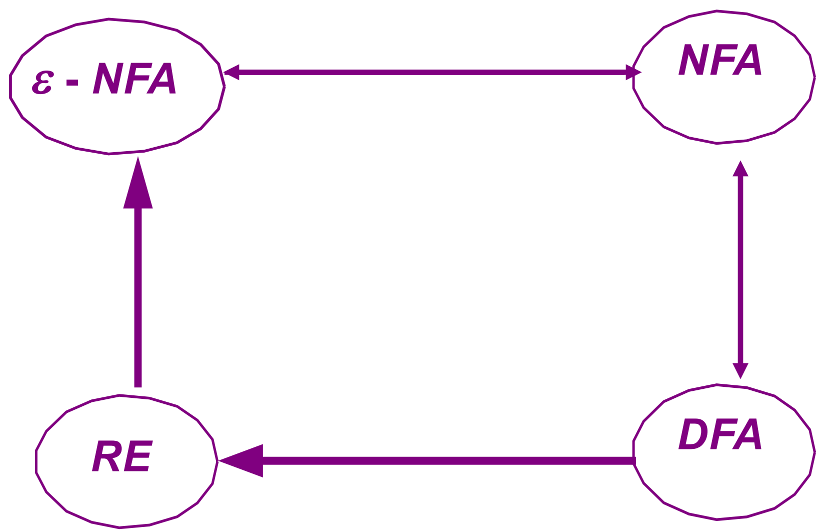
第四讲

- 正则表达式与有限自动机的关系
- 右线性语言与有限自动机的关系
- 右线性语言的性质(part1)

3.7 正则表达式与有限自动机的关系

结论： 有限自动机、右（左）线性文法、正则表达式都定义了同一种语言-- 正则语言 .

证明策略



RE(Regular Expression) --- 正则表达式

从有限自动机构造等价的正则表达式
（状态消去法）

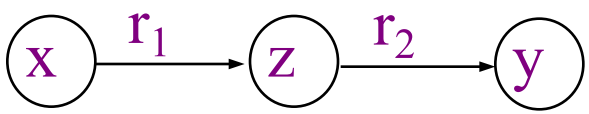
思路:

(1) 扩展自动机的概念，允许正则表达式作为转移弧的标记. 这样，就有可能在消去某一中间状态时，保证自动机能够接受的字符串集合保持不变.

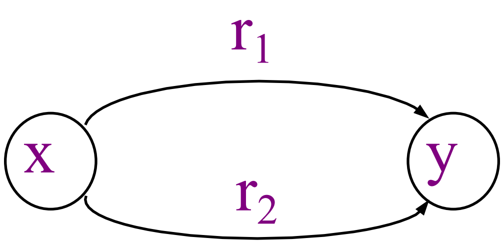
(2) 在消去某一中间状态时，与其相关的转移弧也将同时消去，所造成的影响将通过修改从每一个前趋状态到每一个后继状态的转移弧标记来弥补.

以下分别介绍中间状态的消去与正则表达式构造过程.

