Упражнение 5

1 Регулярни изрази и теорема на Клини

Регулярният израз е още едно средство за описване на формалните езици над дадена азбука Σ. Грубо казано, регулярните изрази използват само буквите от азбуката и *символите* $\oslash$, $\boldsymbol{\epsilon}$, $+$, *, $($ и $)$, за да дадат представяне на езиците над Σ. Формалната дефиниция за *синтаксиса* на регулярните изрази е следната.

Дефиниция 1. Нека е дадена азбука Σ. **Регулярен израз** над Σ дефинираме индуктивно, както следва.

База: $\oslash$ е регулярен израз, $\boldsymbol{\epsilon}$ е регулярен израз и всеки елемент на Σ е регулярен израз.

Стъпка: (1) Ако α и β са регулярни изрази, то и $(\alpha\beta)$ е регулярен израз.
(2) Ако α и β са регулярни изрази, то и $(\alpha + \beta)$ е регулярен израз.
(3) Ако α е регулярен израз, то и $\alpha^\star$ е регулярен израз.

Всеки регулярен израз съответства на един и точно един език над Σ. Формално, релацията между регулярните изрази и езиците, които те представят, се описва чрез функция $\mathscr{L}$, такава че ако α е регулярен израз, то $\mathscr{L}(\alpha)$ е езикът, който α представя. Функцията $\mathscr{L}$ ще наричаме **семантика**, а стойността ѝ в α — **семантика на $\boldsymbol{\alpha}$**.

За да дефинираме формално $\mathscr{L}$ ще трябва да кажем еднозначно какво тя връща за всеки даден регулярен израз α над Σ. Това естествено ще сторим с индукция по конструкцията на α.

База: $\mathscr{L}(\oslash) = \varnothing$, $\mathscr{L}(\boldsymbol{\epsilon}) = \{\epsilon\}$ и $\mathscr{L}(a) = \{a\}$ за всяко $a \in \Sigma$.

Стъпка: (1) Ако α и β са регулярни изрази, то $\mathscr{L}((\alpha\beta)) = \mathscr{L}(\alpha)\mathscr{L}(\beta)$.
(2) Ако α и β са регулярни изрази, то $\mathscr{L}((\alpha + \beta)) = \mathscr{L}(\alpha) \cup \mathscr{L}(\beta)$.
(3) Ако α е регулярен израз, то $\mathscr{L}(\alpha^\star) = \mathscr{L}(\alpha)^*$.

Пример 1. Да намерим $\mathscr{L}(((a + b)^* a))$. Имаме следното.

$$
\begin{aligned}
\mathscr{L}(((a+b)^\star a)) &= \mathscr{L}((a+b)^\star)\mathscr{L}(a) && // \text{ съгласно } (2) \\
&= \mathscr{L}((a+b)^\star)\{a\} && // \text{ съгласно } (1) \\
&= \mathscr{L}((a+b))^*\{a\} && // \text{ съгласно } (4) \\
&= (\mathscr{L}(a) \cup \mathscr{L}(b))^*\{a\} && // \text{ съгласно } (3) \\
&= (\{a\} \cup \{b\})^*\{a\} && // \text{ съгласно } (1) \text{ два пъти} \\
&= (\{a, b\})^*\{a\} \\
&= \{w \in \Sigma^* \mid w \text{ завършва на } a\}
\end{aligned}
$$

Класът на **регулярните езици** над дадена азбука Σ се дефинира да бъде множеството от всички езици L над Σ, такива че $L = \mathscr{L}(\alpha)$ за някой регулярен израз α над Σ. Алтернативно, можем да мислим за класа на регулярните езици като *затварянето* на множеството от езици

$$\{\{\sigma\} \mid \sigma \in \Sigma\} \cup \{\{\epsilon\}\} \cup \{\varnothing\}$$

относно операциите обединение, конкатенация и звезда на Клини.

Основен резултат в теорията на формалните езици е следната теорема за еквивалентност.

Теорема 1(на Клини). Един език L е регулярен тогава и само тогава, когато е автоматен.

Едната посока на доказателството на тази теорема вече на практика сме направили, а именно правата; трябва само да дадем автомати за базовите регулярно езици — $\varnothing$, $\{\epsilon\}$ и $\{a\}$ за $a \in \Sigma$. От там нататък затвореността на автоматните езици относно операциите обединение, конкатенация и звезда на Клини ни дават именно, че регулярността на един език влече неговата автоматност.

