О ЧАСТОТНЫХ ЯЗЫКАХ НА БИГРАММАХ

Петюшко А.А. (Московский Государственный Университет им. М. В. Ломоносова)

petsan@newmail.ru

Пусть A ($|A| < \infty$) - конечный алфавит, а $L \subseteq A^*$ - некоторый язык над этим алфавитом.

По каждому слову α языка L можно построить матрицу биграмм $(n(\alpha))_{a,b\in A}$, такую что $n_{ab}(\alpha)$ - это число рядом рядом стоящих букв ab в слове α. В данной статье решается обратная задача - по матрице $n(\alpha)$ установить некоторые свойства языка $L(n(\alpha))$, то есть множества всех слов, имеющих матрицу биграмм $n(\alpha)$. Полученные языки $L(n(\alpha))$ удается классифицировать.

Пример. Пусть $A = \{0, 1\}$, $\alpha = 01011100$.
Тогда матрица биграмм $n(\alpha) = \begin{pmatrix} n_{00}(\alpha) & n_{01}(\alpha) \\ n_{10}(\alpha) & n_{11}(\alpha) \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 2 & 2 \end{pmatrix}$.

Рассмотрим сначала результат, касающийся регулярности языка, в котором заданы некоторые ограничения на какое-то подмножество элементов матрицы биграмм.

Теорема 1. *Пусть задан набор $k < \infty$ биграмм $\overline{\beta} = (\beta_1, ..., \beta_k)$, где $|\beta_i| = 2, i = 1..k$, а также набор отрезков $\overline{c} = ([c_1^1, c_2^1], ..., [c_1^k, c_2^k])$, где $c_1^i \leq c_2^i, c_j^i \in N \cup \{0\}, i = 1..k, j = 1..2$. Тогда язык $L_{\overline{\beta},\overline{c}} = \{\alpha| \quad n_{\beta_i}(\alpha) \in [c_1^i, c_2^i], i = 1..k\}$ регулярен.*

Более интересный случай, когда мы рассматриваем матрицу биграмм не как абсолютное ограничение, а как задание относительных значений биграмм, то есть языка, в котором сохраняются отношения $n_{ab}(\alpha)/n_{cd}(\alpha) \quad \forall a, b, c, d \in A, n_{cd}(\alpha) > 0$. Для более детального рассмотрения нам потребуется ряд определений.

Определение. *Назовем частотным языком на биграммах, заданным матрицей биграмм $n(\alpha)$, следующий язык при $k \in N$:*

$$F_{\cup n(\alpha)} = \bigcup_{k=1}^{\infty} L(kn(\alpha)).$$

Построим по матрице $n(\alpha)$ ориентированный граф $G_{n(\alpha)}$ на плоскости. Вершинами у этого графа будут все буквы из алфавита A, при этом ребра будут соответствовать биграммам с учетом их кратностей, то есть кратность $n_{ab}(\alpha)$ будет порождать $n_{ab}(\alpha)$ ориентированных ребер $a \to b$. Аналогично, кратность $n_{cc}(\alpha)$ будет порождать $n_{cc}(\alpha)$ петель $c \to c$.

Определение. *Назовем ориентированный граф эйлеровым, если если выполняются следующие условия: 1) Все вершины, являющиеся начальной или конечной вершиной хотя бы одного ребра, лежат в одной компоненте связности соответствующего неориентированного графа; 2) У всех вершин количество входящих ребер равно количеству исходящих ребер.*

Определение. *Назовем ориентированный граф почти эйлеровым, если выполняются следующие условия: 1) Все вершины, являющиеся начальной или конечной вершиной хотя бы одного ребра, лежат в одной компоненте связности соответствующего неориентированного графа; 2) У всех вершин, кроме двух, количество входящих ребер равно количеству исходящих ребер. У оставшихся двух вершин разность количества входящих ребер и количества исходящих ребер равна +1 и -1 соответственно.*

Как показано в [1], в эйлеровом графе существует эйлеров цикл (то есть такой цикл, который содержит все ребра, причем каждое - только один раз), а в почти эйлеровом - эйлеров путь, не являющийся эйлеровым циклом (то есть такой путь, который содержит все ребра, причем каждое - только один раз, и при этом начальная вершина не совпадает с конечной).

