec{\lambda}, \vec{\mu}\right)\big/\partial \lambda_k +$$

$$+ \sum_{k=1}^w \mu_k g_k^2\left(\partial \tilde{L}\left(\vec{\lambda}, \vec{\mu}\right)\big/\partial \mu_k\right)\partial \tilde{L}\left(\vec{\lambda}, \vec{\mu}\right)\big/\partial \mu_k \stackrel{def}{=} \delta\left(\vec{\lambda}, \vec{\mu}\right).$$

Заметим также, что

$$\forall\ C \geq V_1\left(\vec{\lambda}^*, \vec{\mu}^*\right) \rightarrow L_{V_1}(C) \cap \Upsilon$$

- ограниченное множество.

Поскольку существует, притом единственное, решение системы (3), то из отмеченных выше фактов имеем следующее утверждение.

Утверждение 1. Пусть $\left(\vec{\lambda}_0, \vec{\mu}_0\right) \in \mathbb{R}^l \oplus \mathbb{R}^w_{++}$. Тогда

$$\delta\left(\vec{\lambda}\left(t;\left(\vec{\lambda}_0, \vec{\mu}_0\right)\right), \vec{\mu}\left(t;\left(\vec{\lambda}_0, \vec{\mu}_0\right)\right)\right) \xrightarrow[t\to\infty]{} 0 \Rightarrow$$

$$\Rightarrow \left(\vec{\lambda}\left(t;\left(\vec{\lambda}_0, \vec{\mu}_0\right)\right), \vec{\mu}\left(t;\left(\vec{\lambda}_0, \vec{\mu}_0\right)\right)\right) \xrightarrow[t\to\infty]{} \left(\vec{\lambda}^*, \vec{\mu}^*\right).$$

Отсюда имеем глобальную (на $\mathbb{R}^l \oplus \mathbb{R}^w_{++}$) устойчивость положения равновесия $\left(\vec{\lambda}^*, \vec{\mu}^*\right)$ СОДУ (7) и, как следствие (этого факта и устойчивости матрицы Якоби СОДУ (7) в точке $\left(\vec{\lambda}^*, \vec{\mu}^*\right)$), глобальную сходимость итерационного процесса, порожденного СОДУ (7) по схеме Эйлера (4), т.е. $\forall$ êî ì ï àêòà $K \subset \mathbb{R}^l \oplus \mathbb{R}^w_{++}$

1) $\lim\limits_{t\to\infty} h\left(Y\left(t;K\right), \left\{\left(\vec{\lambda}^*, \vec{\mu}^*\right)\right\}\right) = 0$, где

$$Y\left(t;K\right) = \bigcup_{\left(\vec{\lambda}_0, \vec{\mu}_0\right)\in K} \left(\vec{\lambda}\left(t;\left(\vec{\lambda}_0, \vec{\mu}_0\right)\right), \vec{\mu}\left(t;\left(\vec{\lambda}_0, \vec{\mu}_0\right)\right)\right);$$

2) $\exists\ \vec{\gamma}_K > \vec{0}_{l+w}$: $\forall\ \vec{0}_{l+w} < \vec{\gamma} \le \vec{\gamma}_K$

$$\lim_{t\to\infty} h\left(Y_n\left(K\right), \left\{\left(\vec{\lambda}^*, \vec{\mu}^*\right)\right\}\right) = 0,$$

где (аналогично формуле (6))

$$Y_n\left(K\right) = \bigcup_{\left(\vec{\lambda}_0, \vec{\mu}_0\right)\in K} \left(\vec{\lambda}_n\left(\vec{\lambda}_0, \vec{\mu}_0\right), \vec{\mu}_n\left(\vec{\lambda}_0, \vec{\mu}_0\right)\right).$$

3.3. Барьерно - мультипликативный алгоритм решения задачи энтропийно – линейного программирования

Пусть в обозначениях п. 2 функционал имеет энтропийный вид, а ограничения типа неравенств также как и равенств линейные:

$$H\left(\vec{x}\right) = -\sum_{k=1}^{m} x_k \ln\left(\frac{x_k}{e}\right),\ \vec{F}\left(\vec{x}\right) = \vec{d} - G\vec{x} \ge 0,$$

и справедливы следующие **предположения.**

1) $\exists\ \vec{\vec{x}} > \vec{0}_m$: $\vec{q} - T\vec{\vec{x}} = \vec{0}_l$, $\vec{d} - G\vec{\vec{x}} \ge \vec{0}_w$.

2) Строки матриц T и G линейно независимы в совокупности (условие сильной регулярности ограничений).

3) Выполняются условия строгой дополняющей нежесткости.

Тогда

$$\vec{x}\left(\vec{\lambda}, \vec{\mu}\right) = \exp\left(-\sum_{k=1}^{l} t_{kn}\lambda_k - \sum_{k=1}^{w} g_{kn}\mu_k\right), \qquad (9)$$

$$\tilde{L}\left(\vec{\lambda}, \vec{\mu}\right) = \sum_{n=1}^{m} \exp\left(-\sum_{k=1}^{l} t_{kn}\lambda_k - \sum_{k=1}^{w} g_{kn}\mu_k\right) + \left\langle \vec{q}, \vec{\lambda}\right\rangle + \left\langle \vec{d}, \vec{\mu}\right\rangle, \qquad (10)$$

и итерационный процесс, порожденный СОДУ (7) по схеме Эйлера, глобально (на $\mathbb{R}^l \oplus \mathbb{R}^w_{++}$) сходится к решению задачи (2) (см. п. 3.2).

Этот результат, для матриц T, G с неотрицательными компонентами и векторов $\vec{q}$, $\vec{d}$ с положительными, был ранее установлен [3]

- Ю.С. Попковым (1998 г.) при

$$g_k^1\left(\tilde{L}_{\lambda_k}\right) = \ln\left(1 - \frac{1}{q_k}\tilde{L}_{\lambda_k}\right),\ k = 1,...,l;$$

$$g_k^2\left(\tilde{L}_{\mu_k}\right) = -\tilde{L}_{\mu_k},\ k = 1,...,w;$$

- Ю.А. Дубовым, Ш.С. Имельбаевым, Ю.С. Попковым (1983 г.) при

$$g_k^1\left(\tilde{L}_{\lambda_k}\right) = \ln\left(1 - \frac{1}{q_k}\tilde{L}_{\lambda_k}\right),\ k = 1,...,l;\ w = 0;$$

- А.Г. Шлейховским (1946 г.), Л.М. Брэгманом (1967 г.)

$$g_k^1\left(\tilde{L}_{\lambda_k}\right) = \ln\left(1 - \frac{1}{q_k}\tilde{L}_{\lambda_k}\right),\ k = 1,...,l;\ w = 0;$$

матрица $T = \left\|t_{kn}\right\|_{k,n=1}^{l,m}$ специального вида.

Замечание 3. Если справедливо только предположение 1, то единственное решение $\vec{x}\left(\vec{\lambda}^*, \vec{\mu}^*\right)$ задачи энтропийно – линейного программирования дается формулой (9), при этом $\left(\vec{\lambda}^*, \vec{\mu}^*\right)$ - любая из точек, которая доставляет минимум функционалу $\tilde{L}\left(\vec{\lambda}, \vec{\mu}\right)$ (10) на множестве $\left(\vec{\lambda}, \vec{\mu}\right) \in \mathbb{R}^l \oplus \mathbb{R}^w_{+}$.

4. Модель распределения WWW трафика (J.A. Tomlin, IBM, 2003)

Рассмотрим, следуя работе [4], Network flow модель, описывающую распределение WWW трафика.

Представим World Wide Web в виде ориентированного графа $G = \left\langle V, E\right\rangle$. Вершинам V этого графа соответствуют web-страницы, а ребрам E ссылки (запись $\left(i, j\right) \in E$ означает, что на i-ой страницы имеется ссылка на j-ую страницу). Обозначим через y_{ij} - число пользователей WWW, которые в единицу времени переходят по ссылки $\left(i, j\right)$. Будем считать, что WWW функционирует в установившемся режиме, что означает:

I. $\sum\limits_{(i,\ j)\in E} y_{ij} = Y = \text{const}$ -

закон сохранения активности пользователей;

II. $\forall\ i \in V \to \sum\limits_{k:(k,i)\in E} y_{ki} - \sum\limits_{k:(i,k)\in E} y_{ik} = 0$ -

закон сохранения числа посетителей i-ой страницы ($i \in V$ - произвольно).

Заметим, что

$$\forall\ (i,\ j)\in E\ \ \exists\ \breve{p}_{ij}=\frac{\left|\left\{k:\ (k,i)\in E\right\}\right|}{|Y|\cdot\left|\left\{k:\ (i,k)\in E\right\}\right|}\in \mathrm{A}\bigcap \mathbb{R}_{++}^{|E|}$$

(здесь и далее $|M|$ означает мощность множества M), где A - выпуклое множество, заданное I, II.

Заметим также, что $|V|+1$ уравнение, вообще говоря, не однозначно определяет $|E|$ неизвестных, поскольку $|E|\ \square\ |V|$. Для того чтобы однозначно определить y_{ij}, $(i,j)\in E$ введем определение.

Будем понимать под исходами всевозможные инъективные отображения $\theta:\ Y\to E$ (множества всех активных в данную единицу времени пользователей Y на множество ссылок E (ребер графа G)), такие что

$$\left\{y_{ij}=\left|\left\{k:\ \theta(k)=(i,\ j)\right\}\right|\right\}_{(i,\ j)\in E}\in \mathrm{A}.$$

Предположим, что все исходы равновероятны. Тогда вероятность трафика $\left\{y_{ij}\right\}_{(i,\ j)\in E}\in \mathrm{A}\bigcap \mathbb{R}_{+}^{|E|}$ пропорциональна

$$P\left(\left\{y_{ij}\right\}_{(i,\ j)\in E}\right)\ \square\ \left(\prod_{(i,\ j)\in E}y_{ij}!\right)^{-1}.$$

Откуда, с учетом формулы Стирлинга $\ln n!\approx n\ln n$ (обоснование см. ниже), имеем

$$\ln P\left(\left\{y_{ij}\right\}_{(i,\ j)\in E}\right)\ \square\ -\sum_{(i,\ j)\in E}y_{ij}\ln y_{ij}\ .$$

Таким образом, поиск наиболее вероятного трафика сводится к задаче максимизации энтропии

$$H\left(\left\{y_{ij}\right\}_{(i,\ j)\in E}\right)=-\sum_{(i,\ j)\in E}y_{ij}\ln y_{ij}$$

на неотрицательном ортанте $\mathbb{R}_{+}^{|E|}$ при аффинных ограничениях A.

Положим

$$\forall\ (i,\ j)\in E\to p_{ij}=y_{ij}/|Y|. \qquad (11)$$

Считая

$$|Y|\ \square\ |E|\ \square\ |V|\ \square\ 1, \qquad (12)$$

“забудем” далее про рациональность p_{ij}.

Таким образом, приходим к задаче (*):

$$H\left(\left\{p_{ij}\right\}_{(i,\ j)\in E}\right)=-\sum_{(i,\ j)\in E}p_{ij}\ln p_{ij}\to\max_{\left\{p_{ij}\right\}_{(i,\ j)\in E}\in\tilde{\mathrm{A}}\bigcap \mathbb{R}_{+}^{|E|}},$$

$$\tilde{\mathrm{A}}:\ \sum_{(i,\ j)\in E}p_{ij}=1,\ \sum_{k:(k,\ i)\in E}p_{ki}-\sum_{k:(i,\ k)\in E}p_{ik}=0,\ i\in V.$$

Из п. 3.3 и формул (11), (12) следует, что

$$\forall\ (i,\ j)\in E\to y_{ij}\ \square\ 1,$$

т.е. применение асимптотической формулы Стирлинга было законно.

Поясним теперь чем примечательно наиболее вероятное распределение $\left\{p_{ij}^{*}\right\}_{(i,\ j)\in E}$ трафика. Оказывается, в предположении равновероятности всех исходов, возникать будут в основном те распределения $\left\{p_{ij}\right\}_{(i,\ j)\in E}\in\tilde{\mathrm{A}}\bigcap \mathbb{R}_{+}^{|E|}$, которые близки к $\left\{p_{ij}^{*}\right\}_{(i,\ j)\in E}$. Действительно [3], $\forall\ M>0$

$$\forall\ \left\{p_{ij}\right\}_{(i,\ j)\in E}\in\tilde{\mathrm{A}}\bigcap \mathbb{R}_{+}^{|E|}:\ \sum_{(i,\ j)\in E}\frac{\left(p_{ij}-p_{ij}^{*}\right)^{2}}{2\max\left\{p_{ij},\ p_{ij}^{*}\right\}}\geq M$$

$$P\left(\left\{p_{ij}\right\}_{(i,\ j)\in E}\right)\leq \mathrm{e}^{-M}\ P\left(\left\{p_{ij}^{*}\right\}_{(i,\ j)\in E}\right).$$

Осталось заметить, что $|E|\ \square\ 1$. Таким образом, вогнутая функция $P\left(\left\{p_{ij}\right\}_{(i,\ j)\in E}\right)$, на множестве $\tilde{\mathrm{A}}\bigcap \mathbb{R}_{+}^{|E|}$ в точке $\left\{p_{ij}^{*}\right\}_{(i,\ j)\in E}\in\tilde{\mathrm{A}}\bigcap \mathbb{R}_{++}^{|E|}$ имеет острый пик (чем больше $|E|$, тем острее пик).

Вернемся к задаче (*). Для ее решения естественно воспользоваться результатами п. 3.3 (предварительно приведя ограничения A, путем исключения зависимых равенств, в соответствии с требованиями предположения 2). Отметим, что полученное попутно решение двойственной задачи будет иметь самостоятельную ценность. Причина заключается в том, что имеется аналогия при выводе наиболее вероятного трафика и получении статистик Больцмана, Ферми – Дирака, Бозе - Эйнштейна в термодинамике [16], [4]. При этом возникающие в термодинамике множители Лагранжа получают впоследствии интерпретацию и размерность [16] (например, величина обратная к температуре). Используя эту аналогию J.A. Tomlin предложил в 2003 г. [4] ранжировать web-страницы (в поисковой системе) по их “температуре”. Более посещаемые web-страницы “нагреты” сильнее, как следствие, имеют больший ранг.

Отметим, что в поисковой системе Google на момент времени май 2003 г. web-страницы ранжировались исходя из Марковской модели. Опишем вкратце эту модель.

Рассматривается конечная дискретная однородная Марковская цепь (ДМЦ), матрица переходных вероятностей $P=\left\|p_{ij}\right\|_{(i,\ j)\in E}$ которой представляется графом $G=\langle V,\ E\rangle$, причем

$$\forall\ (i,\ j)\in E\to p_{ij}\overset{def}{=}\alpha_{ij}\left|\left\{k:\ (i,k)\in E\right\}\right|^{-1}+\left(1-\alpha_{ij}\right),$$

где $\alpha_{ij}=\alpha\in(0,1)$, если $i=j$ и $\alpha_{ij}=0$ - иначе. Параметр α интерпретируется, как вероятность уйти с web-страницы за промежуток времени, соответствующий одному шагу ДМЦ. В случае, если полученная ДМЦ эргодична [17], финальное распределение вероятностей (на вершинах V)

совпадает со стационарным, вектор с рангами web-страниц отождествляется (с точностью до множителя) со стационарным распределением.

В заключении заметим, что проведенные численные эксперименты, часто показывали сходимость рассмотренных нами алгоритмов со скоростью геометрической прогрессии.

За счет варьирования шага удавалось улучшить скорость сходимости итерационных процессов. Подробно об оценках сложности (полиномиальности) итерационных алгоритмов, решающих задачи максимизации энтропии (энтроийно-подобных функций) при выпуклых (в частности, аффинных) ограничениях написано в работе [2].

Автор выражает благодарность Ю.С. Попкову, В.Г. Жадану и А.И. Голикову за ряд полезных ссылок.

6. Список литературы

[1] А.Дж. Вильсон, *Энтропийные методы моделирования сложных систем*, Москва, Наука, 1978.

[2] S.-C. Fang, J.R. Rajasekera, H.-S.J. Tsao, *Entropy optimization and mathematical programming*, Kluwer's International Series, 1997.

[3] Ю.С. Попков, *Теория макросистем: равновесные модели*, Москва, УРСС, 1999.

[4] J.A. Tomlin, "A new paradigm for ranking pages on the World Wide Web", *Proceedings of the WWW2003*, Budapest, Hungary, pages ACM 1-58113-680-3/03/0005, May 20-24, 2003.
http://www.vldb.org/dblp/db/indices/a-tree/t/Tomlin:John_A=.html

[5] Б.Т. Поляк, *Введение в оптимизацию*, Москва, Наука, 1983.

[6] Г.Г. Магарил-Ильяев, В.М. Тихомиров, *Выпуклый анализ и его приложения*, Москва, УРСС, 2003.

[7] А.Ф. Измаилов, М.В.Солодов, *Численные методы оптимизации*, Москва, ФИЗМАТЛИТ, 2005.

[8] Ю.И. Любич, Г.Д. Майстровский, "Общая теория релаксационных процессов для выпуклых функционалов", *Успехи матем. наук*, т. 25, № 1, 1970, стр. 57-112.

[9] В.З. Беленький, В.А.Волконский и др., *Итерационные методы в теории игр и программировании*, Москва, Наука,1974.

[10] Дж. Ортега, В. Рейнболдт, *Итерационные методы решения нелинейных систем уравнений со многими неизвестными*, Москва, Мир, 1975.

[11] Ю.Г. Евтушенко, В.Г. Жадан, "Применение метода функций Ляпунова для исследования сходимости численных методов", *Журнал вычислительной матем. и матем. физики*, т. 15, № 1, 1975, стр. 101.-112.

[12] Ю.Г. Евтушенко, *Методы решения экстремальных задач и их применение в системах оптимизации*, Москва, Наука, 1982.

[13] В.Г. Жадан, *Численные методы линейного и нелинейного программирования: вспомогательные функции в условной оптимизации*, Москва, ВЦ РАН, 2002.

[14] В.И. Зубов, *Устойчивость движения*, Москва, Высшая школа, 1973.

[15] В.И. Опойцев, "Обращение принципа сжимающих отображений", *Успехи матем. наук*, т. 31, № 4, 1976, стр. 169-198.

[16] E. Jaynes, "Information theory and statistical mechanics", *Physical Review*, v. 106, № 4, 1957, pp. 620-630.

[17] А.В. Булинский, А.Н. Ширяев, *Теория случайных процессов*, Москва, ФИЗМАТЛИТ, 2003.

Сдвиг инвариантов в алгебрах Пуассона алгебр Ли токов на прямой

Л.Г.Рыбников
ИППИ РАН и ИТЭФ, Москва.
leo.rybnikov@gmail.com.

Аннотация

Для любой конечномерной алгебры Ли $\mathfrak{g}$ строится коммутативная подалгебра в алгебре Пуассона алгебры Ли $\mathfrak{g}$-значных токов на прямой. Конструкция использует метод Ленарда–Магри построения интегралов бигамильтоновых систем. В случае полупростой алгебры Ли данная коммутативная подалгебра была получена ранее Бейлинсоном и Дринфельдом, а ее квантовая версия – Фейгиным, Френкелем и Решетихиным при помощи теории представлений соответствующей аффинной алгебры Каца–Муди. Известные коммутативные подалгебры Мищенко–Фоменко в алгебре Пуассона конечномерной алгебры Ли $\mathfrak{g}$, получаемые методом сдвига инвариантов, могут быть получены как проекции нашей коммутативной подалгебры.

1. Введение

Пусть $\mathfrak{g}$ – конечномерная комплексная алгебра Ли. Скобка Пуассона $\{\cdot,\cdot\}$ в симметрической алгебре $S(\mathfrak{g}) = \mathbb{C}[\mathfrak{g}^*]$ определяется следующим образом: на образующих $x, y \in \mathfrak{g}$ имеем $\{x, y\} = [x, y]$. *Метод сдвига инвариантов* позволяет строить коммутативные относительно скобки Пуассона подалгебры в $S(\mathfrak{g})$. Этот метод состоит в следующем. Пусть $ZS(\mathfrak{g}) = S(\mathfrak{g})^{\mathfrak{g}}$ – центр алгебры $S(\mathfrak{g})$ относительно скобки Пуассона и пусть $\mu \in \mathfrak{g}^*$. Тогда подалгебра $A_\mu \subset S(\mathfrak{g})$, порожденная элементами вида $\partial_\mu^n \Phi$, где $\Phi \in ZS(\mathfrak{g})$, (или, что то же самое, центральными элементами алгебры $S(\mathfrak{g}) = \mathbb{C}[\mathfrak{g}^*]$, сдвинутыми на $t\mu$ при всевозможных $t \in \mathbb{C}$) коммутативна относительно скобки Пуассона (см. [7]). Более того, если алгебра Ли $\mathfrak{g}$ полупроста, то подалгебры A_μ имеют максимально возможную степень трансцендентности [7] и являются максимальными (по включению) коммутативными подалгебрами в $S(\mathfrak{g})$ [12]. Подалгебры $A_\mu \subset S(\mathfrak{g})$ названы в [13] *подалгебрами Мищенко–Фоменко.*

Рассмотрим алгебру Ли $\mathfrak{g}[t] := \mathfrak{g} \otimes \mathbb{C}[t]$ $\mathfrak{g}$-значных токов на прямой. Симметрическая алгебра $S(\mathfrak{g}[t])$ есть алгебра многочленов с пространством образующих $\mathfrak{g}[t]$. В данной работе строится коммутативная подалгебра A в алгебре Пуассона $S(\mathfrak{g}[t])$ и показывается, что все подалгебры Мищенко–Фоменко $A_\mu \subset S(\mathfrak{g})$ могут быть получены как проекции этой коммутативной подалгебры.

Данная коммутативная подалгебра в случае полупростой алгебры Ли $\mathfrak{g}$ была получена ранее Бейлинсоном и Дринфельдом, а ее квантовая версия – Фейгиным, Френкелем и Решетихиным при помощи теории представлений соответствующей аффинной алгебры Каца–Муди. Эта подалгебра имеет приложения в теории интегрируемых систем, в частности, в модели Годена и других интегрируемых спиновых цепочках.

Работа выполнена при частичной поддержке грантов РФФИ 05-01-02805-CNRSL-a и 07-01-92214-CNRSL-a.

2. Конструкция

Коммутативная подалгебра в алгебре Пуассона строится при помощи следующей общей конструкции, известной как схема Ленарда–Магри [6,8]. Пусть R – коммутативная алгебра с двумя согласованными скобками Пуассона $\{\cdot,\cdot\}_1$ и $\{\cdot,\cdot\}_2$ (т.е. любая линейная комбинация $\{\cdot,\cdot\}_1$ и $\{\cdot,\cdot\}_2$ есть скобка Пуассона). Пусть Z_u – центр алгебры R относительно скобки $\{\cdot,\cdot\}_1 + u\{\cdot,\cdot\}_2$. Пусть A – подалгебра в R, порожденная всеми Z_u для u общего положения.

Факт 1. *Подалгебра $A \subset R$ коммутативна относительно каждой скобки $\{\cdot,\cdot\}_1 + u\{\cdot,\cdot\}_2$.*

Доказательство. Пусть $a \in Z_{u_1}$, $b \in Z_{u_2}$, причем $u_1 \neq u_2$. Тогда линейная по u функция $\{a, b\}_1 + u\{a, b\}_2$ обращается в нуль в двух различных точках: u_1 и u_2. Следовательно, $\{a, b\}_1 + u\{a, b\}_2 = 0$ для любого u.

Пусть теперь $a, b \in Z_{u_0}$. Тогда (в силу того, что u_0 общего положения) существует непрерывная функция $a(s)$, такая, что $a(u_0) = a$ и в некоторой окрестности $s = u_0$ выполнено $a(s) \in Z_s$. По доказанному имеем $\{a(s), b\}_1 + u\{a(s), b\}_2 = 0$ для s, принадлежащих проколотой окрестности u_0. Следовательно, $\{a, b\}_1 + u\{a, b\}_2 = 0$. □

Для любого элемента $x \in \mathfrak{g}$ мы будем обозначать элемент $x \otimes t^m \in \mathfrak{g}[t]$ через $x[m]$, а подпространство $\mathfrak{g} \otimes t^m \subset \mathfrak{g}[t]$ через $\mathfrak{g}[m]$. Таким образом, имеем $[x[n], y[m]] = [x, y][n + m]$. Зададим новую скобку Пуассона на алгебре $S(\mathfrak{g}[t])$ следующим образом

$$\{g[n], h[m]\}_1 = [g, h][n + m + 1] \qquad \forall g, h \in \mathfrak{g},\ n, m \in \mathbb{Z}_+. \quad (1)$$

Новая скобка Пуассона согласована со старой, т.е. любая линейная комбинация скобок $\{\cdot,\cdot\}$ и $\{\cdot,\cdot\}_1$ удовлетворяет тождеству Якоби (по сути, получаются две согласованные структуры алгебры Ли на пространстве $\mathfrak{g}[t]$). При этом относительно скобки $\{\cdot,\cdot\}_1$ пространство $\mathfrak{g}[t]$ образует алгебру Ли, изоморфную алгебре $\mathfrak{g}[t]_1 := \mathfrak{g} \otimes t\mathbb{C}[t]$ $\mathfrak{g}$-значных токов на прямой, обращающихся в нуль при $t = 0$. Обозначим алгебру Пуассона $S(\mathfrak{g}[t])$ со скобкой $u\{\cdot,\cdot\} + v\{\cdot,\cdot\}_1$ через $S(\mathfrak{g}[t])_{u,v}$.

Рассмотрим пополнение $\mathfrak{g}[[t]]$ алгебры Ли $\mathfrak{g}[t]$, состоящее из формальных степенных рядов с коэффициентами в алгебре Ли $\mathfrak{g}$. Скобки Пуассона $\{\cdot,\cdot\}$ и $\{\cdot,\cdot\}_1$ очевидным образом продолжаются до согласованных скобок на алгебре $S(\mathfrak{g}[[t]])$.

Заметим, что при $u \neq 0$ скобки $u\{\cdot,\cdot\} + v\{\cdot,\cdot\}_1$ задают изоморфные структуры алгебры Пуассона на $S(\mathfrak{g}[[t]])$: изоморфизм $\varphi_{u,v} : S(\mathfrak{g}[[t]])_{1,0} \tilde{\to} S(\mathfrak{g}[[t]])_{u,v}$ задается формулой

$$x[m] \mapsto u^{-1}(x[m] + \sum_{k=1}^{\infty} (-\frac{v}{u})^k x[m + k]). \quad (2)$$

Обратное отображение задается формулой

$$x[m] \mapsto ux[m] + vx[m + 1]. \quad (3)$$

Далее, обе скобки инвариантны относительно присоединенного действия алгебры Ли $\mathfrak{g}$ на алгебре $S(\mathfrak{g}[[t]])$. Таким образом, любая линейная комбинация скобок $u\{\cdot,\cdot\} + v\{\cdot,\cdot\}_1$ ограничивается на подалгебру инвариантов $S(\mathfrak{g}[[t]])^{\mathfrak{g}}$. Заметим, что при $u \neq 0$ действие алгебры Ли $\mathfrak{g}$ гамильтоново: гамильтонианом элемента $x \in \mathfrak{g}$ является $\varphi_{u,v}(x[0]) \in \mathfrak{g}[[t]]$.

Пусть $\Phi \in S(\mathfrak{g}[0])^{\mathfrak{g}}$. Тогда $\varphi_{u,v}(\Phi)$ лежит в центре алгебры Пуассона $S(\mathfrak{g}[[t]])^{\mathfrak{g}}_{u,v}$: в самом деле, любой элемент алгебры Пуассона $S(\mathfrak{g}[[t]])^{\mathfrak{g}}_{u,v}$ коммутирует с любым элементом вида $\varphi_{u,v}(x[0]) \in S(\mathfrak{g}[[t]])$, а значит и с любым полиномиальным выражением от элементов такого вида.

Из факта 1 следует, что все элементы вида $\varphi_{u,v}(\Phi)$ для всевозможных $\Phi \in S(\mathfrak{g}[0])^{\mathfrak{g}}$ попарно коммутируют относительно любой линейной комбинации скобок $u\{\cdot,\cdot\} + v\{\cdot,\cdot\}_1$. Это, в частности, означает, что коэффициенты при степенях v в разложении $\varphi_{1,v}(\Phi)$ коммутируют относительно любой линейной комбинации скобок. Пусть $\varphi_{1,v}(\Phi) = \Phi + \sum_{k=1}^{\infty} v^k \Phi^{(k)}$. Заметим, что $\Phi^{(k)}$ принадлежат алгебре Пуассона алгебры *полиномиальных* токов $S(\mathfrak{g}[t])^{\mathfrak{g}}$. Таким образом, $\Phi^{(k)}$ порождают некоторую коммутативную (относительно $\{\cdot,\cdot\}$ и $\{\cdot,\cdot\}_1$) подалгебру $A \subset S(\mathfrak{g}[t])^{\mathfrak{g}}$.

Опишем построенную нами коммутативную подалгебру $A \subset S(\mathfrak{g}[t])^{\mathfrak{g}}$ более явно. Рассмотрим дифференцирование D алгебры Пуассона $S(\mathfrak{g}[t])$, заданное на образующих следующим образом:

$$D(g[m]) = (m + 1)g[m + 1] \qquad \forall g \in \mathfrak{g}, m = 1, 2, \ldots \quad (4)$$

Предложение 1. *Подалгебра $A \subset S(\mathfrak{g}[t])$, порождена всевозможными элементами вида $D^k\Phi$, где $\Phi \in S(\mathfrak{g}[0])^{\mathfrak{g}}$, $k \in \mathbb{Z}_+$.*

Доказательство. Заметим, что для любого $f \in S(\mathfrak{g}[0])$ выполнено $\exp(-vD)f = \varphi_{1,v}(f)$. Таким образом, $\Phi^{(k)}$ суть коэффициенты при степенях v в разложении $\exp vD(\Phi)$, т. е. пропорциональны $D^k\Phi$. □

Замечание. В случае полупростой алгебры Ли $\mathfrak{g}$ коммутативная подалгебра $A \subset S(\mathfrak{g}[t])$ была описана ранее в [1]. В этом случае подалгебра $A \subset S(\mathfrak{g}[t])$ есть свободная коммутативная алгебра с образующими $D^k\Phi_i$, $i = 1, \ldots, \operatorname{rk}\mathfrak{g}$, $k = 0, 1, 2, \ldots$, где Φ_i, $i = 1, \ldots, \operatorname{rk}\mathfrak{g}$ – (свободные) образующие алгебры инвариантов $S(\mathfrak{g}[0])^{\mathfrak{g}}$. Кроме того, в этом случае подалгебра $A \subset S(\mathfrak{g}[t])$ оказывается централизатором своей квадратичной части относительно скобки $\{\cdot,\cdot\}_1$ (см. [10]). Это, в частности, означает, что она является максимальной (по включению) коммутативной подалгеброй.

3. Связь с методом сдвига инвариантов

Обозначим через $\mathfrak{g}[t]_1$ алгебру Ли, изоморфную $\mathfrak{g}[t]$ как векторное пространство, с коммутатором, задаваемым скобкой $\{\cdot,\cdot\}_1$. Для любого $z \neq 0$ имеется гомоморфизм алгебр Ли

$$\psi_z : \mathfrak{g}[t]_1 \to \mathfrak{g}, \quad x[m] \mapsto z^{m+1}x. \qquad (5)$$

Обозначим через $\mathfrak{g}_{ab}$ алгебру Ли, изоморфную $\mathfrak{g}$ как векторное пространство, но с нулевой операцией коммутатора. Имеется гомоморфизм алгебр Ли

$$\psi_0 : \mathfrak{g}[t]_1 \to \mathfrak{g}_{ab},\ x[0] \mapsto x,\ x[m] \mapsto 0, \quad m = 1, 2, \ldots \qquad (6)$$

Таким образом, для любого $z \neq 0$ имеется гомоморфизм алгебр Пуассона:

$$\psi_{z,0} : S(\mathfrak{g}[t]_1) \to S(\mathfrak{g} \oplus \mathfrak{g}_{ab}), \quad \psi_{z,0}(x[m]) = z^{m+1}x \oplus \delta_{0,m}x. \qquad (7)$$

Положим

$$A(z,0) = \psi_{z,0}(A) \subset S(\mathfrak{g} \oplus \mathfrak{g}_{ab})$$

Каждый элемент $\mu \in \mathfrak{g}^*$ задает гомоморфизм специализации $S(\mathfrak{g} \oplus \mathfrak{g}_{ab}) \to S(\mathfrak{g})$, $x \oplus y \mapsto \langle \mu, y \rangle + x\ \forall x, y \in \mathfrak{g}$ (т.е. элементы $\mathfrak{g}_{ab}$ отображаются в константы), который мы тоже будем обозначать буквой μ.

Предложение 2. *Образом коммутативной подалгебры $A(z,0) \subset S(\mathfrak{g} \oplus \mathfrak{g}_{ab})$ при гомоморфизме специализации $\mu : S(\mathfrak{g} \oplus \mathfrak{g}_{ab}) \to S(\mathfrak{g})$ для любого $\mu \in \mathfrak{g}^*$ является подалгебра $A_\mu \subset S(\mathfrak{g})$.*

Доказательство. Дадим явное описание подалгебры $A(z,0) \subset S(\mathfrak{g} \oplus \mathfrak{g}_{ab})$.

Имеем $\psi_{z,0} \circ \varphi_{1,v}(x[0]) = \frac{z}{1+zv}x \oplus x$ (равенство формальных рядов по v). Для любого $p \in \mathbb{C}$ рассмотрим гомоморфизм

$$i_p : S(\mathfrak{g}) \to S(\mathfrak{g} \oplus \mathfrak{g}_{ab}), \quad x \mapsto px \oplus x,\ \forall x \in \mathfrak{g}.$$

Пусть $\Phi \in S(\mathfrak{g})^{\mathfrak{g}}$. Мы будем обозначать тем же символом соответствующие элементы в $S(\mathfrak{g}[0])^{\mathfrak{g}} \subset S(\mathfrak{g}[t])$ и $S(\mathfrak{g} \oplus 0)^{\mathfrak{g}} \subset S(\mathfrak{g} \oplus \mathfrak{g}_{ab})$. Тогда $\psi_{z,0} \circ \varphi_{1,v}(\Phi) = i_{\frac{z}{1+zv}}(\Phi) \in S(\mathfrak{g} \oplus \mathfrak{g}_{ab})^{\mathfrak{g}}$ (т.е. образующие подалгебры $A(z,0) \subset S(\mathfrak{g} \oplus \mathfrak{g}_{ab})$ получаются подстановкой во всевозможные инварианты $\Phi \in S(\mathfrak{g})^{\mathfrak{g}}$ вместо $x \in \mathfrak{g}$ всевозможных линейных комбинаций $px \oplus x$, $p \in \mathbb{C}$).

Иначе говоря, подалгебру $A(z,0) \subset S(\mathfrak{g} \oplus \mathfrak{g}_{ab})$ можно описать следующим эквивалентным образом. Рассмотрим дифференцирование e алгебры $S(\mathfrak{g} \oplus \mathfrak{g}_{ab})$, заданное на образующих следующим образом:

$$e(x \oplus y) = 0 \oplus x. \qquad (8)$$

Тогда подалгебра $A(z,0) \subset S(\mathfrak{g} \oplus \mathfrak{g}_{ab})$ порождена элементами $e^j(\Phi)$ для всевозможных $\Phi \in S(\mathfrak{g} \oplus 0)^{\mathfrak{g}}$ и целых положительных j.

Очевидно, что для любого $f \in S(\mathfrak{g} \oplus 0)$ выполнено $\mu(e^j(f)) = \partial_\mu^j f$. Отсюда следует, что $\mu(A(z,0)) = A_\mu$. □

Замечание. Можно рассматривать также образы подалгебры $A \subset S(\mathfrak{g}[t]_1)$ в алгебры Пуассона вида $S(\mathfrak{g} \oplus \mathfrak{g} \cdots \oplus \mathfrak{g})$, при отображениях вида

$$\psi_{z_1,\ldots,z_n} : S(\mathfrak{g}[t]_1) \to S(\mathfrak{g} \oplus \mathfrak{g} \cdots \oplus \mathfrak{g}), \qquad (9)$$

где

$$\psi_{z_1,\ldots,z_n}(x[m]) = \sum_{i=1}^{n} z_i^{m+1} 0 \oplus \cdots \oplus x \oplus \cdots \oplus 0. \qquad (10)$$

(в последнем равенстве x стоит на i-ом месте).

Коммутативная подалгебра $\psi_{z_1,\ldots,z_n}(A) \subset S(\mathfrak{g} \oplus \mathfrak{g} \cdots \oplus \mathfrak{g})^{\mathfrak{g}}$ в случае полупростой алгебры Ли $\mathfrak{g}$ имеет максимально возможную степень трансцендентности и известна как алгебра (высших) гамильтонианов модели Годена. В этом случае существуют также квантовые аналоги алгебр A и $\psi_{z_1,\ldots,z_n}(A)$, являющиеся, соответственно, коммутативными подалгебрами в универсальных обертывающих алгебрах $U(\mathfrak{g}[t]_1)$ и $U(\mathfrak{g} \oplus \mathfrak{g} \cdots \oplus \mathfrak{g})$, дающие квантовые гамильтонианы модели Годена (см. [2,3,4,5,9,10,11]).

4. Список литературы

[1] A. Beilinson, V. Drinfeld, *Quantization of Hitchin's integrable system and Hecke eigensheaves*, Preprint, available at www.ma.utexas.edu/∼benzvi/BD.

[2] A. Chervov, D. Talalaev, *Quantum spectral curves, quantum integrable systems and the geometric Langlands correspondence*, preprint hep-th/0604128

[3] B.Feigin, E.Frenkel, N.Reshetikhin, *Gaudin model, Bethe Ansatz and critical level.* Comm. Math. Phys., 166 (1994), pp. 27-62.

[4] B. Feigin, E. Frenkel and V. Toledano Laredo, *Gaudin model with irregular singularities*, Preprint math.QA/0612798.

[5] М.Годен, *Волновая функция Бете*, М. "Мир 1987.

[6] F. Magri, *A simple model of the integrable Hamiltonian equation*, J.Math.Phys, 19(5): 1156–1162, 1978.

[7] А. С. Мищенко, А. Т. Фоменко, *Уравнения Эйлера на конечномерных группах Ли*, Изв. АН СССР Сер. Мат. 42 (1978), No. 2, 396–415, 471.

[8] J. Praught, R. G. Smirnov, *Andrew Lenard: a mystery unraveled.* SIGMA Symmetry Integrability Geom. Methods Appl. 1 (2005), Paper 005, 7 pp. (electronic).

[9] Л. Г. Рыбников, *Метод сдвига инвариантов и модель Годена*, Функ. Ан. Прил. 40 (2006), No. 3, 30–43 (math.RT/0606380).

[10] L. G. Rybnikov, *Uniqueness of higher Gaudin hamiltonians,* math.QA/0608588.

[11] D. Talalaev, *Quantization of the Gaudin system*, Preprint hep-th/0404153.

[12] А. А. Тарасов, *Максимальность некоторых коммутативных подалгебр в алгебрах Пуассона полупростых алгебр Ли*, УМН 57, вып.5, с.165–166.

[13] Э. Б. Винберг, *О некоторых коммутативных подалгебрах универсальной обертывающей алгебры*, Изв. АН СССР, сер. матем., 54 No 1, 3-25 (1990).

Задача одношагового управления

Мещерякова О.В.
ЛГТУ
syperolga@mail.ru

Аннотация

Существует некоторые модели управления над различными алгебраическими структурами, в которых могут не выполняться привычные законы операций. В работе предложен метод исследования и решения уравнений в таких структурах. Этот метод может быть применен для решения задачи одношагового управления.

1. Введение

Рассмотрим простейшую дискретную нестационарную динамическую систему управления

$x[t+1] = (a[t] \bullet u[t]) * x[t],\ t = t_0,\ t_0 = 1, \ldots$

где $x[t]$ – состояние системы в момент времени t

$u[t]$ – управляющее воздействие

$a[t]$ – переменный параметр системы

Система полностью задана, если задан параметр в любой момент t и начальное состояние $x[t_0] = x_0$. Прямая задача для системы заключается в том, что при заданном воздействии $u[t]$ в любой момент времени t можно определить состояние системы в любой момент.

В теории систем управления ставится обратная задача – задача одношагового управления:

$x[t_0] = x_0$ – начальное состояние;

$x[t_0 + 1]$ - желаемое конечное состояние;

$a[t_0] = a$ – значение параметра;

требуется определить значение управления $u[t_0]$, так чтобы под его воздействием система перешла из состояние x_0 в состояние $x[t_0 + 1]$:

$x[t_0 + 1] = (a[t_0] \bullet u[t_0]) * x[t_0]$

В обычных операциях, для которых выполняются привычные законы (ассоциативность, коммутативность, дистрибутивность), такие системы и задача одношагового управления для них являются предельно простыми.

