

# 1 Регулярные языки, диаграммы, конечные автоматы

## 1.1 Языки в общем виде.

$A = \{a_1, a_2, \dots, a_n\}$  — алфавит;  $\alpha = a_{i_1}a_{i_2} \dots a_{i_k}$  — слова, где  $a_{i_j} \in A$ ,  $j = 1, \dots, k$ .

$\Lambda$  — пустое слово (нейтральный элемент по отношению к операции конкатенации, то есть присоединения)<sup>1</sup>. Для любого слова  $\alpha$  выполняются равенства:  $\Lambda\alpha = \alpha\Lambda = \alpha$ .

$A^* = \bigcup_{k \geq 0} A^k$ , где  $A^k$  — множество слов над алфавитом  $A$  длины  $k$ .

Языком  $L$  называется некоторое подмножество всех слов над алфавитом  $A$ , то есть  $L \subset A^*$ .

**Пример 1.** 1. Алфавит  $A = \{a, b, c, d, \dots, z\}$ , язык — множество слов английского языка, для написания которых стандартного алфавита достаточно.

2. Алфавит  $A = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 0\}$ , язык — множество всех натуральных чисел (10-ая система исчисления).

3. Алфавит  $A = \{x, y, z, \text{,}_1 \text{,}_2 \text{,}_3 \text{,}_4 \text{,}_5 \text{,}_6 \text{,}_7 \text{,}_8 \text{,}_9 \text{,}_0, \vee, \&, \neg, (, )\}$ , язык — множество всех формул, реализующих функции алгебры логики (использующие только конъюнкцию, дизъюнкцию и отрицание).

## 1.2 Регулярные языки

Определим операции над языками. Пусть  $A$  — некоторый алфавит,  $L_1 \subset A^*$ ,  $L_2 \subset A^*$ .

1. Объединение  $L_1 \cup L_2 = \{\alpha \in A^* \mid \alpha \in L_1 \text{ или } \alpha \in L_2\}$ .

2. Конкатенация  $L_1 \cdot L_2 = \{\alpha \in A^* \mid \alpha = \alpha_1\alpha_2, \alpha_1 \in L_1, \alpha_2 \in L_2\}$ .

3. Итерация  $L_1^* = \{\Lambda\} \cup \{\alpha \in A^* \mid \alpha = \alpha_1\alpha_2 \dots \alpha_k, \alpha_1 \in L_1, \alpha_2 \in L_1, \dots, \alpha_k \in L_1\}$ .

**Пример 2.** Пусть  $A = \{a, b, c\}$ ,  $L_1 = \{a, ab, abc, cba\}$ ,  $L_2 = \{ab, ac, abc, acb\}$ ,  $L_3 = \{a\}$ ,  $L_4 = \emptyset$ ,  $L_5 = \{b\}$ ,  $L_6 = \{c\}$ ,  $L_7 = \{\Lambda\}$ .

1.  $L_1 \cup L_2 = \{a, ab, abc, cba, ac, acb\}$ ,  $L_3 = \{a\}$ ;

2.  $L_1 \cdot L_2 = \{aab, aac, aabc, aacb, abab, abac, ababc, abacb, abcab, abcac, abcabc, abcacb, cbaab, cbaac, cbaaa\}$ ;

3.  $L_i \cdot L_7 = L_i$ , для  $i = 1, 2, 3, 5, 6, 7$ ; (Что будет в результате конкатенации  $L_4$  и  $L_7$ ?)

4.  $L_3^* = \{\Lambda, a, aa, aaa, \dots, a^k = \underbrace{a \dots a}_k, \dots\}$ .

Отметим, что операция конкатенации не коммутативна.

**Пример 3.** Обозначения языков из примера 2

1.  $L_1 \cdot L_3 = \{aa, aba, abca, cbaa\}$ ;

2.  $L_3 \cdot L_1 = \{aa, aab, aabc, acba\}$ ;

---

<sup>1</sup>Также распространённое обозначение для пустого слова  $\varepsilon$

*Регулярный язык* — множество слов, которое можно получить за конечное число операций объединения, конкатенации и итерации, начиная с  $\emptyset$ ,  $\{a_1\}$ ,  $\{a_2\}$ ,  $\dots$ ,  $\{a_n\}$ ,  $\Lambda$  (пустого языка и языков, содержащих одно однобуквенное слово).

**Пример 4.** Обозначения языков из примера 2

1. Любой язык, содержащий конечное число слов, является регулярным.
2. Язык, состоящий из всех слов над алфавитом  $A$  регулярный, так как это  $(L_3 \cup L_5 \cup L_6)^*$ .
3. Язык, состоящий из всех слов над алфавитом  $A$ , начинающихся с буквы  $a$ , регулярный, так как это  $L_3 \cdot (L_3 \cup L_5 \cup L_6)^*$ .

