Дискретная математика

том 25 выпуск 3 * 2013

УДК 519.713.32 DOI 10.4213/dm1248

О биграммных языках

© 2013 г. А. А. Петюшко

В статье рассматриваются формальные языки, заданные матрицей биграмм. Устанавливается связь различных характеристик этих языков с ориентированными графами и эйлеровыми циклами в них. Приводятся критерии непустоты, конечности и бесконечности языков. Устанавливаются условия регулярности этих языков.

1. Введение и необходимые определения

Еще в начале 20 века выдающимся русским ученым А. А. Марковым был создан аппарат цепей, впоследствии названных цепями Маркова, и опробован [1] на вычислении переходных вероятностей между соседними буквами в поэме А. С. Пушкина „Евгений Онегин“. В дальнейшем этот аппарат получил широкое применение для распознавания и статистического моделирования естественных языков [2]. Тем не менее, в детерминированном случае, за редким исключением прикладных задач (например, для подсчета ДНК-последовательностей [3]), биграммы для исследования формальных языков практически не применялись. В данной статье автор изучает языки, состоящие из слов с фиксированными частотами пар соседних букв.

Пусть A ($|A| < \infty$) — конечный алфавит, A^* — множество всех конечных слов (включая пустое) в данном алфавите.

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

Определение 2. Обозначим через $\theta_\beta(\alpha)$, где $\beta \in A^*, \alpha \in A^*$, причем β — непустое слово, отображение $A^* \to N \cup \{0\}$, сопоставляющее слову α число 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, меньшей длины слова β, значение $\theta_\beta(\alpha)$ положим равным 0. Само же значение $\theta_\beta(\alpha)$ 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 $\Theta(\alpha) = (\theta_{a_i a_j}(\alpha))_{i,j=1}^{|A|}$ размера $|A| \times |A|$ при условии, что все буквы алфавита $A = \{a_1, a_2, ..., a_{|A|}\}$ пронумерованы и нумерация зафиксирована.

Обозначим через Ξ множество квадратных матриц размера $|A| \times |A|$, каждый элемент которых является целым неотрицательным числом. Таким образом, для

каждого $\alpha \in A^*$ имеем $\Theta(\alpha) \in \Xi$. Также, здесь и далее через $\Theta(\alpha)$ будем обозначать матрицу биграмм, построенную по конкретному слову α, а через Θ — просто некоторую матрицу из Ξ, при этом будем считать, что на месте (i,j) матрицы Θ будет стоять значение $\theta_{a_i a_j}$ (т.е. для произвольной матрицы из Ξ мы опустили зависимость от α как для самой матрицы биграмм, так и для отдельных ее элементов).

Определение 3. Назовем языком $L(\Theta)$, порожденным матрицей $\Theta \in \Xi$, множество всех слов, имеющих одну и ту же матрицу биграмм Θ, т.е. $L(\Theta) = \{\beta | \Theta(\beta) = \Theta\}$.

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

2. Свойства матрицы $\Theta(\alpha)$

Пусть $a \in A$. Обозначим через $n_a^{in}(\alpha), \alpha \in A^*$, количество всех двухбуквенных подслов в α, заканчивающихся на a, т.е. $n_a^{in}(\alpha) = \sum\limits_{b \in A} \theta_{ba}(\alpha)$, что будет равно сумме значений матрицы $\Theta(\alpha)$ 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. $n_a^{out}(\alpha) = \sum\limits_{b \in A} \theta_{ab}(\alpha)$, что будет равно сумме значений матрицы $\Theta(\alpha)$ в строке, соответствующей символу a.

Пример 2. Рассмотрим то же слово, что и в предыдущем примере, а именно $A = \{0,1\}$, $\alpha = 01011100$. Тогда
$n_1^{in}(\alpha) = \theta_{01}(\alpha) + \theta_{11}(\alpha) = 2 + 2 = 4$,
$n_1^{out}(\alpha) = \theta_{10}(\alpha) + \theta_{11}(\alpha) = 2 + 2 = 4$,
$n_0^{in}(\alpha) = \theta_{00}(\alpha) + \theta_{10}(\alpha) = 1 + 2 = 3$,
$n_0^{out}(\alpha) = \theta_{00}(\alpha) + \theta_{01}(\alpha) = 1 + 2 = 3$.

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

Лемма 1 (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, где a_1 и a_2 — соответственно первая и последняя буквы этого слова (слово β может быть пустым). Тогда матрица $\Theta(\alpha)$ 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 $i \neq j$).

Доказательство. Пусть сначала $a_1 \neq a_2$. Рассмотрим три случая.

1) $b \neq a_1, b \neq a_2$. Очевидно, что при букве b, не совпадающей ни с начальной, ни с концевой буквой слова α, каждое такое вхождение b в α будет давать вклад 1 как в значение $n_b^{out}(\alpha)$, так и в $n_b^{in}(\alpha)$; в итоге получим $n_b^{out}(\alpha) = n_b^{in}(\alpha)$. При этом $\delta_{ba_1} = \delta_{ba_2} = 0$, и утверждение леммы выполнено ($0 = 0$).

2) $b = a_1$. Тогда каждое вхождение буквы b в α, будет давать вклад 1 либо одновременное в $n_b^{out}(\alpha)$ и $n_b^{in}(\alpha)$ (b стоит не на первом месте), либо только в $n_b^{out}(\alpha)$ (в случае нахождения b на первом месте). Значит, $n_b^{out}(\alpha) = n_b^{in}(\alpha) + 1$, при этом $\delta_{ba_1} = 1$, а $\delta_{ba_2} = 0$, и утверждение леммы выполнено ($1 = 1$).

3) $b = a_2$. В данном случае рассуждение аналогично п. 2).

Если же $a_1 = a_2$, то случай 1) рассматривается аналогично, а случаи 2) и 3) объединяются в один, где мы имеем $n_b^{out}(\alpha) = n_b^{in}(\alpha)$, $\delta_{ba_1} = \delta_{ba_2} = 1$, и утверждение леммы выполнено $(0 = 0)$.