В [1] рассмотрены некоторые модели управления над различными алгебраическими структурами, в которых могут не выполняться привычные законы операций. Это в частности модели над некоторыми кольцами, полукольцами, группами. При формировании модели системы управления используются одна или две алгебраических операции – поэтому необходимо сформулировать модель системы управления над алгебраической структурой с одной или двумя операциями – над группоидом или бигруппоидом, причем с минимальными требованиями к операциям, как крайний случай – вообще без каких-либо требований к операциям.

Для решения задачи одношагового управления в этих структурах необходимо разработать методику исследования и решения уравнений. Если мы введем соответствующие обозначения: $a[t_0] = a$, $u[t_0] = x$, $x[t_0] = b$, $x[t_0 + 1] = c$, мы приведем уравнение, описывающее обратную задачу управления $x[t_0 + 1] = (a[t_0] \bullet u[t_0]) * x[t_0]$ к виду $(a \bullet x) * b = c$.

2. Метод исследования и решения уравнений

В [2] предложен метод исследования и решения односторонних уравнений вида $a \bullet x = b$ в группоидах, а также двусторонних уравнений вида $(a \bullet x) * b = c$ в бигруппоидах. Произвольный группоид определяется заданием операции $\bullet$ над элементами некоторого множества S и обозначается $\tilde{A} = \langle S, \bullet \rangle$. Бигруппоид $B\tilde{A}_{\bullet,*} = \langle S, \bullet, * \rangle$ определяется заданием двух операций.

Для исследования и получения частного решения уравнения $a \bullet x = b$ в группоиде $\tilde{A} = \langle S, \bullet \rangle$ к коэффициенту a следует предъявить требование, связывающее его с элементом $a^{\bullet}$, который является сопровождающим

относительно операции $\bullet$, и элементами $s \in S$ группоида

$$(\nabla a)(\exists a^{\bullet})(\forall s)\{a \bullet (a^{\bullet} \bullet (a \bullet s)) = a \bullet s\}. \quad (1)$$

Предположим уравнение $a \bullet x = b$ разрешимо, то есть существует x_0, такой что $a \bullet x_0 = b$. Применяя (1) с $s = x_0$, получим $b = a \bullet x_0 = a \bullet (a^{\bullet} \bullet (a \bullet x_0)) = a \bullet (a^{\bullet} \bullet b)$. Это есть необходимое и достаточное условие разрешимости уравнения. Это условие является критерием разрешимости уравнения, так указывает частное решение:

$$x_* = a^{\bullet} \bullet b. \quad (2)$$

Для получения общего решения предъявим требование о наличии сопровождающей операции $\times$. Рассмотрим сопровождающий бигруппоид $B\tilde{A}_{\bullet,\times} = \langle S, \bullet, \times \rangle$ и индивидуально-общие требования:

$$(\nabla a)(\nabla a^{\bullet})(\exists a^{\times})(\forall s)\{(a^{\bullet} \bullet (a \bullet s)) \times (a^{\times} \bullet s) = s\} \quad (3)$$

$$(\nabla a)(\exists a^{\times})(\forall t, s)\{a \bullet (t \times (a^{\times} \bullet s)) = a \bullet t\} \quad (4)$$

применение $a^{\times}$ к обеим частям (3) и использование (4) с $t = a^{\bullet} \bullet (a \bullet s)$ дает (1).

Исходя из этих требований общее решение данного уравнения записывается в виде:

$$x = x(f) = (a^{\bullet} \bullet b) \times (a^{\times} \bullet f),\ f \in S \quad (5)$$

где f - произвольный элемент сопровождающего бигруппоида $B\tilde{A}_{\bullet,\times}$.

Рассмотрим уравнение $x * a = b$. Для исследования и решения этого уравнения также необходимо предъявить требование о наличии сопровождающего элемента

$$(\nabla a)(\exists^{*} a)(\forall s)\{((s * a) *^{*} a) * a = a\} \quad (6)$$

Исходя из этого требования, критерий разрешимости уравнения записывается в виде:

$$(b *^{*} a) * a = b \quad (7)$$

частное решение

$$x_* = b *^{*} a. \quad (8)$$

Требования

$$(\nabla a)(\nabla^{*} a)(\exists^{\circ} a)(\forall s)\{(s *^{\circ} a) \circ ((s * a) *^{*} a) = s\} \quad (9)$$

$$(\nabla a)(\exists^{\circ} a)(\forall t, s)\{((s *^{\circ} a) \circ t) * a = t * a\} \quad (10)$$

делают возможным запись общего решения:

$$x = x(k) = (k *^{\circ} a) \circ (b *^{*} a), k \in S \quad (11)$$

где k - произвольный элемент сопровождающего бигруппоида $B\tilde{A}_{*,\circ} = \langle S, *, \circ \rangle$.

Рассмотрим двустороннее уравнение $(a \bullet x) * b = c$ в основном бигруппоиде $BG_{\bullet,*} = \langle S, \bullet, * \rangle$. Двустороннее уравнение делится на два: правостороннее $y * b = c$ и левостороннее $a \bullet x = y$. Аналогично вышеизложенному, производится исследование и находятся решения каждого уравнения.

Критерий разрешимости первого уравнения $(c *^{*} b) * b = c$, его общее решение $y = y(k) = (k *^{\circ} b) \circ (c *^{*} b)$.

Критерий разрешимости второго уравнения:

$$a \bullet (a^{\bullet} \bullet ((k *^{\circ} b) \circ (c *^{*} b))) = (k *^{\circ} b) \circ (c *^{*} b).$$

Критерий разрешимости двустороннего уравнения:

$$(a \bullet (a^{\bullet} \bullet ((k *^{\circ} b) \circ (c *^{*} b)))) * b = ((k *^{\circ} b) \circ (c *^{*} b)) * b = c$$

Общее решение двустороннего уравнения в сопровождающем тетрагруппоиде $\grave{O}\tilde{A}_{\bullet,*,\times,\circ} = \langle S, \bullet, *, \times, \circ \rangle$:

$$x = (a^{\bullet} \bullet ((k *^{\circ} b) \circ (c *^{*} b))) \times (a^{\times} \bullet g) \quad (12)$$

Этот подход позволяет найти решение двустороннего уравнения в различных алгебраический структурах.

Таким образом, мы можем решить обратную задачу управления, определить значение управления $x = u[t_0]$. Критерий разрешимости уравнений интерпретируется как критерий одношаговой управляемости данной системы, общее решение уравнения интерпретируется как синтез алгоритма одношагового управления данной системы.

3. Заключение

В современной прикладной математике, в частности, в искусственном интеллекте, при математическом моделировании объектов и процессов, все чаще встречаются, операции, для которых нарушаются привычные законы. Предлагаемый подход охватывает подобные ситуации и может быть применен для решения задачи одношагового управления.

4. Литература

[1] Ванечек А. Модели: эквивалентность, применения, обобщения // Современные методы идентификации систем. – М.: Мир, 1983. – С. 11 – 73.

[2] Блюмин С Л., Мещерякова О.В. Исследование и решение в группоидах уравнений с коэффициентами, обладающими свойствами обратимости / С Л.Блюмин, О. В.Мещерякова // Вести высш. учебн. заведений Черноземья. – Липецк, ЛГТУ. – 2007.№1. – С. 3-7.

[3] Keedwell A.D. Construction, properties and applications of finite neofields: Comment.Math.Univ.Carolinae 41,2 (2000). – P. 283-297.

Применение идемпотентного принципа соответствия для решения задач теории систем массового обслуживания

Жбанов С. А.
Липецкий государственный технический университет
lisky@lipetsk.ru

Аннотация

Работа посвящена приложению модельной алгебраической структуры R_{max} из семейства идемпотентных полуколец. Идемпотентная (в зарубежных источниках её чаще называют тропическая) математика за короткий срок своего существования сумела обеспечить глубокое понимание и изящное решение многих проблем оптимизации и управления, поэтому применение идемпотентного принципа соответствия для решения задач теории систем массового обслуживания (СМО) отвечает не только требованиям актуальности, но и содержит в себе новизну подхода, новаторство.

1. Введение

Термин “идемпотентность” сейчас известен каждому старшекласснику из учебника по информатике. Этим термином обозначают свойство бинарной операции $*$ такое, что $x * x = x$, где x любое. Самые известные идемпотентные операции – это взятие максимума и минимума.

В основе идемпотентной математики лежит замена обычных арифметических операций новым набором базовых операций (такими как максимум и минимум), при этом числовые поля заменяются идемпотентными полукольцами и полуполями.

Идемпотентная математика за короткий период своего существования уже имеет ряд преимуществ перед традиционным полем вещественных чисел, за счёт которых её использование может быть широко распространено во многих областях человеческих знаний.

Идемпотентный принцип соответствия состоит в том, что многие практически важные задачи, в частности, оптимизационные проблемы, в действительности являются линейными, но не в обычном смысле, а над правильно подобранной идемпотентной алгеброй. Использование этого идемпотентного принципа соответствия позволяет широко применять универсальные алгоритмы линейной алгебры и обеспечивает очень большое богатство приложений [1].

К настоящему времени идемпотентная теория достигла значительного развития. Она включает, в частности, новую теорию интегрирования, новую линейную алгебру, спектральную теорию и функциональный анализ. Её приложения охватывают различные задачи оптимизации, такие как многокритериальное принятие решений, оптимизация на графах, дискретная оптимизация с большим параметром (асимптотические задачи), оптимальное конструирование компьютерных систем и вычислительных сред, оптимальная организация параллельной обработки данных, динамическое программирование [4], приложения к дифференциальным уравнениям, численному анализу, системам дискретных событий, информатике, дискретной математике, математической логике и так далее. Интересные результаты получены для ряда задач теории систем массового обслуживания.

2. Алгебраические структуры

Коммутативным кольцом называется множество S, снабжённое двумя операциями $\oplus$ и $\otimes$ (сложение и умножение соответственно), удовлетворяющее следующим свойствам:

1) операции $\oplus$ и $\otimes$ ассоциативны, то есть $(a \oplus b) \oplus c = a \oplus (b \oplus c)$ и $(a \otimes b) \otimes c = a \otimes (b \otimes c)$, $\forall a, b, c \in S$;

2) операции $\oplus$ и $\otimes$ коммутативны, то есть $a \oplus b = b \oplus a$ и $a \otimes b = b \otimes a$, $\forall a, b \in S$;

3) существует элемент $\mathbf{0}$ (нейтральный элемент для сложения) такой, что $a \oplus \mathbf{0} = a$, $\forall a \in S$;

4) существует элемент $\mathbf{1}$ (нейтральный элемент для умножения) такой, что $a \otimes \mathbf{1} = a$, $\forall a \in S$;

5) операция $\otimes$ дистрибутивна относительно операции $\oplus$, то есть $a \otimes (b \oplus c) = (a \otimes b) \oplus (b \otimes c)$, $\forall a, b, c \in S$;

6) элемент **0** обладает свойством поглощения, то есть $a \otimes \mathbf{0} = \mathbf{0},\ \forall a \in S$;

В данной работе будет рассматриваться структура, в которой, наряду с приведёнными выше шестью свойствами, будет выполняться дополнительное свойство идемпотентности операции $\oplus$, то есть $a \oplus a = a,\ \forall a \in S$. Введём сокращённую запись полукольца в виде ($S, \oplus, \otimes, \mathbf{0}, \mathbf{1}$) [2].

Идемпотентное полуполе R_{max} в данном контексте можно записать в виде $\left(R \cup \{-\infty\}, \max, +, -\infty, 0\right)$, то есть рассматривается множество действительных чисел, к которому присоединён дополнительный наименьший элемент $-\infty$, с операциями $a \oplus b = \max\{a,b\}$ и $a \otimes b = a + b$ ("сложение" и "умножение"). Элементы $-\infty$ и 0 являются нейтральными для "сложения" и "умножения" соответственно. Каждый элемент относительно операции $\otimes$ имеет симметричный к себе элемент (особенность полуполя), то есть $a \otimes a^{-1} = a^{-1} \otimes a = 0,\ \forall a \in R_{\max}$, но ни один элемент (кроме $-\infty$) не имеет симметричного к себе элемента относительно операции $\oplus$.

3. Тропическая алгебра матриц

Прямоугольная таблица из m строк и n столбцов, заполненная элементами идемпотентного полукольца ($S, \oplus, \otimes$, **0**, **1**), называется матрицей над идемпотентным полукольцом размера $m \times n$ или тропической матрицей размера $m \times n$. Рассмотрим основные свойства матриц над полуполем R_{max}.

Квадратная матрица над R_{max}, элементы которой удовлетворяют условию:

$$a_{ij} = \begin{cases} a_{ij} \neq -\infty, & i = j \\ a_{ij} = -\infty, & i \neq j, \end{cases}$$

называется диагональной тропической матрицей.

Диагональная матрица порядка n называется единичной над полуполем R_{max}, если все её элементы главной диагонали равны **0**. Матрица любого размера называется нулевой, если все её элементы равны $-\infty$. Такие матрицы имеют вид:

$$I = \begin{pmatrix} 0 & \cdots & -\infty \\ \cdots & \cdots & \cdots \\ -\infty & \cdots & 0 \end{pmatrix};\ 0 = \begin{pmatrix} -\infty & \cdots & -\infty \\ \cdots & \cdots & \cdots \\ -\infty & \cdots & -\infty \end{pmatrix}.$$

Суммой матриц $A = (a_{ij})$ и $B = (b_{ij})$ размеров $m \times n$ над полуполем R_{max} называется матрица $C = (c_{ij})$ тех же размеров такая, что $c_{ij} = a_{ij} \oplus b_{ij}$ для всех i и j. Таким образом, чтобы сложить матрицы A и B над полуполем R_{max} – нужно взять максимум из элементов, стоящих на одинаковых местах.

Произведение матрицы A на число λ над полуполем R_{max} называется матрица $\lambda A = (\lambda \otimes a_{ij})$, получаемая умножением всех элементов матрицы A на число λ. Таким образом, умножение матрицы на число в R_{max} соответствует простому прибавлению числа ко всем элементам матрицы.

Произведением матрицы $A = (a_{ij})$ размера $m \times k$ и матрицы $B = (b_{jk})$ размера $k \times n$ над полуполем R_{max} называется матрица $C = (c_{ik})$ размера $m \times n$ такая, что $c_{ik} = \bigoplus_{j=1}^{k} a_{ij} \otimes b_{jk}$. Таким образом, элемент произведения матриц A и B, стоящий в i-ой строке и k-ом столбце над полуполем R_{max}, равен максимуму из сумм элементов i-ой строки матрицы A и соответствующих элементов k-ого столбца второй матрицы B.

Операция транспонирования тропических матриц не отличается от обычного транспонирования, и основные свойства не теряют своей силы в R_{max}.

Отличие же между тропическими и обычными матрицами заключается в том, что ни для одной тропической матрицы A (кроме нулевой) не существует противоположной матрицы $(-A)$ такой, что $A \oplus (-A) = 0$. Отсутствие обратной операции к взятию максимума не позволяет решить задачу нахождения обратной матрицы для произвольной квадратной тропической матрицы, так как продолжение построения матричной теории для полуполя R_{max} по стандартному образцу (как для поля R) базируется на существовании операции обратной к "сложению".

4. Постановка задачи

Рассмотрим работу сервисной службы поддержки [5], которая представляет собой многоканальную СМО с отказами. Каждая заявка k $(k = 1,2,\ldots,N)$, поступающая в службу поддержки должна быть обслужена одним из операторов W_i $(i = 1,\ldots,M)$. В данной СМО отсутствует очередь. Следует также отметить, что новая заявка автоматически направляется к свободному оператору с меньшим индексом i. Для того чтобы обслужить заявку k, оператору W_i необходимо $\tau_i(k)$ тактов времени, причём $\tau_i(k)$ распределена по показательному закону.

Теперь обозначим через $x_i(k)$ момент начала работы оператора W_i с заявкой k. Прежде, чем W_i сможет начать обслуживание следующей заявки, должны быть выполнены следующие условия: W_i свободен, W_j занят $(\forall j < i)$.

Предположим теперь, что компания несёт убытки p_i за каждый такт времени содержания операторов. Необходимо определить количество операторов M таким образом, чтобы издержки компании были минимальны, но при этом среднее число отказов не должно превышать d %.

5. Задача о СМО с последовательным обслуживанием заявок

Прежде чем перейти к решению поставленной задачи, рассмотрим существующие подходы к решению задач теории СМО базирующиеся на идемпотентном принципе соответствия.

Рассматривается система массового обслуживания [3], состоящая из четырёх последовательных серверов S_i $(i = 1, \ldots, 4)$ и буфера, ассоциированного с первым сервером (показано на рис. 1).

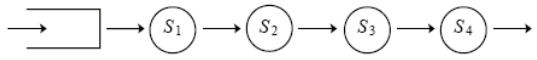


Рис. 1. СМО с последовательным обслуживанием

Каждый клиент k $(k = 1, 2, \ldots, N)$ должен быть обслужен всеми четырьмя серверами в определённой последовательности: S_1, S_2, S_3 и S_4. Ограниченное количество клиентов прибывает в буфер системы массового обслуживания одновременно, и каждый ожидает своей очереди на обслуживание l_k. Через $k[l_k]$ обозначается клиент, который занимает в очереди место $[l_k]$. Для того, чтобы обслужить клиента k, серверу S_i необходимо $\tau_i(k[l_k])$ тактов времени. Между серверами буфера отсутствуют. Следовательно, если сервер S_i, $i = 1, 2, 3$, уже закончил обслуживание клиента $k[l_k]$, а S_{i+1} всё ещё занят с клиентом $k[l_k - 1]$, то S_i не может начать работу с клиентом $k[l_k + 1]$ и должен ждать. Время перехода от сервера к серверу положгается нулевым.

Через $x_i(k[l_k])$ обозначается момент начала работы сервера S_i с клиентом k. Прежде, чем S_i сможет начать обслуживание клиента $k[l_k + 1]$, должны быть выполнены следующие условия:

а) S_i закончил обслуживание клиента $k[l_k]$;

б) S_{i+1} свободен (для $i = 4$ это условие не применяется);

в) S_{i-1} закончил обслуживание клиента $k[l_k + 1]$ (для $i = 1$ это условие не применяется);

Формализовав эти условия, записывается уравнение в терминах идемпотентного анализа, которому удовлетворяет вектор x, состоящий из четырёх компонент:

$$x(k[l_k+1]) = \begin{pmatrix} -\infty & \cdots & -\infty & -\infty \\ \tau_1(k[l_k+1]) & \cdots & -\infty & -\infty \\ \cdots & \cdots & \cdots & \cdots \\ -\infty & \cdots & \tau_3(k[l_k+1]) & -\infty \end{pmatrix} \otimes$$
$$\otimes x(k[l_k+1]) \oplus \begin{pmatrix} \tau_1(k[l_k]) & \cdots & -\infty & -\infty \\ -\infty & \cdots & 0 & -\infty \\ -\infty & \cdots & \tau_3(k[l_k]) & 0 \\ -\infty & \cdots & -\infty & \tau_4(k[l_k]) \end{pmatrix} x(k[l_k]). \quad (1)$$

Здесь, как и прежде, символ $\oplus$ обозначает идемпотентное сложение, а формула (1), которую можно кратко записать

$$x(k[l_k+1]) = A_2(k[l_k+1], k[l_k+1]) \otimes$$
$$\otimes x(k[l_k+1]) \oplus A_1(k[l_k+1], k[l_k]) x(k[l_k]) \quad (2)$$

является неявным уравнением относительно $x(k[l_k+1])$, решение которого имеет вид

$$x(k[l_k+1]) = A_2(k[l_k+1], k[l_k+1])^* \otimes$$
$$\otimes A_1(k[l_k+1], k[l_k]) x(k[l_k]) \quad (3)$$

где $(A_2(k[l_k+1], k[l_k+1]))^*$ соответствует матрице

$$\begin{pmatrix} 0 & \cdots & -\infty & -\infty \\ \tau_1(k[l_k+1]) & \cdots & -\infty & -\infty \\ \tau_1(k[l_k+1])\tau_2(k[l_k+1]) & \cdots & 0 & -\infty \\ \tau_1(k[l_k+1])\tau_2(k[l_k+1])\tau_3(k[l_k+1]) & \cdots & \tau_3(k[l_k+1]) & 0 \end{pmatrix}$$

Приведённый подход применим для достаточно широкого спектра задач теории СМО, где заявки должны обслуживаться несколькими каналами обслуживания последовательно (например, задачи сетевого планирования, конвейерного производства и т. д.). Но для задачи оптимизации работы сервисной службы поддержки такой подход не годится, т. к. заявки обслуживаются распараллеленными каналами (операторами службы).

6. Построение основной модели

Прежде чем перейти к построению модели работы сервисной службы поддержки, следует выявить отличия данной задачи от рассмотренной выше. Первое и самое существенное отличие состоит в том, что работа операторов службы распараллелена, то есть пришедшая заявка может быть обслужена любым

из операторов W_i ($i=1,\ldots,M$), а не должна обслуживаться каждым из M операторов, как в случае с серверами. Также в данной системе нет никаких ассоциированных буферов и накопителей. Это означает, что заявка будет отклонена, если все операторы заняты обслуживанием предыдущих заявок. К тому же в случае с службой поддержки вводится дополнительная характеристика – время между приходами соседних заявок.

Положим, что $x_i(k)$ – момент начала работы оператора W_i с заявкой k, $\tau_i(k)$ – количество тактов времени, которое необходимо потратить оператору W_i на обслуживание заявки k, $\mu(k)$ – время между приходами заявок k и $k-1$, примем $\mu(1)\equiv 0$. В таблице 1 схематически представлена работа такой модели для трёх операторов W_1, W_2 и W_3 при заданных векторах $\mu(k)=(0\ \ 2\ \ 1\ \ 1\ \ 3)$ и $\tau(k)=(5\ \ 4\ \ 3\ \ 7\ \ 4)$, заявки будем обозначать заглавными латинскими буквами A, B, C, D и т. д.

Таблица 1. Схематический вид работы СМО.

Такт времени	Приход заявки	W_1	W_2	W_3
0	*A*	*A*		
1		*A*		
2	*B*	*A*	*B*	
3	*C*	*A*	*B*	*C*
4	*D* $\not\rightarrow$	*A*	*B*	*C*
5			*B*	*C*
6				
7	*E*	*E*		

Обозначение $\not\rightarrow$ означает, что пришедшая новая заявка не может быть обслужена, т. к. все операторы заняты обслуживанием предыдущих заявок.

В терминах идемпотентного анализа работу сервисной службы можно описать следующим уравнением:

$$x_{k+1}(W_i)=\bigotimes_{j=1}^{k+1}\mu(j)\,I\,\Theta\begin{pmatrix}\tau_1(k) & -\infty & -\infty\\ -\infty & \tau_2(k) & -\infty\\ -\infty & -\infty & \tau_2(k)\end{pmatrix}x_k(W_i). \quad (4)$$

Здесь Θ соответствует операции обратной к $\oplus$ и обладает следующими свойствами свойствами (при $a>b$) $a\Theta b=a$, $b\Theta a=\Theta a$, $a\Theta a=a^{\diamond}$. В векторе $x_k(W_i)$ представлено время начала обслуживания текущих заявок всех операторов и если в момент времени прихода заявки $k+1$ какой-либо из операторов стал свободен, то время начала обслуживания заменяется на момент прихода новой заявки, в противном случае $x_{k+1}(W_i)$ даёт "отрицательное" значение в терминах идемпотентного анализа, что соответствует отклонению заявки.

7. Основные результаты

Практическим результатом проведённой работы стало построение математической модели над полуполем с идемпотентным сложением и её программная реализация, моделирующая работу данной системы, которая в зависимости от входных параметров, рассчитывает основные характеристики системы. Далее по характеристикам СМО определяется минимальное число сотрудников (учитывая сезонность и время суток) необходимое для поддержания качественной (критерием качества выступает процент обслуженных заявок) работы службы, и составляется оптимальное расписание.

8. Список литературы

[1] Г Л. Литвинов, *Универсальные алгоритмы и идемпотентная математика*, Компьютерные инструменты в образовании. - СПб.: Изд-во ЦПО "Информатизация образования", 2000, №6, С. 12-18.

[2] S. Gaubert S., R.D. Katz, *Rational semimodules over the max-plus semiring and geometric approach to discrete event systems*. Kybernetika, 40(2):153-180, 2004

[3] F. Baccelli, G. Cohen, G. J. Olsder, J.-P. Quadrat, *Synchronization and Linearity: an algebra for discrete event systems.* – New-York: John Wiley & Sons, 1992. P. 20 – 25.

[4] С.А. Жбанов *Полуполе R_{max} – основа новых подходов к оптимизации*. Сборник тезисов докладов II Международного студенческого форума “Образование, Наука, Производство”. Белгород, 2004. С. 23 – 25.

[5] С.А. Жбанов, *Применение методов идемпотентной математики для задачи оптимизации работы сервисной службы поддержки*. Труды III Всероссийской молодёжной конференции по проблемам управления (ВМКПУ’2008) – М.: ИПУ РАН – 2008. – С. 31 – 32

Байесовские и марковские сети: логико-вероятностный вывод в базах фрагментов знаний с неопределенностью

Александр Львович Тулупьев, Александр Владимирович Сироткин
СПИИРАН
alt@iias.spb.su, avs@iias.spb.su

Аннотация

Основные определения и положения теории алгебраических байесовских сетей, байесовских сетей доверия и марковских сетей излагаются с точки зрения «объединяющего» указанные формализмы логико-вероятностного подхода (в версии Н. Нильссона и его последователей). Указанные виды сетей могут быть использованы как математические модели баз фрагментов знаний с неопределенностью для проектирования и реализации интеллектуальных систем, вывод в которых опирается на стохастические закономерности, выявленные в некоторой предметной области. Источниками таких закономерностей могут служить как результаты «извлечения» знаний из специалистов-экспертов, так и анализ накопленных массивов данных. Считается, что для закономерностей, полученных любым из указанных путей, характерны, с одной стороны, особая структура, допускающая декомпозицию, а с другой, различные виды неопределенности.

1. Введение

Энтузиазм 70–90-х годов, сопровождавший попытки создать экспертные системы (ЭС), на какое-то время поутих; однако исследователи в области искусственного интеллекта (ИИ) не прекратили предпринимать усилия с целью, с одной стороны, осмыслить итоги и причины неудач в попытках перехода к массовой промышленной эксплуатации ЭС, а с другой стороны, развить соответствующий теоретический и алгоритмический аппарат до такой степени, чтобы его практические применения стали более реалистичными. Указанные усилия привели как к некоторым поверхностным изменениям (в частности, появились новые термины «интеллектуальные системы», «интеллектуальные агенты», «многоагентные системы»), к дальнейшей диверсификации направлений исследований в области ИИ, так и к формированию нового понимания возникших проблем и к постановке новых исследовательских задач.

Среди указанных проблем можно выделить два так называемых «узких места»: (1) представление данных и знаний с неопределенностью (uncertain knowledge representation bottleneck), (2) дефицит данных и знаний (knowledge bottleneck). Суть первой проблемы достаточно очевидна: как в формальных системах представить и систематически обрабатывать неопределенность (недоопределенность) наших знаний, с которой зачастую приходится сталкиваться на практике. Тривиальные решения, которые сводятся просто к тому, чтобы отбросить знания (или данные) с неполнотой, непригодны по двум причинам: (1) отнюдь не во всех предметных областях за разумное время можно получить «полные знания», (2) человек способен действовать и в условиях неполноты знаний о складывающейся ситуации (не без того, конечно, чтобы перед принятием решений не попытаться уменьшить степень своей неосведомленности). Вторая проблема заметно менее очевидна, но еще более болезненна. Есть предметные области, где просто не существует человека-эксперта (можно быть уверенным, что еще ни один живой специалист не смог поработать в жерле действующего вулкана или на поверхности Венеры, но роботизированные комплексы в такие места посылать требуется). Кроме того, не все эксперты готовы делиться своим знанием. Наконец, при попытках организовать машинное обучение (machine learning, на самом деле — автоматический подбор некоторых моделей связей между переменными, характеризующими предметную область, а также оценку параметров этих моделей) по данным из баз данных вдруг обнаруживается, что эти данные неполны (то есть имеют пропуски) и частью или зачастую недостоверны (например, неисправное средство измерения (СИ) выдает неправильный результат, перепутаны измерительные каналы, СИ установлено на позиции, которая не отвечает ряду требований, и прочее).

Объектом одного из направлений исследований в области ИИ, которые стремятся к преодо-

лению обозначенных выше «узких мест», являются байесовские и марковские сети.

Основной принцип, лежащий в основе этих формализмов, хорошо известен — это принцип декомпозиции. Он применяется к совокупности знаний о предметной области. Считается, что эксперт может достаточно детально охарактеризовать связи между двумя–тремя–четырьмя утверждениями о предметной области — в каком-то смысле получается «фрагмент знаний» (ФЗ). Таких фрагментов знаний много, они образуют базу фрагментов знаний (БФЗ). К сожалению, напрямую такие фрагменты знаний и их базы нельзя заложить в программный комплекс. Сначала требуется рассмотреть математические модели ФЗ и БФЗ, разработать соответствующие структуры данных, снабдить их алгоритмами обработки и лишь затем пытаться пользоваться полученными результатами при написании программного кода.

С математической точки зрения возникающие объекты могут быть рассмотрены как система случайных элементов, которая, как правило, организована в виде графа со специфическим свойствами или решетки. Случайные элементы в ней могут быть связаны друг с другом: оказывают влияние на означивания других случайных элементов (или эти означивания взаимно зависимы); однако такие связи достаточно редки. Относительно случайных элементов системы (или их совокупностей) требуется ответить на ряд вопросов: непротиворечиво ли их задание, какова вероятностная семантика всей совокупности случайных элементов, какова вероятность того или иного означивания априорно, как она изменится при поступлении свидетельства (то есть при поступлении сведений об означивании части случайных элементов).

Логико-вероятностными моделями баз фрагментов знаний с неопределенностью являются алгебраические байесовские сети (АБС), байесовские сети доверия (БСД), марковские сети (МС)[1]. *Цель настоящей работы* — дать определение АБС и сжато описать алгоритмы логико-вероятностного вывода в них, а также в сформированном контексте охарактеризовать БСД и МС.

2. Базовые объекты и перенумерация

В первую очередь зафиксируем конечное множество атомарных пропозициональных формул (атомов) — алфавит $A=\{x_i\}_{i=0}^{n-1}$. Отметим, что для удобства мы будем вести нумерацию переменных с нуля. Определим над указанными атомами два набора «базовых» пропозициональных формул.

Первый набор формул — идеал цепочек конъюнкций (идеал конъюнктов) — $\left\{x_{i_1}x_{i_2}\ldots x_{i_k} \mid 0\le i_1<\ldots<i_k\le n-1, k\le n\right\}$, где $x_{i_1}x_{i_2}\ldots x_{i_k}$ означает конъюнкцию соответствующих переменных; сам знак конъюнкции мы для удобства будем опускать. Каждому из конъюнктов вида можно сопоставить число $2^{i_1}+2^{i_2}+\ldots+2^{i_k}$ — номер конъюнкта. Более того, если представить полученное число в двоичной записи, то переменная x_i будет входить в конъюнкт тогда и только тогда, когда i-тый бит номера будет равен единице (нумеровать биты предлагается начиная с младшего разряда, считая его нулевым битом).

Для определения второго набора формул — множества квантов — будет полезным следующее обозначение. Литерал (аргументное место) $\tilde{x}_i$ обозначает, что на его месте в формуле может стоять либо x_i, либо его отрицание $\overline{x_i}$. Тогда множество квантов над алфавитом $A=\{x_i\}_{i=0}^{n-1}$ — $Q=\{\tilde{x}_0\tilde{x}_1\ldots\tilde{x}_{n-1}\}$. Иными словами, квант — это конъюнкция, которая для любой атомарной переменной содержит либо ее саму, либо ее отрицание. Для нумерации квантов мы воспользуемся способом, аналогичным нумерации конъюнктов. Выделим «положительную» часть кванта (множество положительно означенных переменных) и рассмотрим ее как конъюнкт. Номер этого конъюнкта и будет номером рассматриваемого кванта. Таким образом, единице в двоичной записи номера соответствует положительное вхождение переменной, а нулю — отрицательное, при этом рассматриваются все n бит (то есть учитываются лидирующие нули).

После введения нумерации квантов и конъюнктов можно определить векторы вероятностей квантов и конъюнктов:

$$\mathbf{P_c}=\begin{pmatrix}1\\ p(c_1)\\ \vdots\\ p(c_{2^n-1})\end{pmatrix} \quad \text{и} \quad \mathbf{P_q}=\begin{pmatrix}p(q_0)\\ p(q_1)\\ \vdots\\ p(q_{2^n-1})\end{pmatrix},$$

где c_i — конъюнкт номер i, а q_i — i-тый квант. Появление единицы в первом случае вполн-

[1] Марковские сети, конечно, известны уже несколько десятилетий и под разными именами (например, марковские случайные поля). Они нашли свое применение (в частности, модель Изинга в изучении магнетизма) еще задолго до начала целенаправленных исследований в области ИИ.

не оправдано, так как согласно определению c_0 — пустой конъюнкт, соответствующий тождественной истине. После введения перенумерации и определения базовых объектов (конъюнктов и квантов) перейдем к определению вероятности над пропозициональными формулами.

3. Оценки вероятностей над пропозициональными формулами

Пусть $F_0(A)$ — множество всех пропозициональных формул над заданным алфавитом $A = \{x_i\}_{i=0}^{n-1}$. Определим множество отличающихся формул как фактор-множество всех формул по отношению эквивалентности: $F(A) = F_0(A)/\equiv$. Над n атомарными переменными таких классов эквивалентности существует ровно 2^{2^n}.

По теореме о совершенной нормальной дизъюнктивной форме любая пропозициональная формула может быть представлена в виде дизъюнкции конечного числа квантов. Так как при любом зафиксированном означивании всех переменных x_0, x_1, …, x_{n-1} никакие два разных кванта не могут быть одновременно истинны, а, с другой стороны, один из них заведомо истинен, можно рассмотреть множество Q как множество элементарных событий. Задав вероятность на квантах, можно, в свою очередь, построить вероятностное пространство, на котором будет определена вероятность любой пропозициональной формулы. За более подробным описанием аксиоматики вероятностной логики можно обратиться, например, к [9, 15].

Определение вероятности на элементах множества Q потребует следующих ограничений:

$$\forall q \in Q \quad p(q) \geq 0; \tag{1а}$$

$$\sum_{q \in Q} p(q) = 1. \tag{1б}$$

Пользуясь введенными выше векторами, можно переписать эти условия следующим образом:

$$\mathbf{P_q} \geq \mathbf{0}; \tag{2а}$$

$$(\mathbf{1}, \mathbf{P_q}) = 1. \tag{2б}$$

В алгебраических байесовских сетях мы, в основном, работаем не с квантами, а с конъюнктами. В монографии [9] показано, что через вероятности конъюнктов, как и через вероятности квантов можно выразить вероятность любой пропозициональной формулы. Переход между этими двумя «базисами» можно выразить следующей формулой:

$$\mathbf{P_q} = \mathbf{I}_n \times \mathbf{P_c}. \tag{3}$$

При этом ограничения (1) принимают вид:

$$\mathbf{P_q} = \mathbf{I}_n \times \mathbf{P_c}. \tag{4}$$

Здесь n — число атомов, а матрица $\mathbf{I}_n$ — матрица перехода от вероятностей квантов к вероятностям конъюнктов.

Матрица $\mathbf{I}_n$ имеет очень четкую структуру, которую удобнее всего описать рекуррентно:

$$\mathbf{I}_1 = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix},$$

$$\mathbf{I}_n = \mathbf{I}_1 \otimes \mathbf{I}_{n-1} = \mathbf{I}_1 \otimes \mathbf{I}_1^{[n-1]} = \mathbf{I}_1^{[n]}$$ например,

$$\mathbf{I}_2 = \mathbf{I}_1 \otimes \mathbf{I}_1 = \begin{pmatrix} 1 & -1 & -1 & 1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Здесь $\otimes$ означает кронекерово (тензорное) произведение матриц. $\mathbf{I}_n$ — это кронекерова степень матрицы $\mathbf{I}_1$, то есть $\mathbf{I}_n = \mathbf{I}_1^{[n]}$ [1]. Кроме рассмотренной матрицы $\mathbf{I}_n$, будет использоваться и обратная ей — $\mathbf{J}_n$, удовлетворяющая условию:

$$\mathbf{P_c} = \mathbf{J}_n \times \mathbf{P_q}. \tag{5}$$

$$\mathbf{J}_1 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},$$

$$\mathbf{J}_n = \mathbf{J}_1 \otimes \mathbf{J}_{n-1} = \mathbf{J}_1 \otimes \mathbf{J}_1^{[n-1]} = \mathbf{J}_1^{[n]};$$

$$\mathbf{J}_2 = \mathbf{J}_1 \otimes \mathbf{J}_1 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

4. Непротиворечивость фрагмента знаний

Основным элементом АБС является фрагмент знаний (ФЗ); исследуем вопросы его непротиворечивости. ФЗ представляет с собой идеал конъюнктов с оценками истинности (рис. 1). При этом можно выделить два основных случая:

1. Оценка каждого конъюнкта задана скалярно (точечно); совокупность всех оценок представлена в виде вектора оценок $\mathbf{P_c}$; тогда для проверки

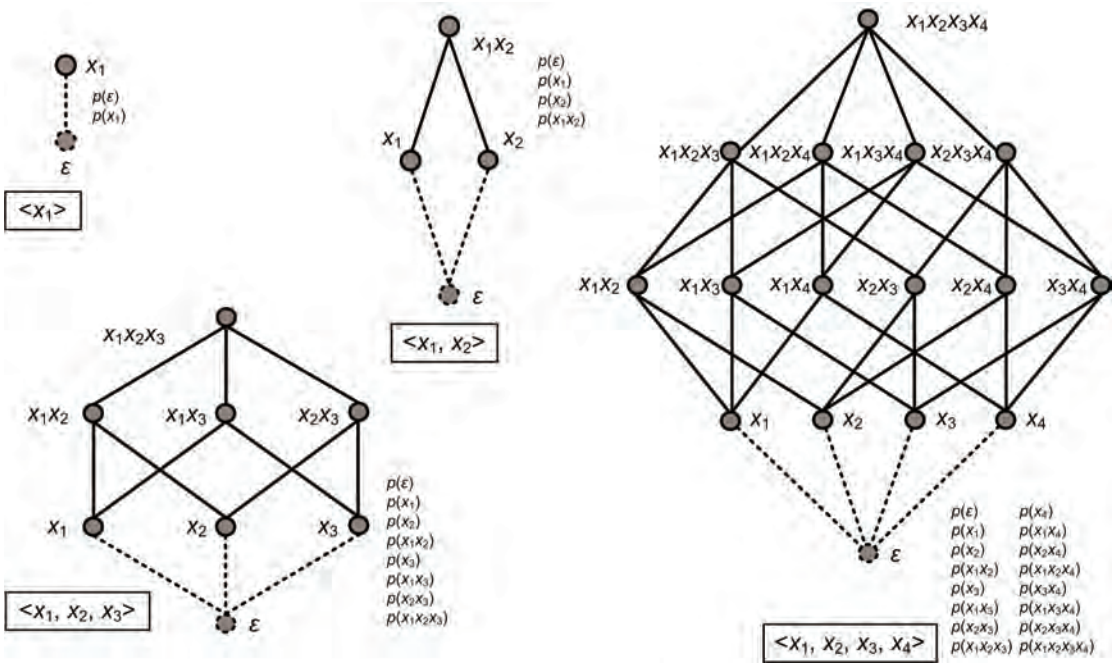


Рис. 1. Примеры фрагментов знаний АБС.

непротиворечивости достаточно проверить выполнение условий (4).

2. Заданы два вектора — верхние и нижние оценки $\mathbf{P}^-$ и $\mathbf{P}^+$; мы говорим, что подобные оценки непротиворечивы, если:

$$\forall i: 1 \le i \le 2^n - 1 \; \forall \varepsilon : \mathbf{P}^-[i] \le \varepsilon \le \mathbf{P}^+[i]$$

$$\exists \mathbf{P}_\mathbf{c} : \left(\mathbf{P}^- \le \mathbf{P}_\mathbf{c} \le \mathbf{P}^+\right) \& \quad .$$

$$\& \left(\mathbf{P}_\mathbf{c}[i] = \varepsilon\right) \& \left(\mathbf{I}_n \times \mathbf{P}_\mathbf{c} \ge \mathbf{0}\right)$$

Здесь запись вида $\mathbf{P}_\mathbf{c}[i]$ означает i-тый элемент вектора $\mathbf{P}_\mathbf{c}$. Если выразить словами, то для любой оценки любого конъюнкта (кроме нулевого, оценка для которого всегда единица), лежащей в границах определенных векторами $\mathbf{P}^-$ и $\mathbf{P}^+$, существует совокупность оценок вероятностей всех остальных конъюнктов, лежащих в границах, определенных $\mathbf{P}^-$ и $\mathbf{P}^+$, и задающих непротиворечивый фрагмент знаний со скалярными оценками.

