

## 2.9 Регулярни изрази. Теорема на Клини

В предишните параграфи установихме, че множеството на автоматните езици е затворено относно операциите обединение, сечение, допълнение, произведение, итерация и разлика. Този факт ни навежда на аналогия с алгебричните изрази в алгебрата, тъй като сумата, произведението и разликата на алгебрични изрази, отново е алгебричен израз. Тази аналогия по естествен начин извежда необходимост от разглеждане на изрази от думи на дадена азбука, които притежават свойства подобни на алгебричните изрази. Така достигаме до понятието за регулярни изрази. Да забележим, че алгебричните изрази винаги се свързват с някакво базово числово множество, в което са дефинирани и също така всеки израз представя някакво подмножество на базовото множество. Например, ако разгледаме алгебричния израз  $2 + x^2$  над множеството на целите числа, то той представя множеството

$$\{2, 3, 6, 11, 18, \dots\},$$

което е множеството от стойности на този израз. Както при алгебричните изрази, така и при регулярните изрази, се разглежда базово множество, което съвпада с множеството  $V^*$  от всички думи над дадена азбука  $V$ . Също така всеки регулярен израз представя някое подмножество на  $V^*$ . След тези бележки определението на регулярни изрази е напълно естествено.

**Определение 2.9.1** 1. *Регулярният израз  $\emptyset$  съответствува на множеството  $\emptyset$ .*

2. *Регулярният израз  $\epsilon$  представя множеството  $\{\epsilon\}$ .*

3. *Регулярният израз  $\omega_1 + \omega_2 + \dots + \omega_n$  представя крайното множество*

$$\{\omega_1, \omega_2, \dots, \omega_n\} \subset V^*.$$

4. *Ако  $z_1$  и  $z_2$  са регулярни изрази, представящи множеств-*

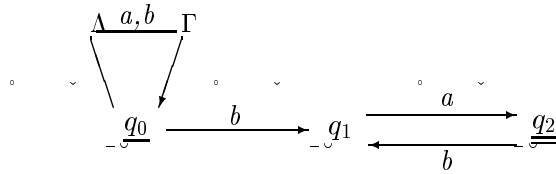
вата  $L_1$  и  $L_2$ , то:

- (i) изразът  $z_1 + z_2$  представя множеството  $L_1 \cup L_2$ ;
- (ii) изразът  $z_1 z_2$  представя множеството  $L_1 L_2$ ;
- (iii) изразът  $z_1^k$  представя множеството  $L_1^k$ , ( $k \geq 0$ );
- (iv) изразът  $z_1^*$  представя множеството  $L_1^*$ .

5. Регулярни са само изразите, описани в точки 1, 2, 3 и 4.

**Пример 2.9.1** Регулярните изрази  $(a+b)^*ba(ba)^*$  и  $(a+ba^*b)^*+b$  над азбуката  $V = \{a, b\}$  представят множествата  $\{a, b\}^* \{ba\} \{ba\}^* = \{a, b\}^* \{ba\}^+$  и  $(\{a\} \cup (\{b\} \{a\}^* \{b\}))^* \cup \{b\}$ .

Обикновено по зададен регулярен израз лесно се построява НКА, които разпознава езика, представен от този израз. Например за езика, представен от изразът  $(a+b)^*ba(ba)^*$  съответният НКА е представен графично на Фиг. 2.9.1



**Фиг. 2.9.1**

Свойствата на регулярните изрази следват непосредствено от свойствата на операциите с множествата, което се вижда от определението на регулярни изрази. Да посочим някои от по-важните свойства на регулярните изрази.

Ако  $Q, R, S$  са регулярни изрази над азбуката  $V$  то:

1.  $Q + Q = Q + \emptyset = Q$
2.  $Q + R = R + Q$
3.  $(Q + R) + S = Q + (R + S)$
4.  $(QR)S = Q(RS)$
5.  $Q\varepsilon = \varepsilon Q = Q + \varepsilon = \varepsilon + Q = Q$
6.  $Q\emptyset = \emptyset Q = \emptyset$
7.  $(R + S)Q = RQ + SQ$
8.  $Q(R + S) = QR + QS$
9.  $(Q^*)^* = Q^*$

10.  $QQ^* = Q^*Q = Q^*$
11.  $\varepsilon^* = \varepsilon, \emptyset^* = \varepsilon$
12.  $Q^* = \varepsilon + Q + Q^2 + \dots + Q^k Q^*, k \geq 1,$
13.  $(Q^* + R^*)^* = (Q^* R^*)^* = (Q^* R)^* Q^* = (Q + R)^*$
14. Ако  $Q = R^* S$ , то  $Q = RQ + S$ .