Обратната посока ще бъде от интерес за нас за целите на това упражнение. Как от детерминиран автомат $A = (Q, \Sigma, \delta, s, F)$ да получим регулярен израз α, такъв че $L(A) = \mathscr{L}(\alpha)$? За начало, да номерираме състоянията на автомата; нека $Q = \{q_1, ..., q_n\}$, като $s = q_1$. За $i, j = 1, ..., n$ и $k = 0, ..., n$, дефинираме $R(i, j, k)$ да бъде множеството от всички думи в Σ^*, с които A може да направи преход от състояние q_i до състояние q_j без да премине през *вътрешни* състояния с индекси по-големи от k — крайните състояния за този преход q_i и q_j могат да имат индекси по-големи от k. За $k = n$, имаме че

$$R(i, j, n) = \{w \in \Sigma^* \mid \hat{\delta}(q_i, w) = q_j\}.$$

Следователно

$$L(A) = \bigcup\{R(1, j, n) \mid q_j \in F\}.$$

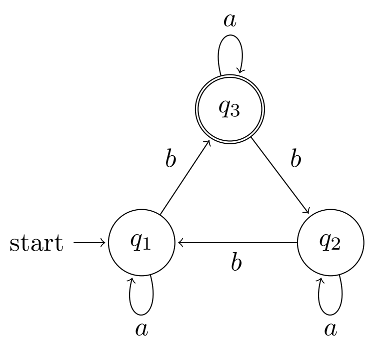
За целите на доказателството следва да се покаже, че езиците $R(i, j, k)$ по всевъзможните i, j и k са регулярни, и следователно такъв е и $L(A)$. Доказателството на този факт се извършва с индукция по k. За нас съществено е изразяването на езика $R(i, j, k)$ чрез езици с по-малка стойност на параметъра k, а именно

$$R(i, j, k) = R(i, j, k-1) \cup R(i, k, k-1)R(k, k, k-1)^*R(k, j, k-1).$$

Това уравнение просто ни казва, че за да стигне от състояние q_i до състояние q_j без да минава през състояние с номер по-голям от k, автоматът А може да стори едно от следните две неща

(1) да отиде от състояние q_i до състояние q_j, минавайки през вътрешни състояния с номера, по-малки от $k-1$; или
(2) да премине първо от q_i до q_k; след което да премине от q_k до q_k нула или повече пъти; след което да премине от q_k до q_j; като през цялото време не минава през *вътрешни* състояния с индекси по-големи от $k-1$.

Пример 2. Да построим регулярен израз за езика $L = \{w \in \Sigma^* \mid w$ има $3k+1$ b-та за някое $k \in \mathbb{N}\}$, разпонат от следния автомат.



Ако се опитаме подробно да проследим конструкцията от доказателството на обратната посока на **Теорема 1** ще трябва да построим всички езици $R(i,j,k)$ по всевъзможните стойностти за i, j и k, които са $3*3*4 = 36$ на брой. Ясно е, че $R(i,j,0) = \{\sigma \in \Sigma \mid \delta(q_i, \sigma) = q_j\}$ за $i, j = 1, 2, 3$.

Нататък за $k = 1$ имаме $R(1,1,1) = \mathscr{L}(a^\star)$,
$R(1,2,1) = \varnothing$, $R(1,3,1) = \mathscr{L}(a^\star b)$, $R(2,1,1) = R(2,1,0)$,
$R(2,2,1) = R(2,2,0)$, $R(2,3,1) = \mathscr{L}(ba^\star b)$, а $R(3,j,1) = R(3,j,0)$,
за $j = 1, 2, 3$.

След това за $k = 2$ имаме $R(1,j,2) = R(1,j,1)$, за $j = 1, 2, 3$,
$R(2,1,2) = \mathscr{L}(a^\star ba^\star)$, $R(2,2,2) = \mathscr{L}(a^\star)$, $R(2,3,2) = \mathscr{L}(a^\star ba^\star b)$,
$R(3,1,2) = \mathscr{L}(ba^\star ba^\star)$, $R(3,2,2) = \mathscr{L}(ba^\star)$, $R(3,3,2) = \mathscr{L}(a + ba^\star ba^\star b)$.

Накрая за $k = 3$ е достатъчно да намерим $R(1,3,3)$, защото съгласно **Теорема 1**, $L(A) = R(1,3,3)$. Имаме

$$\begin{gathered} R(1,3,3) = R(1,3,2) \cup R(1,3,2)R(3,3,2)^* R(3,3,2) = \\ \mathscr{L}(a^\star b) \cup \mathscr{L}(a^\star b)\mathscr{L}(ba^\star ba^\star b)^* \mathscr{L}(ba^\star ba^\star b) = \\ \mathscr{L}(a^\star b + a^\star b(a + ba^\star ba^\star b)^\star (a + ba^\star ba^\star b)) = \\ \mathscr{L}(a^\star b(\boldsymbol{\epsilon} + (a + ba^\star ba^\star b)^\star (a + ba^\star ba^\star b)) = \mathscr{L}(a^\star b(a + ba^\star ba^\star b)^\star). \end{gathered}$$

2 Задачи

3 Решения