Замечание 1. Доказательство этой несложной леммы, кроме того, можно найти в [7], см. также [8].

Замечание 2. Условие неразрывности (1) является необходимым, но не достаточным условием существования хотя бы одного слова α с заданной матрицей биграмм $\Theta \in \Xi$. Например, если $\theta_{00}(\alpha) = \theta_{11}(\alpha) = 1$, $\theta_{01}(\alpha) = \theta_{10}(\alpha) = 0$ в алфавите $A = \{0, 1\}$, то очевидно, что слова α с данным набором кратностей биграмм не существует: где-то в слове должны рядом стоять буквы 1 и 0, и должна быть ненулевая кратность $\theta_{01}(\alpha)$ или $\theta_{10}(\alpha)$, а у нас обе кратности нулевые.

Построим по матрице $\Theta(\alpha)$ (или по произвольной матрице $\Theta \in \Xi$) ориентированный граф $G_{\Theta(\alpha)}$ 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 $\theta_{ab}(\alpha)$ будет порождать $\theta_{ab}(\alpha)$ ориентированных ребер $a \to b$. Аналогично, кратность $\theta_{cc}(\alpha)$ будет порождать $\theta_{cc}(\alpha)$ петель $c \to c$.

Пример 3. $A = \{0, 1\}$, $\alpha = 01011100$.
$\Theta(\alpha) = \begin{pmatrix} \theta_{00}(\alpha) & \theta_{01}(\alpha) \\ \theta_{10}(\alpha) & \theta_{11}(\alpha) \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 2 & 2 \end{pmatrix}$.
Построим граф $G_{\Theta(\alpha)}$ по $\Theta(\alpha)$: см. Рис. 1.

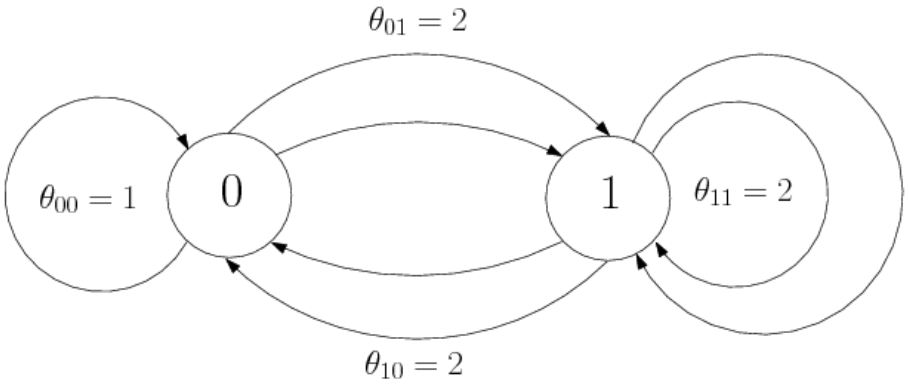

Рис. 1. Граф $G_{\Theta(\alpha)}$, построенный по набору $\Theta(\alpha)$

Напомним несколько широко известных понятий, касающихся эйлеровых путей.

Определение 4. Путем в ориентированном графе называется такая последовательность попарно различных ребер (набор параллельных ребер будем считать состоящим из различных ребер), что конец предыдущего ребра совпадает с началом следующего.

Определение 5. Циклом в ориентированном графе называется такой путь, что начало первого ребра в этом пути совпадает с концом последнего.

Определение 6. Эйлеровым путем в ориентированном графе называется такой путь, который содержит все ребра этого графа.

Определение 7. Эйлеровым циклом в ориентированном графе называется такой цикл, который содержит все ребра этого графа.

Определение 8. Полуэйлеров граф — граф, содержащий эйлеров путь, который не является эйлеровым циклом.

Определение 9. Эйлеров граф — граф, содержащий эйлеров цикл.

Замечание 3. На самом деле в каноническом определении полуэйлерова графа не говорится о том, что эйлеров путь не должен являться эйлеровым циклом. Но, следуя такому определению, несложно заметить, что любой эйлеров граф является также и полуэйлеровым, поэтому каждый раз для разграничения данных понятий пришлось бы дополнять фразой „полуэйлеров граф, не являющийся эйлеровым“.

В [4] доказаны следующие важные теоремы, позволяющие достаточно просто проверять ориентированные графы на наличие эйлеровых путей и циклов:

Теорема 1. *Ориентированный граф является эйлеровым тогда и только тогда, когда выполняются следующие условия:*
1) *все вершины, инцидентные ребрам, лежат в одной компоненте связности соответствующего неориентированного графа;*
2) *у всех вершин количество входящих ребер равно количеству исходящих ребер.*

Теорема 2. *Ориентированный граф является полуэйлеровым тогда и только тогда, когда выполняются следующие условия:*
1) *все вершины, инцидентные ребрам, 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.*

В дальнейшем, там, где будут упоминаться понятия эйлеровых и полуэйлеровых графов, будем иметь в виду, что установить факт, является граф эйлеровым или полуэйлеровым, можно по двум вышеприведенным теоремам.

Замечание 4. Несложно показать, что условие неразрывности (Лемма 1) как раз и является условием 2) в вышеприведенных теоремах.

Лемма 2 (Достаточное условие существования)**.** *Для того, чтобы существовало хотя бы одно слово α с данной матрицей кратностей биграмм $\Theta \in \Xi$, достаточно, чтобы построенный по Θ ориентированный граф G_Θ был либо эйлеровым, либо полуэйлеровым.*

Доказательство. По определению, в эйлером и полуэйлеровом графах существует эйлеров путь, т.е. путь, проходящий по всем ребрам орграфа, причем ровно по одному разу. Пусть такой эйлеров путь задается последовательностью ребер $a_1 \to a_2, a_2 \to a_3, ..., a_{n-1} \to a_n$. Тогда слово $\alpha = a_1 a_2 ... a_{n-1} a_n$ и будет искомым, поскольку в его построении будут участвовать все ребра из G_Θ (а значит, и все ненулевые кратности биграмм из G_Θ), причем с учетом правильных кратностей.

