КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА

УДК 519.007

А. А. Петюшко, аспирант
Московский государственный университет имени М. В. Ломоносова

ЧАСТОТНЫЕ ЯЗЫКИ

Рассматриваются как конечные языки, заданные матрицей биграмм, так и введенные частотные языки на матрице биграмм, характеризумые тем, что язык состоит из всех слов с матрицей биграмм, кратной (т. е. умноженной на натуральное число) исходной матрице биграмм. Устанавливается связь различных характеристик частотных языков с ориентированными графами и эйлеровыми циклами в них. Приводятся необходимые и достаточные условия для непустоты и счетности частотных языков. Рассматривается вопрос зависимости мощности частотного языка от исходной матрицы биграмм. Приведена формула для числа слов в зависимости от матрицы биграмм. Устанавливаются условия регулярности счетных частотных языков.

Ключевые слова: матрица биграмм, частотные языки, регулярность языков, эйлеровы циклы

Необходимые определения

Цель работы – исследовать различные множества слов с ограничениями на кратность входящих в них подслов (а именно биграмм). В частности, будут исследованы такие свойства, как условие пустоты таких языков, количество слов в них, условия регулярности и т. п. Для этого введем ряд необходимых понятий и обозначений.

Пусть A $(|A|<\infty)$ – конечный алфавит.

Определение 1. *Биграммой в алфавите A называется двухбуквенное слово $ab \in A^*, a,b \in A$ (порядок вхождения букв в биграмму имеет значение, например, биграмма ab не равна биграмме ba при $a \neq b$).*

Определение 2. *Обозначим через $n_\beta(\alpha)$, где $\beta \in A^*, \alpha \in A^*$, причем β – непустое слово, отображение $A^* \to N \cup \{0\}$, которое определяется как количество подслов β в слове α, т. е. количество различных разложений слова α в виде $\alpha = \alpha'\beta\alpha''$ (α' и α'' могут быть пустыми). При длине слова α, меньшем, чем длина слова β, значение $n_\beta(\alpha)$ положим равным 0. Само же значение $n_\beta(\alpha)$ при данных β и α назовем кратностью β в слове α.*

С учетом введенных определений, по каждому слову $\alpha \in A^*$ можно построить квадратную матрицу биграмм $(n(\alpha))_{i,j=1}^{|A|}$ размера $|A|\times|A|$, такую, что на месте (i,j) матрицы будет стоять значение $n_{a_i a_j}(\alpha)$ (при условии, что все буквы алфавита $A = \{a_1, a_2, ..., a_{|A|}\}$ пронумерованы).

Определение 3. *Назовем языком* $L(n(\alpha))$, *порожденным матрицей биграмм* $n(\alpha)$, *множество всех слов, имеющих одну и ту же матрицу биграмм* $n(\alpha)$, *т. е.* $L(n(\alpha)) = \{\beta \mid n(\beta) = n(\alpha)\}$.

Пример. Пусть $A = \{0,1\}$, $\alpha = 01011100$.

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

Регулярность языков на биграммах с кратностями

Рассмотрим сначала задачу с достаточно простыми ограничениями на кратности биграмм. Здесь и далее будем использовать обозначение $len(\beta)$ в качестве длины слова β (количества входящих в него букв).

Лемма 1. *Язык* $L^{+}_{\beta,c}$ *с ограничением вида* $n_\beta(\alpha) \ge c$, *где* $c \in N \cup \{0\}$, *а* $len(\beta) = 2$ *– регулярен.*

Лемма 2. *Язык* $L^{-}_{\beta,c}$ *с ограничением вида* $n_\beta(\alpha) \le c$, *где* $c \in N \cup \{0\}$, *а* $len(\beta) = 2$ *– регулярен.*

Замечание. В частности, взяв язык $L^{-}_{\beta,0}$, получим множество всех слов, в которых нет подстроки β.

