ДИСКРЕТНЫЙ АНАЛИЗ И ИССЛЕДОВАНИЕ ОПЕРАЦИЙ
******** 2000. Том 55, № 10. С. 3–29

УДК 999.9

О СВОЙСТВАХ ВЕРОЯТНОСТНЫХ ХАРАКТЕРИСТИК ДЕРЕВЬЕВ ВЫВОДА В РАЗЛОЖИМЫХ СТОХАСТИЧЕСКИХ КС-ГРАММАТИКАХ

Л. П. Жильцова[1], *И. М. Мартынов*[1]

[1] Нижегородский государственный университет им. Н. И. Лобачевского, пр. Гагарина, X, XXXXXX Нижний Новгород, Россия,

E-mail: larzhil@rambler.ru, murbidodrus@gmail.com

Аннотация. !!!ПРОВЕРИТЬ!!! При исследовании возможностей экономного кодирования слов, структурные и вероятностные свойства которых моделируются стохастической КС-грамматикой, в качестве меры эффективности кодирования рассматривается стоимость кодирования, которая определяется на множестве "‘длинных"’ слов. В качестве множества таких слов целесообразно рассматривать множество слов КС-языка, имеющих деревья вывода фиксированной высоты t при $t \to \infty$. При этом возникает необходимость в вычислении математических ожиданий числа применений различных правил КС-грамматики в словах языка, имеющих дерево вывода высоты t.

Ключевые слова: теория кодирования, сжатие данных, КС-грамматики, деревья вывода, ...

Введение

Автором в [4, 5] рассматривались вопросы, связанные с кодированием сообщений, являющихся словами стохастического контекстно-свободного языка (стохастического КС-языка), при условии, что матрица первых моментов грамматики неразложима, непериодична, и ее максимальный по модулю собственный корень (перронов корень) строго меньше единицы (докритический случай). При неразложимой матрице первых моментов нетерминальные символы грамматики образуют один класс. В настоящей работе рассматриваются стохастические КС-грамматики с произвольным числом классов нетерминальных символов.

При исследовании возможностей экономного кодирования слов, структурные и вероятностные свойства которых моделируются стохастической КС-грамматикой, в качестве меры эффективности кодирования рас-

сматривается стоимость кодирования, которая определяется на множестве „длинных“ слов. В качестве множества таких слов целесообразно рассматривать множество слов КС-языка, имеющих деревья вывода фиксированной высоты t при $t \to \infty$. При этом возникает необходимость в вычислении математических ожиданий числа применений различных правил КС-грамматики в словах языка, имеющих дерево вывода высоты t. (добавить)

Основные определения и предварительные сведения

Для изложения результатов о контекстно-свободных языках будем использовать определения КС-языка и стохастического КС-языка из [1, 9].

Стохастической КС-грамматикой называется система $G = \langle V_T, V_N, R, s\rangle$, где V_T и V_N - конечные множества терминальных и нетерминальных символов (терминалов и нетерминалов) соответственно; $s \in V_N$ - аксиома, R – множество правил. Множество R можно представить в виде $R = \cup_{i=1}^{k} R_i$, где k - мощность алфавита V_N и $R_i = \{r_{i1}, \ldots, r_{i,n_i}\}$. Каждое правило r_{ij} из R_i имеет вид

$$r_{ij} : A_i \overset{p_{ij}}{\to} \beta_{ij},\ j = 1, ..., n_i,$$

где $A_i \in V_N, \beta_{ij} \in (V_T \cup V_N)^*$ и p_{ij} - вероятность применения правила r_{ij}, удовлетворяющая следующим условиям:

$$0 < p_{ij} \leqslant 1 \text{ и } \sum_{j=1}^{n_i} p_{ij} = 1.$$

Для слов α и β из $(V_T \cup V_N)^*$ будем говорить, что β непосредственно выводимо из α (и записывать $\alpha \Rightarrow \beta$), если $\alpha = \alpha_1 A_i \alpha_2$, $\beta = \alpha_1 \beta_{ij} \alpha_2$ для некоторых $\alpha_1, \alpha_2 \in (V_T \cup V_N)^*$, и в грамматике G имеется правило $A_i \overset{p_{ij}}{\to} \beta_{ij}$.

Обозначим через $\Rightarrow_*$ рефлексивное транзитивное замыкание отношения $\Rightarrow$. КС-язык, порождаемый грамматикой G, определяется как множество слов $L_G = \{\alpha : s \Rightarrow_* \alpha, \alpha \in V_T^*\}$. В работе рассматриваются бесконечные языки.

Каждому слову α КС-языка соответствует последовательность правил грамматики (вывод), с помощью которой α выводится из аксиомы s. Вероятность вывода определяется как произведение вероятностей правил, образующих вывод. Вероятность слова α определяется как сумма

вероятностей всех различных левых выводов слова α (при левом выводе очередное правило применяется к самому левому нетерминалу в слове).

Грамматика G называется согласованной, если

$$\lim_{N\to\infty} \sum_{\alpha\in L, |\alpha|\leqslant N} p(\alpha) = 1$$

(здесь $|x|$ - длина слова x). В работе рассматриваются согласованные КС-грамматики. Согласованная КС-грамматика G индуцирует распределение вероятностей P на множестве слов порождаемого КС-языка L.

Левому выводу соответствует дерево вывода, которое строится следующим образом. Корень дерева помечается аксиомой s. Пусть при выводе слова α на очередном шаге в процессе левого вывода применяется правило $A_i \overset{p_{ij}}{\to} b_{i_1} b_{i_2} \dots b_{i_m}$, где $b_{i_l} \in V_N \cup V_T \ \ (l = 1, \dots, m)$. Тогда из самой левой вершины-листа дерева, помеченной символом A_i (при обходе листьев дерева слева направо), проводится m дуг в вершины следующего яруса, которые помечаются слева направо символами $b_{i_1}, b_{i_2}, \dots, b_{i_m}$ соответственно. Если правило грамматики имеет вид $A_i \overset{p_{ij}}{\to} \lambda$, где $\lambda-$ пустое слово, в следующий ярус проводится одна дуга, и новая вершина помечается λ.

После построения дуг и вершин для всех правил грамматики в выводе слова языка все листья дерева помечены символами из $A_t \cup \{\lambda\}$ и само слово получается при обходе кроны дерева слева направо. Высотой дерева называется максимальная длина пути от корня к листу. Вероятность дерева вывода определяется как вероятность соответствующего ему левого вывода.

На Рис. 1 показано дерево вывода для одного из слов языка Дика (языка правильных скобочных последовательностей). Грамматика этого языка задаётся алфавитами $V_N = \{S\}$, $V_T = \{(,)\}$, аксиомой S и правилами $S \to (S)S$, $S \to \lambda$. Построенное дерево соответствует слову (())() в задаваемом грамматикой языке.

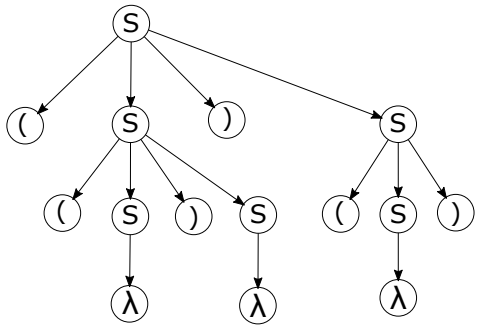

Рис. 1. Дерево вывода

Ярусы дерева будем нумеровать следующим образом. Корень дерева расположен в нулевом ярусе. Вершины дерева, смежные с корнем, образуют первый ярус, и т.д. Дуги, выходящие из вершин $j-$го яруса, ведут к вершинам $(j+1)-$го яруса.

Рассмотрим многомерные производящие функции

$$F_i(s_1, s_2, \ldots, s_k),\ i = 1, \ldots, k,$$

где переменная s_i соответствует нетерминальному символу A_i [8]. Функция $F_i(s_1, s_2, \ldots, s_k)$ строится по множеству правил R_i с одинаковой левой частью A_i следующим образом.

Для каждого правила $A_i \overset{p_{ij}}{\to} \beta_{ij}$ выписывается слагаемое

$$q_{ij}{=}p_{ij} \cdot s_1^{l_1} \cdot s_2^{l_2} \cdot \ldots \cdot s_k^{l_k},$$

где l_m - число вхождений нетерминального символа A_m в правую часть правила $(m = 1, \ldots, k)$. Тогда

$$F_i(s_1, s_2, \ldots, s_k) = \sum_{j=1}^{n_i} q_{ij}.$$

Пусть

$$a_j^i = \frac{\partial F_i(s_1, \ldots, s_k)}{\partial s_j} \mid_{s_1=s_2=\ldots=s_k=1.}$$

Квадратная матрица A порядка k, образованная элементами a_j^i, называется *матрицей первых моментов* грамматики G.

Так как матрица A неотрицательна, существует максимальный по модулю действительный неотрицательный собственный корень (перронов корень) [3]. Обозначим этот корень через r.

Известно необходимое и достаточное условие согласованности стохастической КС-грамматики [9]: стохастическая КС-грамматика при отсутствии бесполезных нетерминалов (т.е. не участвующих в порождении слов языка) является согласованной тогда и только тогда, когда $r \leqslant 1$.

В работе рассматривается докритический случай $r < 1$. Основные результаты относятся к стохастическим КС-грамматикам с разложимой матрицей [3] первых моментов.

