ПРОБЛЕМЫ ПЕРЕДАЧИ ИНФОРМАЦИИ

Том 49 2013 Вып. 3

УДК 621.391.1 : 519.1

© 2013 г. **Л.Г. Митюшин**

МОДЕЛЬ СЛУЧАЙНОГО ОБЪЕДИНЕНИЯ ОТРЕЗКОВ

Рассматривается растущая совокупность отрезков с целочисленными концами на прямой, в которой каждая пара смежных отрезков с вероятностью q порождает новый отрезок – объединение исходных отрезков. В начальный момент присутствуют отрезки длины от 1 до m. Пусть h_n – вероятность возникновения отрезка $[a, a+n]$; критическое значение $q_c(m)$ определяется как $\sup\{q \mid \lim\limits_{n\to\infty} h_n = 0\}$. Получены нижняя и верхняя оценки для $q_c(m)$.

§ 1. Введение

Рассматриваются отрезки с целочисленными концами на прямой, "вступающие в реакцию" друг с другом. Реагировать могут любые два смежных отрезка (т.е. имеющие единственную общую точку). В результате реакции с вероятностью q возникает новый отрезок – объединение исходных отрезков; исходные отрезки при этом сохраняются. Для всех реакций вероятность q одна и та же, и реакции происходят независимо друг от друга. Одинаковые отрезки, возникшие в результате разных реакций, отождествляются. Новые отрезки вступают в реакцию как с отрезками, существовавшими ранее, так и между собой; каждая пара смежных отрезков реагирует один раз.

Данная модель возникла в связи с задачей синтаксического анализа текстов на естественном языке. Некоторым отрезкам текста (частям предложений или целым предложениям) могут быть приписаны синтаксические структуры, представляющие грамматические отношения между словами и/или группами слов в этом отрезке. Чаще всего используются два вида синтаксических структур: деревья составляющих и деревья зависимостей (см., например, [1, 2]). В этих деревьях вершинам приписаны метки грамматических категорий (и, возможно, некоторые другие), а дугам могут быть приписаны метки синтаксических отношений. Синтаксические структуры должны удовлетворять определенным требованиям, называемым "правилами грамматики". Цель синтаксического анализа – найти синтаксическую структуру или структуры, соответствующие заданному отрезку текста. Отметим, что для реальных синтаксических структур характерна неединственность: если на некотором отрезке длины n можно построить хотя бы одну структуру, то на нем "в среднем" можно построить порядка $\exp(Cn)$ различных структур.

Существенным свойством синтаксических структур является возможность строить их рекурсивно. Пусть вначале известны структуры, соответствующие отрезкам длины 1, т.е. отдельным словам. Затем структуры, построенные на смежных отрезках, используются для построения структур на объединении этих отрезков. В случае деревьев составляющих и зависимостей дерево, представляющее структуру на объединении отрезков, включает исходные структуры в качестве поддеревьев, а также содержит дополнительные элементы, объединяющие их в единое дерево.

В работе [3] предложен подход к синтаксическому анализу, основанный на идее локального взаимодействия. Предполагается, что на синтаксических структурах определена функция, оценивающая их лингвистическое качество. Имеется совокупность U построенных синтаксических структур, вначале состоящая из структур на отрезках длины 1. Затем пары содержащихся в U структур на смежных отрезках взаимодействуют друг с другом "как молекулы в пробирке", и если для данной пары можно построить синтаксическую структуру на объединенном отрезке, она включается в U с некоторой вероятностью q, задаваемой внешним управляющим механизмом, причем q тем больше, чем выше значение функции качества для новой структуры. В результате для заданного предложения достаточно быстро строится оптимальная или субоптимальная синтаксическая структура.