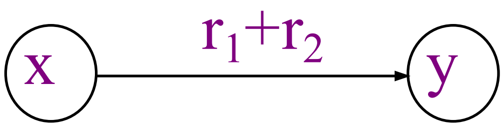
从有限自动机构造等价的正则表达式
（中间状态的消去）

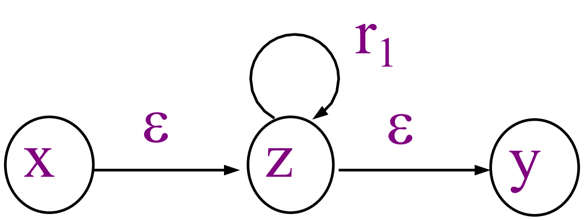


代之以：

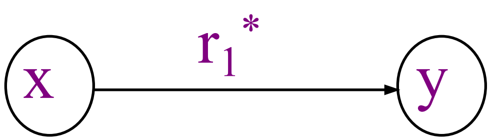


代之以：

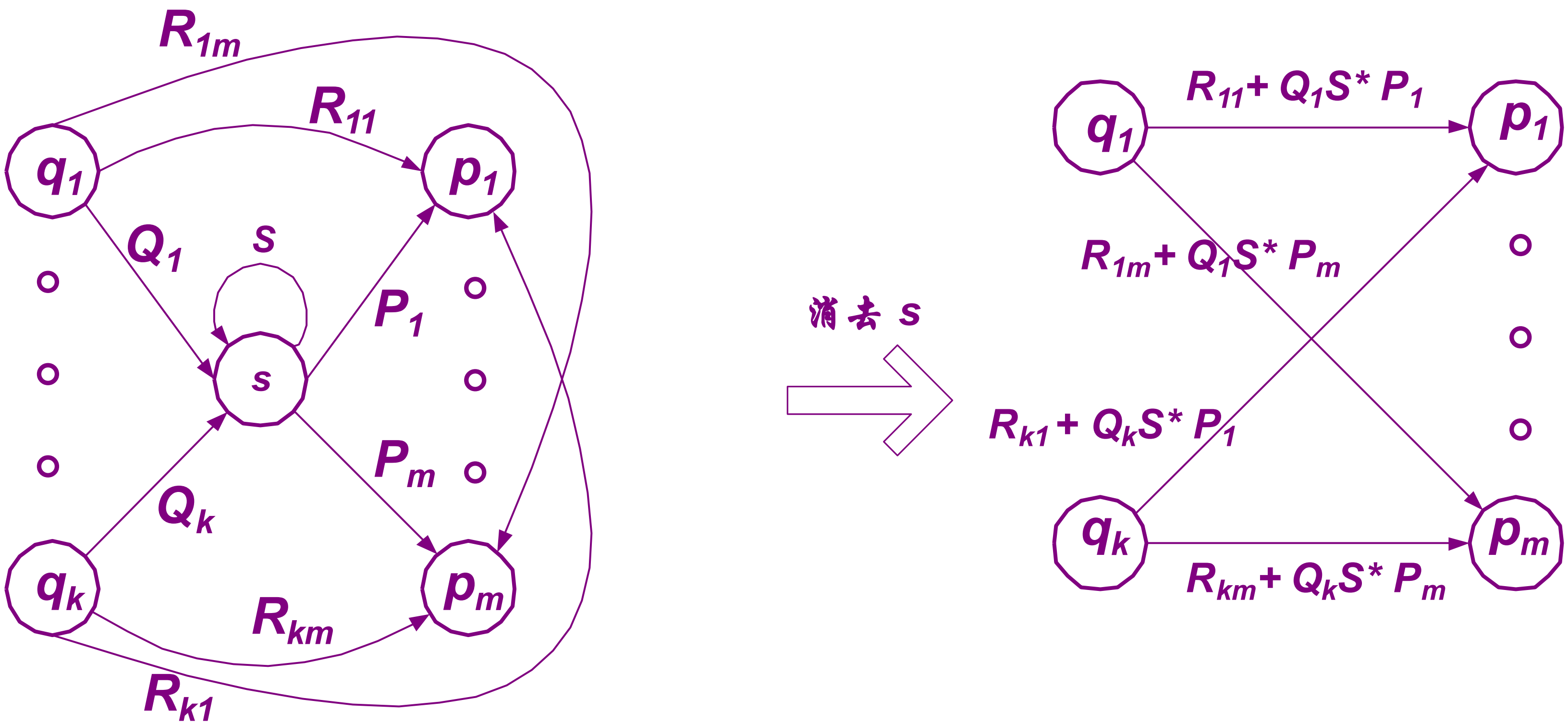




代之以：



从有限自动机构造等价的正则表达式
（中间状态的消去）



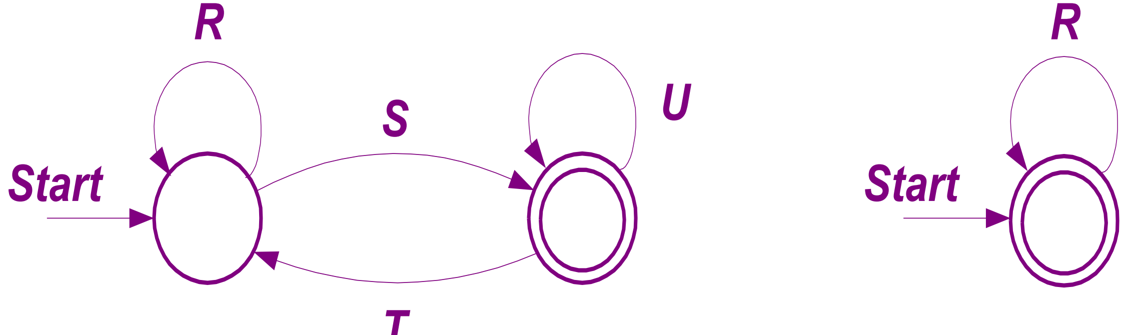
从有限自动机构造等价的正则表达式