### 1.3 Представление языков диаграммами

*Диаграмма*<sup>2</sup>  $D(V, E)$  — конечный ориентированный граф.  $V = \{v_1, \dots, v_n\}$  — множество вершин,  $E = \{e_1, \dots, e_m\}$  — множество рёбер. Выделяется начальная вершина  $v_0 \in V$  и множество финальных (допустимых) вершин  $V' \subset V$  (случаи  $V' = \emptyset$ ,  $v_0 \in V$  также возможны). Каждому ребру приписывается некоторая буква алфавита  $A$  или символ пустого слова:  $\mu(e_i) \in A \cup \{\Lambda\}$  для всех  $e_i \in E$ .

Пусть  $p : v_0 \rightarrow v_k$  — путь из  $v_0$  в  $v_k$ , то есть последовательность рёбер  $e_1, e_2, \dots, e_k$   $((v_0, v_{i_1}), (v_{i_1}, v_{i_2}), \dots, (v_{i_{k-1}}, v_{i_k}))$ . Положим  $\alpha_p = \mu(e_1)\mu(e_2) \dots \mu(e_k)$ . Если длина пути  $p$  равна нулю, то по определению  $\alpha_p = \Lambda$ .

*Язык, представимый диаграммой  $D$*  (обозначается  $[D]$ ) —  $[D] = \{\alpha \in A^* \mid \exists v' \in V', \exists p : v_0 \rightarrow v', \alpha = \alpha_p\}$ .

**Теорема 1.** *Язык  $L$  является регулярным тогда и только тогда, когда существует диаграмма  $D$ , такая что  $L = [D]$ .*

**Лемма 1.** *Для всякого регулярного языка существует диаграмма  $D$ , такая что  $L = [D]$ .*

*Доказательство.* Доказательство будем вести индукцией по числу операций  $n$ , использованных для задания регулярного языка  $L$  (операции объединения, конкатенации, итерации)

1. Язык, не содержащий ни одного слова, реализуется диаграммой с двумя вершинами (начальной и финальной) и без рёбер.

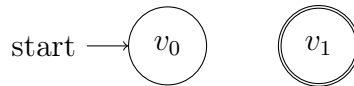


Рис. 1: Язык, не содержащий ни одного слова.

2. Язык, состоящий из одного пустого слова, реализуется диаграммой с одной вершиной (начальной и финальной одновременно) и без рёбер.
3. Язык, одержащий одно однобуквенное слово  $\{a_i\}$ , реализуется диаграммой с двумя вершинами (начальной и финальной) и одним ребром, их соединяющим, и помеченным буквой  $a_i$ .

<sup>2</sup>Также этот объект называется недетерминированным конечным автоматом.

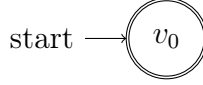


Рис. 2: Язык, состоящий из одного пустого слова.

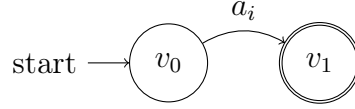


Рис. 3: Язык, содержащий одно слово  $a_i$

4.  $L = L_1 \cup L_2$ . Поскольку для описания языков  $L_1$  и  $L_2$  требуется меньшее число операций  $(\{\cup, \cdot, *\})$ , то по предположению индукции существуют диаграммы  $D_1$  и  $D_2$ , такие что  $L_1 = [D_1]$ ,  $L_2 = [D_2]$ . Пусть  $D_1 = (V_1, E_1)$ ,  $D_2 = (V_2, E_2)$ , начальные вершины —  $v_{01}$  и  $v_{02}$  соответственно, множества допустимых вершин —  $V'_1$  и  $V'_2$  соответственно.

Построим диаграмму  $D = (V, E)$  следующим образом. Положим  $V = V_1 \cup V_2 \cup \{v_0\}$ ,  $E = E_1 \cup E_2 \cup \{(v_0, v_{01}), (v_0, v_{02})\}$ , причём рёбрам  $(v_0, v_{01})$  и  $(v_0, v_{02})$  приписаны символы пустого слова  $\Lambda$ , начальная вершина —  $v_0$ , множество финальных вершин —  $V' = V'_1 \cup V'_2$ . Тогда  $L = [D]$ .