Как уже было сказано, в первом случае нам достаточно проверить, непротиворечива оценка или нет. В случае с интервальными оценками вопрос не только в том, что непротиворечивы они или нет, но и если противоречивы, то можно ли их сузить до непротиворечивых. Решения этих двух вопросов взаимосвязаны. Для проверки непротиворечивости необходимо решить серию задач линейного программирования (ЗЛП) [9]. Переменными данных задач будут точечные вероятности $\mathbf{P}_\mathbf{c}$, а ограничения будут двух типов: $\mathbf{P}^- \le \mathbf{P}_\mathbf{c} \le \mathbf{P}^+$ и $\mathbf{I}_n \times \mathbf{P}_\mathbf{c} \ge \mathbf{0}$. Осталось определить только целевые функции. Целевыми функциями будут максимизация и минимизация $\mathbf{P}_\mathbf{c}[i]$ для каждого $i: 1 \le i \le 2^n - 1$. Решение этой серии ЗЛП позволяет определить, непротиворечив ли ФЗ; в этом случае все ЗЛП будут разрешимы и соответствующие максимумы и минимумы совпадают с заданными границами. Если хотя бы одна из ЗЛП дала результат, отличный от заданных границ, то соответствующие максимумы и минимумы дадут наибольший по включению набор интервальных оценок, задающий непротиворечивый ФЗ и лежащий в указанных границах. А если хоть одна из ЗЛП оказалась неразрешима, то значит такого сужения не существует, а ФЗ — противоречив.

5. Локальный априорный вывод

Сформировав непротиворечивый ФЗ, перейдем к вопросу об оценивании вероятности произвольной пропозициональной формулы, заданной над теми же атомами, что и фрагмент знаний. Суть априорного вывода — построение оценок для некоторой формулы на основе оценок другого набора формул. В контексте АБС наиболее естественно в качестве набора формул, уже имеющих оценки, рассматривать элементы ФЗ. Мы не будем подробно останавливаться на во-

просах обработки формул и переходу к их СДНФ (совершенной дизъюнктивной нормальной форме), за этим можно обратиться к статье [5].

Рассмотрим пропозициональную формулу f, вероятность истинности которой требуется оценить. Обозначим через $\mathbf{L}_f$ вектор, содержащий 2^n элементов, каждый из которых является единицей или нулем в зависимости от того, входит или не входит соответствующий по номеру квант в СДНФ формулы f (нумерация квантов была описана в разделе 2).

Пусть задан ФЗ с оценками истинности, то есть мы имеем ограничения вида: $\mathbf{P}^- \le \mathbf{P_c} \le \mathbf{P}^+$, где $\mathbf{P_c}$ — вектор вероятностей конъюнктов входящих в ФЗ, а $\mathbf{P}^-$ и $\mathbf{P}^+$ — векторы, состоящие из соответствующих нижних и верхних оценок.

Вероятность формулы f можно выразить следующим образом:

$$p(f) = (\mathbf{L}_f, \mathbf{P_q}).$$

Для перехода к вероятностям конъюнктов воспользуемся уже применявшейся выше формулой (3) и получим:

$$p(f) = (\mathbf{L}_f, \mathbf{I_n} \times \mathbf{P_c}) = (\mathbf{I_n}^T \times \mathbf{L}_f, \mathbf{P_c}).$$

Таким образом, мы выразили вероятность произвольной формулы через вероятности конъюнктов. Чтобы оценить $p(f)$, построим задачу линейного программирования.

Переменными нашей ЗЛП, как и в случае поддержания непротиворечивости, будут элементы вектора $\mathbf{P_c}$. Ограничениями будут $\mathbf{P}^- \le \mathbf{P_c} \le \mathbf{P}^+$, заданные в ФЗ, и неравенства $\mathbf{I}_n \times \mathbf{P_c} \ge \mathbf{0}$, заданные аксиоматикой теории вероятностей. Целевая функция описывается $p(f) = (\mathbf{I_n}^T \times \mathbf{L}_f, \mathbf{P_c})$. Найдем максимум и минимум целевой функции — это и будут искомые оценки вероятности формулы f. Если же построенная ЗЛП не будет иметь решения, то это означает, что исходный набор оценок был противоречив.

Кроме такого подхода к априорному выводу, можно рассмотреть еще построение (или уточнение) оценок ФЗ на основе оценок произвольного набора формул, оно происходит аналогично процессу поддержания непротиворечивости, только в ЗЛП добавляется условие $p_f^- \le (\mathbf{I_n}^T \times \mathbf{L}_f, \mathbf{P_c}) \le p_f^+$, где p_f^- и — нижняя и верхняя оценки вероятности формулы f. Более подробное описание ЗЛП в данном случае можно найти в [6, 9].

6. Перестановки, переозначивания и свидетельства

Для описания последнего вида вывода нам потребуется описать логико-вероятностную модель свидетельств в теории АБС, а так же ввести ряд вспомогательных обозначений, которые позволят нам изложить оставшийся материал тоже на матрично-векторном языке.

Под свидетельством мы понимаем «новые» данные, которые поступили во фрагмент знаний, и с учетом которых нам требуется пересмотреть все (или некоторые) оценки. Сама суть апостериорного вывода заключается в оценивании условной вероятности элементов ФЗ относительно поступившего свидетельства.

Свидетельства, применяемые в АБС, разделяются на несколько видов:

- детерминированные свидетельства;
- стохастические свидетельства;
- неточные свидетельства.

В данной статье мы рассмотрим только детерминированное свидетельство, оставшиеся два случая могут быть сведены к пропагации серии детерминированных свидетельств [9]. Мы говорим, что на вход системы поступило детерминированное свидетельство, если новые сведения представимы в виде конъюнкции атомарных переменных и их отрицаний. Примерами таких свидетельств могут быть $\langle x_1 \rangle$, $\langle x_2 \overline{x}_3 \rangle$, $\langle x_1 \overline{x}_3 x_4 \rangle$; угловые скобки используются для обозначения того, что соответствующая конъюнкция рассматривается нами как свидетельство. Заметим, что такое свидетельство можно разбить на «положительный» и «отрицательный» конъюнкты. В первый входят все положительно означенные атомарные переменные свидетельства, а во второй — отрицательно. При этом и положительной и отрицательной части можно сопоставить индекс в соответствии с перенумерацией, приведенной во втором разделе настоящей работы, и наши свидетельства можно будет записать следующими эквивалентными обозначениями:

$$\langle x_1 \rangle = \langle x_1 \rangle = \langle 2;0 \rangle = \langle 00010_2; 00000_2 \rangle,$$

$$\langle x_2 \overline{x}_3 \rangle = \langle x_2, \overline{x}_3 \rangle = \langle 4;8 \rangle = \langle 00100_2; 01000_2 \rangle.$$

Для начала рассмотрим ситуацию, когда детерминированное свидетельство имеет вид $\langle m;0 \rangle$, то есть состоит из конъюнкта. Рассмотрим матрицу $\mathbf{H}^{\langle m;0 \rangle}$, которая действует на вектор $\mathbf{P_q}$ следующим образом: вероятности квантов, которые согласованы со свидетельством $\langle m;0 \rangle$ (содержат его в себе как подконъюнкт) остаются

неизменными, а вероятности всех остальных квантов приравниваются к нулю. Такая матрица состоит из нулей и некоторого количества единиц на главной диагонали:

$$\mathbf{H}^{\langle m;0\rangle}[i,j]=\begin{cases}\delta(i,j), & i\wedge m=m,\\ 0, & i\wedge m\neq m.\end{cases}$$

Здесь δ — дельта Кронекера, а конъюнкцию между индексами следует рассматривать как побитовую. Заданная подобным образом матрица $\mathbf{H}^{\langle m;0\rangle}$ может быть удобно выражена с использованием тензорного произведения:

$$\mathbf{H}^{\langle m;0\rangle}=\tilde{\mathbf{H}}_{n-1}\otimes\tilde{\mathbf{H}}_{n-2}\otimes\ldots\otimes\tilde{\mathbf{H}}_0.$$

В указанном произведении матрица $\tilde{\mathbf{H}}_i$ принимает значение $\mathbf{H}^{+}=\begin{pmatrix}0&0\\0&1\end{pmatrix}$, если i-тый бит числа m равен единице (то есть переменная x_i получила положительное означивание) и значение $\mathbf{H}^{\circ}=\begin{pmatrix}1&0\\0&1\end{pmatrix}$ если не входит (то есть переменная x_i не получила ни какого означивания). В дальнейшем, когда мы будем рассматривать более сложные свидетельства, то появится еще один вариант, матрица $\tilde{\mathbf{H}}_i$ будет принимать значение $\mathbf{H}^{-}=\begin{pmatrix}1&0\\0&0\end{pmatrix}$, если переменная x_i получит отрицательное означивание.

Однако после такого преобразования вектор $\mathbf{H}^{\langle m;0\rangle}\times\mathbf{P_q}$ не будет удовлетворять требованию (1б), поэтому следует рассмотреть нормированный вектор

$$\mathbf{P_q}^{\langle m;0\rangle}=\frac{1}{(\mathbf{1},\mathbf{H}^{\langle m;0\rangle}\times\mathbf{P_q})}\cdot\mathbf{H}^{\langle m;0\rangle}\times\mathbf{P_q}.$$

Для удобства обозначим нормировочный коэффициент через α. В приведенной формуле это будет: $\alpha=\frac{1}{(\mathbf{1},\mathbf{H}^{\langle m;0\rangle}\times\mathbf{P_q})}$, в дальнейшем, где это потребуется, мы будем использовать в формулах α, а после давать расшифровку для конкретного случая.

Возможно, что $(\mathbf{1},\mathbf{H}^{\langle m;0\rangle}\times\mathbf{P_q})=0$; это значит, что вероятность поступившего свидетельства была равна нулю, что в свою очередь значит, что данное свидетельство поступить не могло. В подобной ситуации [5, 7] всем конъюнктам в качестве оценки приписывается интервал $[0;1]$. Отметим, что вектор $\mathbf{P_q}^{\langle m;0\rangle}$ содержит в точности вероятности вида $p\left(\tilde{x}_0\tilde{x}_1\ldots\tilde{x}_{n-1}\,|\,\langle m;0\rangle\right)$. Теперь можно построить вектор

$$\mathbf{P_c}^{\langle m;0\rangle}=\mathbf{J}_n\times\mathbf{P_q}^{\langle m;0\rangle}=\mathbf{J}_n\times\alpha\cdot\mathbf{H}^{\langle m;0\rangle}\times\mathbf{I}_n\times\mathbf{P_c},$$

что также даст нам условные вероятности, но уже для конъюнктов, то есть для элементов исходного ФЗ. Вспомним, что все преведенные в вышестоящей формуле матрицы могут быть представлены в виде произведений Кронекера, которое в свою очередь обладает следующим свойством:

$$(\mathbf{A}_1\otimes\mathbf{B}_1)\times(\mathbf{A}_2\otimes\mathbf{B}_2)=(\mathbf{A}_1\times\mathbf{A}_2)\otimes(\mathbf{B}_1\times\mathbf{B}_2).$$

Тогда $\mathbf{P_c}^{\langle m;0\rangle}=\mathbf{J}_n\times\alpha\cdot\mathbf{H}^{\langle m;0\rangle}\times\mathbf{I}_n\times\mathbf{P_c}=$

$$=\alpha\cdot(\mathbf{J}_1\otimes\ldots\otimes\mathbf{J}_1)\times(\tilde{\mathbf{H}}_{n-1}\otimes\ldots\otimes\tilde{\mathbf{H}}_0)\times$$

$$\times(\mathbf{I}_1\otimes\ldots\otimes\mathbf{I}_1)\times\mathbf{P_c}=\alpha\cdot(\mathbf{J}_1\times\tilde{\mathbf{H}}_{n-1}\times\mathbf{I}_1)\otimes$$

$$\otimes\ldots\otimes(\mathbf{J}_1\times\tilde{\mathbf{H}}_0\times\mathbf{I}_1)\times\mathbf{P_c}=\alpha\cdot\tilde{\mathbf{T}}_{n-1}\otimes\ldots\otimes\tilde{\mathbf{T}}_0.$$

Для удобства записи мы ввели обозначение $\tilde{\mathbf{T}}_i=\mathbf{J}_1\times\tilde{\mathbf{H}}_i\times\mathbf{I}_1$. Как и для $\tilde{\mathbf{H}}_i$ для $\tilde{\mathbf{T}}_i$ у нас будет три возможных «означивания»:

$$\mathbf{T}^{-}=\mathbf{J}_1\times\mathbf{H}^{-}\times\mathbf{I}_1=\begin{pmatrix}1&-1\\0&0\end{pmatrix};$$

$$\mathbf{T}^{\circ}=\mathbf{J}_1\times\mathbf{H}^{\circ}\times\mathbf{I}_1=\begin{pmatrix}1&0\\0&1\end{pmatrix};$$

$$\mathbf{T}^{+}=\mathbf{J}_1\times\mathbf{H}^{+}\times\mathbf{I}_1=\begin{pmatrix}0&1\\0&1\end{pmatrix}.$$

То, что мы проделали, является частным случаем локального апостериорного вывода, лежащим в основе реализации других его случаев [5, 7].

Если детерминированное свидетельство содержит «отрицательную» часть, используется особая замена переменных — переозначивание.

Пусть задан непротиворечивый вектор вероятностей $\mathbf{P_c}$ над множеством атомарных пропозиций $\{x_0,x_1,\ldots,x_{n-1}\}$. Как, имея эти данные, выразить через него вектор $\mathbf{P_c}^{\langle -1\rangle}$, который представляет собой вектор вероятностей, аналогичный $\mathbf{P_c}$, но построенный над множеством атомов $\{\overline{x}_0,x_1,\ldots,x_{n-1}\}$? Верхний индекс вида $\langle -k\rangle$ мы будем использовать для обозначения атомарных пропозиций, получающих отрицательное означивание, при этом само k — максимальный индекс конъюнкта, состоящий только из отрицательно означенных атомов, входящих в свидетельство. Рассмотрим $\mathbf{P_q}=\mathbf{I}_n\times\mathbf{P_c}$ и $\mathbf{P_q}^{\langle -1\rangle}=\mathbf{I}_n\times\mathbf{P_c}^{\langle -1\rangle}$. Заметим, что векторы $\mathbf{P_q}$ и

$\mathbf{P}_{\mathbf{q}}^{\langle -1\rangle}$ отличаются только порядком элементов. Построим соответствующую матрицу переходов

$$\mathbf{A}^{\langle -1\rangle}=\begin{pmatrix}\begin{pmatrix}0&1\\1&0\end{pmatrix}&\mathbf{0}&\mathbf{0}&\mathbf{0}\\ \mathbf{0}&\begin{pmatrix}0&1\\1&0\end{pmatrix}&\mathbf{0}&\mathbf{0}\\ \mathbf{0}&\mathbf{0}&\ddots&\mathbf{0}\\ \mathbf{0}&\mathbf{0}&\mathbf{0}&\begin{pmatrix}0&1\\1&0\end{pmatrix}\end{pmatrix}.$$

В общем случае матрица $\mathbf{A}^{\langle -k\rangle}=\tilde{\mathbf{A}}_{n-1}\otimes\ldots\otimes\tilde{\mathbf{A}}_0$, Каждая из матриц $\tilde{\mathbf{A}}_i$ может принимать одно из двух означиваний:

$\tilde{\mathbf{A}}_i=\mathbf{A}^{\circ}=\begin{pmatrix}1&0\\0&1\end{pmatrix}$, если x_i не меняется и

$\tilde{\mathbf{A}}_i=\mathbf{A}^{-}=\begin{pmatrix}0&1\\1&0\end{pmatrix}$, если x_i меняется на $\overline{x}_i$.

$$\mathbf{P}_{\mathbf{c}}^{\langle -1\rangle}=\mathbf{J}_n\times\mathbf{P}_{\mathbf{q}}^{\langle -1\rangle}=\mathbf{J}_n\times(\mathbf{A}^{\langle -1\rangle}\times\mathbf{P}_{\mathbf{q}})=$$
$$=\mathbf{J}_n\times(\mathbf{A}^{\langle -1\rangle}\times(\mathbf{I}_n\times\mathbf{P}_{\mathbf{c}}))=$$
$$=(\mathbf{J}_n\times\mathbf{A}^{\langle -1\rangle}\times\mathbf{I}_n)\times\mathbf{P}_{\mathbf{c}},$$

а в общем случае:

$$\mathbf{P}_{\mathbf{c}}^{\langle -k\rangle}=\mathbf{J}_n\times\mathbf{P}_{\mathbf{q}}^{\langle -k\rangle}=\mathbf{J}_n\times(\mathbf{A}^{\langle -k\rangle}\times\mathbf{P}_{\mathbf{q}})=$$
$$=\mathbf{J}_n\times(\mathbf{A}^{\langle -k\rangle}\times(\mathbf{I}_n\times\mathbf{P}_{\mathbf{c}}))=$$
$$=(\mathbf{J}_n\times\mathbf{A}^{\langle -k\rangle}\times\mathbf{I}_n)\times\mathbf{P}_{\mathbf{c}}.$$

Вновь воспользовавшись свойством тензорного произведения получим

$$\mathbf{P}_{\mathbf{c}}^{\langle -k\rangle}=(\mathbf{J}_1\times\tilde{\mathbf{A}}_{n-1}\times\mathbf{I}_1)\otimes\ldots\otimes$$
$$\otimes(\mathbf{J}_1\times\tilde{\mathbf{A}}_{n-1}\times\mathbf{I}_1)\times\mathbf{P}_{\mathbf{c}}=$$
$$=(\tilde{\mathbf{V}}_{n-1}\otimes\ldots\otimes\tilde{\mathbf{V}}_0)\times\mathbf{P}_{\mathbf{c}}.$$

Матрицы $\tilde{\mathbf{V}}_i$ могут принимать два значения:

$$\mathbf{V}_i^{\circ}=\mathbf{J}_1\times\mathbf{A}^{\circ}\times\mathbf{I}_1=\begin{pmatrix}1&0\\0&1\end{pmatrix}\text{ и}$$

$$\mathbf{V}_i^{-}=\mathbf{J}_1\times\mathbf{A}^{-}\times\mathbf{I}_1=\begin{pmatrix}1&0\\1&-1\end{pmatrix}.$$

Заметим, что подобная замена обратима, более того $\left(\mathbf{P}_{\mathbf{c}}^{\langle -k\rangle}\right)^{\langle -k\rangle}=\mathbf{P}_{\mathbf{c}}$.

Теперь соединим два полученные выше результата. Пусть поступило свидетельство $\langle m,k\rangle$. Мы считаем что $m\wedge k=0$, то есть в положительную и отрицательную часть свидетельства входят разные атомы. Проведем замену переменных $\langle -k\rangle$, тогда в новых переменных нам надо пропагировать свидетельство $\langle m\vee k\rangle$. Но это мы уже умеем делать. В полученном результате пропагации останется только снова провести замену переменных $\langle -k\rangle$.

Итого получаем:

$$\mathbf{P}_{\mathbf{c}}^{\langle m,k\rangle}=\left(\left(\mathbf{P}_{\mathbf{c}}^{\langle m,k\rangle}\right)^{\langle -k\rangle}\right)^{\langle -k\rangle}=$$
$$=\left(\left(\mathbf{P}_{\mathbf{c}}^{\langle -k\rangle}\right)^{\langle m\vee k,0\rangle}\right)^{\langle -k\rangle}=\mathbf{V}^{\langle -k\rangle}\times\left(\mathbf{P}_{\mathbf{c}}^{\langle -k\rangle}\right)^{\langle m\vee k,0\rangle}=$$
$$=\mathbf{V}^{\langle -k\rangle}\times\alpha\cdot\mathbf{T}^{\langle m\vee k,0\rangle}\times\mathbf{P}_{\mathbf{c}}^{\langle -k\rangle}=$$
$$=\alpha\cdot\mathbf{V}^{\langle -k\rangle}\times\mathbf{T}^{\langle m\vee k,0\rangle}\times\mathbf{V}^{\langle -k\rangle}\times\mathbf{P}_{\mathbf{c}}.$$

В данном случае

$$\alpha=\frac{1}{(\mathbf{1},H^{\langle m\vee k,0\rangle}\times\mathbf{P}_{\mathbf{q}}^{\langle -k\rangle})}=\frac{1}{(\mathbf{1},H^{\langle m\vee k,0\rangle}\times\mathbf{A}^{\langle -k\rangle}\times\mathbf{I}_n\times\mathbf{P}_{\mathbf{c}})}.$$

Как и прежде воспользуемся свойствами тензорного произведения

$$\mathbf{P}_{\mathbf{c}}^{\langle m,k\rangle}=\alpha\cdot(\tilde{\mathbf{V}}_{n-1}\times\tilde{\mathbf{T}}_{n-1}\times\tilde{\mathbf{V}}_{n-1})\otimes\ldots\otimes$$
$$\otimes(\tilde{\mathbf{V}}_0\times\tilde{\mathbf{T}}_0\times\tilde{\mathbf{V}}_0)\times\mathbf{P}_{\mathbf{c}}.$$

Заметим, что $\tilde{\mathbf{V}}_i\times\tilde{\mathbf{T}}_i\times\tilde{\mathbf{V}}_i$ может принимать одно из трех значений $\mathbf{T}^{-}$, $\mathbf{T}^{\circ}$, $\mathbf{T}^{+}$, в зависимости от означивания x_i, а следовательно $\mathbf{P}_{\mathbf{c}}^{\langle m,k\rangle}=\alpha\cdot\mathbf{T}^{\langle m,k\rangle}\times\mathbf{P}_{\mathbf{c}}$. Заметим, что α можно вычислить на самом последнем этапе, воспользовавшись тем, что $\mathbf{P}_{\mathbf{c}}^{\langle m,k\rangle}[0]=1$, а следовательно, вычислив произведение $\mathbf{T}^{\langle m,k\rangle}\times\mathbf{P}_{\mathbf{c}}$, нам достаточно посмотреть на его нулевой элемент. Нормировочный коэффициент α будет равен обратной величине к этому элементу.

Полученная формула дает нам точное выражение для апостериорной вероятности элементов ФЗ при поступившем свидетельстве, если исходные оценки были скалярные (то есть оценки вероятности конъюнктов имели вид $\mathbf{P}_{\mathbf{c}}[i]=p_i$, где p_i — числа из интервала $[0;1]$). Если оценки были не скалярные, а интервальные, то нам потребуется решать серию ЗЛП [9].

Пусть задан ФЗ с оценками $\mathbf{P}^{-}$ и $\mathbf{P}^{+}$, тогда опишем ЗЛП для получения апостериорной вероятности на элементах ФЗ.

Полученная формула однородна по $\mathbf{P}_{\mathbf{c}}$, то есть если подставить вместо него $\lambda\cdot\mathbf{P}_{\mathbf{c}}$, где

$\lambda > 0$, то результат не изменится. Рассмотрим в качестве переменных нашей ЗЛП $\lambda \cdot \mathbf{P_c}[i]$ и обозначим их через $\mathbf{d}[i]$. Очевидно, что $\mathbf{d}[0] = \lambda$. Над новыми переменными можно выписать ограничения вида $\mathbf{d}[i] \le \lambda \mathbf{P}^{+}[i]$ и $\lambda \mathbf{P}^{-}[i] \le \mathbf{d}[i]$, кроме этого мы должны включить ограничения $\mathbf{I}_n \times \mathbf{d} \ge \mathbf{0}$. Кроме указанных ограничений, добавим ограничения $(\mathbf{1}, \mathbf{H}^{\langle m \vee k, 0 \rangle} \times \mathbf{A}^{\langle -k \rangle} \times \mathbf{I}_n \times \mathbf{d}) = 1$ и $\lambda \ge 0$. Внеся это ограничение, мы ничего не потеряем, так как всегда можно подобрать такое $\lambda > 0$, что это будет верно, а в силу однородности мы не исключим не одного возможного значения для элементов $\mathbf{P_c}^{\langle m,k \rangle}$ (подробнее применительно к АБС можно посмотреть в [9], а применительно к общему случаю сведения задачи гиперболического программирования к ЗЛП в [2]). Последнее ограничение позволяет переписать выражение для $\mathbf{P_c}^{\langle m,k \rangle}$, а именно

$$\mathbf{P_c}^{\langle m,k \rangle} = \mathbf{T}^{\langle m,k \rangle} \times \mathbf{d}.$$

Теперь нам осталось только решить полученные ЗЛП и найти при заданных условиях максимумы и минимумы для элементов $\mathbf{P_c}^{\langle m,k \rangle}[i]$, они и будут оценками апостериорной вероятности.

7. Фрагменты знаний в других сетях

В теории байесовских сетей доверия рассматриваются только скалярные оценки вероятностей, причем преимущественно — условных вероятностей.

Математической моделью фрагмента знаний в теории БСД можно считать узел некоторого направленного графа, которому приписан тензор условных вероятностей (рис. 2). В рамках логико-вероятностного подхода будем считать, что в таких узлах стоят случайные бинарные элементы, а тензоры вероятностей характеризуют вероятность приобретения таким СБЭ того или иного означивания, зависящую только от означивания СБЭ в родителях рассматриваемого узла.

Пусть в интересующем нас узле y байесовской сети доверия стоит СБЭ $\hat{y}$, а узлы x_0, …, x_{m-1} являются его родителями.

Тогда тензор условных вероятностей вида $p(\tilde{y} \mid \tilde{x}_0 \dots \tilde{x}_{m-1})$ характеризует зависимость означиваний $\hat{y}$ от означиваний его родителей. Нам достаточно знать скалярные величины $p(y \mid \tilde{x}_0 \dots \tilde{x}_{m-1})$, поскольку

$$p(\overline{y} \mid \tilde{x}_0 \dots \tilde{x}_{m-1}) = 1 - p(y \mid \tilde{x}_0 \dots \tilde{x}_{m-1}).$$

Множество родителей некоторого узла y удобно обозначать как $\mathrm{pa}(y)$; а кванты, образованные над этим множеством, как $\tilde{\wedge}\,\mathrm{pa}(y)$. В этих обозначениях соотвествующий тензор условных вероятностей будет записан как $p(y \mid \tilde{\wedge}\,\mathrm{pa}(y))$. Если $\tilde{\wedge}\,\mathrm{pa}(y) = \varnothing$, то вместо условной вероятности формально рассматривается маргинальная вероятность: $p(y \mid \varnothing) = p(y)$.

Непротиворечивость фрагмента знаний в теории БСД определяется весьма просто:

$$(\forall \tilde{y})(\forall\, \tilde{\wedge}\,\mathrm{pa}(y))\ \ 0 \le p(\tilde{y} \mid \tilde{\wedge}\,\mathrm{pa}(y)) \le 1$$

$$(\forall\, \tilde{\wedge}\,\mathrm{pa}(y))\ \ p(y \mid \tilde{\wedge}\,\mathrm{pa}(y)) + p(\overline{y} \mid \tilde{\wedge}\,\mathrm{pa}(y)) = 1$$

и вопросы, с ней связанные, зачастую не рассматриваются в силу своей тривиальности.

Наконец, оставаясь рамках логико-вероятностного подхода, упомянем фрагмент знаний из теории марковских сетей (МС) (марковских случайных полей) [11].

В самом общем случае марковская сеть состоит из узлов, причем в узле y размещен СБЭ $\hat{y}$.

Узлы организованы в некоторую структуру, в которой для каждого узла y определено множество его соседей $\mathrm{ne}(y)$. Примером такой структуры может служить (рис. 3) ненаправленный граф или решетка (lattice).

То или иное означивание СБЭ $\hat{y}$ характеризуется уровнем «энергии» $u(\tilde{y})$, причем в марковских сетях выполняется условие марковости: уровень энергии означивания узла зависит только от означиваний его соседей. Пользуясь тем же соглашением для $\mathrm{ne}(y)$, что ранее было введено для $\mathrm{pa}(y)$, условие марковости можно записать как $u(\tilde{y}) = u(\tilde{y}, \tilde{\wedge}\,\mathrm{ne}(y))$.

Важно то, что энергия означивания конкретного узла зависит не от всех узлов, которые присутствуют в структуре, а только от соседствующих с этим узлом.

В зависимости от распределения энергии $u(\tilde{y})$ распределяются вероятности означиваний СБЭ $\hat{y}$.

Следовательно, фрагментом знаний в случае марковских цепей выступают функции или таблицы значений: $u(\tilde{y}, \tilde{\wedge}\,\mathrm{ne}(y))$, либо $p(\tilde{y}, \tilde{\wedge}\,\mathrm{ne}(y))$, либо $p(\tilde{y} \mid \tilde{\wedge}\,\mathrm{ne}(y))$.

Все они с формальной точки зрения называются потенциалами.

Можно сказать, что фрагментом знаний в случае марковских сетей называется потенциал, значения которого определены для означиваний некоторого узла и всех его соседей.

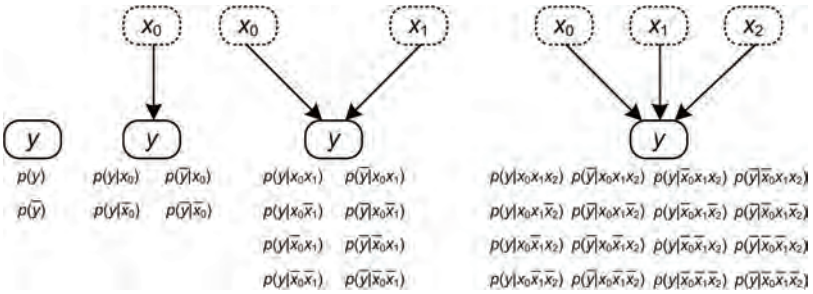


Рис. 2. Примеры ФЗ байесовской сети доверия.

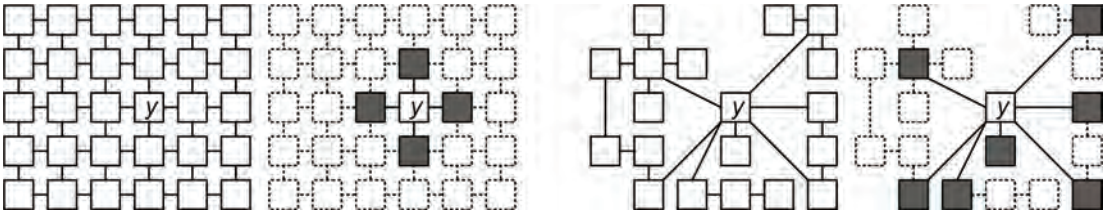


Рис. 3. Примеры Марковских сетей и ФЗ в них.

Алгоритмы логико-вероятностного вывода в марковских сетях [11] опираются преимущественно на метод сэмплирования (порождения означиваний узлов) по Гиббсу. (Это — один из методов Монте-Карло.) Сначала все узлы получают произвольное означивание. После этого определяется распределение энергии среди означиваний каждого узла, а затем — распределение вероятностей над этими означиваниями. Согласно полученным распределениям снова моделируется означивание каждого узла, и начинается новая итерация.

Поступившее свидетельство позволяет фиксировать означивание некоторого небольшого набора узлов. В них уже означивание не моделируется; состояние узлов, входящих в свидетельство, лишь влияет на вероятности означиваний других узлов.

В процессе сэмплирования по Гиббсу подсчитываются частоты означиваний отдельных узлов или их совокупностей, которые интересуют исследователя.

Марковские сети рассматривались [11] в теории магнетизма (модель Изинга), их использовали для моделирования распространения слухов в социальных сетях, изучают возможность их применения для разработки интеллектуальных систем, подсказывающих по набору совершенных покупок, какие покупки еще клиенту может захотеться совершить. Эти подсказки основаны на результатах обработки статистики продаж — а в Интернет-магазинах наборы соответствующих «сырых» данных для анализа формируются автоматически.

8. Непротиворечивость сети

Рассмотрев отдельные фрагменты знаний всех предложенных сетей, перейдем к совокупности фрагментов знаний и распределению вероятностей, которое определяется всеми элементами сети. Такое единое распределение мы будем называть вероятностной семантикой сети. Следует заметить, что вероятностная семантика может быть описана как набор вероятностей всех конъюнктов.

Мы рассмотрим случаи точечных и интервальных оценок, четыре степени непротиворечивости АБС, представленной в виде такого дерева смежности, а также вопросы, связанные с гипотезой условной независимости и апостериорными выводом, учитывающем эту гипотезу. Примеры АБС приведены на рис. 4.

Рассмотрим два идеала конъюнктов, поостренных над атомарными пропозициями из цепочек конъюнкций U и V соответственно. Эти цепочки имеют структуру

$$U = XY, \quad V = YZ .$$

Цепочка Y может быть пустой, а цепочки X, Y и Z не имеют общих элементов. Для удобства введем обозначение $W = XYZ$.

Пусть над квантами вида $\tilde{U}$ и $\tilde{V}$ введены согласованные распределения вероятностей $p_U\left(\tilde{U}\right)$ и $p_V\left(\tilde{V}\right)$; они согласованы на общей подцепочке:

$$\forall \tilde{Y} \; p_U\left(\tilde{Y}\right) = p_V\left(\tilde{Y}\right).$$

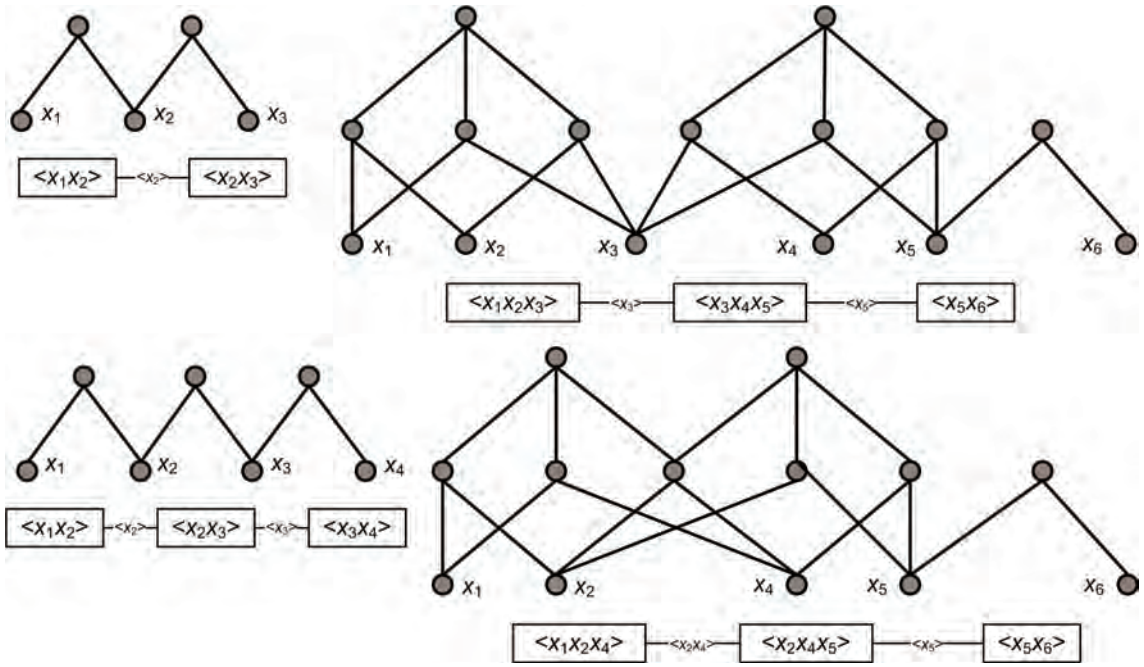


Рис. 4. Примеры алгебраических байесовских сетей.

Тогда над квантами вида $\tilde{W}$ можно построить распределение вероятностей $p_W\left(\tilde{W}\right)$, по формуле:

$$p_W\left(\tilde{W}\right)=p_W\left(\tilde{X}\tilde{Y}\tilde{Z}\right)=$$

$$=\begin{cases}0, & p_V(\tilde{Y})=0,\\ \dfrac{p_V(\tilde{X}\tilde{Y})p_U(\tilde{Y}\tilde{Z})}{p_V(\tilde{Y})} & p_V(\tilde{Y})\neq 0,\end{cases} \quad (6)$$

маргинализующееся над квантами вида $\tilde{U}$ и $\tilde{V}$ к соответствующим исходным:

$$\forall\tilde{U}\ p_U\left(\tilde{U}\right)=p_W\left(\tilde{U}\right),\quad \forall\tilde{V}\ p_V\left(\tilde{V}\right)=p_W\left(\tilde{V}\right).$$

Такое построение композиции распределений случайных бинарных последовательностей (СБП) равносильно утверждению об условной независимости случайных бинарных последовательностей $\hat{X}$ и $\hat{Z}$ при известном означивании $\hat{Y}$. Сама же условная независимость определяется в общем случае как:

$$\forall\tilde{X}\tilde{Y}\tilde{Z}\ \ p\left(\tilde{X}\tilde{Y}\tilde{Z}\right)p\left(\tilde{Y}\right)=p\left(\tilde{X}\tilde{Y}\right)p\left(\tilde{Y}\tilde{Z}\right).$$

Употребление термина условная независимость становиться интуитивно ясным при $p(\tilde{Y})\neq 0$:

$$\frac{p\left(\tilde{X}\tilde{Y}\tilde{Z}\right)p\left(\tilde{Y}\right)}{p\left(\tilde{Y}\right)p\left(\tilde{Y}\right)}=\frac{p\left(\tilde{X}\tilde{Y}\right)p\left(\tilde{Y}\tilde{Z}\right)}{p\left(\tilde{Y}\right)p\left(\tilde{Y}\right)},$$

$$\frac{p\left(\tilde{X}\tilde{Y}\tilde{Z}\right)}{p\left(\tilde{Y}\right)}=\frac{p\left(\tilde{X}\tilde{Y}\right)}{p\left(\tilde{Y}\right)}\cdot\frac{p\left(\tilde{Y}\tilde{Z}\right)}{p\left(\tilde{Y}\right)},$$

$$p\left(\tilde{X}\tilde{Z}\,|\,\tilde{Y}\right)=p\left(\tilde{X}\,|\,\tilde{Y}\right)p\left(\tilde{Z}\,|\,\tilde{Y}\right).$$

Пусть каждая из цепочек конъюнкций X и Z разбита на две произвольные непересекающиеся подцепочки: $X=X_1X_2$ и $Z=Z_1Z_2$. Тогда справедливы равенства

$$p_W\left(\tilde{X}_s\tilde{Y}\tilde{Z}\right)=\begin{cases}0, & p_V(\tilde{Y})=0,\\ \dfrac{p_V(\tilde{X}_s\tilde{Y})p_U(\tilde{Y}\tilde{Z})}{p_V(\tilde{Y})}, & p_V(\tilde{Y})\neq 0,\end{cases}$$

$$p_W\left(\tilde{X}_s\tilde{Y}\tilde{Z}\right)=\begin{cases}0, & p_V(\tilde{Y})=0,\\ \dfrac{p_V(\tilde{X}_s\tilde{Y})p_U(\tilde{Y}\tilde{Z})}{p_V(\tilde{Y})}, & p_V(\tilde{Y})\neq 0,\end{cases}$$

$$p_W\left(\tilde{X}_s\tilde{Y}\tilde{Z}_r\right)=\begin{cases}0, & p_V(\tilde{Y})=0,\\ \dfrac{p_V(\tilde{X}_s\tilde{Y})p_U(\tilde{Y}\tilde{Z}_r)}{p_V(\tilde{Y})}, & p_V(\tilde{Y})\neq 0,\end{cases}$$

$r\in\{1,2\}$, $s\in\{1,2\}$, которые непосредственно вытекают из суммирования левой и правой часть равенства (6) по означиваниям цепочек $\tilde{X}_{3-s}$ и $\tilde{Z}_{3-r}$ соответственно. Кроме того, СБП над любой подцепочкой X и СБП над любой

подцепочкой Z условно независимы при известном означивании $\hat{Y}$.

Отметим, что согласованность распределений на квантах $\tilde{U}$ и $\tilde{V}$ равносильна согласованности распределений на соответствующих идеалах конъюнктов над цепочками U и V. Поэтому если распределения над идеалами согласованы в общем подыдеале над цепочкой Y, то указанным выше способом эти два распределения можно распространить до распределения в объемлющем идеале, построенном над атомарными пропозициями из цепочки W. Если фрагменты знаний (т.е. идеалы с оценками вероятностей — точечными в данном случае) над U и V непротиворечивы, то будет непротиворечив и построенный объемлющий их фрагмент знаний над W.

АБС называется *глобально непротиворечивой*, если ее можно погрузить в объемлющий фрагмент знаний, который непротиворечив. «Погрузить» означает то, что в получившемся объемлющем фрагменте знаний сохраняются неизменными оценки истинности тех конъюнктов, которые входили в исходную АБС.

Определения непротиворечивости фрагмента знаний с точечными оценками (по набору оценок на конъюнктах должно восстанавливаться непротиворечивое распределение на квантах) и с интервальными оценками (если для произвольного элемента ФЗ мы выбираем произвольную точку из интервала его оценки, то тогда в интервалах оценок всех других элементов можно выбрать такие точечные значения, что в совокупности с первым выбранным эти значения задают непротиворечивое точечное распределение на элементах ФЗ) были рассмотрены раньше.