Введем некоторые обозначения. Будем говорить, что нетерминал A_j непосредственно следует за нетерминалом A_i (и обозначать $A_i \to A_j$),

если в грамматике существует правило вида $A_i \stackrel{p_{il}}{\to} \alpha_1 A_j \alpha_2$, где $\alpha_1, \alpha_2 \in (V_T \cup V_N)^*$. Рефлексивное транзитивное замыкание отношения $\to$ обозначим $\to_*$.

Грамматика называется неразложимой, если для любых двух различных нетерминалов A_i и A_j верно $A_i \to_* A_j$. В противном случае она называется разложимой. Классом нетерминалов назовем максимальное по включению подмножество $K \in V_N$, такое, что $A_i \to_* A_j$ для любых $A_i, A_j \in K$.

Для различных классов K_1 и K_2 будем говорить, что класс K_2 непосредственно следует за классом K_1 (и обозначать $K_1 \prec K_2$), если существуют $A_1 \in K_1$ и $A_2 \in K_2$, такие, что $A_1 \to A_2$. Рефлексивное транзитивное замыкание отношения $\prec$ обозначим через $\prec_*$ и назовем отношением следования.

Очевидно, множество классов нетерминалов является разбиением множества V_N, и отношение $\prec_*$ устанавливает на множестве классов нетерминалов частичный порядок.

Будем полагать, что классы нетерминалов перенумерованы числами от 1 до m таким образом, что из $K_i \prec K_j$ и $i \neq j$ следует $i < j$.

Соответствующая разложимой грамматике разложимая матрица [3] первых моментов A имеет следующий вид:

$$A = \begin{pmatrix} A_{11} & A_{12} & \dots & A_{1m-1} & A_{1m} \\ 0 & A_{22} & \dots & A_{2m-1} & A_{2m} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & A_{m-1m-1} & A_{m-1m} \\ 0 & 0 & \dots & 0 & A_{mm} \end{pmatrix}. \qquad (1)$$

Один класс нетерминалов в матрице первых моментов представлен множеством подряд идущих строк и соответствующим множеством столбцов с теми же номерами. Для класса K_i квадратная подматрица, образованная соответствующими строками и столбцами, обозначается A_{ii}. Блоки, расположенные ниже главной диагонали, нулевые в силу упорядоченности классов нетерминалов. Подматрица A_{ij} является нулевой, если $K_i \not\prec K_j$.

Будем считать, что в грамматике нет особых классов, т.е. классов, состоящих из одного нетерминала, для которых $A_{ii} = 0$. Этого всегда можно добиться, применяя метод укрупнения правил грамматики, описанный в [4].

Для каждого класса K_i матрица A_{ii} неразложима. Обозначим через r_i перронов корень матрицы A_{ii}. Для неразложимой матрицы перронов

корень является положительным и простым по теореме Фробениуса [3]. Очевидно, $r = \max_i\{r_i\}$, и $r > 0$.

Пусть $J = \{i_1, i_2, \ldots, i_l\}$ — множество всех номеров i_j классов, для которых $r_{i_j} = r$. Назовем J определяющим множеством.

Зафиксируем пару (l, h), $l, h \in \{1, 2, \ldots, m\}$, и рассмотрим всевозможные последовательности классов $K_{i_1} \prec K_{i_2} \prec \ldots \prec K_{i_s}$, где $i_1 = l, i_s = h$. Среди всех таких последовательностей выберем ту, которая содержит наибольшее число классов с номерами из J. Это число обозначим через s_{lh}. В случае $K_l \not\prec_* K_h$ положим $s_{lh} = 0$.

На Рис. 2 приведён показана схема классов для некоторой грамматики. Вершины графа соответствуют классам грамматики. Вершины, соответствующие классам из множества J, закрашены.

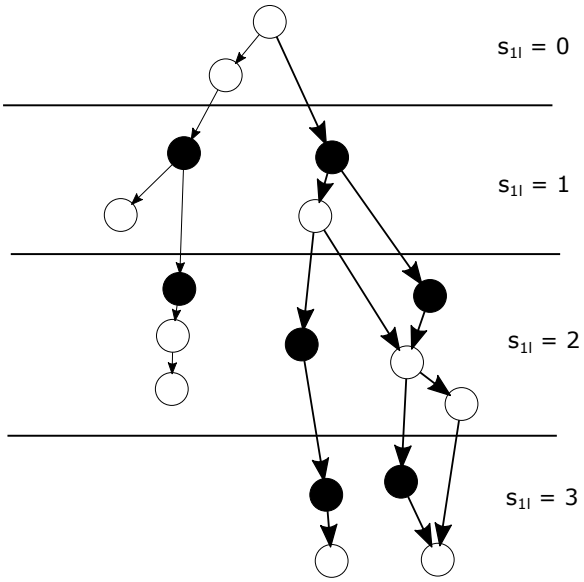

Рис. 2. Максимальные пути

Дополнительно переупорядочим классы по неубыванию величины s_{1l}, причем при одинаковых значениях s_{1l} сначала поставим классы с номерами из множества J.

Среди последовательностей вида

$$K_{i_1} \prec K_{i_2} \prec \ldots \prec K_{i_s}, \tag{2}$$

где $i_1 = l$ и i_s принимает всевозможные значения, выберем последовательности с наибольшим числом классов с номерами из J. Это число

обозначим через q_l. Максимальным путем назовем последовательность вида (2) при $i_1 = 1$, содержащую q_1 классов с номерами из J. Множество всех классов с номерами из J, принадлежащих максимальным путям, обозначим J_{MAX}.

Запись $A_{lh}^{(t)}$ будем применять для обозначения соответствующей подматрицы матрицы A^t.

Теорема 1. *[6]. При* $t \to \infty$:

$$a)\ A_{ij}^{(t)} = H_{ij} \cdot t^{s_{ij}-1} r^t (1 + o(1)) \text{ при } s_{ij} > 0,$$

где H_{ij} *– неотрицательная матрица, не зависящая от* t,

$$b)\ A_{ij}^{(t)} = o(r^t) \text{ при } s_{ij} = 0.$$

Подробное описание строения матрицы первых моментов приведено в [6].

Пусть $G = \langle V_T, V_N, R, s \rangle$ - стохастическая КС-грамматика, где $V_N = \{A_1, A_2, \ldots, A_k\}$. Будем полагать $s = A_1$. Через L_i обозначим язык, порожденный грамматикой G_i, которая получается из G заменой аксиомы на нетерминал A_i. Будем считать, что $L = L_1$ для исходного языка L. Обозначим через D_i множество деревьев вывода слов из L_i и через $Q_l(t)$ - вероятность множества деревьев вывода из D_i, высота которых больше t. Эту вероятность назовем вероятностью продолжения по аналогии с теорией ветвящихся процессов.

Пусть $(A_{j+1}, A_{j+2}, \ldots, A_{j+k_i})$ - последовательность нетерминалов, образующих класс K_i, где k_i – число нетерминалов в K_i, и $j+1$ – номер первого по порядку нетерминала в K_i.

Через $Q^{(i)}(t)$ обозначим вектор вероятностей продолжения $Q^{(i)}(t) = (Q_{j+1}(t), Q_{j+2}(t), \ldots, Q_{j+k_i}(t))^T$.

Теорема 2. *[7] При* $t \to \infty$

$$Q^{(i)}(t) = U^{(i)} \cdot t^{q_i - 1} \cdot r^t \cdot (1 + o(1)),$$

где $U^{(i)}$ *- некоторый положительный вектор.*

Отметим, что в случае $r_i = r$ вектор $U^{(i)}$ пропорционален правому собственному вектору матрицы A_{ii}, соответствующему r. Подробное описание свойств $Q^{(i)}(t)$ содержится в [7].

Обозначим D_l^t множество всех деревьев вывода высоты t для слов из $L_l = L(G_l)$. Вероятность множества D_l^t обозначим через $P_l(t)$. Очевидно, $P_l(t) = Q_l(t-1) - Q_l(t)$.

Из теоремы 2 вытекает

Следствие 1. *Пусть нетерминал $A_l \in K_i$. Тогда*

$$P_l(t) = u_l \cdot t^{q_i - 1} r^{t-1} \cdot (1-r) \cdot (1 + o(1)), \tag{3}$$

где u_l – компонента вектора $U^{(i)}$, соответствующая нетерминалу A_l.

Моменты

Пусть $\Xi = (\xi_1, \ldots, \xi_k)-$ случайный вектор, $\alpha^* = (\alpha_1, \ldots, \alpha_k)$ — фиксированный вектор с целочисленными неотрицательными компонентами и $\alpha = \alpha_1 + \ldots + \alpha_k$. Обозначим

$$\Xi^{[\alpha^*]} = \xi_1^{[\alpha_1]} \ldots \xi_n^{[\alpha_k]},$$

где $x^{[a]} = x(x-1)\ldots(x-a+1)$. Математическое ожидание $M\Xi^{[\alpha^*]}$ будем называть α^*-*моментом* Ξ [8].

Пусть $x_j^i(t)$ — число нетерминалов A_j в дереве вывода из D_i на ярусе t. Через $M_{\alpha^*}^i(t)$ обозначим α^*-момент вектора $X^i(t) = (x_1^i(t), \ldots, x_k^i(t))$.

