

## Глава 2. Лексический анализ

### 2.8. Конечные автоматы с $\epsilon$ -переходами

Недетерминированный конечный автомат  $M = (K, T, \delta, k_0, F)$  имеет  $\epsilon$ -переходы, если функция  $\delta$  определена на множестве  $K \times (T \cup \{\epsilon\})$ , т. е.  $\delta: K \times (T \cup \{\epsilon\}) \rightarrow 2^K$ , где  $2^K$  – булеан множества  $K$  (множество всех подмножеств множества  $K$ , мощность булеана равна  $2^{|K|}$ ).

Пример такого автомата представлен на рис. 2.14.

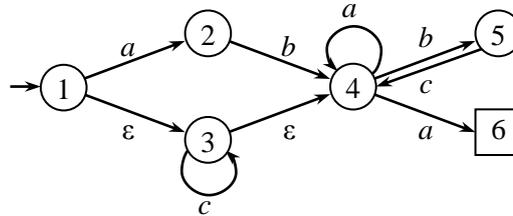


Рис. 2.14. Конечный автомат с  $\epsilon$ -переходами

Наличие  $\epsilon$ -перехода вносит недетерминированность в функционирование конечного автомата, поскольку автомат может переходить из состояния в состояние без чтения входного символа.

Пусть состояния  $k, k' \in K$  такие, что  $k' \in \delta(k, \varepsilon)$ , тогда говорят, что состояние  $k'$   *$\varepsilon$ -достижимо* из состояния  $k$  и записывают  $k \xrightarrow{\varepsilon} k'$ . Таким образом, автомат может в любой момент перейти из состояния  $k$  в состояние  $k'$  без чтения входного символа. Рассмотрим более широкое понятие  $\varepsilon$ -достижимости, а именно  $k \xrightarrow{\varepsilon^*} k'$ , т. е.  $k'$   $\varepsilon$ -достижимо из  $k$ , если существует путь по графу автомата от вершины  $k$  до  $k'$ , состоящий из дуг, помеченных символом  $\varepsilon$  и длиной  $\geq 0$ .

Обозначим через  $R(k)$  множество всех состояний, которые  $\varepsilon$ -достижимы из состояния  $k$ , что формально записывается как

$$R(k) = \{ k' \mid k \xrightarrow{\varepsilon^*} k' \}.$$

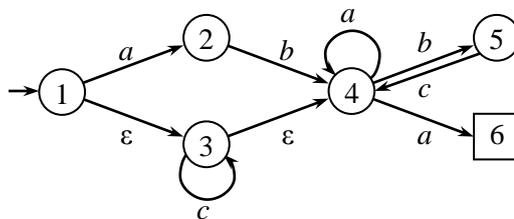
Следует обратить внимание на то, что  $k \in R(k)$ .

Расширим это определение для подмножества  $M \subseteq K$  состояний следующим образом:  $R(M) = \bigcup_{k \in M} R(k)$ .

Для автомата на рис. 2.14 имеем

$$\begin{aligned} R(1) &= \{1, 3, 4\}, & R(4) &= \{4\}, \\ R(2) &= \{2\}, & R(5) &= \{5\}, \\ R(3) &= \{3, 4\}, & R(6) &= \{6\}, \\ R(\{1, 3\}) &= R(1) \cup R(3) = \{1, 3, 4\}. \end{aligned}$$

Рассмотрим, как влияет на функционирование автомата наличие  $\varepsilon$ -переходов. Для состояния 1 имеем  $\delta(1, a) = \{2\}$ , т. е. при чтении входного символа  $a$  автомат должен перейти из состояния 1 в состояние 2. Однако, поскольку имеются  $\varepsilon$ -переходы из состояния 1 в состояние 3, а из состояния 3 в состояние 4, автомат без обработки входного символа мог перейти в любое из этих состояний. Пусть автомат находится в состоянии 4, тогда чтение символа  $a$  может перевести автомат в состояние 4 или 6. Таким образом, при чтении символа  $a$  автомат может перейти из состояния 1 в состояние 2, 4 или 6.



