编译原理与设计

计卫星　　王贵珍

北京理工大学 计算机学院

词法分析：概览

- 两个关键问题
  - 如何定义语言的词法规则
  - 如何识别输入字符串中的单词

Token定义

正规式

DFA构造

Token识别

确定有限状态机

语言设计和定义

语言实现

词法分析：正规式与正规集

- 设Σ为有限字母表，在Σ上的正规式与正规集可递归定义如下：
  - ε和Φ是Σ上的正规式，它们表示的正规集分别为$\{\varepsilon\}$和Φ;
  - 对任何$a\in\Sigma$, a是Σ上的正规式,它的正规集为$\{a\}$;
  - 若r,s都是正规式，它们的正规集分别为R和S，则$(r|s)$、$(r\cdot s)$、$(r)^*$也是正规式，它们分别表示的正规集是：$R\cup S$，RS，R^*。
  - 有限次使用上述三条规则构成的表达式，称为Σ上的正规式，仅由这些正规式表示的集合为正规集。

词法分析：正规式与正规集

- 相关说明
  - 正规式与相应的正规集是等价的，正规集给出了相应正规式所描述的全部单词（句子）；
  - 正规式的运算结果是正规集；
  - 正规式不是集合，其运算结果正规集是集合，*Φ*是特例；
  - 正规式运算优先级从高到低 “ *()*、***、*•*、*/*” ；
  - 同级运算从左到右。

词法分析：正规式与正规集

- 令$\Sigma=\{0, 1\}$，则 0，1，ε和Φ是Σ上的正规式；

正规式	正规集
$0 \mid 1$	$\{0, 1\}$
$0 \cdot 1$	$\{01\}$
$1 \cdot 0$	$\{10\}$
0^*	$\{\varepsilon, 0, 00, 000, \ldots\}$
1^*	$\{\varepsilon, 1, 11, 111, \ldots\}$
$(0\|1)0^*$	$\{0, 1, 00, 000, \ldots, 10, 100, 1000, \ldots\}$
$(0\|1)01$	$\{001, 101\}$

词法分析：正规式与正规集

- 例:令 $\Sigma=\{A, B, 0, 1\}$
  - $(A|B)(A|B|0|1)^*$ => { 标识符 }
  - $(0|1)(0|1)^*$ => { 二进制数字串 }
  - $1(01)^* = (10)^*1$

正规式 r 所表示的正规集 R 是字母表 Σ 上的语言，称为正则语言，用 $L(r)$ 表示，即 $R=L(r)$。$L(r)$ 中的元素为字符串，称为 $L(r)$ 的句子。

若两个正规式 r 和 s 所表示的语言 $L(r)=L(s)$，则称 r, s 等价，记为 $r=s$。

词法分析：正规式与正规集

- C语言有如下单词
  - int、if、else、for、while · · ·
  - 标识符、无符号整数
  - <、<=、>、>=、==
- 对应的正规式描述为
  - ***int | if | else | for | while***
  - ***<字母>(<字母>|<数字>)****
  - <|<=|>|>=|==

词法分析：正规式与正规集

- 正规式的相关性质

公理/定理	描述
$s \mid t = t \mid s$	| 是可交换的
$s \mid (t \mid r) = (s \mid t) \mid r$	| 是可结合的
$(s\ t)r = s\ (t\ r)$	连接是可结合的
$s\ (t \mid r) = s\ t \mid s\ r$	连接对 | 可分配
$(t \mid r)\ s = t\ s \mid r\ s$	连接对 | 可分配
$\varepsilon s = s \quad (s\varepsilon = s)$	ε是连接的恒等元素
$s^* = (s \mid \varepsilon)^*$	*和ε间的关系
$a^{**} = a^*$	*是幂等的

词法分析：有限状态自动机

Token定义

正规式

DFA构造

Token识别

确定有限状态机

语言设计和定义

语言实现

词法分析：有限状态自动机

- 确定的有限自动机(DFA)
  - DFA : Deterministic Finite Automata
  - 五元组定义：$M = (S, \Sigma, f, S_0, Z)$
    - S： 状态的有限集合，每个元素s_i $(s_i \in S)$ 称为一个状态
    - Σ： 输入字符的有限集合(或有穷字母表)。每个元素是一个输入字符。
    - S_0： M的惟一初态（也称开始状态），$S_0 \in S$
    - Z： M的终态集（或接受状态）$Z \subseteq S$
    - f： 状态转换函数：从$S \times \Sigma \rightarrow S$的部分映射

词法分析：有限状态自动机

- DFA的说明
  - DFA是具有离散输入、输出系统的一个纯数学模型；
  - ~~DFA的技巧在于状态的设置；~~
  - DFA映射的唯一性和初态的唯一性
- 例子
  - 计算机系统
  - 电梯控制系统