АБС считается *локально непротиворечивой*, если каждый фрагмент знаний в сети непротиворечив. Этот уровень непротиворечивости не предъявляет требований к согласованности оценок на формулах, общих для двух или нескольких фрагментов знаний.

АБС считается *экстернально непротиворечивой*, если каждый фрагмент знаний в сети непротиворечив, а также оценки истинности каждой формулы, входящей одновременно в два фрагмента знаний или более, совпадают.

АБС считается *интернально непротиворечивой*, если каждый фрагмент знаний в сети непротиворечив, а также для каждой формулы из АБС для любого точечного значения из интервала оценки ее истинности можно выбрать согласованные (т.е. совпадающие на одинаковых формулах) точечные значения во всех фрагментах знаний, так, что все получившиеся ФЗ с точечными оценками будут непротиворечивы.

АБС считается *глобально непротиворечивой*, если ее с имеющимися оценками можно погрузить в непротиворечивый объемлющий фрагмент знаний и при этом оценки на формулах из АБС не изменятся.

Отметим, что в общем случае из глобальной непротиворечивости следует интернальная, из интернальной — экстернальная, а из экстернальной — локальная. Проверка и поддержание степеней непротиворечивости имеют различную вычислительную сложность: самая вычислительно сложная — глобальная непротиворечивость, затем идет интернальная, экстернальная и, наконец, локальная.

«Идеальным» случаем непротиворечивости является глобальная непротиворечивость, поскольку только она гарантирует существование хотя бы одного «всеобщего» распределения вероятностей над квантами, построенными сразу над всеми атомарными пропозициями из АБС. Но проверка глобальной непротиворечивости по ее определению экспоненциально сложна, поэтому важным является рассмотрение отношений степеней непротиворечивостей. В этой связи установлено несколько нетривиальных фактов.

Несовпадение локальной и экстернальной непротиворечивости очевидно: достаточно рассмотреть два пересекающихся ФЗ, у которых не будут совпадать оценки на общих элементах.

Существует пример АБС, которая экстернально непротиворечива, но не является интернально непротиворечивой [4].

Существует пример алгебраической байесовской сети, построенной на цикле трех фрагментов знаний второго порядка, которая является интернально непротиворечивой, но противоречивой глобально [9]. Однако это связано с наличием в ней цикла. В случае ациклических АБС из их интернальной непротиворечивости следует их глобальная непротиворечивость [9].

9. Байесовские сети доверия

Согласно классическому определению [10, 16], байесовская сеть доверия (рис. 5) — это, в первую очередь, ациклический направленный граф G с тензорами условных вероятностей в узлах. Исследуем, как это определение, пока еще неполное и неокончательное, разворачивается в рамках логико-вероятностного подхода.

В этом разделе мы будем опираться на ряд монографий [10, 12, 13, 16], посвященных теории байесовских сетей доверия.

Перечень узлов БСД можно задать ранее использовавшимся алфавитом A. Узлы соединены

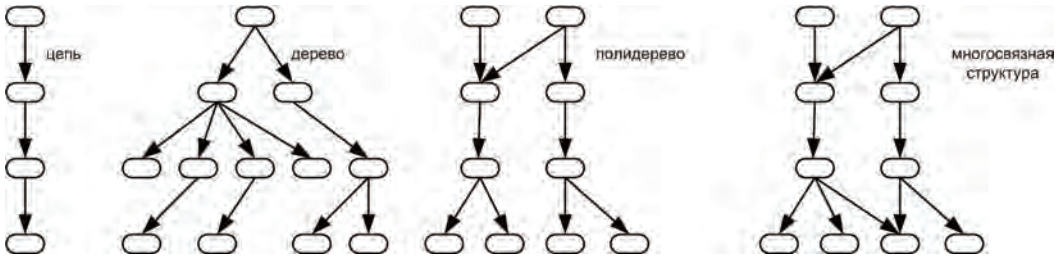


Рис. 4. Примеры структур байесовских сетей доверия.

направленными ребрами вида (x_i, x_j), где $x_i, x_j \in A$.

Ребро направлено от x_i (первого узла пары) к x_j (второму ее узлу). Мы исключаем из рассмотрения графы с петлями, то есть не допускаем ребра вида (x_i, x_i) — они не могут входить во множество ребер графа.

Родителями $\mathrm{pa}(x_i)$ узла x_i считаются те узлы, из которых исходит ребро, направленное в x_i. Предполагается, что число родителей у каждого узла весьма ограничено ($0 \ldots 3$).

Оставаясь в рамках логико-вероятностного подхода, мы считаем, что в байесовской сети доверия каждому узлу взаимно однозначно сопоставлен случайный бинарный элемент $\hat{x}_i$, где $x_i \in A$. Для каждого узла также задан тензор условных вероятностей $p(\tilde{x}_i \mid \tilde{\wedge}\,\mathrm{pa}(x_i))$.

Этот тензор характеризует «причинно-следственную» зависимость означивания узла от означиваний его родителей. Такая зависимость не является жестко детерминированной; ее стохастический, в определенной степени случайный характер как раз и выражается указанным тензором условных вероятностей.

Однако, чтобы дать окончательное определение байесовским сетям доверия, нам необходимо ввести еще ряд понятий и обозначений.

Множество $\mathrm{ds}(x_i)$ состоит из *потомков* узла x_i, то есть всех тех узлов графа G, в которые из x_i ведет направленный путь, содержащий хотя бы одно ребро. Множество $\mathrm{an}(x_i)$ состоит из *предков* узла x_i, то есть всех тех узлов графа G, из которых в x_i ведет направленный путь, содержащий хотя бы одно ребро. Родители являются подмножеством множества предков:

$$\mathrm{pa}(x_i) \subset \mathrm{an}(x_i).$$

Множество $\mathrm{nd}(x_i)$ состоит из *непотомков* узла x_i, то есть всех тех узлов графа G, которые не попали во множество $\mathrm{ds}(x_i)$.

Кроме того, сам узел x_i также не включается во множество $\mathrm{nd}(x_i)$ своих непотомков.

Множество *непотомков* узла x_i, которые также не являются его родителями, обозначим $\mathrm{nn}(x_i)$ и заметим, что $\mathrm{nn}(x_i) = \mathrm{nd}(x_i) \setminus \mathrm{pa}(x_i)$.

Пусть определено совместное распределение вероятностей над всеми случайными элементами, которые можно построить над изначально заданным алфавитом (или, что то же, над всеми случайными элементами, стоящимо в узлах графа G): $p\left(\bigwedge\limits_{x_i \in A} \tilde{x}\right)$. Будем говорить, что пара граф-распределение (G, p) удовлетворяют *условию марковости*, если для любого узла x_i соответствующий случайных элемент $\hat{x}_i$ условно независим от всех своих непотомков, которые также не являются его родителями, $\mathrm{nn}(x_i)$ при заданном означивании своих родителей $\mathrm{pa}(x_i)$.

О том же самом говорят иначе: граф G и распределение p совместно удовлетворяют условию марковости. Следуя ряду источников, в частности [13, 16], рассмотренное отношение условной независимости для отдельного узла x_i будем обозначать как $I_p\left(\{x_i\}, \mathrm{nn}(x_i) \mid \mathrm{pa}(x_i)\right)$.

Если пара (G, p) удовлетворяет условию марковости, то можно доказать [13], что справедливо *цепное правило*, увязывающее совместное распределение $p\left(\bigwedge\limits_{x_i \in A} \tilde{x}\right)$ и индуцированные им условные вероятности $p(x_i \mid \tilde{\wedge}\,\mathrm{pa}(x_i))$ в узлах носителя БСД — графа G:

$$p\left(\bigwedge_{x_i \in A} \tilde{x}_i\right) = \prod_{x_i \in A} p(\tilde{x}_i \mid \tilde{\wedge}\,\mathrm{pa}(x_i)) \qquad (7)$$

следует помнить, что означивания соотвествующих литералов в левой и правой частях формулы связаны, то есть должны совпадать.

Кроме того, известно [13], что если по заданным условным вероятностям строится совместное распределение по формуле (7), то тогда такое совместное распределение и граф G удовлетворяют условию марковости.

После проведенной подготовки, можно окончательно сформулировать определение БСД.

Байесовская сеть доверия — это пара (G, p), где G — ациклический направленный граф, p — совместное распределение вероятностей всех СБЭ, приписанных узлам графа, причем граф G и распределение p совместно удовлетворяют условию марковости.

В теории БСД при таком подходе к определению, как правило, совместное распределение p задается как произведение соотвествующих тензоров условных вероятностей, приписанных узлам графа G; связь устанавливается по формуле (7), то есть согласно цепному правилу.

Отношения независимости и условной независимости между случайными бинарными элементами и их совокупностями — один из самых интенсивно изучаемых аспектов выразительной силы и вероятностной семянтики байесовских сетей доверия. В выявлении системы указанных отношений особую роль играет d*-разделимость* [14, 16] (direction dependent separation) — понятие, на основе которого некоторые авторы выстраивают альтернативное (но эквивалентное предложенному выше) определение БСД.

В графе G случайные бинарные элементы $\hat{x}_i$ и $\hat{x}_j$ условно независимы относительно совокупности E случайных бинарных элементов, получивших означивание:

$$p\left(\tilde{x}_i \tilde{x}_j \mid \bigwedge_{x \in A} \tilde{x}\right) = p\left(\tilde{x}_i \mid \bigwedge_{x \in A} \tilde{x}\right) p\left(\tilde{x}_j \mid \bigwedge_{x \in A} \tilde{x}\right)$$

если в любом *ненаправленном* пути между узлами x_i и x_j существует узел x', не совпадающий ни с x_i, ни с x_j, такой, что выполнено одно из трех условий:

- $x' \in E$, и оба ребра пути *выходят* из x'; такое сочетание ребер в узле x' называется *расходящейся связью*;
- $x' \in E$, и одно ребро пути *выходит* из x', а другое — *выходит*; такое сочетание ребер в узле x' называется *последовательной связью*;
- ни сам x', ни один из его потомков (то есть никакой элемент из $\mathrm{ds}(x')$ не входят в E, а оба ребра пути *входят* в x'; такое сочетание ребер в узле x' называется *сходящейся связью*.

Для облегчения понимания заметим, что получившие означивание СБЭ из E играют роль *свидетельства*.

Когда в отношении какого-либо пути между узлами x_i и x_j удовлетворено одно из перечисленных выше условий, говорят, что x' блокирует этот путь при заданном E. Если все пути между узлами x_i и x_j блокированы, тогда говорят, что E d*-разделяет* узлы x_i и x_j; более того: тогда в байесовской сети доверия случайные бинарные элементы $\hat{x}_i$ и $\hat{x}_j$ условно независимы от E.

Учитывая понятия, введенные выше, можно дать альтернативное определение *байесовским сетям доверия* — это ациклические направленные графы с тензорами условных вероятностей в узлах, причем случайные бинарные элементы, соответствующие d-разделенным узлам, являются условно независимыми при заданных означиваниях случайных бинарных элементов, соответствующим узлам из d-разделяющего множества. Соотношения между d-разделимостью и цепным правилом описываются в специализированных монографиях по байесовским сетям доверия и даже в некоторых книгах более широкой тематики [9, 10, 12, 13, 14, 16].

Вероятностной семантикой БСД при любом определении последней является то единственное распределение вероятностей, которое сеть задает; это распределение подчиняется цепному правилу.

Как указано в обоих вариантах определения, в направленном графе G циклы недопустимы; и тут есть некоторая тонкость в понимании формулировки. Недопустимыми являются *направленные циклы*; в ненаправленном графе-носителе направленного графа циклы допустимы.

Не предложено еще ни одного БСД-исчисления, в котором направленные циклы могли бы быть обработаны [10, 16]. Однако направленный БСД-цикл можно преобразовать во фрагмент знаний (или цепь смежности фрагментов знаний) АБС; такой фрагмент, полученный в результате преобразования, может обрабатываться с помощью алгоритмов логико-вероятностного вывода, известных в теории алгебраических байесовский сетей. Вопросы вероятностной семанти-

ки и обработки циклических паттернов подробно рассматриваются в [8].

Допустимые ненаправленные циклы в байесовских сетях доверия обрабатываются с помощью операции триангуляции: сначала из исходной БСД получается ненаправленное дерево сочленений (оно уже не содержит циклов, а его вероятностная семантика описана в [3]), а потом в этом дереве сочленений выполняются операции логико-вероятностного вывода (первичная пропагация и пропагация свидетельств) [10, 12, 13].

10. Задачи машинного обучения

Machine learning — машинное обучение, автоматическое извлечение «знаний» из совокупностей «данных» — достаточно долго, но все еще формирующийся в искусственном интеллекте ответ на упоминавшуюся проблему дефицита знаний (knowledge bottleneck). Действительно, если мы хотим достичь приемлемой степени автономности интеллектуальной системы, то надо сделать так, чтобы при изменении внешней среды система могла бы дообучиться или переобучиться сама, а не были бы вынуждена всякий раз обращаться к эксперту. Более того, поскольку процесс извлечения знаний из экспертов достаточно дорогостоящий и далеко не всегда конструктивно воспринимаемый последними, хотелось бы, чтобы интеллектуальная система могла бы обучаться по тем наборам сведений, которые накопились и продолжают копиться в различных базах, банках и хранилищах данных.

В теории байесовских сетей доверия задачи машинного обучения весьма конкретны:

- *локальное обучение*: в предположении, что структура сети известна, по имеющимся совместным реализациям всех случайных элементов сопоставить конкретные численные оценки тензорам условных вероятностей вида $p(\tilde{y} \mid \tilde{x}_0 \dots \tilde{x}_{m-1})$ в каждом узле сети;
- *обучение локальной структуры*: отнюдь не всегда требуется задание величин вида $p(y \mid \tilde{x}_0 \dots \tilde{x}_{m-1})$ для всех возможных означиваний $\tilde{x}_0 \dots \tilde{x}_{m-1}$, такие величины могут быть связаны и, таким образом, может быть определен набор параметров, число которых меньше, чем 2^m, но на их основе можно восстановить значения всех $p(y \mid \tilde{x}_0 \dots \tilde{x}_{m-1})$; а чем меньше число параметров, тем быстрее и лучше (в определенном смысле) удается получить оценку их величины;
- *обучение глобальной структуры*: выявление того ациклического направленного графа, который лежит в основе байесовской сети доверия, которая лучше всего (опять в некотором определенном смысле) соответствует выборке (набору известных совместных означиваний всех случайных элементов сети).

В рамках логико-вероятностного подхода, которого мы придерживаемся в настоящей работе, задача локального обучения специфицируется следующим образом: для каждого означивания родителей $\tilde{x}_0 \dots \tilde{x}_{m-1}$ узла y определить оценку вероятности $p(y \mid \tilde{x}_0 \dots \tilde{x}_{m-1}) = \theta$. (Заметим, что $p(\bar{y} \mid \tilde{x}_0 \dots \tilde{x}_{m-1}) = 1 - \theta$ согласно аксиоматике вероятностей).

Для удобства считается, что бета-распределение — априорное распределение вероятностей параметра $\theta \sim B(d, q)$, где гиперпараметры бета-распределения — натуральные числа. Бета-распределение выбирается потому, что оно относится к классу сопряженных распределений: при поступлении новых данных, а значит, при переходе от априорного распределения к апостериорному распределению мы остаемся в том же классе распределений.

Когда происходит локальное обучение, перебираются все известные реализации совместных означивания случайных переменных. Если для заданного означивания родителей $\tilde{x}_0 \dots \tilde{x}_{m-1}$ литерал $\tilde{y}$ приобрел положительное означивание, то от априорного $\theta \sim B(d, q)$ осуществляется переход к апостериорному $\theta \sim B(d+1, q)$, если же отрицательное, то — к $\theta \sim B(d, q+1)$. Таким образом перебирается вся доступная выборка. В конце обучения выбирается наиболее вероятное значение параметра $\theta = \dfrac{d}{d+q}$.

Назначение исходных значений параметров в бета-распределении формально никак не регламентируется; по существу же зависит от знаний о предметной области. Чем больше изначальные значения параметров, тем медленнее будет происходит обучение.

В рамках предложенного подхода делаются попытки обработать ситуацию и с пропущенными данными. Существует несколько вариантов обработки пропусков, но все они в своей основе сводятся к тому, чтобы априорно (но необязательно осознанно или явно) предположить наличие какой-то закономерности и согласно ей выбирать наиболее вероятное значение. В этом отношении алгебраические байесовские сети позволили бы большую гибкость: в случае пропуска значений в реализации от накапливающихся скалярных (точечных) значений можно было бы перейти к интервальным оценкам.

Заметим, что в рамках более широких подходов, чем логико-вероятностный, в байесовской сети доверия вместо случайных бинарных элементов могут быть использованы случайные многозначные (мультиномиальные) переменные. Тогда вместо бета-распределений нам пришлось бы использовать распределения Дирихле, однако принцип локального обучения остался бы тот же самый.

Как уже отмечалось, обучение локальной структуры становится возможным тогда, когда для определения тензора условных вероятностей в узле требуется меньше параметров, чем число элементов самого тензора. Имеется широкий выбор вариантов возможной локальный структуры, мы ограничимся разбором лишь одного частного случая для лог-линейной модели. На самом деле тензор условных вероятностей можно задать в виде

$$\mathrm{logit}[\,p(y \mid \tilde{x}_0\tilde{x}_1)] = a + a_0\tilde{x}_0 + a_1\tilde{x}_1 + a_{01}\tilde{x}_0\tilde{x}_1,$$

где a_* — параметры лог-линейной модели, $\mathrm{logit}[\,p] = \log\left(\frac{p}{1-p}\right)$, а в уравнении литерал $\tilde{x}$ приобретает значение 0, если означен отрицательно, и 1 — если положительно. В полной лог-линейной модели параметров столько же, сколько элементов тензора условных вероятностей. Но если удалось обнаружить, что некоторые коэффициенты нулевые, например $\mathrm{logit}[\,p(y \mid \tilde{x}_0\tilde{x}_1)] = a_0\tilde{x}_0 + a_1\tilde{x}_1$, то тогда считается, что соответствующий тензор условных вероятностей обладает нетривиальной локальной структурой. Она позволяет ограничиться обучением (то есть оцениванием) меньшего числа параметров, что, в конце концов, ведет к более точным результатам или требует менее объемную обучающую выборку.

Обучение глобальной структуры — один из наиболее сложных вопросов теории БСД. Его сложность компенсируется, правда, тем, что в хорошо изученных областях эксперты в заметном числе случаев могут сообщить достаточно сведений либо для построения структуры графаносителя сети, либо для сокращения перебора возможных структур. Однако, как было отмечено, обращаться к экспертам нежелательно, если имеются доступные наборы данных.

На современном этапе исследований цель обучения глобальной структуры БСД примерно такова: нам надо построить как можно меньшую байесовскую сеть доверия, которая как можно более точно соответствует набору реализаций совместных означиваний случайных бинарных элементов в доступной выборке. С одной стороны, утверждение должно быть, конечно, уточнено, с другой стороны, уже на этом этапе понятно, что части утверждения находятся в «конфликте» друг с другом и требуется искать какой-то компромисс.

Такой компромисс достигается заданием одной из возможных метрик (как правило, опирающихся на классические достижения теории информации), которые учитывают как объем информации, требующейся для передачи байесовской сети, так и объем информации, требующийся для передачи «дополнений», которые формально могут позволить восстановить исходный набор данных в точности. Задача становится совсем простой с «идеологической точки зрения»: перебрать все возможные структуры ациклических направленных графов и все возможные параметризации тензоров таких графов с тем, чтобы выбрать тот граф, на котором достигается минимум выбранной метрики. Конечно, такой перебор оказывается настолько вычислительно сложным, что речь о прямом его применении не идет.

Задача разбивается на две части: поиск структуры сети и оценка значений элементов тензоров. Последнюю подзадачу мы разбирали; сосредоточимся на вопросе выявления структуры. Его решение учитывает особенность байесовских сетей доверия, которая позволяет не связывать переменные, которые друг от друга условной не зависят при заданном означивании некоторого набора других переменных.

Самая общая схема алгоритма такова (его название — алгоритм PC [18]). На заданном наборе случайных элементов строится полный граф. Перебираются все пары элементов, соединенных дугой; перебираются все возможные множества (сначала мощности 1, затем 2 и так далее), не содержащие эту пару. Решается вопрос (с помощью критерия хи-квадрат) условной независимости текущей пары элементов относительно текущего множества других элементов. Если такая независимость имеет место, то дуга исключается.

Вопрос о направленности дуг также решается; алгоритм приписывания направления может учитывать какие-то априорные сведения об упорядоченности узлов, а также опираться на принципы выбора причин и следствий, сформированные Гансом Райхенбахом (Hans Reichenbach).

Предложенная схема алгоритма обучения глобальной структуры БСД допускает массу возможных реализаций. Для более подробного ознакомления с вопросами машинного обучения байесовских сетей доверия можно рекомендовать [12, 17]

11. Заключение

Три вида рассмотренных в настоящей работе сетей (алгебраические байесовские сети, байесов-

ские сети доверия и марковские сети) являются составной частью «ответа», формируемого исследователями в области искусственного интеллекта, на проблему представления данных и знаний с неопределенностью и на проблему дефицита таких данных и знаний.

Марковские сети (известные также как марковские случайные поля) обладают наиболее длительной историей исследований. До того, как они нашли свое применение в представлении данных и знаний с вероятностной неопределенностью в приложениях искусственного интеллекта, марковские сети интенсивно применялись в математической физике, математической теории социальных сетей и других традиционных отраслях науки. Однако приемы логико-вероятностного вывода в таких сетях, в основном, являются модификациями сэмплирования по Гиббсу; также в теории марковских сетей возникают нерешенные вопросы, связанные с проверкой непротиворечивости. Наконец, марковские сети не позволяют представлять и обрабатывать интервальные оценки вероятности истинности.

Байесовские сети доверия в настоящее время обладают наиболее развитым теоретическим и алгоритмическим аппаратом; последний реализован в ряде программных комплексов (которые могут в некоторой степени различаться между собой доступной функциональностью). Однако указанный вид сетей не позволяет систематически обрабатывать интервальные оценки вероятностей, а также содержит существенное ограничение на структуру сети — в ней не должно быть направленных циклов.

Теория алгебраических байесовских сетей развита в наименьшей степени; вместе с тем она позволяет (1) обрабатывать как скалярные, так и интервальные оценки истинности, (2) четко различать три основные виды логико-вероятностного вывода: проверка и поддержание непротиворечивости, априорный вывод, апостериорный вывод, (3) моделировать свидетельства с зависимыми компонентами, (4) получить теоретические и эмпирические оценки чувствительности результатов локального априорного и апостериорного выводов. При этом возникающие в процессе различных видов вывода экстремальные задачи сводятся к задачам линейного программирования. Наконец, достижения теории алгебраических байесовских сетей позволяют анализировать вероятностную семантику и осуществлять обработку направленного цикла, запрещенного в теории байесовских сетей доверия. В вероятностной семантике АБС и БСД имеются существенные сходства, что позволяет (в части случаев) разрабатывать и исследовать алгоритмы преобразования сети одного вида в сеть другого вида.

12. Литература

[1] Беллман Р. *Введение в теорию матриц*. М.: Наука, Гл. редакция физико-математической литературы, 1969. 368 с.

[2] Гавурин М. К., Малоземов В. Н. *Экстремальные задачи с линейными ограничениями: учебное пособие.* Л.: Изд-во Ленингр. ун-та, 1984. 175 с.

[3] Сироткин А.В. «Байесовские сети доверия: дерево сочленений и его вероятностная семантика». // *Труды СПИИРАН*. 2006. Вып. 3., т. 1. СПб.:«Наука», 2006. С. 228–239.

[4] Сироткин А.В. «Интернальная и экстернальная степени непротиворечивости ациклических алгебраических байесовских сетей» // X Санкт-Петербургская международная конференция «Региональная информатика - 2006 (РИ-2006)», Санкт-Петербург, 24–26 октября 2006 г.: Материалы конференции. СПб.: СПОИСУ, 2006. С. 52.

[5] Сироткин А. В., Тулупьев А. Л. «Локальный априорный вывод в алгебраических байесовских сетях: комплекс основных алгоритмов» // *Труды СПИИРАН*. Вып. 5. 2007. СПб.: Наука, 2007. С. 100–111.

[6] Тулупьев А. Л. *Алгебраические байесовские сети: локальный логико-вероятностный вывод.* СПб.: СПбГУ; Анатолия. 2007. 80 с. — (Элементы мягких вычислений.)

[7] Тулупьев А. Л. *Алгебраические байесовские сети: глобальный логико-вероятностный вывод.* СПб.: СПбГУ; Анатолия. 2007. 40 с. — (Элементы мягких вычислений.)

[8] Тулупьев А. Л. *Байесовские сети: логико-вероятностный вывод в циклах*. СПб.: Издательство С.-Петербургского университета, 2008. 140 с. — (Элементы мягких вычислений.)

[9] Тулупьев А. Л., Николенко С. И., Сироткин А. В. *Байесовские сети: логико-вероятностный подход*. СПб.: Наука, 2006. 607 с.

[10] Jensen F. V. *Bayesian Networks and Decision Graphs.* NY.: Springer-Verlag, 2001. 268 p.

[11] Kindermann R., Snell J. L. *Markov Random Fields and Their Applications.* Providence, RI: Amer. Math. Soc., 1980. 142 p.

[12] Korb K. B., Nicholson A. E. *Bayesian Artificial Intelligence*. NY.: Chapman and Hall/CRC, 2004. 364 p.

[13] Neapolitan R. E. *Learning Bayesian Networks*. Pearson Prentice Hall, 2003. 674 p.

[14] Nilsson N. J. *Artificial Intelligence: a New Synthesis.* San Francisco, CA: Morgan Kaufmann Publ., 1998. 513 p.

[15] Nilsson N. J. «Probabilistic Logic» // *Artificial Intelligence*. 1986. Vol. 47. Amsterdam: Elsevier Science Publishers B.V., 1986. P. 71–87.

[16] Perl J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* NY etc.: Morgan Kaufmann Publ., 1994. 552 p.

[17] Pelikan M., Sastry K., Goldberg D. E. *iBOA: The Incremental Bayesian Optimization Algorithm.* MEDAL Report No. 2008002. January 2008. [Электронный ресурс] / arXiv:0801.3113v1 [cs.NE] 21 Jan 2008. 18 p.

[18] Verma T. S., Pearl J. «Equivalence and synthesis of causal models» // Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence (San Mateo, CA, 1990), Morgan Kauffmann, 1990. P. 220–227.

Деформация границы простых множеств под действием растягивающей последовательности отображений

Сергей Комеч*
ИППИ РАН
komech@iitp.ru

Аннотация

Изучается соотношение между ростом объема множества и искажением его границы под действием растягивающей последовательности отображений. Полученные результаты находят применение при изучении геометрического смысла энтропии Колмогорова-Синая для диффеоморфизмов.

1. Введение

Предположение о связи между энтропией и скоростью искажения границы простых множеств появилось впервые в физической литературе (см. [1]). Точные результаты были впервые сформулированы в [2] (для символических динамических систем) и в [3] (для автоморфизмов n-мерного тора). При анализе гладких гиперболических систем удобно изучать искажение границы отдельно на локальных неустойчивых слоях. Имея это в виду, мы изучаем в данной работе искажение границы под действием растягивающей последовательности отображений в $\mathbb{R}^n$.

2. Основные результаты

Для произвольной последовательности отображений $f_1, f_2, ...$ обозначим через $f_{[1,k]}$ отображение $f_k \circ ... \circ f_1$.

Определение. Последовательность отображений $f_i : \mathbb{R}^n \to \mathbb{R}^n, i \in \mathbb{N}$, будем называть *растягивающей*, если существуют $\lambda > 1$ и последовательность чисел $r(i) \to \infty$, такие, что

$$d(f_{[1,i]}(x), f_{[1,i]}(y)) \geq \lambda^{r(i)} d(x, y)$$

при всех $i \in \mathbb{N},\ x, y \in \mathbb{R}^n$.

*Эта работа частично поддержана грантом CRDF # RUM1-2693-MO-05.

Фиксируем произвольный репер в $\mathbb{R}^n$. Обозначим через P_δ параллелепипед со стороной δ, построенный на этом репере.

Теорема 1. *Пусть $A_1, A_2, ...$ — растягивающая последовательность линейных отображений $\mathbb{R}^n \to \mathbb{R}^n$. Тогда*

(i) *для всякого $\delta > 0$*

$$\lim_{k\to\infty} \frac{mes(O^\delta(A_{[1,k]}P_\delta))}{mes(A_{[1,k]}P_\delta)} = 1,$$

где $O^\delta(\cdot)$ — δ-окрестность в $\mathbb{R}^n$ и mes — мера Лебега;

(ii) *если функция $k : \mathbb{R}_+ \to \mathbb{Z}_+$ такова, что $\lim_{\varepsilon\to 0} k(\varepsilon) = \infty$, то*

$$\lim_{\varepsilon\to 0} \frac{mes(O^\varepsilon(A_{[1,k(\varepsilon)]}P_\varepsilon))}{mes(A_{[1,k(\varepsilon)]}P_\varepsilon)} = 1.$$

Замечание. Очевидно, что оба утверждения теоремы останутся верными, если в (ii) заменить $A_{[1,k(\varepsilon)]}$ на $A_{k(\varepsilon)} \circ V_{k(\varepsilon)}(\varepsilon) \circ ... \circ A_1 \circ V_1(\varepsilon)$, где $V_i(\varepsilon)$ — сдвиг на произвольный вектор $v_i(\varepsilon) \in \mathbb{R}^n, i \in \mathbb{N}$, и аналогичную замену сделать в (i).

Аналогичное утверждение для нелинейного случая, вообще говоря, неверно. Однако оно будет выполнено при дополнительном ограничении на функцию $k(\cdot)$.

Теорема 2. *Пусть $f_1, f_2, ...$ — растягивающая последовательность отображений $f_i : \mathbb{R}^n \to \mathbb{R}^n$, $f_i \in C^{1+\alpha}$, $\alpha > 0$, $i \in \mathbb{N}$, и $k : \mathbb{R}^+ \to \mathbb{Z}^+$ — функция, удовлетворяющая условиям*

$$\lim_{\varepsilon\to 0} k(\varepsilon) = \infty, \quad \lim_{\varepsilon\to 0} \frac{k(\varepsilon)}{\ln \varepsilon} = 0.$$

Обозначим через $D_x f_i$ дифференциал отображения f_i в точке x. Предположим, что при всех i найдутся такие $\beta_i > 0$ и $\theta_i > 0$, что

$$|f_i(x) - f_i(y)| \leq \beta_i |x - y|,$$

$$||D_y f_i - D_x f_i|| \leq \theta_i |x - y|^\alpha,$$

$$\overline{\lim_{\varepsilon\to 0}}\ \varepsilon^{\alpha}(\sum_{i=1}^{k(\varepsilon)}\theta_i)(\Pi_{i=1}^{k(\varepsilon)}\beta_i)^{1+\alpha} < \infty$$

при всех $x, y \in \mathbb{R}^n, i \in \mathbb{N}$. *Тогда*

$$\lim_{\varepsilon\to 0}\frac{mes(O^{\varepsilon}(f_{[1,k(\varepsilon)]}P_{\varepsilon}))}{mes(f_{[1,k(\varepsilon)]}P_{\varepsilon})} = 1.$$

Замечание. Условие на ограниченность верхнего предела будет выполнено, например, в случае $\theta_i = \theta, \beta_i = \beta$ при всех $i \in \mathbb{N}$. Используя $k(\varepsilon) = o(\ln\varepsilon)$, легко проверить, что

$$\lim_{\varepsilon\to 0}\varepsilon^{\alpha}k(\varepsilon)\theta\beta^{k(\varepsilon)(1+\alpha)} = 0.$$

Список литературы

[1] Г. М. Заславский, *Стохастичность динамических систем.*, Наука, Москва, 1984.

[2] B. M. Gurevich, Geometric Interpretation of Entropy for Random Processes., *AMS Transl.*, Vol 171, 1996, pp. 81-87.

[3] С. А. Комеч, Энтропия и скорость усложнения границ в гиперболических системах, Труды 27 конференции молодых ученых мехмата МГУ им. М.В. Ломоносова, Москва, 2005, стр. 76-80.

Протокол демонстрации K-кратного вхождения строки

Федюкович В.Е. Шарапов В.Г.
GlobalLogic Ukraine НТУУ 'КПИ'
vadym.fedyukovych@globallogic.com vital.evidence@gmail.com

Аннотация

Предложен протокол для демонстрации выполнимости условия о K-кратном вхождении строки-шаблона в строку-текст. Протокол использует обязательства, полученные из обеих строк, и не раскрывает полезной информации о строках, а также о позициях шаблона в тексте. Протокол является аргументом и может иметь свойство нулевого разглашения с честным проверяющим. Протокол создавался как одно из возможных решений задачи обработки персональных данных.

1 Введение

Последовательности символов определенного алфавита (строки) являются основной структурой данных в биоинформатике. Изучение свойств таких строк позволяет делать статистические выводы. Обрабатываемая таким образом информация, а также получаемые выводы в некоторых случаях могут квалифицироваться как персональные данные. Для ряда задач защиты личной информации были предложены специфические протоколы, в том числе электронных монет и демократического электронного голосования.

В этой работе предложен протокол, позволяющий достигнуть противоречивых (на первый взгляд) целей: убедить Проверяющего в выполнимости определенного утверждения о строках, не раскрывая при этом какой-либо полезной информации об этих строках ни на каком этапе такой демонстрации. Предложенный протокол позволяет Проверяющему убедиться в выполнении условия о присутствии K копий шаблона в тексте, так что вероятность обмана Доказывающим является ничтожно малой. Возможности Проверяющего получить какую-либо информацию о строках в процессе такой проверки, а также убедительно демонстрировать выполнение условия любым третьим лицам после завершения проверки могут быть ограничены.

Предложенный протокол является интерактивным аргументом в предположении о вычислительной сложности задачи SRSA, и может иметь специальное свойство нулевого разглашения с честным Проверяющим. Общей информацией являются обязательства из строк и смещений строки-шаблона в строке-тексте. Доказывающий демонстрирует знание значений, из которых ранее были получены обязательства. Доказывающий также демонстрирует выполнение условия о K-кратном вхождении шаблона для строк, из которых были получены обязательства. Свойство корректности аргумента получено из оценки вероятности случайно выбрать корень полинома (лемма Швартца-Зиппеля). В этой работе также предложено полиномиальное представление строк, которое может представлять независимый интерес.

2 Протоколы и обязательства

Пусть имеется два простых числа p,q такие что $p = 2p'+1$, $q = 2q'+1$ и p', q' простые. Пусть $N = pq$. Обозначим Z_N^* мультипликативную группу классов вычетов, Q_N - подгруппу элементов Z_N^*, имеющих квадратный корень в Z_N.

Проблемой *SRSA* [1] называют задачу поиска (e,A), таких что $A^e \equiv B \pmod N$ для случайно выбранных RSA модуля N и $B \in Z_N^*$. В дальнейшем мы будем использовать знак равенства в операциях с модульной арифметикой.

Последовательность, убывающую быстрее любого полинома, будем называть *ничтожно малой*:

$$\forall(p(x))\exists(n = n(p()))\forall(n' > n): \quad \varepsilon(n') < 1/p(n')$$

Схемой обязательства (commitment scheme) называют тройку алгоритмов $(GEN(), CM(), OPEN())$, такую что выполняются свойства связывания (binding) и скрытия (hiding). Говорят что схема обязательства имеет свойство *связывания* если имеется только ничтожная вероятность успешно раскрыть обязательство, полученное из некоторого значения x используя любое альтернативное значение x':

$$\forall(x, x' \neq x): \Pr\begin{bmatrix} P \leftarrow GEN(1^l) \\ C \leftarrow CM(P,x) \\ Accept \leftarrow OPEN(P,C,x') \end{bmatrix} = \varepsilon(l)$$

Говорят, что схема обязательства имеет свойство *скрытия* если $C = CM(P,x)$ и $C' = CM(P,x')$ (статистически, вычислительно) неразличимы для любых x,x'. Говорят что схема обязательства имеет свойство гомоморфизма (homomorphic) если для некоторых операций $\oplus$ и $\otimes$ выполняется

$$\forall(x,x'): \quad C(P,x\oplus x') = C(P,x)\otimes C(P,x')$$

Значительный интерес представляют схемы обязательств, допускающие протокол демонстрации знания значений, открывающих экземпляр обязательства. В этой работе используется вариант схемы обязательств для целых чисел [2], который имеет связывание в предположении о сложности задачи SRSA статистическое скрытие, аддитивный гомоморфизм, и протокол демонстрации знания.

Протоколом доказательства (interactive proof system) [3, 4, 5] для задачи распознавания языка (decision problem) назвают интерактивную пару машин Тьюринга с полиномиальным Проверяющим и неограниченным Доказывающим, так что выполняются свойства полноты (completeness) и корректности (soundness). *Аргументом* [6, 7] или вычислительно корректным доказательством (computationally sound proof) называют протокол с полиномиальным Доказывающим, в котором корректность зависит от сложности решения некоторой задачи. Говорят, что участник протокола является *честными* (honest) если машина участника выполняет программу, предписанную протоколом; также говорят, что участник протокола является *произвольным* (any) в противном случае.

В этой работе рассматривается протокол, в котором общей информацией является экземпляр обязательства, так что слово x, из которого было получено это обязательство, принадлежит языку ВЫПОЛНИМОСТЬ:

$$L = \{x \mid f(x) = 0\}$$

где $f(x)$ - некоторый полином (диофантово уравнение).

Стандартное определение протокола для распознавания языка неудобно, так как множество приемлемых экземпляров обязательства совпадает с множеством значений. Обозначим $\langle \mathbf{P}(x), \mathbf{V}\rangle (P,C)$ состояние машины Проверяющего при завершении протокола с Доказывающим, так что параметры P и экземпляр обязательства C являются общим входным словом, а x является дополнительной информацией машины Доказывающего.

Будем называть *протоколом аргумента* программу для интерактивной пары полиномиальных машин Тьюринга, так что выполняются свойства полноты и корректности.

Полнота: машина Проверяющего всегда переходит в состояние Accept при завершении протокола с честным Доказывающим в случае если слово x, из которого было получено обязательство C, принадлежит языку:

$$\forall(x \in L) : \Pr\begin{bmatrix} P\leftarrow GEN(1^l) \\ C\leftarrow CM(P,x) \\ Accept\leftarrow\langle \mathbf{P}(x),\mathbf{V}\rangle(P,C)\end{bmatrix} = 1$$

Корректность: имеется только ничтожно малая вероятность для машины Проверяющего перейти в состояние Accept при завершении протокола с произвольным Доказывающим в случае если слово x, из которого было получено обязательство C, не принадлежит языку:

$$\forall(x \notin L) : \Pr\begin{bmatrix} P\leftarrow GEN(1^l) \\ C\leftarrow CM(P,x) \\ Accept\leftarrow\langle \mathbf{P}(x),\mathbf{V}\rangle(P,C)\end{bmatrix} = \varepsilon(l)$$

Кроме того, выполняется свойство корректности для стандартного протокола демонстрации знания для схемы обязательств: имеется только ничтожно малая вероятность для машины Проверяющего перейти в состояние Accept при завершении протокола с произвольным Доказывающим в случае если слово x, из которого было получено обязательство C, не совпадает со словом x' на дополнительной ленте Доказывающего.

Стенограммой протокола (protocol transcript) называют множество сообщений, которыми обменялись участники протокола. *Моделирующим алгоритмом* (simulator) называют алгоритм Проверяющего, позволяющий для любого слова из языка получить стенограмму, неотличимую от стенограммы протокола с Доказывающим. Говорят, что протокол с моделирующим алгоритмом имеет свойство *нулевого разглашения* (zero knowledge) [3, 4, 8, 5]. Говорят, что протокол имеет свойство *нулевого разглашения с честным Проверяющим* (honest verifier zero knowledge) [9] если стенограмма, полученная с помощью моделирующего алгоритма, неотличима от всех стенограмм с Доказывающим, имеющих такой же запрос Проверяющего. Говорят, что протокол имеет *специальное свойство нулевого разглашения с честным проверяющим* [10] если моделирующий алгоритм позволяет получить неотличимую стенограмму для любого входного слова в модели с честным Проверяющим.

Протоколы с двоичными запросами Проверяющего предусматривают повторения для достижения приемлемо низкой вероятности обмана произвольным Доказывающим. Некоторые протоколы достигают корректности без повторений, используя запросы Проверяющего, выбранные из множества большой мощности, а также гомоморфизм используемой схемы обязательств. Примером является протокол Шнорра [11], в котором

запросом является элемент конечного поля, а ответом - значение линейного полинома.

В этой работе предложен протокол аргумента для строк, скрытых обязательствами, таких что в строке-тексте имеется K копий строки-шаблона. Протокол проверяет выполнимость диофантового уравнения, эквивалентного условию о K копиях шаблона в тексте. Протокол использует 'большие' запросы и свойство гомоморфизма схемы обязательств, и достигает корректности без повторений.