Обозначим через  $t(k, a)$  множество состояний, которые могут быть достигнуты из состояния  $k$  после чтения входного символа  $a$ . Формально оно определяется следующим образом:

$$t(k, \varepsilon) = R(k);$$

$$t(k, a) = \bigcup_{k' \in R(k)} R(\delta(k', a)) \text{ для всех } k \in K, a \in T.$$

Например, для нашего автомата имеем

$$t(1, \varepsilon) = R(1) = \{1, 3, 4\};$$

$$\begin{aligned} t(1, a) &= \bigcup_{k' \in \{1, 3, 4\}} R(\delta(k', a)) = R(\{2\}) \cup R(\emptyset) \cup R(\{4, 6\}) = \\ &= \{2\} \cup \{4\} \cup \{6\} = \{2, 4, 6\}; \end{aligned}$$

$$t(1, b) = \bigcup_{k' \in \{1, 3, 4\}} R(\delta(k', b)) = R(\emptyset) \cup R(\emptyset) \cup R(\{5\}) = \{5\};$$

$$t(1, c) = \bigcup_{k' \in \{1, 3, 4\}} R(\delta(k', c)) = R(\emptyset) \cup R(\{3\}) \cup R(\emptyset) = \{3, 4\}.$$

Тогда правило построения НКА без  $\varepsilon$ -переходов  $M' = (K, T, \delta', k_0, F')$ , эквивалентного НКА с  $\varepsilon$ -переходами  $M = (K, T, \delta, k_0, F)$ , заключается в следующем:

$$\delta'(k, a) = t(k, a) \text{ для всех } k \in K, a \in T;$$

$F' = F \cup \{k \in K \mid R(k) \cap F \neq \emptyset\}$ , т. е. к множеству конечных состояний добавляются состояния  $k$ , для которых в их множества  $\varepsilon$ -достижимых состояний  $R(k)$  входят конечные состояния исходного автомата.

Построим автомат без  $\epsilon$ -переходов, эквивалентный автомату на рис. 2.14. Это удобно делать по таблице переходов автомата, представленной на рис. 2.15 (для выделения конечное состояние заключено в квадратные скобки).

Вход ( $T$ )	Состояния ( $K$ )					
	1	2	3	4	5	[6]
$a$	{2}	$\emptyset$	$\emptyset$	{4, 6}	$\emptyset$	$\emptyset$
$b$	$\emptyset$	{4}	$\emptyset$	{5}	$\emptyset$	$\emptyset$
$c$	$\emptyset$	$\emptyset$	{3}	$\emptyset$	{4}	$\emptyset$
$\epsilon$	{3}	$\emptyset$	{4}	$\emptyset$	$\emptyset$	$\emptyset$
$R(k)$	{1, 3, 4}	{2}	{3, 4}	{4}	{5}	{[6]}

Рис. 2.15. Таблица переходов НКА, дополненная строкой для  $R(k)$

Множества  $R(k)$  легко можно вычислить транзитивно по строке  $\epsilon$ . Например, при вычислении  $R(1)$  само состояние 1 сразу включается в это множество. В строке  $\epsilon$  для состояния 1 имеем состояние 3, которое добавляется в  $R(1)$ . В строке  $\epsilon$  для состояния 3 имеем состояние 4, которое также добавляется в  $R(1)$ . В строке  $\epsilon$  для состояния 4 содержится  $\emptyset$ , следовательно, процесс вычисления  $R(1) = \{1, 3, 4\}$  завершен.

После вычисления множеств  $R(k)$  для всех состояний  $k \in K$  легко можно вычислить множества  $t(k, a)$ . Рассмотрим состояние  $k$ . Сначала для каждого состояния из  $R(k)$  по таблице переходов определяется множество состояний, в которые может перейти автомат при чтении символа  $a$ . Затем выполняется объединение всех множеств  $R(k)$ , содержащихся в столбцах, соответствующих найденным состояниям. В результате будет получено множество  $t(k, a)$ . Рассмотрим вычисление множеств  $t(k, a)$  на примере состояния 1 при чтении символа  $c$ , т. е. вычислим  $t(1, c)$ . Поскольку  $R(1) = \{1, 3, 4\}$ , согласно определению

$$t(1, c) = R(\delta(1, c)) \cup R(\delta(3, c)) \cup R(\delta(4, c)).$$

По таблице переходов определяем  $\delta(1, c) = \emptyset$ ,  $\delta(3, c) = \{3\}$ ,  $\delta(4, c) = \emptyset$ , следовательно,

$$t(1, c) = R(\emptyset) \cup R(\{3\}) \cup R(\emptyset) = \{3, 4\}.$$

Процесс построения автомата без  $\epsilon$ -переходов завершается после вычисления функции переходов  $t(k, a)$  для всех  $k \in K, a \in T$  и уточнения множества конечных состояний.