Теорема 1. *Пусть задан набор* k *слов* $\bar{\beta} = (\beta_1,...,\beta_k)$, *где* $len(\beta_i) = 2, i = 1..k$, *а также набор отрезков* $\bar{c} = ([c_1^1, c_2^1],...,[c_1^k, c_2^k])$, *где* $c_1^i \le c_2^i, c_j^i \in N \cup \{0\}$, $i = 1..k, j = 1..2$. *Тогда язык* $L^{\bar{-}}_{\bar{\beta},\bar{c}}$ *с* k *ограничениями вида* $n_{\beta_i}(\alpha) \in [c_1^i, c_2^i]$ *– регулярен.*

Следствие 1. *В условиях Теоремы 1 можно изменить некоторые ограничения, заменив принадлежность отрезкам на принадлежность полуоткрытым и открытым интервалам (в т.ч. бесконечным справа), а также на точное равенство числу* $c, c \in N \cup \{0\}$, *при этом регулярность языка* $L^{\bar{-}}_{\bar{\beta},\bar{c}}$ *не нарушится.*

Замечание. Условие Теоремы 1 остается верным даже в том случае, если в качестве подстроки брать одиночный символ, т.е. $len(\beta) = 1$. Пусть $\beta = a, a \in A$. Тогда условие $n_a(\alpha) \ge c$ порождает множество

$$< A \setminus \{a\} > \cdot a \cdot < A \setminus \{a\} > \cdot a \cdot < A \setminus \{a\} > \cdot \ldots \cdot a \cdot < A >$$

(в записи c раз встречается фрагмент $\cdot a \cdot$), которое является регулярным, и, значит, соответствующий ему язык $L^{+}_{a,c}$ тоже. Остальные построения проводятся аналогично оным для случая биграмм.

Свойства матрицы $n(\alpha)$

Пусть $a \in A$. Обозначим за $n_a^{in}(\alpha), \alpha \in A^*$ количество всех двухбуквенных подслов в α, заканчивающихся на a, т. е. $n_a^{in}(\alpha) = \sum_{b \in A} n_{ba}(\alpha)$, что будет равно сумме значений матрицы $n(\alpha)$ в столбце, соответствующем символу a.

Аналогичным образом определим и $n_a^{out}(\alpha), \alpha \in A^*$ как количество двухбуквенных подслов в α, начинающихся на a, т. е. $n_a^{out}(\alpha) = \sum_{b \in A} n_{ab}(\alpha)$, что будет равно сумме значений матрицы $n(\alpha)$ в строке, соответствующей символу a.

Пример. Рассмотрим то же слово, что и в предыдущем примере, а именно $A = \{0,1\}$, $\alpha = 01011100$.

Тогда соответствующие значения будут вычисляться как

$$n_1^{in}(\alpha) = n_{01}(\alpha) + n_{11}(\alpha) = 2 + 2 = 4;$$

$$n_1^{out}(\alpha) = n_{10}(\alpha) + n_{11}(\alpha) = 2 + 2 = 4;$$

$$n_0^{in}(\alpha) = n_{00}(\alpha) + n_{10}(\alpha) = 1 + 2 = 3;$$

$$n_0^{out}(\alpha) = n_{00}(\alpha) + n_{01}(\alpha) = 1 + 2 = 3.$$

Рассмотрим некоторые свойства матрицы $n(\alpha)$.

Лемма 3 (Условие неразрывности). *Пусть задано слово* $\alpha = a_1 \beta a_2$, $a_i \in A, i = 1,2, \beta \in A^*$ *длины не менее 2 (*$len(\alpha) \geq 2$*), где* a_1 *и* a_2 *– соответственно первая и последняя буквы этого слова (слово* β *может быть пустым). Тогда матрица* $n(\alpha)$ *обладает следующим свойством:*

$$\forall b \in A \quad n_b^{out}(\alpha) - n_b^{in}(\alpha) = \delta_{ba_1} - \delta_{ba_2}, \tag{1}$$

где δ_{ij} *– символ Кронекера (*$\delta_{ij} = 1$ *при* $i = j$, $\delta_{ij} = 0$ *при* $i \neq j$*).*

Пример. Продолжим рассматривать все то же слово, а именно $A = \{0,1\}$, $\alpha = 01011100$. Проверим условие неразрывности.