Модель, рассматриваемая в данной статье, представляет собой максимальное упрощение этой ситуации. Она не претендует на какое-либо лингвистическое содержание и исследуется как чисто математический объект. Возможны некоторые естественные обобщения – например, вероятность q порождения нового отрезка может зависеть от его длины n, один из вариантов такой зависимости: $q = \min(1, cn^{-d})$, где $c > 0$ и $0 < d \leqslant 1$ – параметры. Другая возможность – модель с многими типами отрезков, в которой отрезки типа α и β порождают объемлющий отрезок типа γ с вероятностью $q_{\alpha\beta\gamma}$. В этой схеме также может быть введена зависимость вероятности порождения от длины отрезка.

Вернемся к исходной модели. Удобно описывать процесс образования отрезков как происходящий в дискретном времени, где в момент n возникают отрезки длины n. Пусть s_n^a – индикатор события "в системе имеется отрезок длины n с левым концом a". Предположим, что в начальный момент присутствуют все отрезки длины от 1 до m, и только они. Тогда $s_n^a = 1$, $n \leqslant m$, $a \in \mathbb{Z}$; значения s_n^a, $n > m$, определяются следующим образом. Пусть ξ_n^{ab}, $n > 1$, $a, b \in \mathbb{Z}$, $a < b < a + n$, – независимые в совокупности случайные величины, принимающие значение 1 с вероятностью q и 0 с вероятностью $1 - q$. Величина ξ_n^{ab} определяет результат реакции между отрезками $[a, b]$ и $[b, a+n]$. Значения s_n^a определяются последовательно для $n = m+1, m+2, \ldots$ следующим образом:

$$s_n^a = \max_{0<k<n} s_k^a s_{n-k}^{a+k} \xi_n^{a,a+k}, \tag{1}$$

что соответствует $n - 1$ возможности построения отрезка длины n из двух смежных отрезков.

Пусть m фиксировано и $h_n = \mathbf{P}(s_n^a = 1)$ (не зависит от a). Нас интересует предельное поведение h_n при различных значениях q. Легко показать, что h_n как функция q не убывает. Для данного m определим критическое значение $q_c(m) = \sup\{q \mid \lim\limits_{n\to\infty} h_n = 0\}$.

Следующие утверждения дают нижнюю и верхнюю оценки для $q_c(m)$.

Теорема 1. *Имеют место неравенства* $q_c(1) \geqslant \dfrac{1}{4}$ *и*

$$q_c(m) \geqslant \bigl(1 - \sqrt{(m+1)/2m}\bigr)/(m-1), \quad m > 1.$$

Теорема 2. *Имеет место неравенство* $q_c(m) \leqslant 1 - 3^{-1/r} < \dfrac{\ln 3}{r}$, *где* $r = [(m+1)/2]$.

Моделирование методом Монте-Карло показывает, что при больших n зависимость h_n от q близка к ступенчатой функции со значениями 0 и 1. В таблице приводятся данные для $m = 1$; $\overline{h}_{1000}$ – среднее значение s_{1000}^0, полученное в 100 испытаниях.

Гипотеза. *Для любого* m *и* $q > q_c(m)$ *имеет место равенство* $\lim\limits_{n\to\infty} h_n = 1$.

Таблица

q	$\overline{h}_{1000}$
$\leqslant 0{,}34$	0,00
0,35	0,10
0,36	0,59
0,37	0,98
$\geqslant 0{,}38$	1,00

Используя методы, близкие к доказательству теоремы 2, можно получить следующее более слабое утверждение. Как и выше, $r = [(m+1)/2]$.

Т е о р е м а 3. *Если* $q > \dfrac{2}{3}$ *при* $r = 1, 2$ *или* $q > \dfrac{7}{4} \cdot \dfrac{\ln r}{r}$ *при* $r > 2$, *то* $\lim\limits_{n\to\infty} h_n = 1$.

Следующие параграфы содержат доказательства теорем 1 и 2. Доказательство теоремы 3 не приводится.