Функция переходов конечного автомата без  $\epsilon$ -переходов, эквивалентного конечному автомату с  $\epsilon$ -переходами из рис. 2.14, представлена на рис. 2.16, а граф автомата – на рис. 2.17.

Вход ( $T$ )	Состояния ( $K$ )					
	1	2	3	4	5	[6]
$a$	{2, 4, 6}	$\emptyset$	{4, 6}	{4, 6}	$\emptyset$	$\emptyset$
$b$	{5}	{4}	{5}	{5}	$\emptyset$	$\emptyset$
$c$	{3, 4}	$\emptyset$	{3, 4}	$\emptyset$	{4}	$\emptyset$

Рис. 2.16. Таблица переходов автомата без  $\epsilon$ -переходов

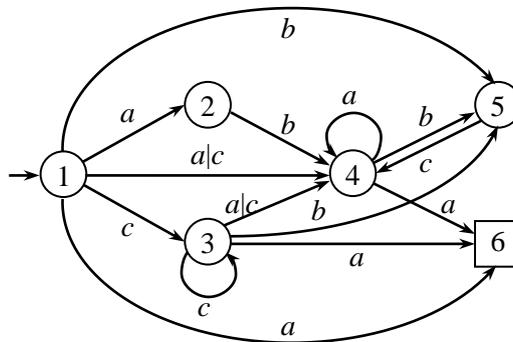


Рис. 2.17. Конечный автомат без  $\epsilon$ -переходов

После исключения из НКА  $\epsilon$ -переходов автомат в общем случае остается недетерминированным. Поскольку использование НКА в качестве распознавателя приводит к существенным потерям времени при лексическом анализе, на следующем этапе необходимо преобразовать его в эквивалентный ДКА.

Для сложных регулярных выражений построенный ДКА может оказаться не минимальным. Поэтому завершающим этапом является минимизация ДКА.

## 2.9. Минимизация конечного автомата

При построении ДКА выгодно, чтобы автомат имел как можно меньше состояний. Для произвольного ДКА можно построить эквивалентный автомат с наименьшим числом состояний (возможно, им будет исходный автомат). Процедуру построения такого автомата называют *минимизацией конечного автомата*, а сам автомат – *минимальным конечным автоматом*.

Для простоты будем рассматривать только ДКА, имеющие полную функцию переходов. Если исходный ДКА имеет неполную функцию переходов, то, добавив новое фиктивное состояние, можно определить новый эквивалентный ДКА с полной функцией переходов.

Автомату без выхода можно поставить в соответствие автомат Мура, у которого дуги, ведущие в конечные состояния, помечаются 1, а все остальные дуги помечаются 0. Тогда задача сводится к задаче минимизации автоматов Мура.

Другой путь – минимизация непосредственно автомата без выхода. Пусть  $M = (K, T, \delta, k_0, F)$  – ДКА с полной функцией переходов, принимающий язык  $L$ . Рассмотрим построение минимального ДКА  $M_L$  исходя только из языка  $L$ .

Минимизация заключается в последовательном разбиении множества состояний на непересекающиеся подмножества неразличимых состояний до тех пор, пока такое разбиение возможно.

Определим на множестве  $K$  отношения  $D_0, D_1, D_2, \dots$  следующим образом:  $kD_0k'$  (состояния  $k$  и  $k'$  различимы по строке длины 0) тогда и только тогда, когда  $k \in F$  и  $k' \notin F$  или  $k \notin F$  и  $k' \in F$ .

Пусть  $i > 0$ ,  $kD_i k'$  (состояния  $k$  и  $k'$  различимы по строке длины  $\leq i$ ) тогда и только тогда, когда  $kD_{i-1} k'$ , т. е. существует  $a \in T$ , такое, что  $\delta(k, a)D_{i-1}\delta(k', a)$ . Говорят, что состояние  $k$  различимо от состояния  $k'$ , если существует такое  $i \geq 0$ , что  $kD_i k'$ . Другими словами,  $kD_i k'$  тогда и только тогда, когда существует такая строка  $x$ ,  $|x| \leq i$ , что либо  $\delta(k, x) \in F$  и  $\delta(k', x) \notin F$ , либо  $\delta(k, x) \notin F$  и  $\delta(k', x) \in F$ .