Имеем $a_1 = a_2 = 0$. Тогда

$$n_0^{out} - n_0^{in} = 3 - 3 = 0,\ \delta_{0a_1} - \delta_{0a_2} = 1 - 1 = 0, \text{ верно;}$$

$$n_1^{out} - n_1^{in} = 4 - 4 = 0,\ \delta_{1a_1} - \delta_{1a_2} = 0 - 0 = 0, \text{ верно.}$$

Замечание. Условие неразрывности (1) является необходимым, но не достаточным условием существования хотя бы одного слова α с заданным набором значений биграмм $n(\alpha)$. Например, если $n_{00}(\alpha) = n_{11}(\alpha) = 1$, $n_{01}(\alpha) = n_{10}(\alpha) = 0$ в алфавите $A = \{0,1\}$, то очевидно, что слова α с данным набором $n(\alpha)$ не существует – хотя бы потому, что где-то в слове должны рядом стоять буквы 1 и 0, и должна быть ненулевая кратность $n_{01}(\alpha)$ или $n_{10}(\alpha)$ (хотя у нас обе кратности нулевые). При этом

$$n_0^{in}(\alpha) = n_0^{out}(\alpha) = n_1^{in}(\alpha) = n_1^{out}(\alpha) = 1,$$

и если бы слово α начиналось и заканчивалось на одну и ту же букву, то условие неразрывности было бы выполнено.

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, т. е. 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 $c \to c$.

Пример. $A = \{0,1\}$, $\alpha = 01011100$.

$$n(\alpha) = \begin{pmatrix} n_{00}(\alpha) & n_{01}(\alpha) \\ n_{10}(\alpha) & n_{11}(\alpha) \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 2 & 2 \end{pmatrix}.$$

Построим граф $G_{n(\alpha)}$ по $n(\alpha)$ – см. рис. 1.

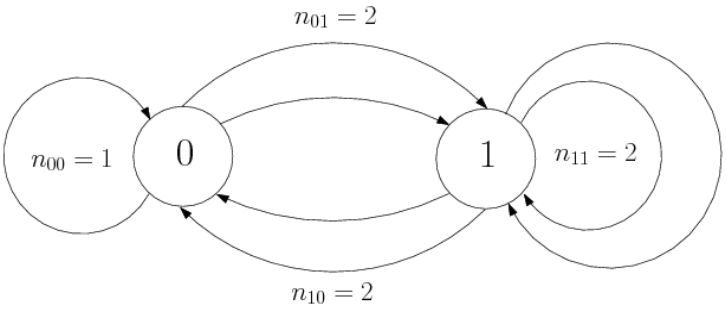

Рис. 1. Граф $G_{n(\alpha)}$, построенный по набору $n(\alpha)$

Введем несколько широко известных понятий, касающихся эйлеровых путей.

Определение 4. *Назовем ориентированный граф 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 количество входящих ребер равно количеству исходящих ребер.*

Определение 5. *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.*

Определение 6. *Путем в ориентированном графе назовем такую последовательность попарно различных ребер (набор параллельных ребер будем считать состоящим из различных ребер), что конец предыдущего ребра совпадает с началом следующего.*

Определение 7. *Циклом в ориентированном графе назовем такой путь, что начало первого ребра в этом пути совпадает с концом последнего.*

Определение 8. *Эйлеровым путем в ориентированном графе назовем такой путь, который содержит все ребра этого графа.*

Определение 9. *Эйлеровым циклом в ориентированном графе назовем такой цикл, который содержит все ребра этого графа.*

Как показано в [1], в эйлеровом графе существует эйлеров цикл, а в почти эйлеровом – эйлеров путь, не являющийся эйлеровым циклом (то есть начальная вершина не совпадает с конечной).

Лемма 4 (Достаточное условие существования). *Для того, чтобы существовало хотя бы одно слово* α *с данной матрицей кратностей биграмм* $n(\alpha)$, *достаточно, чтобы построенный по* $n(\alpha)$ *ориентированный граф* $G_{n(\alpha)}$ *был либо эйлеровым, либо почти эйлеровым.*

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

Построим граф $G_{n(\alpha)}$ по $n(\alpha)$ – см. рис. 1. В этом графе можно без труда найти эйлеров путь, например, $1 \to 0 \to 0 \to 1 \to 1 \to 1 \to 0 \to 1$ (мы получили другое слово 10011101, отличное от изначального $\alpha = 01011100$), при этом все вершины (а именно 0 и 1) находятся в одной компоненте связности. Значит, для данного набора $n(\alpha)$ есть слово с такой матрицей кратностей биграмм.