Ответом Доказывающего является значение характерного для протокола Шнорра линейного полинома, так что аргументом полинома является запрос Проверяющего. Рассматривается полином проверки принадлежности слова языку, в котором ответы Доказывающего используются вместо переменных. Проверяется условие равенства нулю старшего коэффициента такого полинома. Корректность протокола следует из ограниченного количества корней такого полинома.

Алгоритм $GEN()$ используемой схемы обязательств выбирает некоторые приемлемые $p,q\ h \in Q_N$ порядка $p'q'$, некоторыое значение z, получает $g = h^z$. Параметрами является тройка (N,g,h). Алгоритм $CM()$ использует накоторое случайное r и имеет вид:

$$CM(x,r) = g^x h^r \tag{1}$$

Алгоритм $OPEN(C,x,r)$ проверяет $C \stackrel{?}{=} g^x h^r$. Алгоритм $GEN()$ выполняет Проверяющий, $CM()$ - Доказывающий, а вместо $OPEN()$ используется протокол демонстрации знания, а также выполнимости утверждения о строках. Проверяющий не раскрывает p,q, а также $log_h(g)$. Такая схема имеет вычислительно стойкое связывание в предположении о сложности задачи SRSA, и статистически стойкое скрытие. Схема также имеет протокол демонстрации знания значений (x,r), а именно вариант протокола Шнорра с двумя ответами Доказывающего, полученными как значения линейных полиномов над кольцом целых чисел. В отличие от оригинальной схемы обязательств Дамгарда-Фуджисаки, порядок используемой группы не содержит малых сомножителей. Оригинальный протокол демонстрации знания значений, из которых получены обязательства использует выбор дополнительных случайных значений из большого интервала для достижения статистической неотличимости моделируемой стенограммы.

В качестве обязательства из строки предложенный протокол использует множество обязательств, полученных из значения c_j и позиции i_j каждого символа обеих строк:

$$S \to \{(c_j,i_j)\}, \quad j = 1 \ldots |S|$$
$$CM(S,\{(\alpha_j,\beta_j)\}) = \{(g^{c_j}h^{\alpha_j}, g^{i_j}h^{\beta_j})\} \tag{2}$$

Обязательства из строки-шаблона, строки-текста, множество обязательств из смещений шаблона в тексте, количество (K) копий шаблона, а также параметры схемы обязательств (N,g,h) являются информацией, доступной Проверяющему в предложенном протоколе.

Распознавание языка путем проверки выполнимости утверждений о полиномах с использованием ответов протокола Шнорра было предложено для протокола демонстрации знания незначительно отличающихся множеств [12], а также использовалось для протокола демонстрации знания кодового слова кода Гоппы [13].

3 Полиномиальное представление

Каждой строке, представленной в виде множества пар значение-позиция символов, поставим в соответствие *характеристический полином* над кольцом целых чисел:

$$S : \{(c_j,i_j)\} \to F(x,y;S) \tag{3}$$
$$F(x,y;S) = \prod_{j=1}^{|S|}(1 + xc_j + yi_j) \tag{4}$$

Характеристический полином является произведением взаимно простых элементарных линейных полиномов, соответствующих отдельным символам, а отображение строк на кольцо полиномов двух переменных является инъективным.

Полиномиальное представление множества (ранее использовался термин характеристическая функция) было предложено для задачи согласования множеств (set reconciliation) [14, 15]. Модифицированное полиномиальное представление множества использовалось для построения протокола аргумента знания незначительно отличающихся множеств [12]. Было также получено полиномиальное представление ориентированного графа, использованное в протоколах аргумента знания изоморфизма, раскраски вершин и цикла Гамильтона [16].

4 Оценка вероятности

Для оценки вероятности обмана (soundness error) используется следующая лемма (одномерный случай):

Лемма 1 (Schwartz,Zippel [17]). *Вероятность случайно выбрать корень полинома степени d выбирая значение аргумента из некоторого множества $x \in D$ с равномерным распределением*

$$\Pr[f(x) = 0] \le \frac{d}{|D|} \tag{5}$$

Этот простой результат позволил получить ряд альтернативных и принципиально новых протоко-

лов, в том числе аргумент знания слова кода Гоппы, полинома Гоппы и ошибки ограниченного веса [13], а также аргумент $x > 0$ над подмножеством ограниченных сверху целых чисел [18].

5 Проверочный полином

Существование полинома, построенного из запросов Проверяющего и корректных ответов Доказывающего, и позволяющего показать необходимое условие, является решающим при построении протокола. Пусть имеются обязательства, полученные из символов строки (c_j, i_j), смещений o_k и флагов $r_j \in \{0,1\}$, а также корректные ответы (над кольцом целых чисел) $\{C_j\}$, $\{I_j\}$, $\{O_k\}$, $\{R_j\}$ протокола Шнорра для запросов (e,d,s) Проверяющего и начальных случайных значений $\{\alpha_j\}$, $\{\beta_j\}$, $\{\gamma_k\}$, $\{\rho_j\}$:

$$C_j = ec_j + \alpha_j \tag{6}$$
$$I_j = di_j + \beta_j \tag{7}$$
$$O_k = do_k + \gamma_k \tag{8}$$
$$R_j = sr_j + \rho_j \tag{9}$$

Доказывающий использует флаги $r_j = 1$ для всех символов строки-текста, которые являются символами копий строки-шаблона; для всех остальных символов текста $r_j = 0$. Рассмотрим следующие *проверочные полиномы*, построенный из ответов для шаблона S_P и текста S_H:

$$F_P(x,y;S_P) = \prod_{k=1}^{K}\prod_{j=1}^{L_P}(ed + xdC_{Pj} + ye(I_{Pj} + O_k))$$

$$F_H(x,y;S_H) = \prod_{j=1}^{L_H}(eds + R_j(xdC_{Hj} + yeI_{Hj}))$$

Рассмотрим старший коэффициент полинома:

$$F_H(x,y;S) = s^{L_H}F'_H(x,y) + \sum_{t=0}^{L_H-1} s^t v_t$$

$$F'_H(x,y) = (ed)^{L_H - KL_P}\prod_{j=1}^{KL_P}(ed + xdC_j + yeI_j)$$

Обозначим $n = KL_P$. В произведении для $F'_H(x,y)$ присутствуют только символы, отмеченные флагом $'1'$, то есть символы из всех копий шаблона. Рассмотрим старшие коэффициенты:

$$F'_H(x,y) = (ed)^{L_H - n}\sum_{l=0}^{n} e^l a_l(x,y) \tag{10}$$

$$a_n(x,y) = \sum_{m=0}^{n} d^m b_m(x,y) \tag{11}$$

$$b_n(x,y) = \prod_{k=1}^{K}\prod_{j=1}^{L_P}(1 + xc_j + y(i_j + o_k)) \tag{12}$$

Аналогично для проверочного полинома шаблона:

$$F_P(x,y;S) = \sum_{l=0}^{n} e^l a'_l(x,y) \tag{13}$$

$$a'_n(x,y) = \sum_{m=0}^{n} d^m b'_m(x,y) \tag{14}$$

$$b'_n(x,y) = \prod_{k=1}^{K}\prod_{j=1}^{L_P}(1 + xc_j + y(i_j + o_k)) \tag{15}$$

Таким образом имеет место

Лемма 2. *Старшие коэффициенты проверочных полиномов совпадают:*

$$b'_n(x,y) \equiv b_n(x,y) \tag{16}$$

тогда и только тогда, когда в стоке-тексте имеется K копий строки-шаблона.

В рамках предложенного протокола Доказывающий помогает Проверяющему убедиться в том, что

$$b'_n(x_c,y_c) = b_n(x_c,y_c) \tag{17}$$

для значений (x_c, y_c), выбранных Проверяющим в качестве запроса.

6 Протокол

Пусть (N,g,h) - параметры схемы обязательств. Общей информацией Проверяющего и Доказывающего являются:

- параметры схемы;
- обязательства из значений и позиций символов строки-шаблона $\{(W_{Pj}, M_{Pj})\}, j = 1 \ldots L_P$
- обязательства из значений и позиций символов строки-текста $\{(W_{Hj}, M_{Hj})\}, j = 1 \ldots L_H$
- обязательства из смещений $\{Q_k\}, k = 1 \ldots K$

Информацией Доказывающего являются значения c_j и позиции i_j символов обеих строк, смещения o_k, а также дополнительные случайные значения $\theta_{Pj}, \phi_{Pj}, \theta_{Hj}, \phi_{Hj}, \lambda_k$ такие, что:

$$W_{Pj} = g^{c_{Pj}}h^{\theta_{Pj}}, \quad M_{Pj} = g^{i_{Pj}}h^{\phi_{Pj}}$$
$$W_{Hj} = g^{c_{Hj}}h^{\theta_{Hj}}, \quad M_{Hj} = g^{i_{Hj}}h^{\phi_{Hj}}$$
$$Q_k = g^{o_k}h^{\lambda_k}$$

Стороны выполняют следующий протокол:

1. Доказывающий назначает флаги $r_j = 1$ символам текста, которые являются символами любой из K копий шаблона, а также флаги $'0'$ всем остальным символам текста. Доказывающий выбирает дополнительные случайные значения ω_j, получает обязательства

$$L_j = g^{r_j}h^{\omega_j} \qquad j = 1 \ldots L_H \tag{18}$$

и пересылает $\{L_j\}$ Проверяющему.

2. Проверяющий выбирает запрос (x_c, y_c), и пересылает его Доказывающему. Значения x_c, y_c, а также запросы d, e, s выбираются случайным образом, являются ненулевыми, и ограничены сверху порядком используемой группы.

3. Доказывающий выбирает случайные $\{(\beta_{Hj}, \mu_{Hj})\}$, $\{(\beta_{Pj}, \mu_{Pj})\}$ и $\{(\gamma_k, \chi_k)\}$, получает коэффициенты $\{b_m\}$, $\{b'_m\}$:

$$\prod_{j=1}^{n}(z + x_c z c_{Hj} + y_c(z i_{Hj} + \beta_{Hj})) = \sum_{m=0}^{n} z^m b_m$$

так что в произведение входят только символы текста, отмеченные флагом '1' (то есть, символы из всех копий шаблона);

$$\prod_{k=1}^{K}\prod_{j=1}^{L_P}(z + x_c z c_{Pj} + y_c((z i_{Pj} + \beta_{Pj}) + (z o_k + \gamma_k))) = \sum_{m=0}^{n} z^m b'_m$$

Доказывающий выбирает случайные $\zeta_m, \zeta'_m, \psi, \tau, \tau'$, получает обязательства:

$$B_m = g^{b_m} h^{\zeta_m}, \quad B'_m = g^{b'_m} h^{\zeta'_m} \quad m = 0 \ldots n$$
$$Y_{Pj} = g^{\beta_{Pj}} h^{\mu_{Pj}} \qquad j = 1 \ldots L_P$$
$$Y_{Hj} = g^{\beta_{Hj}} h^{\mu_{Pj}} \qquad j = 1 \ldots L_H$$
$$X_k = g^{\gamma_k} h^{\chi_k} \qquad k = 1 \ldots K$$
$$Z = g^{\psi} h^{\tau}$$
$$Z' = g^{\psi} h^{\tau'}$$

Доказывающий пересылает $\{B_m\}$, $\{B'_m\}$, $\{Y_{Pj}\}$, $\{Y_{Hj}\}$, $\{X_k\}$, Z, Z' Проверяющему.

4. Проверяющий выбирает случайный запрос d и пересылает его Доказывающему.

5. Доказывающий получает ответы

$$I_{Hj} = d i_{Hj} + \beta_{Hj}, \quad \Phi_{Hj} = d\phi_{Hj} + \mu_{Hj}, \quad j = 1 \ldots L_H$$
$$I_{Pj} = d i_{Pj} + \beta_{Pj}, \quad \Phi_{Pj} = d\phi_{Pj} + \mu_{Pj}, \quad j = 1 \ldots L_P$$
$$O_k = d o_k + \gamma_k, \quad \Lambda_k = d\lambda_k + \chi_k, \quad k = 1 \ldots K$$
$$\Psi = d b_n + \psi, \quad \Gamma = d\zeta_n + \tau, \quad \Gamma' = d\zeta'_n + \tau'$$

Доказывающий выбирает случайные $\{(\alpha_{Hj}, \eta_{Hj})\}$ и $\{(\alpha_{Pj}, \eta_{Pj})\}$, получает коэффициенты $\{a_n\}$, $\{a'_n\}$:

$$\prod_{j=1}^{KL_P}(z d + x_c d(z c_{Hj} + \alpha_{Hj}) + y_c z(I_{Hj} + d o_k)) = \sum_{l=0}^{n} z^l a_l$$

$$\prod_{k=1}^{K}\prod_{j=1}^{L_P}(z d + x_c d(z c_{Pj} + \alpha_{Pj}) + y_c z(I_{Pj} + O_k)) = \sum_{l=0}^{n} z^l a'_l$$

Доказывающий выбирает случайные π_l, π'_l, получает обязательства:

$$A_l = g^{a_l} h^{\pi_l}, \quad A'_l = g^{a'_l} h^{\pi'_l} \quad l = 0 \ldots (n-1)$$
$$T_{Pj} = g^{\alpha_{Pj}} h^{\eta_{Pj}} \qquad j = 1 \ldots L_P$$
$$T_{Hj} = g^{\alpha_{Hj}} h^{\eta_{Hj}} \qquad j = 1 \ldots L_H$$

Доказывающий пересылает $\{I_{Hj}\}$, $\{\Phi_{Hj}\}$, $\{I_{Pj}\}$, $\{\Phi_{Pj}\}$, $\{O_k\}$, $\{\Lambda_k\}$, Ψ, $\{A_l\}$, $\{A'_l\}$, $\{T_{Pj}\}$, $\{T_{Hj}\}$ Проверяющему.

6. Проверяющий выбирает случайный запрос e и пересылает его Доказывающему.

7. Доказывающий получает ответы:

$$C_{Hj} = e c_j + \alpha_{Hj}, \quad \Theta_{Hj} = e\theta_{Hj} + \eta_{Hj}$$
$$C_{Pj} = e c_j + \alpha_{Pj}, \quad \Theta_{Pj} = e\theta_{Pj} + \eta_{Pj}$$

Доказывающий выбирает случайные (ρ_j, δ_j), получает коэффициенты $\{v_t\}$ и обязательства:

$$\prod_{j=1}^{L_H}(e d z + (z r_j + \rho_j)(x_c d C_{Hj} + y_c e I_{Hj})) = \sum_{t=0}^{L_H} z^t v_t$$

Доказывающий получает коэффициенты

$$(z r_j + \rho_j)(z(r_j - 1) + \rho_j) = u_{1j} z + u_{0j}$$

Доказывающий выбирает случайные $\sigma_t, \nu_{1j}, \nu_{0j}$, получает обязательства:

$$V_t = g^{v_t} h^{\sigma_t} \qquad t = 0 \ldots (L_H - 1)$$
$$N_j = g^{\rho_j} h^{\delta_j} \qquad j = 1 \ldots L_H$$
$$E_{1j} = g^{u_{1j}} h^{\nu_{1j}}, \quad E_{0j} = g^{u_{0j}} h^{\nu_{0j}}$$

Доказывающий пересылает $\{C_{Hj}\}$, $\{\Theta_{Hj}\}$, $\{C_{Pj}\}$, $\{\Theta_{Pj}\}$, $\{V_t\}$, $\{N_j\}$, $\{(E_{1j}, E_{0j})\}$ Проверяющему.

8. Проверяющий выбирает случайный запрос s и пересылает его Доказывающему.

9. Доказывающий получает ответы

$$R_j = s r_j + \rho_j, \quad j = 1 \ldots L_H$$
$$\Omega_j = s\omega_j + \delta_j$$
$$\Xi_j = s\nu_{1j} + \nu_{0j}$$
$$\Delta_P = e^n \sum_{m=0}^{n} d^m \zeta'_m + \sum_{l=0}^{n-1} e^l \pi'_l$$
$$\Delta_H = s^{L_H}(e^n \sum_{m=0}^{n} d^m \zeta_m + \sum_{l=0}^{n-1} e^l \pi_l) + \sum_{t=0}^{L_H - 1} s^t \sigma_t$$

Доказывающий пересылает $\{R_j\}, \{\Omega_j\}, \{\Xi_j\}, \Delta_P, \Delta_H$ Проверяющему.

10. Проверяющий тестирует корректность ответов (значения линейных полиномов):

$$g^{C_{Pj}}h^{\Theta_{Pj}}W_{Pj}^{-e} = T_{Pj} \quad (19)$$

$$g^{I_{Pj}}h^{\Phi_{Pj}}M_{Pj}^{-d} = Y_{Pj} \quad (20)$$

$$g^{C_{Hj}}h^{\Theta_{Hj}}W_{Hj}^{-e} = T_{Hj} \quad (21)$$

$$g^{I_{Hj}}h^{\Phi_{Hj}}M_{Hj}^{-d} = Y_{Hj} \quad (22)$$

$$g^{O_k}h^{\Lambda_k}Q_k^{-d} = X_k \quad (23)$$

$$g^{R_j}h^{\Omega_j}L_j^{-s} = N_j \quad (24)$$

Проверяющий тестирует корректность флагов ($r_j \in \{0,1\}$):

$$g^{-R_j(R_j-s)}h^{-\Xi_j}E_{1j}^{s}E_{0j} = 1 \quad (25)$$

Проверяющий получает значения проверочных полиномов:

$$U_P = \prod_{k=1}^{K}\prod_{j=1}^{L_P}(ed + x_c dC_j + y_c e(I_j + O_k))$$

$$U_H = (ed)^{-(L_H-n)}\prod_{j=1}^{L_H}(eds + R_j(x_c dC_j + y_c eI_j))$$

Проверяющий рассматривает демонстрацию как убедительную если:

$\{A'_l\},\{B'_m\}$ являются обязательствами из коэффициентов проверочного полинома шаблона:

$$g^{-U_P}h^{-\Delta_P}\left(\prod_{m=0}^{n}(B'_m)^{d^m}\right)^{e^n}\prod_{l=0}^{n-1}(A'_l)^{e^l} = 1 \quad (26)$$

$\{A_l\},\{B_m\}$ являются обязательствами из коэффициентов проверочного полинома текста:

$$g^{-U_H}h^{-\Delta_H}\left(\left(\prod_{m=0}^{n}(B_m)^{d^m}\right)^{e^n}\prod_{l=0}^{n-1}(A_l)^{e^l}\right)^{s^{L_H}} \prod_{t=0}^{L_H-1}(V_t)^{s^t} = 1 \quad (27)$$

обязательства B_n и B'_n получены из совпадающих старших коэффициентов ($b_n = b'_n$):

$$g^{\Psi}h^{\Gamma}B_n^{-d} = Z \quad (28)$$

$$g^{\Psi}h^{\Gamma'}(B'_n)^{-d} = Z' \quad (29)$$

7 Свойства протокола

Полнота: нетрудно убедиться, что честный Проверяющий всегда соглашается с аргументом честного Доказывающего в случае выполнения условия о наличии K копий шаблона в тексте.

Корректность: рассмотрим возможности, доступные произвольному Доказывающему в случае невыполнения условия о K копиях шаблона в тексте. Покажем что любой Доказывающий, успешно завершающий протокол с альтернативными (то есть, отличающимися от значений линейного полиномов) ответами также получает решение задачи SRSA. Предположим произвольный Доказывающий получил ответы (C',Θ') такие что $g^{C'}h^{\Theta'}W^{-e} = T$ (19), причем $C' \neq C$, $\Theta' \neq \Theta$ для $W = g^c h^{\theta}$, $T = g^{\alpha}h^{\eta}$, $C = ec + \alpha$, $\Theta = e\theta + \eta$. Тогда $g^{C'-C}h^{\Theta'-\Theta} = 1$. Тогда алгоритм Эвклида дает (a,b) такие что $a(C'-C) + b(\Theta'-\Theta) = gcd(C'-C,\Theta'-\Theta)$, откуда $(g^b h^{-a})^{\frac{\Theta-\Theta'}{gcd(C'-C,\Theta'-\Theta)}} = g$, то есть такой Доказывающий также находит решение задачи SRSA. В дальнейшем будем считать, что все ответы получены как значения соответствующих линейных полиномов.

Рассмотрим произвольного Доказывающего, использующего флаги $r_j \notin \{0,1\}$. Полином $R_r(z) = (zr_j + \rho_j)(z(r_j - 1) + \rho_j)$ является квадратичным. Согласно лемме Шварца-Зиппеля, вероятность $\Pr[R_r(s) = 0] \leq \frac{2}{q}$. Произвольный Доказывающий, успешно завершивший протокол (25) с $R_r(s) \neq 0$, также получает решение задачи SRSA из $R_r(s) = -R_j(R_j - s) + u_{1j}s + u_{0j}$ и $(-\Xi_j + \nu_{1j}s + \nu_{0j})$.

Произвольный Доказывающий, успешно завершивший протокол с $b_n(x_c,y_c) \neq b'_n(x_c,y_c)$, также получает решение задачи SRSA из ответов Ψ,Γ,Γ' в одном из двух уравнений (28, 29).

Можно показать, что оптимальной стратегией такого Доказывающего, дающей максимальную вероятность обмана, является попытка пройти уравнение проверки (27) используя значение b'_n вместо b_n. Полином $F_B(z) = (b_n(x_c,y_c) - b'_n(x_c,y_c))z^n + \sum_{m=0}^{n-1} b_m$ является ненулевым степени n. Согласно лемме Шварца-Зиппеля, вероятность $\Pr[F_B(d) = 0] \leq \frac{n}{q}$ как результат выбора Проверяющим запроса d.

Рассмотрим случай $F_B(d) \neq 0$. Полином $F_A(z) = F_B(d)z^n + \sum_{l=0}^{n-1} a_l z^l$ ненулевой степени n. Согласно лемме Шварца-Зиппеля, вероятность $\Pr[F_A(e) = 0] \leq \frac{n}{q}$ как результат выбора Проверяющим запроса e.

Рассмотрим случай $F_A(e) \neq 0$. Полином $F_V(z) = -U_H(z) + F_A(e)z^{L_H} + \sum_{t=0}^{L_H-1} v_t z^t$ является ненулевым, $\deg(F_s(z)) = L_H$. Согласно лемме Шварца-Зиппеля, вероятность $\Pr[F_V(s) = 0] \leq \frac{L_H}{q}$ как результат выбора Проверяющим запроса s.

Таким образом, суммарная вероятность (по всем вариантам выбора запросов d,e,s) $\Pr[F_V(s) = 0] \leq \frac{L_H+2n}{q}$. Произвольный Доказывающий, успешно завершающий протокол с $F_V(s) \neq 0$,

также получает решение задачи SRSA.

Таким образом, вероятность обмана честного Проверяющего произвольным Доказывающим не превышает $\frac{L_H+2KL_P}{q}$, при условии сложности задачи SRSA для Доказывающего.

Знание: как было показано, ответы Проверяющего являются значениями линейных полиномов при условии сложности задачи SRSA. Следовательно, экстрактор протокола Шнорра также является экстрактором предложенного протокола.

Нулевое разглашение: стенограммой протокола являются сообщения $\{L_j\}$; (x_c, y_c); $\{B_m\}$, $\{B'_m\}$, $\{Y_{Pj}\}$, $\{Y_{Hj}\}$, $\{X_k\}$, Z, Z'; d; $\{I_{Hj}\}$, $\{\Phi_{Hj}\}$, $\{I_{Pj}\}$, $\{\Phi_{Pj}\}$, $\{O_k\}$, $\{\Lambda_k\}$, Ψ, Γ, Γ', $\{A_l\}$, $\{A'_l\}$, $\{T_{Pj}\}$, $\{T_{Hj}\}$; e; $\{C_{Hj}\}$, $\{\Theta_{Hj}\}$, $\{C_{Pj}\}$, $\{\Theta_{Pj}\}$, $\{V_t\}$, $\{N_j\}$, $\{(E_{1j}, E_{0j})\}$; s; $\{R_j\}, \{\Omega_j\}, \{\Xi_j\}, \Delta_P, \Delta_H$.

Рассмотрим следующий алгоритм, который может быть выполнен честным Проверяющим. Исходными данными является общая информация (q, g, h), $\{(W_{Pj}, M_{Pj})\}$, $\{(W_{Hj}, M_{Hj})\}$, $\{Q_k\}$, и запросы (x_c, y_c, d, e, s). Проверяющий :

1. выбирает произвольные элементы группы $\{E_{1j}\}$, $\{B'_m\}$, $\{A'_l\}_{l=1..\,n-1}$, $\{B_m\}$, $\{A_l\}$, $\{V_t\}_{t=1...L_H-1}$
2. выбирает произвольные числа из соответствующих интервалов $\{C_{Pj}\}$, $\{\Theta_{Pj}\}$, $\{I_{Pj}\}$, $\{\Phi_{Pj}\}$, $\{C_{Hj}\}$, $\{\Theta_{Hj}\}$, $\{I_{Hj}\}$, $\{\Phi_{Hj}\}$, $\{O_k\}$, $\{\Lambda_k\}$, $\{R_j\}$, $\{\Omega_j\}$, Δ_P, Δ_H, Ψ, Γ, Γ'
3. получает

$$T_{Pj} = g^{C_{Pj}} h^{\Theta_{Pj}} W_{Pj}^{-e}$$
$$Y_{Pj} = g^{I_{Pj}} h^{\Phi_{Pj}} M_{Pj}^{-d}$$
$$T_{Hj} = g^{C_{Hj}} h^{\Theta_{Hj}} W_{Hj}^{-e}$$
$$Y_{Hj} = g^{I_{Hj}} h^{\Phi_{Hj}} M_{Hj}^{-d}$$
$$X_k = g^{O_k} h^{\Lambda_k} Q_k^{-d}$$
$$N_j = g^{R_j} h^{\Omega_j} L_j^{-s}$$
$$E_{0j} = g^{R_j(R_j-s)} h^{\Xi_j} E_{1j}^{-s}$$
$$U_P = \prod_{k=1}^{K} \prod_{j=1}^{L_P} (ed + x_c d C_j + y_c e (I_j + O_k))$$
$$U_H = (ed)^{-(L_H-n)} \prod_{j=1}^{L_H} (eds + R_j (x_c d C_j + y_c e I_j))$$
$$A'_0 = g^{U_P} h^{\Delta_P} \left(\prod_{m=0}^{n} (B'_m)^{d^m} \right)^{-e^n} \prod_{l=1}^{n-1} (A'_l)^{-e^l}$$
$$V_0 = g^{U_H} h^{\Delta_H} \left(\left(\prod_{m=0}^{n} (B_m)^{d^m} \right)^{e^n} \prod_{l=0}^{n-1} (A_l)^{e^l} \right)^{-s^{L_H}} \prod_{t=1}^{L_H-1} (V_t)^{-s^t}$$
$$Z = g^{\Psi} h^{\Gamma} B_n^{-d}$$
$$Z' = g^{\Psi} h^{\Gamma'} (B'_n)^{-d}$$

Алгоритм допускает выбор независимых значений $\Xi_j, \Delta_P, \Delta_H, \Gamma, \Gamma'$ из определенных достаточно больших интервалов, так что альтернативная стенограмма, полученная этим алгоритмом, может быть статистически неотличима от стенограммы протокола с Доказывающим.

8 Обсуждение

Предполагается, что задача демонстрации K копий строки-шаблона в строке-тексте может быть удобной моделью управления доступом к информации в некоторых задачах биоинформатики.

Список литературы

[1] Ronald Cramer and Victor Shoup. Signature schemes based on the strong RSA assumption. *ACM Trans. Inf. Syst. Secur.*, 3(3):161–185, 2000.

[2] Ivan Damgård and Eiichiro Fujisaki. A statistically-hiding integer commitment scheme based on groups with hidden order. In *ASIACRYPT*, pages 125–142, 2002.

[3] Shafi Goldwasser, Silvio Micali, and Charles Rackoff. The knowledge complexity of interactive proof systems. *SIAM J. Comput.*, 18(1):186–208, 1989.

[4] Oded Goldreich, Silvio Micali, and Avi Wigderson. Proofs that yield nothing but their validity for all languages in NP have zero-knowledge proof systems. *J. ACM*, 38(3):691–729, 1991.

[5] Ященко В.В. *Введение в криптографию*. МЦНМО, 2003.

[6] Gilles Brassard, David Chaum, and Claude Crépeau. Minimum disclosure proofs of knowledge. *J. Comput. Syst. Sci.*, 37(2):156–189, 1988.

[7] Gilles Brassard, Claude Crépeau, and Moti Yung. Everything in *NP* can be argued in perfect zero-knowledge in a bounded number of rounds. In *ICALP*, pages 123–136, 1989.

[8] Варновский Н.П. Типы нулевого разглашения. Рукопись, 2002.

[9] Mihir Bellare, Silvio Micali, and Rafail Ostrovsky. The (true) complexity of statistical zero knowledge. In *STOC*, pages 494–502, 1990.

[10] Ronald Cramer, Ivan Damgård, and Berry Schoenmakers. Proofs of partial knowledge and simplified design of witness hiding protocols. In *CRYPTO*, pages 174–187, 1994.

[11] Claus-Peter Schnorr. Efficient identification and signatures for smart cards. In *CRYPTO*, pages 239–252, 1989.

[12] Федюкович В.Е. Изменчивые ключи подписи. In *Безопасность информации в информационно-телекоммуникационных системах (Киев)*, 2007. Имеются слайды.

[13] Vadym Fedyukovych. Argument of knowledge of a bounded error. In *Coding Theory Days (accepted)*, 2008. Presentation available.

[14] Y. Minsky, A. Trachtenberg, and R. Zippel. Set reconciliation with nearly optimal communication complexity. In *International Symposium on Information Theory*, page 232, 2001. `http://citeseer.ist.psu.edu/minsky00set.html`.

[15] Yaron Minsky, Ari Trachtenberg, and Richard Zippel. Set reconciliation with nearly optimal communication complexity. *IEEE Transactions on Information Theory*, 49(9):2213–2218, 2003.

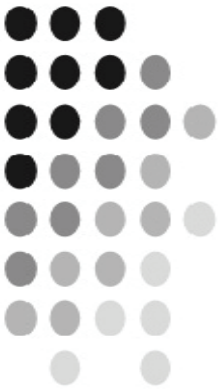
[16] Vadym Fedyukovych. Protocols for graph languages. In *(submitted)*, 2008.

[17] J. T. Schwartz. Fast probabilistic algorithms for verification of polynomial identities. *J. ACM*, 27(4):701–717, 1980.

[18] Федюкович В.Е. Построение интерактивных протоколов доказательства свойств полиномов и их приложения. In *Безопасность информации в информационно-телекоммуникационных системах (Киев)*, 2008. Имеются слайды.

Раздел 8

Распознавание образов и обработка изображений



Локализация семантических блоков на растровых изображениях без предварительного обучения

С. В. Кулешов, А. А. Зайцева
Санкт-Петербургский институт информатики и автоматизации РАН
kuleshov@iias.spb.su

Аннотация

Рассматривается метод объектной локализации семантических блоков на растровых изображениях, особенностью которого является отсутствие необходимости обучения системы на выделение конкретных видов сигнатур.

В основе предлагаемого метода лежит анализ результатов адаптивного пирамидального представления растровых изображений.

Особенностью алгоритма, формирующего пирамидальное представление, является связь величины значений последнего уровня иерархии и уровня семантической значимости, что позволяет находить области, представляющие наибольший интерес при восприятии их человеком.

Предложенный алгоритм способен успешно решать как задачи выделения неизвестных объектов на изображениях одной сцены, полученных при неизвестных условиях, так и задачи поиска известных объектов на неизвестном фоне.

1. Введение

Результатом распознавания лиц, идентификации речи часто является идентификация некоторого определенного объекта. Человеку в процессе идентификации совсем не обязательно явным образом определять характерные признаки объекта. Имеет значение, в сущности, только окончательный результат процесса наблюдения, восприятия и распознавания. Автоматические системы, решающие проблемы распознавания посредством органов чувств человека, должны осуществлять такую же идентификацию, как и человек, сталкивающийся с определенным объектом. Однако распознающие системы обычно используют характерные признаки объекта, и задаются в том же формате, что и сам объект. При этом для принятия решения о факте локализации необходима некоторая процедура вычисления схожести двух представлений.

2. Описание метода

В основе предлагаемого метода объектной локализации семантических блоков лежит анализ результатов адаптивного пирамидального представления растровых изображений [1, 2, 3].

Основной принцип пирамидального представления приведен на рисунке 1 и заключается в том, что для минимизации битовой размерности передаваемых элементов имеет смысл передавать разность не относительно пустого изображения, а относительно похожего изображения, так как в этом случае отличия будут минимальны.

Особенностью алгоритма, формирующего пирамидальное представление, является связь величины значений последнего уровня иерархии (для элементов с размером 1×1) и уровня семантической значимости, что позволяет находить области, представляющие наибольший интерес при восприятии их человеком.

Элементом разложения (на заданном уровне разложения) назовем неделимый фрагмент изображения, воспринимаемый алгоритмом как одно целое, имеющее яркостные и цветовые характеристики, определяемые как средние значения соответствующих характеристик входящих в него пикселей. Исходное изображение (первый уровень разложения) воспринимается как один элемент размером $W\times H$ где W — ширина изображения, H — высота изображения в пикселях.

На каждом последующем уровне разложения происходит разбиение каждого элемента на элементы, имеющие линейный размер в 2 раза меньше предыдущего, пока не достигается размер элементов в 1×1 пиксель. Такое последовательное разложение назовем пирамидальным представлением изображения.

Алгоритм, показанный на рисунке 1, одновременно с формированием пирамидального

представления определяет уровни, на которых находятся терминальные зоны. Выделению подлежат участки, соответствующие терминальным зонам с наибольшим номером уровня.

Назовем терминальной зоной элемент уровня разбиения, последующая детализация которого не дает дополнительной информации.

Терминальная зона отвечает критерию

$$\sum_i d_i > \varepsilon \text{ ,}$$

где d_i — величина изменения элемента на следующем иерархическом уровне, ε — величина порога неразличимости.

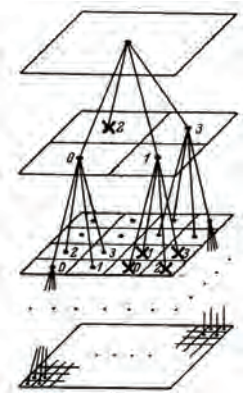


Рисунок 1. Иллюстрация алгоритма локализации областей при пирамидальном представлении (знаком «x» отмечены терминальные зоны).

Недостатком алгоритма является привязанность границ выделения к сетке разбиения изображения. Для компенсации этого недостатка возможно многократное повторение алгоритма со сдвигом центра первого уровня разбиения и формирование границ выделения по совокупности результатов каждого из разбиений. Основным применением такого алгоритма могут стать алгоритмы предобработки, компрессии с потерями и передачи по линиям с фиксированной пропускной способностью с динамическим изменением качества фрагментов изображения для компенсации психофизиологического фактора восприятия этих изображений.

3. Результаты работы алгоритма

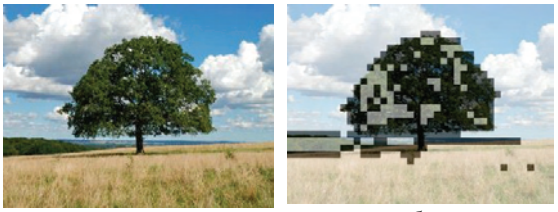
Примеры работы, приведенные на рис. 2, 3 показывают устойчивое выделение области интереса (фрагментов, имеющих повышенную сложность) на фотографиях с достаточно сложным фоном. На рисунках 2*а* и 3*а* приведены исходные изображения, на рисунках 2*б* и 3*б* — сумма исходного изображения и маски, определяющей область интереса.

По сравнению с примером на рисунке 2, на котором объект, задающий область интереса находится на однородном фоне, пример на рисунке 3 показывает наличие дополнительных объектов на заднем фоне, усложняющих задачу локализации семантически значимых областей.

Предложенный алгоритм способен успешно решать как задачи выделения неизвестных объектов на изображениях одной сцены, полученных при неизвестных условиях, так и задачи поиска известных объектов на неизвестном фоне.

а *б*

Рисунок 2. Пример работы алгоритма на простом изображении



а *б*

Рисунок 3. Пример работы алгоритма на сложном изображении

Особенностью предложенного алгоритма в отличие от алгоритмов поиска границ является его нечувствительность к шуму, т.е. при его использовании выделения шума не наблюдается.

Основное применение метода – динамическое управление степенью компрессии [4]. Результаты экспериментов показывает, что визуально изображения с динамическим управлением степенью компрессии с потерями (lossy compression) лучше изображения с фиксированным значением компрессии при сравнимом битовом объеме.

4. Литература

[1] Александров В. В., Горский Н. Д. *Представление изображений. Рекурсивный подход*. — Л.: Наука. 1985. 190 с.

[2] Кулешов С. В., Зайцева А. А., Аксенов А. Ю. *Ассоциативно-пирамидальное представление данных*. — Информационно-измерительные и управляющие системы. —2008. №4.

[3] Александров В. В., Кулешов С. В., Цветков О. В. *Цифровая технология инфокоммуникации. Передача, хранение и семантический анализ текста, звука, видео*. — СПб.: Наука. 2008. 244 с.

[4] Александров В. В., Кулешов С. В., Цветков О. В. *Концепция программируемой технологии цифровой теории связи: от герц к бит/с* // Информационно-измерительные и управляющие системы. 2007. № 6. — С. 62-72.

Быстрый алгоритм совмещения изображений документов в произвольной геометрической модели

Сергей Усилин
Московский физико-технический институт (государственный университет)
usilin.sergey@gmail.com

Дмитрий Николаев
Институт проблем передачи информации имени А. А. Харкевича РАН
dimonstr@iitp.ru

Василий Постников
Cognitive Technologies
pvv@cognitive.ru

Аннотация

В работе описан новый алгоритм определения типа документа, основанный на динамическом совмещении изображения образца с исследуемым изображением. Ключевой особенностью алгоритма является использование стандартизованного описания произвольной геометрической модели документа. Это делает алгоритм применимым для большого количества различных классов документов. Второй особенностью алгоритма является медленный (сублинейный) рост вычислительных затрат при увеличении количества типов документов, к которым может принадлежать образец.

1. Введение

В современном мире огромное количество бумажных документов ежедневно переводится в электронную форму. Хранение электронной копии документа обладает рядом преимуществ по сравнению с хранением бумажного оригинала: сохранность внешнего вида документа на протяжении всего времени хранения, простота обмена документами, мгновенный поиск информации.

Наиболее распространенной формой создания электронных образов документов является сканирование и распознавание бумажных оригиналов [1]. С внедрением систем электронного документооборота в коммерческих организациях и появлением электронных библиотек потребность в системах распознавания документов значительно возросла. Стоит отметить, что ввод документа не сводится к распознаванию текста. Сначала необходимо проанализировать графический образ документа: определить тип, провести разметку станицы (выделить на изображении документа графические элементы, текстовые блоки и пр.).

Задача понимания графического образа документа (Document Image Analysis) активно исследуется более 20 лет. Однако разработанные на сегодняшний день алгоритмы не обладают необходимыми качествами: высокой производительностью, устойчивостью к искажениям и шуму, инвариантностью к геометрической трансформации формы. Стоит также отметить, что большинство работ, посвященных методам обработки изображений документов, обладают общей особенностью: задачи ставятся и решаются изолированно от контекста, то есть не используют информацию о том, что анализируемое изображение - это изображение документа. Как оказывается, такой подход не всегда оправдан, а универсальность подобных методов может носить мнимый характер. Например, бинаризация изображения является составляющей частью практически любой процедуры предобработки изображения, однако не существует алгоритма бинаризации одинаково хорошо обрабатывающего различные типы изображения [2].

Наибольшую популярность в современных системах распознавания изображений документов получил подход, основанный на выделении и сравнении заранее определенных графических примитивов. Большое количество работ посвящено этому подходу [3 - 6]. После предварительной обработки изображения документа (поворота, компенсирования искажений, исключения областей текста, которые могут повлиять на поиск графических примитивов) производится поиск простых графических элементов: линий, таблиц, специально нанесенных на оригинальный документ служебных символов. Исследуя взаиморасположение найденных графических примитивов можно сделать вывод о принадлежности исследуемого документа к тому или иному типу. Стоит отметить, что изложенный подход, несмотря на большую популярность,

обладает существенным недостатком: опорные графические примитивы могут отсутствовать на документе.

Существует другой класс алгоритмов, основывающихся на сопоставлении так называемых компонентных блоков [7, 8]. На подготовленном изображении документа выделяются компонентные блоки – области изображения, содержащие значимую информацию. Примерами таких блоков могут служить области текста, рукописных пометок, графические элементы и др. Найденные компонентные блоки группируются в определенные структуры данных, которые впоследствии сравниваются с соответствующими структурами, построенными для эталонных изображений. Типом исследуемого документа считается тип эталона, который совместился с исследуемым изображением наилучшим образом. Однако рассмотренный подход имеет существенный недостаток: геометрическая модель в нем фиксирована. Документ считается топологически упорядоченным. Это может вызвать проблему при вводе набора документов близких по топологической структуре, но различающихся по информационной сути. К тому же, наличие «неоговоренных» помарок на документе может привести к ложному определению дополнительных блоков компонент, что может повлечь за собой неправильную идентификацию документа.