Пример 4. $A = \{0, 1\}$. Пусть $\Theta = \begin{pmatrix} \theta_{00} & \theta_{01} \\ \theta_{10} & \theta_{11} \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 2 & 2 \end{pmatrix}$.
Построим граф G_Θ по Θ — 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 Θ есть слово с такой матрицей кратностей биграмм.

Как итог, получаем следующее важное следствие:

Следствие 1 (Алгоритмическая разрешимость)**.** *Задача определения того, существует ли хотя бы одно слово α с заданной матрицей биграмм Θ, алгоритмически разрешима.*

Доказательство. Для доказательства данного утверждения можно воспользоваться результатом Леммы 2: построить по набору значений Θ ориентированный граф G_Θ и проверить, существует ли в нем эйлеров путь. Очевидно, эта задача алгоритмически разрешима.

3. 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, т.е. случай языка, в котором отношения $\theta_{ab}(\alpha)/\theta_{cd}(\alpha)$ зависят только от букв $a, b, c, d \in A$, $\theta_{cd}(\alpha) > 0$, но не зависят от слова α из этого языка. Определим такой язык.

Определение 10. Назовем частотным языком на биграммах с кратностями, заданным матрицей биграмм $\Theta \in \Xi$, язык

$$F_\Theta = \bigcup_{k=1}^{\infty} L(k\Theta),$$

т.е. язык, состоящий из всех таких слов β, что набор кратностей этих слов $\Theta(\beta)$ кратен набору Θ, а именно, $F_\Theta = \{\beta | \Theta(\beta) = k\Theta, k \in N\}$, где умножение k на Θ понимается как умножение скаляра на матрицу.

Рассмотрим, какие дополнительные ограничения накладывает условие „частотности“ на матрицу $\Theta(\alpha)$.

Лемма 3 (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 ($\alpha'$ может быть пустым). Для того, чтобы в частотном языке $F_{\Theta(\alpha)}$ 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), с матрицей биграмм $\Theta(\beta) = k\Theta(\alpha), k \in N, k > 1$, 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 $b_1 = b_2$).

Доказательство. Из утверждения Леммы 1 получаем, что $n_b^{out}(\beta) - n_b^{in}(\beta) = \delta_{bb_1} - \delta_{bb_2}$ для каждого $b \in A$. Если существует такое слово β, что $\Theta(\beta) = k\Theta(\alpha), k \in N, k > 1$, то, очевидно, $n_b^{out}(\beta) = kn_b^{out}(\alpha)$ и $n_b^{in}(\beta) = kn_b^{in}(\alpha)$ для каждого $b \in A$. Значит, $k\left(n_b^{out}(\alpha) - n_b^{in}(\alpha)\right) = \delta_{bb_1} - \delta_{bb_2}$ при $k > 1$. Поскольку $\delta_{bb_1} - \delta_{bb_2} \in \{-1, 0, 1\}$, то последнее равенство выполняется только в случае $n_b^{out}(\alpha) - n_b^{in}(\alpha) = 0$ (и, следовательно, $kn_b^{out}(\alpha) = n_b^{out}(\beta) = n_b^{in}(\beta) = kn_b^{in}(\alpha)$), для каждого $b \in A$ и $b_1 = b_2$. Также из условия $0 = n_b^{out}(\alpha) - n_b^{in}(\alpha) = \delta_{ba_1} - \delta_{ba_2}$ для каждого $b \in A$ следует, что $a_1 = a_2$.

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

Доказательство. По Лемме 3 слово α представимо в виде $\alpha = a\alpha' a, a \in A, \alpha' \in A^*$ (α' может быть пустым). Тогда искомое слово $\beta_k = a\underbrace{\alpha' a\alpha' a \ldots \alpha' a}_{k-1 \quad \alpha' a}$, где подслово $\alpha' a$ приписано $k-1$ раз справа от изначального слова α. Пусть второй буквой слова α является b (возможно, она одновременно является и последней буквой при пустом α'). Тогда $\Theta(\alpha) = \Theta(ab) + \Theta(\alpha' a)$ („+“ понимается как сложение матриц), $\Theta(\beta_k) = \Theta(ab) + \Theta(\alpha' a) + \ldots + \Theta(ab) + \Theta(\alpha' a) = k\left(\Theta(ab) + \Theta(\alpha' a)\right) = k\Theta(\alpha) = k\Theta$.

Как итог, приведем достаточный признак существования такого β_k, что для любого натурального k выполняется $\Theta(\beta_k) = k\Theta$.

Лемма 4 (Достаточное условие существования частотных языков)**.** *Для того, чтобы для каждого $k \in N$ существовало такое слово β_k, что для заданного набора кратностей биграмм Θ выполнялось $\Theta(\beta_k) = k\Theta$, достаточно, чтобы построенный по Θ ориентированный граф G_Θ являлся эйлеровым графом.*

Доказательство. Возьмем в качестве слова α, имеющего матрицу биграмм Θ, некоторое слово β_1 из условия данной леммы.

Поскольку, согласно Лемме 3, первая и последняя буквы α должны совпадать, и при обходе нам нужно пройти через каждое ребро ровно один раз, то для существования α с заданным набором биграмм Θ достаточно, чтобы в графе G_Θ существовал эйлеров цикл (ср. с Леммой 2). А поскольку ориентированный граф G_Θ эйлеров, это гарантирует наличие в нем эйлерова цикла. При этом, согласно Следствию 2, для получения β_k при $k \in N$ достаточно пройти по этому циклу k раз.

