编译原理

Compiler Construction Principles

朱 青

信息学院计算机系，

中国人民大学，

zqruc2012@aliyun.com

第2章:词法分析 (Lexical Analysis)

- ⌘2.1 词法分析程序的功能
- ⌘2.2 词法分析器的设计
- ⌘2.3 正规表达式 (Regular Expression)
- ⌘2.4 有限自动机
- ⌘2.5 词法分析器的自动生成

2.4 有限自动机

- ⌘2.4.1 确定有限自动机（DFA）
- ⌘2.4.2 非确定有限自动机（NFA)确定化
- ⌘2.4.3 具有ε-转移的NFA M确定化
- ⌘2.4.4 DFA的化简
- ⌘2.4.5 正规式与有限自动机的等价性
- ⌘2.4.6 正规文法与有限自动机

2.4.1 确定有限自动机（DFA）

有限自动机

正规表达式 ==== 有限自动机

DFA NFA 正规式

2.4.1 确定有限自动机（DFA）

- DFA的定义：一个确定的有限自动机（DFA）

M是一个五元式

$$M=(S, \Sigma, f, s_0, Z)$$

其中

1　S是一个有限集，它的每个元素称为一个状态；

2　Σ 是一个有穷字母表，它的每个元素称为一个输入字符。

3　f 是一个从S X Σ 至S的（单值）部分映照　f(s，a）= s’。表示：当前状态是s，输入字符是a时，下一个状 态是s’，s’叫s 的后继状态。

4　$s_0 \in S$，是唯一的初态。

5　$Z \subseteq S$，是终态集（可空）。

- 一个DFA可以表示为一个矩阵。

f (s , a) 值-------状态转换矩阵值 。

行-------状态,

列-------输入字符。

- **一个DFA也可以表示为一张（确定的）状态转换图。**

例题 2.4-1 DFA M 可用一个矩阵表示

DFA M=（{0， 1， 2， 3}，{a, b}, f , 0 , {3}).

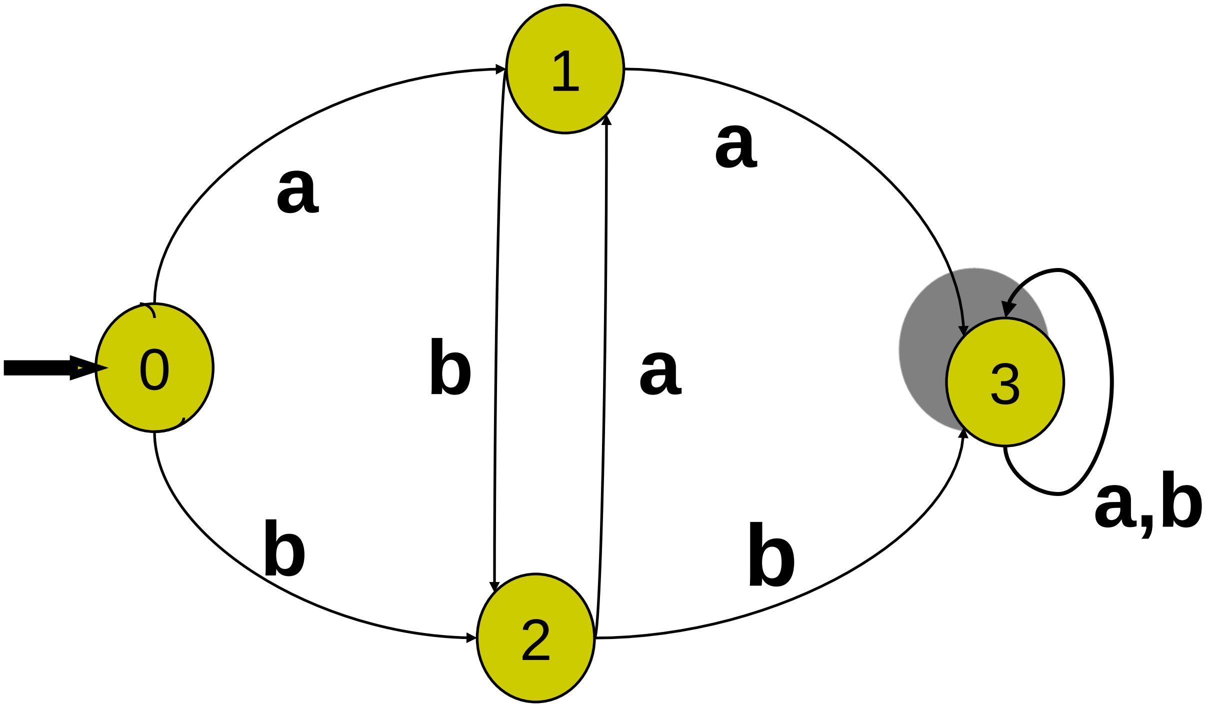
其中f:

f(0,a)=1	f(0,b)=2
f(1,a)=3	f(1,b)=2
f(2,a)=1	f(2,b)=3
f(3,a)=3	f(3,b)=3

矩阵表示：

状态 \ 字符	a	b
0	1	2
1	3	2
2	1	3
3	3	3

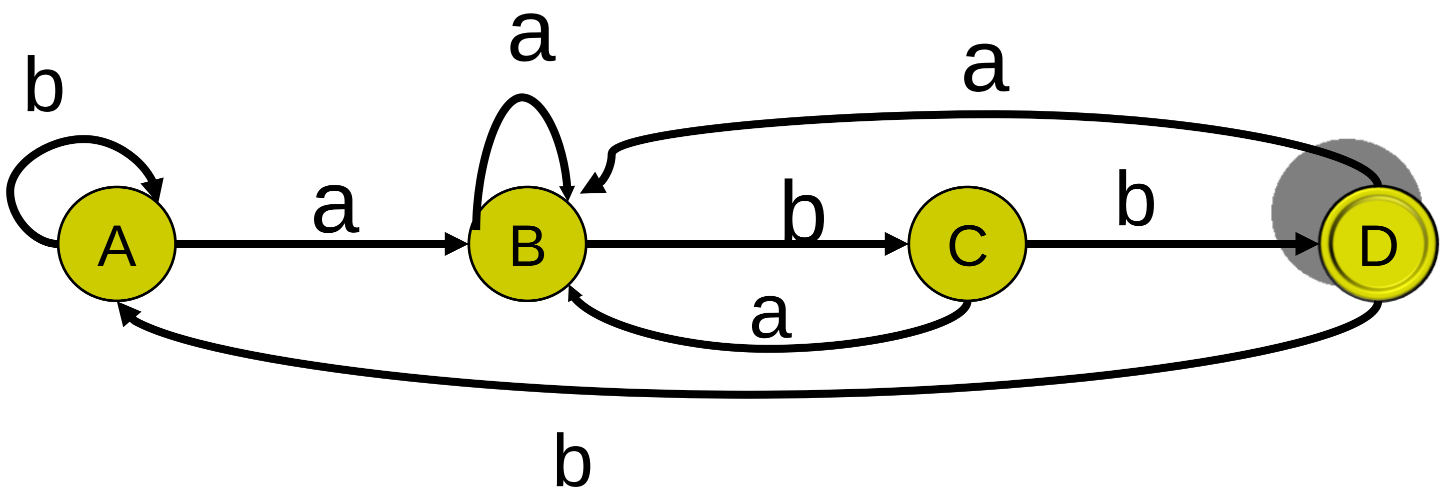
9


1
a
a
0
b
a
3
a,b
b
b
2

例题 2.4-2 DFA M 可用一张（确定的）状态转换图表示。

DFA M=({A,B,C,D},{a,b},Θ,A,{D})

Θ(A,a)=B	Θ(A,b)=A
Θ(B,a)=B	Θ(B,b)=C
Θ(C,a)=B	Θ(C,b)=D
Θ(D,a)=B	Θ(D,b)=A


b
a
a
a
b
b
A
B
C
D
a
b

用DFA识别单词符号：

在DFA状态转换图中，存在一条从初态到终态的通路。该路上各弧的标记字符依此连结构成的字与S相同，则称S能被该DFA识别。

一个DFA M 所能识别的所有的字的集合记为：L（M）。

定理

- Σ上的一个 字集 V $\subset$Σ* 是正规的，当且仅当存在 Σ上的DFA M，使得V=L（M）。

- DFA 的确定性表现在 f 是一个从 S X Σ 至 S的单值函数。即唯一确定了下一个状态。

2.4.2 非确定有限自动机（NFA)确定化

- **NFA的定义：**

  一个非确定的有限自动机（**NFA**）**M**

是一个五元式

$$M=（S，\Sigma，f，S_0，Z）$$

其中

1 S是一个有限集，它的每个元素称为一个状态；

2 Σ 是一个有穷字母表，它的每个元素称为一个输入字符。

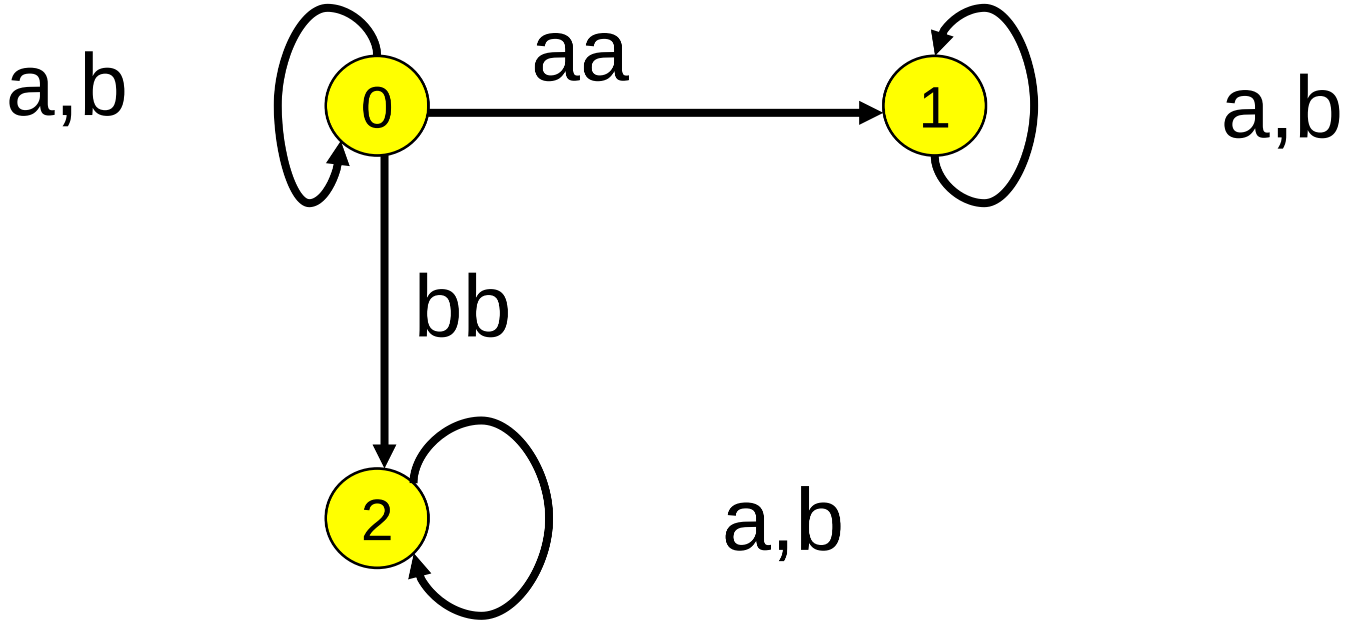
3 f是一个从S X Σ 至S的子集映照。

即 $f: S \times \Sigma^* \longrightarrow 2^S$。

4 $S_0 \subset S$，是一个非空初态集。

5 $Z \subset S$，是终态集（可空）。

NFA实例：



- 一个 NFA 可以表示为一个矩阵。

  f (s , a) 值------状态转换矩阵。

  行------状态,

  列------输入字符

- 一个NFA也可以表示为一张（确定的）状态转换图。

例题 2.4-3 NFA M 可用一 张状态转换图表示。

NFA M=({A, B, C, D}, {a, b}, δ, A, {D})

