Регулярные выражения

Определение 1. Регулярное выражение над алфавитом Σ определяется рекурсивно следующим образом: 0 является регулярным выражением; 1 является регулярным выражением; если $a \in \Sigma$, то а является регулярным выражением; если e и f являются регулярными выражениями, то $(e + f), (e \cdot f)$ и e^* тоже являются регулярными выражениями. Для экономии скобок будем считать, что операция $*$ связывает сильнее (то есть имеет более высокий приоритет), чем умножение, а умножение связывает сильнее, чем сложение. Вместо $e \cdot f$ часто пишут просто ef.

Определение 2. Каждое регулярное выражение е над алфавитом Σ задаёт (denotes, represents) некоторый язык над алфавитом Σ (обозначение $L(e)$), определяемое рекурсивно следующим образом:

$L(a) \leftrightharpoons \{a\}$, если $a \in \Sigma$,
$L(0) \leftrightharpoons \emptyset$,
$L(1) \leftrightharpoons \{\varepsilon\}$,
$L(e + f) \leftrightharpoons L(e) \cup L(f)$,
$L(e \cdot f) \leftrightharpoons L(e) \cdot L(f)$,
$L(e^*) \leftrightharpoons L(e)^*$.

Заметим, что в правой части последнего выражения символом $*$ обозначена итерация языка. Вместо $L(e)$ часто пишут просто e.

Пример. Пусть $\Sigma = \{a, b\}$. Регулярное выражение $(ab)^* \cdot (1 + a)$ задаёт язык $\{(ab)^n \mid n \geq 0\} \cup \{(ab)^n\, a \mid n \geq 0\}$.

Определение 4. Язык L называется регулярным, если он задаётся некоторым регулярным выражением.

Определение 5. Пусть e — регулярное выражение. Тогда $e^+ \leftrightharpoons e^*e$.

Свойства регулярных выражений

Лемма. Для любых регулярных выражений e, f и g выполняются следующие тождества:

1. $e + f = f + e$;
2. $e + 0 = e$;
3. $(e + f) + g = e + (f + g)$;
4. $e \cdot 1 = e$;
5. $1 \cdot e = e$;
6. $(e \cdot f) \cdot g = e \cdot (f \cdot g)$;
7. $e \cdot (f + g) = e \cdot f + e \cdot g$;
8. $(f + g) \cdot e = f \cdot e + g \cdot e$;
9. $e \cdot 0 = 0$;
10. $0 \cdot e = 0$;
11. $e + e = e$;
12. $(1 + e + ee + \cdots + e^{n-1})(e^n)^* = e^*$ для любого $n \geq 1$;
13. $(e^*f)^*e^* = (e + f)^*$;
14. $1 + e(fe)^*f = (ef)^*$.

Равенство понимается как равенство языков, задаваемых регулярными выражениями.

Пример. Чтобы упростить регулярное выражение$((a^*b)^*(bc+ca))^*(a^*b)^*$, применим пункт 13 леммы, подставив вместо e выражение $(a*b)$, а вместо f выражение $bc+ca$. Получаем, что исходное выражение равно регулярному выражению $(a^*b+bc+ca)^*$.

Теорема Клини

Определение. Назовём обобщённым конечным автоматом аналог конечного автомата, где переходы помечены не словами, а регулярными выражениями. Метка пути такого автомата — произведение регулярных выражений на переходах данного пути. Слово w допускается обобщённым конечным автоматом, если оно принадлежит языку, задаваемому меткой некоторого успешного пути.

Замечание. Каждый конечный автомат можно преобразовать в обобщённый конечный автомат, допускающий те же слова. Для этого достаточно заменить всюду в метках переходов пустое слово на 1, а каждое непустое слово — на произведение его букв.

Теорема Клини. Язык L является регулярным тогда и только тогда, когда он является автоматным.

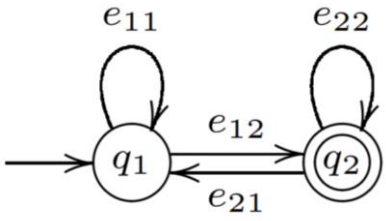
Доказательство. Пусть e — регулярное выражение. Индукцией по построению e легко показать, что задаваемый им язык является автоматным.