Ще докажем последното свойство 14, чрез следните равенства, в които са използвани свойствата 2, 12, 5, 7.

$$Q = R^* S = (RR^* + \varepsilon)S = RR^* S + S = RQ + S$$

Като вземем предвид, че  $\emptyset, \{\varepsilon\}$  и  $\{a\}$ , за всяка  $a \in V$  са автоматни езици следва, че всеки регулярен израз  $R$  над  $V^*$ , представя множество  $L$ , получено с операциите обединение, произведение и итерация на тези езици и следователно  $L$  е автоматен език. От начина, по който дефинирахме регулярните изрази следва, че всеки регулярен израз над  $V^*$  определя еднозначно автоматен език, представен от този израз.

Не по-малко интересен е и въпросът за това дали всеки автоматен език  $L$  се представя с някои регулярен израз. Ясно е, че ако отговорът е утвърдителен, то съответният регулярен израз не е единствен. Например изразите  $(a^* + b^*)^*$  и  $(a^* b^*)^*$  представят един и същи език  $\{a, b\}^*$ . Следващата теорема дава отговор на поставения по-горе въпрос и е известна като Теорема на Клини. Важността ѝ се състои в това, че чрез нея можем да разглеждаме автоматните езици като регулярни изрази над дадена азбука.

**Теорема 2.9.1 (Клини)** *Всеки автоматен език над дадена азбука  $V$  се представя с регулярен израз над  $V^*$ .*

**Доказателство:** Нека  $L$  е произволен автоматен език и  $A = \langle K, V, \delta, q_0, F \rangle$  е ДКА, който разпознава езика  $L$ , т.е.  $L = L(A)$ . Да предположим, че  $K = \{q_0, q_1, \dots, q_n\}$  и  $F = \{q_{t_1}, q_{t_2}, \dots, q_{t_m}\}$ . Тогава  $L = \bigcup_{i=1}^m L_i$ , където  $L_i = L(A_i)$  и  $A_i = \langle K, V, \delta, q_0, \{q_{t_i}\} \rangle$  за  $i = 1, 2, \dots, m$ .

Понеже на операцията обединение на езици, съответствува операцията събиране (+) на регулярни изрази, то достатъчно е да покажем, че всеки от езиците  $L_i$  се представя с регулярен израз над  $V^*$ . Без ограничение на общността (тъй като винаги можем да пренумерираме състоянията в  $K$ ) да докажем, че автоматът  $A_n = \langle K, V, \delta, q_0, \{q_n\} \rangle$  разпознава език, който се представя с регулярен израз.

Да означим с  $L_{ij}^l$  за  $0 \leq i \leq j \leq n$ ,  $0 \leq l \leq n$  езикът, който се състои от думите  $\omega$  такива, че  $\delta(q_i, \omega) = q_j$  и за всеки собствен префикс  $y$  на  $\omega$  е в сила  $\delta(q_i, y) \in \{q_p | p \leq l\}$ . С други думи  $\omega \in L_{ij}^l$ , ако  $\omega$  привежда автомата  $A_n$  от състояние  $q_i$  в състояние  $q_j$  чрез път, който не съдържа вътрешно състояние  $q_p$ , чийто индекс  $p$  надвишава  $l$ . Ще докажем с индукция по  $l$ , че  $L_{ij}^l$  е език, които се представя с регулярен израз и понеже  $L(A_n) = L_{0n}^n$  от това следва, верността на теоремата.

