УДК 519.7 DOI: 10.18698/1812-3368-2017-4-4-15

О ЧИСЛОВЫХ ХАРАКТЕРИСТИКАХ ФОРМАЛЬНЫХ ЯЗЫКОВ

Р.С. Исмагилов[1] ismagil@bmstu.ru
А.А. Мастихина[1] anmast@bmstu.ru
Л.Е. Филиппова[2]

[1] МГТУ им. Н.Э. Баумана, Москва, Российская Федерация
[2] Московский институт электроники и математики им. А.Н. Тихонова, Национальный исследовательский университет «Высшая школа экономики», Москва, Российская Федерация

Аннотация

Рассмотрена задача подсчета числа слов регулярного языка заданного состава. Задан алфавит из символов. Состав слова определен как вектор. Введены функции числа слов заданного состава. Определены простые оценки указанных функций для языков, полученных из языков с помощью обычных операций (объединение, конкатенация, итерация). Изучены ряды для языков, порожденных автоматами

Ключевые слова

Формальный язык, регулярный язык, граф, производящая функция, порядок, состав слов

Поступила в редакцию 29.09.2016

Введение. Задача подсчета числа слов с некоторой характеристикой широко исследована в дискретной математике и теории вероятности. Рассмотрены, в частности, характеристика наличия в последовательности серии символов (повторения символа некоторое число раз), а также число таких серий. Этой проблеме посвящены, например, работы [1–3]. Следует отметить, что в указанных работах для подсчета числа слов использованы производящие функции.

Отдельный интерес представляет подсчет числа слов в регулярном языке, т. е. в языке, порожденном некоторым конечным автоматом. Способ нахождения числа слов заданной длины по матрице автомата изложен в работе [4]. В частности, описание функций, которые могут служить функциями роста числа слов данной длины, дано в работе [5]; показано, что регулярный язык может иметь в качестве функции числа слов длиной n любой неотрицательный целочисленный полином.

В предложенной работе исследована функция числа слов заданного состава. Таким образом, функция числа слов заданной длины может быть получена из функции числа слов данного состава.

С одной стороны, исследование опирается на построение регулярного языка из простейших языков с помощью операций. Получены оценки того, как меняется функция числа слов заданного состава при применении операции над языками. С другой стороны, использовано графовое представление регулярного языка. Введено понятие эквивалентности для исследуемых функций числа слов, показано, как соотносятся классы эквивалентности сильно связных компонент графа.

Постановка задачи. Краткое описание результатов. Будем рассматривать языки в алфавите $A=\{\xi_1,...,\xi_N\}$; пустое слово обозначим через Λ. Язык L задается либо явным списком (упорядоченным каким-либо образом) входящих в него слов, либо регулярным выражением, либо ориентированным графом (автоматом). (Необходимые сведения о формальных языках можно почерпнуть из работ [6–8]). Для любого слова $\alpha\in L$ рассмотрим две простейшие характеристики: 1) длину слова $|\alpha|$; 2) состав слова — так назовем вектор $(m_1,...,m_N)$, в котором координата m_i — число вхождений буквы ξ_i в слово α. Соответственно, введем две числовые функции: функцию $\varphi_L(n)$ — число слов длиной n и функцию $\psi_L(m_1,...,m_N)$ — число слов состава $(m_1,...,m_N)$ в языке L. *Цель настоящей работы* — изучение этих двух функций и, в частности, асимптотик при больших значениях аргументов.

Для изучения величин $\psi_L(m_1,...,m_N)$ используем производящую функцию (точнее, формальный ряд) $\Psi_L(x)=\sum\limits_m \psi_L(m)x^m$, где $(x_1,...,x_N)=x$, $(m_1,...,m_N)=$ $=m$, $m_i\in Z_+$, $x_1^{m_1}\dots x_N^{m_N}=x^m$. Введем также аналогичный ряд $\Phi_L(t)=\sum\limits_{n=0}^{\infty}\varphi_L(n)t^n$, $n\in Z_+$. Назовем эти ряды ψ- и ϕ-рядом; коэффициенты этих рядов назовем ψ- и ϕ-коэффициентами.

Если в ряде $\Psi_L(x)$, $x\in R^N$, отождествить все переменные (т. е. принять $x_i=t$ для всех i), то получится $\Phi_L(t)$. Если язык задан явным списком входящих в него слов (упорядоченных каким-либо образом), то для получения ψ-ряда достаточно в каждом слове провести замену $\xi_i\to x_i$ и соединить полученные одночлены знаком суммирования.

Далее определим простые оценки указанных функций для языков, полученных из некоторых языков с помощью операций объединения, конкатенации, итерации. Затем изучим ψ-ряды для языков, порожденных автоматами; в частности, здесь отметим роль сильно связных автоматов. Приведем пример вычисления асимптотики ψ-коэффициентов для языка в алфавите, состоящем из двух букв.

