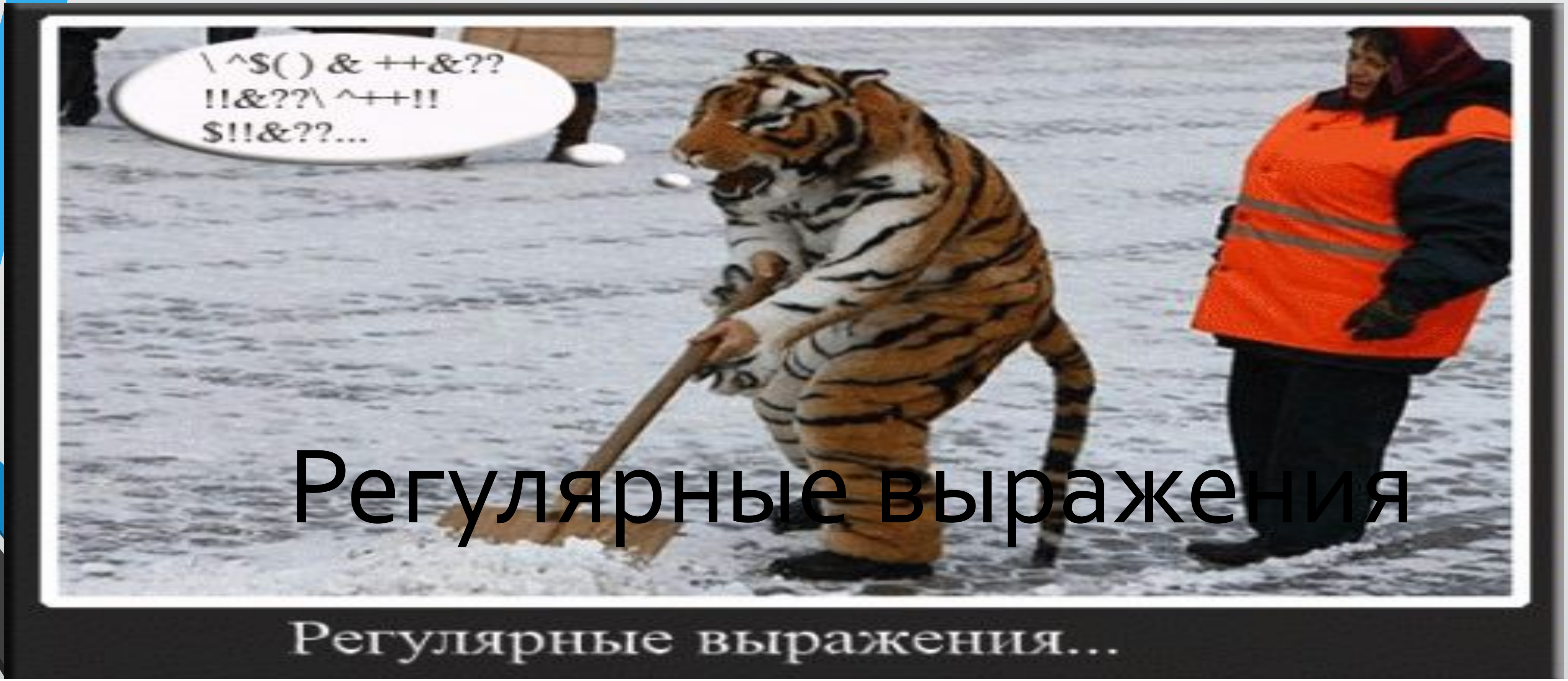
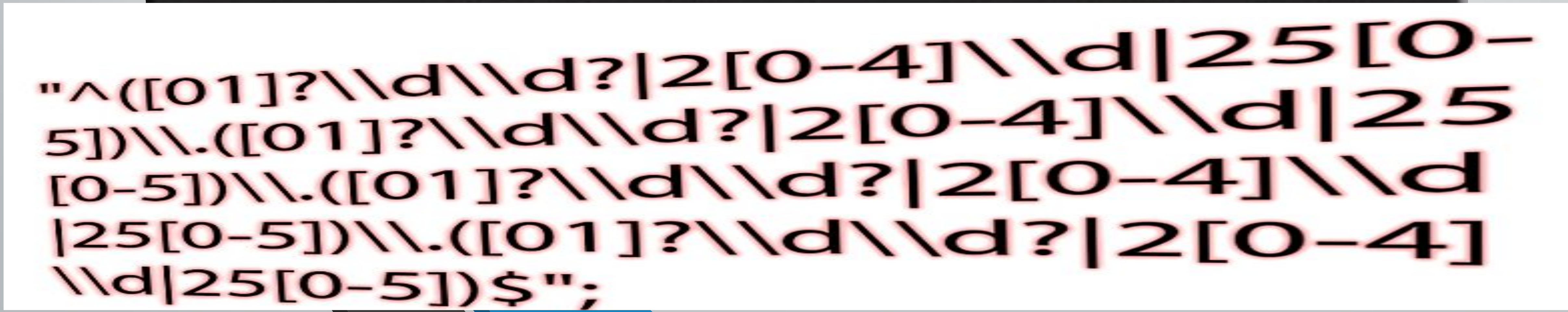
Регулярные выражения





```
"^([01]?\\d\\d?|2[0-4]\\d|25[0-5])\\.([01]?\\d\\d?|2[0-4]\\d|25[0-5])\\.([01]?\\d\\d?|2[0-4]\\d|25[0-5])\\.([01]?\\d\\d?|2[0-4]\\d|25[0-5])$";
```

Регулярные выражения

Регулярные выражения над алфавитом $\Sigma = \{a_1, a_2, \ldots, a_n\}$ — это формулы, которые определяются над множеством констант

1. $\mathbf{0}, \mathbf{1}$,
2. $\mathbf{a_1}, \mathbf{a_2}, \ldots, \mathbf{a_n}$,

и множеством операций, состоящим из

1. двухместных операций $\cdot, +$,
2. одноместной операции $*$.

Примеры регулярных выражений.

$(\mathbf{0}+\mathbf{1})$,
$(\mathbf{a}_1+\mathbf{a}_2)$,
$((\mathbf{a}_1\cdot\mathbf{a}_2)+(\mathbf{a}_2\cdot\mathbf{a}_1))$,
$(((\mathbf{a}_1\cdot((\mathbf{a}_1+\mathbf{a}_2)^*))+\mathbf{a}_2)^*)$.

Для упрощения записи введем приоритет операций: высокий для $*$, средний для $\cdot$, низкий для $+$. Будем опускать некоторые скобки. Например, запись

$\mathbf{a}_1\cdot\mathbf{a}_2{}^*+\mathbf{1}^*\cdot\mathbf{a}_1$

обозначает регулярное выражение

$((\mathbf{a}_1\cdot(\mathbf{a}_2{}^*))+((\mathbf{1}^*)\cdot\mathbf{a}_1))$.

Значением каждого регулярного выражения R является формальный язык $L(R)$, $L(R) \subseteq \Sigma^*$, который определяется по следующим правилам:

1. $L(\mathbf{0}) = \varnothing$,
2. $L(\mathbf{1}) = \{\varepsilon\}$,
3. $L(\mathbf{a_i}) = \{a_i\}$, $1 \le i \le n$,
4. $L(R_1 \cdot R_2) = L(R_1)L(R_2)$ (конкатенация),
5. $L(R_1 + R_2) = L(R_1) \cup L(R_2)$ (объединение),
6. $L(R_1^*) = L^*(R_1)$ (итерация).

Регулярный язык

Язык назыввется регулярным, если он является значением некоторого регулярного выражения.

Например, язык L, состоящий из всех слов четной длины над алфавитом $\Sigma = \{a, b\}$, является регулярным, поскольку $L = L(((a + b) \cdot (a + b))^*)$.

Регулярные множества и выражения

- Замечание:
- В современной логике и программировании вместо «+»

решено применять более точное по отношению ко множествам обозначение «|»(или).

(a+b)(a+bc)* перепишется (a|b)(a|bc)*

Регулярные выражения формально можно записать:

- S=a|SS|S+S|(S)|S*
- S=a|SS|(S|S)|(S)|S*

Регулярные выражения

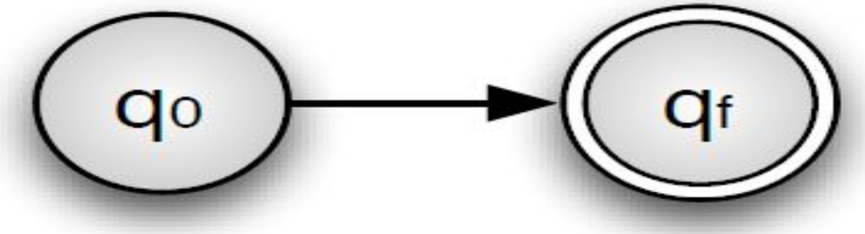
- $S=a|SS|(S|S)|(S)|S^*$

Регулярные выражения и конечные автоматы

Регулярные выражения это шаблон для некоторого языка,
конечный автомат – распознаватель.