§ 2. Доказательство теоремы 1

Из (1) следует

$$s_n^a \leqslant \sum_{k=1}^{n-1} s_k^a s_{n-k}^{a+k} \xi_n^{a,a+k}.$$

В каждом слагаемом в правой части сомножители s_k^a, s_{n-k}^{a+k} и $\xi_n^{a,a+k}$ являются независимыми случайными величинами, так как образование отрезков $[a, a+k]$ и $[a+k, a+n]$ определяется непересекающимися наборами вспомогательных случайных величин ξ_i^{cd}, и эти наборы не включают $\xi_n^{a,a+k}$. Поэтому при $n > m$

$$h_n = \mathbf{E}\, s_n^a \leqslant \sum_{k=1}^{n-1} \mathbf{E}\, s_k^a s_{n-k}^{a+k} \xi_n^{a,a+k} = \sum_{k=1}^{n-1} \mathbf{E}\, s_k^a\, \mathbf{E}\, s_{n-k}^{a+k}\, \mathbf{E}\, \xi_n^{a,a+k} = \sum_{k=1}^{n-1} h_k h_{n-k} q.$$

При $1 \leqslant n \leqslant m$ имеем $h_n = 1$.

Пусть величины g_n задаются равенствами

$$g_n = \sum_{k=1}^{n-1} g_k g_{n-k} q, \quad n > m, \tag{2}$$

$$g_n = 1, \quad 1 \leqslant n \leqslant m. \tag{3}$$

Тогда $h_n \leqslant g_n$, $n \geqslant 1$. Индукцией по n легко показать, что $g_n \leqslant C_{n-1}$, где $\{C_n\}$, $n \geqslant 0$, – числа Каталана, удовлетворяющие соотношениям $C_0 = 1$ и $C_{n+1} = \sum\limits_{k=0}^{n} C_k C_{n-k}$, $n \geqslant 0$.

Пусть $f(x) = \sum\limits_{n=1}^{\infty} g_n x^n$ – производящая функция последовательности $\{g_n\}$. Поскольку $C_n \leqslant 4^n$, этот ряд имеет радиус сходимости не меньше $\dfrac{1}{4}$. Из (2) получаем

$$g_n x^n = q \sum_{k=1}^{n-1} (g_k x^k)(g_{n-k} x^{n-k}), \quad n > m. \tag{4}$$

Из (3) получаем

$$g_n x^n = q \sum_{k=1}^{n-1} (g_k x^k)(g_{n-k} x^{n-k}) + (1 - q(n-1)) x^n, \quad 1 \leqslant n \leqslant m. \tag{5}$$

Суммируя (4) и (5) для всех $n \geqslant 1$, получаем для $f(x)$ уравнение

$$f(x) = qf^2(x) + w(x),$$

где $w(x) = \sum\limits_{k=1}^{m} (1 - q(k-1))x^k$. Отсюда

$$f(x) = \frac{1 - \sqrt{1 - 4qw(x)}}{2q} \tag{6}$$

(минус перед корнем выбран потому, что $f(0) = 0$).

Далее мы считаем, что $q \leqslant 1/(m-1)$ при $m \geqslant 3$, тогда коэффициенты в $w(x)$ неотрицательны и $\max\limits_{|x| \leqslant 1} |w(x)| = w(1)$. Если $4qw(1) < 1$, то функция, определяемая равенством (6), является аналитической в некотором круге радиуса $R > 1$, откуда следует, что в этом круге ряд $\sum\limits_{n=1}^{\infty} g_n x^n$ сходится, и следовательно, $\lim\limits_{n\to\infty} g_n = 0$ и $\lim\limits_{n\to\infty} h_n = 0$. Поскольку $4qw(1) = 4qm - 2q^2m(m-1)$, мы получаем, что в случае $m = 1$ для сходимости h_n к 0 достаточно $q < 1/4$, а в случае $m > 1$ для этого достаточно $q < q_0$, где $q_0 = (1 - \sqrt{(m+1)/2m})/(m-1)$ – меньший из двух корней уравнения $4qm - 2q^2m(m-1) = 1$.