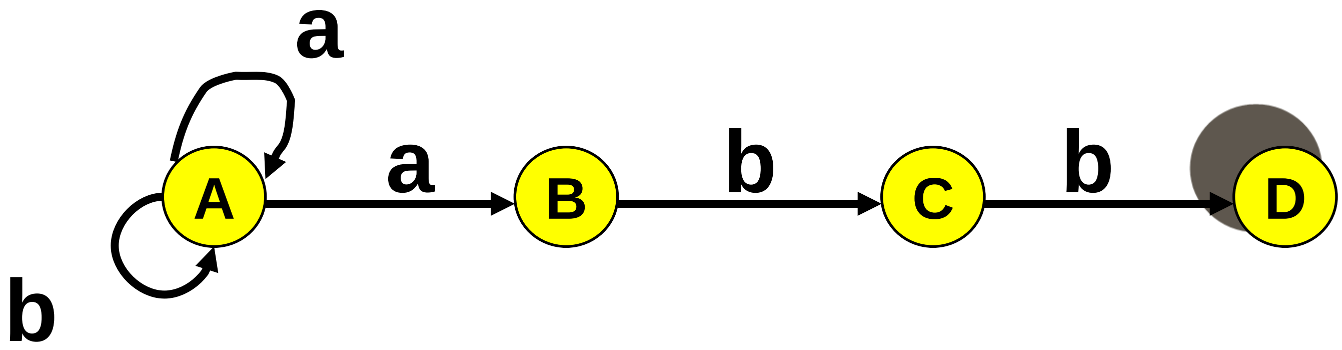
$\delta(A, a)=\{A, B\}$ $\delta(A, b)=\{A\}$

$\delta(B, a)=\Phi$ $\delta(B, b)=\{C\}$

$\delta(C, a)=\Phi$ $\delta(C, b)=\{D\}$

$\delta(D, a)=\Phi$ $\delta(D, b)= \Phi$

状态图：

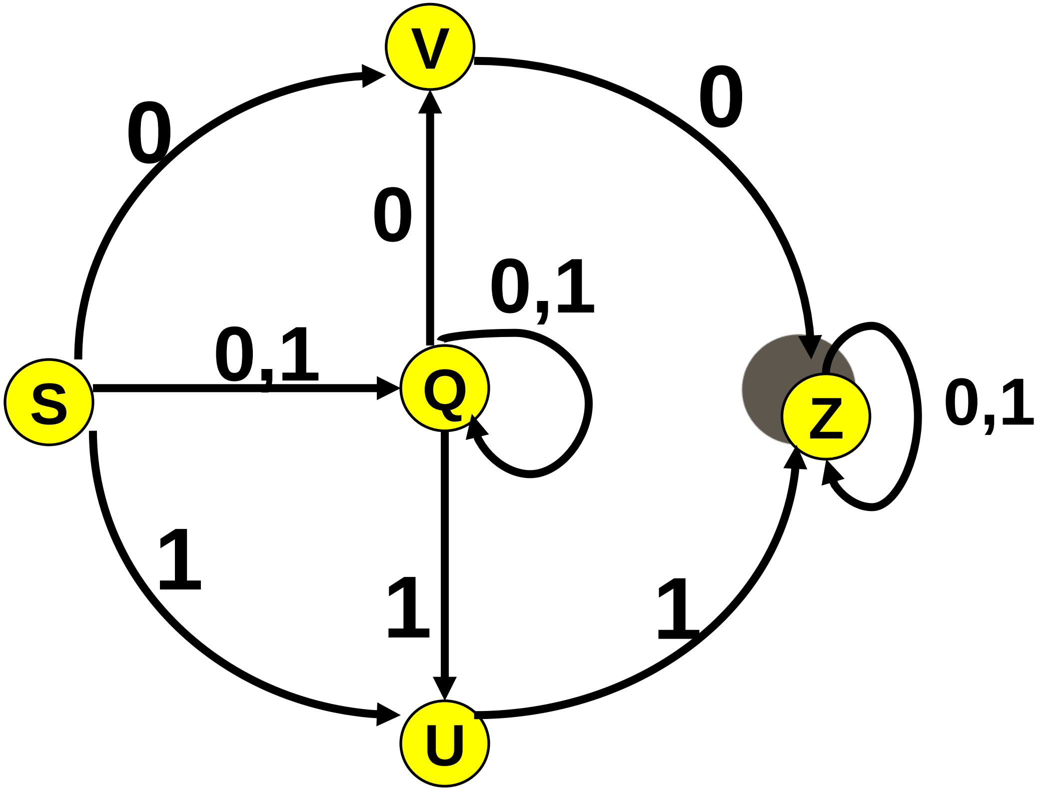


例题 2.4-4 NFA M 实例 。

NFA $M=(\{S,Q,U,V,Z\},\{0,1\},\delta,\{S\},\{Z\});$

$\delta(S,0)=\{V,Q\}$	$\delta(S,1)=\{U,Q\}$
$\delta(U,0)=\Phi$	$\delta(U,1)=\{Z\}$
$\delta(V,0)=\{Z\}$	$\delta(V,1)=\Phi$
$\delta(Q,0)=\{V,Q\}$	$\delta(Q,1)=\{U,Q\}$
$\delta(Z,0)=\{Z\}$	$\delta(Z,1)=\{Z\}$

NFA的状态转换图：



图中某些状态射出两条具有相同标记的弧S,Q.

识别单词:

在NFA中识别ε有两种情况:

（1） 终态与初态是同一点.

（2） 从初态到终态,所走的弧上都是ε.

定理：对任何一个NFA M，都存在一个DFA M’，使得L（M’）=L（M）。

证明的思想是由M出发构造等价的M’，办法是让M’的状态对应于M的状态集合。

即若 f(q,a)={q1,q2,…,qk},

我们将集合 {q1,q2,…,qk}作为一个整体看作M’中的一个状态，即S’中的一个元素。

例题2.4-5：设：

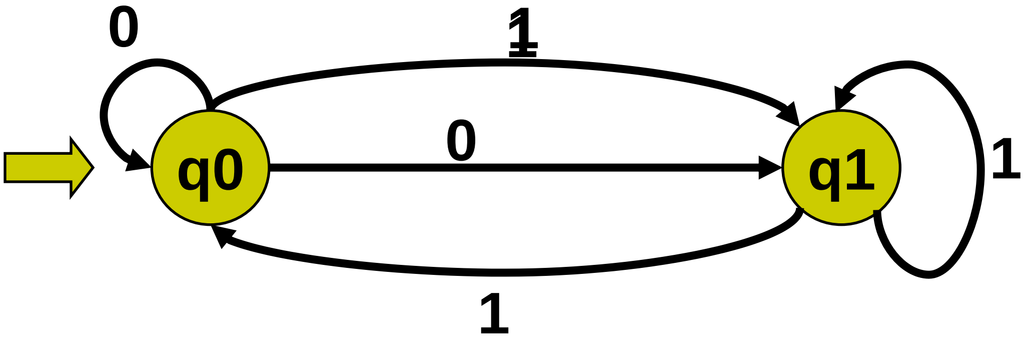
NFA M=({0,1},{q0,q1},f,q0,{q1}).

f(q0,0)={q0,q1}   f(q0,1)={q1}

f(q1,0)= Φ   f(q1,1)={q0,q1}

构造与M 等价的DFA M’.