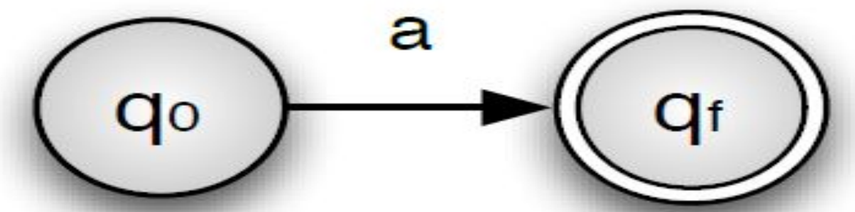
Задача-
построить по шаблону распознаватель

Построение автомата для регулярных выражений



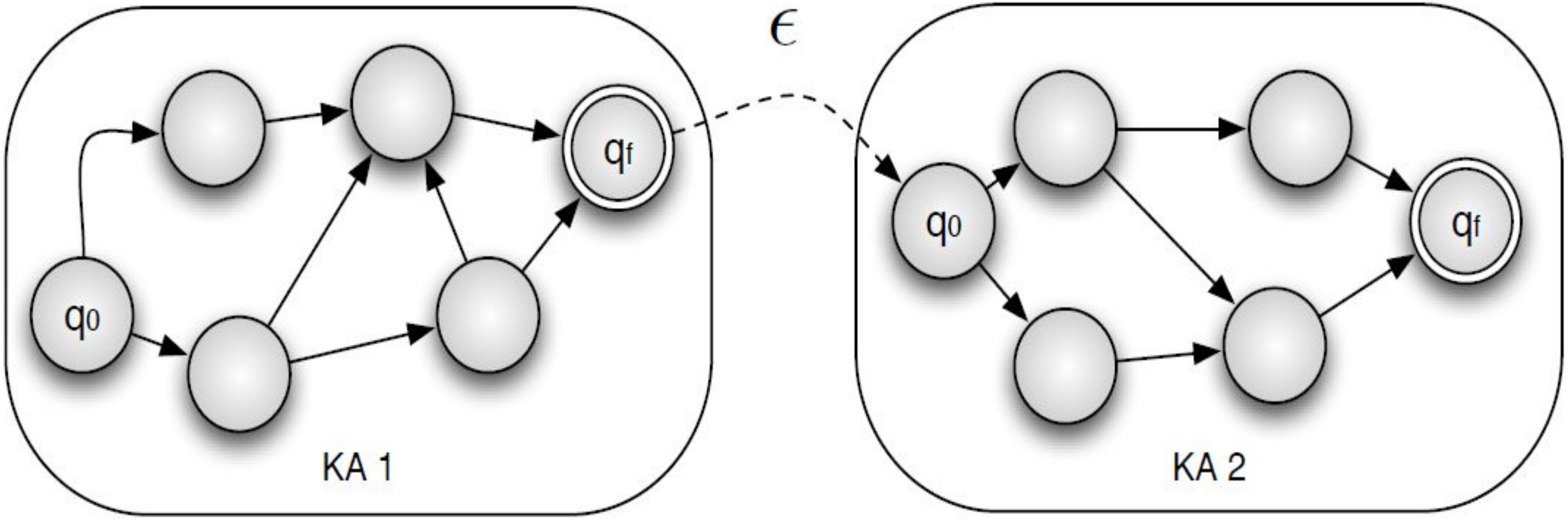
$$r = \epsilon$$

$$r = \varnothing$$



$$r = a$$

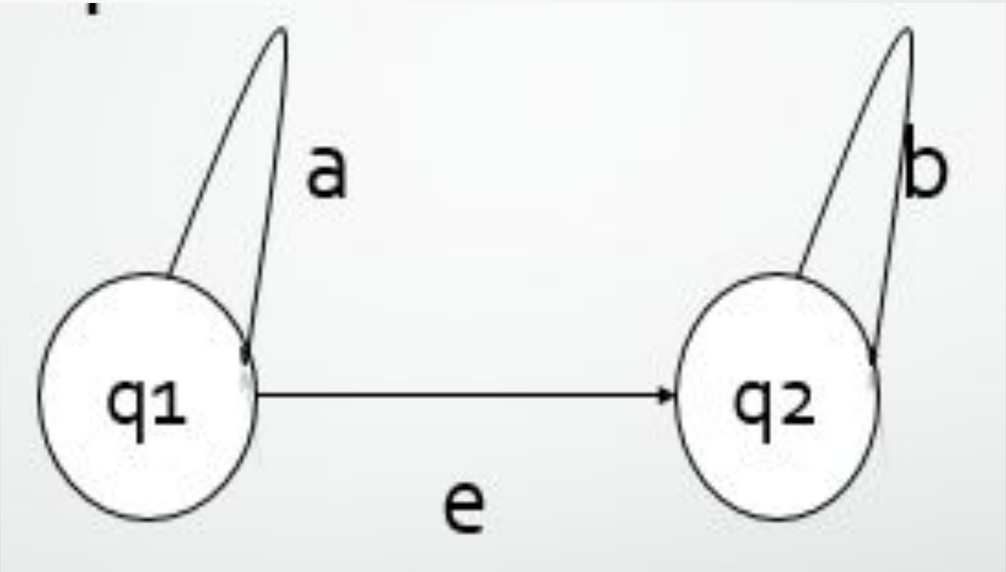
S=a|SS|(S|S)|(S)|S*



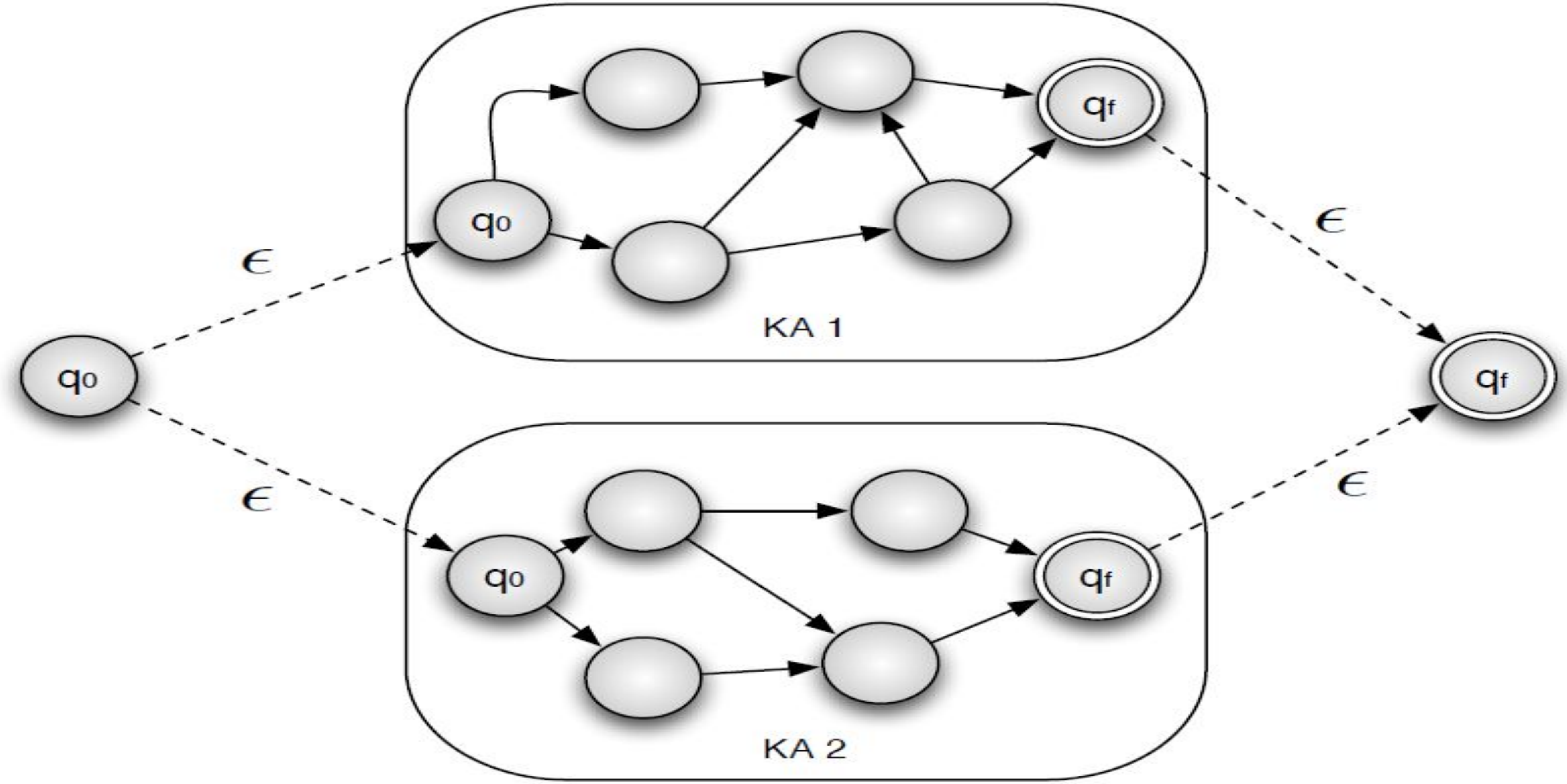
Последовательное соединение двух конечных автоматов

