РЕГУЛЯРНЫЕ ЯЗЫКИ И КОНЕЧНЫЕ АВТОМАТЫ

1. Алфавит, слово, язык

Алфавит — это произвольное *непустое конечное множество* $V = \{a_1, \ldots, a_n\}$, *элементы* которого называют **буквами**, или **символами**.

Определение 6.1. **Словом**, или **цепочкой**, в алфавите V называют произвольный кортеж из множества V^k (k-ой *декартовой степени* алфавита V) для различных $k = 0, 1, 2, \ldots$

Например, если $V = \{a, b, c\}$, то (a), (b), (c), (a, b), (a, b, c), (c, b, a, a, c) и т. д. есть слова в V.
При $k = 0$ получаем *пустой кортеж*, называемый в данном контексте **пустым словом**, или **пустой цепочкой** и обозначаемый λ.
Множество всех слов в алфавите V обозначают V^*, а множество всех непустых слов в V — V^+. Слова будем записывать без угловых скобок и запятых. Так, для записанных выше слов получим: a, b, c, ab, abc, $cbaac$.
Пустое слово λ — это слово, не имеющее символов.
Длину слова w можно понимать как число составляющих это слово букв.

Определение 6.2. Языком в алфавите V называется произвольное подмножество множества V^*.

Поскольку языки есть множества слов, к языкам применимы теоретико-множественные операции $\cup$, $\cap$, $\setminus$, $\triangle$, $\overline{}$ и т.д.

Определение 6.3. Соединением языков L_1 и L_2 называют язык L_1L_2, состоящий из всех возможных соединений слов xy, в которых слово x принадлежит первому, а слово y — второму языку, т.е.

$$L_1L_2 = \{xy | x \in L_1 \text{ и } y \in L_2\}.$$

Пример 1. Если $V = \{a, b, c\}$, $L_1 = \{ab, bcc, cab\}$, $L_2 = \{ca, bcc\}$, то

$$L_1L_2 = \{abca, abbcc, bccca, bccbcc, cabca, cabbcc\}.$$

Формально можно записать:

$$(ab + bcc + cab)(ca + bcc) = abca + abbcc + bccca + bccbcc + cabca + cabbcc.$$

Соединение языков не коммутативно. Например,

$$L_2L_1 = \{caab, cabcc, cacab, bccab, bccbcc, bcccab\} \neq L_1L_2.$$

Операция соединения языков позволяет определить операцию возведения языка в произвольную натуральную степень: для любого $L \subseteq V^*$ $L^0 = \{\lambda\}$, а для любого $n > 0$ $L^n = L^{n-1}L$.
Итерацией языка L называют объединение всех его степеней:

$$L^* = \bigcup_{n=0}^{\infty} L^n.$$

Рассматривая объединение всех степений языка L, начиная с первой, получим **позитивную итерацию**

$$L^+ = \bigcup_{n=1}^{\infty} L^n.$$

Основное алгебраическое свойство множества всех языков в алфавите V сформулировано в следующей теореме.

Теорема 1. Алгебра

$$\mathcal{L}(V) = (2^{V^*}, \cup, \cdot, \varnothing, \{\lambda\})$$

есть *замкнутое полукольцо*.

В *замкнутом полукольце* $\mathcal{L}(V)$ всех языков в алфавите V рассмотрим *подалгебру, порожденную множеством*, состоящим из пустого языка, языка $\{\lambda\}$ и всех однобуквенных языков $\{a\}$, $a \in V$, и *замкнутую относительно итерации*. Эта подалгебра, обозначаемая $\mathcal{R}(V)$, является *полукольцом с итерацией*.
Элементы полукольца $\mathcal{R}(V)$ называются **регулярными множествами**, или **регулярными языками**.

Определение 6.4. Пусть фиксирован некоторый алфавит V. Тогда:

1) Пустое множество $\varnothing$, множество $\{\lambda\}$ (состоящее из одной пустой цепочки) и множество $\{a\}$ для каждого $a \in V$ является регулярным языком (множеством) в алфавите V.

2) Если P и Q — регулярные языки в алфавите V, то *объединение* $P \cup Q$ и *соединение* PQ — регулярные языки в алфавите V.

3) Если P — регулярный язык в алфавите V, то итерация P^* — регулярный язык в алфавите V.

4) Никаких других регулярных языков, кроме определенных в пп. (1) — (3), не существует.

Алгебраические операции над регулярными множествами удобно представлять с помощью так называемых **регулярных выражений**.

Каждое регулярное выражение представляет (или обозначает) некоторое однозначно определяемое регулярное множество, причем:

1) регулярные выражения $\varnothing$, λ и a обозначают регулярные множества $\varnothing$, $\{\lambda\}$ и $\{a\}$ соответственно $(a \in V)$;

2) если регулярное выражение p обозначает регулярное множество P, а q обозначает Q, то регулярные выражения $(p+q)$, (pq) и (p^*) обозначают регулярные множества $P \cup Q$, PQ и P^* соответственно.