Теорема 3 (О числе слов в частотных языках)**.** *Пусть задан набор биграмм $\Theta \in \Xi$. Тогда:*
1) *если ориентированный граф G_Θ является эйлеровым, то в частотном языке F_Θ счетное множество слов;*
2) *если ориентированный граф G_Θ является полуэйлеровым, то частотный язык F_Θ совпадает с $L(\Theta)$ и в нем конечное ненулевое множество слов, имеющих одинаковую длину;*
3) *если ориентированный граф G_Θ не является ни эйлеровым, ни полуэйлеровым, то в частотном языке F_Θ нет ни одного слова.*

Доказательство. 1) Воспользуемся Леммой 4. Для каждого $k \in N$ будет существовать хотя бы одно слово β_k с $\Theta(\beta_k) = k\Theta$, и, следовательно, лежащее в F_Θ. Так

как для каждого k количество таких β_k не более чем конечно, а объединение счетного множества конечных множеств счетно, то отсюда следует первое утверждение теоремы.

2) По Лемме 2 существует хотя бы одно слово α с набором кратностей $\Theta(\alpha) = \Theta$ (т.е. язык $L(\Theta)$ непуст). При этом, если существует более одного такого слова, то все они будут имеют одинаковую длину l_Θ, которую легко вычислить, поскольку длина любого непустого слова на 1 больше суммы значений кратностей всех биграмм:

$$l_\Theta = 1 + \sum_{a,b \in A} \theta_{ab}.$$

Очевидно, что таких слов конечной длины также конечное число. С другой стороны, по Лемме 3 для того, чтобы при каждом целом $k > 1$ существовало такое β_k, что $\Theta(\beta_k) = k\Theta$, необходимо, чтобы $n_b^{out}(\alpha) = n_b^{in}(\alpha)$ для всех $b \in A$, что очевидным образом влечет за собой равенство количества входящих и исходящих ребер у всех вершин орграфа G_Θ, а это противоречит условию теоремы. Значит, включение $L(\Theta) \subseteq F_\Theta$ в случае полуэйлерова графа G_Θ влечет за собой равенство непустых языков $L(\Theta) = F_\Theta$; при этом в частотном языке F_Θ конечное ненулевое число слов, имеющих длину l_Θ.

3) В этом случае по Лемме 2 нет ни одного слова α с матрицей биграмм Θ. Согласно Теоремам 1 и 2 граф G_Θ либо не связен, либо в нем есть вершина с разным числом входящих и исходящих ребер. Тогда для любого целого $k > 1$ граф $G_{k\Theta}$ либо не связен, либо имеет вершину, у которой разность чисел входящих и исходящих ребер по модулю не меньше k. Значит, граф $G_{k\Theta}$ в любом случае при $k > 1$ не является ни эйлеровым, ни полуэйлеровым, и по Лемме 2 не существует ни одного слова β_k с матрицей биграмм $k\Theta$.

Пример 5. $A = \{0, 1\}$. Пусть $\Theta = \begin{pmatrix} \theta_{00} & \theta_{01} \\ \theta_{10} & \theta_{11} \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 2 & 2 \end{pmatrix}$.
Построим граф G_Θ по Θ - 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_Θ эйлеров, т.е. в Теореме 3 выполняется условие 1), и, соответственно, в частотном языке F_Θ счетное множество слов.

Следствие 3 (Алгоритмическая разрешимость для частотных языков)**.** *Задача определения того, пуст, конечен или счетен частотный язык F_Θ с заданной матрицей частот биграмм $\Theta \in \Xi$, алгоритмически разрешима.*

Доказательство. Используем предложенный в Следствии 1 вариант доказательства, основанный на рассмотрении эйлеровых путей в графе.

Достаточно построить граф G_Θ и выяснить, существует ли в нем эйлеров цикл, тогда язык счетен, а в отсутствие такового — существует ли эйлеров путь, не являющийся циклом, тогда язык конечен. В остальных случаях язык F_Θ пуст.

4. Регулярность счетных частотных языков на биграммах с кратностями

В данном разделе рассмотрим вопрос о регулярности бесконечных языков F_Θ с заданным набором Θ. Рассмотрим сначала общий случай с произвольным числом букв

в алфавите A.

Определение 11. Назовем две ненулевые матрицы Θ_1 и Θ_2 из Ξ неколлинеарными, если не существует ненулевых действительных коэффициентов $c_1, c_2 \in R, (c_1, c_2) \neq (0,0)$, для которых $c_1\Theta_1 + c_2\Theta_2 = 0$.

Определение 12. Назовем элементарной матрицей, соответствующей ненулевой матрице биграмм Θ в алфавите $A, |A| < \infty$, матрицу $\widehat{\Theta} = \Theta/\text{НОД}(\{\theta_{ab}|\theta_{ab} > 0, a, b \in A\})$, где под НОД$(M)$ имеется в виду наибольший общий делитель натуральных чисел из множества M.

Теорема 4. *Пусть $A, |A| < \infty$ — некоторый конечный алфавит. Далее, пусть задана такая матрица биграмм Θ, что соответствующий ей ориентированный граф G_Θ является эйлеровым. Тогда:*
1) *если существует такое разложение Θ в сумму двух ненулевых неколлинеарных матриц $\Theta = \Theta_1 + \Theta_2$, что обе матрицы Θ_1 и Θ_2 задают ориентированные эйлеровы графы G_{Θ_1} и G_{Θ_2}, то язык F_Θ нерегулярен;*
2) *в противном случае язык F_Θ регулярен. При этом, если l — число ненулевых элементов в матрице Θ, то для каждого $k \in N$ существуют ровно l таких слов $\beta_{k,i}, i = 1, \ldots, l$, что $\Theta(\beta_{k,i}) = k\Theta$.*

Доказательство. 1) Пусть существует разложение Θ в сумму двух ненулевых неколлинеарных матриц $\Theta = \Theta_1 + \Theta_2$ и обе матрицы Θ_1 и Θ_2 задают ориентированные эйлеровы графы G_{Θ_1} и G_{Θ_2}. На языке графов это значит, что изначальный эйлеров цикл графа G_Θ распадается в сумму двух различных циклов, соответствующих графам G_{Θ_1} и G_{Θ_2}. При этом, поскольку граф G_Θ связен (с точностью до изолированных вершин), то графы G_{Θ_1} и G_{Θ_2} имеют хотя бы одну общую вершину. Пусть этой вершиной будет $a \in A$.