Пример использования единичного выражения для конкатенации

- $L=\{a^* b^*\}$



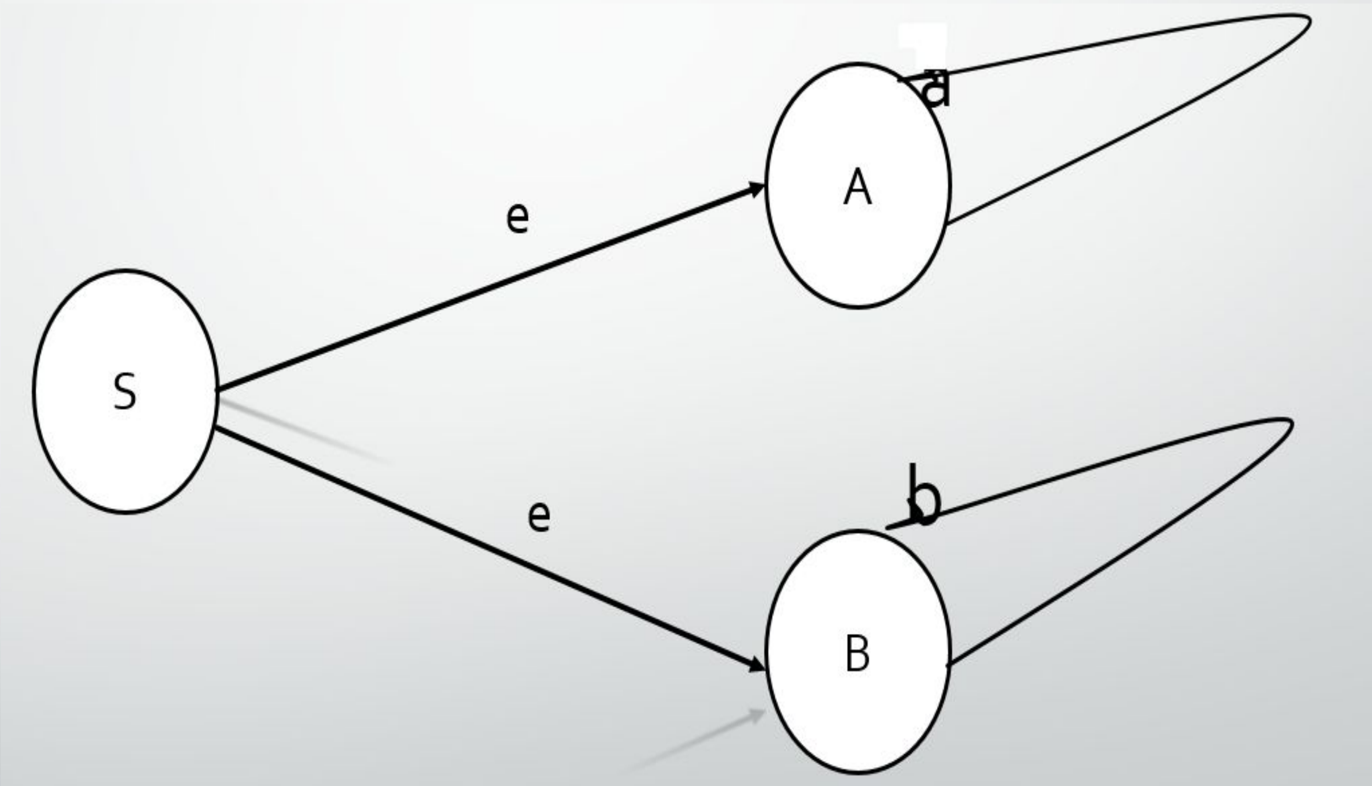
S=a|SS|(S|S)|(S)|S*



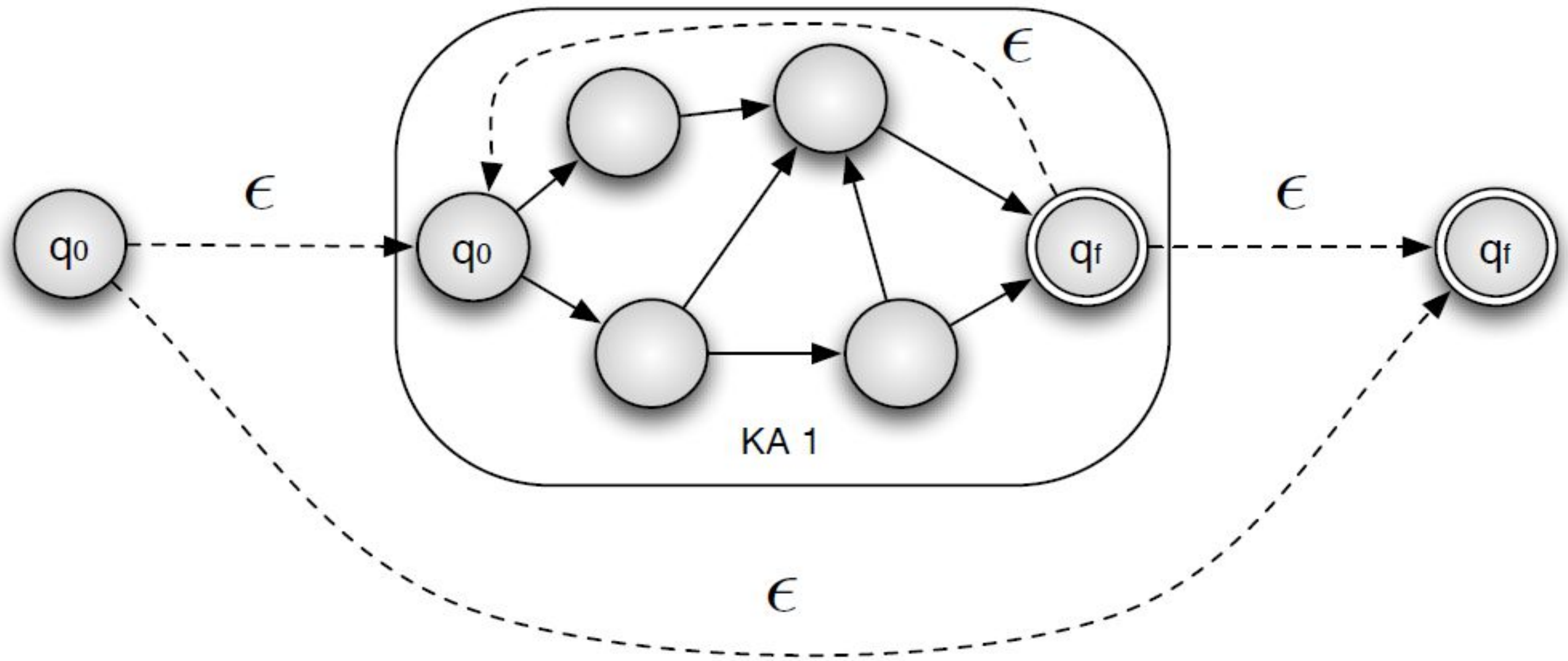
Объединение двух конечных автоматов

Пример использования единичного выражения для «или»