Теорема 2. *Пусть задана матрица биграмм $n(\alpha)$. Тогда:*
1) Если ориентированный граф $G_{n(\alpha)}$ является эйлеровым, то в частотном языке $F_{\cup n(\alpha)}$ счетное число слов;
2) Если ориентированный граф $G_{n(\alpha)}$ является почти эйлеровым, то в частотном языке $F_{\cup n(\alpha)}$ конечное ненулевое число слов, имеющих одинаковую длину;
3) Если ориентированный граф $G_{n(\alpha)}$ не является ни эйлеровым, ни почти эйлеровым, то в частотном языке $F_{\cup n(\alpha)}$ нет ни одного слова.

Очевидно, что если выполняются условия 2) или 3) Теоремы 2, то язык $F_{\cup n(\alpha)}$, в котором не более чем конечное число слов, будет регулярным. Поэтому интересен вопрос, когда он будет являться регулярным при условии 1).

Определение. *Назовем две ненулевые матрицы n_1 и n_2 одинакового размера неколлинеарными, если не существует ненулевых действительных коэффициентов $c_1, c_2 \in R, (c_1, c_2) \neq (0, 0)$, таких,* *что верно* $c_1 n_1 + c_2 n_2 = 0$.

Теорема 3. *Пусть* $A, |A| < \infty$ *- некоторый конечный алфавит. Далее, пусть задана матрица биграмм* $n(\alpha)$ *такая, что соответствующий ей ориентированный граф* $G_{n(\alpha)}$ *является эйлеровым. Тогда:*
1) Если существует такое разложение $n(\alpha)$ *в сумму двух ненулевых неколлинеарных матриц* $n(\alpha) = n(\alpha_1) + n(\alpha_2)$ *такое, что обе матрицы* $n(\alpha_1)$ *и* $n(\alpha_2)$ *задают ориентированные графы* $G_{n(\alpha_1)}$ *и* $G_{n(\alpha_2)}$*, которые являются эйлеровыми, то язык* $F_{\cup n(\alpha)}$ *нерегулярен;*
2) В противном случае язык $F_{\cup n(\alpha)}$ *регулярен.*

Однако данная теорема дает слишком общие условия на матрицу биграмм. Рассмотрим частный, но часто используемый на практике случай двухбуквенного алфавита.

Теорема 4. *Пусть* $A = \{0, 1\}$. *Далее, пусть задана матрица биграмм* $n(\alpha)$ *такая, что соответствующий ей ориентированный граф* $G_{n(\alpha)}$ *является эйлеровым. Тогда:*
1) Язык $F_{\cup n(\alpha)}$ *нерегулярен, если* $\exists i, i \in \{0, 1\}$ *такое, что* $n_{ii}(\alpha) > 0$, *и при этом* $\exists u \neq v, u, v \in \{0, 1\}$ *такие, что* $n_{uv}(\alpha) > 0$;
2) Язык $F_{\cup n(\alpha)}$ *регулярен, если* $\exists i, i \in \{0, 1\}$ *такое, что* $n_{ii}(\alpha) > 0$, *и при этом* $\forall u, v \in \{0, 1\}, (i, i) \neq (u, v)$ *выполняется* $n_{uv}(\alpha) = 0$;
3) Язык $F_{\cup n(\alpha)}$ *регулярен при* $n_{00}(\alpha) = n_{11}(\alpha) = 0$.

Отметим, что для доказательства двух последних теорем напрямую использовалась теорема Клини о представимости регулярных событий в автомате (см. [2]).

Автор выражает благодарность своему научному руководителю, д. ф.-м. н., профессору Бабину Д. Н., за постановку задачи и ценные указания.

Литература

1. Оре О. Теория графов. – М.: Наука, 1980.

2. Кудрявцев В. Б., Алешин С. В., Подколзин А. С. Введение в теорию автоматов. – М.: Наука, 1985.