5.  $L = L_1 \cdot L_2$ . Поскольку для описания языков  $L_1$  и  $L_2$  требуется меньшее число операций  $(\{\cup, \cdot, *\})$ , то по предположению индукции существуют диаграммы  $D_1$  и  $D_2$ , такие что  $L_1 = [D_1]$ ,  $L_2 = [D_2]$ . Пусть  $D_1 = (V_1, E_1)$ ,  $D_2 = (V_2, E_2)$ , начальные вершины —  $v_{01}$  и  $v_{02}$  соответственно, множества допустимых вершин —  $V'_1$  и  $V'_2$  соответственно.

Построим диаграмму  $D = (V, E)$  следующим образом. Положим  $V = V_1 \cup V_2$ ,  $E = E_1 \cup E_2 \cup E'$ , где  $E'$  — множество рёбер, соединяющих вершины из множества  $V'_1$  с вершиной  $v_{02}$ , помеченные символом пустого слова  $\Lambda$ , то есть  $E' = \{(v, v_{02}) \mid v \in V'_1\}$ . Начальная вершина —  $v_{01}$ , множество допустимых вершин —  $V'_2$ . Тогда  $L = [D]$ .

6.  $L = L_1^*$ . Поскольку для описания языка  $L_1$  требуется меньшее число операций  $(\{\cup, \cdot, *\})$ , то по предположению индукции существует диаграмма  $D_1$ , такая что  $L_1 = [D_1]$ . Пусть  $D_1 = (V_1, E_1)$ , начальная вершина —  $v_{01}$ , множество допустимых вершин —  $V'_1$ .

Построим диаграмму  $D = (V, E)$  следующим образом. Положим  $V = V_1 \cup \{v_0\}$ ,  $E = E_1 \cup E'$ , где  $E'$  — множество рёбер, соединяющих вершины из множества  $V'_1$  с вершиной  $v_0$ , помеченные символом пустого слова  $\Lambda$ , а также ребро, соединяющее вершины  $v_0$  и  $v_{01}$ , также помеченное символом пустого слова  $\Lambda$ . Таким образом,  $E = E_1 \cup \{(v, v_{01}) \mid v \in V'_1\} \cup \{(v_0, v_{01})\}$ . Начальная вершина —  $v_0$ , множество допустимых вершин —  $v_0$ . Тогда  $L = [D]$ .

□

*В качестве упражнения рекомендуется сделать картинку к доказательству*

**Лемма 2.** Пусть  $D$  — диаграмма,  $L = [D]$ . Тогда  $L$  — регулярный.

*Доказательство.*  $D = (V, E)$ ,  $v_0$  — начальная вершина,  $V'$  — множество допустимых вершин. Положим  $q = |E|$  (то есть  $q$  — это число рёбер в диаграмме). Доказательство будем вести индукцией по  $q$ .

База индукции  $q = 0$ . Если  $v_0 \in V'$ , то  $[D] = \{\Lambda\}$ . Если  $v_0 \notin V'$ , то  $[D] = \emptyset$ .

Шаг индукции. Пусть  $q > 0$ , причём для всех диаграмм с меньшим числом рёбер утверждение доказано.  $v_0$  — начальная вершина. Если нет рёбер, исходящих из  $v_0$ , то случай аналогичен случаю  $q = 0$ .

Пусть есть рёбра, исходящие из вершины  $v_0$ . Выделим одно из таких рёбер. Пусть это ребро  $e = (v_0, v_1)$ , помеченное символом  $a$ . Рассмотрим следующие диаграммы:

$D_1 = (V, E \setminus \{e\})$ ,  $v_0$  — начальная вершина,  $V'$  — множество допустимых вершин.

$D_2 = (V, E \setminus \{e\})$ ,  $v_1$  — начальная вершина,  $V'$  — множество допустимых вершин.

$D_3 = (V, E \setminus \{e\})$ ,  $v_0$  — начальная вершина,  $\{v_0\}$  — множество допустимых вершин.

$D_4 = (V, E \setminus \{e\})$ ,  $v_1$  — начальная вершина,  $\{v_0\}$  — множество допустимых вершин.

Поскольку диаграммы  $D_1, D_2, D_3, D_4$  содержат меньше, чем  $q$  рёбер, то по предположению индукции существуют языки  $L_1, L_2, L_3, L_4$  соответственно, представимые этими диаграммами. Тогда  $L = L_1 \cup (L_3 \cup \{a\}L_4)^*\{a\}L_2$ .  $\square$

## 1.4 Конечные автоматы

Конечный автомат<sup>3</sup> — это пятёрка  $\{A, Q, G, q_0, Q'\}$ , где

$A = \{a_1, \dots, a_m\}$  — входной алфавит;

$Q = \{q_0, \dots, q_n\}$  — множество состояний;

$G : A \times Q \rightarrow Q$  — функция перехода, которая по каждой паре «текущее состояние» и «входной символ» показывает, в какое новое состояние должен перейти автомат;

$q_0 \in Q$  — начальное состояние;

$Q' \subseteq Q$  — множество финальных состояний<sup>4</sup>.