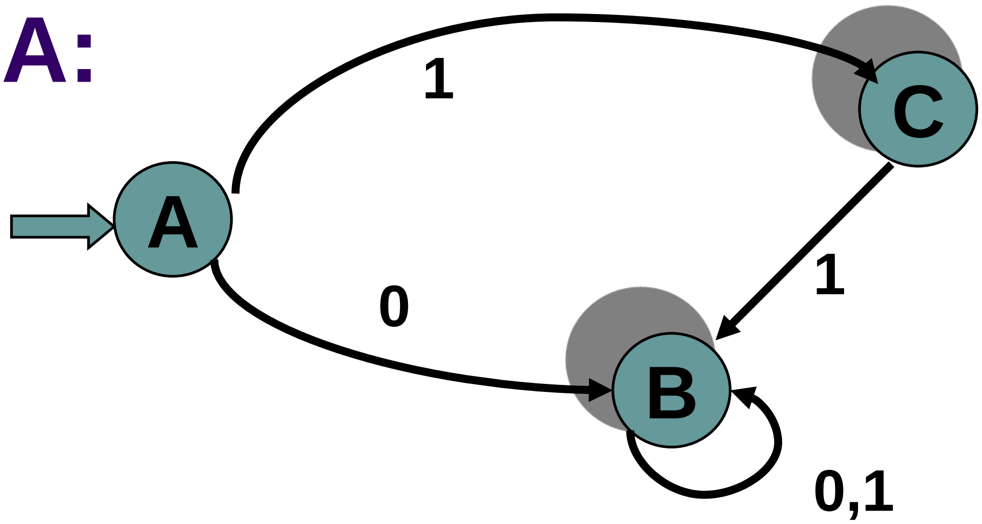
状态图：



解：令DFA M’ =（{0, 1}, Q’, f’, q0’. F’）

	0	1
A.{q0}	{q0,q1}	{q1}
B.{q0,q1}	{q0,q1}	{q0,q1}
C.{q1}	Φ	{q0,q1}

DFA:

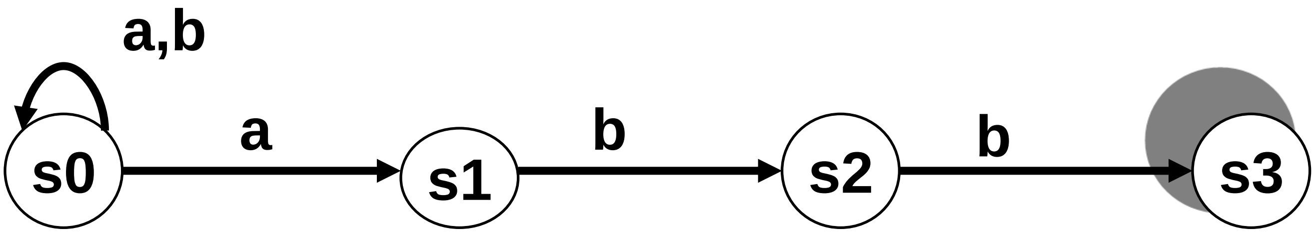


28

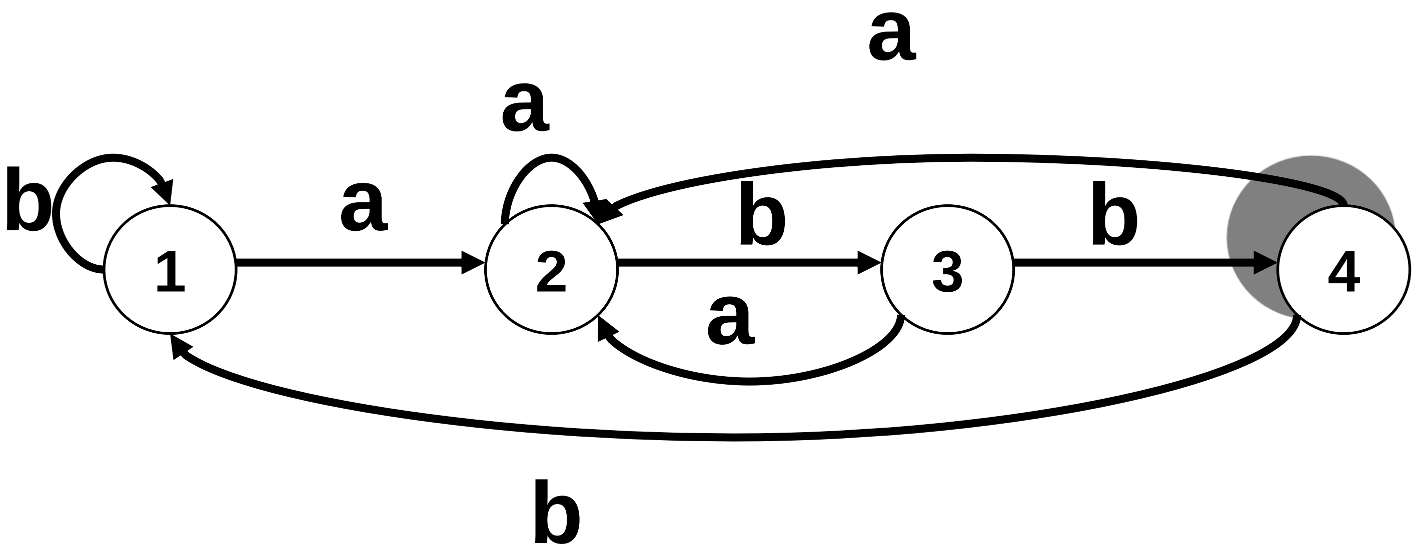
例2.4-6: 设有

NFA M=({S0,S1,S2,S3},{a,b},f,S0,{S3})

状态转换图:



	a	b
1. {s0}	{s0,s1}	{s0}
2. {s0,s1}	{s0,s1}	{s0,s2}
3. {s0,s2}	{s0,s1}	{s0,s3}
4. {s0,s3}	{s0,s1}	{s0}


a
a
b
a
b
b
1
2
3
4
a
b

2.4.3 具有ε–转移的NFA M确定化

定义1：状态集S的子集I的ε -闭包，

即：ε–closure(I):

(1) 若s∈I，则 s ∈ ε-closure(I).