词法分析：有限状态自动机

- DFA表示

形式定义

状态转换图

状态矩阵

$M = (\{0,1,2,3\}, \{a,b\}, f, 0, \{3\})$

$f(0, a) = 1$ $\quad$ $f(0, b) = 2$

$f(1, a) = 3$ $\quad$ $f(1, b) = 2$

$f(2, a) = 1$ $\quad$ $f(2, b) = 3$

$f(3, a) = 3$ $\quad$ $f(3, b) = 3$

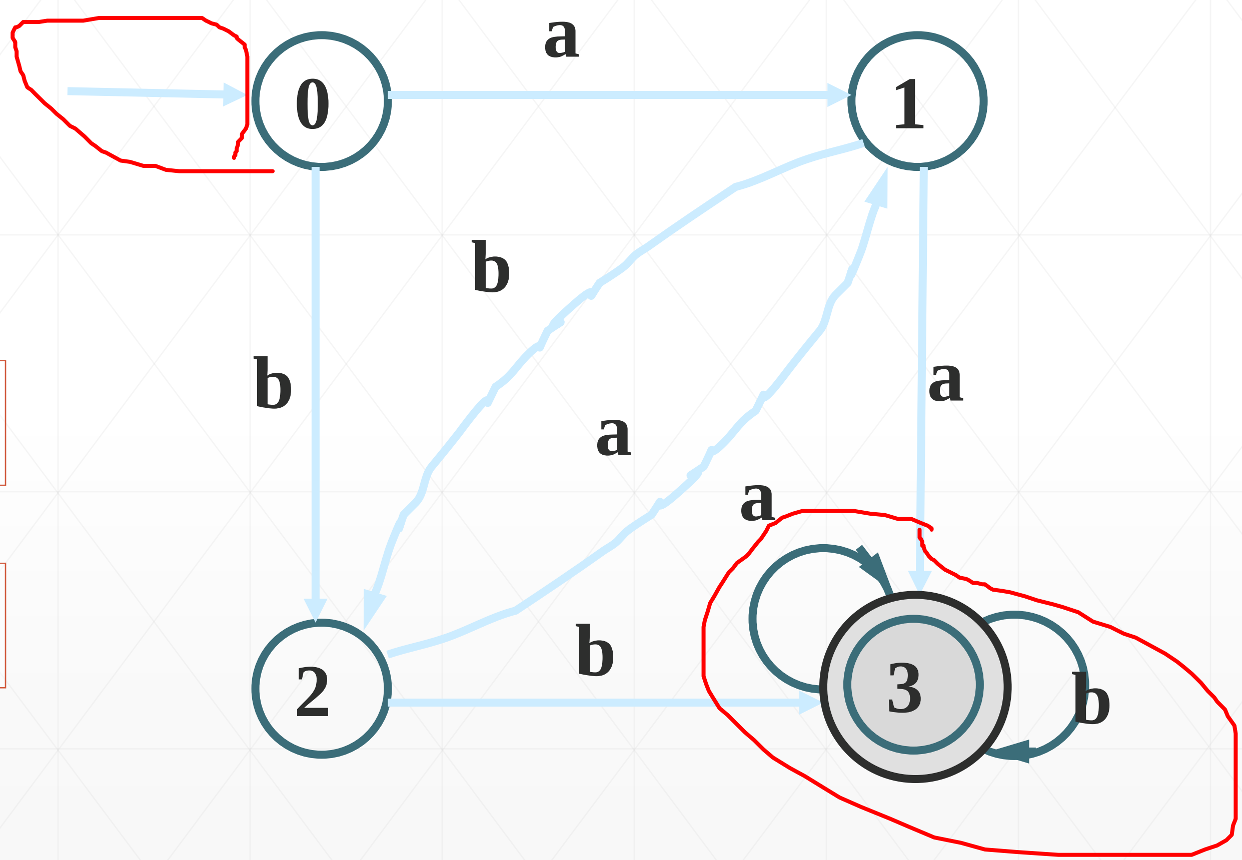
词法分析：有限状态自动机

- DFA表示

形式定义

状态转换图

状态矩阵



词法分析：有限状态自动机

- DFA表示

形式定义

状态转换图

状态矩阵

S

S_0

Z

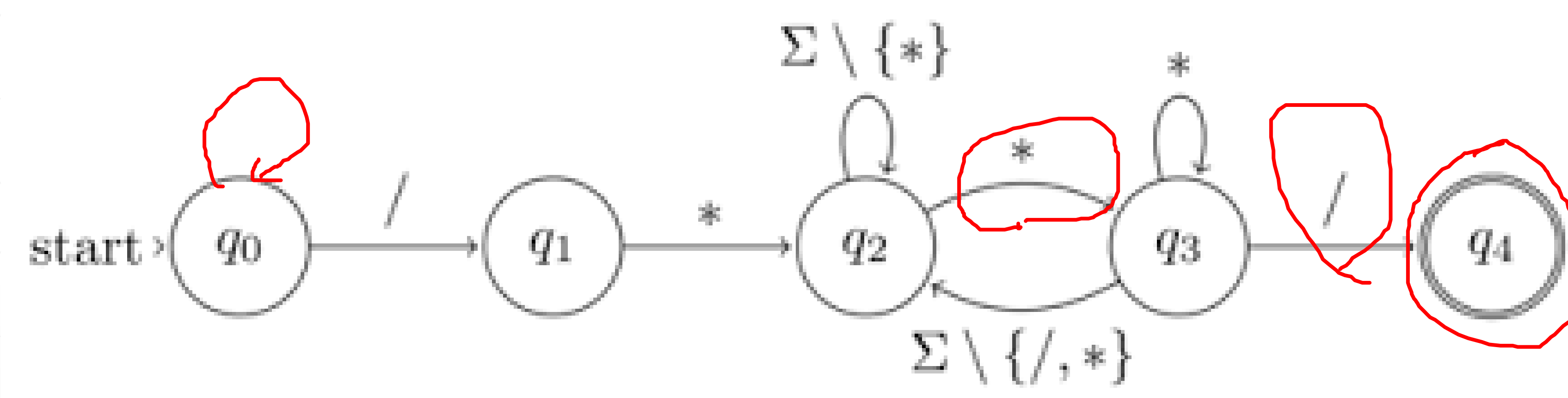
Σ

	a	b
0	1	2
1	3	2
2	1	2
3*	3	3

f

词法分析：有限状态自动机

- DFA:识别C语言块注释的DFA

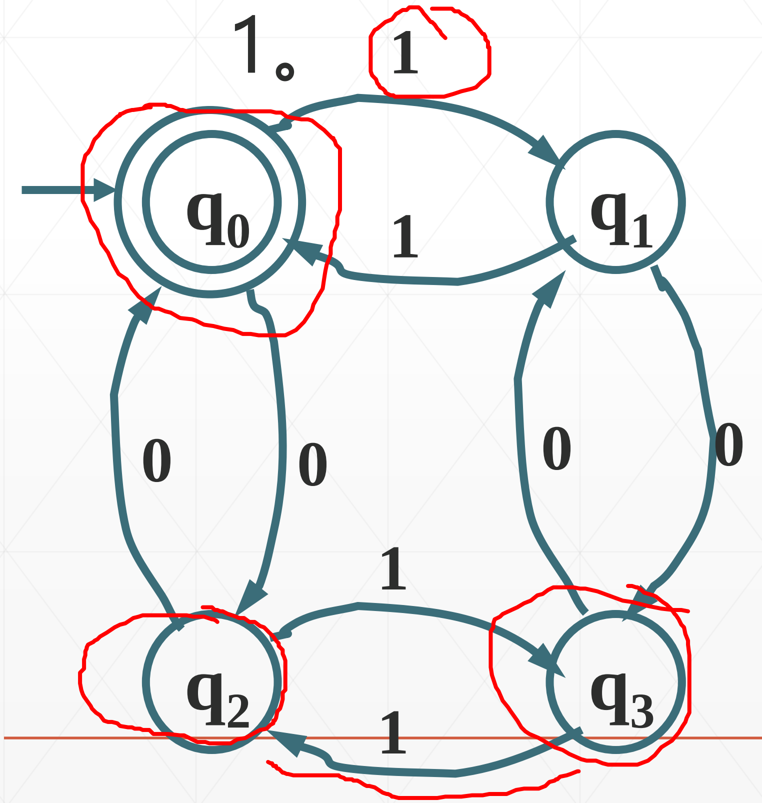



```java
package bit.minisys.minicc.parser;

import java.util.ArrayList;

import bit.minisys.minicc.MiniCCCfg;
import bit.minisys.minicc.util.MiniCCUtil;

/*
 * PROGRAM     --> FUNC_LIST
 * FUNC_LIST   --> FUNC FUNC_LIST | e
 * FUNC        --> TYPE ID '(' ARGUMENTS ')' CODE_BLOCK
 * TYPE        --> INT
 * ARGS        --> e | ARG_LIST
 * ARG_LIST    --> ARG ',' ARGLIST | ARG
 * ARG         --> TYPE ID
 * CODE_BLOCK  --> '{' STMTS '}'
 * STMTS       --> STMT STMTS | e
 * STMT        --> RETURN_STMT
 *
 * RETURN STMT --> RETURN EXPR ';'
 *
 * EXPR        --> TERM EXPR'
 * EXPR'       --> '+' TERM EXPR' | '-' TERM EXPR' | e
 *
 * TERM        --> FACTOR TERM'
 * TERM'       --> '*' FACTOR TERM' | e
 *
 * FACTOR      --> ID
 *
 */
```

```java
//PROGRAM --> FUNC_LIST
public TreeNode program() {
    TreeNode p = new TreeNode(TreeNodeType.TN_TYPE_PROGRAM);
     TreeNode fl = funcList();
     if(fl != null) {
         p.getSubNodes().add(fl);
     }

    return p;
}
```

词法分析：有限状态自动机

- DFA:识别机制
  - 例DFA M1可以识别偶数个0或(和)偶数个1。



对1010:

$$q_0 \xrightarrow{1} q_1 \xrightarrow{0} q_3 \xrightarrow{1} q_2 \xrightarrow{0} \underline{q_0}$$

对11001:

$$q_0 \xrightarrow{1} q_1 \xrightarrow{1} q_0 \xrightarrow{0} q_2 \xrightarrow{0} q_0 \xrightarrow{1} \underline{q_1}$$

