Глава 2. Лексический анализ

2.7. Регулярные выражения

Альтернативой регулярным грамматикам для формального описания регулярных языков являются регулярные выражения. Регулярные выражения эквивалентны регулярным грамматикам и широко используются на практике.

Регулярное выражение r над алфавитом T описывает язык $L(r)$, который рекурсивно определяется на основании языков, описываемых подвыражениями r.

Базисные регулярные выражения образованы тремя правилами:

1. Символ $\varnothing$, представляющий пустое множество, является регулярным выражением, а $L(\varnothing)$ представляет собой пустой язык.

2. Символ пустой строки ε является регулярным выражением, а $L(\varepsilon)$ представляет собой множество $\{\varepsilon\}$, т. е. язык, единственный член которого – пустая строка.

3. Если $a \in T$, то a представляет собой регулярное выражение, а $L(a)$ представляет множество $\{a\}$, т. е. язык с одной строкой единичной длины с символом a.

Имеют место правила, посредством которых регулярные выражения строятся из подвыражений. Пусть r и s являются регулярными выражениями, описывающими соответственно языки $L(r)$ и $L(s)$.

1. $r \mid s$ (объединение) – регулярное выражение, описывающее язык $L(r) \cup L(s)$. Вместо символа $'|'$ можно использовать символ $'+'$, т. е. записи $r \mid s$ и $r + s$ эквивалентны.

2. rs (конкатенация) – регулярное выражение, описывающее язык $L(r)L(s)$.

3. r^* (итерация, т. е. нуль или более повторений r) – регулярное выражение, описывающее язык $(L(r))^*$.

Все операции левоассоциативны. Подразумевается следующая система приоритетов: унарная операция итерации обладает наивысшим приоритетом, за ним следует операция конкатенации, а затем следует операция объединения. Приоритеты можно изменять с помощью использования скобок.

Одним из расширений регулярных выражений является унарная операция r^+ (один или более повторений r). Эта операция имеет тот же приоритет и ассоциативность, что и операция итерации. Имеют место алгебраические законы: $r^* = r^+ \mid \varepsilon$ и $r^+ = rr^* = r^* r$.

Регулярное выражение генерирует регулярное множество. Например, регулярное выражение $(a \mid b)c$ генерирует регулярное множество (регулярный язык) $\{ac, bc\}$, а регулярное выражение $(aab \mid ab)^*$ – множество $\{aab, ab\}^*$, включающее строки

ε
aabaabab
ababaab
abaababababaab и т. п.

Регулярное выражение, описывающее классическое определение идентификатора, имеет вид $l(l \mid d)^*$, где l и d обозначают соответственно букву и цифру. Если в определение идентификатора добавить маркер конца ввода лексемы $\perp \notin \{l, d\}$, то есть любой символ кроме буквы и цифры, то регулярное выражение для него примет вид $l(l \mid d)^* \perp$.

Существует алгебра регулярных выражений, позволяющая выполнять эквивалентные преобразования выражений. Основные алгебраические свойства регулярных выражений (q, r, s – некоторые регулярные выражения) [14]:

1) $q \mid q = q$, $q \mid \varnothing = q$
2) $q \mid r = r \mid q$, свойство коммутативности
3) $(q \mid r) \mid s = q \mid (r \mid s) = q \mid r \mid s$ – ассоциативность
4) $(qr)s = q(rs) = qrs$ – ассоциативна, но не коммутативна
5) $q\varepsilon = \varepsilon q = q$, $q\varnothing = \varnothing q = \varnothing$
6) $(r \mid s)q = rq \mid sq$
7) $q(r \mid s) = qr \mid qs$
8) $q^* q^* = q^*$
9) $(q^*)^* = q^*$
10) $qq^* = q^* q$
11) $\varepsilon^* = \varepsilon$, $\varnothing^* = \varepsilon$
12) $(q^* \mid r^*)^* = (q^* r^*)^* = (q \mid r)^*$
13) $(rq)^* r = r(qr)^*$
14) $(q^* r)^* q^* = (q \mid r)^*$
15) $(q^* r)^* = (q \mid r)^* r \mid \varepsilon$
16) $(qr^*)^* = q(q \mid r)^* \mid \varepsilon$

Для любого регулярного выражения можно определить недетерминированный конечный автомат, который принимает регулярный язык, соответствующий заданному регулярному выражению. Как и для регулярных грамматик, существует процедура построения конечного автомата-распознавателя по заданному регулярному выражению.

Конечные автоматы для базисных регулярных выражений ∅, ε, *a* представлены на рис. 2.11, где k_0 и k_f – начальное и конечное состояния автомата соответственно.

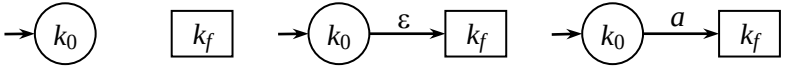


Рис. 2.11. Конечные автоматы для базисных регулярных выражений:
а – ∅; *б* – ε; *в* – *a*

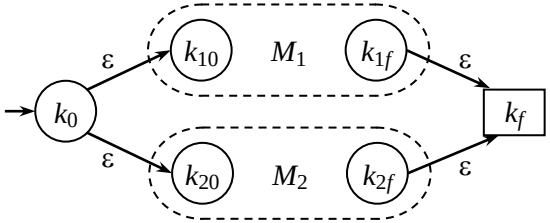
Обратите внимание, что в графах переходов автоматов могут появиться переходы, помеченные символом ε (пустая строка). Такие переходы называются ε-*переходами*. Таким образом, функция переходов δ определена на множестве $K \times (T \cup \{\varepsilon\})$, т. е. $\delta: K \times (T \cup \{\varepsilon\}) \to 2^K$.