Примем специальные обозначения для моментов первых четырех порядков. Факториальные моменты первого порядка будем обозначать через $a_j^i(t)$. Для факториальных моментов второго порядка введем обозначения $b_{jn}^i(t)$. Таким образом, $b_{jj}^i(t) = Mx_j^i(t)(x_j^i(t)-1)$ и $b_{jn}^i(t) = Mx_j^i(t)x_n^i(t)$ при $j \neq n$. Для факториальных моментов третьего и четвертого порядков введем обозначения $c_{jnq}^i(t)$ и $f_{jnql}^i(t)$ соответственно.

Нетрудно заметить, что $a_j^i(1)$ – элементы матрицы первых моментов, для которых мы ввели ранее обозначения a_j^i. Будем также применять далее обозначения b_{jn}^i для $b_{jn}^i(1)$.

Нас интересуют оценки для первых четырех моментов.

Свойства первых моментов исследованы в [6], так как $a_j^i(t)$ - элемент матрицы A^t. Для вторых моментов известна следующая формула из [8]:

$$b_{jn}^i(t) = \sum_{\tau=1}^{t} \sum_{l,m,s} a_l^i(t-\tau) b_{ms}^l a_j^m(\tau-1) a_n^s(\tau-1). \tag{4}$$

Пусть a_l^i принадлежит подматрице $A_{h_ih_l}$, a_j^m — подматрице $A_{h_mh_j}$, и a_n^s — подматрице $A_{h_sh_n}$. Подставим в (4) представление для первых моментов:

$$b_{jn}^i(t) = \sum_{\tau=1}^{t} \sum_{l,m,s} c_{il} \cdot \left((t-\tau)^{\delta_1} \cdot r^{t-\tau} + o\left((t-\tau)^{\delta_1} \cdot r^{t-\tau} \right) \right) \cdot b_{ms}^l \cdot c_{mj} \times$$

$$\left((\tau-1)^{\delta_2} r^{\tau-1} + o\left((\tau-1)^{\delta_2} \cdot r^{\tau-1} \right) \right) \cdot c_{sn} \cdot \left((\tau-1)^{\delta_3} \cdot r^{\tau-1} + o\left((\tau-1)^{\delta_3} \cdot r^{\tau-1} \right) \right).$$

Здесь $\delta_1 = s_{h_ih_l} - 1$, $\delta_2 = s_{h_mh_j} - 1$, и $\delta_3 = s_{h_sh_n} - 1$, и c_{il}, c_{mj} и c_{sn} — коэффициенты в соответствующих элементах матрицы A^t.

Проведя несложные преобразования в полученном равенстве, получим:

$$b_{jn}^i(t) = r^t \cdot \sum_l c_{il}\, t^{\delta_1} \cdot \left(\sum_{m,s} b_{ms}^l c_{mj} c_{sn} \right) \times$$

$$\sum_{\tau=1}^{t} \left(\left(1 - \frac{\tau}{t} \right)^{\delta_1} + o\left(\left(1 - \frac{\tau}{t} \right)^{\delta_1} \right) \right) \cdot \left((\tau-1)^{\delta_2+\delta_3} r^{\tau-2} \cdot (1 + o(1)) \right)$$

Ряд $\sum_{\tau=1}^{\infty} (\tau-1)^{\delta_2+\delta_3} r^{\tau-2}$ сходится, поэтому величина

$$c_{il} \left(\sum_{m,s} b_{ms}^l c_{mj} c_{sn} \right) \cdot \sum_{\tau=1}^{t} \left(\left(1 - \frac{\tau}{t} \right)^{\delta_1} + o\left(\left(1 - \frac{\tau}{t} \right)^{\delta_1} \right) \right) \cdot \left((\tau-1)^{\delta_2+\delta_3} r^{\tau-2} \cdot (1 + o(1)) \right)$$

ограничена сверху. Обозначим ее значение через $g_{jn}^i(l)$. Отметим, что $g_{jn}^i(l) > 0$ в тех случаях, когда существуют m и s такие, что $b_{ms}^l > 0$,

$$K_{h_i} \prec_* K_{h_l} \prec_* K_{h_m} \prec_* K_{h_j} \text{ и } K_{h_i} \prec_* K_{h_l} \prec_* K_{h_s} \prec_* K_{h_n}.$$

Условие $b_{ms}^l > 0$ выполняется тогда и только тогда, когда в грамматике существует правило с нетерминалом A_l в левой части, содержащее в правой части оба нетерминала A_m и A_s.

При $t \to \infty$

$$b_{jn}^i(t) = \sum_l g_{jn}^i(l) \cdot t^{\delta_1} r^t (1 + o(1)),$$

где суммирование ведется по тем l, для которых $g_{jn}^i(l) > 0$.

Очевидно, определяющими в этой сумме являются слагаемые с теми значениями l, для которых δ_1 имеет наибольшее значение. Обозначим его через δ_{jn}^i. Поэтому формулу для $b_{jn}^i(t)$ можно записать в следующем виде:

$$b_{jn}^i(t) = g_{jn}^i t^{\delta_{jn}^i} r^t \cdot (1 + o(1)). \tag{5}$$

Здесь $g^i_{jn} = \sum_l g^i_{jn}(l)$, где суммирование ведется по значениям l, удовлетворяющим перечисленным выше условиям. Так как $l \leqslant j$ и $l \leqslant n$, то $\delta^i_{jn} \leqslant \max\{s_{h_ih_j} - 1, s_{h_ih_n} - 1\}$. Поэтому $b^i_{jn}(t) \leqslant O\left(a^i_j(t) + a^i_n(t)\right)$.

Используя результаты из [8], запишем формулу для третьего момента:

$$c^i_{jnq}(t) = \sum_{\tau=1}^{t} \sum_l a^i_l(t-\tau) \cdot z^l_{jnq}(\tau - 1).$$

В этой формуле $z^l_{jnq}(\tau-1)$ состоит из конечного числа слагаемых двух типов. Слагаемые первого типа имеют вид: $Ca^s_q(\tau - 1) \cdot a^m_n(\tau - 1) \cdot a^l_j(\tau - 1)$ для некоторых s, m, l, где C — некоторая константа, зависящая от слагаемого; слагаемые второго типа имеют вид: $Ca^l_j(\tau-1) \cdot b^m_{nq}(\tau-1)$ для некоторых l, m и константы C.

Поэтому вычисление $c^i_{jnq}(t)$ сводится к вычислению конечного числа сумм вида

$$S_1(t) = \sum_{\tau=1}^{t} a^i_l(t-\tau) \cdot a^l_j(\tau-1) \cdot a^s_q(\tau-1) \cdot a^m_n(\tau-1)$$

и вида

$$S_2(t) = \sum_{\tau=1}^{t} a^i_l(t-\tau) \cdot a^s_j(\tau-1) \cdot b^m_{nq}(\tau-1)$$

для некоторых значений l, m, s.

Оценим $S_1(t)$ и $S_2(t)$, используя оценки $a^i_j(t) = O(t^{s_{ij}-1}r^t)$, $b^i_{jl}(t) \leqslant O\left(a^i_j(t) + a^i_l(t)\right)$. Применяя очевидное неравенство $s_{ij} \leqslant w$, где $w = \max_{i,j}\{s_{ij}\}$, получим, что $S_1(t) \leqslant O(t^{w-1}r^t)$ и $S_2(t) \leqslant O(t^{w-1}r^t)$. Поэтому

$$c^i_{jnq}(t) \leqslant O\left(t^{w-1}r^t\right).$$

Закономерности в деревьях вывода слов стохастического КС-языка

Для доказательства основного результата раздела предварительно докажем лемму.

Через $R_X(n)$ обозначим выражение

$$\prod_{j=1}^{k} (1 - Q_j(n))^{x_j} - \prod_{j=1}^{k} (1 - Q_j(n-1))^{x_j}, \tag{6}$$

где $X = (x_1, \ldots, x_k)$ — целочисленный неотрицательный вектор, $Q_j(n)$ — вероятности продолжения $(j = 1, \ldots, k)$, k — общее число нетерминалов в грамматике.

Лемма 1. *При* $n \to \infty$

$$R_X(n) = (1 + \psi_X(n)) \sum_{j=1}^{k} x_j P_j(n),$$

где $-\tilde{c}_1 n^{q_1 - 1} r^n \cdot \sum x_j \leqslant \psi_X(n) \leqslant \tilde{c}_2 n^{q_1 - 1} r^n$, *и* $\tilde{c}_1$ *и* $\tilde{c}_2$ *– положительные константы.*

Доказательство.

Заметим, что

$$\prod_{j=1}^{k} (1 - Q_j(n))^{x_j} = 1 - \sum_{j=1}^{k} x_j Q_j(n) + \sum_{j,l=1}^{k} x_j x_l Q_j(n) Q_l(n) + \sum_{j=1}^{k} (C_{x_j}^2 Q_j(n)^2) - \ldots$$

В каждом из слагаемых выделим $\sum x_j Q_j(n)$:

$$1 - \sum_{j=1}^{k} x_j Q_j(n) + \sum_{j=1}^{k} x_j Q_j(n) \sum_{l=1}^{k} x_l Q_l(n) + \sum_{j=1}^{k} \frac{(x_j - 1)}{2} Q_j(n) - \cdots =$$

Аналогично преобразуя $\prod_{j=1}^{k} (1 - Q_j(n-1))^{x_j}$, и подставляя в $R_X(n)$, получаем:

$$R_X(n) = \sum_{j=1}^{k} x_j Q_j(n) \left(1 - \sum_{l=1}^{k} x_l Q_l(n) - \ldots \right)$$

Число слагаемых в скобках, а также константы перед величинами $Q(.)$, полностью определяются вектором X, и не зависят от n. Обозначим через c максимальный модуль таких констант. Каждое слагаемое можно оценить как $O(n^{q_1 - 1} r^t)$. Тогда

$$R_X(n) = \sum_{j=1}^{k} x_j Q_j(n)(1 + \psi_X(n))$$

где $|\psi_X(n)| \leqslant \tilde{c} n^{q_1 - 1} r^t$.