Обратно, пусть язык L распознаётся некоторым конечным автоматом с одним начальным состоянием и одним заключительным состоянием. Существует эквивалентный ему обобщённый конечный автомат $\langle Q,\Sigma,\Delta,\{q_1\},\{q_2\}\rangle$, где $q_1 \neq q_2$. Если есть несколько переходов с общим началом и общим концом (такие переходы называются параллельными), заменим их на один переход, используя операцию $+$.

Устраним по очереди все состояния, кроме q_1 и q_2. При устранении состояния q нужно для каждого перехода вида $\langle p_1, f_1, q\rangle$, где $p_1 \neq q$, и для каждого перехода вида $\langle q, f_2, p_2\rangle$, где $p_2 \neq q$, добавить переход $\langle p_1, f_1 g^* f_2, p_2\rangle$, где регулярное выражение g — метка перехода из q в q (если нет перехода из q в q, то надо добавить переход $\langle p_1, f_1 f_2, p_2\rangle$), и снова всюду заменить параллельные переходы на один переход, используя операцию $+$.

После устранения всех состояний, кроме q_1 и q_2, получится обобщённый конечный автомат $\langle\{q_1,q_2\},\Sigma,\Delta',\{q_1\},\{q_2\}\rangle$, где

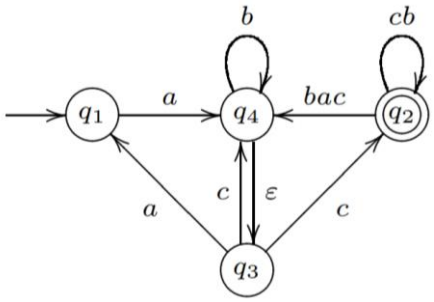
$$\Delta' = \{\langle q_1, e_{11}, q_1\rangle, \langle q_1, e_{12}, q_2\rangle, \langle q_2, e_{21}, q_1\rangle, \langle q_2, e_{22}, q_2\rangle\}.$$



Очевидно, что $L = L(e_{11}^* e_{12}(e_{22} + e_{21}e_{11}^* e_{12})^*)$.

Пример. Рассмотрим язык, распознаваемый конечным автоматом $M = \langle\{q_1, q_2, q_3, q_4\}, \Sigma, \Delta, \{q_1\}, \{q_2\}\rangle$, где $\Sigma = \{a, b, c\}$ и

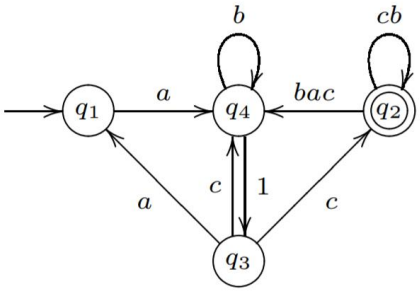
$\Delta = \{\langle q_1, a, q_4\rangle, \langle q_2, cb, q_2\rangle, \langle q_2, bac, q_4\rangle, \langle q_3, a, q_1\rangle, \langle q_3, c, q_2\rangle, \langle q_3, c, q_4\rangle, \langle q_4, \varepsilon, q_3\rangle, \langle q_4, b, q_4\rangle\}$.



Тот же язык порождается обобщённым конечным автоматом

$M_1 = \langle\{q_1, q_2, q_3, q_4\}, \Sigma, \Delta_1, \{q_1\}, \{q_2\}\rangle$, где

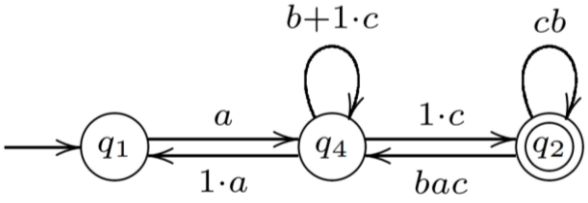
$\Delta_1 = \{\langle q_1, a, q_4\rangle, \langle q_2, cb, q_2\rangle, \langle q_2, bac, q_4\rangle, \langle q_3, a, q_1\rangle, \langle q_3, c, q_2\rangle, \langle q_3, c, q_4\rangle, \langle q_4, 1, q_3\rangle, \langle q_4, b, q_4\rangle\}$.