Для регулярного выражения $\alpha\alpha^*$ или $\alpha^*\alpha$ будем использовать обозначение α^+ и называть это выражение позитивной итерацией выражения α.

2. Вычисление языка, допускаемого КА

Определение 6.5. Конечный автомат — это *орграф, размеченный над полукольцом* $\mathcal{R}(V)$ *регулярных языков* в *алфавите* V, с выделенной *вершиной* q_0, которая называется **начальной** и выделенным подмножеством вершин F, каждый элемент которого называется **заключительной** вершиной.

На *функцию разметки* при этом накладываются следующие ограничения: *метка* каждой *дуги* есть либо язык $\{\lambda\}$, либо непустое подмножество *алфавита* V.

Вершины графа называют обычно в этом случае **состояниями конечного автомата**, начальную вершину — **начальным состоянием**, заключительную вершину — **заключительным состоянием конечного автомата**.

Если $e = (q, r)$ — *дуга* автомата M, и ее метка $\varphi(e)$ есть регулярное выражение λ, то в этом случае будем говорить, что в автомате M возможен **переход из состояния** q **в состояние** r **по пустой цепочке** и писать $q \to_\lambda r$. Дугу с меткой λ будем называть λ-**переходом** (или **пустой дугой**).

Если же метка дуги e есть множество, содержащее входной символ a, то будем говорить, что в автомате M возможен **переход из состояния** q **в состояние** r **по символу** a и писать $q \to_a r$.

Согласно общему определению *метки пути в размеченном орграфе метка пути* в конечном автомате есть *соединение*меток входящих в этот путь дуг (в порядке их прохождения). Таким образом, метка любого *пути конечной длины* в конечном автомате есть регулярный язык.

Если цепочка $x \in \varphi(W)$, где W — некоторый *путь, ведущий из вершины* q *в вершину* r конечного автомата M, то говорят, что **цепочка** x **читается на пути** W в M. Пишем $q \Rightarrow_x^* r$, если x читается на некотором пути из q в r.

Стоимость прохождения из состояния q в состояние r есть (согласно общему определению этого понятия в размеченных орграфах) *объединение* меток всех путей, ведущих из q в r, т. е. множество всех таких x, что $q \Rightarrow_x^* r$.

Язык $L(M)$ **конечного автомата** M есть множество всех цепочек во входном алфавите, читаемых в M на некотором пути из начального состояния в какое-либо из заключительных.

Чтобы найти язык конечного автомата, надо вычислить сумму тех элементов матрицы стоимостей автомата, которые находятся на пересечении строки, соответствующей начальному состоянию q_0 и в столбцов, соответствующих всем заключительным состояниям $q_f \in F$.
Чтобы практически вычислить язык конечного автомата, достаточно решить систему уравнений

$$\xi = A\xi + \beta, \tag{1}$$

где A - квадратная матрица n-ого порядка, элемент a_{ij} которой есть регулярное выражение, служащее меткой дуги из вершины (состояния) q_i в вершину (состояние) q_j, если такая дуга существует, и есть регулярное выражение $\varnothing$, если нет дуги из q_i в q_j;

β — столбец, компоненты с номерами $t_1, \ldots, t_m, ,$ соответствующих заключительным состояниям, равны единице полукольца (λ), а все остальные компоненты равны нулю полукольца ($\varnothing$).
(Ко всем уравнениям системы, соответствующим заключительным состояниям, добавляется слагаемое λ.)

Решение системы (1) будет иметь вид:

$$\xi = A^*\beta = A^* \begin{pmatrix} \varnothing \\ \vdots \\ \varnothing \\ \lambda \\ \varnothing \\ \vdots \\ \varnothing \\ \lambda \\ \varnothing \\ \vdots \\ \varnothing \end{pmatrix} \begin{matrix} \\ \\ \\ t_1 \\ \\ \\ \\ t_m \\ \\ \\ \end{matrix} \qquad (2)$$

(элементы λ находятся в строках с номерами $t_1, \ldots, t_m$). Умножая в (2) матрицу A^*, равную матрице C стоимостей, на столбец β, получим столбец, s-я компонента которого x_s будет равна произведению s-ой строки матрицы C $(c_{s1}, \ldots, c_{st_1}, \ldots, c_{st_m}, \ldots c_{sn})$ на столбец β , т.е.

$$x_s = c_{st_1} + \ldots + c_{st_m},$$

Полученное регулярное выражение описывает язык конечного автомата.