Как итог, получаем следующее важное следствие:

Следствие 2 (Алгоритмическая разрешимость). *Задача определения по набору значений* $n(\alpha)$, *существует ли хотя бы одно слово* α, *имеющее эту матрицу биграмм, алгоритмически разрешима.*

Напоследок рассмотрим вопрос о том, сколько существует слов с данным набором $n(\alpha)$. В данной статье рассмотрим рассмотрим случай двухбуквенного алфавита.

Теорема 2. *Для алфавита* $A = \{0,1\}$ *и матрицы биграмм* $n(\alpha)$, *задающей эйлеров или почти эйлеров ориентированный граф* $G_{n(\alpha)}$, *число слов* $N_{n(\alpha)}$ *с заданной матрицей биграмм* $n(\alpha)$:

1) *При* $n_{01}(\alpha) > n_{10}(\alpha)$ *количество* $N_{n(\alpha)} = C^{n_{11}}_{n_{11}+n_{10}} C^{n_{00}}_{n_{00}+n_{10}}$;

2) *При* $n_{01}(\alpha) < n_{10}(\alpha)$ *количество* $N_{n(\alpha)} = C^{n_{11}}_{n_{11}+n_{01}} C^{n_{00}}_{n_{00}+n_{01}}$;

3) *При* $n_{01}(\alpha) = n_{10}(\alpha)$ *количество*

$$N_{n(\alpha)} = C^{n_{00}}_{n_{00}+n_{01}} C^{n_{11}}_{n_{11}+n_{01}} \left(\frac{n_{01}}{n_{00}+n_{01}} + \frac{n_{01}}{n_{11}+n_{01}} \right);$$

(*здесь под* C^k_n *понимается число сочетаний из* n *по* k, *т. е.* $C^k_n = \frac{n!}{k!(n-k)!}$, *а* $n_{ij} = n_{ij}(\alpha), i, j = 0,1$).

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

Так как $n_{01}(\alpha) = n_{10}(\alpha)$, то искомое число слов с данным набором по доказанной выше теореме равно

$$N_{n(\alpha)} = C^{n_{00}}_{n_{00}+n_{01}} C^{n_{11}}_{n_{11}+n_{01}} \left(\frac{n_{01}}{n_{00}+n_{01}} + \frac{n_{01}}{n_{11}+n_{01}} \right) = C^1_3 C^2_4 \left(\frac{2}{3} + \frac{1}{2} \right) = 12 + 9 = 21.$$

Это действительно так, поскольку с данным набором $n(\alpha)$ существует ровно 21 слово: 11100101, 11001101, 11001011, 10011101, 10011011, 10010111, 11101001, 11011001, 11010011, 10111001, 10110011, 10100111, 00111010, 00110110, 00101110, 01110010, 01100110, 01001110, 01110100, 01101100, 01011100.

Частотные языки на биграммах с кратностями

Более интересен случай, когда мы рассматриваем матрицу биграмм не как абсолютное ограничение, а как задание относительных значений (пропорций) биграмм, то есть языка, в котором сохраняются отношения $n_{ab}(\alpha)/n_{cd}(\alpha)$ $\forall a,b,c,d \in A, n_{cd}(\alpha) > 0$. Определим такой язык.

Определение 10. *Назовем частотным языком на биграммах с кратностями, заданным матрицей биграмм* $n(\alpha)$*, следующий язык при* $k \in N$*:*

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

т. е. язык, состоящий из всех таких слов β, т.ч. набор кратностей этих слов $n(\beta)$ кратен набору $n(\alpha)$, а именно $F_{\cup n(\alpha)} = \{\beta \mid n(\beta) = kn(\alpha), k \in N\}$, где умножение k на $n(\alpha)$ понимается как умножение скаляра на матрицу.

Рассмотрим, какие дополнительные ограничения накладывает условие «частотности» на матрицу $n(\alpha)$.