После устранения состояния q_3 получается обобщённый конечный автомат

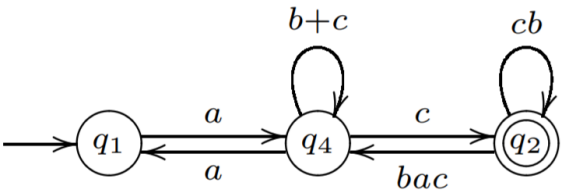
$M_2 = \langle\{q_1, q_2, q_4\}, \Sigma, \Delta_2, \{q_1\}, \{q_2\}\rangle$, где

$\Delta_2 = \{\langle q_1, a, q_4\rangle, \langle q_2, cb, q_2\rangle, \langle q_2, bac, q_4\rangle, \langle q_4, 1 \cdot a, q_1\rangle, \langle q_4, 1 \cdot c, q_2\rangle, \langle q_4, b + 1 \cdot \mathrm{c}, q_4\rangle\}$.



Можно упростить регулярные выражения и получить

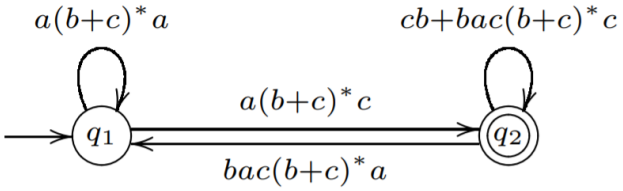
$\Delta_2' = \{\langle q_1, a, q_4\rangle, \langle q_2, cb, q_2\rangle, \langle q_2, bac, q_4\rangle, \langle q_4, a, q_1\rangle, \langle q_4, c, q_2\rangle, \langle q_4, b + \mathrm{c}, q_4\rangle\}$.



После устранения состояния q_4 и упрощения регулярных выражений получается обобщённый конечный автомат

$M_3 = \langle\{q_1, q_2\}, \Sigma, \Delta_3, \{q_1\}, \{q_2\}\rangle$, где

$\Delta_3 = \{\langle q_1, a(b+c)^*a, q_1\rangle, \langle q_1, a(b+c)^*c, q_2\rangle, \langle q_2, bac(b+c)^*a, q_1\rangle, \langle q_2, cb + bac(b+c)^*c, q_2\rangle\}$.



Следовательно, язык $L(M)$ задаётся регулярным выражением

$(a(b+c)^*a)^*a(b+c)^*c \cdot (cb + bac(b+c)^*c + bac(b+c)^*a(a(b+c)^*a)^*a(b+c)^*c)^*$.

Звёздная высота

Определение. Звёздная высота (star-height) регулярного выражения (обозначение $sh(e)$) определяется рекурсивно следующим образом:

$$sh(a) \leftrightharpoons 0,$$

$$sh(0) \leftrightharpoons 0,$$

$$sh(1) \leftrightharpoons 1,$$

$$sh(e+f) \leftrightharpoons \max\big(sh(e), sh(f)\big),$$

$$sh(e \cdot f) \leftrightharpoons \max\big(sh(e), sh(f)\big),$$

$$sh(e^*) \leftrightharpoons 1 + sh(e).$$

Пример. Пусть $\Sigma = \{a, b, c\}$. Тогда $sh((a^* + b^* + ab)^* + (ab^*c)^*) = 2$.

Определение. Звёздной высотой регулярного языка L (обозначение $sh(L)$) называется минимум звёздных высот регулярных выражений, задающих этот язык.

Пример. Пусть $\Sigma = \{a, b\}$ и $L = \{w \in \Sigma^* | (|w|_a - |w|_b) \vdots 4\}$. Тогда $sh(L) = 2$. Действительно, язык L задаётся регулярным выражением $(ab + ba + (aa + bb)(ab + ba)^*(aa + bb))^*$ и не задаётся никаким регулярным выражением меньшей звёздной высоты.