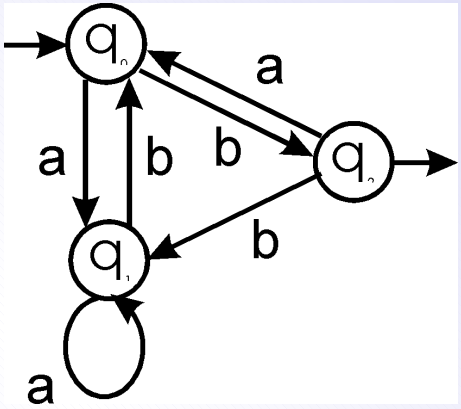


Рис. 1

Пример 2. Найдем язык конечного автомата, изображенного на рис. 1. Запишем для этого автомата систему уравнений:

$$\begin{aligned} x_0 &= ax_1 + bx_2, \\ x_1 &= bx_0 + ax_1, \\ x_2 &= ax_0 + bx_1 + \lambda, \end{aligned}$$

Слагаемое λ добавлено в уравнение для x_2, так как вершина q_2 является заключительной.

Исключая x_0, получим

$$x_1 = b(ax_1 + bx_2) + ax_1,$$
$$x_2 = a(ax_1 + bx_2) + bx_1 + \lambda.$$

Отсюда

$$x_1 = (ba + a)^* b^2 x_2,$$
$$x_2 = (a^2 + b)(ba + a)^* b^2 x_2 + abx_2 + \lambda.$$

Тогда

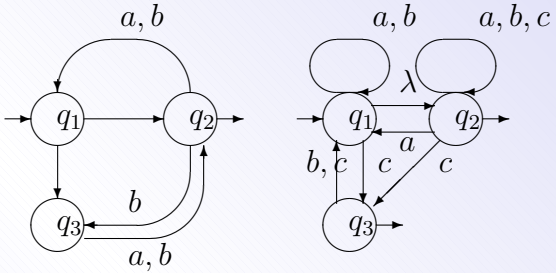
$$x_2 = ((a^2 + b)(ba + a)^* b^2 + ab)^*,$$
$$x_1 = (ba + a)^* b^2((a^2 + b)(ba + a)^* b^2 + ab)^*.$$

Отсюда получаем регулярное выражение, обозначающее язык КА, как значение переменной x_0:

$$x_0 = a(ba + a)^* b^2((a^2 + b)(ba + a)^* b^2 + ab)^* +$$
$$+ b((a^2 + b)(ba + a)^* b^2 + ab)^*.$$

Полученное регулярное выражение достаточно сложно, и найти его, не располагая заранее разработанным алгоритмом, было бы затруднительно.

Рис. 2



Задачи

6.1. Доказать, что язык $L^{+k} = \bigcup\limits_{i=k>0}^{\infty} L^i$ регулярен для любого k при условии регулярности L.

6.2. Привести примеры слов в алфавите $\{a, b, c\}$, которые задаются следующими регулярными выражениями:

(а) $(a+b)^*c^*(b+c)$;

(б) $((ab)^+(ca)^*)^*$;

(в) $(a^+(b+c)^*a + b^+(a+b)^*bc)^*$;

6.3. Найти языки, допускаемые конечными автоматами, заданными на рис. 2.

6.4. Найти язык, допускаемый конечным автоматом:

(а) вход: q_1 ; выходы: q_2 , q_3 ; дуги: (q_1, q_2, a) , (q_1, q_4, a, b) , (q_2, q_4, a, b) , (q_3, q_4, λ) , (q_4, q_3, a, b) , (q_3, q_2, a, b) , (q_4, q_2, b) ;

(б) входы: q_0 , q_1 ; выходы: q_2 , q_1 ; дуги: (q_0, q_2, a, b, c) , (q_0, q_1, a) , (q_1, q_0, a, b) , (q_1, q_2, a, c) , (q_2, q_0, c) , (q_2, q_2, a, b) .

6.5. Решить систему линейных уравнений с регулярными коэффициентами:

$$\begin{cases} x_1 = (01^* + 1)x_1 + x_2, \\ x_2 = 1x_1 + 00x_3 + 11, \\ x_3 = x_1 + x_2 + \lambda. \end{cases}$$

Для регулярного выражения, задающего компоненту решения x_3 , построить допускающий его конечный автомат.

6.6. Доказать, что линейное уравнение $x = \alpha x + \beta$ с регулярными коэффициентами:

(а) имеет единственное решение при $\lambda \notin \alpha$;

(б) имеет бесконечно много решений при $\lambda \in \alpha$, причем общее решение можно записать в виде $x = \alpha^*(\beta + L)$, где L — произвольный язык.

6.7. Инверсия цепочки $x \in V^*$ — зто цепочка символов x^R , полученная переписыванием x справа налево, т.е. если $x = a_{j1} \dots a_{jn}$, то $x^R = a_{jn} \dots a_{j1}$. Доказать, что если L — регулярный язык, то $L^R = \{y \,|\, y = x^R\,,\ x \in L\}$ — регулярный язык. Дайте два варианта доказательства:

1) используя только определение регулярного языка;

2) используя конечные автоматы.