Лемма 5 (Условие неразрывности для частотных языков). *Пусть задано слово* $\alpha = a_1 \alpha' a_2$, $a_i \in A, i = 1,2, \alpha' \in A^*$ *длины не менее 2 (*$len(\alpha) \geq 2$, α' *может быть пустым). Для того, чтобы в частотном языке* $F_{\cup n(\alpha)}$ *существовало хотя бы одно слово* $\beta = b_1 \beta' b_2$, $b_i \in A, i = 1,2, \beta' \in A^*$ *(*β' *может быть пустым), т. ч.* $n(\beta) = kn(\alpha), k \in N, k > 1$*, необходимо:*

$$\forall b \in A \quad n_b^{out}(\alpha) = n_b^{in}(\alpha), \quad n_b^{out}(\beta) = n_b^{in}(\beta), \tag{2}$$

при этом как в α*, так и в* β *первая буква должна совпадать с последней (*$a_1 = a_2$ *и* $b_1 = b_2$*).*

Следствие 3 (Корректность определения). *Если существует хотя бы одно слово* α*, соответствующее набору* $n(\alpha)$*, и при этом выполняется условие неразрывности для частотных языков (2), то существует слово* β_k *для любого натурального* k*, т.ч.* $n(\beta_k) = kn(\alpha)$.

Ну и как итог, приведем достаточный признак существования такого β_k, что для любого натурального k выполняется $n(\beta_k) = kn(\alpha)$.

Лемма 6 (Достаточное условие существования для частотных языков). *Для того, чтобы* $\forall k \in N$ *существовало слово* β_k*, т.ч. для заданного набора кратностей биграмм* $n(\alpha)$ *выполнялось* $n(\beta_k) = kn(\alpha)$*, достаточно, чтобы построенный по* $n(\alpha)$ *ориентированный граф* $G_{n(\alpha)}$ *являлся эйлеровым графом.*

Теорема 3 (О числе слов в частотных языках). *Пусть задан набор 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.*

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

Построим граф $G_{n(\alpha)}$ по $n(\alpha)$ – см. рис. 1. В этом графе в вершину 0 входит 3 ребра, исходит тоже 3, в вершину 1 входит 4 ребра и исходит тоже 4. Все вершины лежат в одной компоненте связности. Получается, что граф $G_{n(\alpha)}$ эйлеров, т. е. в Теореме 3 выполняется условие 1), и, соответственно, в частотном языке $F_{\cup n(\alpha)}$ счетное число слов.

Следствие 4 (Алгоритмическая разрешимость для частотных языков). *Задача определения по набору значений* $n(\alpha)$, *пуст, конечен или счетен частотный язык* $F_{\cup n(\alpha)}$, *алгоритмически разрешима.*

Регулярность частотных языков на биграммах с кратностями

В данном разделе попытаемся установить, регулярны ли бесконечные языки $F_{\cup n(\alpha)}$ с заданным набором $n(\alpha)$. Рассмотрим сначала общий случай с произвольным числом букв в алфавите A.

Определение 11. *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* $c_1 n_1 + c_2 n_2 = 0$.

Определение 12. *Назовем элементарной матрицей, соответствующей набору биграмм* $n(\alpha)$ *в алфавите* $A, |A| < \infty$, *матрицу* $\hat{n}(\alpha) = n(\alpha) / НОД(n_{ab}(\alpha), a, b \in A)$.

Теорема 4. *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. При этом для* $\forall k \in N$ *существуют ровно* l *слов* $\beta_{k,i}, i = 1..l$, *т.ч.* $n(\beta_{k,i}) = kn(\alpha)$, *а* l *– число ненулевых элементов в матрице* $n(\alpha)$.

Доказательство.

1) Доказывается от противного с использованием теоремы Клини [2].

2) В данном случае матрице биграмм будет соответствовать элементарная матрица с единицами на ненулевых позициях, неразложимая в сумму двух ненулевых неколлинеарных матриц, каждая их которых задает эйлеров граф, что и будет означать один цикл в графе с точностью до выбора начальной вершины.

Конец доказательства.