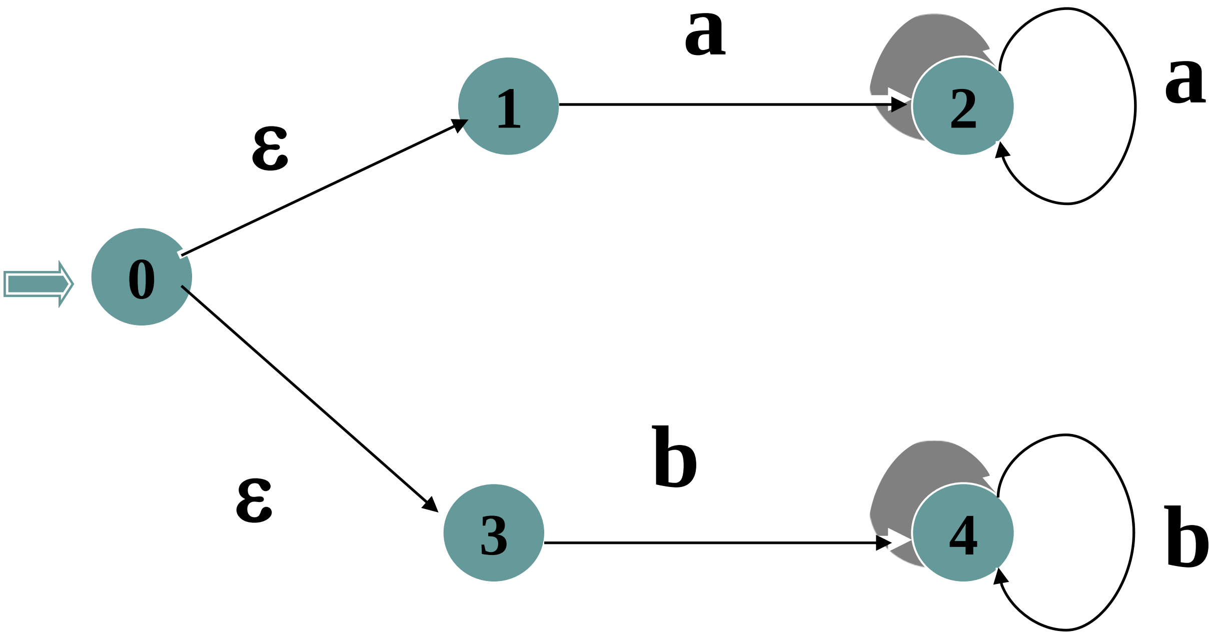
(2) 若s ∈ I，则从s出发经过若干条ε弧所到达的状态s’，s ’ ∈ ε-closure(I).

例题2.4-7 具有ε转移的识别正规式：
aa*|bb*的非确定有限自动机.

NFA M=({0,1,2,3,4},{a,b},f,{0},{2,4})

f(1,a)={2}　　f(3,b)={4}
f(2,a)={2}　　f(4,b)={4}
f(0,ε)={1,3}

状态转换图：



ε-closure(0) ={0, 1, 3}

定理1：对任何一个具有ε–转移的NFA M，一定存在一个不具有ε–转移的 NFA M’, 使

L（M’）=L（M）。

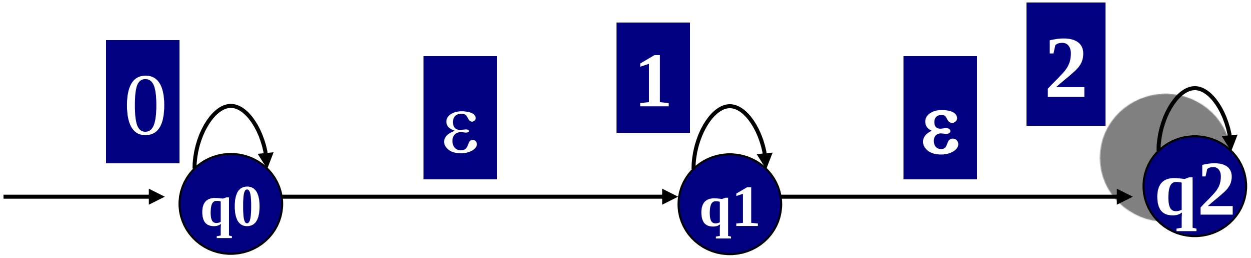
例题2.4-8：

设NFA M=（Σ, Q, f, q0, F）

Σ={0,1,2}, Q={q0,q1,q2},

F={q2}

状态图：



从M出发构造一个不具有 ε– 转移的 NFA M’，使得 L（M’）= L（M）。

解：令M’=（ Σ, Q’, f’,q0, F’），

其中 Σ, Q’,q0 的意义同M中完全一样。

1) F'包含M的终态集F，其次若M中从q0出发有一条到达某终态的 ε 道路，则将q0 加在F'中。

$$F' = \begin{cases} F \cup \{q0\}, & \text{若 } \varepsilon\text{-closure}(q0) \text{ 包含F的一个状态} \\ F, & \text{否则。} \end{cases}$$

终态集　$F' = \{q0,q1,q2\}$,

ε-closure$(q0)=\{q0,q1,q2\}$

ε -closure$(q1)=\{q1,q2\}$

2)