Ряды $\Psi_L(x_1,...,x_N)$ и $\Phi_L(t)$ для языков, заданных регулярным выражением. Будем записывать регулярные выражения так, как это сделано в работе [6]. Таким образом, регулярное выражение содержит алфавитные буквы, символ пустого слова, символ итерации, знаки сложения, умножения и скобки. Зафиксируем биекцию между алфавитом A и набором переменных $\{x_1,...,x_N\}$, $\xi_i\leftrightarrow x_i$. Пусть дано регулярное выражение R в алфавите $A=\{\xi_1,...,\xi_N\}$. Построим по нему алгебраическое выражение с переменными x_i посредством следующих замен. Каждую букву $\xi_i\in A$ заменим буквой x_i, символ пустого слова — единицей 1, итерацию $(\alpha)^*=\bigcup_{i=0}^{\infty}(\alpha)^i$ — выражением $(1-\alpha)^{-1}$ (здесь используем формулу для «суммы геометрической прогрессии», полагая α «малой величиной»). Переменные x_i считаются коммутирующими, что позволяет преобразовать полученное выражение согласно обычным правилам. В результате приходим к выражению, которое обозначим через $A(R)$.

Поясним, как связаны выражение $A(R)$ и ψ -ряд $\Psi_L(x_1,...,x_N)$. Регулярное выражение позволяет перечислить язык L, т. е. построить последовательность слов, в которой содержатся все слова языка — возможно, с повторениями. Для получения такой последовательности следует выполнить действия над языками, диктуемые данным регулярным выражением. Если в этой последовательности нет повторяющихся слов, то выражение $A(R)$ совпадает с ψ -рядом (с точностью до порядка слагаемых). В противном случае, выражение $A(R)$ и ψ-ряд не совпадают.

Так, возьмем язык L, заданный регулярным выражением $R=(\xi_1\xi_2\cup\xi_1)\times$ $\times(\Lambda\cup\xi_2)$. С одной стороны, имеем $L=\{\xi_1, \xi_1\xi_2, \xi_1\xi_2\xi_2\}$, отсюда $\Psi(x_1,x_2)=$ $=x_1x_2+x_1+x_1x_2x_2$. С другой стороны, $A(R)=(x_1x_2+x_1)(1+x_2)=2x_1x_2+x_1+$ $+x_1x_2x_2$. Следовательно, $\Psi_L(x_1,x_2)$ и $A(R)$ не совпадают.

Операции над языками и формальные ряды. Введем следующие термины. Назовем язык L *кодовым*, если равенство $\alpha_1\ldots\alpha_r=\beta_1\ldots\beta_s$ возможно только при $r=s$, $\alpha_i=\beta_i$. Это условие выполняется, если ни одно слово языка не является началом другого слова.

Аналогично определим кодовую цепочку языков $L_1,...,L_r$ (здесь необходимо полагать, что $\alpha_i,\beta_i\in L_i$).

Отметим некоторые простые свойства введенных рядов:

а) если языки $L_1,...,L_r$ попарно не пересекаются, то $\Psi_{L_1\cup...\cup L_r}=\Psi_{L_1}+...+\Psi_{L_r}$;

б) если цепочка языков $L_1,...,L_r$ является кодовой, то $\Psi_{L_1...L_r}=\Psi_{L_1}\ldots\Psi_{L_r}$;

в) если язык L — кодовый, то $\Psi_{L^*}=(1-\Psi_L)^{-1}$.

Здесь не будем останавливаться на доказательствах этих свойств, ибо они достаточно просты.

Применим изложенные простые соображения для получения оценок величин $\varphi_L(n)$ и $\psi_L(m_1,...,m_N)$.

Будем выражать оценки указанных величин через соответствующие формальные ряды. Далее обозначим через $\tilde{x}$ и $\tilde{m}$ наборы переменных $(x_1,...,x_N)$ и $(m_1,...,m_N)$. Имея формальные ряды $\Psi^s=\sum\limits_{\tilde{m}}a^s_{\tilde{m}}\,x_1^{m_1}\ldots x_N^{m_N}$, $s=1,2$, запишем $\Psi^1\le\Psi^2$, если каждый коэффициент первого ряда не превосходит соответствующего коэффициента второго ряда. Имея ряды Ψ^j, $1\le j\le M$, определяем ряд $\max_j\Psi^j$, взяв в качестве его коэффициентов максимум соответствующих коэффициентов данных рядов.

Теперь обратимся к неравенствам для рядов.

В следующей лемме использован линейный оператор D^p, действующий на функцию $x_1^{m_1}\ldots x_N^{m_N}$ следующим образом:

$$D^p(x_1^{m_1}\ldots x_N^{m_N})=\frac{1}{C_{n-1}^{p-1}}x_1^{m_1}\ldots x_N^{m_N},$$

где $n=m_1+\ldots+m_N$.

Для простоты ограничимся случаем, когда рассматриваемые в лемме языки L_i, L не содержат пустого слова Λ.