Помимо алгоритмов, выполняющих глобальное сопоставление изображение, существуют алгоритмы, которые используют локальное сопоставление (то есть независимо сопоставляют отдельные области изображений), а также алгоритмы, комбинирующие в себе эти два подхода. В качестве примера последнего можно привести алгоритм, основанный на совмещении гистограмм изображений [9]. После предварительной подготовки, изображение разбивается на области заранее заданного размера. Определенным образом строятся гистограммы количества черных пикселей по горизонтали и по вертикали для всего изображения и отдельно для каждой области. Построенные гистограммы сравниваются с соответствующими гистограммами, построенными для каждого эталона, после чего выбирается эталон, наилучшим образом совместившийся с исследуемым изображением. Можно указать два недостатка этого подхода. Во-первых, два скана одного и того же документа после бинаризации могут существенно отличаться (не существует процедур бинаризации робастных к параметрам сканирования [2, 10]). Во-вторых, данный подход, как и в предыдущих случаях, обладает фиксированной геометрической моделью.

Приведенный краткий обзор методов сопоставления изображений документов показывает, что эта задача не нова и довольно глубоко проработана. Однако все рассмотренные методы обладают существенными недостатками. В частности, каждый из рассмотренных методов рассматривает документы с жесткой геометрической моделью.

Оказывается, задача идентификации типа документа по изображению может быть поставлена как задача сопоставления изображений. В этой постановке предполагается, что модель документа – это (может быть, синтетическое) изображение идеального документа.

В данной работе описан новый подход к решению проблемы, основанный на идее совмещения изображений, в рамках которого мы попытались удовлетворить следующим требованиям:

- возможность идентификации документов, не содержащих предопределенных графических примитивов;
- поддержка различных геометрических моделей документа;
- частичная инвариантность к различным параметрам сканирования;
- инвариантность к заполнению документа.

2. Модель документа

Каждый документ содержит статические элементы (области, одинаковые для всех экземпляров данного типа документа) и поля заполнения (области, отличные на каждом экземпляре типа документа). Например, на Рис. 1 синим и зеленым выделены статические элементы, красным – изменяемые.

Графические образы однотипных документов могут отличаться заполнением и взаимо-расположением статических элементов (области документа, одинаковые для всех экземпляров данного типа документа) за счет наличия сдвига и наклона документа на стекле сканера, и, возможно, других допустимых трансформаций. Например, развороты паспорта геометрически различны из-за неточности переплетения, общий масштаб анкеты зависит от настроек принтера, а расстояния между секциями форм InfoPath зависят от числа строк заполнения (см. Рис. 1).

Рис. 1. Пустая и заполненная формы InfoPath. Синим и зеленым указаны статические элементы, красным – изменяемые

Моделью таких документов является набор обязательных статических элементов совокупно с множеством допустимых трансформаций. Без ограничения общности будем задавать статические элементы прямоугольными областями.

3. Алгоритм совмещения изображений

Рассмотрим изображение документа, тип которого требуется определить. Процесс совмещения его с некоторым эталонным изображением будем рассматривать как процесс последовательного поиска статических элементов, характерных для данного эталона, на исследуемом изображении. При поиске элемента могут появиться альтернативные решения. Возникает переборная задача: отбор вариантов расположения элементов, соответствующих допустимой трансформации. В процессе перебора параметры трансформации ограничены набором уже сопоставленных элементов. Такую «ограниченную трансформацию» можно всегда задавать отображением из точки в область. Такое универсальное описание позволяет разделить задачу совмещения изображений на две независимые подзадачи:

1. определение области поиска очередного элемента в соответствии с возможными трансформациями статических элементов;
2. поиск элемента в указанной области.

Ясно, что для решения первой подзадачи достаточно лишь знать геометрическую модель изображения документа, которая определяет множество допустимых трансформаций статических элементов, тогда как вторая подзадача может быть решена с помощью алгоритмов поиска элемента в заданной области изображения, независимо от геометрической модели.

Рассмотрим принцип действия переборного алгоритма и решение каждой подзадачи более подробно.

3.1. Переборный алгоритм

Позицией перебора назовём набор сопоставленных статических элементов и накопленный штраф. В процессе перебора позиции образуют дерево, порождая дочерние позиции с расширенным набором сопоставленных элементов с возможными вариантами сопоставления и увеличенным штрафом. Начальная позиция (корень дерева) - позиция без найденных элементов со штрафом 0. Переборный алгоритм представим в виде последовательности шагов.

Шаг 1. Поместим начальную позицию в кучу [11].

Шаг 2. Выберем из кучи наилучшую позицию P. Наилучшей считается позиция с наименьшим штрафом.

Шаг 3. Выберем из очереди несопоставленных элементов очередной элемент E.

Шаг 4. В соответствии с геометрической моделью изображения документа и множеством уже найденных элементов в позиции P определим область поиска элемента E (см. раздел 3.2).

Шаг 5. С помощью технологии поиска элемента определим варианты возможного расположения элемента E на исследуемом изображении (см. раздел 3.3). Ошибка расположения каждого варианта определяется величиной штрафа . Допустимыми вариантами расположения считаются варианты, штраф которых не превышает заданного порогового значения.

Шаг 6. Если допустимых вариантов нет или их количество слишком велико (больше заданного значения), то элемент E считается несопоставимым на позиции P. При этом, если элемент еще не был помечен, то помечаем его как несопоставимый, передвигаем в конец очереди несопоставленных элементов, затем переходим на Шаг 3; иначе переходим к следующему шагу.

Шаг 7. Каждый допустимый вариант расположения элемента E порождает новую позицию, в которой множество найденных элементов расширено на элемент E и штраф увеличен на величину . Новые позиции положим в кучу взамен P.

Шаг 8. Проверяем состояние позиций. В случае если появилась позиция, в которой все статические элементы эталона найдены, то объявляем ее результатом совмещения

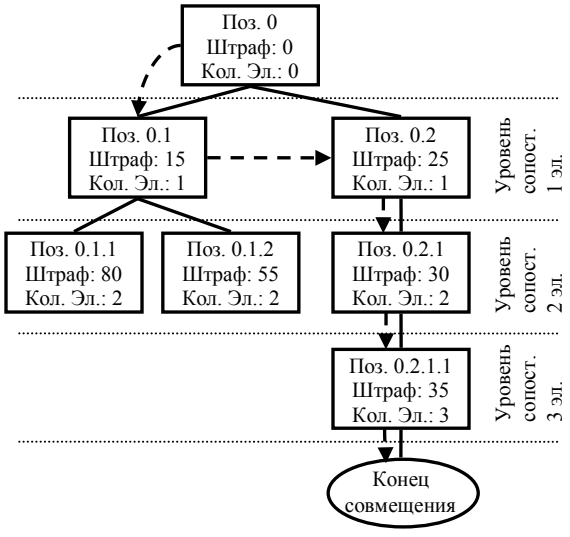


Рис. 2. Дерево перебора позиций при поиске 3-х статических элементов

изображений. Если все позиции находятся в статусе несопоставимости с некоторым статическим элементом, то рассматриваемое изображение считается несовместимо с данным эталоном. В противном случае переходим на Шаг 2 и производим поиск следующего статического элемента.

На Рис. 2 схематично изображено дерево перебора позиций при поиске трех элементов. Прямоугольниками отмечены соответствующие позиции перебора, а стрелками – последовательность выбора наилучшей позиции.

3.2. Геометрическая модель

Зная геометрическую модель изображения документа и расположение уже найденных элементов, можно «предсказать» область расположения остальных элементов. Например, для изображения документа, представленного на Рис. 1, зная расположение элемента 1 можно утверждать, что элемент 2 должен быть расположен строго ниже элемента 1, но не ближе, чем некоторое минимальное расстояние.

Рассмотрим теперь геометрические модели характерные для обширного числа документов, часто встречающихся на практике с соответствующими алгоритмами построения «предсказания» расположения элементов.

3.2.1 «Сдвиг и поворот». Такая модель подходит для перевода в электронный формат документов строго определенной формы или документов типографского происхождения. Примерами могут служить водительское удостоверение, регистрационные карточки, стандартизованные квитанции. Расположения статических элементов на различных графических образах таких документов могут отличаться только лишь по причине неаккуратного расположения документа на стекле сканера. Отсюда следует, что множество допустимых трансформаций: одинаковые для всех элементов угол поворота и смещение относительно расположения элементов на эталонном изображении.

Рассмотрим задание такой модели с математической точки зрения. Пусть на эталонном изображении выделено n статических элементов с координатами центров $(x_i^e, y_i^e), i = \overline{,n}$. Тогда расположение элементов на исследуемом изображении $(x_i^t, y_i^t), i = \overline{,n}$ можно выразить следующим образом:

$$\begin{cases} x_i^t = x_i^e \cdot \cos\alpha \ - y_i^e \cdot \sin\alpha \ - x_0 \\ y_i^t = x_i^e \cdot \sin\alpha \ - y_i^e \cdot \sin\alpha \ - y_0 \end{cases}, i = \overline{1,n} \qquad (1)$$

Здесь α и (x_0, y_0) соответственно угол поворота и сдвиг элементов на исследуемом изображении относительно элементов на эталоне.

Рассмотрим, как с помощью такой модели определяется область поиска очередного элемента. Пусть требуется определить координаты $k+1$ элемента при известных координатах первых k элементов. Решим задачу определения сдвига и угла поворота, которые минимизируют сумму квадратов ошибок расположения найденных k элементов от предположения модели:

$$\begin{cases} x_i^t = x_i^e \cdot \cos\alpha \ - y_i^e \cdot \sin\alpha \ - x_0 + \xi \\ y_i^t = x_i^e \cdot \sin\alpha \ - y_i^e \cdot \sin\alpha \ - y_0 + \xi \end{cases}, i = \overline{,k} \qquad (2)$$

$$\sum_{i=}^{k} \xi \ \tilde{} + \xi \ \tilde{} \rightarrow \ \text{in}$$

Определив сдвиг и угол поворота из задачи (2), можно вычислить предположительные координаты $k+1$ элемента. Тогда область поиска задается по формуле:

$$U^{k+} = U^0_{d(k)}(x^t_{k+}, y^t_{k+}) \qquad (3)$$

В формуле (3) U^{k} и $U^0_{d(k)}(x^t_{k+}, y^t_{k+})$ соответственно область поиска $k+1$ элемента и некоторая заданная область с координатами в точке (x^t_{k+1}, y^t_{k+1}) и характерным размером $d(k)$, монотонно убывающим с ростом k.

Отметим, что на практике угол α по абсолютному значению не превышает 5 градусов. Для таких углов справедливо $\sin\alpha \approx \alpha$ и $\cos\alpha \approx 1$. Эти приближения позволяют существенно облегчить решение задачи (2).

3.2.2. «Сдвиг, поворот, линейное сжатие (общее аффинное преобразование)». Данная модель подходит для перевода в электронный формат документов, распечатанных на принтерах. В силу наличия у большинства принтеров непечатных полей, в процессе печати документ может быть подвержен линейному сжатию. Математически такую модель можно представить следующим образом:

$$\begin{cases} x_i^t = a \cdot (x_i^e \cdot \cos\alpha \ - y_i^e \cdot \sin\alpha) + x_0 \\ y_i^t = b \cdot (x_i^e \cdot \sin\alpha \ - y_i^e \cdot \cos\alpha) + y_0 \end{cases}, i = \overline{1,n} \qquad (4)$$

Здесь a и b - коэффициенты линейного сжатия документа по горизонтали и вертикали соответственно. Так же, как и ранее, считаем угол α малым, поэтому заменив значение косинуса и синуса линейным приближением и раскрыв скобки в формуле (4) получаем:

$$\begin{cases} x_i^t = a \cdot x_i^e - a \cdot y_i^e \cdot \alpha \ - x_0 \\ y_i^t = b \cdot x_i^e \cdot \alpha \ - b \cdot y_i^e + y_0 \end{cases}, i = \overline{1,n} \qquad (5)$$

На практике оказывается проще использовать общее аффинное преобразование вместо преобразования (5):

$$\begin{cases} x_i^t = a \cdot x_i^e + b \cdot y_i^e + x_0 \\ y_i^t = c \cdot x_i^e + d \cdot y_i^e + y_0 \end{cases}, i = \overline{1,n} \qquad (6)$$

Задача определения области поиска очередного элемента решается аналогично предыдущему случаю.

3.2.3 Другие модели. Существует множество различных документов, изображение которых не укладываются в рассмотренные выше геометрические модели. Примером может служить изображение паспорта (см. Рис. 3). Статические элементы каждой страницы расположены друг относительно друга жестко (т к. паспорта печатаются на строго определенных бланках), однако из-за неточности переплета возможно нежесткое взаиморасположение отдельных страниц (на Рис. 3 верхняя страницы повернута относительно нижней страницы на некоторый угол). Для таких документов уместно использовать геометрическую модель, учитывающую общее смещение и наклон всего документа на изображении (как и ранее из-за неточности расположения документа на стекле сканера), а также взаимное расположение отдельных страниц паспорта. Благодаря выделению геометрической модели изображения документа в отдельную независимую подзадачу, можно добавлять новые геометрические модели, не изменяя остальные части алгоритма.

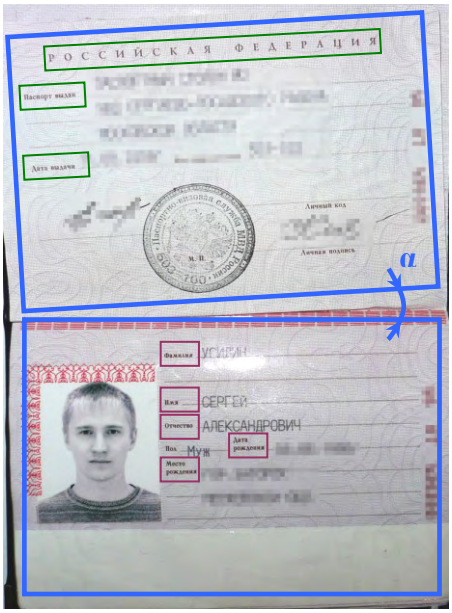


Рис. 3. Скан паспорта РФ. Верхняя страница повернута относительно нижней

3.3. Поиск элементов

Задача поиска элементов сводится к задаче поиска областей изображения, задающих эталонные элементы, на исследуемом изображении. В качестве детектора разумно использовать метод корреляции изображений [12]. Изложим основные моменты. Пусть есть изображение $f(x,y)$ с размерами $M \times N$. Будем искать на нем эталонный элемент $w(x,y)$ с размерами $J \times K$. Запишем функцию корреляции:

$$c(x,y) = \frac{M[f_{xy} \cdot w] - M[f_{xy}] \cdot M[w]}{\sqrt{D[f_{xy}] \cdot D[w]}} =$$

$$= \frac{D[f_{xy}] + D[w] + M[f_{xy}] - M[w])^2}{2 \cdot \sqrt{D[f_{xy}] \cdot D[w]}} - \qquad (7)$$

$$- \frac{M[(f_{xy} - v)^2]}{2 \cdot \sqrt{D[f_{xy}] \cdot D[w]}}$$

В формуле (7) $M[$, $D[$ – среднее и дисперсия изображения, а f_{xy} – прямоугольная область на f с центром в (x,y) и размером с эталонный элемент w. Положение максимумов функции $c(x,y)$ указывает те точки, где область $w(x,y)$ лучше всего согласуется с изображением $f(x,y)$.

Вычисление корреляционной функции достаточно трудоемкая задача из-за необходимости пересчета слагаемого $M[(f_{xy} - v)^2]$ для каждой точки изображения. Оказывается, на практике с большой точностью квадратичную функцию $(f_{xy} - v)^2$ можно

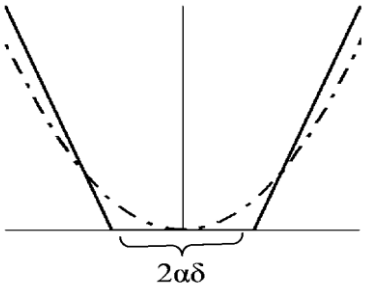


Рис. 4. Аппроксимация квадратичной функции кусочно-линейной

аппроксимировать кусочно-линейной $\max(0,|f_{xy}-w|-$, как показано на Рис. 4. Это позволяет приблизить функцию корреляции $c(x,y)$ величиной, которая быстро вычисляется на процессорах семейства x86 с технологией MMX.

Другим приемом, ускоряющим работу коррелятора, является использование кратномасштабного подхода [12]. Поиск элемента начинается на верхнем уровне пирамиды, и постепенно спускаясь на нижеследующие уровни, уточняется истинное расположение элемента. Стоит также отметить, что штраф за неточность расположения элемента возрастает по мере спуска по пирамиде, что помогает заранее (не спускаясь до самого нижнего уровня) отбросить ошибочные варианты расположения элемента.

4. Идентификация документов

Простейший процесс идентификации документа представляет собой последовательное совмещение исследуемого изображения с каждым эталонным изображением. За тип исследуемого документа принимается тип эталона, который наилучшим образом совместился с исследуемым изображением. Сложность такого алгоритма растет линейно с ростом числа эталонов.

Архитектура данного алгоритма позволяет решать задачу идентификации сублинейно от числа эталонов, то есть, не производя до конца совмещение исследуемого изображения с каждым эталоном. Действительно, алгоритм совмещения изображений устроен таким образом, что в процессе работы штраф наилучшей позиции на текущем шаге не меньше штрафа наилучшей позиции на предыдущем шаге. Опираясь на это свойство алгоритма совмещения изображений, опишем пошаговую работу алгоритма одновременного конкурентного совмещения исследуемого изображения с набором эталонов.

4.1. Алгоритм конкурентного совмещения изображения с несколькими эталонами

Позицией совмещения со штрафом не превосходящим назовем позицию перебора алгоритма совмещения изображений с максимальным количеством совмещенных элементов, штраф которой не превосходит . За начальную позицию совмещения примем соответствующие начальные позиции перебора алгоритма совмещения изображений.

Шаг 1. Положим в кучу начальные позиции совмещения исследуемого изображения с каждым эталоном.

Шаг 2. Выберем из кучи наилучшую позицию совмещения. Наилучшей считается такая позиция совмещения, штраф соответствующей позиции перебора которой минимален.

Шаг 3. Предадим наилучшей позиции совмещения приращение по штрафу на величину , то есть заменим текущую позицию совмещения, штраф которой не превосходит , на позицию совмещения, штраф которой не превосходит . Новую позицию совмещения положим в кучу взамен текущей.

Шаг 4. Проверим статус позиций совмещения. Если существует позиция совмещения, в которой изображения совместились, то завершаем процесс идентификации, принимаем за тип исследуемого изображения тип соответствующего эталона. Если во всех позициях совмещения изображения несовместимы, то завершаем процесс идентификации, тип исследуемого изображения объявляем неопределенным. В противном случае переходим на Шаг 2.

Альтернативный вариант алгоритма подразумевает использование расширенного алгоритма перебора 3.1, в котором общая куча позиций перебора используется для поиска на нескольких эталонах.

5. Заключение

В работе предложен новый алгоритм идентификации типа документа, основанный на методе совмещения изображений. Архитектура алгоритма позволяет определить тип документа, не содержащего предопределенные графические примитивы. Благодаря разделению геометрической модели трансформаций и процесса детектирования элементов оказалось возможным одновременная поддержка различных геометрических моделей документов. Так как совмещение изображений осуществляется на основании совмещения статических элементов, заполнение полей документа не оказывает влияния на определение типа документа. А использование корреляции в качестве детектора делает алгоритм устойчивым к различным параметрам сканирования.

На основании предложенного алгоритма разработана прикладная программа, которая была испытана для распознавания типа паспорта. В качестве возможных образцов были предложены внутренние и заграничные паспорта Российской Федерации, Украины и Казахстана с указанными характерными идентификационными зонами. В качестве тестовых изображений был предложен набор из 300 различных паспортов указанных государств, полученный с помощью различных сканеров. Алгоритм надежно определял и обрабатывал случайные смещения документов на

изображении, связанные с неаккуратным сканированием, а также показал высокую устойчивость к различным искажениям и шуму.

Программа внедрена в промышленную систему распознавания документов Cognitive Forms компании Cognitive Technologies Ltd.

6. Список литературы

[1] В. Л. Арлазаров, О. А. Славин, "Алгоритмы распознавания и технологии ввода текстов в ЭВМ", *Информационные технологии и вычислительные системы №1*, 1996, стр. 48-54.

[2] Oivind Due Trier, Torfinn Taxt, "Evaluation of Binarization Methods for Document Images", *IEEE Trans. Pattern Analysis and Machine Intelligence*, 1995, стр. 312-315.

[3] В. В. Постников, "Автоматическая идентификация и распознавание структурированных документов", *Автореф. дис. ... канд. тех. наук*, Москва, 2001.

[4] В. В. Постников, "Идентификация и распознавание документов заранее известной структуры", *Сборник трудов 6-ой Международной конференции "Распознавание образов и анализ изображений: новые информационные технологии"*, Великий Новогород, 2002.

[5] Kuo-Chin Fan, Yuan-Kai Wang, Mei-Lin Chang, "Form Document Identification Using Line Structure Based Features", *Proc. ICDAR,* 2001, стр. 704-708.

[6] Lin Yu Tseng, Rung-Ching Chen, "The Recognition of Form Documents Based on Three Types of Line Segments", *Proc. ICDAR*, 1997, стр. 71-75.

[7] Hanchuan Peng, Fuhui Long, Wan-Chi Siu, Zheru Chi, David Dagan Feng, "Document image matching based on component blocks", *Proc. ICIP*, 2000, стр. 601-604.

[8] Hanchuan Peng, Fuhui Long, Zheru Chi, "Document Image Recognition Based on Template Matching of Component Block Projections", *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2003, стр. 1188-1192.

[9] Ming Ye , M. Bern, D. Goldberg, "Document image matching and annotation lifting", *Proc. ICDAR*, 2001, стр. 753-760.

[10] Ю. В. Титов, "Об искажении символов при сканировании", *Труды ИСА РАН "Системный подход к управлению информацией"*, 2006, стр. 260-288.

[11] Т. Кормен., Ч. Лейзерсон., Р. Ривест, *Алгоритмы: построение и анализ*, Москва, 1990.

[12] Р. Гонсалес, Р. Вудс, *Цифровая обработка изображений*, Москва, 2006.

Использование методов анализа неподвижных изображений для поиска событий в видеопоследовательностях

И.Б. Кондратенков, Л.И. Рубанов
Лаборатория методов анализа и цифровой обработки изображений
Институт Проблем Передачи Информации им. А.А. Харкевича (РАН), Москва, Россия
ivankond@mail.ru

Аннотация

В работе рассмотрены возможности использования методов анализа неподвижных изображений в пространственной и частотной областях для распознавания интересующих событий в видеопоследовательности. Экспериментальная часть базировалась на обработке и анализе цифровых изображений на ПК с последующей численной и визуальной оценкой результатов при различных уровнях качества видеопоследовательности. Для тестирования использовались видеофрагменты, взятые с различными видеорядами, из которых в дальнейшем для исследования выделялось определенное подмножество ключевых кадров алгоритмов видеокомпрессии, которые обычно применяются в системах охранного телевидения. Полученные в результате обработки этих опорных кадров векторы признаков в дальнейшем предполагается обрабатывать с применением методов временного анализа. Для программной реализации различных методов использовались системы математического моделирования Mathcad и MATLAB.

1. Введение

Обработка видеоизображений, представленных последовательностью оцифрованных кадров, в настоящее время широко используется в различных сферах человеческой деятельности. Одной из наиболее сложных и актуальных задач обработки видеоизображений является проблема выделения и распознавания интересующих событий в условиях действия различного рода помех и возмущений.

Построение описания видеоизображения на основе его представления с использованием признаков — едва ли не самая сложная задача при построении любой системы распознавания. При этом, если в рамках некоторых математических моделей удалось формализовать процесс классификации, то процесс выбора признаков до сих пор остается процедурой эвристической, зависящей как от предметной области, так и от разработчика. В то же время определенный опыт, накопленный за годы использования средств распознавания образов и обработки изображений для решения практических задач, позволяет выделить ряд основных групп признаков, которые успешно используются для описания и распознавания изображений.

Постановка задачи: исследование методов анализа неподвижных изображений для выделения признаков. Для выделения признаков в видеопоследовательностях предлагается использовать методы в пространственной (с использованием модуля градиента) и частотной (с использованием фазовой составляющей двумерного Фурье спектра) областях.

2. Анализ в пространственной области. Модуль градиента

Исследованиями установлено, что с точки зрения распознавания и анализа объектов на изображении наиболее информативными являются не значения яркостей объектов, а характеристики их границ – контуров. Другими словами, основная информация заключена не в яркости отдельных областей, а в их очертаниях.

Как правило, граница предмета на фотографии отображается перепадом яркости между двумя сравнительно однотонными областями. Здесь могут быть и исключения, т к. перепад яркости может быть вызван также текстурой предмета, тенями, бликами, перепадами освещенности, и т п.

Будем называть контуром изображения совокупность его пикселов, в окрестности которых наблюдается скачкообразное изменение функции яркости. Так как при цифровой обработке изображение представлено как функция целочисленных аргументов, то контуры представляются линиями шириной, как минимум, в один пиксел.

Если исходное изображение, кроме областей постоянной яркости, содержит участки с плавно меняющейся яркостью, то непрерывность контурных линий не гарантируется. С другой стороны, если на “кусочно-постоянном” изображении присутствует шум, то могут быть обнаружены “лишние” контуры в точках, которые не являются границами областей.

Мерой изменения некоторой величины является её производная. На изображении величина яркости изменяется в пространстве. Пространственная производная – это градиент, который кроме величины имеет ещё и направление, т.е. представляет собой вектор.

Градиентом двумерной функции *f*(*x*, *y*) называется следующий вектор[1]:

$$\nabla \mathbf{f} = \begin{bmatrix} G_x \\ G_y \end{bmatrix} = \begin{bmatrix} \dfrac{\partial f}{\partial x} \\ \dfrac{\partial f}{\partial y} \end{bmatrix} \qquad (1)$$

Модуль (длина) этого вектора равен

$$\nabla f = |\nabla \mathbf{f}| = \left[G_x^2 + G_y^2 \right]^{1/2} = \left[\left(\frac{\partial f}{\partial x} \right)^2 + \left(\frac{\partial f}{\partial y} \right)^2 \right]^{1/2} \qquad (2)$$

Часто эту величину можно аппроксимировать с помощью суммы абсолютных величин

$$\nabla f \approx |G_x| + |G_y| \qquad (3)$$

Такое приближение позволяет избежать возведения в квадрат и извлечения квадратного корня, но при этом оно ведет себя как производная (т.е. оно равно нулю на областях с постоянным цветом (или яркостью) пикселов, и его величина пропорциональна степени изменения цвета на неоднородных областях).

Основное свойство вектора градиента состоит в том, что он направлен в сторону максимального роста *f* в точке с координатами (*x*, *y*). Угол этого направления равен

$$\alpha(x, y) = \operatorname{arctg}\left(\frac{G_x}{G_y} \right) \qquad (4)$$

Производные принято аппроксимировать разностями значений пикселов в малой окрестности изображения. На Рисунке. 1, *а*) показана окрестность размерами 3x3, где буквы *z* обозначают значения пикселов [3, 6].

z_1	z_2	z_3
z_4	z_5	z_6
z_7	z_8	z_9

(а)

−1	−2	−1
0	0	0
1	2	1

(б)

−1	0	1
−2	0	2
−1	0	1

(в)

Рисунок 1: а – окрестность (маска). б, в – маски Собела, используемые при вычислении частных производных, соответственно, по направлению *x* (вертикальному) и *y* (горизонтальному) относительно центральной точки этой окрестности

Приближение частной производной в (вертикальном) направлении *x* по отношению к центральной точке окрестности (т.е. z_5) определяется разностной формулой

$$G_x = (z_7 + 2z_8 + z_9) - (z_1 + 2z_2 + z_3) \qquad (5)$$

Аналогично, частная производная по направлению *y* приближается разностным выражением

$$G_y = (z_3 + 2z_6 + z_9) - (z_1 + 2z_4 + z_7) \qquad (6)$$

Обе эти величины вычисляются во всех точках изображения с помощью свертки изображения с двумя масками, приведенными на Рисунке 1, *б*) и *в*). После этого приближения градиентное изображение получается сложением абсолютных величин двух фильтрованных изображений.

(а) (б)

Рисунок 2: Выделение контуров с помощью модуля градиента. а – исходное изображение, б-градиент

Далее, градиентное изображение подвергается бинаризации по половинному порогу Рисунок 3, после чего двоичный вектор признаков строится путем обхода изображения из центра по прямоугольной спирали с увеличивающимся шагом; длина такого вектора составила в экспериментах от 80 до 240 байтов.

Градиентная обработка подчеркивает мелкие дефекты, которые нелегко заметить на полутоновом изображении. Способность подчеркнуть мелкие неоднородности на ровном сером поле есть еще одна важная характеристика градиента. Оборотной стороной этого эффекта является сильная чувствительность градиентных методов к уровню шума изображения, что заставляет

предъявлять повышенные требования к качеству сенсора и уровню освещенности объекта.

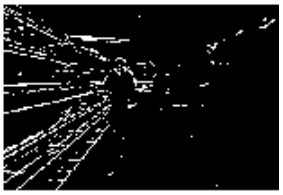


Рисунок 3: Бинарное изображение, полученное из градиентного изображения

3. Анализ в частотной области. Двумерное дискретное преобразование Фурье

Методы обработки в *частотной области* основываются на модификации сигнала, формируемого путем применения к изображению преобразования Фурье[5].

Пусть $f(x, y)$ при $x = 0, 1, 2,\ldots, M-1$ и $y = 0, 1, 2,\ldots, N-1$ обозначает изображение MxN. Двумерное *дискретное преобразование Фурье* (DFT, Discrete Fourier Transform) изображения f, которое обозначается $F(u, v)$, задается равенством

$$F(u,v)=\sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x,y)e^{-j2\pi(ux/M+vy/N)} \qquad (7)$$

при $u = 0, 1, 2,\ldots, M-1$ и $v = 0, 1, 2,\ldots, N-1$ [2]. Частотной областью называется координатная система, задающая аргументы $F(u, v)$ частотными переменными u и v. Здесь можно обнаружить аналогию с заданием аргументов $f(x, y)$ пространственными переменными x и y. Прямоугольную область размера MxN, задаваемую при $u = 0, 1, 2,\ldots, M-1$ и $v = 0, 1, 2,\ldots, N-1$, принято называть *частотным прямоугольником*. Видно, что частотный прямоугольник имеет те же размеры, что и исходное изображение.

Основной метод визуального анализа этого преобразования заключается в вычислении его *спектра* (т.е. абсолютной величины $F(u, v)$) и его отображения на дисплее. Для повышения визуальной информативности при этом обычно приходится прибегать к сжатию большого динамического диапазона Фурье-спектра при его преобразовании в доступный диапазон яркостей, с помощью некоторой функции с замедленным ростом, например логарифмической или корня небольшой степени. Пусть $R(u, v)$ и $I(u, v)$ обозначают вещественную и мнимую компоненты $F(u, v)$, тогда спектр Фурье задается выражением

$$|F(u,v)|=\left[R^2(u,v)+I^2(u,v)\right]^{1/2} \qquad (8)$$

Фазовый угол (угол сдвига фазы) преобразования задается формулой

$$\phi(u,v)=\operatorname{arctg}\left[\frac{I(u,v)}{R(u,v)}\right] \qquad (9)$$

Визуальный анализ спектра упрощается, если переместить начало координат преобразования в центр спектрального прямоугольника. Это можно сделать, умножив $f(x, y)$ на $(-1)^{x+y}$ до вычисления преобразования Фурье. Используя свойства экспонент, нетрудно показать, что

$$\mathsf{F}\left[f(x,y)(-1)^{x+y}\right]=F(u-M/2,v-N/2) \qquad (10)$$

где $\mathsf{F}[\ ]$ обозначает преобразование Фурье своего аргумента. Это равенство означает, что начало координат для Фурье-преобразования функции $f(x, y)$ $(-1)^{x+y}$ находится в точке с координатами $u=M/2$, $v=N/2$ [4].

(а) (б)

Рисунок 4: Вычисление и визуализация двумерного ДПФ. а – простое изображение, б - центрированный спектр Фурье, улучшенный функцией log для большей наглядности

Для спектрального изображения вектор признаков строится путем выбора последовательных отсчетов от центра спектрального прямоугольника в направлении главной диагонали первой четверти изображения, т.е. для фиксированного значения фазы; число таких отсчетов выбиралось в интервале от 32 до 80.

К достоинствам Фурье-признаков относятся существование быстрых алгоритмов их формирования, полнота описания изображения (преобразование обратимо), помехоустойчивость. К недостаткам — отсутствие инвариантности к целому ряду геометрических преобразований, в том числе к преобразованиям подобия.

4. Результаты

Предложенные методы были реализованы в среде Matlab 7.6.0. Для тестирования были взяты видеофрагменты с различными видеорядами, из которых выделялось подмножество ключевых стоп-кадров. Данные стоп-кадры обрабатывались с помощью методов, описанных выше, в результате чего формировались вектора признаков. На рис. 2-4 представлены некоторые результаты тестирования. Предварительный анализ полученной

системы признаков показывает, с одной стороны, отсутствие корреляции между ними, т.е. такие признаки реально являются независимыми; с другой стороны, наблюдается заметная корреляция векторов признаков с изменением видеоряда. Это позволяет рассчитывать, что размерности векторов признаков удастся снизить до разумных величин, не потеряв при этом их дискриминантных свойств.

5. Заключение

В данной работе предложены методы анализа неподвижных изображений, позволяющие выделять признаки в видеопоследовательностях. В дальнейшем планируется обрабатывать полученные вектора признаков с применением методов временного анализа.

6. Библиография

[1] Гонсалес Р., Вудс Р. “Цифровая обработка изображений”, 2006, страниц 1072.

[2] Rafael C. Gonzalez, Richard E. Woods, Steven L. Eddins “Digital Image Processing using MATLAB”, 2004, 12 chapters, 624 pages.

[3] Дьяконов В. П. MATLAB 6.5 SP1/7/7 SP1+Simulink 5/6. Работа с изображениями и видеопотоками. – М.: СОЛОН-Пресс. 2005: – 400 стр.

[4] Прэтт У. Цифровая обработка изображений. Кн.2./ М.: Мир, 1982. – 784 с.

[5] Gwilym M. Jenkins, Donald G. Watts “Spectral Analysis and its applications”, 1969, 525 pages.

[6] Ziou, D. and Tabbone, S.: Edge Detection Techniques An Overview, International Journal of Pattern Recognition and Image Analysis, 8(4):537--559, 1998.

Обработка данных в системах создания виртуального окружения

А.В. Торчигин
Институт проблем информатики РАН
v_torchigin@mail.ru

Аннотация

Рассматривается подход к созданию виртуального окружения для многих зрителей, при котором зрители наблюдают отражения множества модулируемых по яркости светодиодов в колеблющихся зеркалах. Анализируются требования к электронным средствам, обеспечивающим требуемые изменения яркости светодиодов во времени.

1. Введение

Развитие технических средств, в первую очередь, высокопроизводительных вычислительных систем достигло в настоящее время такого уровня, что появляются системы, позволяющие перенести зрителей в несуществующий в природе виртуальный мир, созданный воображением разработчиков этих систем. Известно, что человек воспринимает объемную картину реального мира благодаря бинокулярности его зрения. Левый и правый глаз видят немного различающиеся картинки, и это помогает нам оценивать расстояния до ближайших предметов, определять взаимное расположение объектов. Описание бинокулярного зрения человека встречается еще в работах Леонардо да Винчи, но без какого-либо практического применения. Первооткрывателем стереоскопии считается архитектор Чарльз Уэтстон, который 21 июня 1833 года на собрании Королевского общества в Лондоне прочитал лекцию о своих исследованиях в области стереоскопии и продемонстрировал первый стереоскоп. Фотография к тому времени еще не была изобретена и стереопары были рисованные.

В настоящее время в России наиболее совершенной является система, созданная в ИММВС РАН коллективом под руководством М.И. Михайлюка для тренировки космонавтов, в частности, для получения опыта при стыковке космического корабля с космической станцией. Эта система использует два проектора, которые проектируют изображения соответственно для левого и правого глаз на экран. Поляризация света, излучаемого одним проектором, ортогональна поляризации света, излучаемого другим проектором. В качестве экрана используется материал, который сохраняет поляризацию отраженного света. Наблюдатель смотрит на экран через очки, которые пропускают в левый глаз свет с одной поляризацией, а в правый – с другой. Таким образом, каждый глаз видит сформированную для него картину.

Основной недостаток подобных систем состоит в том, что картина, которую видит наблюдатель, ограничена размером отражающего экрана. Большие экраны требуют больших помещений и мощных проекторов. Кроме того, должны использоваться дорогостоящие экраны из материала, сохраняющего поляризацию света при отражении от экрана.

В последнее время получают распространение так называемые виртуальные шлемы, являющиеся усовершенствованием первоначальных стереоскопов. Вместо статических изображений виртуальные шлемы дают возможность разглядывать динамические изображения, предназначенные для левого и правого глаз. Эти изображения формируются с помощью жидкокристаллических матриц [1].

Важной особенностью таких систем является наличие так называемой системы виртуальной ориентации, которая отслеживает движение головы и в соответствии с ним корректирует изображение на экранах. При повороте головы в одну сторону панорамное изображение на ЖК матрицах прокручивается в противоположном направлении, благодаря чему возникает иллюзия стабильности и неповторимое ощущение реальности окружающей картины (так, как это происходит в реальной жизни).

Однако система виртуальной реальности представляет собой сложный и дорогостоящий набор аппаратно-программных средств, который включает в себя датчики угловых положений и координат головы наблюдателя. По поступающий от этих датчиков информации определяется тот фрагмент общей картины, который должен

видеть наблюдатель при заданном положении головы. Весьма существенным недостатком такой системы является то обстоятельство, что для каждого наблюдателя должен быть обеспечен свой собственный полный комплект аппаратных средств. Кроме того, в реальном мире зритель может видеть различные фрагменты этого мира и без поворота головы, просто путем изменения угла зрения зрачков. Это свойство отсутствует в виртуальных шлемах.

Как известно, в настоящее происходит очередная революция в светотехнике. На смену лампам накаливания и газоразрядным лампам приходят светодиоды, приборы, являющиеся продуктом высоких технологий [2,3,4]. Наряду с многими весьма ценными свойствами, такими как высокий кпд, высокая светоотдача, чрезвычайно высокая надежность, способность переносить ударные перегрузки, светодиоды имеют весьма ценные свойства для формирования изображений. В отличие от существующих источников света яркость светодиодов может модулироваться с чрезвычайно высокой частотой. Широко используемые в настоящее время предназначенные для модуляции светодиодов интегральные схемы позволяют модулировать яркость с частотой до 25 МГц [5]. Кроме того, стоимость отдельного светодиода уже в настоящее время на порядки ниже стоимости лампы накаливания. Это позволяет использовать в системе десятки и сотни тысяч светодиодов.

Появившиеся в последнее время красные, зеленые и синие яркие светодиоды позволяют осуществить совершенно новый подход к формированию объемных цветных изображений для больших зрительских аудиторий, при котором естественным образом реализуется система виртуальной ориентации. При этом отпадает необходимость в больших экранах и мощных проекторах, обычно используемых для погружения в виртуальную реальной большие зрительские аудитории. Отпадает необходимость в датчиках угловых положений и координат головы наблюдателя и в соответствующем дорогостоящем наборе аппаратно-программных средств, используемых в виртуальных шлемах. Более того, при таком подходе зритель без всякой задержки может видеть при повороте зрачков другой фрагмент формируемого изображения.

2. Принцип работы

Рассмотрим принцип формирования изображений при этом подходе. Пусть имеется ряд светодиодов, расположенных вдоль вертикальной прямой линии на одинаковом расстоянии друг от друга (вертикальная линейка светодиодов). Пусть имеется плоское зеркало, которое может вращаться вокруг вертикальной оси, проходящей через отражающую поверхность зеркала. При вращении зеркала, зритель, наблюдающий за изображением светодиодов в зеркале, будет видеть, что каждый светодиод перемещается по окружности.