Чтобы вычислить отношение  $D$ , необходимо последовательно вычислить отношения  $D_0, D_1, D_2, \dots$ , и если в процессе этих вычислений на некотором шаге  $r$  окажется, что  $D_{r+1} = D_r$ , то это будет означать, что итерационный процесс окончен и  $D = D_r$ .

Если  $m = |K|$ , то существует только  $(m^2 - m)$  пар  $(k_i, k_j)$ , где  $i \neq j$ . В худшем случае всякое  $D_{i+1}$  отличается от  $D_i$  двумя такими парами, тогда  $D = D_r$ , где  $r < (m^2 - m) / 2$ , т. е. процесс всегда конечен. В результате будут найдены все неразличимые состояния автомата, которые образуют множество состояний минимального автомата  $M_L$ .

Пусть  $K' \subset K$  – одно из множеств неразличимых состояний. функция переходов автомата  $M_L$  определяется следующим образом:  $\delta_L(K', a) = K''$ , где  $K''$  – множество неразличимых состояний, содержащее состояние  $\delta(k, a)$  для всех  $k \in K'$ . Тогда начальным состоянием автомата  $M_L$  является множество неразличимых состояний, содержащее начальное состояние исходного автомата  $M$ . Конечными состояниями автомата  $M_L$  являются те множества неразличимых состояний, которые содержат конечные состояния автомата  $M$ .

Рассмотрим пример минимизации ДКА, граф переходов которого приведен на рис. 2.18.

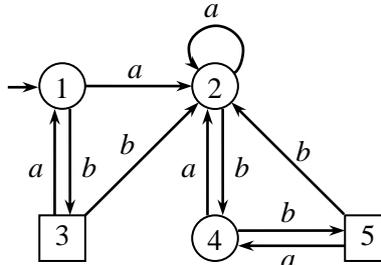


Рис. 2.18. Граф переходов минимизируемого ДКА

Отношение  $D_i$  можно представить в виде булевой матрицы размером  $5 \times 5$ , в которой значение элемента  $(j, k)$  равно  $T$  (true), если  $jD_i k$ , и равно  $F$  (false) противном случае. Поскольку отношение  $D_i$  симметрично и никогда не выполняется отношение  $kD_i k$  (элементы главной диагонали всегда равны  $F$ ), для построения матрицы достаточно определить только те элементы, которые находятся над главной диагональю.

Последовательность матриц, соответствующих отношениям  $D_0, D_1$ , будет следующей:

$D_0$	1	2	3	4	5
1		$F$	$T$	$F$	$T$
2			$T$	$F$	$T$
3				$T$	$F$
4					$T$
5					

$D_1$	1	2	3	4	5
1		$T$	$T$	$F$	$T$
2			$T$	$T$	$T$
3				$T$	$F$
4					$T$
5					

Например, элемент  $(1, 2)$  равен  $F$ , поскольку оба состояния 1 и 2 не являются конечными (неразличимы по строке длины 0), элемент  $(1, 3)$  равен  $T$ , поскольку состояние 1 не является конечным, а состояние 3 – конечное (различимы по строке длины 0), т. е.  $1D_03$ . Аналогично определяются остальные значения элементов матрицы для отношения  $D_0$ .

Для отношения  $D_1$  имеем  $1D_12$ , поскольку  $\delta(1, b)D_0\delta(2, b)$ , и  $2D_14$ , поскольку  $\delta(2, b)D_0\delta(4, b)$ .

Отношение  $D_2$  будет полностью совпадать с отношением  $D_1$ , т. е. процесс минимизации завершен. В результате имеем, что состояния 1 и 4 неразличимы между собой, а также состояния 3 и 5 также неразличимы. Таким образом, выполнено разбиение множества состояний на непересекающиеся подмножества неразличимых состояний:  $\{1, 4\}$ ,  $\{2\}$ ,  $\{3, 5\}$ . Начальным состоянием минимального автомата будет состояние  $\{1, 4\}$  (содержит начальное состояние 1 исходного автомата), конечным состоянием будет  $\{3, 5\}$  (содержат конечные состояния 3 и 5 исходного автомата). Переобозначим эти состояния через 1, 2 и 3 соответственно. Тогда граф переходов полученного минимального ДКА примет вид, приведенный на рис. 2.19.

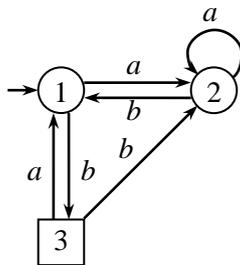


Рис. 2.19. Минимальный ДКА