Лемма 1. *Справедливы неравенства*

$$\max_{1\le i\le p}\Psi_{L_i}(\tilde{x})\le\Psi_{L_1\cup\ldots\cup L_p}(\tilde{x})\le\sum_{i=1}^{p}\Psi_{L_i}(\tilde{x}); \tag{1}$$

$$\frac{\sum_{i=1}^{p}\Psi_{L_i}(\tilde{x})}{p}\le\Psi_{L_1\cup\ldots\cup L_p}(\tilde{x}); \tag{2}$$

$$D^p\prod_{i=1}^{p}\Psi_{L_i}(\tilde{x})\le\Psi_{L_1\ldots L_p}(\tilde{x})\le\prod_{i=1}^{p}\Psi_{L_i}(\tilde{x}); \tag{3}$$

$$\frac{1}{1-\Psi_L(\tilde{x}/2)}\le\Psi_{L^*}(\tilde{x})\le\frac{1}{1-\Psi_L(\tilde{x})}. \tag{4}$$

◄ Неравенства (1) и (2) очевидны. Рассмотрим неравенство (3) при условии $\Lambda\notin L$. Слова состава $\tilde{m}$, входящие в множество $L_1\ldots L_p$, получаются следующим образом. Сначала рассматриваются всевозможные наборы векторов $\tilde{m}_k$, сумма которых есть $\tilde{m}$, и для каждого такого вектора составляются конкатенации слов состава $\tilde{m}_k$, взятых (последовательно) из языков L_k. Затем из полученного набора слов выделяются наборы одинаковых слов, такой набор заменяется одним словом, взятым из него. Первый из этих шагов приводит к оценке сверху в неравенстве (3). Для получения оценки снизу необходимо учесть, сколько раз можно получить одно и то же слово как конкатенации слов из языков L_i. Это число не превосходит числа способов разделить последовательность длиной $n=m_1+m_2+\ldots+m_N$ на p непустых частей. Решая комбинаторную задачу, получаем, что это число есть C_{n-1}^{p-1}. Неравенство (3) доказано.

Докажем неравенство (4). По определению итерации, $L^*=\Lambda\cup L\cup\cup L^2\cup\ldots\cup L^i\cup\ldots$ Тогда из неравенств (1) и (3) следует, что $\Psi_{L^*}(\tilde{x})\le$ $\le 1+\Psi_L(\tilde{x})+\Psi_L^2(\tilde{x})+\ldots=\dfrac{1}{1-\Psi_L(\tilde{x})}$. Причем эта оценка точна для кодовых языков. Оценка сверху доказана.

Оценка снизу получается из следующего наблюдения.

Возьмем слово $\alpha=\alpha_1\ldots\alpha_r$ из языка L^*, состав которого можно представить в виде $(m_1,\ldots,m_N)=\sum_{i=1}^{r}(m_1^i,\ldots,m_N^i)$, где m^i — состав слова α_i. Взяв конкатенации $\alpha_1\ldots\alpha_r$ таких слов, получаем $\sum\prod_{i=1}^{r}\psi_L(m_1^i,\ldots,m_N^i)$ слов (сумма распространяется на все целые неотрицательные m_i). Среди таких слов могут быть совпадающие. Ясно,

что число различных слов не меньше, чем $\frac{1}{2^{m_1+\ldots+m_N}} \sum\limits_{m_1,\ldots,m_N} \prod\limits_{i=1}^{r} \psi_L(m_1^i,\ldots,m_N^i)$, так как слово состава $(m_1,\ldots,m_N)$ имеет длину $m_1+\ldots+m_N$, поэтому его можно разбить на части $2^{m_1+\ldots+m_N}$ способами. Это дает оценку снизу в неравенстве (4). ►

Следствие. Справедливо неравенство

$$\frac{1}{1-\Phi_L(x/2)} \le \Phi_{L^*}(x) \le \frac{1}{1-\Phi_L(x)}.$$

До сих пор речь шла о рядах для рассматриваемых языков. Теперь обратимся к их коэффициентам. Здесь ограничимся величинами $\Phi_{L^*}(n)$, полагая язык L кодовым. Отметим, что эти величины являются коэффициентами ряда $(1-\Phi_L(x))^{-1}$. Таким образом, для их вычисления следует найти n-производную последней функции $(1-\Phi_L(x))^{-1}$ для любого n, для чего используем формулу Фаа ди Бруно [9] для производной сложной функции:

$$\frac{d^n}{dx^n}(g(f(x))) = \sum_{m_1,\ldots,m_n} \frac{n!}{m_1!\cdots m_n!} g^{(m_1+\ldots+m_n)}(f(x)) \prod_{j=1}^{n} \left(\frac{f^{(j)}(x)}{j!} \right)^{m_j},$$