- L={a* |b*}



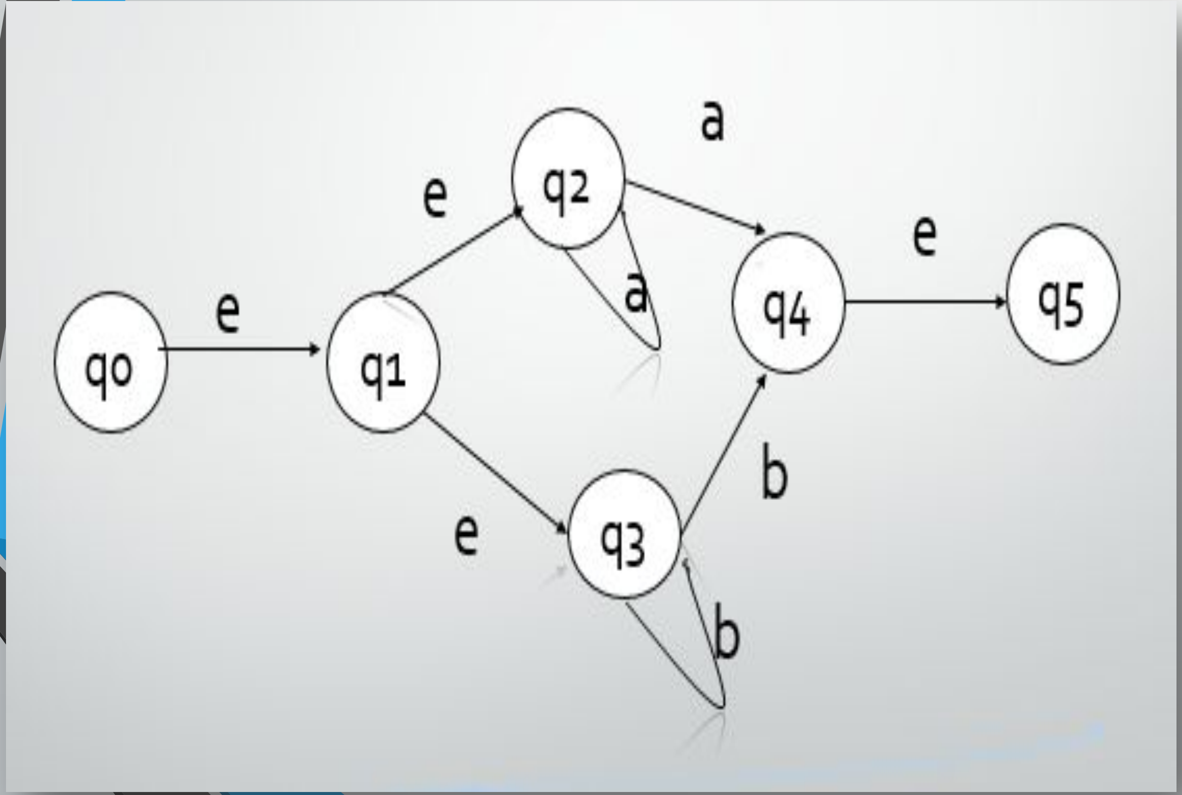
S=a|SS|(S|S)|(S)|S*



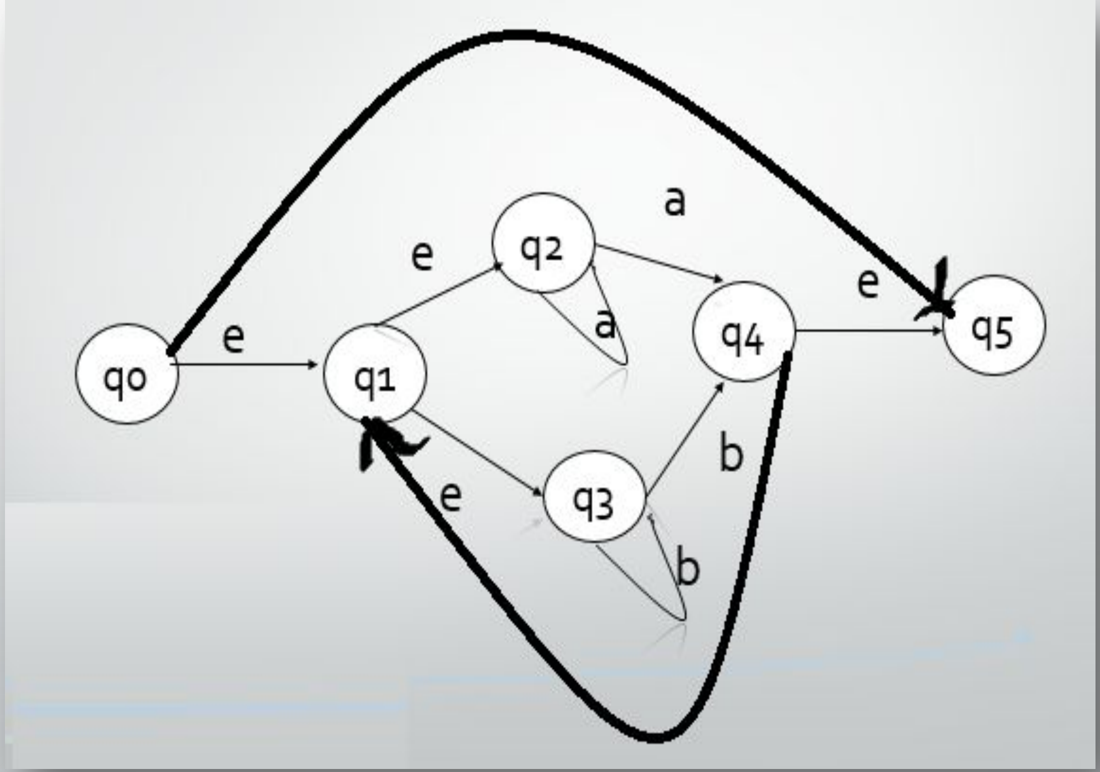
Замыкание конечного автомата

Пример использования единичного выражения для «*»

L={a+|b+}



L={(a+|b+)*}



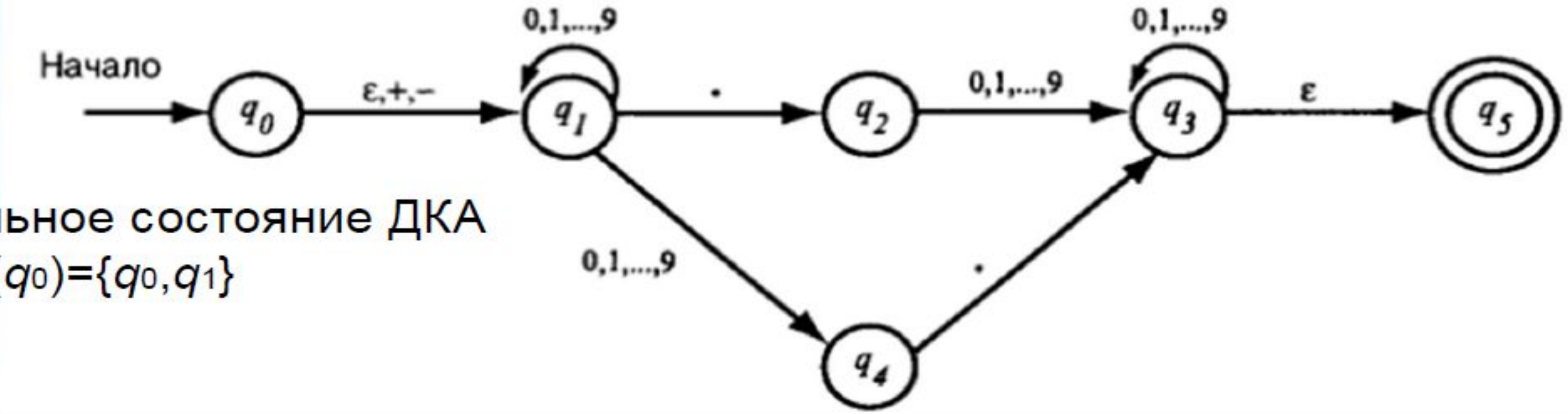
Построение минимального ДКА по регулярному выражению

список операций РВ в порядке их приоритетности:

- итерация (замыкание Клини), с помощью символа "*"
- конкатенация задается с помощью пробела или пустой строки (например: ab)
- Объединение(+), с помощью символа "|"

Пример, регулярное выражение числа

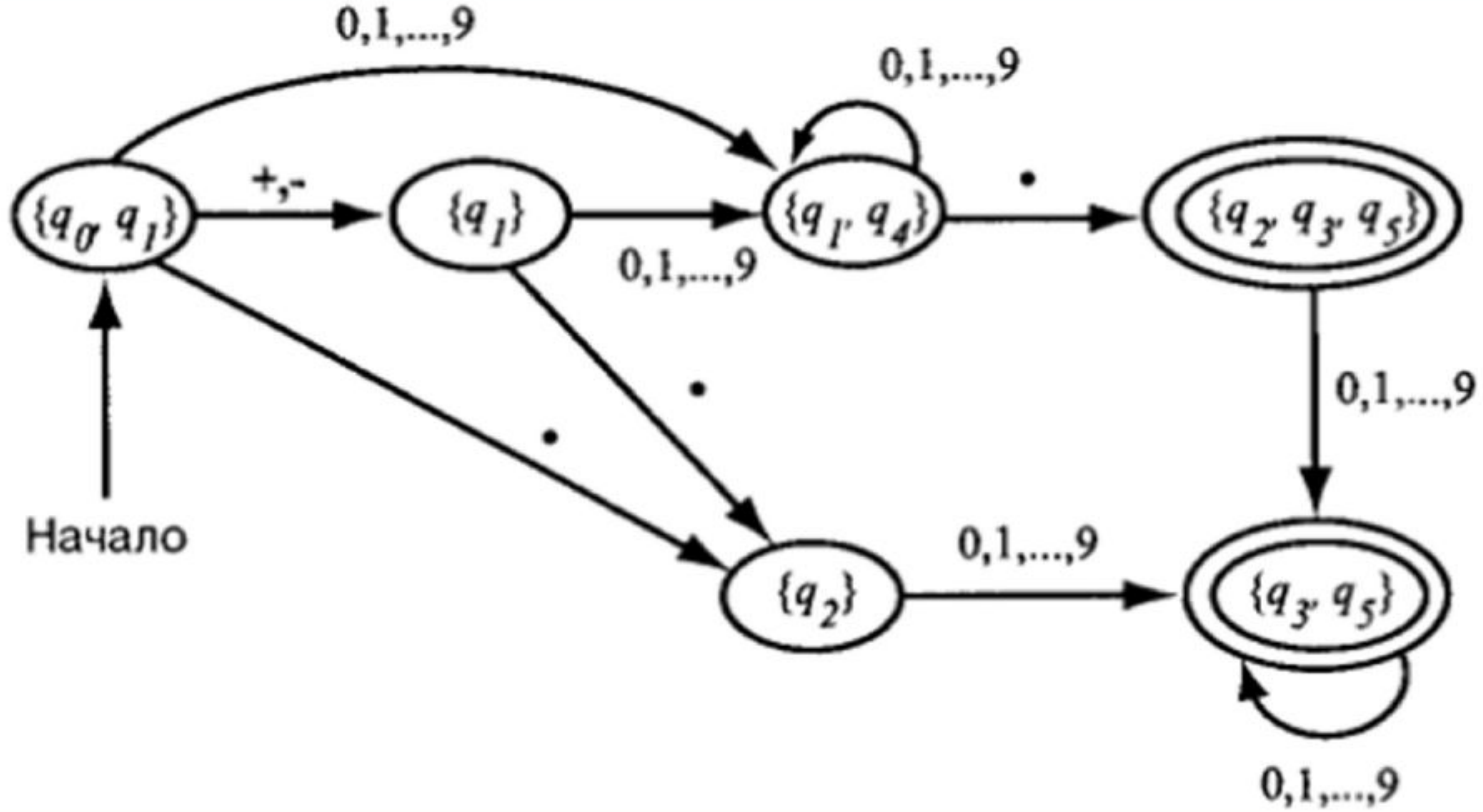
- (+|-|e) ((dd*.d*)|(d*.dd*)))



1. Начальное состояние ДКА

ECLOSE(q_0)={q_0,q_1}

	+	-	.	0-9
->{q_0,q_1}	{q_1}	{q_1}	{q_2}	{q_1,q_4}
{q_1}	Ø	Ø	{q_2}	{q_1,q_4}
{q_2}	Ø	Ø	Ø	{q_3,q_5}
{q_1,q_4}	Ø	Ø	{q_2,q_3,q_5}	{q_1,q_4}
*{q_3,q_5}	Ø	Ø	Ø	{q_3,q_5}
*{q_2,q_3,q_5}	Ø	Ø	Ø	{q_3,q_5}



0,1,...,9
0,1,...,9
+,-
{q0, q1}
{q1}
{q1, q4}
{q2, q3, q5}
0,1,...,9
.
.
.
0,1,...,9
Начало
{q2}
0,1,...,9
{q3, q5}
0,1,...,9