词法分析：有限状态自动机

- DFA:识别机制
  - 对于Σ上的任何字α，如果存在一条从初态到某一终态结点的路径，且该路径上所有弧的标记符连接成的字等于α，则称α为DFA M所识别(接受)。
  - 若DFA仅一个状态结点，该状态结既是初态又是终态，则空字集合$\{\varepsilon\}$为DFA M所接受。
  - 一个DFA M所能接受的字的全体记为$L(M)$。

词法分析：有限状态自动机

- DFA:语言和等价关系

∴ 有限状态自动机M所接受的语言为：

$$L(M) = \{\alpha \mid f(S_0, \alpha) \in Z \ \& \ \cancel{\alpha \in V_T^*}\}$$

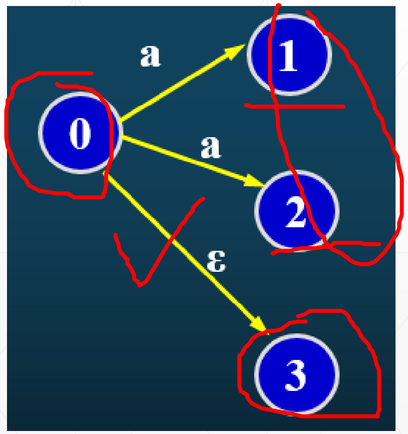
设有FA M 和 FA M'，若$L(M) = L(M')$，则称 M 和 M' **等价**。

词法分析：有限状态自动机

- 实际问题中映射函数往往是多值函数。
  - ~~C语言~~：+/++/+=
- NFA：定义
  - 非确定的有限自动机M (NFA M)五元组

$$M = (S, \Sigma, f, S_0, Z)$$

  - S，Σ^*，Z，S_0同DFA
  - F：状态转换函数：从$S \times \Sigma^* \rightarrow 2^S$的映射



词法分析：有限状态自动机

- NFA：表示

形式定义

状态转换图

状态矩阵

$M=(\{q_0, q_1, q_2, q_3, q_4\}, 0, f, q_0, \{q_2, q_4\})$

$f(q_0, 0) = q_0$ $\quad$ $f(q_0, 0) = q_3$

$f(q_0, 1) = q_0$ $\quad$ $f(q_0, 1) = q_1$

$f(q_1, 0) = \Phi$ $\quad$ $f(q_1, 1) = q_2$

$f(q_2, 0) = q_2$ $\quad$ $f(q_2, 1) = q_2$

$f(q_3, 0) = q_4$ $\quad$ $f(q_3, 1) = \Phi$

$f(q_4, 0) = q_4$ $\quad$ $f(q_4, 1) = q_4$

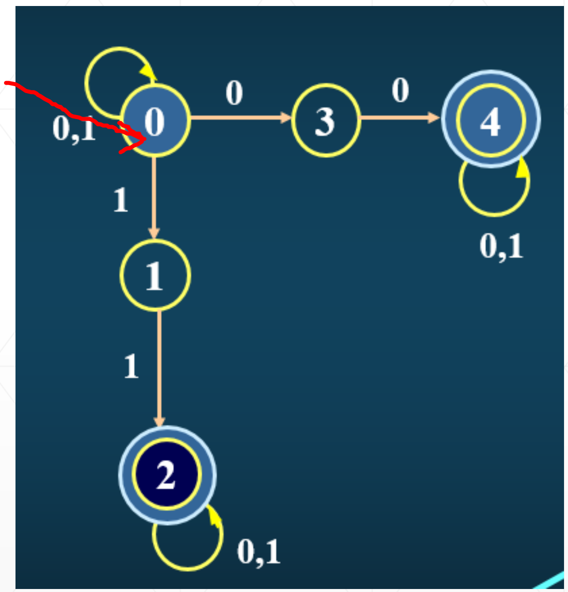
词法分析：有限状态自动机

- NFA：表示

形式定义

状态转换图

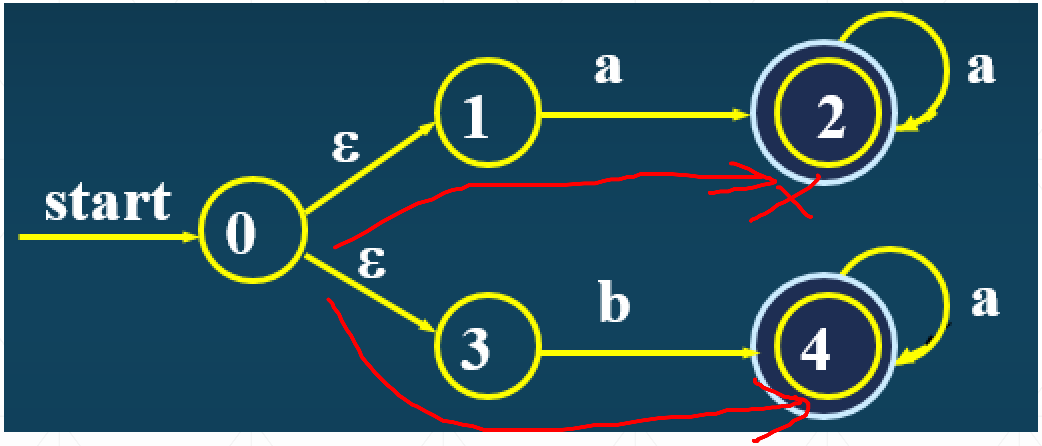
状态矩阵



	0	1
0	{0,3}	{0,1}
1		{2}
2*	{2}	{2}
3	{4}	
4*	{4}	{4}

词法分析：有限状态自动机

- NFA：识别机制



对字符串aaa的接受路径为0，1，2，2，2，接受

路径中边的标记是ε，a，a，a，它们的连接为字符串aaa，ε在连接中消失。

词法分析：有限状态自动机

- DFA与NFA区别

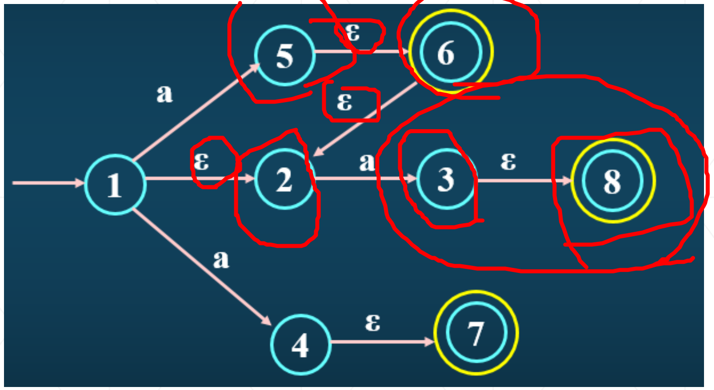
DFA	NFA
Σ	Σ^* ($\varepsilon\in\Sigma^*$)
$f(S\times\Sigma)$	$f(S\times\Sigma^*)$
$f(S\times\Sigma)\rightarrow S$	$f(S\times\Sigma^*)\rightarrow 2^S$

- DFA与NFA等价性

对任何一个NFA *M*，都存在一个DFA *M'*，使 *L(M')=L(M)*。

词法分析：有限状态自动机

- NFA确定化：子集法
  - 消去ε弧：ε-closure(I)
    - ε-closure({5}) = {5, 6, 2}
    - ε-closure({1,5})={1,2,5,6}