где $1\cdot m_1 + 2\cdot m_2 + \ldots + n\cdot m_n = n.$

В рассматриваемом случае $g(x) = \frac{1}{1-x}$, $f(x) = \Phi_L(x)$. Применим формулу Фаа ди Бруно:

$$\frac{d^n}{dx^n}(g(f(x))) = \sum_{m_1,\ldots,m_n} \frac{n!}{m_1!\cdots m_n!} \frac{(-1)^{m_1+\ldots+m_n}(m_1+\ldots+m_n)!}{\left(1-\sum a_i x^i\right)^{m_1+\ldots+m_n}} \prod_{j=1}^{n} \left(\frac{(1-\sum a_i x^i)^{(j)}}{j!} \right)^{m_j}.$$

Полагая $x = 0$, получаем

$$\left.\frac{d^n}{dx^n}(g(f(x)))\right|_{x=0} = \sum_{m_1,\ldots,m_n} \frac{n!(-1)^{m_1+\ldots+m_n}(m_1+\ldots+m_n)!}{m_1!\ldots m_n!} \prod_{j=1}^{n} (-\varphi_L(j))^{m_j} =$$

$$= \sum_{m_1,\ldots,m_n} \frac{n!(m_1+\ldots+m_n)!}{m_1!\cdots m_n!} \prod_{j=1}^{n} (\varphi_L(j))^{m_j} = n!\varphi_{L^*}(n).$$

Таким образом,

$$\varphi_{L^*}(n) = \sum_{m_1,\ldots,m_n} \frac{(m_1+\ldots+m_n)!}{m_1!\cdots m_n!} \prod_{j=1}^{n} (\varphi_L(j))^{m_j}.$$

Язык, заданный автоматом. Почти-детерминированный автомат. Рассмотрим ориентированный граф, на каждом ребре которого написана буква алфавита. Полагаем, что в графе допустимы петли и кратные ребра (обычно определяется как псевдограф).

Перенумеровав вершины числами $1,...,n$, рассмотрим матрицу $A=(a_{ik})$, где a_{ik} — буква алфавита, стоящая на ребре (i,k); $a_{ik}=0$, если отсутствует ребро из i в k. Возьмем вершины u, v и рассмотрим язык $L(u,v)$, состоящий из слов, прочитываемых на всевозможных путях, которые ведут из вершины u в вершину v. Матрица, составленная из языков $L(u,v)$, — это $A^*=E+A+...$ (E — единичная матрица). Цель — получить ψ-ряд для каждого языка $L(u,v)$; матрицу, составленную из этих рядов, назовем матричным ψ -рядом (аналогично определяется матричный ϕ-ряд). Опишем эти ряды для графов со следующим свойством: для любой вершины буквы на исходящих из нее ребрах попарно различны. Назовем такой граф почти-детерминированным.

Такие графы обладают следующим свойством: на двух разных путях, ведущих из вершины u в вершину v, прочитываются разные слова. Таким образом, если взять все пути из u в v и для каждого пути записать прочитываемое на этом пути слово, то получим все слова языка $L(u,v)$, выписанные по одному разу. Следовательно, для получения матричного ψ -ряда достаточно взять в формуле $A^*=E+A+...$ коммутирующие переменные x_i вместо букв алфавита ξ_i и преобразовать полученное выражение по обычным правилам преобразования многочленов и формальных рядов.

С учетом изложенного выше, матричный Ψ-ряд получается разложением в ряд функции $(E-A)^{-1}$.

Далее, для каждого языка $L(u,v)$ величина $\varphi_L(n)$ (число слов длиной n) имеет два описания. Во-первых, это число путей длиной n, ведущих из вершины u в вершину v. Во-вторых, это матричный элемент матрицы A_1^n, стоящий на позициии (u,v), где A_1 — матрица графа. Для матрицы «общего положения» (все корни характеристического уравнения действительны) имеем асимптотическое соотношение $\varphi_L(n)\simeq c\lambda^n$, где λ — наибольший из модулей собственных значений матрицы A, $c=\text{const}$.

Роль сильно связных графов. Напомним, что ориентированный граф $G=(V,E)$ называется сильно связным, если для любых его вершин u,v существует путь $u\Rightarrow v$, а также путь $v\Rightarrow u$. Выразим языки $L(u,v)$, u, $v\in V$, через языки, связанные с сильно связными компонентами графа. Это позволит выразить ψ- и ϕ-ряды для языков $L(u,v)$ через аналогичные ряды для сильно связных компонент указанных графов.

Пусть $\overline{G}=(\overline{V},\overline{E})$ — конденсация графа G; это граф, вершины которого — сильно связные компоненты графа G. Обозначим их через $V(\alpha),V(\beta),...$ Согласно определению конденсации графа, имеем ребро, идущее от компоненты $V(\alpha)$ к компоненте $V(\beta)$, если в исходном графе существует ребро, начало которого лежит в компоненте $V(\alpha)$, а конец — в компоненте $V(\beta)$; назовем это ребро мостом из $V(\alpha)$ в $V(\beta)$. Поскольку допустимы кратные ребра, одному мосту в исходном графе может соответствовать несколько ребер. Имеем есте-

ственное отображение $V \to \overline{V}$, которое порождает отображение, переводящее любой путь в G в путь в $\overline{G}$.