$f'(q,a)=\{q' \mid q'$为从q出发先经若干 ε 箭弧，接着经一个标记为a的箭弧，再经若干ε 箭弧组成的道路所 能到达的状态。}

f’(q0 ,0) = {q0,q1,q2}

f’(q0 ,1) = {q1,q2}

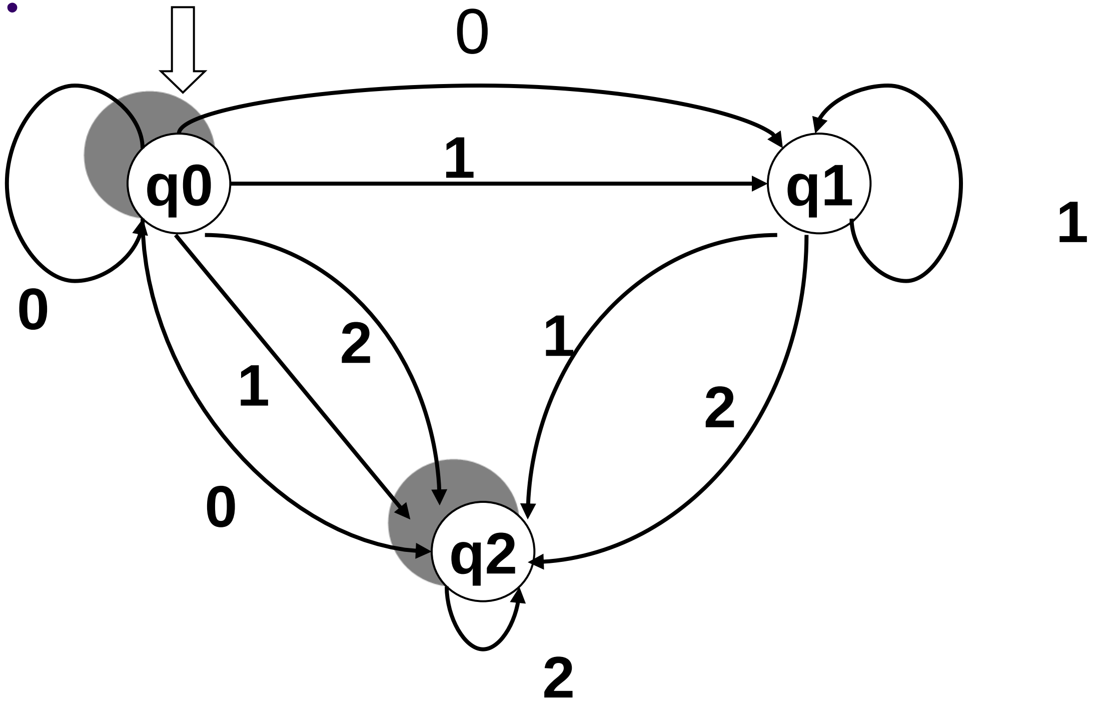
f’(q0 ,2) = {q2}

f’(q1 ,1) = { q1,q2}

f’(q1 ,2) = { q2}

f’(q2 ,2) = { q2}

状态图：

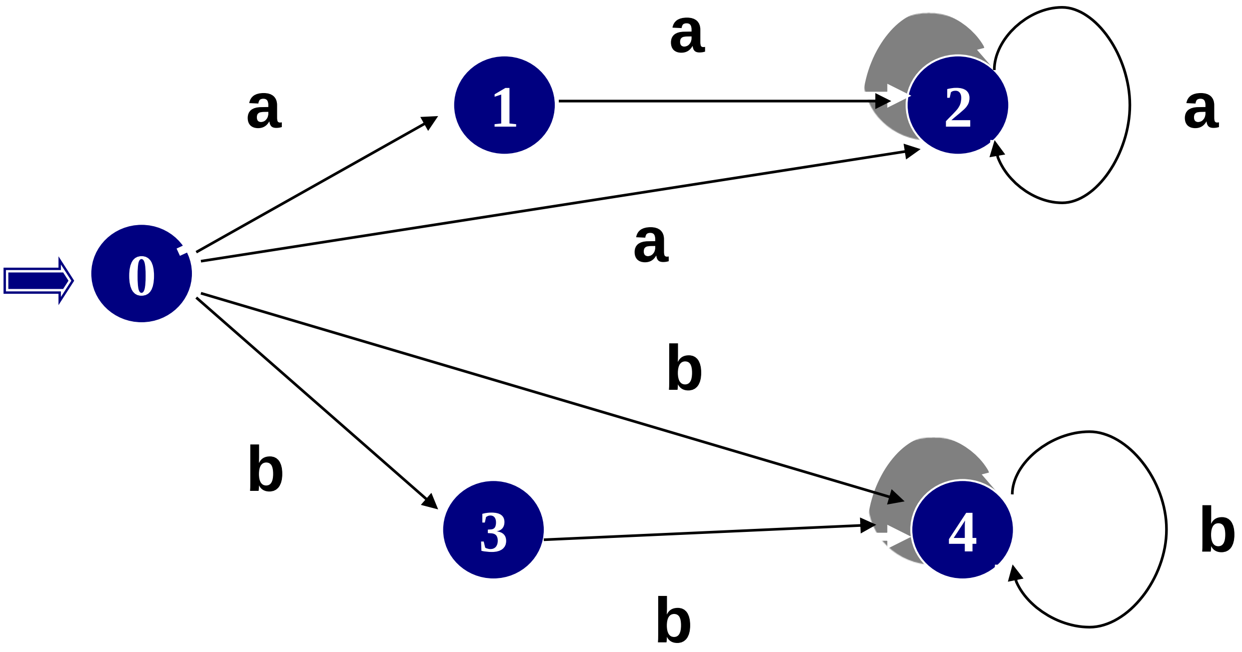


例题2.4-7 转化为不具有ε转移的非确定有限自动机.

NFA M’=({0,1,2,3,4},{a,b},f’,{0},{2,4})

f’(0,a)={1,2}	f’(0,b)={3,4}
f’(1,a)={2}	f’(3,b)={4}
f’(2,a)={2}	f’(4,b)={4}

状态转换图：NFA



定理2：对于字母表 Σ 上任何一个具

有ε–转移的NFA M，一定存

在 一个的 DFA M’,

使得：L（M’）=L（M）。

将 ε-NFA 转化为等价的 DFA

扩充 ε-closure(q)。假设T

是NFA状态的一个集合，

ε-closure(T):

表示所有那些可以从T中的元素出发经过一条ε 道路所能到达的NFA 的状态的全体所组成的集合。

NFA M = （S, Σ, f, S0, Z'）

用构造ε-closure(T) 的方法

实现DFA M' 的转换：

DFA M = （S', Σ, f', q0, Z'）

基本思想：1)首先从S0出发，仅

经过任意条ε箭弧所能到达的状态所组成

的集合作为M'的初态q0.