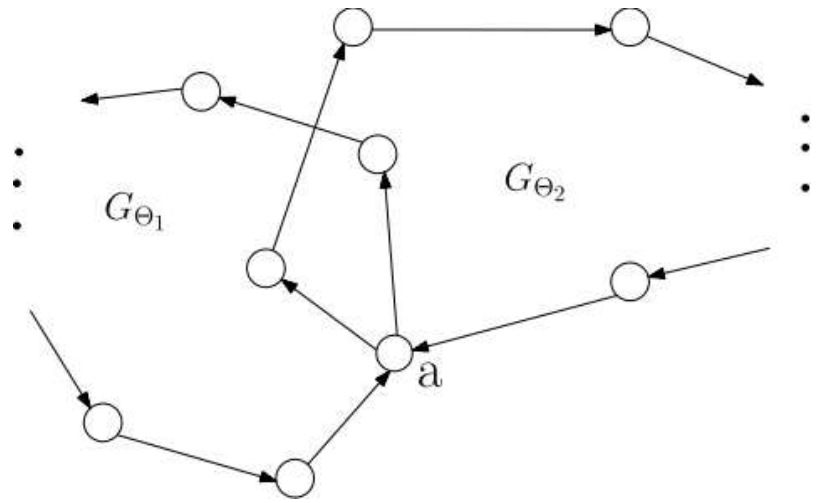

Рис. 2.

Пусть в графе G_{Θ_1} эйлеров цикл с началом (и соответственно концом) в общей точке a задается словом $\alpha_1 = a\alpha_1'$, $\alpha_1' \in A^*$, при этом, очевидно, $\Theta(\alpha_1) = \Theta_1$. Аналогично, пусть в графе G_{Θ_2} эйлеров цикл с началом (и соответственно концом) в

общей точке a задается словом $\alpha_2 = a\alpha_2'$, $\alpha_2' \in A^*$, при этом $\Theta(\alpha_2) = \Theta_2$. Тогда слово $\alpha' = a\alpha_1'\alpha_2'$ будет задавать эйлеров цикл в графе G_Θ, т.е. $\Theta(\alpha') = \Theta$. Отметим, что, поскольку матрицы Θ_1 и Θ_2 ненулевые, то и α_1', и α_2' — непустые слова.

Допустим противное утверждению п. 1) данной теоремы: язык F_Θ регулярен, тогда по теореме Клини [5] он представим в некотором конечно-детерминированном инициальном автомате $V_{q_0} = (A, Q, B, \varphi, \psi, q_0)$, где A — входной алфавит, Q — алфавит состояний, B — выходной алфавит (будем считать, что $B = \{0, 1\}$), $\varphi : Q \times A \to Q$ — функция переходов, $\psi : Q \times A \to B$ — функция выходов, $q_0 \in Q$ — начальное состояние. Полезно расширить функцию переходов на входные слова (а именно, задать $\varphi : Q \times A^* \to Q$) следующим образом [6]: полагаем по определению $\varphi(q, \Lambda) = q, \varphi(q, \alpha a) = \varphi(\varphi(q, \alpha), a)$, где Λ — пустое слово, для любых $q \in Q, \alpha \in A^*, a \in A$. Тогда согласно определению представимости в инициальном автомате V_{q_0} имеем $\beta = \widetilde{\beta}b \in F_\Theta \Leftrightarrow \psi(\varphi(q_0, \widetilde{\beta}), b) = 1$, где $\widetilde{\beta} \in A^*, b \in A$.

Пусть мощность алфавита состояний $|Q| = p$. По Лемме 4 для любого $k \in N$ найдется слово β_k с $\Theta(\beta_k) = k\Theta$. Зафиксируем некоторое $s > p$ и возьмем слово $\beta = a\underbrace{\alpha_1'...\alpha_1'}_{s}\underbrace{\alpha_2'...\alpha_2'}_{s}$ с $\Theta(\beta) = s\Theta$. Это значит, что нужно сначала пройти s раз по эйлерову циклу графа G_{Θ_1} с началом в общей вершине a, после чего s раз по эйлерову циклу графа G_{Θ_2} с началом все в той же общей вершине a.

Обозначим через $q = \varphi(q_0, a)$ состояние, в которое мы попадем при подаче на вход инициального автомата V_{q_0} первой буквы a слова β. Запишем в ряд состояния, в которые мы будем переходить при последовательной подаче букв слова α_1' (как подслова слова β): $q_1 = \varphi(q, \alpha_1'), q_2 = \varphi(q, \alpha_1'\alpha_1') = \varphi(q_1, \alpha_1'), ..., q_s = \varphi(q, \underbrace{\alpha_1'...\alpha_1'}_{s}) = \varphi(q_{s-1}, \alpha_1') = \varphi(q_0, a\underbrace{\alpha_1'...\alpha_1'}_{s})$. Поскольку $s > p$, в этом ряду длины s будут по меньшей мере два повторяющихся состояния, т.е. $q_i = q_j$ при некоторых $i, j \in N, 1 \leqslant i < j \leqslant s$. Значит, для каждого $m \in N$ верно тождество $q_j = \varphi(q, \underbrace{\alpha_1'..\alpha_1'}_{i}\underbrace{\alpha_1'..\alpha_1'}_{m(j-i)})$, так как, подавая одно и то же слово α_1', мы будем ходить по циклу по одним и тем же состояниям $q_i, q_{i+1}, ..., q_j = q_i$.