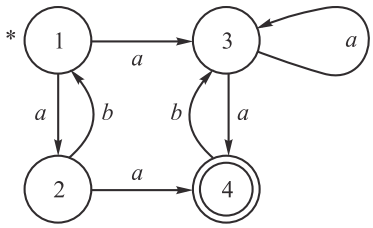
Теперь опишем язык $L(u,v)$. Возьмем в $\overline{G} = (\overline{V}, \overline{E})$ любой путь из $\overline{u}$ в $\overline{v}$. Пусть этот путь имеет вид $V(\alpha_1) \to \ldots \to V(\alpha_r)$, $u \in V(\alpha_1)$, $v \in V(\alpha_r)$. Выберем мосты (x_i, y_i), ведущие из $V(\alpha_{i-1})$ в $V(\alpha_i)$. Возникают языки $L(y_{i-1}, x_i)$, а мост (x_i, y_i) порождает язык, состоящий из одного однобуквенного слова (это буква алфавита, написанная на этом ребре). При упомянутом выше отображении указанный путь в $\overline{G} = (\overline{V}, \overline{E})$ порождает путь в $G = (V, E)$, ведущий из вершины $y_1 = u$ в вершину $x_{r+1} = v$, а этот путь дает слово. Множество таких слов и есть язык $L(u,v)$. Видим, что язык представляет собой объединение конкатенаций языков вида $L(y_1, x_2)L(x_2, y_2)\ldots L(y_{i-1}, x_i)L(x_i, y_i)\ldots L(y_r, x_{r+1})$, каждый из которых определяется цепочкой мостов (x_i, y_i), ведущих из u в v.

Следовательно, $L(u,v)$ — объединение языков, каждый из которых представляет собой конкатенацию языков, заданных сильно связными компонентами $V(\alpha), V(\beta), \ldots$, и однобуквенных языков, заданных мостами. Этот способ вывода регулярного выражения для языка, заданного автоматом, нередко приводит к цели быстрее, чем обычно употребляемый метод Макнотона — Ямады.

Наконец, приходим к следующему выводу: ψ-ряд для $L(u,v)$ можно оценить через аналогичные ψ-ряды, составленные для сильно связных компонент графа. В некоторых случаях можно получить точные выражения (вместо оценок).

Пример. Граф

$$G = (V, E),\ V = \{1,2,3,4\},\ E = \{1 \to 2, 2 \to 1, 3 \to 4, 4 \to 3, 3 \to 3, 1 \to 3, 2 \to 4\}.$$



Граф

На ребрах написаны (последовательно) буквы a, b, a, b, a, a, b. Начальная вершина — 1, заключительная — 4 (рисунок). Связные компоненты — $V_1 = (1,2)$ и $V_2 = (3,4)$. Имеем два моста $1 \to 3$ и $2 \to 4$, связывающих компоненты. Найдем, например, L_{14}, применим описанное выше представление языка через языки, связанные со связными компонентами. Получаем $L_{14} = L_{11}aL_{34} \cup L_{12}bL_{44}$. Далее

$$L_{11} = (ba)^*,\ L_{34} = a^*b(aa^*b)^*,\ L_{12} = (ba)^*b,$$
$$L_{44} = (aa^*b)^*,\ L_{14} = (ab)^*a(aa^* \cup b)(baa^*)^*.$$

Откуда следует, что ψ-ряд получается разложением функции $\dfrac{b(a+b-ba)}{(1-ba)(1-a-ba)}$.

Работа с классами эквивалентности формальных рядов. Рассмотрим еще одну характеристику языков, заданных графами, которая родственна с ψ-рядами.

Предварительно введем некоторые отношения между формальными рядами, используя отношение неравенства между рядами. Возьмем формальные ряды $A(x)=\sum a(m)x^m$, $B(x)=\sum b(m)x^m$, $x=(x_1,...,x_N)$, $m=(m_1,...,m_N)$. Запишем $A(x)\simeq B(x)$ (отношение эквивалентности), если существуют такие полиномы $\gamma_1(x)$, $\gamma_2(x)$, $\gamma_1'(x)$, $\gamma_2'(x)$, что $A(x)\le\gamma_1(x)B(x)+\gamma_2(x)$ и $B(x)\le\gamma_1'(x)A(x)+\gamma_2'(x)$. Класс эквивалентности, содержащий ряд $A(x)=\sum a(m)x^m$, обозначим через $\tilde{A}$. Произведение классов $\tilde{A}$, $\tilde{B}$ определим как класс, содержащий ряд AB. Это определение корректно, ибо, как легко убедиться, из соотношений $A_1(x)\simeq B_1(x)$ и $A_2(x)\simeq B_2(x)$ вытекает соотношение $A_1(x)A_2(x)\simeq B_1(x)B_2(x)$. Однако определить аналогичным образом сумму двух классов невозможно; определим сумму формально (записывая два класса со знаками «+» между ними). Другими словами, образуем свободную абелеву группу, образованную классами.