### Способы задания автомата

1. Диаграмма переходов — граф  $(W, E)$ , где  $W = Q$ , а вершины  $q_i$  и  $q_j$  соединяются направленным ребром, помеченным символом  $a_k$ , если  $G(a_k, q_i) = q_j$ , начальная вершина помечаются (стрелочкой или звёздочкой). Множество конечных вершин также выделяется (мы будем обводить их двойным контуром).

2. Канонические уравнения.

Функция перехода  $G$  определяется не только на буквах алфавита, но и на словах следующим образом:

$$\begin{aligned} G(\Lambda, q) &= q; \\ G(\alpha a, q) &= G(a, G(\alpha, q)). \end{aligned}$$

Таким образом из любого состояния по любому слову можно однозначно сказать, в какое состояние перейдёт автомат.

### Представимые языки

<sup>3</sup>Также называется детерминированный конечный автомат

<sup>4</sup>Недетерминированный конечный автомат можно задать аналогичным образом, однако функция перехода в этом случае имеет вид  $G : A^* \times Q \rightarrow P(Q)$ , где  $P(Q)$  — множество подмножеств  $Q$ , а область определения  $G$  может не совпадать с  $A^* \times Q$ .

Множество слов, представимых автоматом  $V - T(V) = \{\alpha \in A^* \mid G(\alpha, q_0) \in Q'\}$ . (также слова называются распознаваемыми, принимаемыми, допустимыми).

Язык  $L$  называется представимым, если найдётся автомат  $V$ , такой что  $L = T(V)$ .

**Теорема 2** (Клини). *Язык  $L$  является регулярным тогда и только тогда, когда он является представимым.*

*Доказательство.* Если язык является  $L$  представимым, то по определению существует автомат  $V$ , такой что  $L = T(V)$ . Пусть  $D$  — диаграмма переходов автомата  $V$  с соответствующими начальной и множеством финальных вершин. Тогда  $L = [D]$ . По теореме 1 язык является регулярным.

Пусть теперь язык является регулярным. Покажем, что он является представимым. Для этого построим автомат  $V$ , такой что  $L = T(V)$ . Поскольку язык регулярный, то по теореме 1 существует диаграмма  $D = (W, E)$ , такая что  $L = [D]$ . Пусть  $W = \{v_1, \dots, v_n\}$ ,  $v_1$  — начальная вершина,  $W' \subset W$  — множество допустимых вершин.

Обозначим через  $q_1, q_2, \dots, q_{2^n}$  все возможные подмножества множества  $W$ . Положим  $Q = \{q_1, \dots, q_{2^n}\}$ . Пусть  $v \in W$ . Определим функцию  $\theta$  (из  $A^* \times W$  в  $Q$ ):

$$\theta(\alpha, v) = \{v' \in W \mid \text{существует путь } P : v \rightarrow v', \alpha_P = \alpha\}.$$

Без ограничения общности будем считать  $q_1 = \Theta(\Lambda, v_1)$ . Положим

$$Q' = \{q_i \mid q_i \cap W' \neq \emptyset\},$$

$$G(a, q) = \bigcup_{v \in q} \theta(a, v)$$

Рассмотрим автомат  $V = (A, Q, G, q_1, Q')$ . Покажем, что язык, представимый автоматом  $V$  совпадает с языком  $L$ . Доказательство будем вести индукцией по длине  $k$  слова  $\alpha$ .

Пусть  $k = 0$ ,  $\alpha = \Lambda$ . Тогда  $G(\Lambda, q_1) = q_1 = \theta(\Lambda, v_1)$ .

Пусть  $k = 1$ ,  $\alpha = a$ , где  $a \in A$ . Тогда  $G(a, q_1) = \theta(a, q_1)$ .

Пусть  $k > 1$ ,  $\alpha = \alpha'a$ . Тогда

$$G(\alpha, q_1) = G(\alpha'a, q_1) = G(a, G(\alpha', q_1)) = G(a, \theta(\alpha', v_1)) = \bigcup_{v \in \theta(\alpha', v_1)} \theta(a, v) = \theta(\alpha'a, v_1).$$

Следовательно,  $\alpha \in T(V)$  тогда и только тогда, когда  $\alpha \in [D]$ . А значит,  $L = T(V)$ .  $\square$

## 1.5 Замкнутость семейства регулярных языков относительно теоретико-множественных операций

Пусть  $E_1, E_2$  — регулярные языки (над алфавитом  $A$ ). Тогда  $E_1 \cup E_2, E_1 \cap E_2, CE_1 = A^* \setminus E_1, E_1 \setminus E_2$  — также регулярные языки.