（状态消去法）

步骤:

(1) 对每一终态q，依次消去除 q 和初态 q_0 之外的其它状态;

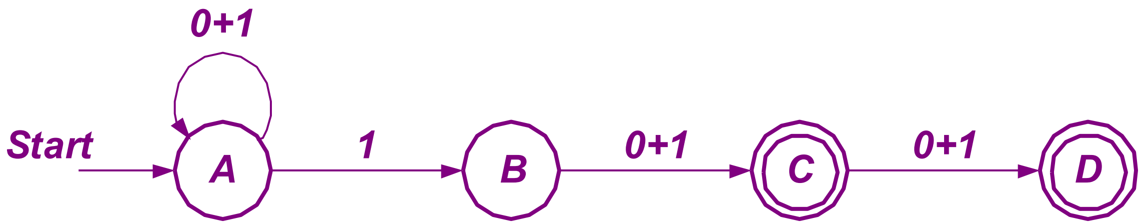
(2) 若 $q \neq q_0$，最终可得到一般形式如下左图两状态自动机，该自动机对应的正则表达式可表示为 $(R+SU^*T)^*SU^*$.

(3) 若 $q = q_0$，最终可得到如下右图的自动机，它对应的正则表达式可以表示为 R^*.

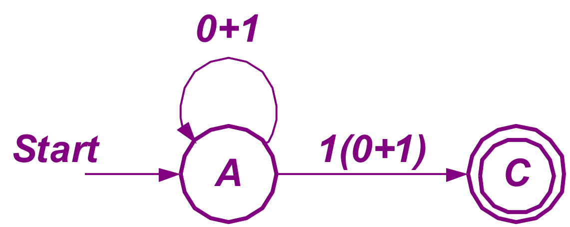


(4) 最终的正则表达式为每一终态对应的正则表达式之和（并）.