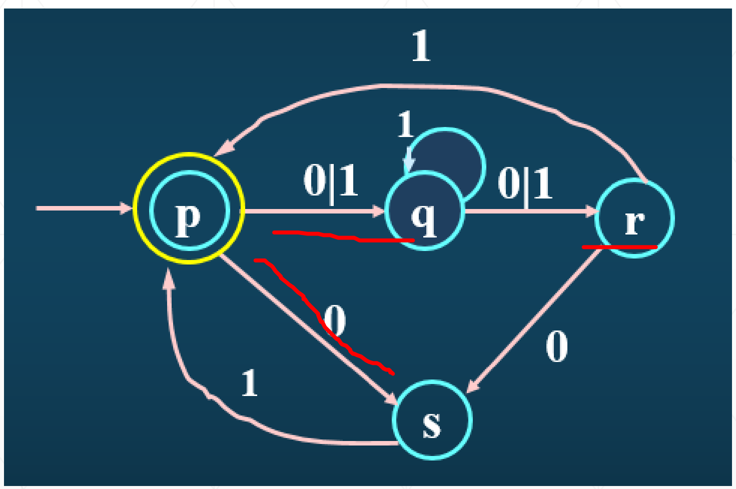


  - 解决映射不唯一问题：求I_a
    - $I_a=\{2, 5\}_a = \{3,8\}$
    - $I_a=\{1\}_a = \{2, 3, 4, 5, 6,7,8\}$

{1, 2}

词法分析：有限状态自动机

- NFA确定化：子集法



	I		I_0	I_1	
0*	{ p }	1	{q, s}	{ q }	2
1	{q, s}	3	{ r }	{ q,r,p }	4
2	{ q }	3	{ r }	{ q,r }	5
3	{ r }	6	{ s }	{ p }	0
4*	{ q,r,p }	7	{ q,r,s }	{ q,r,p }	4
5	{ q,r }	8	{ r,s }	{ q,r,p }	4
6	{ s }			{ p }	0
7	{ q,r,s }	8	{ r,s }	{ q,r,p }	4
8	{ r,s }	6	{ s }	{ p }	0

词法分析：有限状态自动机

- NFA确定化：子集法

	I		I_0	I_1	
0*	{ p }	1	{q, s}	{ q }	2
1	{q, s}	3	{ r }	{ q,r,p }	4
2	{ q }	3	{ r }	{ q,r }	5
3	{ r }	6	{ s }	{ p }	0
4*	{ q,r,p }	7	{ q,r,s }	{ q,r,p }	4
5	{ q,r }	8	{ r,s }	{ q,r,p }	4
6	{ s }			{ p }	0
7	{ q,r,s }	8	{ r,s }	{ q,r,p }	4
8	{ r,s }	6	{ s }	{ p }	0

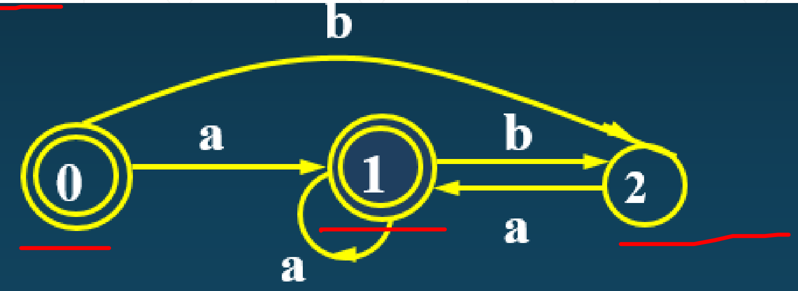
state	0	1
0*	1	2
1	3	4
2	3	5
3	6	0
4*	7	4
5	8	4
6		0
7	8	4
8	6	0

词法分析：有限状态自动机

- DFA最小化：划分法
  - **无关状态**：如果从DFA M的初态开始，任何输入序列都不能到达的那些状态称为无关状态。
  - **等价状态**：设DFA M的两个不同状态 q1，q2，如果对任意输入字符串ω，从q1，q2状态出发，总是同时到达接收状态或拒绝状态之中，称q1，q2是等价的。
  - 如果DFA M既没有无关状态，且没有彼此等价的状态，则称DFA M是规约的(即最小的DFA M)。

词法分析：有限状态自动机

- DFA最小化：划分法



Step1: 形成初始划分。划分为终态集和非终态集。

$P_0 = (\{0,1\}, \{2\})$ 基本划分

考察：$\{0,1\}_a = \{1\} \subset \{0,1\}$

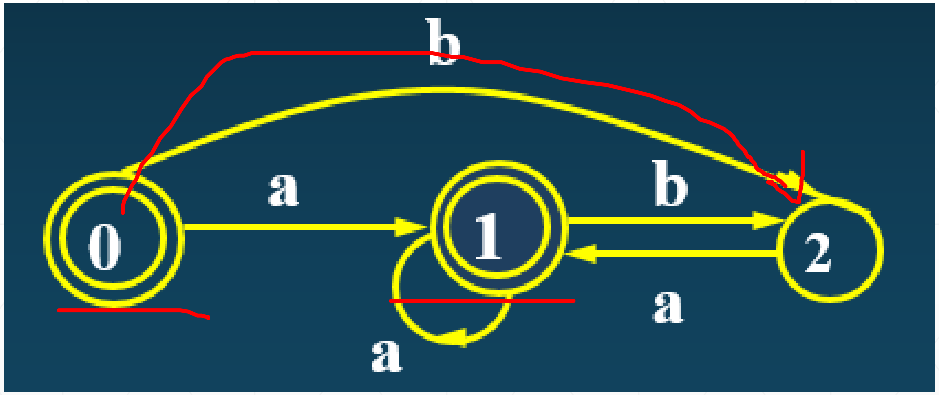
$\{0,1\}_b = \{2\} \subset \{2\}$

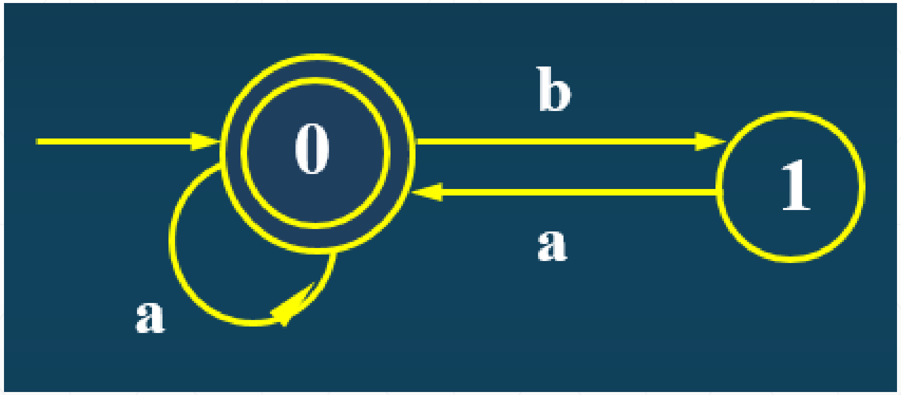
不可对{ 0,1 }再分

Step2: 重新命名。令 {0,1}为0，令{2}为1。

词法分析：有限状态自动机

- DFA最小化：划分法





词法分析：有限状态自动机

- DFA最小化：划分法

step1: 形成初始划分

$\pi = \{ Z, K—Z \}$ // K是M的所有状态

step2: 对当前的划分 $\pi = \{ I_1, I_2, \ldots, I_m \}$中的每个状态集$I_i$考察是否可区分，可区分则进行划分，形成新的划分$\pi_{new}$。

step3：若$\pi_{new} \neq \pi$，则将π_{new}作为π重复step2;