Лемма 1 доказана.

Будем полагать, как и ранее, что аксиомой исходной грамматики G является нетерминал A_1. Рассмотрим D_1^t — множество деревьев из D_1

высоты t. Для $d \in D_1^t$ через $p_t(d)$ будем обозначать условную вероятность дерева d, т.е. $p_t(d) = \frac{p(d)}{P(D_1^t)}$.

Через $M_i(t,\tau)$ обозначим условное математическое ожидание числа вершин на ярусе τ, помеченных нетерминалом A_i, в деревьях вывода высоты t.

Для нетерминала $A_l \in K_j$ положим $q'_l = q_j$ и $s'_{1l} = s_{1j}$.

Теорема 3. *Пусть G — стохастическая КС-грамматика с разложимой матрицей первых моментов, для которой $r < 1$.*

Тогда для любого $i \in \{1,\dots,k\}$ при $\tau \to \infty$ и $t - \tau \to \infty$ выполняется асимптотическое равенство

$$M_i(t,\tau) \sim \frac{f_i \cdot (t-\tau)^{q'_i-1} \cdot \tau^{s'_{1i}-1}}{t^{q_1-1}} + \sum_{l=1}^{k} \frac{f_{il} \cdot (t-\tau)^{q'_l-1} \cdot \tau^{\delta^1_{il}}}{t^{q_1-1}},$$

в котором f_i, f_{il} - неотрицательные константы и δ^1_{il} определено в (5).

Доказательство.

Представим $M_i(t,\tau)$ в виде

$$M_i(t,\tau) = \sum_{d \in D_1^t} p_t(d) z_i(d,\tau) = \frac{1}{P(D_1^t)} \sum_{d \in D_1^t} p(d) z_i(d,\tau),$$

где $z_i(d,\tau)$ – число вершин на ярусе τ дерева d, помеченных нетерминалом A_i.

Рассмотрим неотрицательный целочисленный вектор $X = (x_1,\dots,x_k)$, который будем называть далее вектором нетерминалов. Используя вектор X, мы можем записать, что

$$M_i(t,\tau) = \frac{1}{P(D_1^t)} \sum_{X \neq 0} \Delta_X,$$

где Δ_X – вклад в математическое ожидание тех деревьев вывода из D_1^t, которые на ярусе τ содержат x_j вершин, помеченных нетерминалом A_j $(j = 1,\dots,k)$. Множество таких деревьев обозначим через $D_X^t(\tau)$.

Пусть $d \in D_X^t(\tau)$. Выделим в d поддерево d_0 и последовательность поддеревьев $(d_1, d_2, \dots d_n)$, где $n = \sum_{l=1}^{k} x_l$. Поддерево d_0 получено из d удалением всех вершин на ярусах $\tau+1$, $\tau+2,\dots,t$ и инцидентных им дуг. Последовательность $(d_1, d_2, \dots d_n)$ образуют все поддеревья, корни которых расположены на ярусе τ дерева d. При этом корни поддеревьев

d_1, $d_2, \ldots d_m$ расположены в дереве d последовательно в порядке обхода вершин яруса τ слева направо, и каждое дерево d_l $(l = 1, \ldots, n)$ содержит все дуги и вершины дерева d, лежащие на путях от корня d_l к листьям дерева d.

Выделим в $D_X^t(\tau)$ множество деревьев, имеющих в качестве поддерева d_0 одно и то же дерево. Обозначим это множество через D_0. Нетрудно понять, что

$$P(D_0) = p(d_0) \cdot \left(\prod_{l=1}^{k} (1 - Q_l(t-\tau))^{x_l} - \prod_{l=1}^{k} (1 - Q_l(t-\tau-1))^{x_l} \right), \quad (7)$$

где $Q_l(n)$ — суммарная вероятность деревьев из множества D_l, высота которых больше n, и, следовательно, $(1 - Q_l(n))$ – суммарная вероятность деревьев из D_l, высота которых не превосходит n.

Обозначим через $\delta_1(X)$ выражение $\prod_{l=1}^{k}(1-Q_l(t-\tau))^{x_l}$ и через $\delta_2(X)$ — выражение $\prod_{l=1}^{k}(1 - Q_l(t-\tau-1))^{x_l}$.

В (7) величина $p(d_0) \cdot \delta_1(X)$ есть суммарная вероятность деревьев, определяемых поддеревом d_0, высота которых не превосходит t, так как каждое поддерево с корнем на ярусе τ имеет высоту, не превосходящую $(t - \tau)$.

Вторая величина $p(d_0) \cdot \delta_2(X)$ есть суммарная вероятность деревьев, определяемых поддеревом d_0, высота которых не превосходит $(t-\tau-1)$.

Разность этих величин равна, очевидно, суммарной вероятности деревьев высоты t, определяемых поддеревом d_0, и значение $\delta_1(X) - \delta_2(X)$ не зависит от порядка следования вершин на ярусе τ, помеченных нетерминалами.

Выражение для $\delta_1(X) - \delta_2(X)$ в обозначениях леммы 1 есть $R_X(n)$ при $n = t - \tau$. Поэтому

$$P(D_X^t(\tau)) = \sum_{d_0} (p(d_0) \cdot R_X(t-\tau)) = R_X(t-\tau) \sum_{d_0} p(d_0),$$

где суммирование ведется по всем возможным поддеревьям d_0 деревьев из $D_X^t(\tau)$.

Через $D_X(\tau)$ обозначим множество всех деревьев вывода из D_1, которым на ярусе τ соответствует вектор нетерминалов X. Рассмотрим дерево из $D_X(\tau)$. Для каждой вершины этого дерева, помеченной некоторым нетерминалом A_l, суммарная вероятность возможных деревьев с корнем в этой вершине и листьями, помеченными только символами из $V_T \cup \{\lambda\}$, равна $P(D_l)$. Ввиду согласованности исходной грамматики

$P(D_l) = 1$ для любого l. Поэтому $\sum_{d_0} p(d_0)$ равна вероятности деревьев вывода из D_1, имеющих x_l вершин на ярусе τ, помеченных нетерминалом A_l $(l = 1, \ldots, k)$:

$$\sum_{d_0} p(d_0) = \sum_{d_0} p(d_0) \cdot P(D_1)^{x_1} \cdot P(D_2)^{x_2} \cdot \ldots \cdot P(D_k)^{x_k} = P\left(D_X(\tau)\right).$$

Далее будем обозначать $P\left(D_X(\tau)\right)$ через $P_X(\tau)$. Таким образом,

$$M_i(t,\tau) = \frac{1}{P(D_1^t)} \sum_{X \neq 0} \left(P_X(\tau) \cdot R_X(t-\tau) \cdot x_i\right).$$

Применяя лемму 2 для представления $R_X(t-\tau)$, получим:

$$M_i(t,\tau) = \frac{1}{P(D_1^t)} \sum_{X \neq 0} \left(P_X(\tau) \cdot x_i \cdot (1 + \psi_X(t-\tau)) \sum_{l=1}^{k} x_l P\left(D_l^{t-\tau}\right) \right) =$$

$$\sum_{l=1}^{k} \frac{P(D_l^{t-\tau})}{P(D_1^t)} \sum_{X \neq 0} \left(P_X(\tau) \cdot x_i \cdot x_l \cdot (1 + \psi_X(t-\tau))\right).$$

Отдельно вычислим $S_1 = \sum_{X \neq 0} \left(P_X(\tau) \cdot x_i \cdot x_l\right)$ и $S_2 = \sum_{X \neq 0} \left(P_X(\tau) \cdot x_i \cdot x_l \cdot \psi_X(t-\tau)\right)$.

Используя первые и вторые моменты, мы можем записать, что $S_1 = b_{il}^1(\tau)$ при $i \neq l$ и $S_1 = b_{ii}^1(\tau) + a_i^1(\tau)$ при $l = i$.

Учитывая оценку из леммы 2 для $\psi_X(n)$ и используя первые три момента, получим нижнюю и верхнюю оценки для S_2 :

$$S_2 = \sum_{X \neq 0} \left(P_X(\tau) \cdot x_i \cdot x_l \cdot \psi_X(t-\tau)\right) \geqslant$$

$$-\tilde{c}_2 \tau^{q_1 - 1} r^{\tau} \sum_{X \neq 0} \left(P_X(\tau) \cdot x_i \cdot x_l \cdot \sum_j x_j \right) = -\tilde{c}_2 \tau^{q_1 - 1} r^{\tau} \sum_j c_{ilj}^{1*}(\tau),$$

где

$$c_{ilj}^{1*}(\tau) = c_{ilj}^1(\tau) \quad \text{при } i \neq l,\ i \neq j \quad \text{и } j \neq l,$$

$$c_{iii}^{1*}(\tau) = c_{iii}^1(\tau) + 3b_{ii}^1(\tau) - a_i^1(\tau),$$

и

$$c_{ijj}^{1*}(\tau) = c_{ijj}^1(\tau) + b_{ij}^1(\tau),\ c_{iij}^{1*}(\tau) = c_{iij}^1(\tau) + b_{ij}^1(\tau).$$

Применяя оценки для первых трех моментов, получим, что $S_2 \geqslant -c \cdot \tau^{2q_1 - 2} r^{2\tau}$, где c — некоторая положительная константа.