Рассмотрим любой связный граф G, возьмем в нем такие вершины u, v что вершина v достижима из u. Рассмотрим язык $L(u,v)$, его ψ-ряд $\Psi_{(u,v)}(x)$ и класс эквивалентности $\tilde{\Psi}_{(u,v)}$ этого ряда. Итак, любой паре u,v поставлен в соответствие класс эквивалентности. Это и есть обещанная характеристика языка, заданного графом.

Теорема 1. *Если граф G сильно связный, то все ряды $\Psi_{(u,v)}(x)$ попарно эквивалентны.*

◄ Возьмем две пары вершин u_1,v_1 и u_2,v_2 и слова α, β, прочитываемые на фиксированных путях, ведущих из u_1 в u_2 из v_2 в v_1 соответственно. Ясно, что язык $L_{u_1v_1}$ содержит в себе язык $\alpha L_{u_2v_2}\beta$. Откуда $\Psi(\alpha L_{u_2v_2}\beta)\subset\Psi_{u_1v_1}$; следовательно, взяв мономы $m_1=\Psi_\alpha$, $m_2=\Psi_\beta$, получаем $m_1\Psi_{u_2,v_2}m_2\le\Psi_{u_1,v_1}$. Аналогично находим такие мономы m_1', m_2', что $m_1'\Psi_{u_1,v_1}m_2'\le\Psi_{u_2,v_2}$. Итак, $\Psi_{u_1,v_1}\simeq\Psi_{u_2,v_2}$.►

Следовательно, для сильно связного графа G определен класс эквивалентных рядов. Обозначим этот класс через $\tilde{\Psi}^G$.

Рассмотрим граф G, в котором возьмем вершины u, v (вершина v достижима из u). Для любого пути α в конденсации графа получаем формальную сумму классов $\tilde{\Psi}\alpha$, относящихся к компонентам, которые встречаются на этом пути.

Теорема 2. *Справедливо равенство* $\tilde{\Psi}_{(u,v)}=\sum\limits_{\alpha}\tilde{\Psi}^{G\alpha}$.

◄ Для доказательства отметим, что можем выразить ψ-ряд для графа G через ψ-ряды, составленные для сильно связных компонент графа G, и слагаемые, порожденные мостами. Можем отбросить эти слагаемые, что не меняет класса эквивалентности ряда. Откуда вытекает утверждение теоремы. ►

Пример вычисления асимптотики ψ-коэффициентов. Вычисление асимптотики величины $\psi_L(m_1,...,m_N)$ (напомним, что это число слов данного языка L, в которые буква ξ_i входит m_i раз) — достаточно нетривиальная задача. Ограничимся алфавитом $\{a,b\}$ и языком $L=(a^2\cup b^2\cup ab)$. Этот язык — кодовый.

Найдем коэффициенты ряда Φ_{L^*}. Известно, что указанные ψ-коэффициенты возникают как коэффициенты разложения в степенной ряд функции $(1-a^2-b^2-ab)^{-1}$. Имеем

$$(1-a^2-b^2-ab)^{-1}=\sum_{i,j,k}\frac{(i+j+k)!}{i!j!k!}a^{2i+k}b^{2j+k},$$

отсюда получаются искомые коэффициенты $\psi(m,n)$ при одночленах $a^m b^n$. Ясно, m,n числа одинаковой четности. Рассмотрим сначала случай четных чисел m,n; при этом легко получается равенство $\psi(2m,2n)=\sum\frac{(m+n)!}{(m-k)!(n-k)!(2k)!}$. Найдем асимптотику этой величины при $m,n\to\infty$, изучим случай $m\le n$.

Преобразуем выражение $\psi(2m,2n)$ следующим образом (в этом месте заменим формулы словесным описанием):

а) заменим все факториалы величинами, которые возникают из формулы Стирлинга (см. работу [10], главу 3);

б) заменим суммирование по целочисленной переменной k интегрированием по переменной $k\in[0,+\infty)$;

в) выполним замены $k=mt,\ n=\alpha m$.

Разумеется, следует доказать, что при переходе от суммирования к интегрированию погрешность пренебрежима. Необходимые для этого оценки можно найти в главе 3 работы [11], здесь эту проверку рассматривать не будем. В результате получим следующее:

$$\psi(2m,2n)\simeq\frac{1}{2\pi m}\int_0^1\sqrt{\frac{1+\alpha}{(1-t)(\alpha-t)}}(w(t))^m dt,$$

где $w(t)=(1+\alpha)^{1+\alpha}(1-t)^{t-1}(\alpha-t)^{t-\alpha}(2t)^{-2t}$.