Чтобы убедиться в этом, построим

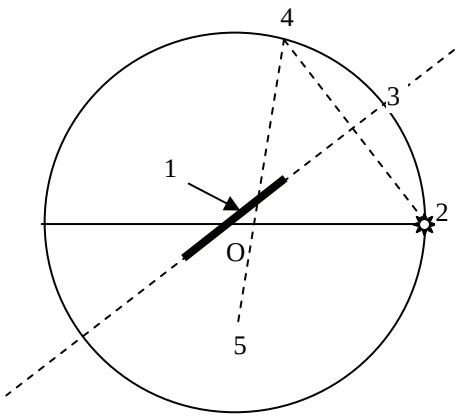


Рис. 1. Траектория 3 перемещения мнимого изображения 4 источника 2 при вращении зеркала 1

изображение светодиода при различных угловых положениях вращающегося зеркала, показанного жирной линией на рисунке 1. Имея в виду, что изображение светодиода в зеркале расположено в точке, симметричной светодиоду относительно плоскости зеркала, получим, что изображение светодиода 2 при положении зеркала 1 окажется в точке 4. При вращении зеркала 1 изображение светодиода 2 описывает окружность 3 с центром О на оси вращающегося зеркала 1. Радиус этой окружности равен расстоянию от оси вращающегося зеркала 1 до источника 2. Однако наблюдатель, расположенный в некоторой точке 5, может не увидеть всю окружность по нескольким причинам.

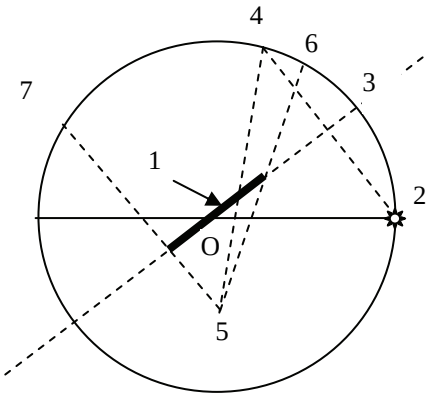


Рис. 2. Дуга, которую видит наблюдатель 5 во вращающемся зеркале 1 ограниченных размеров

Во-первых, угол обзора ограничен строением глаз человека, который, не поворачивая головы, может видеть в секторе около 90 градусов.

Во-вторых, угол обзора изображения ограничен размерами зеркала и расстоянием от оси вращения зеркала до наблюдателя. Для того чтобы изображение источника было видно наблюдателю, необходимо, чтобы прямая линия, соединяющая изображение источника в точке 4 с точкой 5, где находится наблюдатель, пересекала зеркало 1.

На рисунке 2 показан в качестве примера для конкретного положения наблюдателя и размера зеркала фрагмент окружности между точками 6 и 7, по которому перемещается изображение светодиода. Только этот фрагмент виден наблюдателю, расположенному в точке 5. Чем ближе наблюдатель расположен к вращающемуся зеркалу и чем больше горизонтальный размер 2R вращающегося зеркала, тем больший фрагмент окружности может видеть наблюдатель.

На рисунках показана лишь одна окружность, создаваемая одним источником. В том случае, если источники расположены на прямой, параллельной оси вращения, то изображение каждого источника при вращении зеркала будет описывать идентичную окружность, расположенную в плоскости, параллельной плоскости рассмотренной окружности. Набор таких окружностей формирует изображение, подобно тому, как набор прямых линий в виде строк формирует изображение на экране телевизора [7].

Использование зеркал, совершающих вращательные колебания вокруг оси вращения (колеблющихся зеркал), имеет несколько преимуществ перед использованием вращающихся зеркал. Во-первых, при использовании вращающихся зеркал картина формируется только в течение небольшого времени, когда положение зеркал таково, что лучи света от модулируемых источников попадают в глаза наблюдателя. Обычно это время менее 10% от периода вращения. Это приводит к тому, что большую часть времени свет от модулируемых источников не может попадать в глаза наблюдателя, что снижает максимальную яркость формируемого изображения. Во-вторых, если в помещении находятся другие источники света, например, окна, лампы освещения, то при определенном положении вращающихся зеркал свет от них попадает в глаза наблюдателя. Это приводит к тому, что изображение формируется на фоне неконтролируемого светлого фона. В-третьих, колеблющиеся зеркала можно располагать ближе к зрачкам наблюдателя, чем вращающиеся. Это позволяет увеличить максимальный телесный угол, в котором может формироваться изображение.

Формируемое изображение имеет несколько существенных отличий от изображений, формируемых на экране телевизора. Во-первых, видимая наблюдателем картина зависит от его положение относительно экрана. На рисунке 3 приведены две фотографии, сделанные фотоаппаратом, находящемся на разных расстоянии от колеблющегося зеркала. При приближении фотоаппарата к зеркалу виден больший фрагмент логотипа. В отличие от привычных изображений на экране телевизора, где изображение расположено в плоскости

Рис. 3. Фотографии одного и того же изображения, формируемого в прикрепленном к стене колеблющемся зеркале линейкой модулируемых по яркости светодиодов. Правая фотография сделана фотоаппаратом, находящемся на большем расстоянии от зеркала, чем левая

экрана, изображение в колеблющемся зеркале расположено в его глубине на расстоянии, равном расстоянию от зеркала до линейки светодиодов. Колеблющееся зеркало воспринимается наблюдателем в виде отверстия, через которое он может рассматривать виртуальный мир. Свойства картины виртуального мира сходны со свойствами картины более знакомого зеркального мира, который может наблюдаться в обычном неподвижном зеркале ограниченных размеров. Такое зеркало представляется наблюдателю в виде окна, через которое он может рассматривать зеркальное отражение реального мира. Он может смотреть на зеркало под разными углами и видеть различные фрагменты реального мира. При этом сохраняется объемность изображений. В том случае, если телесный угол, под которым наблюдатель мог бы разглядывать изображение, больше максимального телесного угла, под которым человек может видеть реальный мир, наблюдатель путем поворота головы или зрачков глаз может разглядывать различные фрагменты реального мира. Эти свойства сохраняются и при разглядывании зеркального мира. Свойства картины виртуального мира, формируемого в колеблющемся зеркале, совпадают со свойствами картины зеркального мира. Исключением является объемность изображения. Для получения объемности необходимо использовать несколько линеек светодиодов, расположенных на разных расстояниях от колеблющегося зеркала, и соответствующие программные средства, позволяющие формировать согласованные изображения, находящиеся на переднем, заднем и нескольких средних планах. В этом случае наблюдатель, без каких бы ни было очков получает возможность наблюдать в колеблющемся зеркале псевдо объемное изображение.

Что касается необходимых аппаратных и программных средств, управляющих яркостью каждого светодиода, то эти средства во многом схожи со средствами, используемыми в светодиодном табло типа бегущая строка. Такое табло также можно рассматривать как набор линеек из светодиодов. Количество таких линеек равно количеству светодиодов в строке. Точно так же аппаратные средства должны поочередно включать светодиоды в каждой вертикальной линейке табло. Отличие лишь в том, что в рассматриваемом случае включаются одни и те же светодиоды в одной линейке, а в табло последовательно включаются светодиоды в каждой из линеек. Таким образом, проблема управления яркостью светодиодов в настоящее время полностью решена в светодиодных табло [6]. Необходимо лишь обеспечить, чтобы очередной цикл показа изображения в виде набора светящихся линеек светодиодов начинался при одном и том же положении зеркала. В рассматриваемом подходе количество светодиодов сокращается в сотни раз. При достаточно большом размере зеркала возникают технические проблемы с обеспечением необходимой амплитуды колебаний. Эти проблемы могут быть устранены при использовании нескольких вращающихся треугольных призм с зеркальными боковыми гранями. Подобный прием используется в рекламных щитах, обеспечивающих поочередный показ трех изображений при повороте призм на 120 градусов.

Для того чтобы обеспечить максимальный угол обзора изображения, ось колеблющегося зеркала можно закрепить на платформе, жестко связанной с головой наблюдателя, например, на шлеме или очках наблюдателя. Телесный угол, в котором формируется изображение виртуального мира, определяется отношением s/r, где s – размер зеркала в горизонтальном направлении, r – расстояние от центра зеркала до зрачка наблюдателя. Это означает, что при приближении зеркала к зрачку наблюдателя, поперечные размеры зеркала могут уменьшаться. При этом телесный угол формируемого изображения может превосходить телесный угол, в котором глаза наблюдателя могут видеть изображение. Такое устройство может быть выполнено в виде очков, в которых линзы заменены небольшими колеблющимися зеркалами. При повороте зрачков глаз будет виден другой фрагмент круговой панорамы. Если наблюдатель не может наблюдать полное изображение из-за того, что его угол зрения в горизонтальной плоскости ограничен, он может повернуть голову вокруг вертикальной оси и его взору откроется невидимая прежде часть формируемого изображения. Это связано с тем, что при повороте головы на некоторый угол φ на этот же угол поворачивается колеблющееся зеркало. Как следует из рисунка 2, на угол φ сдвигается также видимая наблюдателем дуга, по которой перемещается изображение источника света. Аналогичная ситуация имеет место и при разглядывании реального мира, который предстает перед наблюдателем в виде объемного изображения с углом обзора в 360 градусов.

При повороте головы вверх или вниз вокруг оси параллельной прямой, проходящей через уши наблюдателя, он может разглядывать верхнюю или нижнюю части формируемого изображения. Однако при наклоне головы влево или вправо изображение будет также наклоняться. Это связано с тем, что при этом изменяется направление оси, вокруг которой вращаются зеркала. Если обеспечить параллельность осей

вращения вертикальной линейке светодиодов при любых наклонах головы, то изображение виртуального мира будет иметь те же свойства, что и изображение зеркального. У наблюдателя создается впечатление, что он со всех сторон окружен виртуальным миром. Как известно, движение зрачков глаз человека происходит непроизвольно и достаточно быстро. Это движение не регистрируется известными системами создания виртуальной реальности, которые обычно реагируют только на поворот головы.

До настоящего времени речь шла об одном наблюдателе. Однако все вышесказанное справедливо для любого другого наблюдателя, который разглядывает светодиоды через свою систему зеркал. Таким образом, в виртуальный мир может быть одновременно погружено сколь угодно много наблюдателей, каждый из которых имеет перед своими глазами свой фрагмент виртуального мира, точно так же, как многие наблюдатели в реальном мире видят каждый свой фрагмент, который им виден из их местоположения. Однако, так же как и в реальном мире, иногда наблюдатели могут мешать друг другу, закрывая собой некоторые фрагменты виртуальной картины.

3 Создание виртуального окружения

Имея в виду рассмотренные выше способы получения стереоизображений, рассмотренный подход просто приспособить для этой цели. Схема получения объемного изображения показана на рисунке .4. Здесь источники 1 формируют изображение для левого глаза 3, в первый интервал времени, когда свет от источников попадает в левый глаз, отражаясь от зеркала 8, совершающего вращательные колебания. При этом свет от тех же источников не попадает в правый глаз, так как свет от зеркала 9 отражается в другую сторону (Рис.4а) Во второй момент времени через полпериода колебаний зеркал зеркала принимают положение, показанное на рисунке 4б. В этом случае ситуация изменяется на противоположную. Свет от модулируемых источников, отражаясь от колеблющегося зеркала 9, попадает в правый глаз и не попадает в левый при отражении от зеркала 8. Таким образом, каждый глаз видит формируемую для него картину, а у зрителя создается впечатление, что он видит стереоизображение.

Отметим, что при рассматриваемом подходе при желании могут быть использованы все те средства, которые используются в существующих системах. Действительно, можно использовать 2 комплекта светодиодов, предназначенных соответственно для левого и правого глаз. Один комплект светодиодов покрыт прозрачной пленкой, пропускающей поляризацию, предназначенную для левого глаза, а другой комплект покрыт пленкой, пропускающей поляризацию, предназначенную для правого глаза. Дополнительно к очкам с колеблющимися зеркалами добавляются обычные поляризационно чувствительные очки. В этом случае интенсивность формируемой картины увеличивается вдвое, так как оба глаза воспринимают одновременно предназначенные для них изображения. Отметим, что в этом случае отпадает необходимость в дорогостоящем экране, сохраняющем поляризацию отраженного света, и мощных двух проекторах, обеспечивающих формирование на экране двух изображений.

Возможны и другие варианты получения необходимых изображений для левого и правого

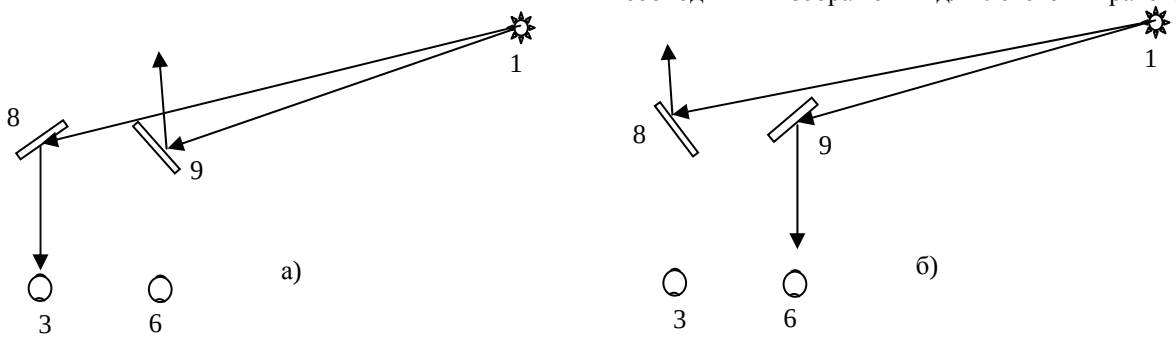


Рис.4. Схема формирования различных изображений для левого и правого глаза от одного комплекта светодиодов. При первом положении зеркал, показанном на рис.а), светодиоды видны левым глазом и не видны правым. При втором положении зеркал, показанном на рис. б), светодиоды видны правым глазом и не видны левым

глаз [7].

Следует обратить внимание, что при изменении местоположения зрителя картина виртуального мира, которую он видит, изменяется. Таким образом, если несколько зрителей снабжены очками с колеблющимися зеркалами, то каждый зритель оказывается погруженным в виртуальное окружение. Это обстоятельство позволяет погружать в виртуальное окружение большие зрительские аудитории. При этом, в отличие от существующих систем, требующих больших помещений, экранов большой площади, мощных дорогостоящих проекторов, и также требующих от зрителей, чтобы те надевали очки, в рассматриваемом подходе виртуальное окружение может быть создано в любом пространстве. Для этого достаточно лишь поместить в любую комнату вертикальную линейку светодиодов размером с высоту комнаты, содержащую около 1000 RGB светодиодов. Из любого места, откуда видна эта линейка, зритель, надев очки, погружается в виртуальное окружение.

Некоторые сомнения у потенциальных потребителей вызывает то обстоятельство, что в очках используется механическое движение в виде колеблющихся зеркал. Такой тип механического движения рассматривался исключительно в целях наглядности. Сплошное плоское зеркало может быть заменено на уже существующие и широко используемые в DLP проекторах DMD (digital mirror device) матрицы [1]. Эти матрицы представляют собой множество плоских микрозеркал, угол поворота каждого из которых задается соответствующим электрическим сигналом. Такие зеркала в состоянии совершать колебательно вращательное движение с частотой в десятки килогерц. В рассматриваемом применении условия работы таких зеркал весьма простые. Все они должны синхронно совершать одинаковые вращательно-колебательные движения с частотой около 100 Гц.

Более того, необходимое изменение угла отражения лучей от зеркала может быть получено и без использования механических перемещений. С этой целью колеблющиеся зеркала могут быть заменены неподвижными дифракционными решетками, период которых управляется электрическими сигналами. Угол отклонения проходящих через дифракционную решетку лучей, зависит от ее периода. При изменении периода решетки изменяется угол отклонения лучей и у наблюдателя создается впечатление, что лучи попадают в его глаза из разных направлений.

Заключение

Из проведенного рассмотрения следует, что в настоящее время развитие технических средств достигло такого уровня, что появляется возможность создания совершенно новых устройств отображения, которые могут использоваться в различных областях применения. В частности, могут быть созданы телевизоры, в которых изображение не ограничено размерами экрана. Подобно тому, как в окно относительно небольшого размера зритель может разглядывать значительный фрагмент реального мира, так и через экран такого телевизора зритель может разглядывать значительный фрагмент виртуального мира. В отличие от громоздких и дорогостоящих широкоформатных телевизоров на основе жидкокристаллических и плазменных панелей, зритель получает возможность разглядывать значительно бо́льшие по размерам изображения, наблюдая в колеблющемся зеркале за расположенной в углу комнаты вертикальной линейкой светодиодов. Колеблющиеся зеркала могут быть вмонтированы в обычные очки.

Большие перспективы использования рассмотренного подхода в компьютерных играх. При этом игрок может быть полностью погружен в виртуальный мир, восприятие которого во многом сходно с восприятием реального.

Список Литературы

Попов С.Н. Аппаратные средства мультимедиа. Видеосистема PC, БХВ-Петербург, Артлит, 2000 г

2. *Кононов В.* Рекордно яркие LED от Nichia http://www.3dnews.ru/news/rekordno_yarkie_led_ot_nichia/

3. *Бурняшев А* Современные мощные светодиоды и их оптика // Современная электроника. – 2006.-№1.- С. 24-27.

4. Мощные светодиоды различных производителей
www.leds.ru
www.lumileds.com/products
www.osram-os.com/goldendragon
www.osram.ru/index.php
www.upec.com
www.laminaceramics.com

5. Драйверы для светодиодов www.mblock.com

6. *Редькин П.* Микроконтроллеры ARM7 семейства LPC2000. –М.: Додэка-XXI, 2007 450 с.

7. *Торчигин А.В.* Способ формирования изображений и устройство для его осуществления. Патент РФ №2328024 от 29 декабря 2003 г.

Расчет количества текстурных кластеров для сегментации изображений на основе Марковских случайных полей

Е.И. Патана

Таганрогский технологический институт Южного федерального университета

Helena.Patana@gmail.com

Аннотация

В работе предлагается метод определения количества текстур, присутствующих на изображении. Информация, получаемая из данного метода необходима для проведения сегментации изображений, т.к. для решения этой задачи нужно заранее знать количество текстур. Сегментация имеет большое прикладное значение. Она может использоваться для анализа аэрофотосъемки, мониторинга движения морских льдов, определения нефтяных пятен на водной поверхности и т.п. Подход основан на расчете коэффициентов линейной зависимости пикселей изображения, которое рассматривается как реализация Марковского случайного поля.

Сегментация является одной из сложных задач анализа текстур по причине отсутствия априорной информации об их количестве и типе, а также о принадлежности конкретной текстуры к некоторой области. Для того чтобы провести сегментацию не обязательно знать, какие специфические текстуры находятся на исходном изображении. Фактически, требуется способ, позволяющий определять, различаются ли текстуры в смежных областях изображения.

Существуют два основных подхода к сегментации изображений. Первый основан на выделении областей изображения, которые имеют однородную текстуру. Данные локальные области объединяются, исходя из подобия текстурных свойств. Изображение, на котором выделены области, имеющие различные текстурные характеристики, предполагается сегментированным. Этот метод имеет преимущество, заключающееся в том, что границы областей всегда замкнуты и поэтому области с отличными текстурами всегда хорошо разделимы. Недостатком является то, что при использовании данного подхода, необходима априорная информация о количестве текстур. Второй подход основан на выделении границ. В этом случае не требуется знать количество присутствующих текстур. Однако использование данного метода не гарантирует замкнутость границ, в результате чего области с различными текстурами могут остаться неразделенными.

В статье для проведения сегментации выбирается подход, основанный на выделении областей, поэтому ставится и решается задача об автоматическом определении количества текстур на исходном изображении.

Метод расчета количества текстур, предложенный в работе, опирается на значения коэффициентов линейной зависимости пикселей изображения, которое рассматривается как реализация Марковского случайного поля.

Введем необходимые определения, которые используются, в частности, при идентификации текстур. Пусть S некоторое конечное множество позиций или множество *узлов*; для каждого узла $s \in S$ через X_s обозначим некоторое конечное *пространство состояний* x_s. Декартово произведение $\mathrm{X} = \prod_{s \in S} \mathrm{X}_s$ назовем пространством конечных *конфигураций* $x = (x_s)_{s \in S}$. Рассмотрим *вероятностную меру* (*распределение*) P на X, т.е. вектор $P = (P(x))_{x \in \mathbf{X}}$ такой, что $P(x) \geq 0$ и $\sum_{x \in \mathbf{X}} P(x) = 1$. Строго положительную вероятностную меру P на X ($P(x) > 0$ для всех $x \in \mathrm{X}$) будем называть *стохастическим* или *случайным полем*. Условные вероятности вида: $P(X_A = x_A \mid X_{S \setminus A} = x_{S \setminus A})$, $A \subset S$, $x_A \in \mathbf{X}_A$, $x_{S \setminus A} \in \mathbf{X}_{S \setminus A}$, являются *локальными характеристиками* [1].

Семейство $N = \{N_s \mid \forall s \in S\}$ подмножеств S называется *системой окрестностей*,

если $s \in N_t$ тогда и только тогда, когда $t \in N_s$. Окрестности различают по порядку. В системе окрестностей первого порядка каждая позиция имеет четыре смежные с ней позиции, второго — восемь. Порядок окрестности однозначно определяет порядок модели МСП [2].

Подмножество C из S будем называть *кликой*, если любые два различных элемента из C являются соседями.

Перейдем к определению Марковского случайного поля. X называется Марковским случайным полем на сетке S по отношению к системе окрестностей N тогда и только тогда, когда выполняются следующие условия:

1. Положительность: $P(x) > 0, \quad \forall x$.

2. Марковость: $P(x_s \mid x_{S-s}) = P(x_s \mid x_{N_s})$.

Марковость описывает локальные характеристики изображения X, т.е. позиция взаимосвязана только с соседними позициями [2].

Вышеуказанное определение Марковского поля дано в терминах локальных характеристик; глобальной характеристикой случайного поля является распределение Гиббса. Вероятностные меры, представимые в виде

$$P(x) = \frac{\exp(-U(x))}{\sum_{z \in \mathbf{X}} \exp(-U(x))} = Z^{-1} \exp(-U(x)),$$

всегда строго положительны и, следовательно (согласно определению), являются случайными полями. Поле P в этом случае называют *гиббсовским случайным полем (ГСП)*, порожденным *энергетической функцией* $U = \sum_{c \in C} V_c$ (сумму потенциальных функций V_c рассчитываем по всевозможным кликам). Знаменатель в представлении (1) называют *статус-суммой*. Иногда для описания энергии распределения Гиббса используется сумма нескольких слагаемых, каждое относится к кликам определенного размера, таким образом:

$$U(x) = \sum_{s \in C_1} V_1(x_s) + \sum_{\{s,s'\} \in C_2} V_2(x_s, x_{s'}) +$$

$$+ \sum_{\{s,s',s''\} \in C_3} V_3(x_s, x_{s'}, x_{s''}) + \ldots$$

Чем меньше значение энергетической функции $U(x)$, тем больше вероятность $P(x)$ появления конфигурации x.

Подходы, основанные на локальных и глобальных характеристиках эквивалентны в соответствии с теоремой Хаммерсли-Клиффорда: X является МСП на множестве S относительно N тогда и только тогда, когда X является ГСП на множестве S [1].

Наибольший интерес из всех моделей представляют *совместные авторегрессионные процессы*, которые удовлетворяют уравнению:

$$X_s = \mu_s + \sum_{t \in N_s} \beta_{st} (X_t - \mu_t) + \varepsilon_s, \qquad (2)$$

где ε - белый шум с дисперсией σ^2. Данные поля имеют плотность, пропорциональную

$$\exp\left(-\frac{1}{2\sigma^2} (x-\mu)^* B^* B (x-\mu) \right),$$

где B - матрица размером $|S| \times |S|$ с единицами на диагонали и с элементами $-\beta_{st}$ вне диагонали (если s и t не являются соседями, то $\beta_{st} = 0$). Матрица ковариаций для авторегрессинного поля имеет вид: $\sigma^2 (B^* B)^{-1}$

Модели изображений, полученные на основе авторегрессионных процессов, обладают высокой точностью, поэтому задачу оценки коэффициентов взаимосвязи в работе предлагается решать именно для них. Полученные результаты могут быть использованы в важнейших приложениях анализа изображений. Например, для анализа аэрофотоснимков или изображений, полученных в радиочастотном диапазоне.

Опишем процесс нахождения оценок коэффициентов взаимосвязи, предлагаемый в статье, необходимых для расчета количества текстурных кластеров. Сеточная система S разбивается на два подмножества позиций: внутреннего множества S_I и граничного множества S_B. Внутреннее множество определяется следующим образом: для любого пикселя из S_I все смежные с ним пиксели принадлежат сеточной системе S, а граничное - $S_B = S - S_I$. На рис. 1 изображено множество S, светло-серым цветом показано множество S_I. Каждая его точка имеет соседнюю точку, принадлежащую множеству S. Окрестно-

сти первого порядка на рис. 1 показаны темно-серым цветом, все их точки принадлежат S.

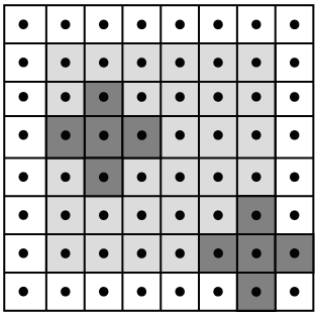


Рис. 1. Графическое представление множества S_I и S_B.

Для оценки коэффициентов рассмотрим уравнение (2) для каждой точки множества S_I, получим СЛАУ: $X\beta = Y$ (3). Матрица системы X представляет собой следующее: в каждой строке находятся значения пикселей, смежных к пикселям некоторой подобласти исходного изображения, т.е. первая строка – соседи для первого пикселя подызображения, вторая строка – для второго и т.д. Столбцом свободных членов Y являются значения пикселей указанной подобласти. На рис. 2 схематично показано расположение пикселей, которые образуют матрицу X и столбец Y.

В данном случае, рассматривая окрестности первого порядка, внутреннее и граничное множества представляют собой следующее:

$$S_I = \{x_7, x_8, x_9, x_{12}, x_{13}, x_{14}\},$$

$$S_B = \{x_1, x_2, x_3, x_4, x_5, x_6, x_{10}, x_{11}, x_{15}, x_{16}, x_{17},$$

$$x_{18,} x_{19}, x_{20}\}.$$

x_1	x_2	x_3	x_4	x_5
x_6	x_7	x_8	x_9	x_{10}
x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
x_{16}	x_{17}	x_{18}	x_{19}	x_{20}

Рис. 2. Пример подызображения для расчета коэффициентов

Матрица системы X и столбец свободных членов Y:

$$X = \begin{pmatrix} x_2 & x_6 & x_8 & x_{12} \\ x_3 & x_7 & x_9 & x_{13} \\ x_4 & x_8 & x_{10} & x_{14} \\ x_7 & x_{11} & x_{13} & x_{17} \\ x_8 & x_{12} & x_{14} & x_{18} \\ x_9 & x_{13} & x_{15} & x_{19} \end{pmatrix} \qquad Y = \begin{pmatrix} x_7 \\ x_8 \\ x_9 \\ x_{12} \\ x_{13} \\ x_{14} \end{pmatrix}.$$

В случае если $\det(X) \neq 0$, решением СЛАУ (3) является $\hat{\beta} = X^{-1}Y$, но в практических приложениях часто возникают ситуации, когда $\det(X) = 0$ или матрица системы X является прямоугольной, поэтому для нахождения решения $\hat{\beta}$ СЛАУ (3) необходимо найти псевдообратную матрицу X^+. Псевдорешение является наилучшим приближением решения матричного уравнения (в смысле метода наименьших квадратов): оно минимизирует величину среднеквадратичного отклонения между правой и левой частью СЛАУ, что обеспечивает корректный выбор коэффициентов β в авторегрессионной модели.

Наиболее распространенным методом вычисления псевдообратной матрицы является скелетное разложение, но его основным недостатком является высокая вычислительная сложность. Поэтому в работе предлагается использовать метод Гревилля последовательного нахождения псевдообратной матрицы X^+ [3].

Далее исследуем точность полученных оценок. Вначале покажем, что оценка $\hat{\beta}$ является несмещенной, т.е. $E\hat{\beta} = \beta$.

$$E\hat{\beta} = E\left(X^+Y\right) = X^+E\left(Y\right) = X^+E\left(X\beta + \varepsilon\right)$$
$$= X^+X\beta + X^+E\left(\varepsilon\right) = \beta.$$

Затем подсчитаем матрицу ковариаций оценки $\hat{\beta}$:

$$V\left(\hat{\beta}\right) = V\left(X^+Y\right) = X^+V\left(Y\right)\left(X^+\right)^T =$$

$$= X^+V\left(X\beta + \varepsilon\right)\left(X^+\right)^T = X^+\sigma^2 I\left(X^+\right)^T =$$

$$= \sigma^2 X^+\left(X^+\right)^T = \sigma^2 X^+\left(X^T\right)^+ = \sigma^2\left(X^TX\right)^+$$

Таким образом, получаем что $\hat{\beta} \in N\left(\beta, \sigma^2 \left(X^T X\right)^+\right)$, если $\varepsilon \in N\left(0, \sigma^2 I_n\right)$. Обозначим через q^{ii} - i-ый диагональный элемент матрицы $\left(X^T X\right)^+$, тогда

$$\hat{\beta}_i \in N\left(\beta, \sigma^2 q^{ii}\right) \quad \text{и} \quad \frac{\hat{\beta}_i - \beta_i}{\sigma_{\beta_i}} \in N(0,1),$$

где $i = 1,...,k$, k - количество точек в окрестности, $\sigma_{\beta_i}^2 = \sigma^2 q^{ii}$.

Приведем пример расчета коэффициентов на основе радиолокационного изображения, полученного с помощью SAR (синтезированной апертуры), которое представлено на рис. 3. Для участка отмеченного белым цветом были рассчитаны оценки $\hat{\beta}$. При расчетах учитывалось, что порядок модели (порядок окрестностей) равен двум. В результате расчетов получен следующий вектор оценок:

$$\hat{\beta} = (0.59 \;\; 0.51 \;\; 0.56 \;\; 0.57 \;\; -0.33 \;\; -0.28 \;\; -0.33 \;\; -0.29)^T.$$

Проверим найденные оценки коэффициентов $\hat{\beta}$ на значимость, т.е. отличность значения коэффициента от нуля. Для этого рассмотрим гипотезу $H_0: \; \beta_j = 0$ и альтернативную ей гипотезу $H_\alpha: \; \beta_j \neq 0$. Рассчитаем величину $t_{набл} = \frac{\beta_j}{\tilde{S}_{\beta_j}}$, где $\tilde{S}_{\beta_i} = S\sqrt{q^{ii}}$ и критическую область $G = \left\{|t| > t_{1-\alpha/2}\left(n-k\right)\right\}$, $t_{1-\alpha/2}$ - квантиль распределения Стьюдента с $n-k$ степенями свободы, где $n = |S|$, k - количество точек в окрестности. H_0 отвергается, если $t_{набл.}$ не попадает в критическую область G.

В результате численного эксперимента на оценках $\hat{\beta}$, рассчитанных на основе предлагаемого метода, было получено, что все они являются значимыми, т.е. гипотезу $H_0: \; \beta_j = 0$ отвергаем в пользу альтернативной $H_\alpha: \; \beta_j \neq 0$, $j = 1,...,k$.

Для реализации метода расчета количества текстур, присутствующих на изображении, необходимо данное изображение разбить на непересекающиеся области. В каждой области нужно рассчитать коэффициенты линейной взаимосвязи пикселей изображения по предложенному выше алгоритму.

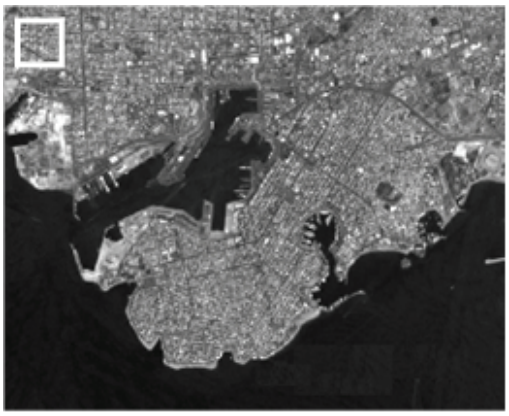


Рис. 3 Исходное изображение

Далее по известным значениям оценок коэффициентов авторегрессионной модели МСП строится гистограмма их распределения, как представлено на рис. 4.

Анализ гистограммы показывает, что наблюдаются ярко выраженные максимумы и некоторые шумы в распределении текстурной характеристики. Количество четких максимумов соответствует количеству текстур на исходном изображении. Каждая текстура обладает определенным значением текстурной характеристики, но по причине зашумленности реальных изображений эти значения имеют статистический разброс. Чтобы избавиться от шумов в распределении, каждое значение гистограммы усредняется по соседним значениям. Результат усреднения представлен на рис. 4.

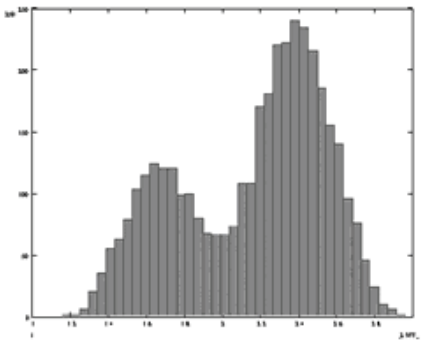


Рис. 4 Гистограмма распределения значений коэффициентов взаимосвязи

Таким образом, по усредненной гистограмме можно определять количество максимумов, которое будет соответствовать количеству текстур на рассматриваемом изображении. Для решения данной задачи разработан алгоритм и написана программа, которая выполняет сглаживание гистограммы, а затем производит поиск максимумов [4].

Разработанный метод носит достаточно универсальный характер. Он подразумевает использование текстурных характеристик, полученных не только с помощью коэффициентов линейной взаимосвязи пикселей изображения, но и какого-либо другого «оконного» алгоритма, предполагающего расчет локальных текстурных характеристик.

Список литературы:

[1]. Г. Винклер «Анализ изображений, случайные поля и динамические методы Монте-Карло» Новосибирск. Изд-во СО РАН, 2002.

[2]. Bing Yue, SAR Sea Ice Recognition Using Texture Methods PhD, Waterloo, Ontario, Canada, 2002

[3]. Ф. Р. Гантмахер «Теория матриц». М., 1966. Изд-во «Наука».

[4]. Патана Е.И. Метод расчета количества текстур для выполнения сегментации изображений. // Интегрированные модели и мягкие вычисления в искусственном интеллекте. Сборник трудов IV-й Международной конференции, 2007. — С. 220-224.

Использование цифрового нейронного кортекса на примере распознавания образов

Лазаренко А.В.
МФТИ
Lazarenko.Andrey@gmail.com

Михайлов А.М.
ИПУ РАН
alx_mikh@yahoo.com

Аннотация

В работе рассматривается система иерархической индексной организации памяти, которая называется цифровой нейронный кортекс(DNC). DNC может быть использован как для индексирования больших объемов текстовой информации, так и для индексирования изображений. Поскольку DNC – индексная система, то не производится поэлементного сравнения, что является серьезной проблемой при больших количествах элементов. Обобщение основных свойств приводит к устойчивости свойств, таким образом достигается устойчивость к шумам и изменениям масштабов и деформациям форм. До настоящего времени на основе DNC построена немало систем. В данной статье обсуждается не теория DNC, а пример конкретного применения.

1. Введение

Процесс распознавания объектов или сцен связан с обработкой большого количества информации. Для идентификации объектов изображённых на рисунки необходимо сравнить его с тысячами или миллионами шаблонов хранящихся в памяти. Поэтому при большем количестве информации для повышения эффективности алгоритмов выделение и идентификации объектов, необходимо разрабатывать методы или подходы, позволяющие быстро находить и сравнивать накопленную информацию.

В работе рассматривается алгоритм распознавания сцен построенный полностью на обработке информации с помощью индексов. Процесс обобщения основан на идентификации общих свойств объектов и опирается на иерархию индексов. То есть результаты обобщения по индексу нижнего уровня учитываются как свойства индексов более высокого уровня: сначала идентифицируются линии и кривые, потом они объединяются в фигуры, которые, в свою очередь, объединяются в ландшафты. Ниже будет рассмотрен алгоритм, который, опираясь на подходы обработки информации из пункта 3, позволит последовательно выделить объекты на изображение, идентифицировать их и сравнить полученную сцену объектов со сценами, распознанными ранее.

Для иллюстрации описываемого алгоритма будем использовать изображение (рисунок 1).

.

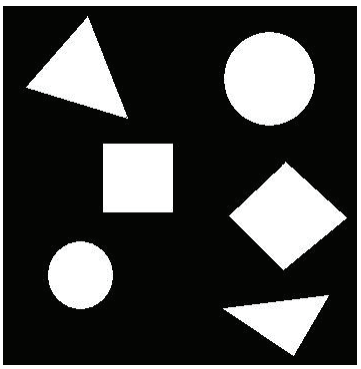


Рисунок 1. Бинарное изображение.

2. Цифровой нейронный кортекс.

DNC не принадлежит классу традиционных систем искусственного интеллекта. Он, скорее, использует невзвешенную нейронную парадигму(WNN), которая была введена в [1] и усовершенствована в [2], [3], [4] и других работах. Разница между DNC и WNN следующая: а) в отличие от WNN, DNC использует иерархическую структуру данных; б) WNN структура данных являет собой матрицу, тогда как DNC структура данных использует списки; в) в отличие от WNN, DNC – самообучающаяся система.

2.1. Определение индекса.

Под понятием индекса будем понимать колонковый индекс или книжный индекс. Формальное определение индекса: пусть X – множество элементов и пусть для каждого

$x \in X$ сопоставлено множество элементов Y(x), тогда множество всех пар (x, Y(x)) будем называть индексом. Введём следующее обозначение для индекса: $\{y\}_x$. Другими словами, индекс можно определить как список элементов, где каждый элемент имеет список ссылок на другие элементы. Как пример можно вспомнить книжный индекс, который встречается практически в любой книге. Книжный индекс содержит список слов, где для каждого слова сопоставлен список страниц, на которых встречается данное слово (Индекс 1).

43				43
190	98	190		190
32	34	43	100	32
Дом	Кот	Машина	Собака	Мышь

Индекс 1. Книжный индекс.

2.2. Итеративный индекс.

Рассмотрим индекс, у которого элементы и номера колонок одной природы, то есть значения колонок указывают на номера колонок. $\forall y \in \{y\}_x, x \in X \to y \in X$.

Как пример можно рассмотреть граф: где номерами колонок будут номера вершин, а элементами колонок будут вершины, с которыми связана данная вершина.

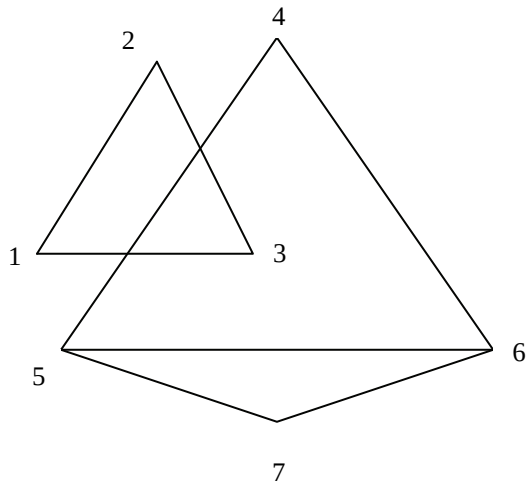


Рисунок 2. Два связных графа.

				4	4	
2	1	1	5	6	5	5
3	3	2	6	7	7	6
1	2	3	4	5	6	7

Индекс 2. $\{x\}_x$

Как показано выше, для графа, изображенного на рисунке 2, индексом будет Индекс 3. То есть мы получили индекс $\{x\}_x$, или, другими словами, - “итеративный индекс”. Если посмотреть внимательно на получившийся индекс, то можно заметить, что в нём существуют группы элементов, не связанные друг с другом. Для Индекса 2: {1,2,3} и {4, 5, 6, 7}.

На основание вышеуказанного примера возникает задача кластеризации индексируемых элементов итеративного индекса. То есть на основании индекса $\{x\}_x$ получить индекс $\{x\}_g$, где $g \in G$ – список кластеров.

Для Индекса 2 $\{x\}_g$ будет иметь следующий вид:

	4
1	5
2	6
3	7
g1	g2

Индекс 3. $\{x\}_g$

Понятие итеративного индекса и кластеризации будет применяться для обобщения множества простых элементов во множество более сложных элементов. Как можно было увидеть на рассмотренном примере, из списка вершин мы выделили две несвязанные друг с другом фигуры.

2.3. Макроколонки.

Известно, что в коре головного мозга колки нейронов образуют макроколонки, но назначение этих макроколонок пока неизвестно. Сделаем предположение, что макроколонки нужна для того, чтобы при активации одной колонки в коре головного мозга, также происходит активация соседних колонок. Смоделировать этот механизм с помощью индексов, впоследствии он понадобится для обобщения элементов.

Пусть задан индекс $\{r\}_n$, где r будет соответствовать колонке в коре головного мозга, а n – абстрактный номер колонки. Также допустить, что есть некоторый индекс, который характеризует размещения колонки в коре головного мозга – $\{x\}_n$. В рамках данных условий, задача будет звучать так: необходимо определить список индексов колонок, которые будут активизированы при активизации колонки с индексом k.