С другой стороны,

$$S_2 \leqslant c_1 \tau^{q_1-1} r^{\tau} \sum_{X \neq 0} (P_X(\tau) \cdot x_i \cdot x_l) = c_1 \tau^{q_1-1} r^{\tau} \cdot S_1.$$

Так как $S_1 = b_{il}^1(\tau)$ при $i \neq l$ и $S_1 = b_{ii}^1(\tau) + a_i^1(\tau)$, то, с учетом оценок для моментов, получаем, что $S_2 = O(\tau^{2q_1-2} r^{2\tau})$.

Вернемся к вычислению $M_i(t,\tau)$:

$$M_i(t,\tau) = \sum_{l \neq i} \frac{P(D_l^{t-\tau})}{P(D_1^t)} b_{il}^1(\tau) + \frac{P(D_i^{t-\tau})}{P(D_1^t)} \left(b_{ii}^1(\tau) + a_i^1(\tau)\right) + \sum_{l=1}^{k} \frac{P(D_l^{t-\tau})}{P(D_1^t)} \cdot O\left(\tau^{2q_1^*-2} r^{2\tau}\right).$$

Раскрывая моменты и используя лемму 2, после несложных преобразований получим:

$$M_i(t,\tau) =$$

$$\sum_{l=1}^{k} \frac{d_l \cdot (1-r) \cdot (t-\tau)^{q_l'-1} \cdot r^{t-\tau-1} \cdot (1+\phi_l(t-\tau))}{d_1 \cdot (1-r) \cdot t^{q_1-1} \cdot r^{t-1} \cdot (1+\phi_1(t))} \cdot \left(g_{il}^1 \cdot r^{\tau} \cdot \tau^{\delta_{il}^1} (1+\psi_{il}(\tau))\right) +$$

$$\frac{d_i \cdot (1-r) \cdot (t-\tau)^{q_i'-1} r^{t-\tau-1} (1+\phi_i(t-\tau)) \cdot c_{1i} \cdot \tau^{s_{1i}'-1} r^{\tau} (1+\varphi_{1i}(n)(\tau))}{d_1 \cdot (1-r) \cdot t^{q_1-1} r^{t-1} \cdot (1+\phi_1(t))} + O\left(\tau^{2q_1^*-2} r^{2\tau}\right),$$

где $\phi_i(n) = o(1)$, $\psi_{il}(n) = o(1)$, $\varphi_{1i}(n) = o(1)$, $q_l' = q_j$ для $A_l \in K_j$, и $s_{1i}' = s_{1m}$ для $A_i \in K_m$.

Отсюда следует, что

$$M_i(t,\tau) = \frac{1}{d_1} \left(\sum_{l=1}^{k} \frac{d_l \cdot g_{il}^1 \cdot (t-\tau)^{q_l'-1} \cdot \tau^{\delta_{il}^1}}{t^{q_1-1}} + \right. \tag{8}$$

$$\left. \frac{d_i \cdot c_{1i} \cdot (t-\tau)^{q_i'-1} \cdot \tau^{s_{1i}'-1}}{t^{q_1-1}} \right) (1 + \xi(\tau, t-\tau)),$$

где $\xi(\tau, t-\tau) \to 0$ при $\tau, t-\tau \to \infty$. Теорема доказана.

Рассмотрим подробнее слагаемые в (8). Определяющими в сумме являются те значения l, для которых $g_{il}^1 > 0$ и $q_l' + \delta_{il}^1 = q_1$. Равенство справедливо при одновременном выполнении следующих условий:

1) нетерминал A_l принадлежит классу K_{j_1} с $j_1 \in J_{MAX}$,

2) $A_i \in K_{j_2}$, для которого $K_{j_1} \prec_* K_{j_2}$.

Обозначим N_i множество номеров l, для которых выполняются условия 1) и 2).

Отметим, что слагаемое $\frac{d_i \cdot c_{1i} \cdot (t-\tau)^{q'_i-1} \cdot \tau^{s'_{1i}-1}}{t^{q_1-1}}$ влияет на значение $M_i(t,\tau)$ при $s'_{1i} + q'_i - 1 = q_1$. Это равенство выполняется в случае, если $A_i \in K_{j_2}$, где $j_2 \in J_{MAX}$. Поэтому равенство (8) при $N_i \neq \varnothing$ можно записать в виде

$$M_i(t,\tau) = \left(\sum_{l \in N_i} \frac{f_{il} \cdot (t-\tau)^{q'_l-1} \cdot \tau^{\delta^1_{il}}}{t^{q_1-1}} + \frac{f_i \cdot (t-\tau)^{q'_i-1} \cdot \tau^{s'_{1i}-1}}{t^{q_1-1}} \right) (1 + \xi(\tau, t-\tau)),$$

где $f_{il} = \frac{d_l \cdot g^1_{il}}{d_1}$, $f_i = \frac{d_i \cdot c_{1i}}{d_1}$ и $\xi(\tau, t-\tau) \to 0$ при $\tau, t-\tau \to \infty$. Очевидно, $M_i(t,\tau) \leqslant O(1/t)$ при $N_i = \varnothing$. Поэтому справедливо

Теорема 3 доказана.

Следствие 2.

$$1) M_i(t,\tau) = \left(\sum_{l \in N_i} \frac{f_{il} \cdot (t-\tau)^{q'_l-1} \cdot \tau^{\delta^1_{il}}}{t^{q_1-1}} + \frac{f_i \cdot (t-\tau)^{q'_i-1} \cdot \tau^{s'_{1i}-1}}{t^{q_1-1}} \right) (1 + \xi(\tau, t-\tau))$$

при $N_i \neq \varnothing$;

2) $M_i(t,\tau) \leqslant O(1/t)$ *при* $N_i = \varnothing$.

Пусть r_{ij} — произвольное правило грамматики G. Через $s^{(ij)}_l$ обозначим число нетерминалов A_l в правой части правила r_{ij}. Условное математическое ожидание числа применений правила r_{ij} в деревьях вывода высоты t на ярусе τ будем обозначать через $M_{ij}(t,\tau)$.

Теорема 4. *Пусть G — стохастическая КС-грамматика с разложимой матрицей первых моментов, для которой перронов корень $r < 1$, и D^t_1 – множество деревьев вывода высоты t.*

Тогда при $\tau \to \infty$ и $t - \tau \to \infty$ выполняется следующее асимптотическое равенство:

$$M_{ij}(t,\tau) \sim \frac{p_{ij}}{t^{q_1-1}} \left(\sum_{l=1}^{k} f_{il} \cdot (t-\tau)^{q'_l-1} \cdot \tau^{\delta^1_{il}} + \frac{1}{r} \sum_{m=1}^{k} f_m \cdot s^{(ij)}_m \cdot (t-\tau)^{q'_m-1} \tau^{s'_{1i}-1} \right).$$

В формулировке теоремы p_{ij} — вероятность правила r_{ij}, задаваемая в исходной грамматике, $s^{(ij)}_m$ – число нетерминалов A_m в правой части правила r_{ij}, а величины q'_l, δ^1_{il}, f_{il} и f_m имеют тот же смысл, что и в теореме 3.

Доказательство. Обозначим $z_{ij}(d,\tau)$ число вершин на ярусе τ дерева d, помеченных нетерминалом A_i, к которым применено правило

r_{ij}. Используя неотрицательный целочисленный вектор $X = (x_1, \ldots, x_k)$, можно записать, что

$$M_{ij}(t,\tau) = \sum_{X \neq 0} \sum_{d \in D_X^t(\tau)} p_t(d) z_{ij}(d,\tau),$$

где $D_X^t(\tau)$ введено в доказательстве теоремы 3.

Представим $z_{ij}(d,\tau)$ в виде суммы случайных величин $I_1 + I_2 + \ldots + I_{x_i}$, где $I_m = 1$, если к m-й по порядку вершине среди вершин, помеченных нетерминалом A_m на ярусе τ, применено правило r_{ij}, и $I_m = 0$ в противном случае ($m = 1, 2, \ldots, x_i$). Тогда

$$M_{ij}(t,\tau) = \sum_{X \neq 0} \sum_{d \in D_X^t(\tau)} p_t(d) \cdot (I_1 + I_2 + \ldots + I_{x_i}).$$

Очевидно, что случайные величины I_m ($m = 1, 2, \ldots, x_i$) – одинаково распределены на $D_X^t(\tau)$, поэтому

$$M_{ij}(t,\tau) = \frac{1}{P(D_1^t)} \sum_{X \neq 0} P\left(D_{X,1}^t(\tau)\right) \cdot x_i,$$

где $P(D_{X,1}^t(\tau))$ – суммарная вероятность тех деревьев из $D_X^t(\tau)$, в которых правило r_{ij} применено к первой по порядку вершине на ярусе τ, помеченной A_i.