§ 3. Доказательство теоремы 2

Пусть сначала $m = 1$. Рассмотрим процесс образования отрезков, аналогичный исходному, но в котором разрешаются только такие реакции, где длина хотя бы одного из отрезков равна 1. Пусть t_n^a – индикатор события "в системе имеется отрезок длины n с левым концом a" для этого процесса. Тогда $t_1^a = 1$ при всех a; как и s_n^a, величины t_n^a при $n > 1$ можно определить рекуррентно как функции от вспомогательных случайных величин ξ_i^{cd}:

$$t_n^a = \max\bigl(t_{n-1}^{a+1}\xi_n^{a,a+1}, t_{n-1}^a \xi_n^{a,a+n-1}\bigr). \tag{7}$$

Легко доказать индукцией по n, что $s_n^a \geqslant t_n^a$ как функции от $\{\xi_i^{cd}\}$ при всех a и n, откуда следует, что $\mathbf{P}(s_n^a = 1) \geqslant \mathbf{P}(t_n^a = 1)$.

Положим $\zeta_{n0}^a = \xi_n^{a,a+n-1}$ при $n > 1$ и $\zeta_{n1}^a = \xi_n^{a,a+1}$ при $n > 2$. Тогда (7) можно переписать в виде

$$t_n^a = \max\bigl(t_{n-1}^a \zeta_{n0}^a, t_{n-1}^{a+1}\zeta_{n1}^a\bigr), \quad n > 2, \quad \text{и} \quad t_2^a = \zeta_{20}^a.$$

Фактически мы имеем дело с моделью перколяции по дугам на графе с множеством вершин $\mathbb{Z}^2$, где из каждой вершины (a, n) идут дуги в вершины $(a, n-1)$ и $(a+1, n-1)$. Будем считать, что независимые случайные величины $\zeta_{n\,d}^a$ со значениями 0 и 1 определены при всех $a, n \in \mathbb{Z}, d \in \{0, 1\}$, и $\mathbf{P}(\zeta_{n\,d}^a = 1) = q$. Дуга между (a, n) и $(a+d, n-1)$ считается открытой, если $\zeta_{n\,d}^a = 1$, и закрытой в противном случае. Индукцией по n легко показать, что $t_n^a = 1$ эквивалентно тому, что существует путь по открытым дугам, соединяющий точку (a, n) и некоторую точку $(b, 2)$, при дополнительном условии, что дуга от $(b, 2)$ к $(b, 1)$ является открытой.

Рассматриваемый граф изоморфен ориентированной квадратной решетке $\overrightarrow{\mathbb{Z}}^2$. Известно, что при $q > \dfrac{2}{3}$ на $\overrightarrow{\mathbb{Z}}^2$ с вероятностью $v(q) > 0$ существует бесконечный путь по открытым дугам, начинающийся в заданной точке (см. [4], а также [5, гла-

ва 6, теорема 6]). Следовательно, если $q > \frac{2}{3}$, то

$$h_n = \mathbf{P}(s_n^a = 1) \geqslant \mathbf{P}(t_n^a = 1) \geqslant qv(q) > 0, \quad n > 1,$$

т.е. $q \geqslant q_c(1)$. Отсюда следует $q_c(1) \leqslant \frac{2}{3}$, а также $q_c(2) \leqslant \frac{2}{3}$ (так как $q_c(m)$ не возрастает по m).

Доказательство при $r = [(m+1)/2] > 1$ следует той же схеме, но вместо решетки на $\mathbb{Z}^2$ рассматривается "укрупненная" решетка, вершинами которой являются параллелограммы $\Delta \subset \mathbb{Z}^2$, содержащие r^2 точек. Для $(c,k) \in \mathbb{Z}^2$, $k > 0$, положим

$$\Delta_k^c = \big\{(a,n) \in \mathbb{Z}^2 \mid cr \leqslant a < (c+1)r,\ (c+k)r \leqslant a+n < (c+k+1)r\big\} =$$
$$= \Delta_1^0 + r(c, k-1).$$

Для $(a,n) \in \Delta_k^c$, $k > 1$, положим (см. рисунок)

$$D_0(a,n) = \{(b,l) \in \Delta_{k-1}^c \mid b = a\},$$
$$D_1(a,n) = \{(b,l) \in \Delta_{k-1}^{c+1} \mid b + l = a + n\}.$$

Аналогично, для $(b,l) \in \Delta_{k-1}^c$, $k > 1$, положим

$$D_0'(b,l) = \{(a,n) \in \Delta_k^c \mid a = b\}$$

и для $(b,l) \in \Delta_{k-1}^{c+1}$, $k > 1$,

$$D_1'(b,l) = \{(a,n) \in \Delta_k^c \mid a + n = b + l\}.$$

Легко видеть, что $(b,l) \in D_0(a,n)$ тогда и только тогда, когда $(a,n) \in D_0'(b,l)$, и аналогично для D_1 и D_1'.

