# Конечные автоматы, регулярные выражения, нерегулярные языки

14 января 2014 г.

## 1 Конечные автоматы и регулярные языки

*Языком* будем называть множество строк над некоторым алфавитом.

**Определение 1.** *Конечный автомат* — это пятерка $(Q, \Sigma, \delta, q_0, F)$, где

$Q$ — конечное множество *состояний*,

$\Sigma$ — конечное множество, называемое *алфавитом*,

$\delta : Q \times \Sigma \to Q$ — *функция переходов*,

$q_0 \in Q$ — *начальное состояние*,

$F \subseteq Q$ — множество *заключительных состояний*.

Автомат $M = (Q, \Sigma, \delta, q_0, F)$ *допускает* строку $w = w_1 w_2 \dots w_n$, где $w_i \in \Sigma$, если существует последовательность состояний $r_0, r_1, \dots, r_n$, такая что

1. $r_0 = q_0$,
2. $\delta(r_i, w_{i+1}) = r_{i+1}$ для $i = 0, \dots, n-1$,
3. $r_n \in F$.

Автомат $M$ *распознает язык* $A = \{w \mid M \text{ допускает } w\}$.

**Определение 2.** Язык называется *регулярным*, если его распознает некоторый конечный автомат.

*Недетерминированный конечный автомат* отличается от детерминированного функцией переходов, которая имеет следующий вид:

$$\delta : Q \times (\Sigma \cup \{\epsilon\}) \to \mathfrak{P}(Q),$$

где $\mathfrak{P}(Q)$ — множество всех подмножеств $Q$, а $\epsilon$ — пустая строка. Переход по $y \in \Sigma \cup \{\epsilon\}$ из состояния $r_i$ в состояние $r_j$ разрешен, если $r_j \in \delta(r_i, y)$.

Для любого недетерминированного автомата можно построить детерминированный автомат, распознающий тот же язык.

Регулярные языки замкнуты относительно следующих операций:

**Объединение:** $A \cup B = \{w \mid w \in A \text{ или } w \in B\}$

**Пересечение:** $A \cap B = \{w \mid w \in A \text{ и } w \in B\}$

**Дополнение:** $\overline{A} = \{w \in \Sigma^\star \mid w \notin A\}$

**Обращение:** $A^R = \{a_1 \dots a_k \mid a_k \dots a_1 \in A\}$

**Конкатенация:** $A \circ B = \{vw \mid v \in A \text{ и } w \in B\}$

**Замыкание Клини:** $A^* = \{w_1 \dots w_k \mid k \geq 0 \text{ и } w_i \in A \text{ для всех } i\}$

# 2 Регулярные выражения

**Определение 3.** $R$ называется *регулярным выражением* над алфавитом $\Sigma$, если $R$ — это

1. $a$, где $a \in \Sigma$, или
2. $\epsilon$, где $\epsilon$ — пустая строка, или
3. $\varnothing$, или
4. $(R_1 \cup R_2)$, где $R_1$ и $R_2$ — регулярные выражения, или
5. $(R_1 R_2)$, где $R_1$ и $R_2$ — регулярные выражения, или
6. $(R_1^*)$, где $R_1$ — регулярное выражение.

$a \in \Sigma$ соответствует языку $\{a\}$;

$\epsilon$ соответствует языку $\{\epsilon\}$;

$\varnothing$ соответствует языку $\varnothing$.

Если $R_1$ и $R_2$ — регулярные выражения, соответствующие языкам $L_1$ и $L_2$, то

$(R_1 \cup R_2)$ соответствует языку $L_1 \cup L_2$;

$(R_1 R_2)$ соответствует языку $L_1 \circ L_2$;

$(R_1^*)$ соответствует языку $L_1^\star$.

$L(R)$ — язык, которому соответствует регулярное выражение $R$.

Регулярное выражение $R$ *допускает* строку $w$ ($w$ *соответствует* $R$), если $w \in L(R)$.

**Пример 1.** *Строка 01010 соответствует выражению* $(01)^*0$.

Если, как в предыдущем примере, скобки опускают, то подразумевается следующий приоритет операций: $R^\star$, $R_1 R_2$, $R_1 \cup R_2$.

**Пример 2.**

$$R_1^* R_2 \cup R_3 = ((R_1^\star) R_2) \cup R_3$$

**Пример 3.** *Пусть* $\Sigma = \{0, 1\}$. *Язык*

$$\{w \mid w \text{ содержит ровно одно вхождение } 1\}$$

*соответствует выражению*

$$0^*10^*.$$

*Выражение* $\varnothing^\star$ *описывает язык* $\{\epsilon\}$. *Язык*

$\{w \mid w$ *содержит не менее трех символов, причем третий символ —* $0\}$

*соответствует выражению*

$$(0 \cup 1)(0 \cup 1)0(0 \cup 1)^*.$$

*Язык*

$$\{w \mid \text{на нечетных позициях в } w \text{ находится } 1\}$$

*соответствует выражению*

$$(1(0 \cup 1))^*(1 \cup \epsilon).$$

**Теорема 1.** *По регулярному выражению $R$ можно построить недетерминированный конечный автомат $N$, распознающий язык $L(R)$.*

*Доказательство.* Докажем по индукции. Базис: $R$ содержит один символ. В этом случае $R$ имеет вид: $\epsilon$, или $\varnothing$, или $a$, где $a \in \Sigma$. Несложно построить автомат, распознающий $L(R)$.