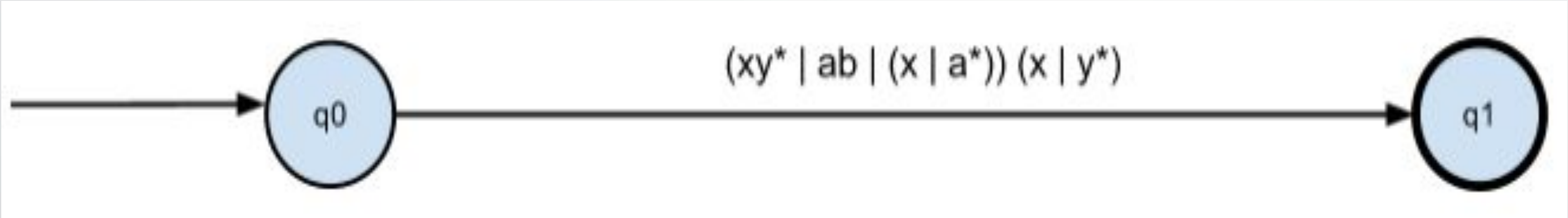
Рассмотрим пример, дано регулярное выражение:

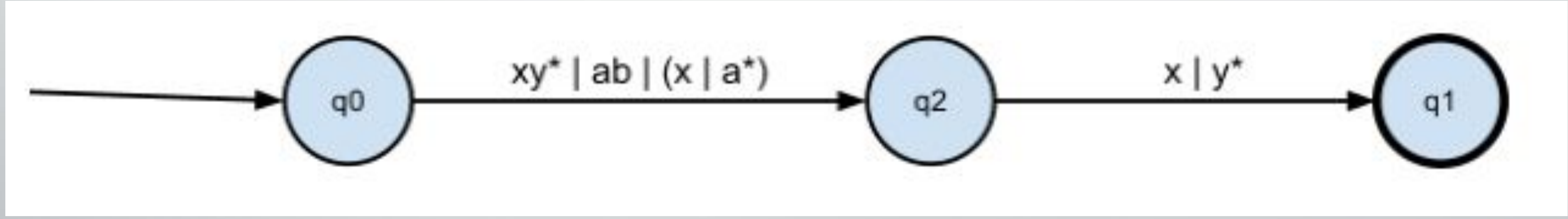
$$xy^*\,(x \mid y^*) \mid ab\,(x \mid y^*) \mid (x \mid a^*)\,(x \mid y^*)$$

- Для начала упростим данное РВ:

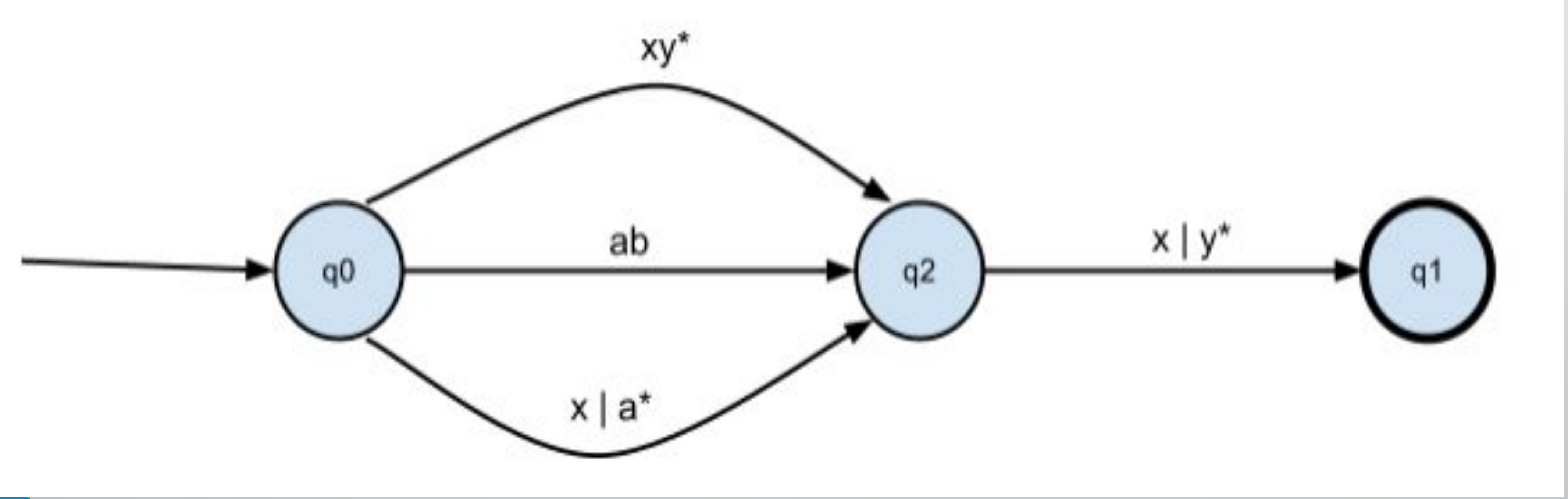
$$(xy^* \mid ab \mid (x \mid a^*))\,(x \mid y^*)$$

Теперь строим автомат по данному РВ:

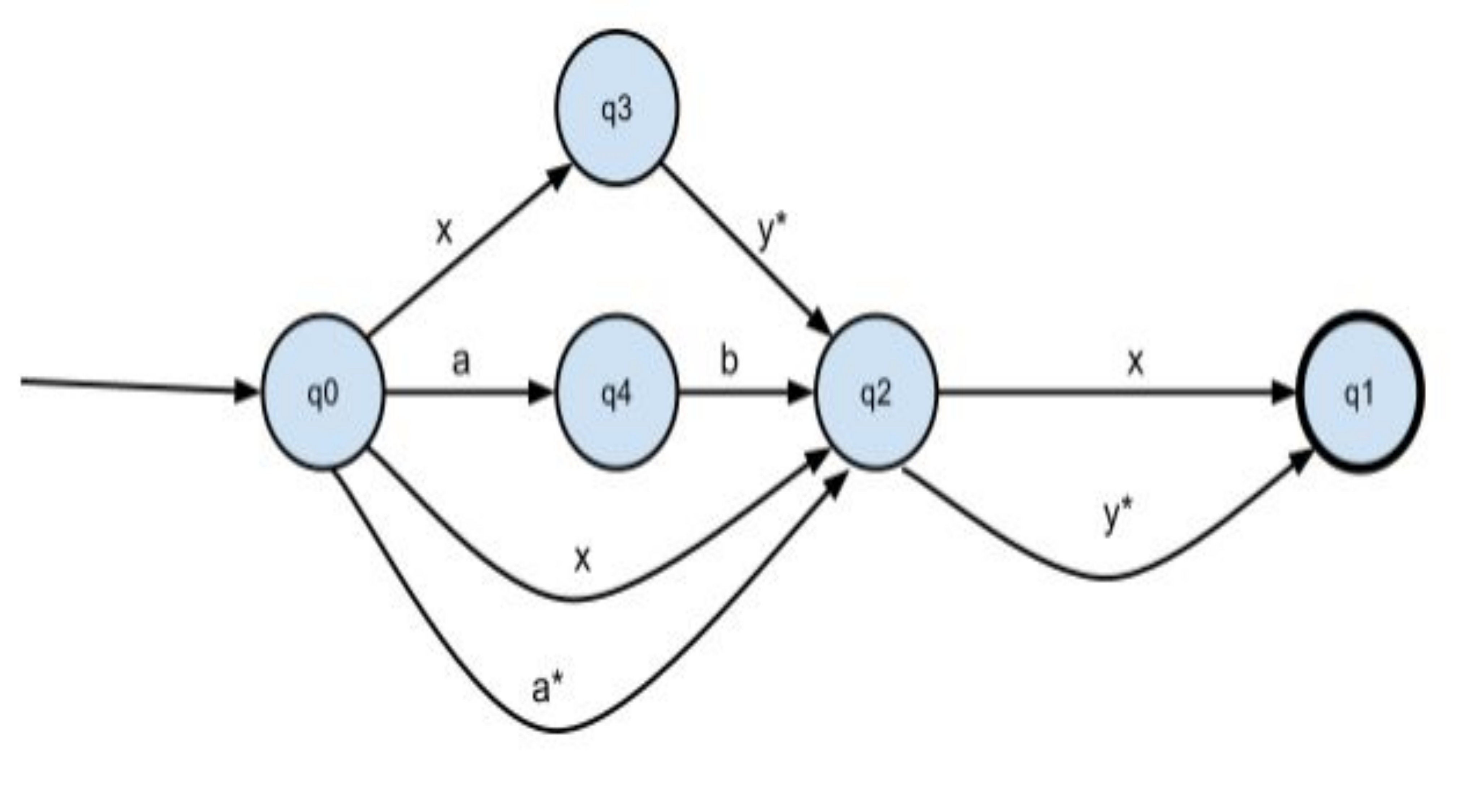




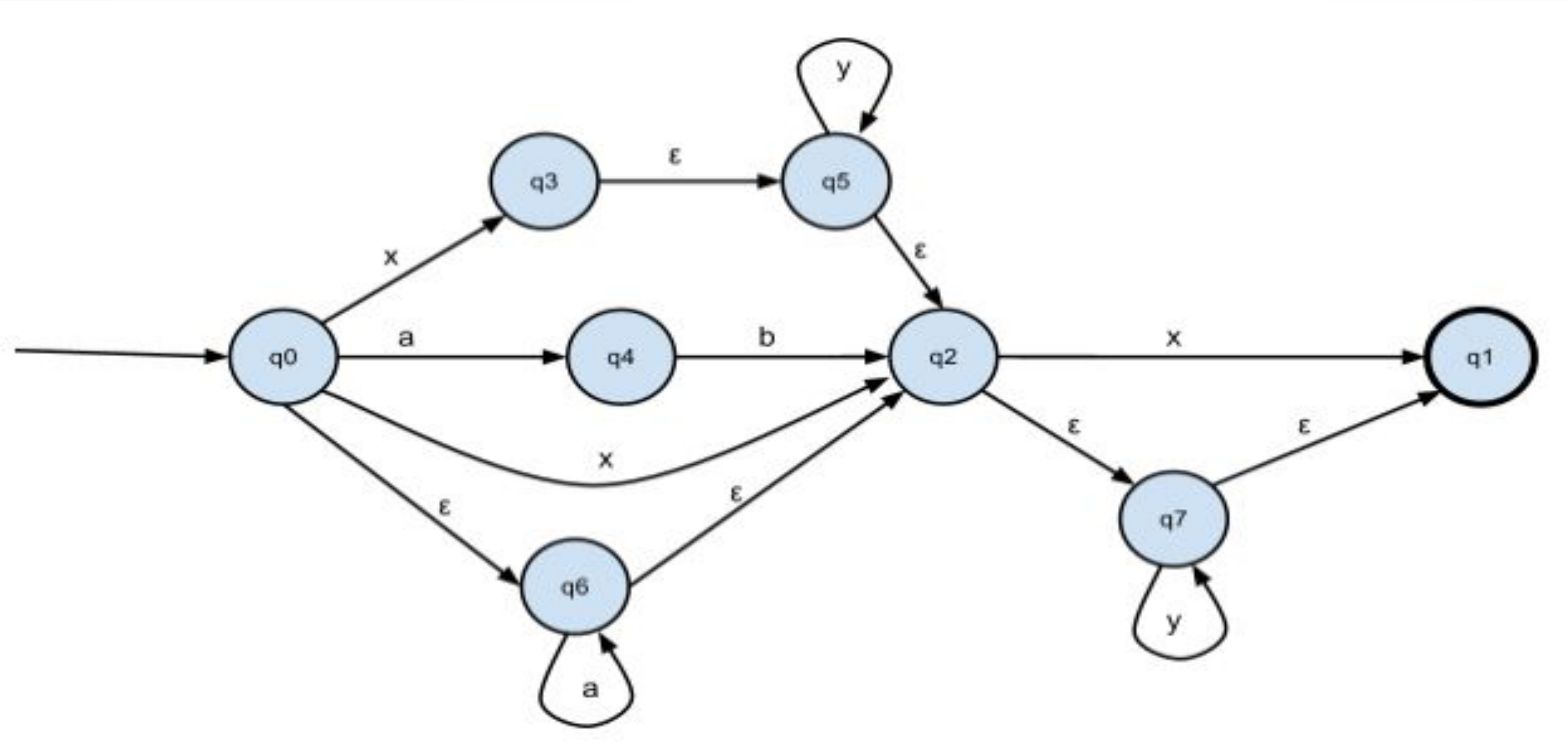
Продолжаем
(xy* | ab | (x | a*)) (x | y*)



Продолжаем (xy* | ab | (x | a*)) (x | y*)

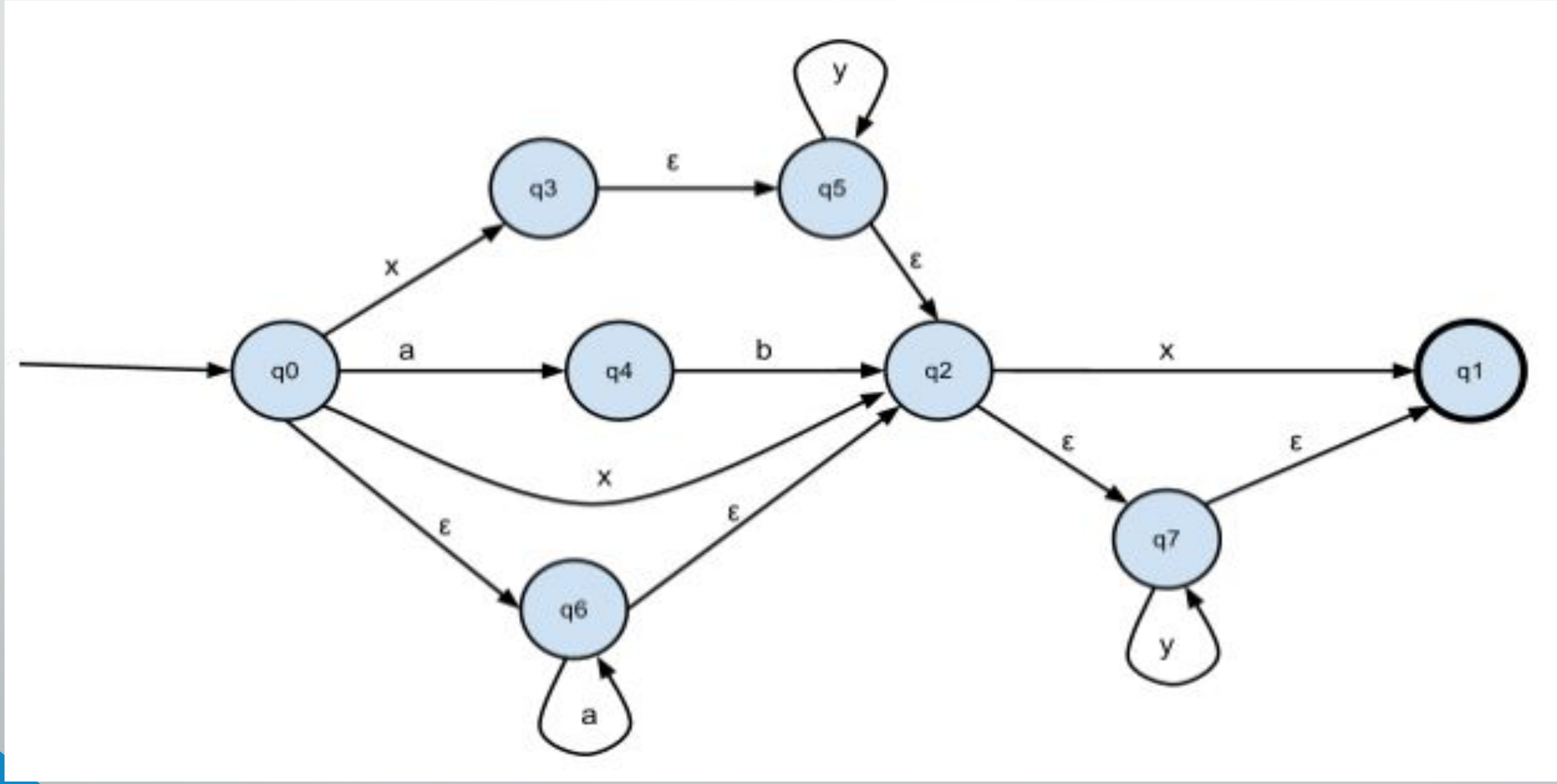


Продолжаем (xy* | ab | (x | a*)) (x | y*)

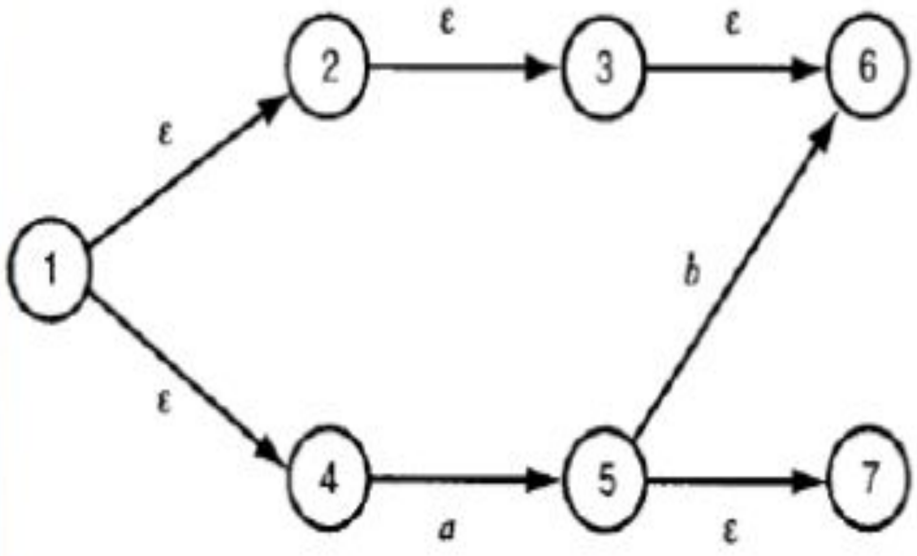


Преобразование недетерминированного автомата в детерминированный автомат

Избавляемся от ε-переходов



Определим ε-замыкание состояния q как множество состояний, доступных из q только по ε- переходам.