Очевидно, ако  $i \neq j$ , то  $L_{ij}^0$  се представя от регулярния израз  $a_1 + a_2 + \dots + a_z$ , където  $a_1, a_2, \dots, a_z$  са букви от  $V$ , които са тегла на ребрата, свързващи  $q_i$  с  $q_j$ . Ако такива ребра няма, то на  $L_{ij}^0$  отговаря изразът  $\emptyset$ . Ако  $i = j$ , то на  $L_{ij}^0$  съответствува изразът  $\varepsilon + a_1 + a_2 + \dots + a_s$ , като тук  $a_1, a_2, \dots, a_s$  са тегла на ребрата, свързващи  $q_i$  със себе си. Очевидното равенство

$$L_{ij}^l = L_{ij}^{l-1} + L_{il}^{l-1} (L_{ll}^{l-1})^* L_{lj}^{l-1}$$

и принципът на математическата индукция показват, че  $L_{ij}^l$  се представя с регулярен израз над  $V^*$  за всяко  $l = 0, 1, \dots, n$ . Следователно  $L(A_n)$  и  $L(A)$  са езици, които могат да бъдат представени от регулярни изрази.  $\square$

### З а д а ч и

1. Да се определи ДКА, които разпознава езика, представен от следните регулярни изрази:

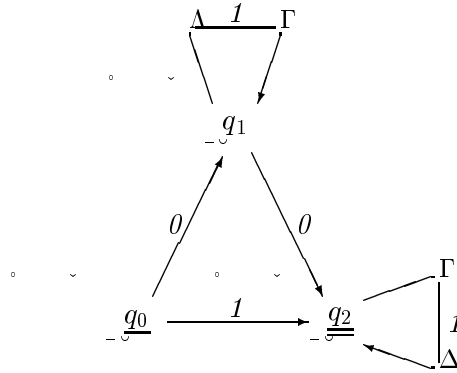
$$\text{а) } a(ba + b)^* + a; \quad \text{б) } (ab + a^*)^* ba; \quad \text{в) } ((b^* a)^*) a^* + b^*.$$

2. Да се докаже, че ако  $R_1, R_2, \dots, R_n$  са регулярни изрази и  $f(R_1, R_2, \dots, R_n)$  е полином, то:

$$\text{а) } f(R_1, R_2, \dots, R_n) + (R_1 + R_2 + \dots + R_n)^* = (R_1 + R_2 + \dots + R_n)^*;$$

$$б) f(R_1^*, R_2^*, \dots, R_n^*)^* = (R_1 + R_2 + \dots + R_n)^*.$$

3. Да се намери регулярен израз, който представя езикът  $L(A)$ , където  $A$  е ДКА зададен, чрез следния граф:



4. Да се построи минимален ДКА, който разпознава езика, представен с регулярния израз:

а)  $(ab)^* + a(ba + a)^*$ ;

б)  $(a + b)^* ca^*$ ;

в)  $(01)^* + (1 + 01^*)0^*$ ;

5. Да се докаже, че следният език е автоматен :

$$L = \{\omega \in \{0, 1\}^* \mid \omega \text{ е двоичното представяне на числата от вида } n = 2^m(2^k - 1) + 1, \text{ } m + k - \text{четно}\}.$$

**У п ъ т в а н е:** Докажете, че  $L = L_{1o}L_{0e}1 \cup L_{1e}L_{0o}1$ , където

$$L_{1o} = \{1^k \mid k - \text{нечетно}\},$$

$$L_{1e} = \{1^k \mid k - \text{четно}\},$$

$$L_{0o} = \{0^k \mid k - \text{нечетно}\} \quad \text{и}$$

$$L_{0e} = \{0^k \mid k - \text{четно}\}.$$

6\*. Да се докаже, че следният език не е автоматен:

$$Q = \{\omega \in \{0, 1\}^* \mid \omega \text{ е двоичното представяне на } n^2, \text{ } n - \text{естествено число}\}.$$

**У п ъ т в а н е:** Разгледайте езика

$$M = \{\omega \in \{0, 1\}^* \mid \omega \text{ е двоичното представяне на числата}$$

от вида  $(2^{k+1} - 1)^2$ ,  $k$  – естествено число}.

Понеже  $(2^{k+1} - 1)^2 = (2^k - 1)2^{k+2} + 1$  следва, че  $M \subset L$ , където  $L$  е езикът от предишната задача. Докажете, че  $1M = T1$ , където  $T = \{1^n 0^n \mid n \geq 1\}$ . Тъй като  $1M$  не е автаматен език (защо?) следва, че  $M$  също не е автоматен. Докажете, че  $M = L \cap Q$ .