Шаг индукции: допустим, что каждое регулярное выражение длины меньше $k > 1$ соответствует некоторому регулярному языку и рассмотрим выражение $R$ длины $k$. Три варианта:

$R = R_1 \cup R_2$**:** По предположению индукции $L_1 = L(R_1)$ и $L_2 = L(R_2)$ — регулярные языки. Тогда $L(R) = L(R_1 \cup R_2) = L_1 \cup L_2$ является регулярным языком, так как множество регулярных языков замкнуто относительно объединения.

$R = R_1 R_2$**:** По предположению индукции $L_1 = L(R_1)$ и $L_2 = L(R_2)$ — регулярные языки. Тогда $L(R) = L(R_1 R_2) = L_1 \circ L_2$ является регулярным языком, так как множество регулярных языков замкнуто относительно конкатенации.

$R = (R_1)^*$**:** Аналогично.

□

**Теорема 2.** *Любому регулярному языку соответствует некоторое регулярное выражение.*

*Доказательство.* Доказательство состоит в преобразовании детерминированного автомата, распознающего язык $L$: будем постепенно удалять состояния и заменять метки на дугах регулярными выражениями. То, что получается в результате таких преобразований, называют *обобщенным недетерминированным автоматом.*

Добавим к конечному автомату уникальные и не совпадающие начальное и заключительное состояния. Свяжем новое начальное состояние с начальным состоянием исходного автомата и все заключительные состояния исходного автомата — с новым заключительным состоянием, пометив соответствующие дуги регулярным выражением $\epsilon$. Каждую дугу исходного автомата, помеченную символами $c_1, \ldots, c_k$ пометим регулярным выражением $(c_1 \cup \cdots \cup c_k)$. Если между двумя состояниями исходного автомата отсутствует дуга, соединим их в новом автомате дугой, помеченной регулярным выражением $\varnothing$.

Пока у автомата более двух состояний, выбираем какое-либо внутреннее состояние, удаляем его и помечаем дуги регулярными выражениями, соответствующими путям, проходившим через удаленное состояние. Обозначим за $R(s,t)$ регулярное выражение на дуге из $s$ в $t$. При удалении состояния $q$ дуга из состояния $i$ в состояние $j$ помечается следующим регулярным выражением:

$$(R(i,q)R(q,q)^*R(q,j) \cup R(i,j)).$$

□

Таким образом, регулярные языки $\Leftarrow$ по определению $\Rightarrow$ детерминированные конечные автоматы $\Longleftrightarrow$ недетерминированные конечные автоматы $\Longleftrightarrow$ регулярные выражения.

# 3 Нерегулярные языки

Язык

$$\{w \mid w \text{ содержит одинаковое количество 1 и 0}\}$$

не является регулярным, но язык

$$\{w \mid w \text{ содержит одинаковое число вхождений 01 и 10}\}$$

является регулярным.

**Лемма 1** (о разрастании)**.** *Если язык $A$ является регулярным, то для некоторого числа $p$ верно, что любая строка $s \in A$, содержащая не менее $p$ символов, может быть разделена на три части: $s = xyz$ таким образом, что*

1. *$xy^iz \in A$ для всех $i \geq 0$;*
2. *$|y| > 0$;*
3. *$|xy| \leq p$.*

Воспользуемся леммой о разрастании для доказательства нерегулярности языка $B = \{0^n1^n \mid n \geq 0\}$. Допустим, что $B$ — регулярный язык. Пусть $w = 0^p1^p$. Так как $B$ — регулярный язык, $w$ можно представить в виде $w = xyz$, где $|y| > 0$ и $xy^iz \in B$ для любого $i \geq 0$. Если $y$ целиком

состоит из нулей, то $xyyz$ содержит больше нулей, чем единиц. Если $y$ целиком состоит из единиц, то $xyyz$ содержит больше единиц, чем нулей. Если $y$ содержит как единицы так и нули, то $xyyz$ не соответствует $0^*1^*$ и, следовательно, не соответствует $0^n1^n$. Получили противоречие.

Докажем, что язык $B = \{1^{n^2} \mid n \geq 0\}$ не является регулярным. Допустим, что $B$ — регулярный язык. Пусть $w = 1^{p^2}$. Так как $B$ — регулярный язык, $w$ можно представить в виде $w = xyz$, где $|y| > 0$, $|xy| \leq p$ и $xy^iz \in B$ для любого $i \geq 0$. Тогда $xyyz = 1^{p^2+|y|} \in B$. Поскольку $0 < |y| \leq p$, получаем

$$p^2 + |y| \leq p^2 + p < p^2 + 2p + 1 = (p+1)^2$$

и

$$p^2 < p^2 + |y| < (p+1)^2.$$

Следовательно, $p^2 + y$ не является квадратом никакого числа и $xyyz = 1^{p^2+|y|} \notin B$. Значит, вопреки нашему предположению, $B$ не является регулярным.