ECLOSE(1)={1,2,3,4,6}
ECLOSE(2)={2,36}
ECLOSE(5)={5,7}
ECLOSE(4)={4}

Детерминированный автомат, эквивалентный данному недетерминированному с ε-переходами, строится следующим образом:

- • множество состояний – множество всех подмножеств состояний исходного автомата;
- • множество входных символов такое же, как у исходного автомата;
- • функция переходов принимает в качестве аргументов состояние (множество

состояний исходного автомата) **q** и символ алфавита **c**,

значение функции –

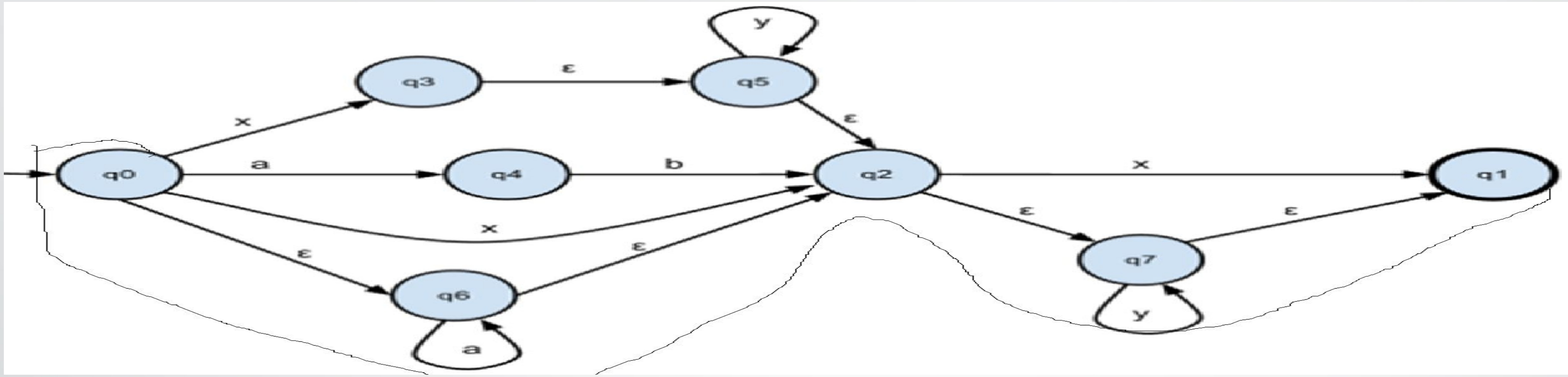
- состояние, соответствующее следующему множеству состояний исходного автомата,
- в которые можно перейти по символу **c** из ε-замыкания множества состояний **q**;

- • начальное состояние – ε-замыкание начального состояния исходного автомата;

• допускающие состояния – все множества состояний исходного автомата,

содержащие допускающие состояния.

(.0|-0)*0



№	Замкнутое Состояние	x	y	a	b
	q0	q3q2q1	q7	q4q6	
s0	q0q6q2q7q1	s1	s2	s3	
	q3q2q1	q1	q7q5		
s1	q3q2q1q5q7	s7	s4		
	q7		q7		
s2	q7q1		s2		
	q4q6	q1	q7	q6	q2
s3	q4q6q2q7q1	s7	s2	s5	s6
	q7q5	q1	q7q5		
s4	q7q5q2q1	s7	s4		
	q6	q1	q7	q6	
s5	q2q7q1	s7	s2	s5	
	q2	q1	q7		
s6	q2q7q1	s7	s2		
	q1				
s7					

Алгоритм

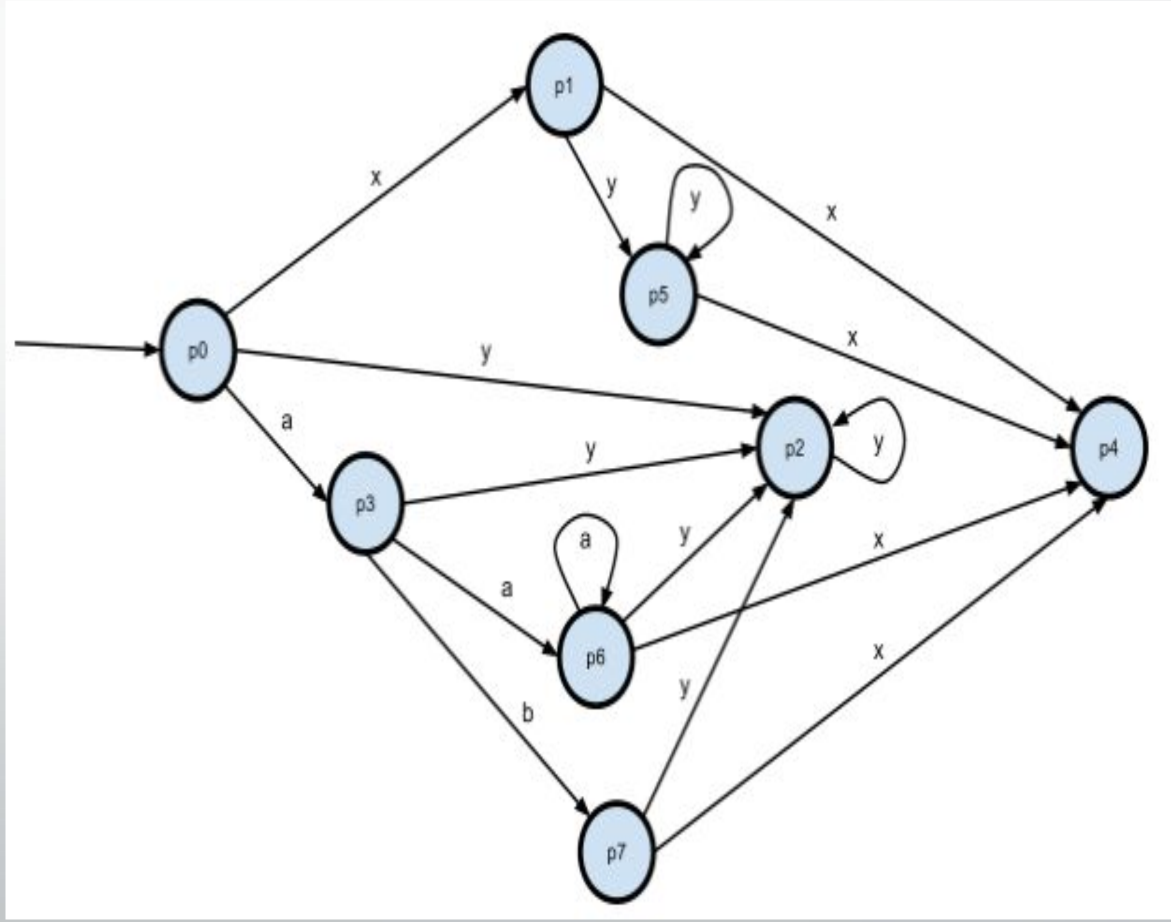
Начальное состояние – ε-замыкание начального состояния исходного автомата;

Создание строки для q

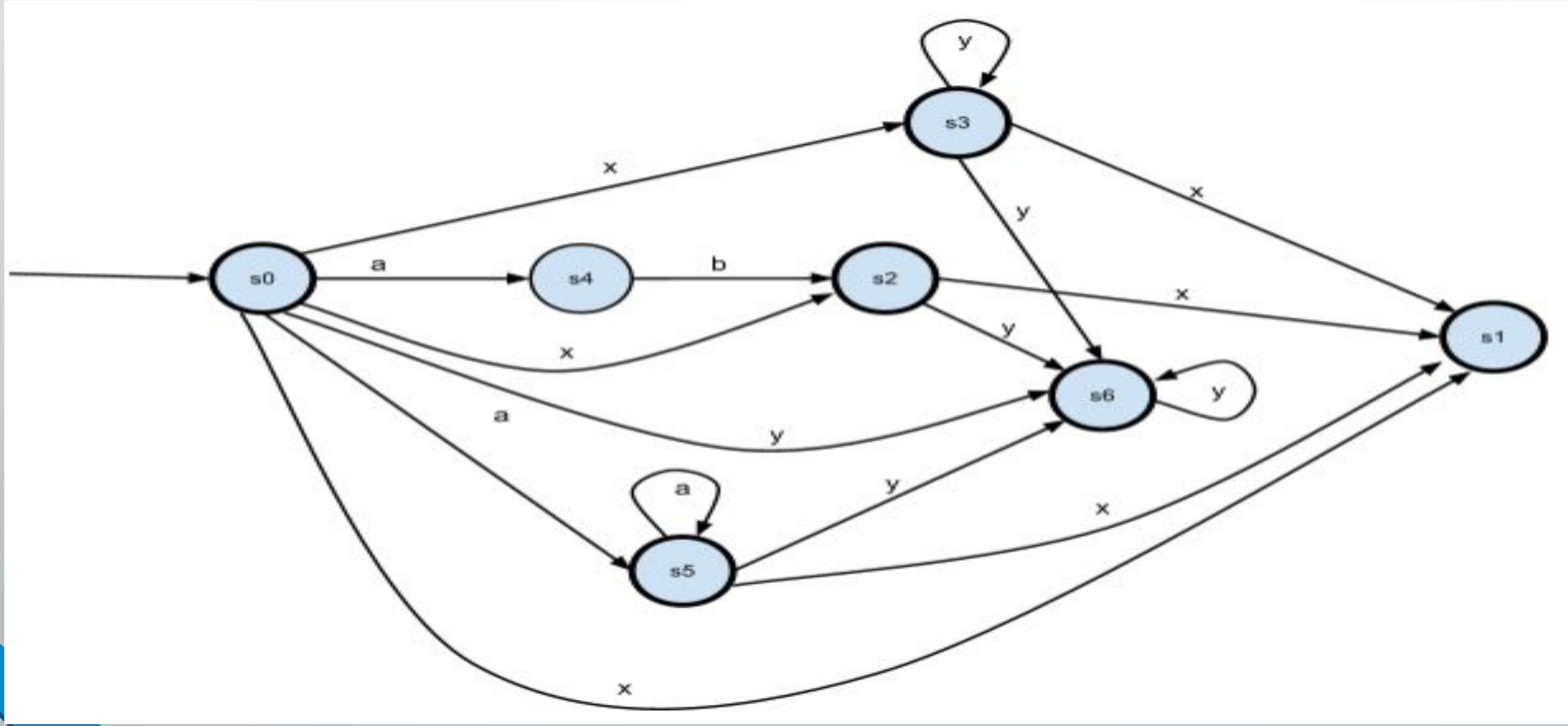
1. ε-Замыкание (q)
2. Поиск перехода((замыкание q),x)
3. Поиск перехода ((замыкание q),y)
4. Поиск перехода ((замыкание q),a)
5. Поиск перехода ((замыкание q),b)