step4：对所得的最后划分π重新命名。

词法分析：正规式与FA等价性

Token定义

正规式

DFA构造

Token识别

确定有限状态机

语言设计和定义

语言实现

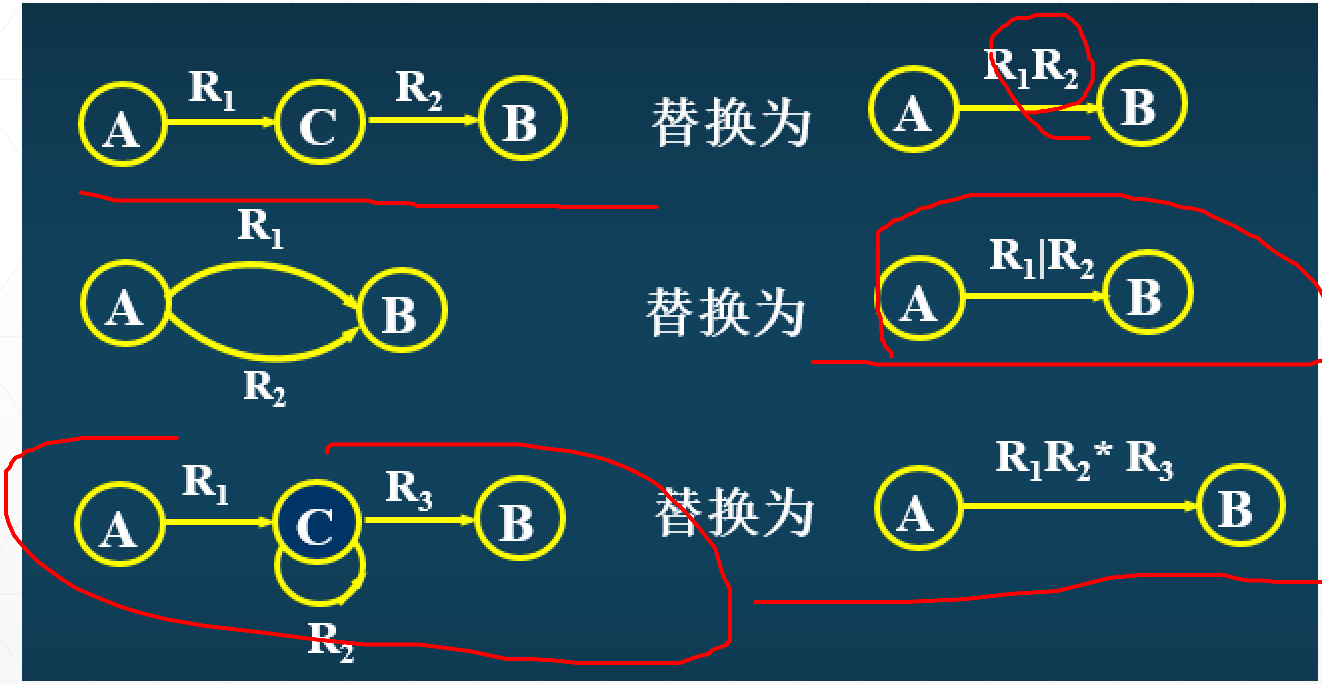
词法分析：正规式与FA等价性

- 定理
  - 字母表∑上的确定的有限自动机M所接受的语言L(M)是∑上的一个正规集；
  - 对于∑上的每一个正规式 r，存在一个∑上的非确定有限自动机M，使得：L(M)=L(r)。