Находим асимптотику интеграла методом Лапласа, полагая, что $m\to+\infty$. Элементарное изложение метода Лапласа содержится в главе 2 работы [12]. Напомним, что этот метод рекомендует следующий способ оценки интеграла, зависящего от параметра.

Пусть $u(t),v(t),t\in[a,b]$ — гладкие функции и $t_0\in(a,b)$, точка максимума функции v. Тогда

$$\int_{[a,b]}u(t)\exp(mv(t))dt\simeq\sqrt{\pi}u(t_0)\exp(mv(t_0))\sqrt{\frac{2}{-mv''(t_0)}},\ \ m\to+\infty.$$

В рассматриваемом случае $v(t)=(t-1)\ln(1-t)+(t-\alpha)\ln(\alpha-t)-2t\ln(2t)$, $w(t)=\exp(v(t))$, и максимум функции $v(t)$ достигается в точке $t_0=$ $=(-\alpha-1+\sqrt{\alpha^2+14\alpha+1})/6$. В результате получаем

$$\psi(2m,2n)\simeq Cm^{-3/2}D^m,$$

где

$$C = \left(\frac{(1+\alpha)t_0}{(2\alpha+2)t_0 - 4\alpha} \right)^{1/2}; \quad D = (1+\alpha)^{1+\alpha}(\alpha - t_0)^{1-\alpha} / 4\pi^2, m \le n, \alpha = n/m, m \to +\infty.$$

Это асимтотическое соотношение выполняется равномерно при $\alpha \in (\alpha_0, \beta_0)$ для фиксированных $\alpha_0, \beta_0,$ удовлетворяющих условию $1 < \alpha_0 < \beta_0$.

В случае нечетных чисел *m*, *n* требуются некоторые изменения в вычислениях, но асимптотика оказывается такой же.

Заключение. Задача нахождения асимптотики функций числа слов заданного состава достаточно трудоемка. При наложении некоторого ограничения на множество (регулярность языка) и сведении к более удобным случаям (кодовые языки) можно находить оценки для искомой функции, в том числе, и точные. Предложен метод получения таких оценок и введен аппарат, который может быть использован в дальнейших исследованиях. Следует отметить, что в качестве следствий для утверждений о числе слов заданного состава получаются утверждения о числе слов заданной длины, которые можно сравнить с результатами, полученными ранее.

ЛИТЕРАТУРА

1. *Гончаров В.Л.* Из области комбинаторики // Известия АН СССР. Сер. матем. 1944. Т. 8. Вып. 1. С. 3–48.

2. *Коршунов А.Д.* Об асимптотике числа бинарных слов с заданной длиной максимальной серии I // Дискретный анализ и исследование операций. 1997. Т. 4. № 4. С. 13–46.

3. *Косточка А.В., Мазуров В.Д., Савельев Л.Я.* Число *q*-ичных слов с ограничением на длину максимальной серии // Дискретная математика. 1998. Т. 10. Вып. 1. С. 10–19.
DOI: https://doi.org/10.4213/dm413

4. *Хомский М., Миллер Д.* Языки с конечным числом состояний // Сер. Киб. сборник. Вып. 4. М.: ИЛ, 1962. С. 233–255.

5. *Строгалов А.С.* О регулярных языках с полиномиальным ростом числа слов // Дискретная математика. 1990. Т. 2. Вып. 3. С. 145–152.

6. *Белоусов А.И., Ткачев С.Б.* Дискретная математика. М.: Изд-во МГТУ им. Н.Э. Баумана, 2004. 744 с.

7. *Трахтенброт Б.А., Бардзин Я.М.* Конечные автоматы (поведение и синтез). М.: Наука, 1970. 400 с.

8. *Мотвани Р., Ульман Дж., Хопкрофт Дж.* Введение в теорию автоматов, языков и вычислений. М.–СПб.–Киев: Вильямс, 2002. 528 с.

9. *Архипов Г.И., Садовничий В.А., Чубариков В.Н.* Лекции по математическому анализу. М.: Дрофа, 2003. 640 с.

10. *Копсон Э.* Асимптотические разложения. М.: Мир, 1966. 160 с.

11. *Адельсон-Вельский Г.М., Кузнецов О.П.* Дискретная математика для инженера. М.: Энергоатомиздат, 1988. 480 с.

12. *Федорюк М.В.* Асимптотика, интегралы и ряды. М.: Наука, 1987. 544 с.

Исмагилов Раис Сальманович — д-р физ.-мат. наук, профессор кафедры «Высшая математика» МГТУ им. Н.Э. Баумана (Российская Федерация, 105005, Москва, 2-я Бауманская ул., д. 5, стр. 1).

Мастихина Анна Антоновна — канд. физ.-мат. наук, доцент кафедры «Высшая математика» МГТУ им. Н.Э. Баумана (Российская Федерация, 105005, Москва, 2-я Бауманская ул., д. 5, стр. 1).