2)分别把从 q0出发，经过对输入符a∈Σ 的状态转移所能到达的状态（包括转移后再经 ε– 箭弧所能到达的状态）所组成的集合作为M’ 的状态， 如此继续，直到不再有新的状态为止。

2.4.4 DFA的化简

DFA的化简是指：寻找一个状态数比M少的DFA M’，

使得L（M）=L（M’）

- 状态S与T是等价：
- 两个状态是可区分的：
- DFA M状态最少化：

<u>最小化：</u>是把状态集S分割成一些不相交的子集，不同子集中的状态是可区分的，而同一子集中的状态是等价的，用一个状态代表一个子集，并消去该子集中的其它状态，从而得到化简的DFA。

对M的状态集S进行分化的步骤：

1）把S的终态与非终态分开，生成两个子集，形成基本分割Π。

2）若某一时刻分割Π已包含M个子集Π={I1，I2，。。。，IM}。检查每个子集Ii，看其是否可再分割。

$I_i=\{S_1, S_2, ..., S_k\}$

设S1，S2对任意一个输入字符a,其后继状态t1,t2,不完全包含在某个Ij 中，则有s1,s2是可区分的。

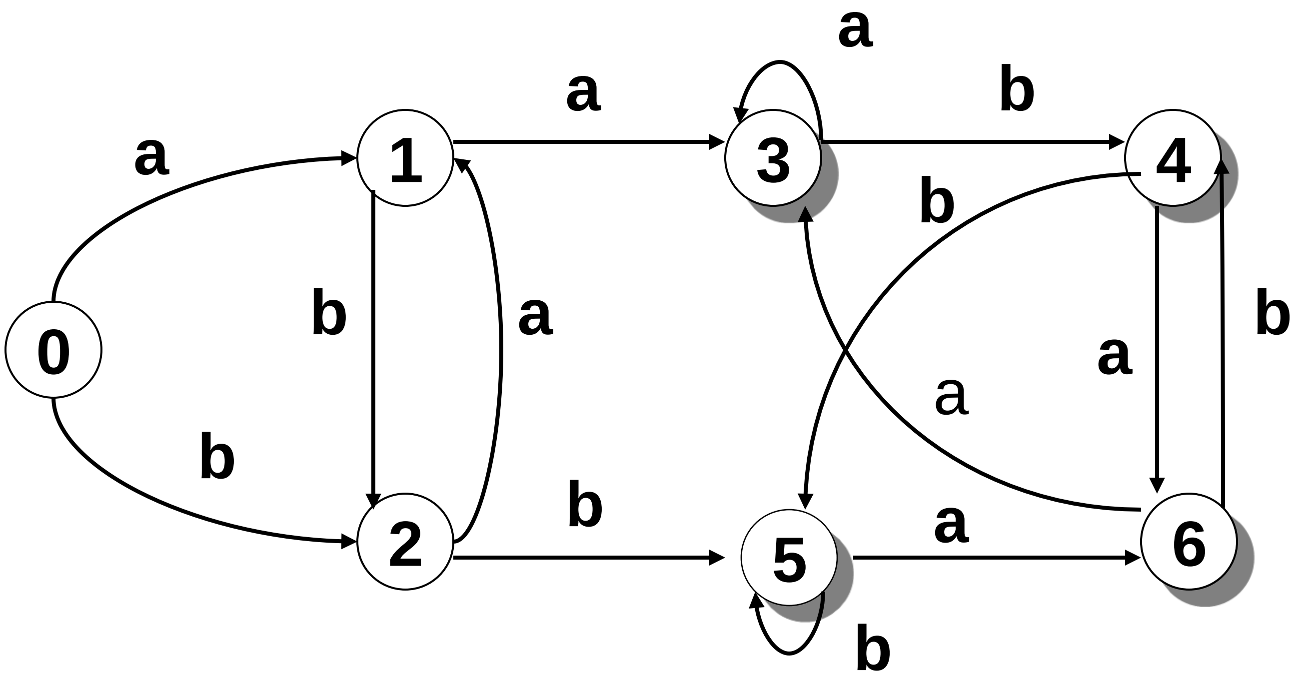
所以，把Ii一分为二，使其一半包含S1，另一半包含S2。

重复上述过程，直至 Π 所含子集不再增加为止。

对于最后分割的每个子集，在该子集中选出一个状态的代表。

若该子集含有原来的初（终）态，则该状态为新的初（终）态。

例题2.4-9：未化简的DFA M（P51）



解：1）{3 , 4, 5, 6}　　,{0, 1, 2}

2)　{3, 4, 5, 6} a ={ 3, 6}

{3, 4, 5,6} b ={4, 5}属于

{3, 4, 5, 6}不能再分。

{0，1，2} a ＝ $\{1, 3\}$

3) {1,3} 所生成的集合没有完

全包含在 {3,4,5,6}和{ 0,1,2}

中，故{0,1,2}一分为二。

1 经a 弧到3，3 ∈ {3，4，5，6}；
0，2 经a 弧到1，1 ∈ {0，1，2}；
将{0，1，2}分成{0，2}，{1}。

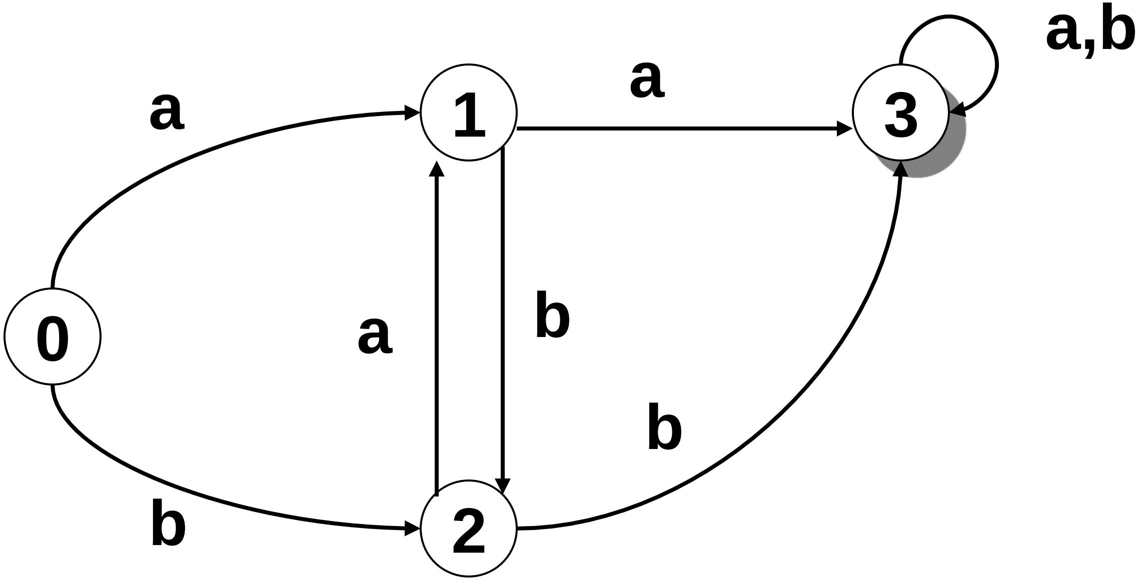