Рассмотрим построение автоматов для операций над регулярными выражениями. Пусть M_1 и M_2 – конечные автоматы, распознающие языки, представленные регулярными выражениями r и s соответственно, причем их множества состояний не пересекаются. Обозначим через k_{10} и k_{20} начальные состояния этих автоматов, через k_{1f} и k_{2f} – их конечные состояния. Тогда конечный автомат M с начальным состоянием k_0 и конечным состоянием k_f, который представляет регулярное выражение q как результат регулярной операции над r и s строится следующим образом (рис. 2.12).

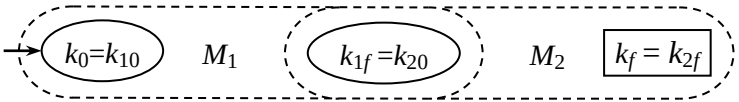
1. $q = r \mid s$ (или $r + s$). Автомат M строится параллельным соединением автоматов M_1 и M_2 (рис. 2.12, *а*). Добавляются новые состояния k_0 и k_f, добавляются ε-переходы из k_0 в k_{10} и k_{20}, а также из k_{1f} и k_{2f} в k_f. Любой путь из k_0 в k_f должен пройти либо исключительно через M_1, либо исключительно через M_2.

2. $q = rs$. Автомат M строится последовательным соединением автоматов M_1 и M_2 (рис. 2.12, *б*). Начальным состоянием M объявляется k_{10}, конечным состоянием – k_{2f}. Состояния k_{1f} и k_{20} объединяются в одно состояние со всеми входящими и исходящими переходами обоих состояний. Путь из k_0 в k_f должен пройти сначала через M_1, а затем через M_2.

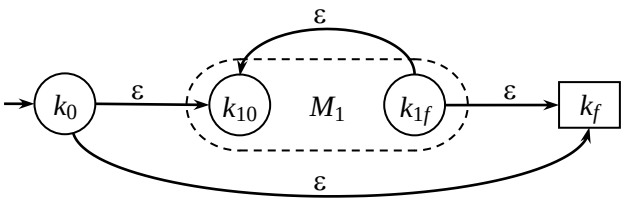
3. $q = r^*$. Автомат M строится зацикливанием автомата M_1 (рис. 2.12, *в*). Добавляются новые состояния k_0 и k_f, добавляются также ε-переходы из k_0 в k_{10} и k_f, из k_{1f} в k_{10}, из k_{1f} в k_f. Для достижения k_f из k_0 необходимо пройти либо по ε-переходу от k_0 к k_f, соответствующей пустой строке, либо перейти к начальному состоянию k_{10} автомата M_1, пройти его и вернуться в k_{10} нуль или более раз.



а



б



в

Рис. 2.12. Графы переходов автоматов для регулярных выражений:
а – $r \mid s$; *б* – rs; *в* – r^*

Заметим, что в этом методе на любом шаге построения начальные состояния не имеют входящих дуг, а конечные состояния – исходящих.

Наличие ε-перехода вносит недетерминированность в функционирование конечного автомата, поскольку автомат может переходить из состояния в состояние без чтения входного символа. Подробнее о НКА с ε-переходами изложено в разделе 2.8.

В качестве примера построим конечный автомат для регулярного выражения $(a \mid b^*)b$. Автоматы для *a*, *b* и *c* изображены на рис. 2.11, *в*. С помощью конструкции на рис. 2.12, *в* построим автомат для b^*, как показано на рис. 2.13, *а*. Затем с помощью конструкции на рис. 2.12, *а* построим автомат для $a \mid b^*$, как показано на рис. 2.13, *б*. Завершаем построение автомата с помощью конструкции на рис. 2.12, *б* для $(a \mid b^*)b$, получив автомат, изображенный на рис. 2.13, *в*.

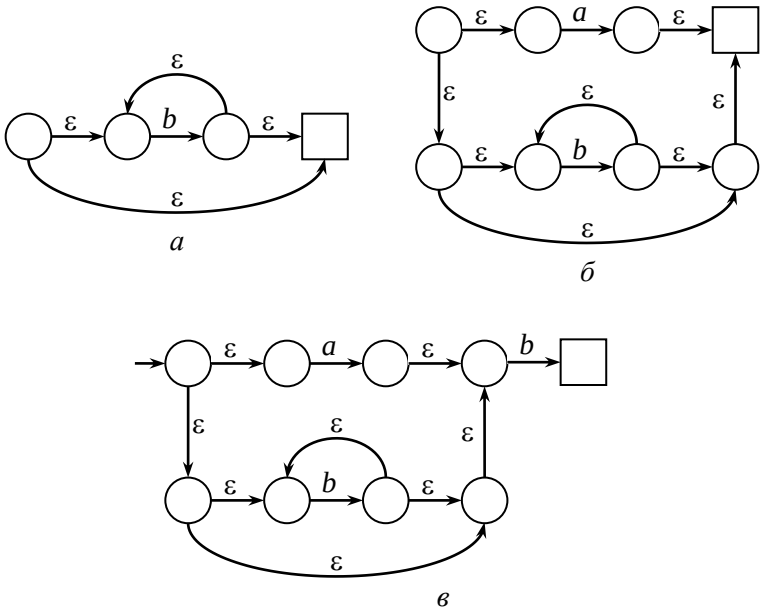


Рис. 2.13. Построение автомата для выражения $(a \mid b^*)b$:
а – для b^*; *б* – для $a \mid b^*$; *в* – для $(a \mid b^*)b$