Филиппова Лариса Евгеньевна — доцент департамента прикладной математики Московского института электроники и математики им. А.Н. Тихонова, Национальный исследовательский университет «Высшая школа экономики» (Российская Федерация, 101000, Москва, Мясницкая ул., д. 20).

Просьба ссылаться на эту статью следующим образом:

Исмагилов Р.С., Мастихина А.А., Филиппова Л.Е. О числовых характеристиках формальных языков // Вестник МГТУ им. Н.Э. Баумана. Сер. Естественные науки. 2017. № 4. С. 4–15. DOI: 10.18698/1812-3368-2017-4-4-15

ON NUMERICAL CHARACTERISTICS OF FORMAL LANGUAGES

R.S. Ismagilov[1] ismagil@bmstu.ru
A.A. Mastikhina[1] anmast@bmstu.ru
L.E. Filippova[2]

[1] Bauman Moscow State Technical University, Moscow, Russian Federation
[2] Higher School of Economics Tikhonov Moscow Institute of Electronics and Mathematics, Moscow, Russian Federation

Abstract

The purpose of this study was to count the number of layers of a regular language of a given composition. First, we specified the alphabet from the characters and defined the composition of the word as a vector. Then, we introduced the functions of the number of words of a given composition. Finally, we defined simple estimates of these functions for languages derived from languages using ordinary operations (union, concatenation, iteration). We also studied series for languages generated automatically

Keywords

Formal language, regular language, graph, generating function, order, composition of words

REFERENCES

[1] Goncharov V.L. From combinatorial theory. *Izvestiya AN SSSR. Ser. Matem.*, 1944, vol. 8, iss. 1, pp. 3–48 (in Russ.).

[2] Korshunov A.D. On the asymptotics of the number of binary words with a given length of a maximal series. I. *Diskretnyy analiz i issledovanie operatsiy*, 1997, vol. 4, no. 4, pp. 13–46 (in Russ.).

[3] Kostochka A.V., Mazurov V.D., Savelyev L.Ya. The number of q-ary words with restrictions on the length of a maximal series. *Discrete Mathematics and Applications*, 1998, vol. 8, no. 2, pp. 109–118. DOI: https://doi.org/10.1515/dma.1998.8.2.109

[4] Chomsky N., Miller G.A. Finite state languages. In: Information and control. 1958, vol. 1, no. 2, pp. 91–112.

[5] Strogalov A.S. Regular languages with polynomial growth in the number of words. *Discrete Mathematics and Applications*, 1992, vol. 2, no. 3, pp. 285–292.

[6] Belousov A.I., Tkachev S.B. Diskretnaya matematika [Descrete mathematics]. Moscow, Bauman MSTU Publ., 2004. 744 p.

[7] Trakhtenbrot B.A., Bardzin Ya.M. Konechnye avtomaty (povedenie i sintez) [Finite-state machines: Behaviour and synthesis]. Moscow, Nauka Publ., 1970. 400 p.

[8] Hopcroft J.E., Rajeev Motwani, Ullman J.D. Introduction to automata theory, languages, and computation. Addison Wesley, 2001. 537 p.

[9] Arkhipov G.I., Sadovnichiy V.A., Chubarikov V.N. Lektsii po matematicheskomu analizu [Lectures on mathematical analysis]. Moscow, Drofa Publ., 2003. 640 p.

[10] Copson E.T. Asymptotic expansions. Cambridge Univ. Press, 2004. 120 p.

[11] Adel'son-Vel'skiy G.M., Kuznetsov O.P. Diskretnaya matematika dlya inzhenera [Descrete mathematics for an engineer]. Moscow, Energoatomizdat Publ., 1988. 480 p.

[12] Fedoryuk M.V. Asimptotika, integraly i ryady [Asymptotics, integrals and raws]. Moscow, Nauka Publ., 1987. 544 p.

Ismagilov R.S. — Dr. Sc. (Phys.-Math.), Professor of Higher Mathematics Department, Bauman Moscow State Technical University (2-ya Baumanskaya ul. 5, str. 1, Moscow, 105005 Russian Federation).

Mastikhina A.A. — Cand. Sc. (Phys.-Math.), Assoc. Professor of Higher Mathematics Department, Bauman Moscow State Technical University (2-ya Baumanskaya ul. 5, str. 1, Moscow, 105005 Russian Federation).

Filippova L.E. — Assoc. Professor, School of Applied Mathematics, Higher School of Economics Tikhonov Moscow Institute of Electronics and Mathematics (Myasnitskaya ul. 20, Moscow, 101000 Russian Federation).

Please cite this article in English as:

Ismagilov R.S., Mastikhina A.A., Filippova L.E. On Numerical Characteristics of Formal Languages. *Vestn. Mosk. Gos. Tekh. Univ. im. N.E. Baumana, Estestv. Nauki* [Herald of the Bauman Moscow State Tech. Univ., Nat. Sci.], 2017, no. 4, pp. 4–15.
DOI: 10.18698/1812-3368-2017-4-4-15