Рассмотрим семейство величин t_n^a со значениями 0 и 1, которые определяются аналогично (1), но максимум берется не по всем членам, присутствующим в (1). А именно положим при $n > m$

$$t_n^a = \max\Big(\max_{\{k \mid (a,k) \in D_0(a,n)\}} t_k^a t_{n-k}^{a+k} \xi_n^{a,a+k}, \max_{\{k \mid (a+k,n-k) \in D_1(a,n)\}} t_k^a t_{n-k}^{a+k} \xi_n^{a,a+k}\Big). \tag{8}$$

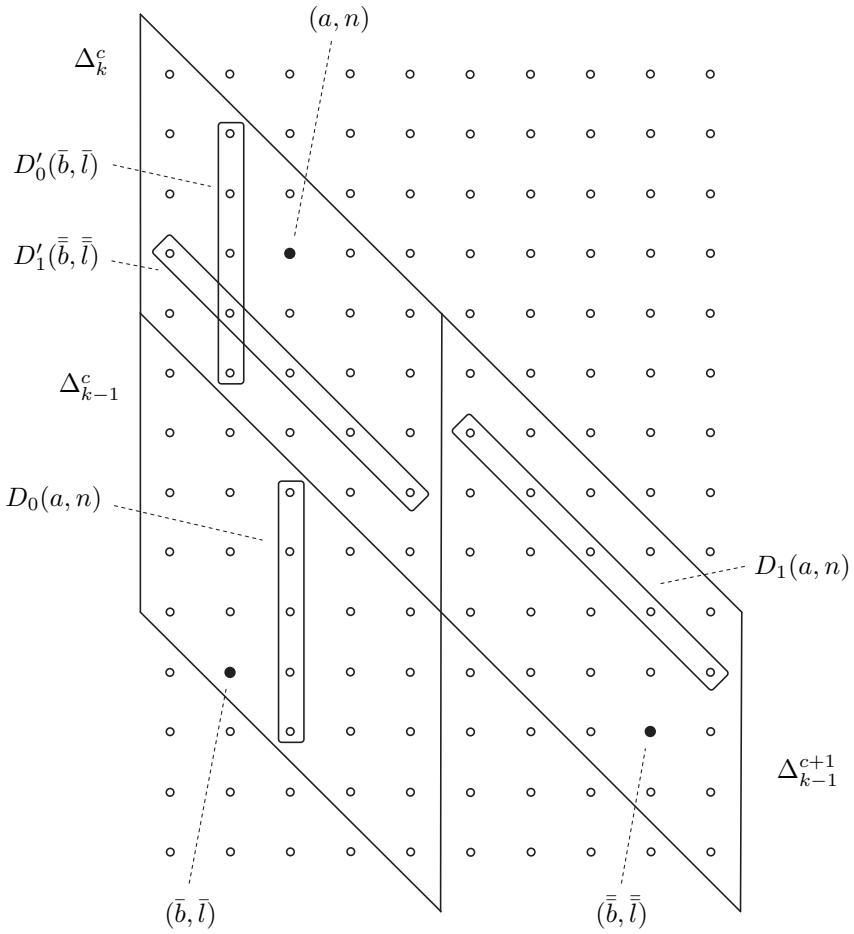
При $n \leqslant m$ полагаем $t_n^a = s_n^a = 1$. Индукцией по n (начиная с $n = m+1$) доказывается, что $s_n^a \geqslant t_n^a$ при всех a и n.