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

Теорема 5. *Пусть* $A = \{0,1\}$. *Далее, пусть задан такой набор* $n(\alpha)$, *что соответствующий ориентированный граф* $G_{n(\alpha)}$ *является эйлеровым. Тогда:*

1) *Если матрица биграмм* $n(\alpha)$ *имеет вид, не совпадающий ни с одним из перечисленных* $M_1 = \begin{pmatrix} c_1 & 0 \\ 0 & 0 \end{pmatrix}, M_2 = \begin{pmatrix} 0 & 0 \\ 0 & c_2 \end{pmatrix}, M_3 = \begin{pmatrix} 0 & c_3 \\ c_3 & 0 \end{pmatrix}$, *где* $c_i \in N, i = 1..3$, *то язык* $F_{\cup n(\alpha)}$ *нерегулярен;*

2) *Если матрица биграмм* $n(\alpha)$ *имеет вид, совпадающий с* M_1 *или* M_2, *то язык* $F_{\cup n(\alpha)}$ *регулярен. При этом для* $\forall k \in N$ *существует единственное* β_k, *т.ч.* $n(\beta_k) = kn(\alpha)$;

3) *Если матрица биграмм* $n(\alpha)$ *имеет вид, совпадающий с* M_3, *то язык* $F_{\cup n(\alpha)}$ *регулярен. При этом для* $\forall k \in N$ *существуют ровно два слова* β_k *и* γ_k, *т.ч.* $n(\beta_k) = n(\gamma_k) = kn(\alpha)$.

Доказательство.

1) Пусть матрица биграмм $n(\alpha)$ имеет вид, не совпадающий ни с одним из M_1, M_2 и M_3, и при этом соответствующий ориентированный граф $G_{n(\alpha)}$ является эйлеровым. С учетом Леммы 5 для двухбуквенного алфавита получаем $n_{01}(\alpha) = n_{10}(\alpha) = c > 0$. Значит, матрица биграмм $n(\alpha)$ имеет вид $n(\alpha) = \begin{pmatrix} d_1 & c \\ c & d_2 \end{pmatrix}$, где хотя бы один из элементов d_1 и d_2 больше нуля. Пусть, для определенности, $d_1 > 0$ (случай $d_2 > 0$ рассматривается аналогично).

Тогда можем представить исходную матрицу биграмм $n(\alpha)$ в виде суммы двух неколлинеарных матриц $n(\alpha) = \begin{pmatrix} d_1 & c \\ c & d_2 \end{pmatrix} = \begin{pmatrix} d_1 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & c \\ c & d_2 \end{pmatrix} = C_1 + C_2$. Заметим, что матрицы C_1 и C_2 задают эйлеровы графы.

Получается, что исходная матрица биграмм раскладывается в сумму двух неколлинеарных матриц, т.ч. соответствующие им ориентированные графы – эйлеровы. Значит, верно условие 1) Теоремы 4, и, следовательно, язык $F_{\cup n(\alpha)}$ нерегулярен.

2) Пусть для определенности $n_{00}(\alpha) = c_1 > 0$, и при этом $\forall u, v \in \{0,1\}, (u,v) \neq (0,0)$ выполняется $n_{uv}(\alpha) = 0$, т. е. матрица биграмм имеет вид M_1 (случай $n_{11}(\alpha) = c_2 > 0$ и вид матрицы биграмм M_2 рассматривается аналогично). Тогда язык $F_{\cup n(\alpha)}$ должен состоять только из таких слов β_k, $n(\beta_k) = kn(\alpha)$, т.ч. слово β_k $\forall k \in N$ имеет вид: $\beta_k = \underbrace{0...0}_{1+k*n_{00}(\alpha)}$. Очевидно, что все такие слова задаются одним регулярным выражением:

$$F_{\cup n(\alpha)} = 0 \cdot \underbrace{0 \cdot \ldots \cdot 0}_{n_{00}} \cdot < \underbrace{0 \cdot \ldots \cdot 0}_{n_{00}} >$$

и, значит, язык $F_{\cup n(\alpha)}$ регулярен.

Для визуализации воспользуемся представлением слова в обобщенном источнике, которые отличается от обычного источника возможностью помечать ребра пустым словом Λ, но при этом позволяется иметь только одну финальную вершину. Обобщенный источник, в котором начальная вершина помечена символом «*», а конечная – «**», изображен на рис. 2.