Подсчитаем вероятность $P\left(D_{X,1}^t(\tau)\right)$:

$$P\left(D_{X,1}^t(\tau)\right) = p_{ij} \cdot P_X(\tau) \times \left[\prod_{m=1}^{k} (1 - Q_m(t-\tau))^{x'_m} \cdot \prod_{m=1}^{k} (1 - Q_m(t-\tau-1))^{s_m^{(ij)}} - \right.$$

$$\left. \prod_{m=1}^{k} (1 - Q_m(t-\tau-1))^{x'_m} \cdot \prod_{m=1}^{k} (1 - Q_m(t-\tau-2))^{s_m^{(ij)}} \right]. \qquad (9)$$

Здесь $X' = (x'_1, \ldots, x'_k) = (x_1, \ldots, x_{i-1}, x_i - 1, x_{i+1}, \ldots, x_k)$ и $S = (s_1^{(ij)}, \ldots, s_k^{(ij)})$, где $s_m^{(ij)}$ равно числу нетерминалов A_m в правой части правила r_{ij} ($m = 1, \ldots, k$). Величина $P_X(\tau)$ имеет тот же смысл, что и в доказательстве теоремы 3. Выражение в квадратных скобках в (9) аналогично выражению $R_X(t-\tau)$. При этом с помощью множителей $(1 - Q_m(t-\tau-1))^{s_m^{(ij)}}$ и $(1 - Q_m(t-\tau-2))^{s_m^{(ij)}}$ учитывается тот факт, что к первому нетерминалу A_i на ярусе τ применено правило r_{ij}, которому на ярусе $\tau + 1$ соответствует $s_m^{(ij)}$ вершин, помеченных нетерминалом A_m ($m = 1, \ldots, k$).

Проведем несложные преобразования в (9):

$$P\left(D^t_{X,1}(\tau)\right) = p_{ij}\cdot P_X(\tau)\cdot\frac{1}{1-Q_i(t-\tau)}\cdot\prod_{m=1}^{k}(1-Q_m(t-\tau-1))^{s_m^{(ij)}}\times$$

$$\left[\prod_{m=1}^{k}(1-Q_m(t-\tau))^{x_m}-\frac{1-Q_i(t-\tau)}{1-Q_i(t-\tau-1)}\times\right.$$

$$\left.\prod_{m=1}^{k}\left((1-Q_m(t-\tau-1))^{x_m}\cdot\frac{(1-Q_m(t-\tau-2))^{s_m^{(ij)}}}{(1-Q_m(t-\tau-1))^{s_m^{(ij)}}}\right)\right].$$

Очевидно, что

$$\frac{1-Q_i(t-\tau)}{1-Q_i(t-\tau-1)} = 1+\frac{Q_i(t-\tau-1)-Q_i(t-\tau)}{1-Q_i(t-\tau-1)} =$$

$$1+\frac{P(D_i^{t-\tau})}{1-Q_i(t-\tau-1)} = 1+P(D_i^{t-\tau})+\frac{P(D_i^{t-\tau})\cdot Q_i(t-\tau-1)}{1-Q_i(t-\tau-1)}.$$

Применим теорему 2 и следствие из нее для оценки $Q_i(t-\tau-1)$ и $P(D_i^{t-\tau})$. Получим, что

$$\frac{1-Q_i(t-\tau)}{1-Q_i(t-\tau-1)} = 1+P(D_i^{t-\tau})+O((t-\tau)^{2(q_1-1)}\cdot r^{2(t-\tau)}).$$

Проводя аналогичные преобразования и учитывая, что $s_m^{(ij)}$ – константа, определяемая правой частью правила r_{ij}, мы можем записать, что

$$\prod_{m=1}^{k}\frac{(1-Q_m(t-\tau-2))^{s_m^{(ij)}}}{(1-Q_m(t-\tau-1))^{s_m^{(ij)}}} = 1-\sum_{m=1}^{k}s_m^{(ij)}\cdot P\left(D_m^{t-\tau-1}\right)+O\left((t-\tau)^{2(q_1-1)}\cdot r^{2(t-\tau)}\right).$$

Поэтому

$$P\left(D^t_{X,1}(\tau)\right) = p_{ij}\cdot P_X(\tau)\cdot\left(1+O\left((t-\tau)^{q_1-1}\cdot r^{t-\tau}\right)\right)\left[\prod_{m=1}^{k}(1-Q_m(t-\tau))^{x_m}-\right.$$

$$\left.\prod_{m=1}^{k}(1-Q_m(t-\tau-1))^{x_m}\cdot\left(1+P(D_i^{t-\tau})+O((t-\tau)^{2(q_1-1)}\cdot r^{2(t-\tau)})\right)\times\right.$$

$$\left(1-\sum_{m=1}^{k} s_m^{(ij)}\cdot P\left(D_m^{t-\tau-1}\right)+O\left((t-\tau)^{2(q_1-1)}\cdot r^{2(t-\tau)}\right)\right)\Bigg] =$$

$$p_{ij}\cdot P_X(\tau)\left[R_X(t-\tau)+\prod_{m=1}^{k}\left(1-Q_m(t-\tau-1)\right)^{x_m}\times\right.$$

$$\left.\left(\sum_{m=1}^{k} s_m^{(ij)}P(D_m^{t-\tau-1})-P(D_i^{t-\tau})\right)\right]\cdot\left(1+O\left((t-\tau)^{q_1-1}\cdot r^{t-\tau}\right)\right).$$

(Здесь $R_X(t-\tau)$ – величина, рассмотренная в лемме 4.)

Вернемся к вычислению $M_{ij}(t,\tau)$, учитывая оценку

$$\prod_{m=1}^{k}\left(1-Q_m(t-\tau-1)\right)^{x_m}=1-O\left((t-\tau)^{q_1-1}\cdot r^{t-\tau}\sum_{m=1}^{k}x_m\right),$$

следующую из (5). Тогда

$$M_{ij}(t,\tau)=\frac{1}{P(D_1^t)}\left[\sum_{X\neq 0} p_{ij}\cdot P_X(\tau)\cdot R_X(t-\tau)\cdot x_i+\left(\sum_{m=1}^{k} s_m^{(ij)}P\left(D_m^{t-\tau-1}\right)-P\left(D_i^{t-\tau}\right)\right)\times\right.$$

$$\left.\sum_{X\neq 0} p_{ij}\cdot P_X(\tau)\cdot x_i\cdot\left(1-O\left((t-\tau)^{q_1-1}\cdot r^{t-\tau}\sum_{m=1}^{k}x_m\right)\right)\right]\cdot\left(1+O\left((t-\tau)^{q_1-1}\cdot r^{t-\tau}\right)\right).$$

Величина

$$\frac{1}{P(D_1^t)}\cdot\sum_{X\neq 0} P_X(\tau)\cdot R_X(t-\tau)\cdot x_i$$

есть $M_i(t,\tau)$ из теоремы 3, и

$$\sum_{X\neq 0} P_X(\tau)\cdot x_i=a_i^1(\tau)=c_{1i}\cdot\tau^{s'_{1i}-1}r^{\tau}(1+o(1)),$$

где $a_i^1(\tau)$ – элемент матрицы A^τ, A – матрица первых моментов, и s'_{1i} имеет тот же смысл, что и в теореме 3.

Кроме того,

$$\sum_{X\neq 0} P_X(\tau)\cdot x_i\cdot O\left((t-\tau)^{q_1-1}\cdot r^{t-\tau}\sum_{m=1}^{k}x_m\right)=$$

$$O\left((t-\tau)^{q_1-1}\cdot r^{t-\tau}\right)\sum_{m=1}^{k} b_{im}^{1}(\tau) = O\left(\tau^{q_1-1}\cdot(t-\tau)^{q_1-1}r^{t}\right),$$

где $b_{im}^{1}(\tau)$ – вторые моменты. Следовательно,

$$M_{ij}(t,\tau) = \left(M_i(t,\tau)\cdot p_{ij} + \frac{p_{ij}\cdot c_{1i}\cdot\tau^{s'_{1i}-1}r^{\tau}\cdot(1+o(1))}{P(D_1^t)}\times\right.$$

$$\left.\left(\sum_{m=1}^{k} s_m^{(ij)}\cdot P(D_m^{t-\tau-1}) - P(D_i^{t-\tau})\right)\right)\cdot\left(1+O\left((t-\tau)^{q_1-1}r^{t-\tau}\right)\right).$$

Применяя теорему 3 к $M_i(t,\tau)$ и формулу (3) к $P(D_m^n)$, после проведения несложных преобразований $M_{ij}(t,\tau)$ можем представить в следующем виде:

$$M_{ij}(t,\tau) = \frac{p_{ij}}{t^{q_1-1}}\left(\sum_{l=1}^{k} f_{il}\cdot(t-\tau)^{q'_l-1}\cdot\tau^{\delta_{il}^1} + \frac{1}{r}\sum_{m=1}^{k} f_m\cdot s_m^{(ij)}\cdot(t-\tau)^{q'_m-1}\tau^{s'_{1i}-1}\right) +$$

$$\xi_{ij}^1(t) + \xi_{ij}^2(\tau) + \xi_{ij}^3(t-\tau),$$

где $\xi_{ij}^1(t) = o(1)$, $\xi_{ij}^2(\tau) = o(1)$ и $\xi_{ij}^3(t-\tau) = o(1)$.