Если $(a,k) \in D_0(a,n)$, то $n - k \leqslant m$ и $t_{n-k}^{a+k} = 1$. Аналогично, если $(a+k, n-k) \in \in D_1(a,n)$, то $k \leqslant m$ и $t_k^a = 1$. Поэтому (8) можно переписать в виде

$$t_n^a = \max\Big(\max_{\{k \mid (a,k) \in D_0(a,n)\}} t_k^a \xi_n^{a,a+k}, \max_{\{k \mid (a+k,n-k) \in D_1(a,n)\}} t_{n-k}^{a+k} \xi_n^{a,a+k}\Big) =$$
$$= \max\Big(\max_{(b,l) \in D_0(a,n)} t_l^b \xi_n^{a,a+l}, \max_{(b,l) \in D_1(a,n)} t_l^b \xi_n^{ab}\Big).$$

Значения t_n^a при $(a,n) \in \Delta_k^c$ определяются через значения t_l^b при $(b,l) \in \Delta_{k-1}^c \cup \cup \Delta_{k-1}^{c+1}$ и значения вспомогательных случайных величин ξ_n^{ad}. Случайные величины ξ_n^{ad} независимы в совокупности, откуда следует, что случайные величины t_n^a, $(a,n) \in \Delta_k^c$, условно независимы в совокупности при известных t_l^b, $(b,l) \in \Delta_{k-1}^c \cup \cup \Delta_{k-1}^{c+1}$.

Обозначим через $T_j^d = \{t_n^a \mid (a,n) \in \Delta_j^d\}$ состояние подмножества Δ_j^d, через Z – пространство значений T_j^d, т.е. множество r^2-мерных векторов с компонентами 0 и 1, и через $\mathbf{0} \in Z$ – вектор, все компоненты которого равны 0.



Множества Δ, D и D' $(r=5)$

Если $T_{k-1}^c = x \in Z$, где $x \neq \mathbf{0}$, то $t_l^b = 1$ для некоторого $(b,l) \in \Delta_{k-1}^c$ и $\mathbf{P}(t_n^a = 1) \geqslant$ $\geqslant \mathbf{P}(\xi_n^{a,a+l} = 1) = q$ для каждого $(a,n) \in D_0'(b,l)$. Поскольку $|D_0'(b,l)| = r$, получаем $\mathbf{P}(T_k^c \neq \mathbf{0} \mid T_{k-1}^c = x) \geqslant 1-(1-q)^r = q_0$. Аналогично, если $T_{k-1}^{c+1} = y \neq \mathbf{0}$, то $\mathbf{P}(T_k^c \neq \mathbf{0} \mid T_{k-1}^{c+1} = y) \geqslant q_0$.

Рассмотрим для каждого $x \in Z$ случайную величину $\eta_x = \{t_{n0}^a \mid (a,n) \in \Delta_k^c\}$ со значениями в Z, где $t_{n0}^a = \max\limits_{(b,l)\in D_0(a,n)} t_l^b \xi_n^{a,a+l}$ – первая компонента правой части (9) при $T_{k-1}^c = x$. Аналогично, рассмотрим случайную величину $\theta_y = \{t_{n1}^a \mid (a,n) \in \Delta_k^c\}$ со значениями в Z, где $t_{n1}^a = \max\limits_{(b,l)\in D_1(a,n)} t_l^b \xi_n^{ab}$ – вторая компонента правой части (9) при $T_{k-1}^{c+1} = y$. Тогда для заданных x и y получаем $T_k^c = \max(\eta_x, \theta_y)$, где max обозначает покомпонентный максимум двух r^2-мерных векторов.