*Доказательство.* 1.  $E_1 \cup E_2$  — по определению регулярных языков.

2. Построим автомат, представляющий язык  $CE_1$ . Поскольку язык  $E_1$  регулярный, то существует автомат  $V = (A, Q, G, q_0, Q')$ , задающий язык  $E_1$ . Тогда для любого слова  $\tilde{\alpha}$  из языка  $E_1$  состояние  $G(\tilde{\alpha}, q_0)$  (то состояние, в которое придёт автомат

из начального при прочтении слова  $\tilde{\alpha}$ ) является финальным. Для любого слова  $\tilde{\beta}$  из языка  $C \setminus E_1$ , то есть слова над алфавитом  $A$ , не принадлежащего языку  $E_1$ , состояние  $G(\tilde{\beta}, q_0)$  (то состояние, в которое придёт автомат из начального при прочтении слова  $\tilde{\beta}$  не является финальным. Следовательно, язык  $CE_1$  задается автоматом  $\hat{V} = (A, Q, G, q_0, Q \setminus Q')$ , а значит является регулярным.

3.  $E_1 \cap E_2 = C(CE_1 \cup CE_2)$ .

4.  $E_1 \setminus E_2 = C(CE_1 \cup E_2)$ .

□

## 1.6 Пример нерегулярного языка

Пусть  $A = \{a, b\}$ . Рассмотрим язык  $L$ , состоящий из всех слов вида  $\underbrace{a \dots a}_k \underbrace{b \dots b}_k$ ,  $k \in \mathbb{N}$ , и только из них. Покажем, что этот язык не является регулярным.

Предположим, что язык  $L$  является регулярным. Тогда существует автомат  $\{A, Q, G, q_0, Q'\}$ , такой что язык  $L$  представим этим автоматом. Положим  $t = |Q|$ , то есть  $t$  — это число состояний автомата. Пусть  $m > t$ . Рассмотрим слово  $\tilde{\alpha} = \underbrace{a \dots a}_m \underbrace{b \dots b}_m$ . По определению языка  $L$  слово  $\tilde{\alpha}$  содержится в языке  $L$ . Следовательно,  $G(\tilde{\alpha}, q_0) \in Q'$  (то есть применяя функцию перехода к слову  $\tilde{\alpha}$ , автомат придёт в одно из допустимых состояний). Рассмотрим следующее подмножество состояний автомата

$$\begin{aligned} q_{i_1} &= G(a, q_0), \\ q_{i_2} &= G(aa, q_0), \\ &\dots \\ q_{i_s} &= G(\underbrace{a \dots a}_s, q_0), 1 \leq s \leq m, \\ &\dots \\ q_{i_m} &= G(\underbrace{a \dots a}_m, q_0). \end{aligned}$$

Поскольку  $m > t$ , то найдутся числа  $p, r, p < r$ , такие что  $q_{i_p} = q_{i_r}$ . А именно,

$$G(\underbrace{a \dots a}_p, q_0) = q_{i_p} = q_{i_r} = G(\underbrace{a \dots a}_r, q_0). \tag{1}$$

Рассмотрим слово  $\tilde{\beta} = \underbrace{a \dots a}_{m+r-p} \underbrace{b \dots b}_m$ . По определению языка  $L$  слово  $\tilde{\beta}$  не содержится в языке  $L$ . С другой стороны, применяя равенство 1, получаем

$$\begin{aligned} G(\tilde{\beta}, q_0) &= G(\underbrace{a \dots a}_{m+r-p} \underbrace{b \dots b}_m, q_0) = G(\underbrace{a \dots a}_r \underbrace{a \dots a}_{m-p} \underbrace{b \dots b}_m, q_0) = \\ &= G(\underbrace{a \dots a}_{m-p} \underbrace{b \dots b}_m, q_{i_r}) = G(\underbrace{a \dots a}_{m-p} \underbrace{b \dots b}_m, q_{i_p}) = \\ &= G(\underbrace{a \dots a}_p \underbrace{a \dots a}_{m-p} \underbrace{b \dots b}_m, q_0) = G(\underbrace{a \dots a}_m \underbrace{b \dots b}_m, q_0) = G(\tilde{\alpha}, q_0) \in Q'. \end{aligned}$$

Получили противоречие. Следовательно, язык  $L$  не является регулярным.