∑上的单词集V ∈∑*是正规的，当且仅当存在∑上的DFA M，使得V=L(M)。

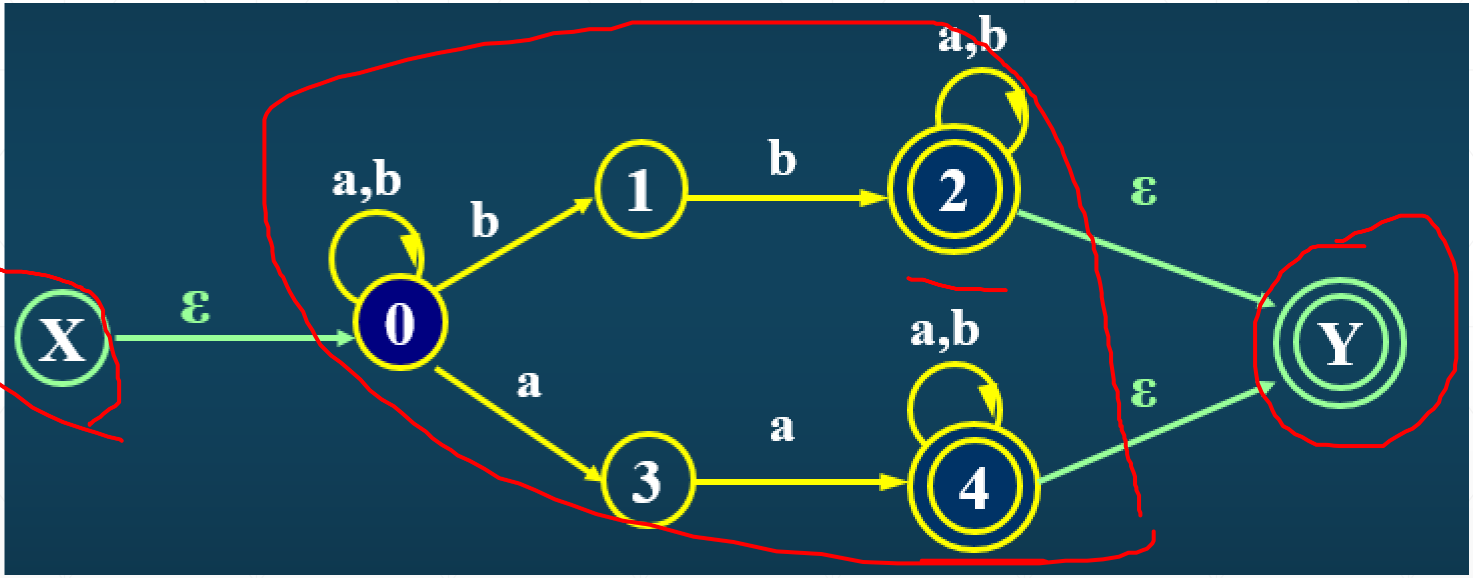
词法分析：正规式与FA等价性

- FA转RE：规则



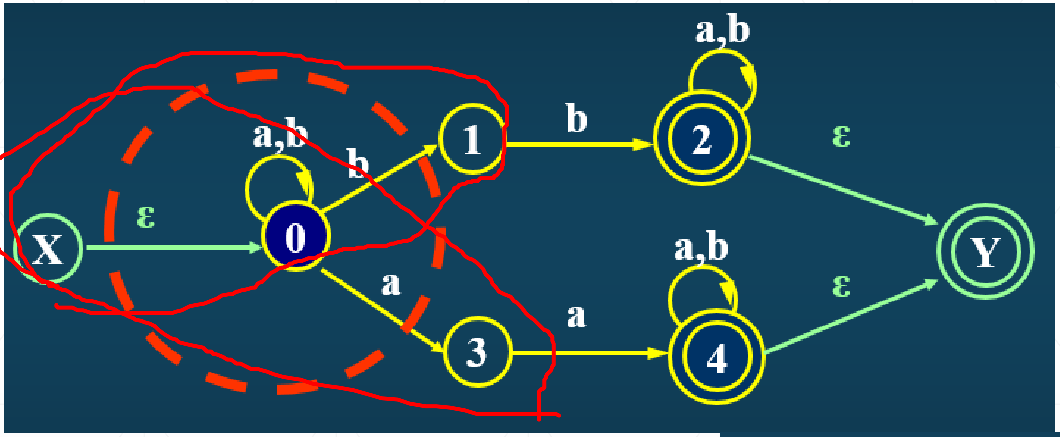
词法分析：正规式与FA等价性

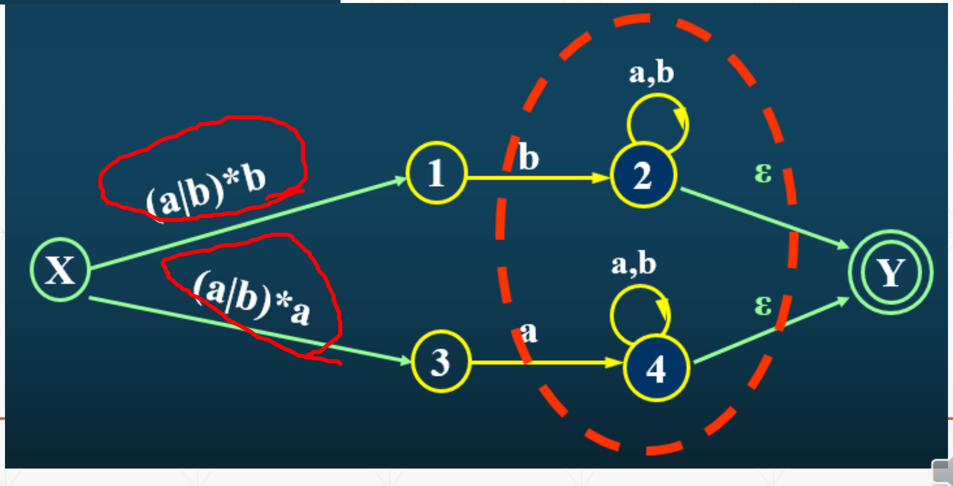
- FA转RE：拓广



词法分析：正规式与FA等价性

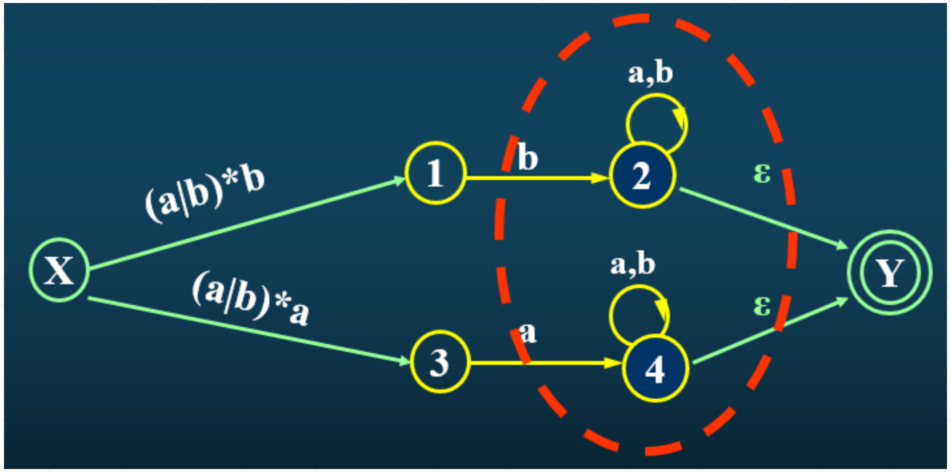
- FA转RE：替换

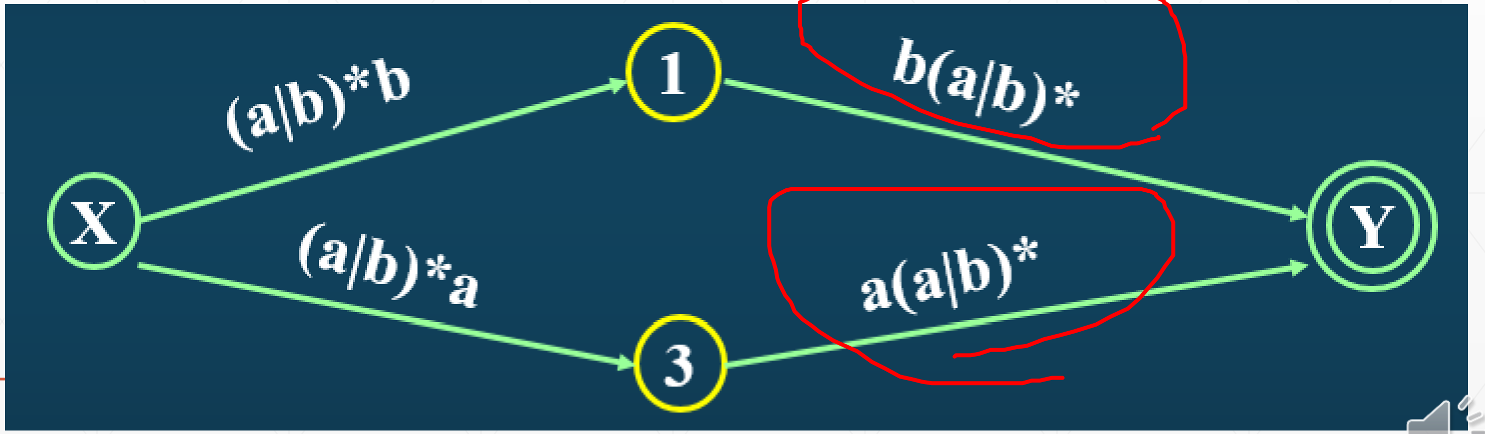




