К построению показателей участия автора в системе научной кооперации

Паринов С.И., д.т.н., РАНХиГС и ЦЭМИ РАН

Преамбула

- Мы представляем результаты проекта Сиртек, финансируемого РАНХиГС
- Проект реализуется уже 3 года
- Его цель на 2020 г. - построение на базе контекстов цитирований из научных публикаций комплекса показателей, характеризующих участие автора в системе научной кооперации
- Руководитель проекта – Оксана Медведева, РАНХиГС

Сиртек за 3 года: расширение и развитие данных

план на 2020 – построение системы показателей, характеризующих развитие научного знания

2019 – группировки показателей для авторов

Предложена концепция изучения участия автора в научной кооперации, основанной на публикациях.
Созданы группы показателей для:
- характеристики публикаций автора
- характеристики связей автора с «поставщиками»

2018 - построение показателей на основе контекстов цитирований

Созданы показатели:
1) источники из списков литературы
2) авторы источников,
3) со-цитируемые источники и авторы
4) классификация контекстов цитирований:
- фразы, профессиональные термины и лексические клише
- классы тональности (нейтральный, позитивный, негативный)
5) распределение цитирований по 5-ти равным фрагментам публикаций

2017 - извлечение данных о цитированиях

Создан открытый и пополняемый массив данных о цитированиях научных публикаций

Созданная архитектура системы Сиртек

Что содержат контексты цитирований?

Экономическая литература, в которой так или иначе затрагивается тема мобильности капитала, весьма обширна. Это объясняется тем, что потоки капитала оказывают влияние на самый широкий круг макроэкономических процессов, в том числе на валютные рынки, рынки ценных бумаг, международную торговлю и т. д. [Tamirisa, 1999]. Их необходимо учитывать при выборе денежно-кредитной и бюджетной политики [Obstfeld, Taylor, 1997; Rodrik, 2011]. Значительный отток капитала зачастую становится существенным фактором развития кризисных процессов [Calvo, 1998;

```
<intextref>
 ▼<Prefix>
    Это объясняется тем, что потоки капитала оказывают влияние на самый широкий круг макр
    числе на валютные рынки, рынки ценных бумаг, международную торговлю и т. д. [
  </Prefix>
 ▼<Suffix>
    ]. Их необходимо учитывать при выборе денежнокредитной и бюджетной политики [Obstfe
    2011]. Значительный отток капитала зачастую становится существенным фактором развития
    Dornbusch, 2001].
  </Suffix>
  <Start>6778</Start>
  <End>6791</End>
  <Exact>Tamirisa, 1999</Exact>
  <Reference start="6778" end="6791" exact="Tamirisa, 1999">75</Reference>
</intextref>
<intextref>
```

```
<reference num="75" start="83672"
end="83798" author="Tamirisa"
title="Exchange and capital controls as
barriers to trade International Monetary
Fund Staff Papers 1999 No 1" year="1999">
 ▼<from_pdf>
    Tamirisa N. Exchange and capital
    controls as barriers to trade.
    International Monetary Fund. Staff
    Papers. 1999. No 1. P. 69–88.
  </from_pdf>
</reference>
```

Отношения между участниками научной кооперации: «поставщики» - автор – «потребители»

Группировки исходных данных и показателей по трем группам:

1) собственные публикации автора;

2) публикации цитируемые автором + характеристики его связей с «поставщиками»;

3) публикации цитирующие автора + характеристики связей автора с «потребителями»;

