О вероятностных свойствах деревьев вывода для разложимой стохастической КС-грамматики, имеющей вид «цепочки»

Л.П. Жильцова, И.М. Мартынов (Нижний Новгород)

В работе исследуются вероятностные свойства деревьев вывода высоты t при $t \to \infty$ для стохастической КС-грамматики специального вида. Рассматривается случай, когда матрица первых моментов A грамматики разложима и имеет перронов корень равный 1. Целью работы ялвляется обобщение результатов, полученных в [2], на случай неограниченного числа классов нетерминалов.

Стохастической КС–грамматикой называется система $G = \langle V_N, V_N, R, s\rangle$, где V_T и V_N — конечные алфавиты терминальных и нетерминальных символов соответственно, $s \in V_N$ — аксиома, $R = \cup_{i=1}^{k} R_i$, где k — мощность алфавита V_N и R_i — множество правил с одинкаовой левой частью вида

$$r_{ij} : A_i \xrightarrow{p_{ij}} \beta_{ij},\ j = 1, 2, \ldots, n,$$

где $A_i \in V_N$, $\beta_{ij} \in (V_T \cup V_N)^*$ и p_{ij} — вероятность применения правила r_{ij}, причём $0 < p_{ij} \leqslant 1$ и $\sum_{j=1}^{n_i} p_{ij} = 1$.

Применение правила грамматики к слову состоит в замене вхождения нетерминала из левой части правила на слово, стоящее в его правой части.

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

По стохастической КС-грамматике строится матрица A первых моментов. Для неё элемент a_j^i определяется как $\sum_{l=1}^{n_i} p_{il} s_{il}^j$, где величина s_{il}^j равна числу нетерминальных символов A_j в правой части правила r_{il}. Перронов корень матрицы A обозначим через r.

Введём 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 $A_i \xrightarrow{p_{il}} \alpha_1 A_j \alpha_2$, где $\alpha_1, \alpha_2 \in (V_T \cup V_N)^*$. Рефлексивное транзитивное замыкание отношения $\to$ обозначим $\to_*$.

Классом нетерминалов назовём 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 нетерминалов 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 $\prec_*$.

Пусть $\mathfrak{K} = \{K_1, K_2, \ldots, K_m\}$ — множество классов нетерминалов грамматики, $m \geqslant 2$. Будем полагать, что классы нетерминалов перенумерованы таким образом, что $K_i \prec K_j$ тогда и только тогда, когда $i < j$.

Будем говорить, что грамматика имеет вид «цепочки», если её матрица первых моментов A имеет вид

$$A = \begin{pmatrix} A_{11} & A_{12} & 0 & \cdots & 0 & 0 \\ 0 & A_{22} & A_{23} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & A_{n-1,n-1} & A_{n-1,n} \\ 0 & 0 & 0 & \cdots & 0 & A_{n,n} \end{pmatrix}$$

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

Для каждого класса K_i матрица A_{ii} неразложима. Без ограничения общности будем считать, что она строго положительна и непериодична. Обозначим через r_i перронов корень матрицы A_{ii}. Для неразложимой матрицы перронов корень является вещественным и простым [1]. Очевидно, $r = \max_i\{r_i\}$.

Пусть $J = \{i_1, i_2, \ldots, i_l\}$ — множество всех номеров i_j классов, для которых $r_{ij} = 1$.

Рассмотрим всевозможные подцепочки классов $K_\mu, K_{\mu+1}, \ldots, K_\nu$. Число классов с номерами из J в такой подцепочке обозначим через $s_{\mu\nu}$.

Через $P_i(t)$ обозначим вероятность множества деревьев вывода высоты t, корень которых помечен нетерминалом A_i.

Теорема 1 *Пусть $r_j = 1$ для любого j. Пусть, кроме этого, некоторый нетерминал A_q принадлежит классу K_μ, и в грамматике имеется некоторое правило r_{ij}, такое что $A_i \in K_\nu$, и $\mu \leqslant \nu$. Тогда при $t \to \infty$ математическое ожидание числа применений правила r_{ij} в деревьях вывода высоты t, корень которых помечен нетерминалом A_q, выражается следующим образом:*

$$M_{ij}^q(t) = \frac{\mathcal{M}_\mu u^{(\mu)}_{q-k_{\mu-1}}}{P_q(t)} \cdot t^{-1+\left(\frac{1}{2}\right)^{m-\nu}\left(2-\left(\frac{1}{2}\right)^{\nu-\mu}\right)}(1+o(1)), \qquad (1)$$

где $\mathcal{M}_\mu$ — некоторая константа, соответствующая номеру класса μ, k_μ — число нетерминалов в классах $K_1, \ldots, K_{\mu-1}$, и $u^{(\mu)}$ — правый собственный вектор матрицы $A_{\mu,\mu}$.

Список литературы

[1] Гантмахер Ф.Р. **Теория матриц.** — 5-е изд., — М.: ФИЗМАТЛИТ, 2010 — 560 с. — ISBN 978-5-9221-0524-8

[2] Борисов А.Е. Закономерности в словах стохастических контекстно-свободных языков, порождённых грамматиками с двумя классами нетерминальных символов. Вопросы экономного кодирования.