Из сказанного выше следует, что если $x \neq \mathbf{0}$, то $\mathbf{P}(\eta_x \neq \mathbf{0}) \geqslant q_0$. Это позволяет представить η_x в виде $\zeta_0 \bar{\eta}_x + (1-\zeta_0)\bar{\bar{\eta}}_x$, $x \in Z$, где ζ_0 – не зависящая от x случайная величина со значениями 0 и 1, для которой $\mathbf{P}(\zeta_0 = 1) = q_0$, а $\bar{\eta}_x$ и $\bar{\bar{\eta}}_x$ – случайные величины со значениями в Z, причем $\mathbf{P}(\bar{\eta}_x = \mathbf{0}) = 0$ при $x \neq \mathbf{0}$. Аналогично, θ_y можно

представить в виде $\zeta_1\bar{\theta}_y + (1-\zeta_1)\bar{\bar{\theta}}_y$, $y \in Z$, где ζ_1 не зависит от y и имеет такое же распределение, как ζ_0, и $\mathbf{P}(\bar{\theta}_y = \mathbf{0}) = 0$ при $y \neq \mathbf{0}$.

Рассматриваемые случайные величины различны для различных k и c. Равенство для T_k^c принимает вид

$$T_k^c = \max\bigl(\zeta_{k0}^c\bar{\eta}_{kx}^c + (1-\zeta_{k0}^c)\bar{\bar{\eta}}_{kx}^c, \zeta_{k1}^c\bar{\theta}_{ky}^c + (1-\zeta_{k1}^c)\bar{\bar{\theta}}_{ky}^c\bigr), \tag{9}$$

причем все случайные величины $\zeta_{k0}^c, \bar{\eta}_{kx}^c, \bar{\bar{\eta}}_{kx}^c, \zeta_{k1}^c, \bar{\theta}_{ky}^c, \bar{\bar{\theta}}_{ky}^c$ независимы в совокупности.

Рассмотрим решетку с вершинами Δ_k^c, в которой вершина Δ_k^c связана дугами с Δ_{k-1}^c и Δ_{k-1}^{c+1}. Будем считать дугу между Δ_k^c и Δ_{k-1}^{c+d} открытой, если $\zeta_{kd}^c = 1$, и закрытой в противном случае. Из (10) следует, что если $\zeta_{k0}^c = 1$ и $T_{k-1}^c \neq \mathbf{0}$ или $\zeta_{k1}^c = 1$ и $T_{k-1}^{c+1} \neq \mathbf{0}$, то $\mathbf{P}(T_k^c \neq \mathbf{0}) = 1$. Отметим, что $T_2^c \neq \mathbf{0}$ для всех c. Индукцией по k легко показать, что для того чтобы $T_k^c \neq \mathbf{0}$ выполнялось почти наверное, достаточно, чтобы существовал путь по открытым дугам между Δ_k^c и некоторым Δ_2^b, где $b \in \mathbb{Z}$.

При $q_0 > \dfrac{2}{3}$ вероятность существования такого пути не меньше $v(q_0) > 0$, откуда $\mathbf{P}(T_k^c \neq \mathbf{0}) \geqslant v(q_0)$ при всех k, и $\lim\limits_{n\to\infty} h_n = 0$ не выполняется. Поэтому равенство $q_0 = 1-(1-q)^r = \dfrac{2}{3}$ дает требуемую верхнюю оценку для $q_c(m)$.

Автор благодарен рецензенту за полезные замечания.

СПИСОК ЛИТЕРАТУРЫ

1. *Гладкий А.В.* Синтаксические структуры естественного языка в автоматизированных системах общения. М.: Наука, 1985.
2. *Manning C.D., Schütze H.* Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press, 1999.
3. *Kempen G., Vosse T.* Incremental Syntactic Tree Formation in Human Sentence Processing: a Cognitive Architecture Based on Activation Decay and Simulated Annealing // Connect. Sci. 1989. V. 1. № 3. P. 273–290.
4. *Liggett T.M.* Survival of Discrete Time Growth Models, with Applications to Oriented Percolation // Ann. Appl. Probab. 1995. V. 5. № 3. P. 613–636.
5. *Bollobás B., Riordan O.* Percolation. Cambridge: Cambridge Univ. Press, 2006.

Митюшин Леонид Григорьевич
Институт проблем передачи информации
им. А.А. Харкевича РАН
mit@iitp.ru

Поступила в редакцию
13.11.2012
После переработки
06.02.2013