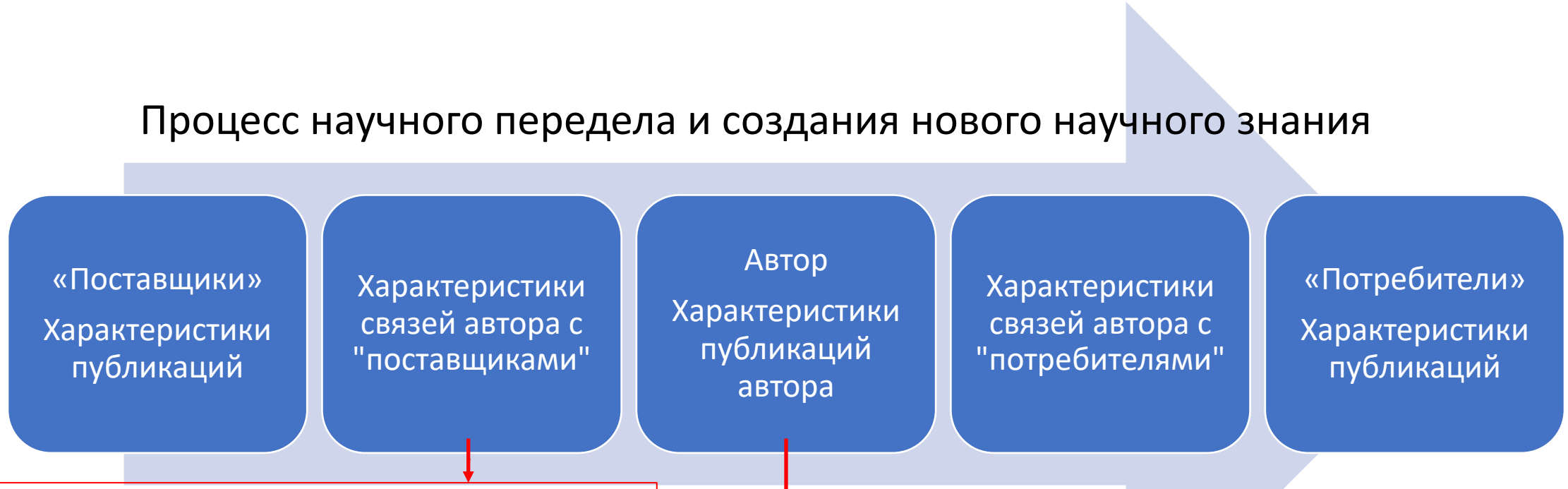
Нужна «теория» для интерпретации получаемых результатов

- Полученные данных характеризуют отношения между авторами цитирующих и цитируемых публикаций
- Для корректной интерпретации данных об этих отношениях нужно ответить на вопрос: Зачем ученым публикации и цитирования?
- Фундаментальная причина – для использования результатов друг друга независимо от географической удаленности ученых, т.е. как инструмент научной кооперации на уровне научного сообщества в целом
- В проекте Сиртек мы создаем комплекс показателей, характеризующих участие ученого в глобальной научной кооперации, основанной на публикациях

Почему такой подход важен?

- Позволяет визуализировать сеть цитирования и кооперации, а также связать ученых цитирующих/использующих результаты друг друга
- Позволяет исследовать развитие научной кооперации (например, какие средства нужны для повышения ее эффективность), включая случаи где публикация является только частным случаем (цифровизация научных коммуникаций и кооперации)
- Позволяет искать комплексное решение для проблемы низкого качества данных о цитированиях (цифровизация цитирований)
- Позволяет анализировать процесс научного передела и, возможно, найти решение проблемы «безбилетника» в науке
- Позволяет радикально улучшить метод оценки научной результативности на базе характеристик участия ученого в научной кооперации и исправить, таким образом, «плохие» мотивации

Показатели научной кооперации автора на основе анализа публикаций



Как различаются цитирования автора:
- кто цитируется и со-цитируется чаще всего
- кто с кем со-цитируется
- по употребляемым терминам и топикам
- по по распределению тональности цитирований
- по распределение цитирований по разделам

Как различаются публикации автора:
- по цитированиям и со-цитированиям
- по употребляемые терминам и топикам
- по тональности цитирований
- по распределению цитирований по разделам

Методы построения показателей

- Источники – группировка одинаковых источников по «автор + год + название» по всем публикациям в Сиртек
- Авторы источников – группировка по фамилиям (однофамильцев пока не различаем) по всем публикациям в Сиртек
- Со-цитирования источников и авторов – выделение рядом стоящих сносок на источники (рядом = любые 7 символов между ними) по всему Сиртеку
- Повторяющиеся фразы – n-gramms для выделения фраз в 2-6 слов в исходном и леммитизированном виде для всех контекстов цитирований
- Топики – тематическое моделирование для 20 топиков, каждый топик- 5 леммитизированных слов, для всех контекстов цитирований
- Тональность – три основных класса (позитивный, нейтральный, негативный) определены сравнением контекстов цитирований с размеченным словарем с ресурса http://linis-crowd.org/
- Распределение по «телу» публикации – полный текст от начала до списка литературы разбивается на 5 равных частей, определяется в какую часть попадает каждый контекст цитирования

Характеристики публикации заданного автора по следующим параметрам:

- а) источники, распознанные в списках литературы заданного автора, которые очищены от повторов и для которых посчитана различная статистика;
- б) распознанные авторы источников, включая со-цитируемых авторов, которые очищены от повторов и для них посчитана различная статистика;
- в) повторяющиеся фразы, включая профессиональные термины и лексические клише из контекстов цитирований, которые выделены из 3-х групп публикаций, связанных с заданным автором;
- г) топики, построенные методом тематического моделирования на основе содержания контекстов цитирований, которые выделены из 3-х групп публикаций, связанных с заданным автором;
- д) классы тональности (нейтральный, позитивный, негативный), распознанные на основе содержания контекстов цитирований;
- е) распределение цитирований по 5-ти равным фрагментам публикаций заданного автора
- Это позволяет охарактеризовать научный продукт автора, включая временные изменения, тематическое развитие автора (например, из самоцитирований) и т.п.

Примеры характеристик публикаций

- http://onir2.ranepa.ru:8081/prl/after_table_edited/content.html

- Основные задачи:
  - создать комплексное представление научной продукции автора, включая изменения во времени (список поставщиков и их влияние, основные профессиональные термины и топики, тематическая траектория автора, характер его научного передела и т.п.)
  - создать возможности для углубленного анализа показателей научной продукции для отдельного автора, сравнение этих показателей для группы авторов и т.п.

Характеристики цитирований заданного автора

- Эти показатели характеризуют связи заданного автора с публикациями и их авторами, которые являются для него “поставщиками” научной продукции, использованной (процитированной) им в процессе создания своих публикаций
- Показатели группируются разными способами на основе их принадлежности к одинаковым контекстам цитирований. Способы группировки показателей основываются на переборе возможных сочетаний друг с другом их основных типов:
  - источники и их авторы, включая со-цитируемых авторов;
  - повторяющиеся фразы и топики;
  - классы тональности;
  - распределение по 5-ти фрагментам
- Это позволяет охарактеризовать связи автора с поставщиками, включая степень и характер влияния поставщиков на научной продукцию автора

Примеры характеристик цитирований

- http://onir2.ranepa.ru:8081/prl/after_table_edited/content.html

- Основные задачи:
  - Создание комплексного представления влияния поставщиков на продукцию автора, включая изменение характеристик этого влияния во времени (кто имеет/имел наибольшее влияние, какой характер этого влияния, как он менялся во времени и т.п.)
  - Создание возможностей для углубленного анализа связей автора с поставщиками, сравнение этих показателей для группы авторов и т.п.

Значимые для научного сообщества перспективы развития данного подхода

- Развитие средств научной кооперации, основанной на публикациях (как можно повысить ее эффективность)
- Анализ процесса научного передела. Решение проблемы «безбилетника» в науке
- Оценка научной результативности на основе характеристик участия ученого в научной кооперации. Создание «правильных» мотиваций для ученых

Перспективы развития: цифровизация процесса создания нового научного знания

Предложения к сотрудничеству

- Мы ищем партнеров для развития идей, методов и результатов проекта Сиртек, а также для их превращения в значимые для всего научного сообщества инновации
- В частности, предлагаем использовать полученные в Сиртек данные для выполнения курсовых и диссертационных работ студентов, для проведения научных исследований и т.д.
- Мы ищем специалистов, в частности, по методам компьютерной лингвистики, для оплачиваемого (по договорам подряда) выполнения некоторых работ по проекту Сиртек

Актуальные задачи

- Улучшение распознавания списков литературы, включая разбор референсов и сносок на референсы, более точное определение авторов (например, методом «именованных сущностей») и т.п.
- Классификации контекстов цитирований с использованием семантической разметки корпуса русского языка, разработанной в Секторе теоретической семантики Института русского языка РАН
- Улучшение распознавания структуры публикаций, включая ее заголовки, разделы, нетекстовые элементы и т.п.

Нужны специалисты для участия в проекте

- Обработка текстовых данных (повышение качества данных о цитированиях в Сиртек)
- Развитие базы данных Сиртек с показателями по авторам и создание требуемых запросов и выборок
- Классификация текстов публикаций, включая контексты цитирований (с помощь семантической разметки корпуса русского языка и др.)
- Конструирование характеристик и показателей на основе данных о цитированиях и содержания публикаций для различных типов потребителей
- Разработка веб представлений для созданных показателей, включая средства их анализа
- Развитие методологии проекта и его приложений, в т.ч. для совершенствования оценки научной результативности и т.п.

Контакты

- Оксана Медведева, руководитель проекта Сиртек - oxana.medvedeva.1984@gmail.com

- Сергей Паринов, руководитель группы разработчиков Сиртек – sparinov@gmail.com