Результат(xy* | ab | (x | a*)) (x | y*)

Состояние	x	y	a	b
s0	s1	s2	s3	
s1	s7	s4		
s2		s2		
s3	s7	s2	s5	s6
s4	s7	s4		
s5	s7	s2	s5	
s6	s7	s2		
s7				



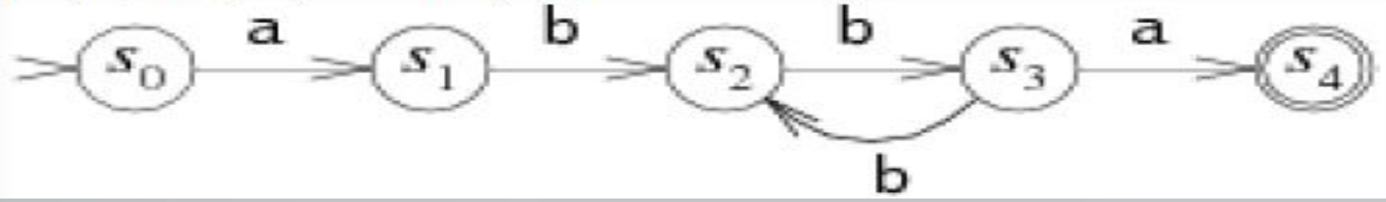
В данном НКА состояния s3 и s5 эквивалентны, так как $\delta(s3, x) = \delta(s5, x) = s1$ и $\delta(s3, y) = \delta(s5, y) = s5, s7$. Переименовываем состояния s6 -> s5 и s7 -> s6:



Регулярные выражения и конечные автоматы

1. описываем шаблон(регулярное выражение)
2. строим по шаблону НКА
3. НКА преобразуем а КДА
4. Минимизируем КДА
5. ДКА используем как распознаватель

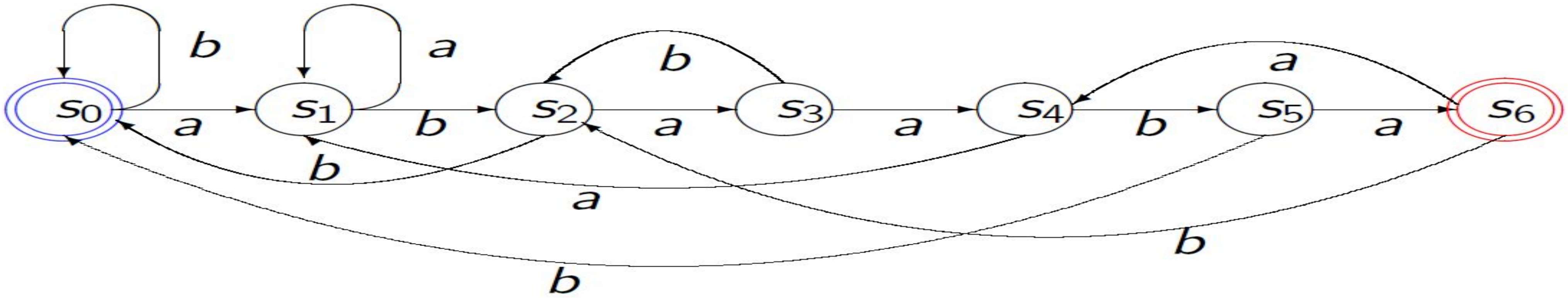
На рисунке приведен пример распознавания машиной шаблона, который соответствует регулярному выражению A(BB)+A:



Пример известного алгоритма

АЛГОРИТМ АХО—КОРАСИК

Если $\Sigma = \{a, b\}$ и $w = abaaba$, то минимальный детерминированный автомат $\mathcal{A}_w$, распознающий язык $(a+b)^*abaaba$ имеет следующий вид.



Регулярное множество

Таким образом,

$L(0+1) = \{\varepsilon\}$,

$L(\mathbf{a}_1 + \mathbf{a}_2) = \{a_1, a_2\}$,

$L(\mathbf{a}_1 \cdot \mathbf{a}_2 + \mathbf{a}_2 \cdot \mathbf{a}_1) = \{a_1a_2, a_2a_1\}$,

$L((\mathbf{a}_1 \cdot ((\mathbf{a}_1 + \mathbf{a}_2)^*) + \mathbf{a}_2)^*) = \{a_1, a_2\}^* = L((\mathbf{a}_1 + \mathbf{a}_1)^*)$

Алгебра регулярных выражений

Утверждение .1. Для регулярных выражений справедливы следующие тождества

1. $F + G = G + F$;
2. $F + \mathbf{0} = F$;
3. $F + (G + H) = (F + G) + H$;
4. $F \cdot \mathbf{1} = F$;
5. $\mathbf{1} \cdot F = F$;
6. $F \cdot (G \cdot H) = (F \cdot G) \cdot H$;

РЕГУЛЯРНЫЕ ВЫРАЖЕНИЯ

Регулярным выражением называется всякая формула, которая удовлетворяет следующим требованиям:

1. каждая константа является регулярным выражением;
2. если формулы R_1 и R_2 являются регулярными выражениями, то формулы
   $(R_1 \cdot R_2)$,
   $(R_1 + R_2)$,
   (R_1^*)
   также являются регулярными выражениями.

Утверждение 1. (Продолжение)
Для регулярных выражений справедливы следующие тождества

7 $F \cdot (G + H) = (F \cdot G) + (F \cdot H)$;

8 $(G + H) \cdot F = (G \cdot F) + (H \cdot F)$;

9 $F \cdot \mathbf{0} = \mathbf{0}$;

10 $\mathbf{0} \cdot F = \mathbf{0}$;

11 $F + F = F$;

Доказательство. Самостоятельно

Упростим регулярное выражение

$$(a+b)(a+b)+aa+bb=$$

$$(a+b)\cdot a+(a+b)\cdot b+aa+bbb=$$

$$aa+ba+ab+bb+aa+bb=$$

$$aa+aa+bb+bb+ba+ab=$$

$$aa+bb+ba+ab=$$

$$(a+b)(a+b)$$

Утверждение .2. Для регулярных выражений справедливы следующие тождества

1. $F \cdot F^* = F^* \cdot F$;
2. $(F^*)^* = F^*$;
3. $F^* = \mathbf{1} + F \cdot F^*$;
4. $F^* = (\mathbf{1} + F + FF + \cdots + F^{n-1}) \cdot (F^n)^*$;
5. $(F + G)^* = (F^* \cdot G)^* \cdot F^*$;
6. $(F \cdot G)^* = \mathbf{1} + F(G \cdot F)^* G$.

Доказательство. Самостоятельно

Используя указанные тождества, можно упрощать регулярные выражения и решать уравнения над регулярными выражениями.

Пример. Упростить регулярное выражение

$$(a^*b)^* + (b^*a)^*$$

$$(a^*b)^* + (b^*a)^* = \mathbf{1} + a^*b(a^*b)^* + \mathbf{1} + b^*a(b^*a)^* =$$

$$\mathbf{1} + \mathbf{1} + (a^*b)^*a^*b + (b^*a)^*b^*a =$$

$$\mathbf{1}+(a+b)^*b+(b+a)^*a=\mathbf{1}+(b+a)^*b+(b+a)^*a=$$

$$\mathbf{1} + (b+a)^*(b+a)=\mathbf{1} + (b+a)(b+a)^*=(b+a)^*$$

Упростить регулярное выражение 0^* .

Упростить регулярное выражение $(a+b+ab)^*$.

Упростить регулярное выражение $(a^*b)^*+(b^*a)^*$.

Упростить регулярное выражение $((b^+a)^*b+1)b^*$.

Упростить регулярное выражение $((ab+aab)^*a^*)^*$.

Упростить регулярное выражение $(abbaab+abbaaba)^*$.

Упростить регулярное выражение $(a+b)^*(a(a+b)^*a+b(a+b)^*b)$.

Упростить регулярное выражение $(eb^*(1+c(d+ab^*c)^*a)b^*f)^*eb^*c(d+ab^*c)^*$.

Найти регулярное выражение для языка $\{w \in \{a,b\}^* \mid (|w|_a - |w|_b)4\}$

Найти регулярное выражение для языка $L_1 \cap L_2 \cap L_3$, где $L_1 = (aaab+c+d)^*$, $L_2 = (a^*ba^*ba^*bc+d)^*$

$L_3 = ((a+b)^*c(a+b)^*cd)^*$

Найти регулярное выражение для дополнения языка $(a+b)^*bbb(a+b)^*$ в алфавите $\{a,b\}$.

Найти регулярное выражение для дополнения языка $(ab+ba)^*(1+a+b)$ в алфавите $\{a,b\}$.

Найти регулярное выражение для дополнения языка $(a+b)^*(aab+abaa+abb)(a+b)^*$ в алфавите $\{a,b\}$.

Найти регулярное выражение для дополнения языка $(aa(ab)^*bb(ab)^*)^*$ в алфавите $\{a,b\}$.