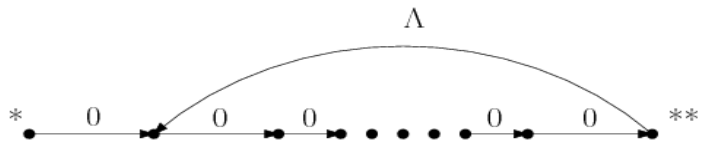

Рис. 2. Обобщенный источник для случая 2

3) Имеем $n_{01}(\alpha) = n_{10}(\alpha) = c_3 > 0$, $n_{00}(\alpha) = n_{11}(\alpha) = 0$, т. е. матрица биграмм имеет вид M_3. Значит, язык $F_{\cup n(\alpha)}$ должен состоять только из таких слов β_k, $n(\beta_k) = kn(\alpha)$, т. ч. слово $\beta_k \ \forall k \in N$ имеет один из двух возможных видов: либо $\beta_k = 1\underbrace{01...01}_{k*n_{01}(\alpha)}$, либо $\beta_k = 0\underbrace{10...10}_{k*n_{10}(\alpha)}$. В любом случае все такие слова задаются одним регулярным выражением:

$$F_{n^2(\alpha)} = 1 \cdot \underbrace{0 \cdot 1 \cdot \ldots \cdot 0 \cdot 1}_{n_{01}(\alpha)} \cdot < \underbrace{0 \cdot 1 \cdot \ldots \cdot 0 \cdot 1}_{n_{01}(\alpha)} > \bigcup 0 \cdot \underbrace{1 \cdot 0 \cdot \ldots \cdot 1 \cdot 0}_{n_{10}(\alpha)} \cdot < \underbrace{1 \cdot 0 \cdot \ldots \cdot 1 \cdot 0}_{n_{10}(\alpha)} >$$

и, значит, язык $F_{\cup n(\alpha)}$ регулярен.

Обобщенный источник, определяющий то же событие, что и приведенное выше регулярное выражение, изображен на рис. 3. Начальная вершина помечена символом «*», конечная – «**».

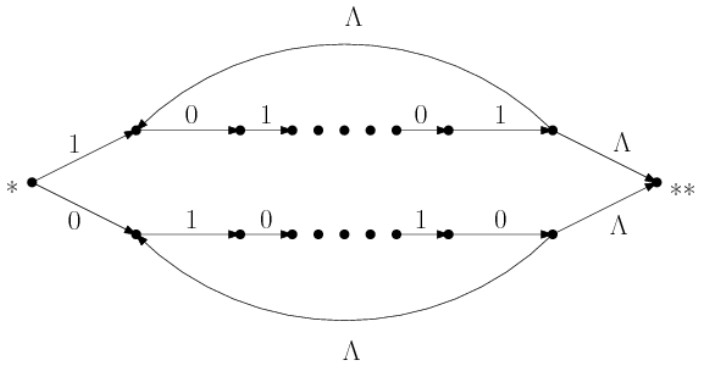

Рис. 3. Обобщенный источник для случая 3

Конец доказательства.

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

Эта матрица задает эйлеров граф $G_{n(\alpha)}$, как было установлено в одном из предыдущих примеров. При этом вид матрицы биграмм $n(\alpha)$ не совпадает ни с одним из M_1, M_2 и M_3. Значит, по Теореме 5 выполняется условие 1), и, соответственно, частотный язык $F_{\cup n(\alpha)}$ в данном случае нерегулярен.

Библиографические ссылки

1. *Оре О.* Теория графов. – 2-е изд. – М. : Наука, 1980. – 336 с.

2. *Кудрявцев В. Б., Алешин С. В., Подколзин А. С.* Введение в теорию автоматов. – М. : Наука, 1985. – 320 с.

***

A. A. Petyushko, Post-graduate, Lomonosov Moscow State University

Frequency languages

This paper deals with finite languages specified by a bigrams matrix along with introduced frequency languages on bigrams, which consist of all words with the bigrams matrix multiple of (i.e. multiplied by a positive integer) a given bigrams matrix. Connection between different characteristics of frequency languages and directed graphs and Eulerian circuits in it is established. Necessary and sufficient conditions of non-emptyness and countability of frequency languages are presented. The question of dependence of frequency language cardinality on a given bigrams matrix is analyzed. The formula for the words quantity for a specified bigrams matrix is established. Conditions of frequency languages to be regular are obtained.

Keywords: bigrams matrix, frequency languages, regularity of languages, Euler circuits

Получено: 10.05.12