词法分析：正规式与FA等价性

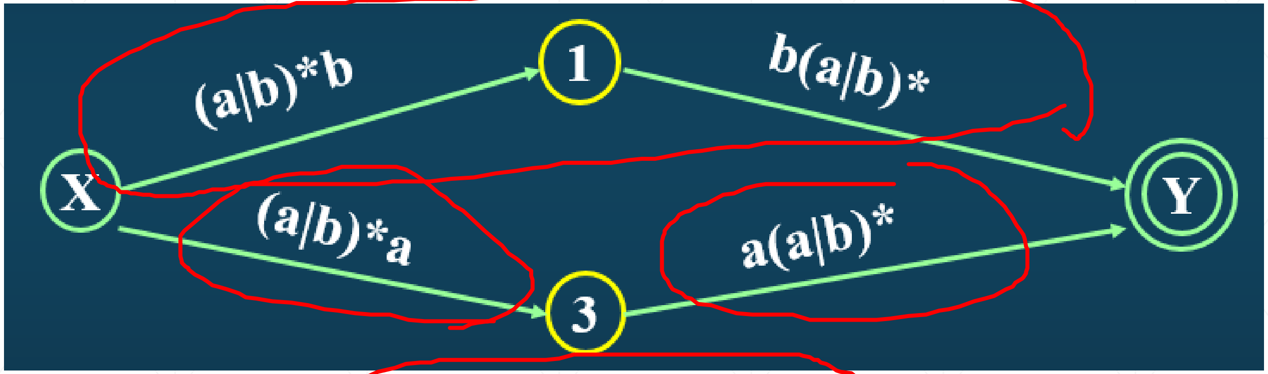
- FA转RE：替换

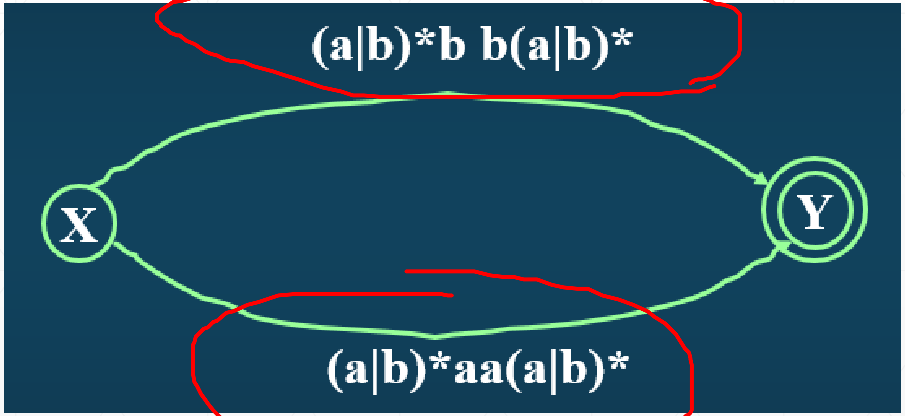




词法分析：正规式与FA等价性

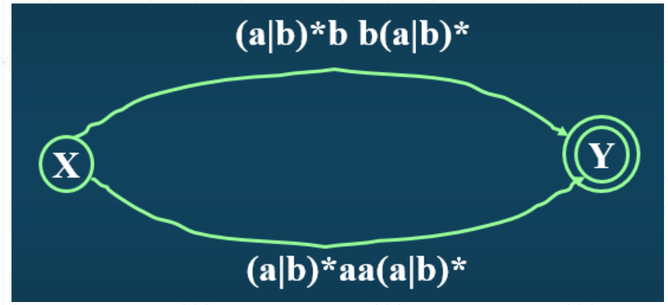
- FA转RE：替换

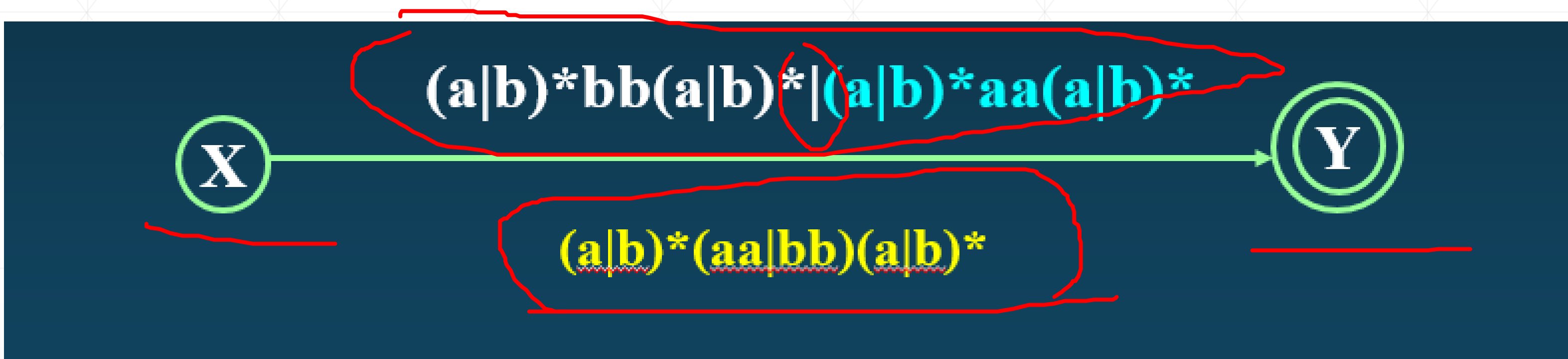




词法分析：正规式与FA等价性

- FA转RE：替换





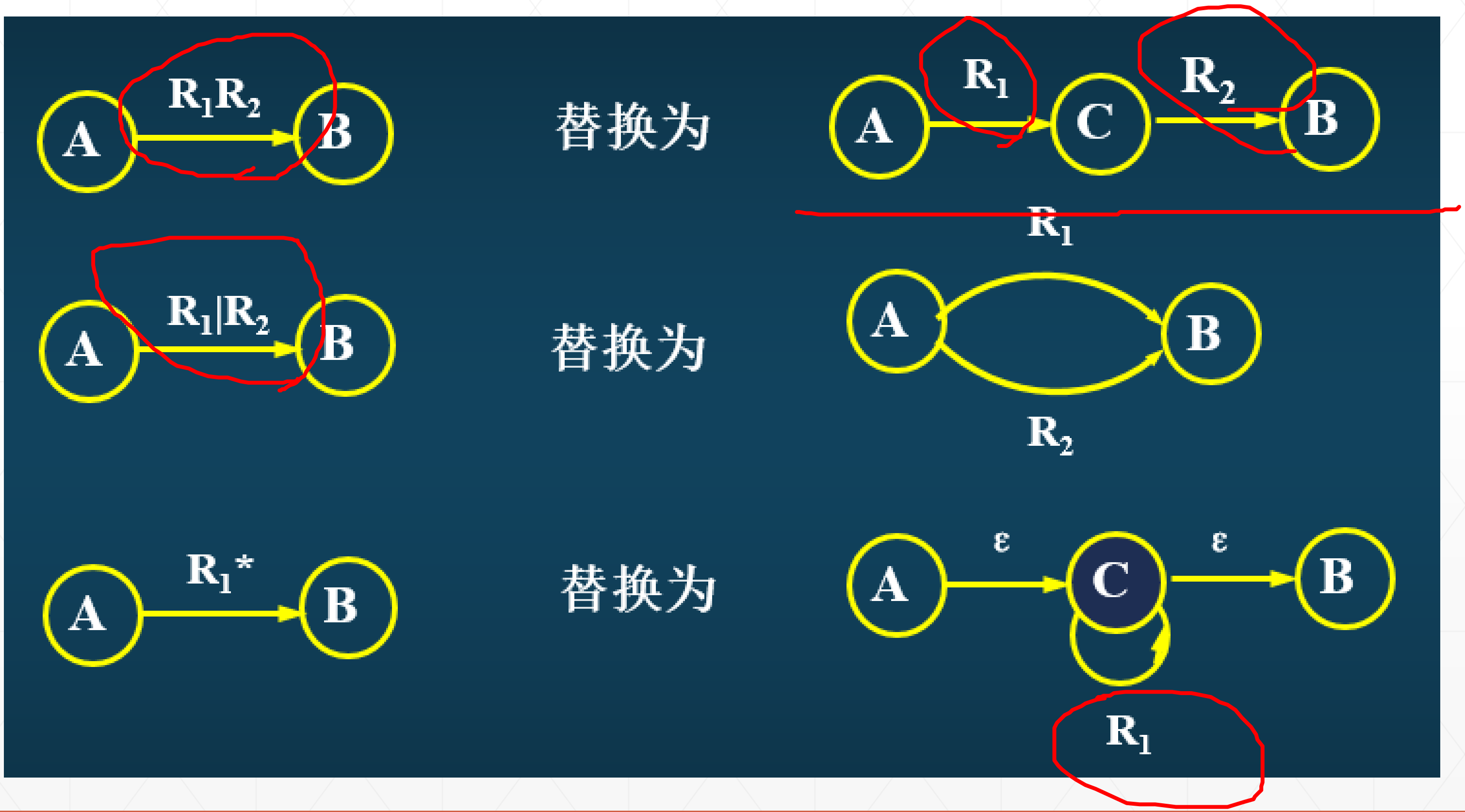
注意：(对消弧、消结过程)