Для решения поставленной задачи введём некоторую функцию $g(x_1,x_2)<1$, которая позволит на основании индекса, характеризующего

размещения колонки в коре головного мозга, построить критерий, определяющий индексы колонок, которые должны активизироваться. Пусть индекс $\{n\}_x$ обратный для $\{x\}_n$, тогда критерий будет иметь следующий вид.

Критерий 1:

$$\{n\}_k = \bigcup_{\forall i: g(\{x\}_k,\{x\}_i)<1} \{n\}_i$$, где $\{n\}_k$ – список номеров колонок, которые будут активизированы при активизации колонки *k*.

Также в данном разделе рассмотрим какой вид будет иметь критерий в случае, когда в качестве индекса характеризующего размещение колонки в коре головного мозга будет выступать не один, а несколько индексов: $\{x_1\}_n,..,\{x_m\}_n$. Построим для них обратные индексы: $\{n\}_{x_1},..,\{n\}_{x_m}$, тогда получим новый вид критерия.

Критерий 2:

$$\{n\}_k = \bigcap_{j\in[1..m]} \bigcup_{\forall i: g(\{x_j\}_k,\{x_j\}_i)<1} \{n\}_i$$, где

$\{n\}_k$ – список номеров колонок, которые будут активизированы при активизации колонки *k*.

3. Индекс первого уровня. Картина углов

В результате обработки изображения с помощью метода “кортексного окна” на выходе будут получен индекс углов $\{a\}_n$, где $n \in N$ – номера пикселей. Этот индекс определяет параметр угла в каждой точке изображения. Для наглядности приводятся результаты численного моделирования (рисунок 3). Фигуры закрашены цветами, которые соответствуют значениям параметров углов пикселей.

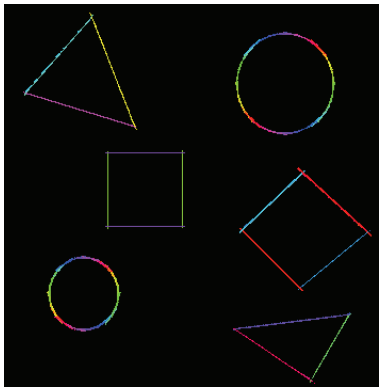


Рисунок 3. Результат обработки на первом уровне.

4. Индекс второго уровня. Распознавание линий

Для объединения пикселей в линии воспользуемся методами макроколонок и кластеризации, которые были описаны выше.

Сначала применим алгоритм макроколонок, чтобы получить индекс $\{n\}_n$, где $n \in N$ – номера пикселей. Для этого построим обратные индексы для $\{x\}_n,\{y\}_n,\{a\}_n$. В результате получим $\{n\}_x$, $\{n\}_y$ и $\{n\}_a$ соответственно. Затем воспользуемся критерием 2 из раздела “Макроколонки”, задав функцию g для параметров x, y, a:

$$g(x_i,x_j) = k_x * \left|x_i - x_j\right|$$

$$g(y_i,y_j) = k_y * \left|y_i - y_j\right|$$

$$g(a_i,a_j) = k_a * \left|a_i - a_j\right|$$

В результате последовательного вычисления $\{n\}_k$ для каждого пикселя получим индекс $\{n\}_n$. Далее применим алгоритм кластеризации, который выделит из пикселей линии. Результатом кластеризации будет индекс $\{n\}_l$, где $l \in L$ – номер линии. Результаты моделирования показаны на рисунке 4, где разными цветами закрашены линии, которые были получены после обработки индекса первого уровня, изображенного на рисунке 3. При расчете использовались параметры k_x =3, k_y =3, k_a =2.

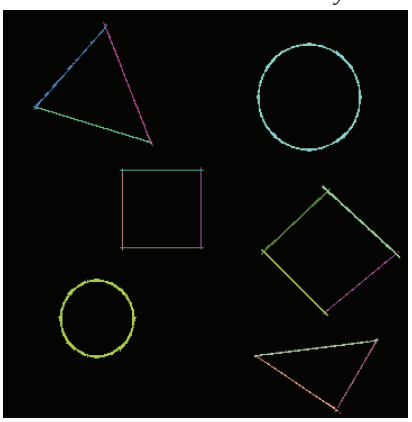


Рисунок 4. Результат обработки на втором уровне.

5. Индекс третьего уровня. Распознавание фигур

На этом этапе происходит объединение линий в фигуры. В данном разделе нам

понадобится результат обработки второго уровня: индекс $\{n\}_l$, объединяющий пиксели в линии, а также индексы $\{n\}_x$, $\{n\}_y$ обратные для $\{x\}_n$, $\{y\}_n$.

Построим индекс $\{l\}_x$ и $\{l\}_y$,используя индекс $\{l\}_n$,обратный для $\{n\}_l$:

$$\{l\}_x = \bigcup_{n \in \{n\}_x} \{l\}_n$$

$$\{l\}_y = \bigcup_{n \in \{n\}_y} \{l\}_y$$

Воспользуемся критерием 2 из раздела “Макроколонки”, задав функцию g для параметров x, y, a:

$$g(x_i, x_j) = k_x * |x_i - x_j|$$

$$g(y_i, y_j) = k_y * |y_i - y_j|$$

После последовательного вычисления $\{n\}_k$ для каждой линии получим индекс $\{l\}_l$. Далее применим алгоритм кластеризации, результатом которого будет индекс $\{l\}_f$, где $f \in F$ – номера фигур. Результаты приведены на рисунке 5, где разными цветами закрашены фигуры, которые были выделены в результате обработки индекса второго уровня, изображенного на рисунке 4. При обработке использовались параметры k_x =4, k_y =4.

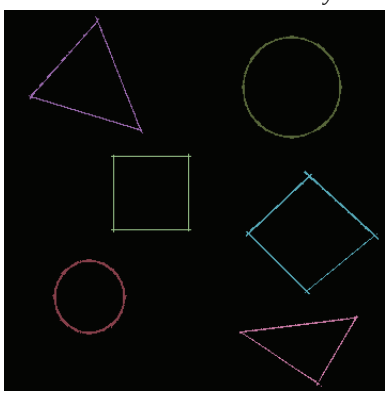


Рисунок 5. Результат обработки на третьем уровне.

6. Индекс четвёртого уровня. Распознавание объектов.

После выделения фигур на изображении необходимо найти способ их классификации. Например: квадраты, прямоугольники можно объединить в один класс - четырёхугольники. Классификация фигур может проводиться с помощью выделения общих свойств. Одним их таких свойств можно считать структуру фигуры, то есть связь линий, из которых состоит фигура.

В пункте 6 будут рассмотрены три типа выделения свойств объектов, которые описывают не только структурные свойства, но и свойства линий, из которых состоит фигура. Это означает, что каждая фигура описывается тремя свойствами. На основании полученных свойств производится классификация фигур.

Рассмотрим случай, когда у фигуры есть n - свойств. Это означает, что существует n - индексов: $\{t_1\}_f \ldots \{t_n\}_f$, где t_i - значение i - свойства для соответствующей фигуры. Причём под t_i может подразумеваться как само значение свойства, так и номер более сложной идентифицирующей формулы. Случай, когда каждой фигуре сопоставлено единственное значение для каждого свойства (то есть высота колонок в индексах $\{t_1\}_f \ldots \{t_n\}_f$ будет равна единице. В индексном представлении - это может быть набор значений.

Также должны существовать n - индексов: $\{t_1\}_c \ldots \{t_n\}_c$, где t_i - значение i - свойства для соответствующего класса объектов(типа объекта). То есть значения, хранящиеся в этих индексах, задают параметры, которым должна удовлетворять фигура, чтобы принадлежать данному классу. Формирование индексов $\{t_1\}_c \ldots \{t_n\}_c$ не требует какого-то предварительного обучения, как это обычно делается в методах классифицирования. Заполнение индексов со значениями свойств класса производится в момент распознавания изображения.

Общая схема при классифицировании состоит из двух этапов: по полученным свойствам фигуры производится поиск данного типа объекта в уже имеющемся наборе классов и, в случае отсутствия, дополнение этого набора новым типом объекта.

Процедура поиска типа объекта для некоторой распознанной фигуры f' осуществляется следующим образом:

$$\{c\}_{f'} = \bigcap_{i=1..n} \{c\}_{t_i},$$

где $\{c\}_{t_i}$ - индекс, обратный для индекса $\{t_i\}_c$. В случае отрицательного результата в индексы $\{t_1\}_c \ldots \{t_n\}_c$ необходимо занести новый класс c', со значениями $\{t_1\}_f \ldots \{t_n\}_f$ для фигуры f'.

В результате описанного выше алгоритма получим индекс $\{c\}_f$, который сопоставляет каждой фигуре тип объекта, и дополненный набор индексов $\{t_1\}_c \ldots \{t_n\}_c$,описывающий характеристики каждого класса.

7. Индекс пятого уровня. Распознавание сцен.

Результатом четвертого уровня был индекс $\{c\}_f$, сопоставляющий номер фигуры с номером типа фигуры и три вспомогательных индекса, содержащих информацию о типе фигур $\{i\}_t, \{s\}_i, \{c\}_i$.

На данном уровне необходимо провести идентификацию сцены, представленной на изображении, и сравнить её со сценами, распознанными ранее. То есть определить, была ли похожая сцена распознана ранее, и определить её формальное название (номер сцены).

Для идентификации сцен будем использовать тот же способ, что и для идентификации типов фигур. Определим способ формирования идентифицирующей формулы и способ её хранения в индексах.

Формирование формулы будет выполняться по следующим правилам:

Построим обратный индекс $\{f\}_c$ для $\{c\}_f$, который содержит списки фигур, имеющих одинаковые типы.

На основании $\{f\}_c$ рассчитаем для каждого типа количество фигур, соответствующих ему. Результатом вычисления будет индекс $\{s\}_t$, где c – тип фигуры, s – количество фигур.

Построим обратный индекс $\{c\}_s$ для $\{s\}_c$.

Из полученного индекса $\{t\}_c$ построим формулу сцены, которая будет иметь следующий вид: $c1_{s1}c2_{s2}c3_{s3}...cN_{sN}$.

Для хранения формулы сцены в индексах воспользуемся тем же подходом, что и в случае с формулой типа объекта. Поэтапно опишем алгоритм разбора формулы и занесения информации в индексы. Информацию о формуле будем хранить в индексах $\{s\}_e, \{c\}_e, \{e\}_t$ и обратных к ним $\{e\}_s, \{e\}_c, \{t\}_e$. Они необходимы для поиска информации о уже сохраненной информации о формуле.

Разбиваем формулу сцены $c1_{s1}c2_{s2}c3_{s3}...cN_{sN}$ на составные части $c1_{s1}, c2_{s2}, \ldots, cN_{sn}$.

Для каждой части ci_{si} пытаемся найти эквивалент в индексах $\{s\}_e, \{c\}_e$. $e = \{e\}_{s1} \bigcap \{e\}_{c1}$, если результат отрицательный, то заносим информацию в $\{s\}_e, \{c\}_e$ о ce_{se} под новым номером e'.

В результате выполнения пунктов 1 и 2 формула сцены $c1_{s1}c2_{s2}c3_{s3}...cN_{sN}$ преобразуется в набор номеров $\{e1, e2, .., eN\}$.

На следующем шаге необходимо проверить, была ли сохранена похожая сцена ранее. Номер сцены можно определить следующим образом: $sc = \bigcap_{e \in \{e1, e2, .., eN\}} \{sc\}_e$. Если результат отрицательный, заносим информацию о новой сцене в индекс $\{e\}_{sc}$ и $\{sc\}_e$.

Ниже приводится рисунок 6 на котором изображён результат обработки.

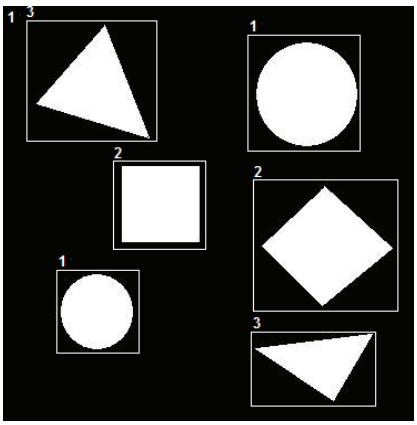


Рисунок 6. Результат обработки

8. Список использованных источников.

[1] W. Bledsoe, I. Browning, Pattern Recognition and Reading by Machine, 1959.

[2] I. Alexander, T. Stonham, A Guide to Pattern Recognition Using RAM, IEE J. Digital Systems and Computers, Vol.2., No.1., 1979.

[3] M. Morciniec, R. Rohwer, The n-Tuple Classifier: Too Good to Ignore, Intellix A/S Articles, Aston University, 1995.

[4] T. Jorgenson, Classification of Hand written Digits Using a RAM Neural Net Architecture

Метод идентификации типа документа по структуре проекций его изображения на координатные оси

Павел Безматерных
Московский институт стали и сплавов
bezmpavel@gmail.com

Дмитрий Николаев
Институт проблем передачи информации имени А. А. Харкевича РАН
dimonstr@iitp.ru

Василий Постников
Cognitive Technologies
pvv@cognitive.ru

Аннотация

В работе предложен новый подход к решению проблемы идентификации типа документа по структуре проекций его изображения на координатные оси. Подход основан на предположении о том, что структура проекций изображения на координатные оси неизменна для каждого типа документа. Для ориентации осей предполагается использовать преобразование Хафа, а для сравнения структуры гистограмм с соответствующими эталонами – алгоритм динамической трансформации временной шкалы (Dynamic time warping, DTW). Предложенный подход по построению является устойчивым к ряду геометрических искажений, в том числе к нелинейному масштабированию изображения. В работе также предлагается метрика, обеспечивающая устойчивость алгоритма DTW по отношению к вариациям заполнения документов-анкет.

1. Введение

В повседневной жизни людям часто необходимо заполнять различные анкеты, бланки, квитанции и прочие документы, которые затем подвергаются специальной обработке. Времена, когда оператору приходилось вручную верифицировать заполненные документы и извлекать из них данные, практически прошли. В настоящее время повсеместно используются технологии автоматизированного ввода документов. Подобные технологии позволяют ускорить обработку документов в десятки раз: современные сканеры позволяют сканировать со скоростью до нескольких страниц в секунду, а специализированные программы распознавания позволяют обрабатывать содержимое документа за время сканирования.

При этом распознавание документов вовсе не сводится к распознаванию букв, как может показаться на первый взгляд. Например, одна и та же распознанная «сумма прописью» может оказаться как доходной, так и расходной статьёй, причём эта неопределённость иногда не может быть снята вне контекста пакета документов. Для того чтобы как можно раньше снять неопределённость интерпретации содержащейся в документе информации, а также (в целях оптимизации) ограничить распознаваемую область документа, на одном из первых этапов обработки изображения документа определяется *тип документа* [1].

Определим теперь некоторые термины, которыми будем пользоваться в дальнейшем. *Статическими элементами* документа некоторого типа будем называть те фрагменты текста и элементы оформления, которые не изменяются от образца к образцу (например, фрагмент «Код подразделения» в паспорте). *Заполнением* будем называть элементы, составляющие собственно содержание документа и меняющееся от экземпляра к экземпляру. *Бланком* документа будем называть совокупность статических элементов.

Методы идентификации типа документа различаются в зависимости от класса трансформаций бланка документа, допустимого внутри типа. Как правило, этот класс включает смещение и поворот, т к. их практически невозможно избежать при сканировании графического образа документа. Типы документов, бланки которых испытывают только такие деформации, часто называют документами *с фиксированной геометрией*. Примером такого документа являются водительские права. По аналогии, документами *с нефиксированной геометрией* будем называть все прочие типы документов (примером может быть счёт-фактура).

Наибольшее распространение среди методов идентификации документов на сегодняшний момент получил метод, основанный на выделении предопределённых графических примитивов и сравнении их взаимного расположения с эталонным [2-4]. Примитивами могут являться, например, линии, таблицы, специальные ориентирующие символы. Известен также подход, основанный на поиске ключевых слов (т.н. атрибутов документа) в результатах предварительного распознавания текста на изображении документа. Этот подход характерен крайней устойчивостью к геометрическим трансформациям бланка документа [2].

Однако у алгоритмов типизации документов, как с использованием графических примитивов, так и основанных на атрибутивном поиске, есть общий недостаток. И выделение примитивов, и распознавание требуют корректной бинаризации изображения (классификации каждого пикселя как принадлежащего фону или распознаваемому элементу). К сожалению, опыт показывает, что для корректной бинаризации документа требуется знание его типа. Продемонстрируем данное утверждение на Рис. 1-2. Как известно [5], одним из самых универсальных методов бинаризации является метод Ниблэка. На Рис. 1 показаны результаты применения этого алгоритма к изображению карточки пенсионного страхования. Столь плачевный результат связан с тем, что карточки обладают текстурным фоном с контрастом, близким к контрасту заполнения. При этом, если воспользоваться знанием о том, что фон в данном случае строго периодичен, можно достичь гораздо лучшего результата (Рис. 2). К сожалению, не существует методов бинаризации, дающих одинаково хорошие результаты для всех типов документов.

Рис. 1. Стандартная бинаризация заполнения пенсионной карточки

Рис. 2. Оптимальная бинаризация заполнения пенсионной карточки

Отсюда очевидно, что в некоторых случаях задача идентификации типа документа требует алгоритмов, не требующих выделения объектов на изображении, а опирающихся по возможности на интегральные характеристики изображения.

В данной работе предложен новый подход к решению проблемы типизации документа по изображению, опирающийся на сравнение структуры проекций изображения на координатные оси. Подход основан на предположении о том, что структура проекций изображения на координатные оси неизменна для каждого типа документа.

2. Типизация документов на основе алгоритма DTW

Итак, рассмотрим класс документов, такой что 1) не существует метод бинаризации, позволяющий уверенно выделять элементы документов на всех изображениях; 2) класс допустимых искажений ограничен сдвигом, поворотом, масштабированием в пределах 10% и локальной деформацией. К этому классу относятся, например: развороты паспортов (деформация за счёт неточного совмещения листов в тетради), анкеты, распечатываемые из файлов формата PDF (случайное масштабирование в пределах 10% в зависимости от драйвера принтера), «жировки» ЕИРЦ (большое количество разнотиражных бланков, отличающихся шириной зазоров в таблице).

Задача типизации документов, принадлежащих указанному классу, очевидно, может быть сведена к задаче оптимальной деформации объекта до соответствия одному из набора *эталонов*, задающих типы документов. К сожалению, даже лучшие из известных алгоритмов эластичного совмещения изображений имеют сложность не менее $O\left(n^6\right)$ [6].

К счастью, большинство печатных документов обладает, по сути, табличной структурой. Это позволяет надеяться, что совмещение двух проекций (горизонтальной и вертикальной) яркости изображения достаточно для приведения изображения к эталону. Вычислительная сложность такого алгоритма определяется сложностью одномерного варианта DTW и составляет $O\left(n^2\right)$.

Однако, эти рассуждения верны, только если нам известна ориентация координатных осей на изображении документа. К сожалению, ошибка углового позиционирования документа на сканере может достигать 5˚. На Рис. 3-6 приведен пример влияния наклона образца на получаемые проекции (в данном случае яркость проецируется на ось Y). Как можно увидеть из сравнения Рис. 4 и 5, такой наклон может полностью разрушить яркостные проекции. Для того чтобы избавиться от этого эффекта, следует использовать какой-либо алгоритмом автоматического определения угла наклона документа. Можно, например, воспользоваться весьма эффективным алгоритмом, основанным на поиске строки преобразования Хафа с максимальной дисперсией [7].

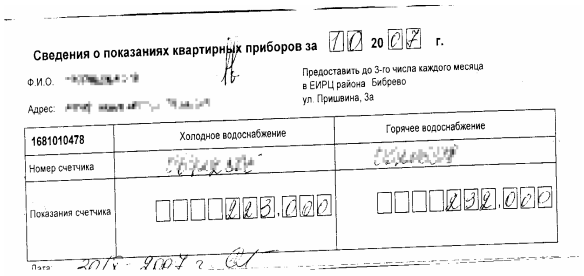


Рис. 3. Пример изображения неточно ориентированного документа

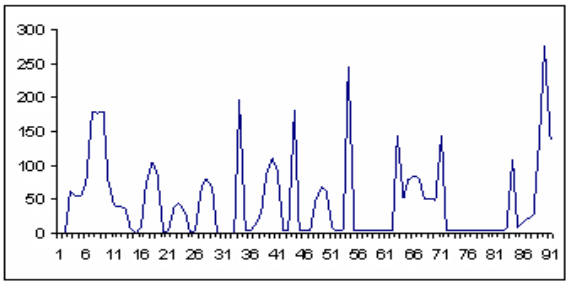


Рис. 4. Эталон проекции яркости на ось Y для документа, изображённого на Рис. 3

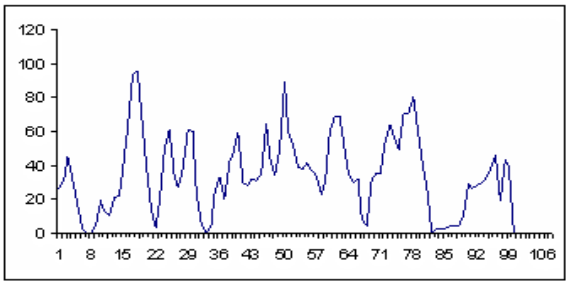


Рис. 5. Проекции яркости на ось Y документа, изображённого на Рис. 3

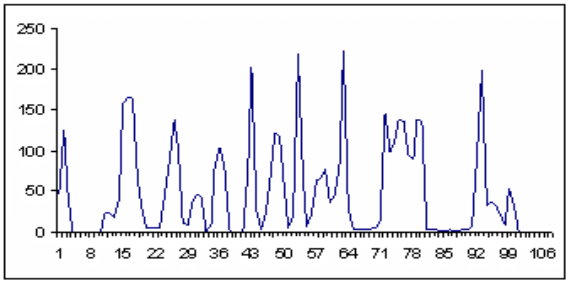


Рис. 6. Проекции яркости документа, изображённого на Рис. 3, после доворота.

Существует ещё одна, более серьёзная проблема. Каждый документ имеет своё индивидуальное заполнение. Влияние заполнения на вид проекций яркости может оказаться сравнимым с влиянием статических элементов бланка. Рассмотрим Рис. 7. Изображённые на нём квитанции имеют различное заполнение в области, выделенной красным прямоугольником. На Рис. 8. видно, как это влияет на структуру проекций яркости. Для того, чтобы добиться устойчивости относительно вариации заполнения, мы предлагаем ввести весовую функцию в алгоритме DTW:

$$D_{ij} = \frac{\|e_i - s_j\|}{\sqrt{\sigma_i^2 + \sigma_0^2}}, \quad (1)$$

где D_{ij} – штраф несоответствия значению проекции образца s_j в точке j эталонному значению e_i в точке i, σ_i^2 – оценка дисперсии значения e_i по выборке документов, а $\sigma_0^2 > 0$ – экспериментально подбираемый параметр. В результате алгоритм DTW будет фактически игнорировать невязки сопоставления в зонах заполнения документа.

Итак, предлагаемый алгоритм использует в качестве эталона каждого типа документов 4 массива (e_X, σ_X, e_Y и σ_Y), и состоит из следующих шагов:

Шаг 1. Автоматическая ориентация образца документа.

Шаг 2. Вычисление проекций яркости исследуемого изображения на координатные оси.

Шаг 3. Вычисление расхождения полученных проекций с эталонными (для каждого типа документов, обрабатываемых в потоке).

Шаг 4. Выбор наиболее подходящего эталона (давшего минимальную меру расхождения) и соответствующего ей типа документа.

Шаг 5. Выделение областей распознавания, указанных на эталоне, согласно карте соответствия координат, полученных на шаге 3 алгоритмом DTW.

Шаг 6. Распознавание выделенных областей.

3. Автоматическое построение эталона по набору образцов

Рассмотрим теперь важную задачу автоматического построения эталона типа документа по набору образцов изображений документов этого типа. Предлагается следующий алгоритм:

Шаг 1. Сориентируем каждый образец.

Шаг 2. Для всевозможных пар образцов оценим взаимное соответствие образцов алгоритмом DTW (σ_i считается равной 0) и найдём наилучшую пару.

Шаг 3. Усредним образцы с учётом взаимной деформации.

Шаг 4. Среди оставшихся образцов найдём наближайший к усреднённому алгоритмом DTW (σ_i считается равной 0). Если число усреднённых образцов меньше k, перейдём к Шагу 3.

Шаг 5. Усредним образцы с учётом взаимной деформации и оценим σ_i.

Шаг 6. Если набор не исчерпан, среди оставшихся образцов найдём наиближайший к усреднённому и перейдём к Шагу 5.

Таким образом будет автоматически получен эталон типа изображения.

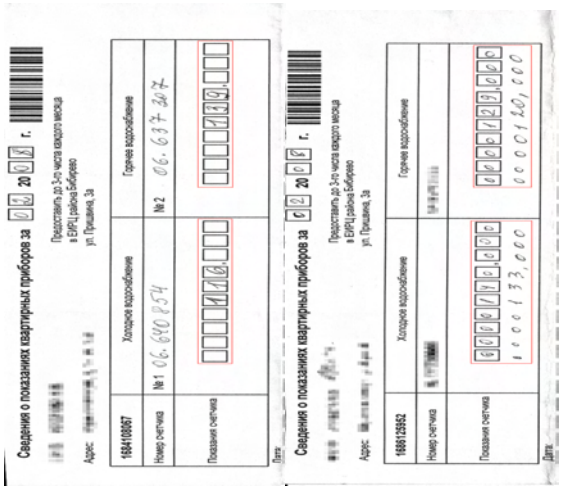


Рис. 7. Примеры квитанций ЕИРЦ об оплате водоснабжения (красным выделены области с варьирующимся заполнением)

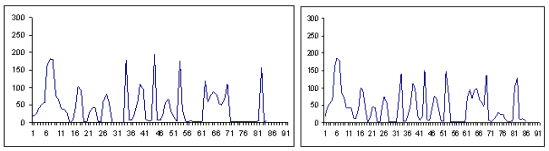


Рис. 8. Проекции яркости на ось Y для квитанций (Рис. 7)

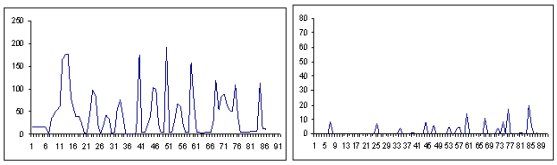


Рис. 9. Среднее и стандартное отклонение по множеству образцов для данного типа документов.

4. Заключение

В работе предложен новый алгоритм идентификации типа документа, основанный на анализе структур проекций яркости на координатные оси. Данный алгоритм позволяет определить тип документа в потоке документов фиксированной геометрии, подверженных небольшим деформациям. Алгоритм не требует предварительной бинаризации изображений и не требует ни выделения графических примитивов, ни предварительного распознавания.

Для рассмотренного алгоритма предложена процедура автоматического построения эталонов документов по набору образцов.

На основании предложенного алгоритма разработан демонстрационный прототип программы идентификации, позволяющий выделять квитанции ЕИРЦ об оплате водоснабжения из смешанного потока документов. Тестирование проводилось на наборе из 2000 изображений документов, 500 из которых составляли искомые квитанции. Из них правильно идентифицированы были 498 документов. Случаев ложной идентификации не произошло. Среднее время обработки составило около 0.1 секунды на ПК класса Pentium IV. Финальное распознавание документов производилось системой Cognitive Forms компании Cognitive Technologies.

8. Список литературы

[1] В. Л. Арлазаров “Управление информационными потоками в системе автоматического ввода документов”, *Труды ИСА РАН “Управление информационными потоками”*, 2002, стр. 17-29.

[2] В. В. Постников, “Автоматическая идентификация и распознавание структурированных документов”, *Автореф. дис. ... канд. тех. наук*, Москва, 2001.

[3] В. В. Постников, “Идентификация и распознавание документов заранее известной структуры”, *Сборник трудов 6-ой Международной конференции “Распознавание образов и анализ изображений: новые информационные технологии”*, Великий Новогород, 2002.

[4] Ming Ye, M. Bern, D. Goldberg, “Document image matching and annotation lifting”, *Proc. ICDAR*, 2001, стр. 753-760.

[5] P. D. Trier and T. Taxt. «Evaluation of binarization methods for document images», *IEEE Trans. Pattern Analysis and Machine Intelligence*, 17(3):312-315, March 1995.

[6] Hansheng Lei; Govindaraju, V. «Direct Image Matching by Dynamic Warping», *Computer Vision and Pattern Recognition Workshop*, 2004, 76-76

[7] D. Nikolaev, S. Karpenko, I. Nikolaev, P. Nikolayev, «Hough transform: underestimated tool in the computer vision field», *Proceedings of the 22th European Conference on Modelling and Simulation*, 2008, 238-246

Метод восстановления изображений, подверженных радиальной дисторсии

Дмитрий Николаев
Институт проблем передачи информации имени А.А. Харкевича РАН
dimonstr@iitp.ru

Сергей Гладилин
Институт проблем передачи информации имени А.А. Харкевича РАН
gladilin@iitp.ru

Семен Карпенко
Институт проблем передачи информации имени А.А. Харкевича РАН
simon.karpenko@gmail.com

Аннотация

В работе обсуждается новый метод восстановления изображений, подверженных радиальным оптическим искажениям. Метод основан на быстром вычислении полного преобразования Хафа (Hough transform) исходного изображения, не требует вмешательства человека, может применяться к изображениям, полученным от неизвестных фото- и видеокамер без предварительной калибровки. Суть метода заключается в подборе параметров искажения, исправление которого привело бы к образованию на изображении прямых линий максимально возможной общей протяжённости. Эта величина оценивается по результатам преобразования Хафа с помощью разработанного авторами функционала. Надежность метода иллюстрируется на реальных изображениях.

1. Введение

Почти каждая современная книга по обработке изображений находит место для рассказа о преобразовании Хафа (Hough transform) [1]. Данное преобразование является дискретным вариантом преобразования Радона – интегрирования изображения вдоль всевозможных прямых на плоскости. Несмотря на столь широкую известность преобразования Хафа, его практическое применение для решения задач компьютерной обработки изображений до сих пор незначительно. По-видимому, это связано с двумя факторами:

1. Считается, что вычисление преобразования Хафа малоэффективно и требует значительных вычислительных мощностей
2. Для ускорения вычислений преобразование Хафа обычно используется как M-estimator, т.е. вычисляется не весь Хаф-образ изображения, а оценивается, где на нем находится точка максимального значения (соответствующая самой длинной прямой линии на исходном изображении) и результат преобразования вычисляется только в ней. Использование в дальнейшем информации только о данной точке значительно снижает возможности предлагаемых алгоритмов.

На самом деле, проблема низкой эффективности вычисления преобразования Хафа весьма преувеличена [2] – известен алгоритм его вычисления за $O(n \cdot \log n)$ аддитивных операций, причем алгоритм хорошо оптимизируется для современных (с набором команд, включающем SIMD инструкции) процессоров.

Таким образом, существует возможность применения всего результата преобразования Хафа, а не только точки, где он достигает максимума. На наш взгляд, разработка подобных алгоритмов позволит решить ряд актуальных на данный момент задач обработки изображений и компьютерного зрения, один из примеров чему и будет продемонстрировано ниже.

В данном докладе рассматривается задача восстановления изображений, подверженных радиальным оптическим искажениям (дисторсии). Такая задача характерна для систем обработки видеоинформации, использующих недорогие широкоугольные фотоаппараты и видеокамеры, что является широко распространённой практикой.

2. Постановка задачи

Модель камеры с радиальным искажением будем задавать уравнениями, описывающими искажения двумерного изображения, получаемого ею, относительно двумерного изображения, которое было бы получено в тех же условиях идеальной камерой-обскурой.

Простейшая однопараметрическая модель дисторсии [3] описывается следующей зависимостью координат точки на идеальном

снимке (x_u, y_u) от координат на реальном снимке с дисторсией (x_d, y_d):

$$x_u = x_d + k\,(x_d - c_x)\,(x_d^2 + y_d^2),$$
$$y_u = y x_d + k\,(y_d - c_y)\,(x_d^2 + y_d^2), \quad (1)$$

где (c_x, c_y) – координаты центра радиального искажения, k – параметр дисторсии.

Таким образом, задача поиска параметров восстановления заключается в поиске такого значения k, при котором обратное к (1) преобразование восстановит изображение максимально близким к тому, которое было бы получено в аналогичных условиях идеальной камерой.

В ситуации неизвестных параметров камеры единственным методом определения параметра радиального искажения (для последующего исправления) является анализ полученных от работы камеры изображений. Предлагаемый нами метод позволяет подобрать параметр на основе анализа единственного изображения в предположении, что протяженные линии, которые оно содержит – в большинстве своем были до искажения прямыми.

3. Алгоритм восстановления на основе преобразования Хафа

Суть предлагаемого метода восстановления заключается в том, что выполняется несколько пробных восстановлений изображения для разных предполагаемых значений параметра дисторсии. Затем среди результатов восстановления выбирается наилучший. Наилучшим считается тот параметр, восстановление с которым приводит к образованию на изображении прямых линий максимально возможной общей протяжённости. Протяженность измеряется путем вычисления разработанного авторами функционала.

Перед вычислением преобразования Хафа происходит дифференцирование изображения (вычисления модуля градиента в каждой точке). Это связано с тем, что типичными линиями на реальных изображениях являются не линии высокой интенсивности, а *границы* предметов. В то же время преобразование Хафа (за счёт суммации) позволяет выделить на изображении *линии* (хребты). Дифференцирование изображения «превращает» границы в такие линии.

Вычисление функционала от результата преобразования Хафа происходит в несколько стадий. На первой стадии вычисляется «псевдолапласиан» – из Хаф-образа картинки вычитается он же, но размытый гауссовой свёрткой. Такое действие позволяет выделить на изображении именно отдельные прямые линии, отфильтровав равномерно окрашенные области (те, в которых небольшое смещение прямой, вдоль которой интегрируют, не изменяет значительно результат интегрирования). Также такая процедура позволяет исключить изогнутые линии – они на преобразовании Хафа порождают размазанные, а не резкие локальные максимумы.

Выбрав теперь в каждой строчке преобразования Хафа максимум, мы определим меру протяженности и «прямизны» (резкости после преобразования) линий, идущих в данном направлении.

На реальных изображениях крайне маловероятно, что все прямые будут строго параллельны – даже если они параллельны в пространстве, за счёт проективного преобразования на изображении они превратятся в пучок прямых. Таким образом, выбирая максимум в каждой сточке преобразования Хафа и исключая таким образом информацию о количестве параллельных прямых, мы для случая реальных изображений, никакой ценной информации не теряем.

Теперь среди самых длинных и прямых линий, расположенных под разными углами на изображении выберем только те, которые достаточно длинны – т.е. из выбранных значений преобразования Хафа выберем только превышающие определенный порог. Значения в выбранных точках сложим между собой. Полученное число и будет результатом вычисления функционала, позволяющего оценить длину прямых линий на изображении.

Среди восстановленных с разными параметрами дисторсии изображений то, для которого данный функционал дает наибольшее значение, и будем считать результатом исправления радиального искажения.

4. Эксперимент

Для эксперимента были взяты несколько реальных изображений, снятых разными широкоугольными камерами. Для каждого изображения строились зависимости значения функционала от параметра дисторсии (Рис. 1). Алгоритм справился с восстановлением всех изображений при фиксированном наборе параметров.

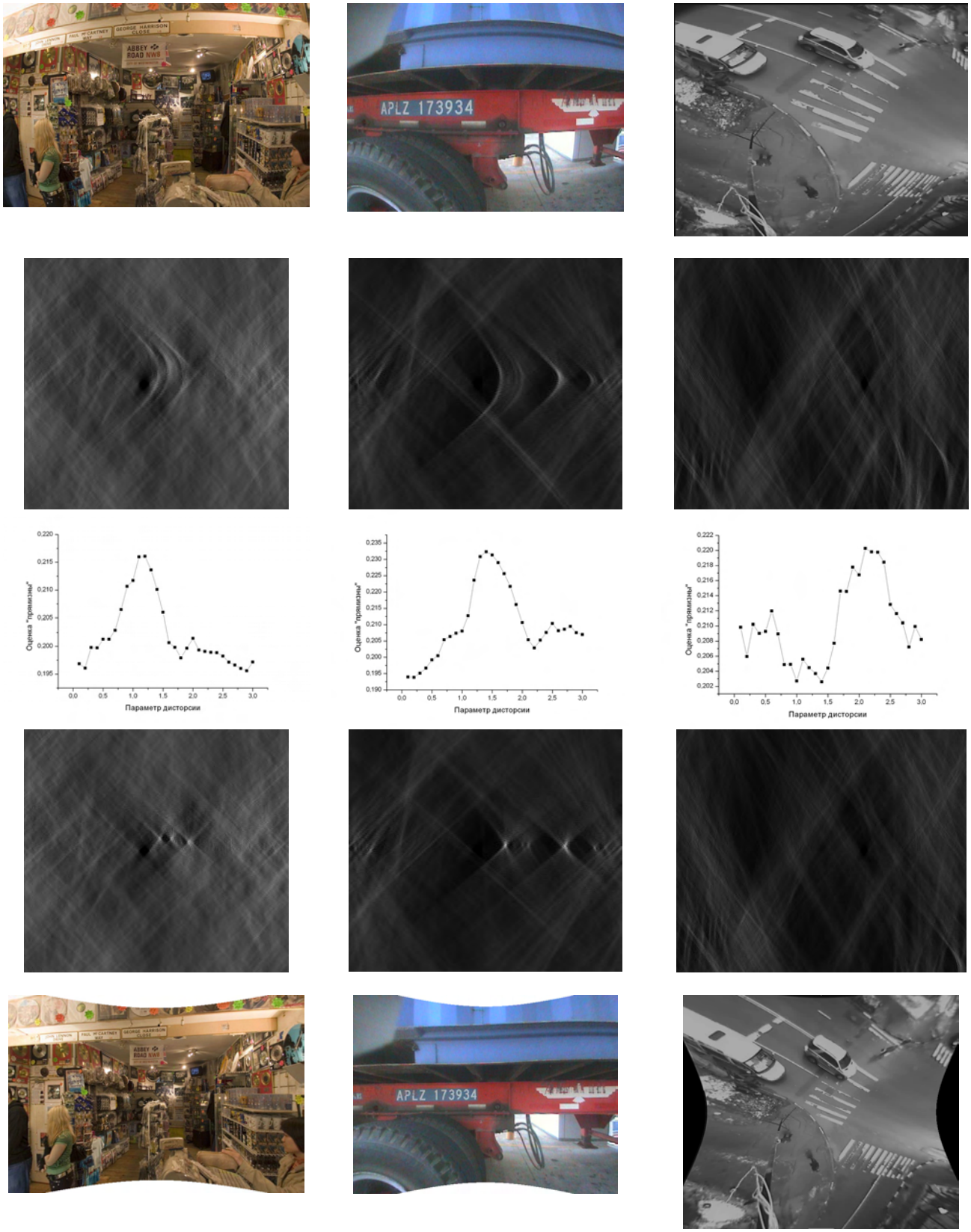


Рис. 1. Результаты компьютерного эксперимента. Все три исходных изображения были получены с разных камер (слева – камера в мобильном телефоне; посередине – камера системы распознавания номеров грузовых платформ на КПП, справа – ч/б камера слежения). Сверху вниз: исходное изображение; его преобразование Хафа; график зависимости максимизируемого функционала от параметра дисторсии; преобразование Хафа в точке максимума; восстановленное изображение.

5. Заключение

В статье предлагается алгоритм восстановления изображения, подверженного радиальным оптическим искажениям (дисторсии) на основе анализа только самого изображения. Приведены примеры применения данного алгоритма к реальным изображениям с неизвестными параметрами дисторсии.

В своей работе алгоритм использует оценку качества восстановлений на основе «прямизны» линий на изображении. На наш взгляд, такая оценка качества восстановления соответствует субъективному восприятию человека, также ориентирующегося при оценке качества восстановления на «прямизну» линий.

Предложенный критерий оценки «прямизны» для нескольких различных видов изображений, используемых в тестах, показал хорошие результаты.

Однопараметрическая модель дисторсии для тех изображений, на которых апробировалась работа алгоритма, оказалась вполне достаточной (можно наблюдать хорошее восстановление изображений), но не идеальной – на изображениях все же заметно, что при восстановлении не все линии «распрямились». Дальнейшее развитие алгоритма может заключаться в использовании более сложных моделей дисторсии, а также в замене полного перебора в пространстве параметров итерационным алгоритмом спуска.

6. Список литературы

[1] Р. Гонсалес, Р. Вудс, *Цифровая обработка изображений*, Москва, 2006.

[2] D. Nikolaev, S. Karpenko, I. Nikolaev, P. Nikolayev, «Hough transform: underestimated tool in the computer vision field», *Proceedings of the 22th European Conference on Modelling and Simulation*, 2008, 238-246

[3] Z. Zhang. “A flexible new technique for camera calibration”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330-1334, 2000.