Обозначим сумму $\xi_{ij}^1(t) + \xi_{ij}^2(\tau) + \xi_{ij}^3(t-\tau)$ через $\xi_{ij}(t,\tau)$. Очевидно, $\xi_{ij}(t,\tau)\to 0$ при $\tau\to\infty$ и $t-\tau\to\infty$. Поэтому

$$M_{ij}(t,\tau) =$$

$$\frac{p_{ij}}{t^{q_1-1}}\left(\sum_{l=1}^{k} f_{il}\cdot(t-\tau)^{q'_l-1}\cdot\tau^{\delta_{il}^1} + \frac{1}{r}\sum_{m=1}^{k} f_m\cdot s_m^{(ij)}\cdot(t-\tau)^{q'_m-1}\tau^{s'_{1i}-1}\right)+\xi_{ij}(t,\tau). \tag{10}$$

Теорема 4 доказана.

Сделаем несколько выводов из теоремы 4.

1. $M_{ij}(t,\tau)$ ограничено константой при $\tau\to\infty$, $t-\tau\to\infty$.

2. Величина $\sum_{m=1}^{k} f_m\cdot s_m^{(ij)}\cdot(t-\tau)^{q'_m-1}\tau^{s'_{1i}-1}$ имеет большее значение для тех правил, которые содержат в правой части большее количество нетерминальных символов.

3. Величина $\sum_{l\in N_i} f_{il}\cdot(t-\tau)^{q'_l-1}\cdot\tau^{\delta_{il}^1}$ имеет одно и то же значение для всех правил грамматики с одинаковой левой частью A_i.

Пусть $S_{ij}(t) = q_{ij}(t,0)+q_{ij}(t,1)+\ldots+q_{ij}(t,t-1)$, где $q_{ij}(t,\tau)$ — число правил r_{ij} на ярусе τ в дереве из D_1^t; $S_{ij}(t)$ — число правил r_{ij} в дереве вывода из D_1^t.

Рассмотрим случайную величину $\frac{S_{ij}(t)}{t}$ — среднее число правил r_{ij}, приходящееся на один ярус дерева вывода из D_1^t.

Теорема 5. Пусть G — стохастическая КС-грамматика с разложимой матрицей первых моментов, для которой перронов корень $r < 1$, и D_1^t – множество деревьев вывода высоты t.

Тогда при $t \to \infty$ выполняется следующее асимптотическое равенство:

$$M\left(\frac{S_{ij}(t)}{t}\right) \sim w_{ij},$$

где w_{ij} - константа, определяемая грамматикой G.

Доказательство.

Разобьем $S_{ij}(t)$ на три части:

$$S_{ij}(t) = S_{ij}^{(1)}(t) + S_{ij}^{(2)}(t) + S_{ij}^{(3)}(t),$$

где

$$S_{ij}^{(1)}(t) = q_{ij}(t,0) + \ldots + q_{ij}(t,\tau_0 - 1),$$

$$S_{ij}^{(2)}(t) = q_{ij}(t,\tau_0) + \ldots + q_{ij}(t,t-\tau_0 - 1),$$

$$S_{ij}^{(3)}(t) = q_{ij}(t,t-\tau_0) + \ldots + q_{ij}(t,t-1),$$

и положим $\tau_0 = \lfloor \log\log t \rfloor$ (здесь и далее логарифм берется по основанию 2). Число слагаемых в $S_{ij}^{(1)}(t)$ и в $S_{ij}^{(3)}(t)$ равно $\lfloor \log\log t \rfloor$, а в $S_{ij}^{(2)}(t)$ равно $t - 2\lfloor \log\log t \rfloor$.

Найдем математические ожидания $M\left(S_{ij}^{(1)}(t)\right)$, $M\left(S_{ij}^{(2)}(t)\right)$ и $M\left(S_{ij}^{(3)}(t)\right)$.

Величину $M\left(S_{ij}^{(1)}(t)\right)$ можно представить в следующем виде:

$$M\left(S_{ij}^{(1)}(t)\right) = M_{ij}(t,0) + M_{ij}(t,1) + \ldots + M_{ij}(t,\tau_0 - 1).$$

Число правил r_{ij} на ярусе τ в дереве из D_1^t обозначим $q_{ij}(t,\tau)$. Оценим $q_{ij}(t,\tau)$ для $\tau < \tau_0$. Обозначим через k_{max} максимальное число нетерминалов в правой части правил грамматики G. Тогда $q_{ij}(t,\tau) \leqslant k_{max}^{\tau} < k_{max}^{\tau_0}$. Поэтому $M_{ij}(t,\tau) < k_{max}^{\tau_0}$ и

$$M\left(S_{ij}^{(1)}(t)\right) \leqslant k_{max}^{\tau_0}\tau_0 \leqslant k_{max}^{\log\log t}\log\log t = \log^{c_1} t\,\log\log t \leqslant \log^{c_2} t,$$

где $c_1 = \log k_{max},\ c_2 = c_1 + 1$.

Для $t-\tau_0 \leqslant \tau < t$ имеем:

$$M_{ij}(t,\tau) \leqslant M_i(t,\tau) = \frac{1}{P(D^t)} \sum_X P_X(\tau) R_X(t-\tau) x_i \leqslant \frac{1}{P(D^t)} \sum_X P_X(\tau) x_i =$$

$$\frac{1}{P(D^t)} a_i^1(\tau) \leqslant O\left(\frac{\tau^{q_1-1}}{t^{q_1-1} \cdot r^{t-\tau}}\right) \leqslant O\left(\frac{1}{r^{t-\tau}}\right).$$

Поэтому

$$M\left(S_{ij}^{(3)}(t)\right) \leqslant \sum_{t-\tau_0}^{t-1} O\left(\frac{1}{r^{t-\tau}}\right) = O\left(\frac{\tau_0}{r^{\tau_0}}\right) = O\left(\frac{\log\log t}{r^{\log\log t}}\right) = O\left(\log^{c_3} t\right)$$

для некоторой константы $c_3 > 0$.

Для τ, удовлетворяющего условию $\tau_0 \leqslant \tau \leqslant t-\tau_0-1$, применим теорему 4:

$$M\left(S_{ij}^{(2)}(t)\right) = \sum_{\tau=\lfloor\log\log t\rfloor}^{t-\lfloor\log\log t\rfloor-1} \frac{p_{ij}}{t^{q_1-1}} \left(\sum_{l=1}^{k} f_{il} \cdot (t-\tau)^{q_l'-1} \cdot \tau^{\delta_{il}^1} + \right.$$

$$\left. \frac{1}{r} \sum_{m=1}^{k} f_m \cdot s_m^{(ij)} \cdot (t-\tau)^{q_m'-1} \tau^{s_{1i}'-1}\right) + \sum_{\tau=\lfloor\log\log t\rfloor}^{t-\lfloor\log\log t\rfloor-1} \xi(t,\tau).$$

Оценим величину $\delta = \frac{1}{t^{n_1+n_2}} \cdot \sum_{\tau=\lfloor\log\log t\rfloor}^{t-\lfloor\log\log t\rfloor-1} (t-\tau)^{n_1} \cdot \tau^{n_2}$:

$$\delta = \sum_{\tau=\lfloor\log\log t\rfloor}^{t-\lfloor\log\log t\rfloor-1} \left(1-\frac{\tau}{t}\right)^{n_1} \left(\frac{\tau}{t}\right)^{n_2} =$$

$$\sum_{\tau=\lfloor\log\log t\rfloor}^{t-\lfloor\log\log t\rfloor-1} \sum_{n=0}^{n_1} (-1)^n C_{n_1}^n \left(\frac{\tau}{t}\right)^{n+n_2} = \left(\sum_{n=0}^{n_1} (-1)^n C_{n_1-1}^n \cdot \frac{t}{n+n_2+1}\right) \cdot (1+o(1)).$$

Очевидно, величина $\sum_{n=0}^{n_1} (-1)^n C_{n_1-1}^n \cdot \frac{1}{n+n_2+1}$ является константой, зависящей от n_1 и n_2, обозначим ее $\alpha(n_1,n_2)$. Применяя обозначение $\alpha(n_1,n_2)$, мы можем записать:

$$\delta = \alpha(n_1,n_2) \cdot t \cdot (1+o(1)).$$

Применим полученную оценку к вычислению $M\left(S_{ij}^{(2)}(t)\right)$, учитывая равенства $q_l' + \delta_{il}^1 = q_1$ и $q_i' + s_{1i}' - 1 = q_1$:

$$M\left(S_{ij}^{(2)}(t)\right) = p_{ij} \cdot \left[\sum_{l=1}^{k} f_{il} \cdot \alpha(q_l'-1, \delta_{il}^1) + \frac{1}{r} \sum_{m=1}^{k} f_m \cdot s_m^{(ij)} \cdot \alpha(q_m'-1, s_{1i}'-1)\right] t \cdot (1+o(1)).$$

Константу в квадратных скобках обозначим w_{ij}.

Применяя полученные оценки для $M\left(S_{ij}^{(1)}(t)\right)$, $M\left(S_{ij}^{(2)}(t)\right)$ и $M\left(S_{ij}^{(3)}(t)\right)$, находим, что при $t \to \infty$

$$M\left(\frac{S_{ij}(t)}{t}\right) = w_{ij} + o(1) + O\left(\frac{\log^{c_2} t}{t}\right) + O\left(\frac{\log^{c_3} t}{t}\right) = w_{ij} + o(1).$$

Теорема 5 доказана.

Энтропия и нижняя оценка стоимости кодирования

Пусть L - стохастический язык, т.е. язык, на множестве слов которого задано распределение вероятностей.