实际是正规式的**合成**过程，其优先级：

*，•，|　　（高 ⟶ 低）

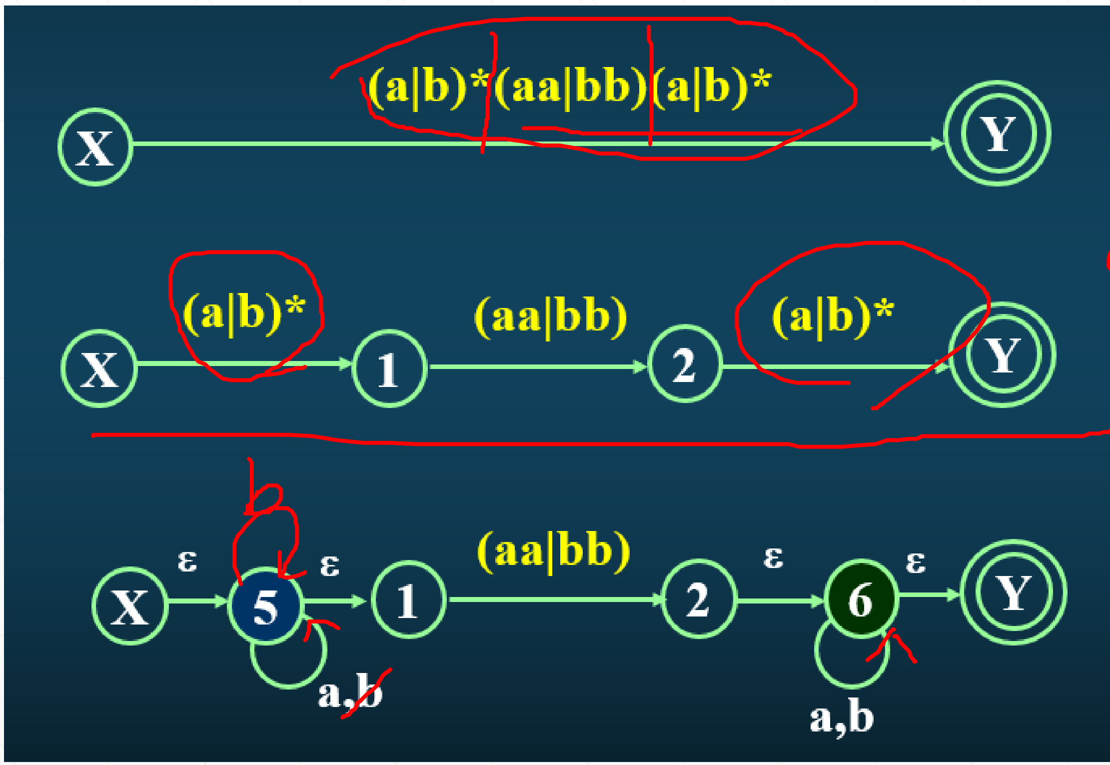
词法分析：正规式与FA等价性

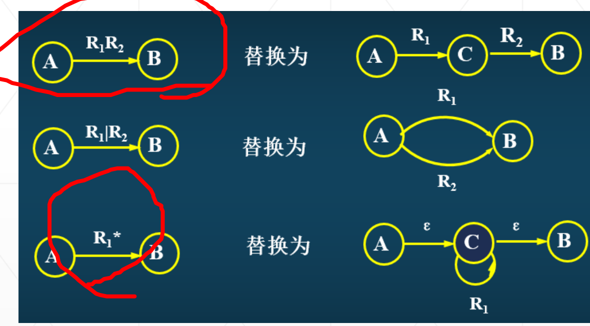
- RE转FA：规则



词法分析：正规式与FA等价性

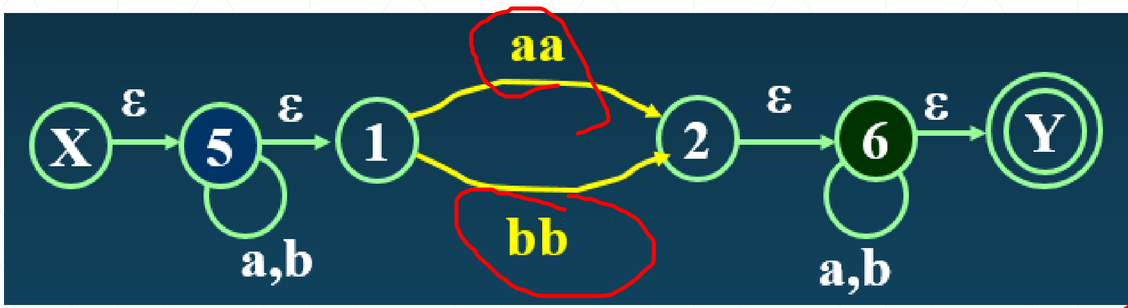
- RE转FA：拓广与替换

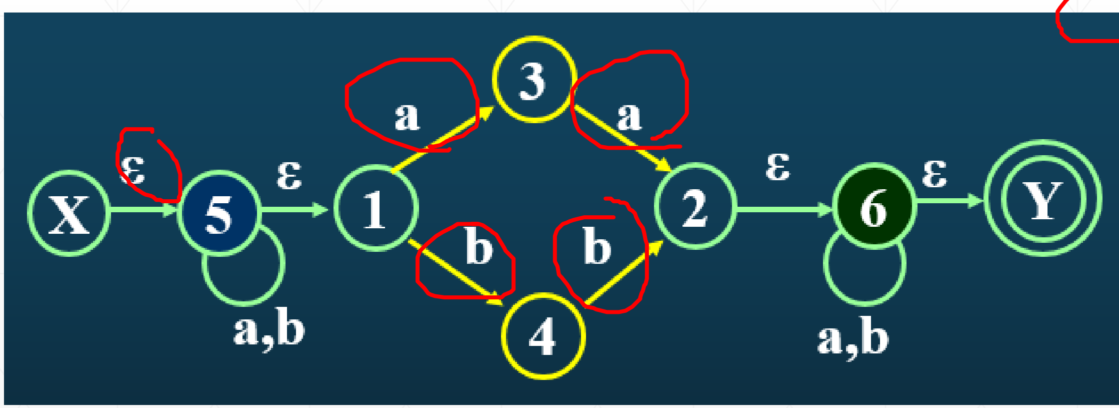




词法分析：正规式与FA等价性

- RE转FA：拓广与替换





词法分析：正规式与FA等价性

- RE转FA：确定化

	I		I_a	I_b	
0	{X,1,5 }	1	{5, 3, 1}	{5, 4, 1 }	2
1	{5, 3, 1 }	3	{5, 3, 1,2,6,Y }	{5, 4, 1 }	2
2	{5, 4, 1}	1	{5, 3, 1 }	{5,4, 1,2,6,Y }	4
3*	{5, 3, 1,2,6,Y }	3	{5, 3, 1,2,6,Y }	{5,4, 1,6,Y }	5
4*	{5,4, 1,2,6,Y }	6	{5,3, 1,6,Y }	{5,4, 1,2,6,Y }	4
5*	{5,4, 1,6,Y }	6	{5,3, 1,6,Y }	{5,4, 1,2,6,Y }	4
6*	{5,3, 1,6,Y }	3	{5, 3, 1,2,6,Y }	{5,4, 1,6,Y }	5

词法分析：正规式与FA等价性

- RE转FA：最小化

state	a	b
0	1	2
1	3	2
2	1	4
3*	3	5
4*	6	4
5*	6	4
6*	3	5

state	a	b
0	1	2
1	3	2
2	1	3
3*	3	3