Обозначим через $\beta_m', m \in N$, слово $\beta_m' = a\underbrace{\alpha_1'..\alpha_1'}_{s+(m-1)*(j-i)}\underbrace{\alpha_2'..\alpha_2'}_{s}$. Тогда $\varphi(q, \underbrace{\alpha_1'..\alpha_1'}_{j}\underbrace{\alpha_1'..\alpha_1'}_{s-j}) = \varphi(q, \underbrace{\alpha_1'..\alpha_1'}_{i}\underbrace{\alpha_1'..\alpha_1'}_{m(j-i)}\underbrace{\alpha_1'..\alpha_1'}_{s-j})$ для каждого $m \in N$, и, значит,

$$\varphi(q_0, \beta) = \varphi(\varphi(q_0, a\underbrace{\alpha_1'..\alpha_1'}_{s}), \underbrace{\alpha_2'..\alpha_2'}_{s}) = \varphi(\varphi(q_0, a\underbrace{\alpha_1'..\alpha_1'}_{s+(m-1)(j-i)}), \underbrace{\alpha_2'..\alpha_2'}_{s}) = \varphi(q_0, \beta_m').$$

Следовательно, и $\varphi(q_0, \widetilde{\beta}) = \varphi(q_0, \widetilde{\beta_m'})$, где под $\widetilde{\delta}$ понимается слово δ без последней буквы, так как, начиная с первого вхождения непустого подслова α_2' в слова β и β_m', мы начнем двигаться с одного и того же состояния состояния q_s, при этом на вход автомату V_{q_0} будут подаваться одни и те же буквы.

Значит, $\psi(\varphi(q_0, \widetilde{\beta_m'}), a) = \psi(\varphi(q_0, \widetilde{\beta}), a) = \psi(q_0, \beta) = 1$ и, следовательно, $\beta_m' \in F_{\Theta(\alpha)}$ (здесь мы использовали то, что последней буквой слов $\beta_m', \beta, \alpha_1'$ и α_2' является a, так как в эйлеровом цикле последняя вершина совпадает с начальной, которая в нашем случае является общей вершиной графов G_{Θ_1} и G_{Θ_2}).

При ненулевых и неколлинеарных матрицах $\Theta(\alpha_1)$ и $\Theta(\alpha_2)$, $\Theta(\alpha_1) + \Theta(\alpha_2) = \Theta_1 + \Theta_2 = \Theta$, имеем: существуют такие $a_1, a_2, a_3, a_4 \in A, (a_1, a_2) \neq (a_3, a_4)$, что $\theta_{a_1a_2}(\alpha_1) > 0$ и $\theta_{a_3a_4}(\alpha_2) > 0$ (что, в свою очередь, означает $\theta_{a_1a_2}(\alpha) = \theta_{a_1a_2}(\alpha_1) + \theta_{a_1a_2}(\alpha_2) > 0$ и $\theta_{a_3a_4}(\alpha) = \theta_{a_3a_4}(\alpha_1) + \theta_{a_3a_4}(\alpha_2) > 0$), а также не существует двух таких коэффициентов $c_1, c_2 \in R, (c_1, c_2) \neq (0, 0)$, что

$c_1\theta_{a_1a_2}(\alpha_1) + c_2\theta_{a_1a_2}(\alpha_2) = 0,$
$c_1\theta_{a_3a_4}(\alpha_1) + c_2\theta_{a_3a_4}(\alpha_2) = 0,$

т.е. определитель

$$\begin{vmatrix} \theta_{a_1a_2}(\alpha_1) & \theta_{a_1a_2}(\alpha_2) \\ \theta_{a_3a_4}(\alpha_1) & \theta_{a_3a_4}(\alpha_2) \end{vmatrix} \neq 0 \tag{3}$$

Так как для любого слова $\gamma \in F_\Theta$ существует такое $k \in N$, что $\Theta(\gamma) = k\Theta$, то равенство $\frac{\theta_{a_1a_2}(\gamma)}{\theta_{a_3a_4}(\gamma)} = \frac{k\theta_{a_1a_2}}{k\theta_{a_3a_4}} = \frac{\theta_{a_1a_2}}{\theta_{a_3a_4}} = c = const > 0$ выполняется для всех $\gamma \in F_\Theta$. Рассчитаем отношение для $\beta'_m \in F_{\Theta(\alpha)}$:

$$\frac{\theta_{a_1a_2}(\beta'_m)}{\theta_{a_3a_4}(\beta'_m)} = \frac{(s + (m-1)(j-i))\theta_{a_1a_2}(\alpha_1) + s\theta_{a_1a_2}(\alpha_2)}{(s + (m-1)(j-i))\theta_{a_3a_4}(\alpha_1) + s\theta_{a_3a_4}(\alpha_2)}.$$

Получается, что это отношение имеет вид отношения двух линейных функций $\frac{ux+v}{zx+t}$ от переменной $x = s + (m-1)(j-i)$. Очевидно, что это отношение будет константным, если постоянные коэффициенты в числителе u, v будут прямо пропорциональны коэффициентам z, t в знаменателе. Значит, существует такое $d \in R, d > 0$, что

$\theta_{a_1a_2}(\alpha_1) = d\theta_{a_3a_4}(\alpha_1),$
$\theta_{a_1a_2}(\alpha_2) = d\theta_{a_3a_4}(\alpha_2).$

Подставляя эти выражения для расчета определителя (3), получим противоречие (две пропорциональные строки в детерминанте, значит, он нулевой). Значит, $\beta'_m \notin F_\Theta$, противоречие с предположением о существовании автомата, представляющего множество F_Θ и, таким образом, регулярности F_Θ.

2) Пусть не существует такого разложения Θ в сумму двух ненулевых неколлинеарных матриц $\Theta = \Theta_1 + \Theta_2$, что обе матрицы Θ_1 и Θ_2 задают ориентированные эйлеровы графы G_{Θ_1} и G_{Θ_2}. Докажем, что в этом случае граф G_Θ будет представлять собой либо множественную петлю (см. Рис. 3 a)), либо набор „параллельных“ простых (т.е. без самопересечений) циклов (см. на Рис. 3 b)).

Предположим, что граф G_Θ имеет отличный от изображенных на Рис. 3 a) и Рис. 3 b) вид. Значит, это либо простой цикл с петлями (см. Рис. 4 a)), либо самопересекающийся цикл (см. на Рис. 4 b)), либо комбинация этих двух случаев — самопересекающийся цикл с петлями.