Под энтропией стохастического языка L будем понимать величину

$$H(L) = -\lim_{N\to\infty} \sum_{\alpha\in L, |\alpha|\leqslant N} p(\alpha)\log p(\alpha).$$

Если энтропия конечна, будем применять запись $H(L) = -\sum_{\alpha\in L} p(\alpha)\log p(\alpha)$.

Кодированием языка L назовем инъективное отображение

$$f : L \to \{0,1\}^+.$$

В качестве L рассмотрим язык, порождаемый стохастической КС-грамматикой с однозначным выводом, т.е. грамматикой, в которой каждое слово из L имеет единственное дерево вывода. Через L^t обозначим множество всех слов из L, каждое из которых имеет дерево вывода высоты t. Для $\alpha \in L^t$ через $p_t(\alpha)$ обозначим условную вероятность появления слова α, т.е. $p_t(\alpha) = \frac{p(\alpha)}{P(L^t)}$. В силу однозначности вывода $P(L^t) = P(D_1^t)$.

Стоимостью кодирования f назовем величину

$$C(L,f) = \lim_{t\to\infty} \frac{\sum_{\alpha\in L^t} p_t(\alpha)\cdot|f(\alpha)|}{\sum_{\alpha\in L^t} p_t(\alpha)\cdot|\alpha|} \tag{11}$$

(здесь $|x|$ -длина последовательности x).

Величина $C(L,f)$ характеризует число двоичных разрядов, приходящихся на кодирование одного символа слова языка.

Через $F(L)$ обозначим класс всех инъективных отображений из L в $\{0,1\}^+$, для которых существует $C(L,f)$.

Стоимостью оптимального кодирования языка L назовем величину

$$C_0(L) = \inf_{f\in F(L)} C(L,f).$$

Предварительно получим асимптотическую формулу для энтропии множества слов L^t. По определению имеем

$$H(L^t) = -\sum_{\alpha \in L^t} p_t(\alpha) \log p_t(\alpha).$$

Следовательно,

$$H(L^t) = -\sum_{\alpha \in L^t} p_t(\alpha) \left(\log p(\alpha) - \log P(L^t)\right) =$$

$$\frac{1}{P(L^t)} \cdot \left(-\sum_{\alpha \in L^t} p(\alpha) \log p(\alpha) \right) + \log P(L^t).$$

Для слова α обозначим через $q_{ij}(\alpha)$ число применений правила r_{ij} при его выводе. Вероятность слова α равна $p(\alpha) = \prod_{i=1}^{k} \prod_{j=1}^{n_i} (p_{ij})^{q_{ij}}$. Следовательно, $\log p(\alpha) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} q_{ij}(\alpha) \log p_{ij}$. Поэтому

$$H(L^t) = \frac{1}{P(L^t)} \cdot \left(-\sum_{\alpha \in L^t} p(\alpha) \cdot \sum_{i=1}^{k} \sum_{j=1}^{n_i} q_{ij}(\alpha) \log p_{ij} \right) + \log P(L^t) =$$

$$\frac{1}{P(L^t)} \cdot \left(-\sum_{i=1}^{k} \sum_{j=1}^{n_i} \log p_{ij} \cdot \sum_{\alpha \in L^t} p(\alpha) q_{ij}(\alpha) \right) + \log P(L^t).$$

Очевидно, что $\sum_{\alpha \in L^t} p(\alpha) q_{ij}(\alpha) = P(L^t) \cdot M(S_{ij}(t))$. Используя теорему 5, выражение для энтропии можно переписать в виде

$$H(L^t) = -t \cdot (1 + o(1)) \sum_{i=1}^{k} \sum_{j=1}^{n_i} w_{ij} \log p_{ij} + \log P(L^t).$$

Ввиду однозначности вывода, с использованием (3), имеем

$$\log P(L^t) = \log P(D_1^t) = t \log r + O(\log t).$$

Поэтому

$$H(L^t) = t \cdot \left(\log r - \sum_{j=1}^{n_i} w_{ij} \log p_{ij} \right) + o(t).$$

Полученный результат сформулируем в виде следующей теоремы.

Теорема 6. *Пусть G — однозначная стохастическая КС-грамматика с разложимой матрицей первых моментов, для которой перронов корень $r < 1$, и L^t - множество всех слов из L, порождаемого G, с деревьями вывода высоты t. Тогда*

$$H(L^t) = t \cdot \left(\log r - \sum_{j=1}^{n_i} w_{ij} \log p_{ij} \right) + o(t),$$

где w_{ij} определяются теоремой 5.

Таким образом, энтропия $H(L^t)$ линейно зависит от высоты t дерева вывода, как и в неразложимом случае [5].

Используя энтропию, оценим стоимость оптимального кодирования $C_0(L)$. Обозначим через f^* кодирование множества L^t, минимизирующее величину

$$M_t(f) = \sum_{\alpha \in L^t} p_t(\alpha) \cdot |f(\alpha)| .$$

Очевидно, для любого кодирования $f \in F(L)$ верно неравенство $M_t(f) \geqslant M_t(f^*)$. Оценим $M^*(L^t) = M_t(f^*)$, используя следующую теорему, доказанную в [2].

Теорема 7. Пусть L_k – последовательность стохастических языков, для которой $H(L_k) \to \infty$ при $k \to \infty$. Тогда

$$\lim_{k \to \infty} \frac{M^*(L_k)}{H(L_k)} = 1.$$

Поскольку $H(L)^t \to \infty$ при $t \to \infty$, из теоремы 7 следует, что $M_t(f^*)/H(L^t) \to 1$ при $t \to \infty$.

Найдем величину $\sum_{\alpha \in L^t} p_t(\alpha) \cdot |\alpha|$. Пусть правило r_{ij} содержит в правой части l_{ij} терминальных символов. Очевидно, $|\alpha| = \sum_{ij} q_{ij}(\alpha) \cdot l_{ij}$. Поэтому

$$\sum_{\alpha \in L^t} p_t(\alpha) \cdot |\alpha| = \sum_{ij} l_{ij} M(S_{ij}(t)) = t \cdot \sum_{ij} l_{ij} w_{ij} + o(t).$$

Следовательно, справедлива

Теорема 8. *Пусть L - стохастический КС-язык, порожденный разложимой стохастической КС - грамматикой с однозначным выводом, для*

которой перронов корень r матрицы первых моментов меньше 1. *Тогда стоимость любого кодирования $f \in F(L)$ удовлетворяет неравенству*

$$C(L, f) \geqslant C_0(L) = \frac{\log r - \sum_{ij} w_{ij} \log p_{ij}}{\sum_{ij} l_{ij} w_{ij}}.$$

ЛИТЕРАТУРА

1. **Ахо А., Ульман Дж.** Теория синтаксического анализа, перевода и компиляции. Том 1. М.: Мир, 1978.
2. **Борисов А. Е.** Кодирование слов стохастического КС-языка, порожденного разложимой грамматикой с двумя нетерминалами // Вестник Нижегородского университета им. Н.И. Лобачевского. Серия Математика, 2004. Выпуск 1(2) С. 18–28.
3. **Гантмахер Ф. Р.** Теория матриц. М.: Наука, 1967.
4. **Жильцова Л. П.** Закономерности применения правил грамматики в выводах слов стохастического контекстно-свободного языка // Математические вопросы кибернетики. М.: Наука. 2000. Вып.9. С. 101–126.
5. **Жильцова Л. П.** О нижней оценке стоимости кодирования и асимптотически оптимальном кодировании стохастического контекстно-свободного языка // Дискретный анализ и исследование операций. 2001. Серия 1. Том 8, N3. Новосибирск: Издательство Института математики СО РАН. С. 26–45.
6. **Жильцова Л. П.** О матрице первых моментов разложимой стохастической КС-грамматики // Ученые записки Казанского государственного университета. Физико-математические науки. Том 151, книга 2, 2009. С. 80–89.
7. **Жильцова. Л. П.** О вероятностях продолжения деревьев вывода в разложимых стохастических КС-грамматиках. Докритический случай // Вестник Нижегородского университета им. Н.И. Лобачевского, № 4, 2012, № 4,. С. 217–224.
8. **Севастьянов В. А.** Ветвящиеся процессы. М.: Наука, 1971.
9. **Фу К.** Структурные методы в распознавании образов. М.: Мир, 1977.

Жильцова Лариса Павловна
Мартынов Игорь Михайлович

Статья поступила
** ** 20** г.

Исправленный вариант —
** ** 20** г.

DISKRETNYI ANALIZ I ISSLEDOVANIE OPERATSII
******** 2000. Volume 55, No. 10. P. 3–29

UDC 999.9

ON PROPERTIES OF PROBABILISTIC CHARACTERISTICS OF DERIVATION TREES IN STOCHASTIC CF-GRAMMARS

L. P. Zhiltsova[1], *I. M. Martynov*[1]

[1]Sobolev Institute of Mathematics,
4 Acad. Koptyug Ave., 630090 Novosibirsk, Russia
[2] Novosibirsk State University,
2 Pirogov St., 630090 Novosibirsk, Russia

E-mail: larzhil@rambler.ru, murbidodrus@gmail.com

Abstract. !!!TRANSLATE PROPERLY!!! Coding coding data compression CF-grammars bla-bla-bla...

Keywords: coding theory, CF-grammar, derivation tree, spectral radius, ...

Larisa P. Zhiltsova
Igor M. Martynov

Received
** ** 20**

Revised
** ** 20**