4）再检查{0，2}。

{0，2} b = {2，5}

{2，5}不完全落在{0，2}，{3，4，5}中，故{0,2}再分成{0}，{2}。

最后得到：{0}，{1}，{2}，{3，4，5，6}。

令：3代表{3，4，5，6}，得到：



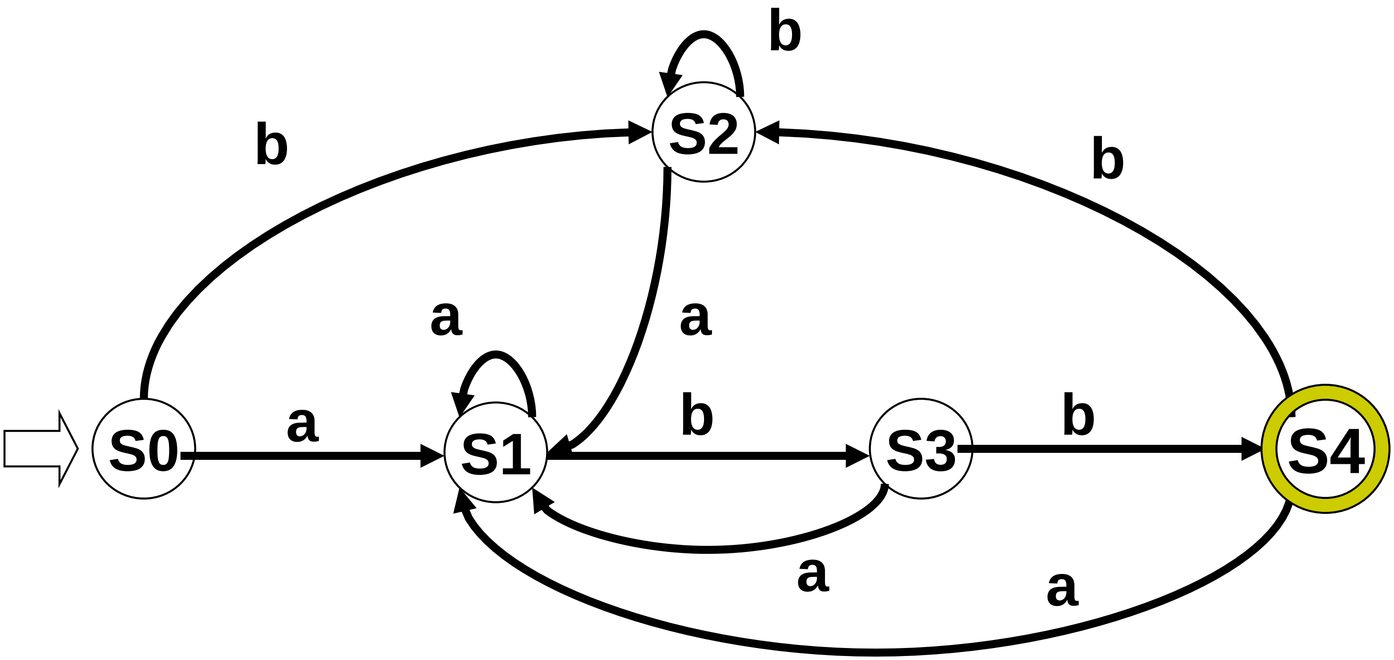
$$M = (\{0,1,2,3\} , \{a,b\} , f , 0 , \{3\})$$

$f(0,a) = 1$	$f(0,b)=2$
$f(1,a) = 3$	$f(1,b)=2$
$f(2,a) = 1$	$f(2,b)=3$
$f(3,a) = 3$	$f(3,b)=3$

例题2.4-10 将下面的DFA最小化：

M=({S0,S1,S2,S3,S4}, {a,b}, f, S0 , {S4})

	a	b
S0	S1	S2
S1	S1	S3
S2	S1	S2
S3	S1	S4
S4	S1	S2



b
S2
b
b
a
a
a
b
b
S0
S1
S3
S4
a
a

解：

1) 初始划分：　{S0,S1,S2,S3}，{S4}

2) 考察 {S0,S1,S2,S3}，

$\{S0,S1,S2,S3\}a=\{S1\}\subset\{S0,S1,S2,S3\}$

$\{S0,S1,S2\}b=\{S2,S3\}$, $\{S3\}b=\{S4\}$

{S0,S1,S2,S3} 不包含在同一子集中。

一分为二：

NEW: {S0,S1,S2},{S3},{S4}

3)考察 {S0,S1,S2},

{S0,S1,S2}a={S1}$\subset${S0,S1,S2}

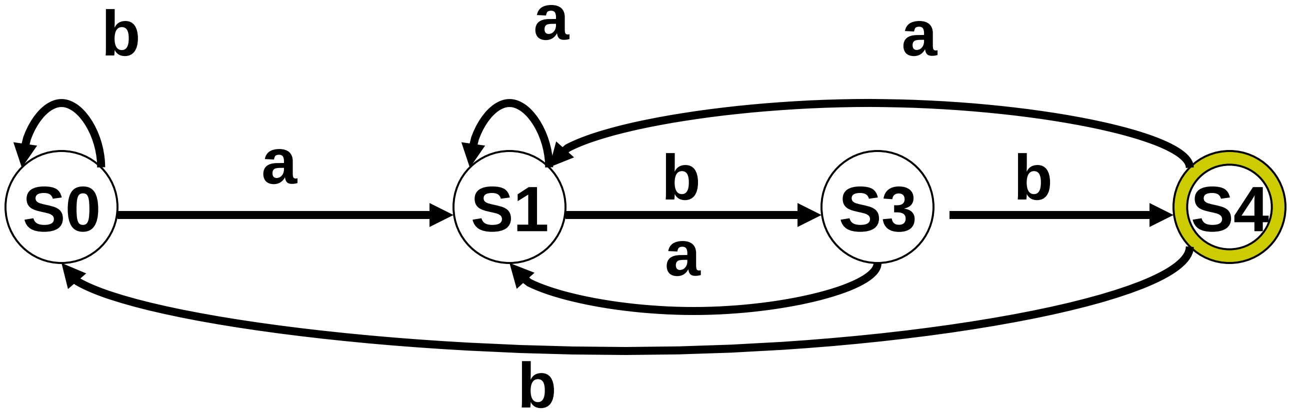
{S0,S2}b={S2}, {S1}b={S3}

{S0,S1,S2} 不包含在同一子集中。一分为二:

NEW :{S0,S2},{S1},{S3},{S4}

…….. 直到NEW不再改变。

4)S0 作为 {S0,S2} 的代表。


b
a
a
a
b
b
S0
S1
S3
S4
a
b