В первом случае он разлагается на сумму простого цикла и петли (и поэтому имеем противоречие с условием 2) теоремы), во втором — на сумму двух циклов с общей вершиной в точке пересечения (и опять имеем противоречие о невозможности разложения в два различных цикла). Значит, предположение неверно. Более того, по условию неразрывности для частотных языков (2), количество входящих в любую вершину ребер равно количеству исходящих ребер, что дает для случая простого цикла на Рис. 3 b) одинаковое количество ребер между любыми двумя соединенными ребрами вершинами. Следовательно, граф G_Θ представляет собой один из двух видов, представленных на Рис. 3.

Заметим, что при умножении матрицы Θ на любое $k \in N$ вид графа $G_{k\Theta}$ останется тем же, что и был для G_Θ, поскольку ребер, которые соединяют ранее не связанные вершины, при такой операции не появится, и при этом количество входящих

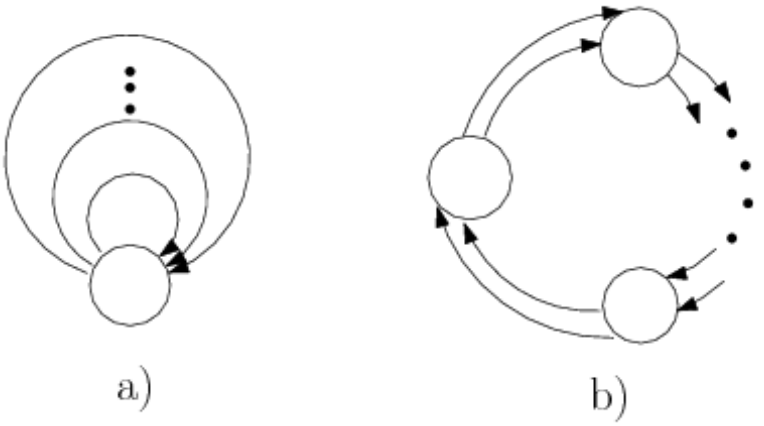

Рис. 3.

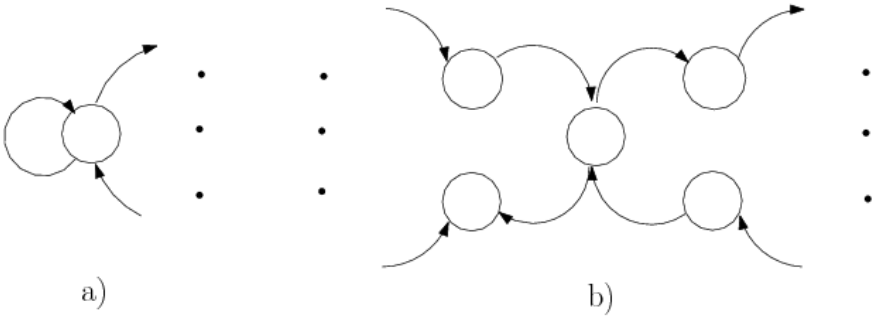

Рис. 4.

в любую вершину ребер останется равным количеству исходящих ребер, поскольку при этом увеличиваются в k раз изначально равные числа входящих и исходящих ребер для любой вершины графа G_Θ.

Значит, в матрице Θ все ненулевые элементы равны между собой, а любой эйлеров цикл в $G_{k\Theta}$ для любого $k \in N$ представляет собой m раз повторенный эйлеров цикл для элементарной матрицы $\widehat{\Theta}$ (т.е. в данном случае матрицы, в которой на всех ненулевых местах единицы), где число повторений $m = k * \text{НОД}(\{\theta_{ab} | \theta_{ab} > 0, a, b \in A\})$ (см. пример на Рис. 5).

При этом длина цикла для элементарной матрицы $\widehat{\Theta}$ (под длиной цикла будем понимать количество ребер в нем) будет равна в точности числу ненулевых элементов в данной матрице (и, следовательно, в матрице Θ), поскольку разные элементы матрицы соответствуют разным ребрам цикла, и наоборот.

Очевидно, что для любого натурального k количество таких различных слов β_k, что $\Theta(\beta_k) = k\Theta$, будет совпадать с количеством способов выбрать первую букву (и, следовательно, начальное ориентированное ребро в цикле) в соответствующем элементарной матрице $\widehat{\Theta}$ цикле, так как остальные буквы уже будут однозначно

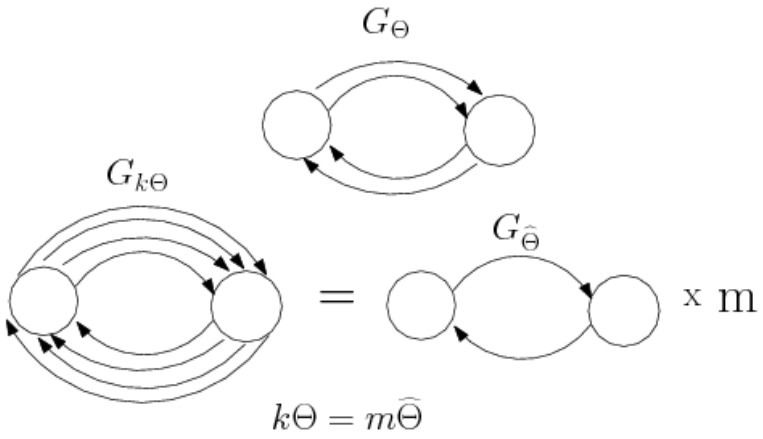

Рис. 5. Простой эйлеров цикл для k = 2, m = 4, НОД($\{\theta_{ab}|\theta_{ab} > 0, a, b \in A\}$) = 2

определены самим эйлеровым циклом (при этом ни k, ни НОД($\{\theta_{ab}|\theta_{ab} > 0, a, b \in A\}$) в этом выборе не играют никакой роли). Следовательно, для любого $k \in N$ существуют ровно l таких слов $\beta_{k,i}, i = 1, \ldots, l$, что $\Theta(\beta_{k,i}) = k\Theta$, а l — число ненулевых элементов в наборе Θ.

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 Θ, что соответствующий ориентированный граф G_Θ является эйлеровым. Тогда:*
1) *если матрица биграмм Θ 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 $c_i \in N, i = 1, \ldots, 3$, то язык F_Θ нерегулярен;*
2) *если матрица биграмм Θ имеет вид, совпадающий с M_1 или M_2, то язык F_Θ регулярен. При этом для каждого $k \in N$ существует единственное β_k с $\Theta(\beta_k) = k\Theta$;*
3) *Если матрица биграмм Θ имеет вид, совпадающий с M_3, то язык F_Θ регулярен. При этом для каждого $k \in N$ существуют ровно два слова β_k и γ_k с матрицами кратностей биграмм $\Theta(\beta_k) = \Theta(\gamma_k) = k\Theta$.*

Доказательство. 1) Пусть матрица биграмм Θ имеет вид, не совпадающий ни с одним из M_1, M_2 и M_3, и при этом соответствующий ориентированный граф G_Θ является эйлеровым. С учетом Леммы 3 для двухбуквенного алфавита получаем $\theta_{01} = \theta_{10} = c > 0$. Значит, матрица биграмм Θ имеет вид $\Theta = \begin{pmatrix} d_1 & c \\ c & d_2 \end{pmatrix}$, где хотя бы одно из чисел 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 Θ в виде суммы двух неколлинеарных матриц $\Theta = \begin{pmatrix} d_1 & c \\ c & d_2 \end{pmatrix} = \begin{pmatrix} d_1 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & c \\ c & d_2 \end{pmatrix} = C_1 + C_2$.

Заметим, что матрица C_1 задает ориентированный граф, который представляет собой множественную петлю у вершины „0“, и, следовательно, является эйлеровым (d_1 проход по петле $0 \to 0$ и даст необходимый эйлеров цикл). Матрица C_2 задает ориентированный граф, который представляет собой простой цикл между вершинами „0“ и „1“, возможно, со множественной петлей у вершины „1“ (при $d_2 > 0$). Значит, можно построить эйлеров цикл следующим образом: сначала d_2 прохода по петле $1 \to 1$ при ее наличии, а затем c проходов по циклу $1 \to 0 \to 1$.

Получается, что исходная матрица биграмм раскладывается в сумму таких двух неколлинеарных матриц, что соответствующие им ориентированные графы — эйлеровы. Значит, верно условие 1) Теоремы 4, и, следовательно, язык F_Θ нерегулярен.

2) Пусть для определенности $\theta_{00} = c_1 > 0$, и при этом для любых $u, v \in \{0,1\}, (u,v) \neq (0,0)$, выполняется $\theta_{uv} = 0$, т.е. матрица биграмм имеет вид M_1 (случай $\theta_{11} = c_2 > 0$ и вид матрицы биграмм M_2 рассматривается аналогично). Тогда язык F_Θ должен состоять только из таких слов β_k, $\Theta(\beta_k) = k\Theta$, что слово β_k имеет вид $\beta_k = \underbrace{0...0}_{1+k*\theta_{00}}$ для любого $k \in N$. Очевидно, что все такие слова задаются одним регулярным выражением (где под „$\cdot$“ понимается конкатенация, а под „<M>“ - итерация, т.е. конкатенация одного или более элемента из M):

$$F_\Theta = 0\cdot < \underbrace{0 \cdot ... \cdot 0}_{\theta_{00}} >$$

и, значит, язык F_Θ регулярен.

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

$$F_\Theta = 1\cdot < \underbrace{0 \cdot 1 \cdot ... \cdot 0 \cdot 1}_{\theta_{01}} > \bigcup 0\cdot < \underbrace{1 \cdot 0 \cdot ... \cdot 1 \cdot 0}_{\theta_{10}} >$$

и, значит, язык F_Θ регулярен.

Пример 6. $A = \{0,1\}$. Пусть $\Theta = \begin{pmatrix} \theta_{00} & \theta_{01} \\ \theta_{10} & \theta_{11} \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 2 & 2 \end{pmatrix}$.
Эта матрица задает эйлеров граф G_Θ, как было установлено в одном из предыдущих примеров. При этом вид матрицы биграмм Θ 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_Θ в данном случае нерегулярен.

Работа выполнена на кафедре МаТИС МГУ им. М. В. Ломоносова под руководством проф. Д. Н. Бабина.

Список литературы

1. Марков А. А., Пример статистического исследования над текстом „Евгения Онегина“, иллюстрирующий связь испытаний в цепь. *Известия Императорской Академии наук. Серия 6* (1913) **7**, №3, 153–162.

2. Essen U., Steinbiss V., Cooccurrence smoothing for stochastic language modeling. *IEEE International Conference on Acoustics, Speech, and Signal Processing* (1992) **1**, 161–164.

3. Hutchinson J. P., Wilf H. S., On eulerian circuits and words with prescribed adjacency patterns. *J. Combinatorial Theory. Ser. A* (1975) **18**, 80–87.

4. Оре О., *Теория графов*. Наука, Москва, 1980.

5. Kleene S. C., Representation of events in nerve nets and finite automata. В сб.: *Automata Studies* (Shannon C., McCarthy J., ред.). Princeton University Press, Princeton, 1956, с. 3–41.

6. Кудрявцев В. Б., Алешин С. В., Подколзин А. С., *Введение в теорию автоматов*. Наука, Москва, 1985.

7. Смирнов Н. В., Сарманов О. В., Захаров В. К., Локальная предельная теорема для чисел переходов в цепи Маркова и ее применения. *ДАН СССР* (1966) **167**, №6, 1238–1241 (см. также Смирнов Н. В., Теория вероятностей и математическая статистика. *Избранные труды* (1970). Наука. Москва. 260–264).

Статья поступила 09.11.2012.