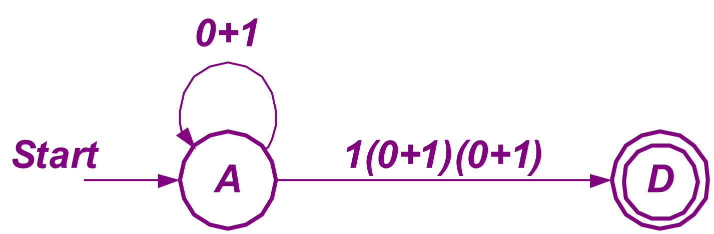
状态消去法举例



对于终态C



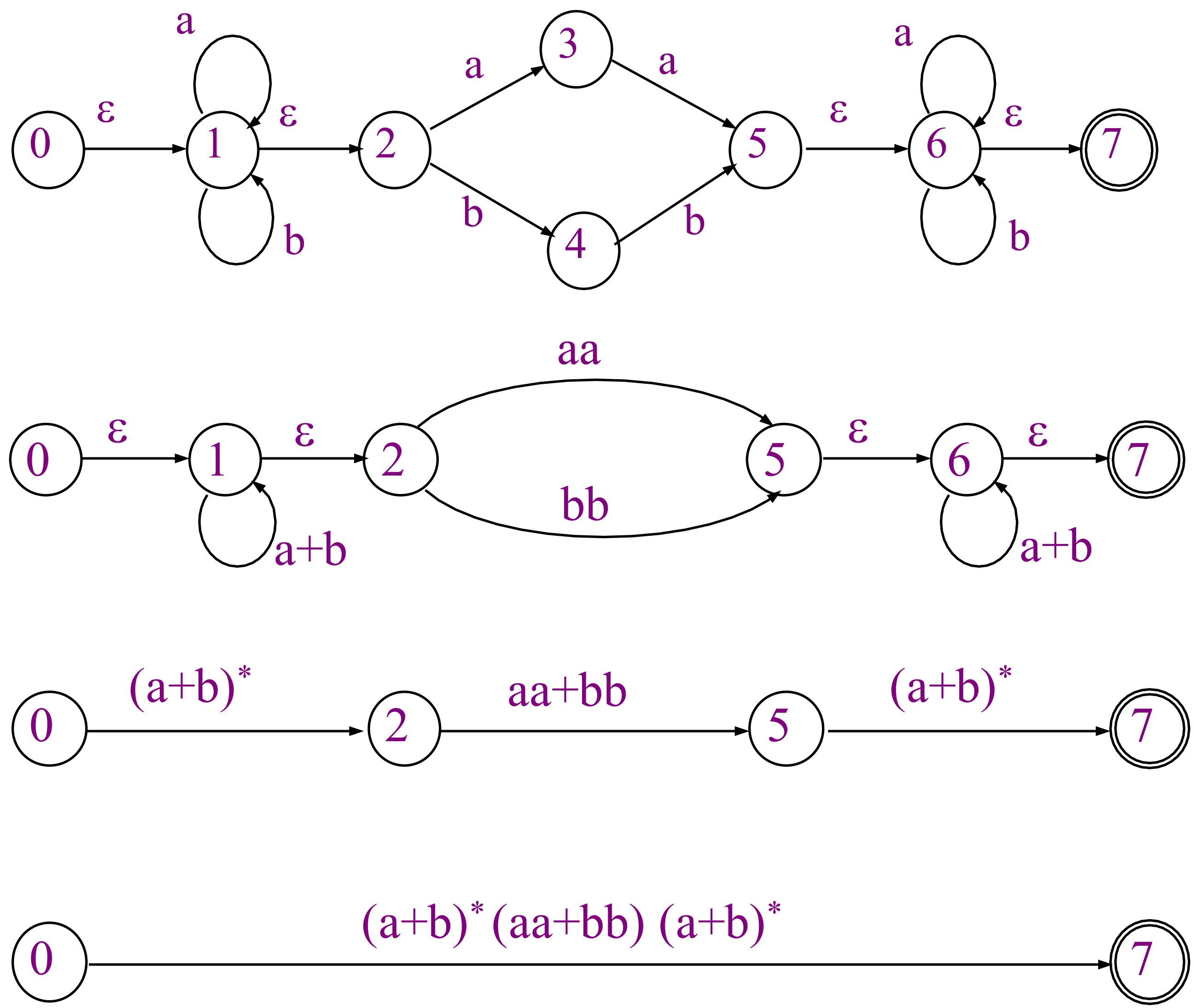
对于终态D



等价的正则表达式

$$(0+1)^*1(0+1)+(0+1)^*1(0+1)(0+1)$$

状态消去法举例



从正则表达式构造等价的ε - NFA

定理: *L* 是正则表达式 *R* 表示的语言, 则存在一个ε - *NFA* *E* ,满足 *L(E) = L(R) = L*.

证明: 构造性证明. 可以通过结构归纳法证明从 *R* 可以构造出与其等价的, 满足如下条件的ε - *NFA* :

(1) 恰好一个终态;

(2) 没有弧进入初态;

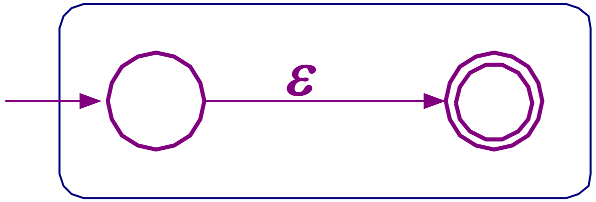
(3) 没有弧离开终态;

□

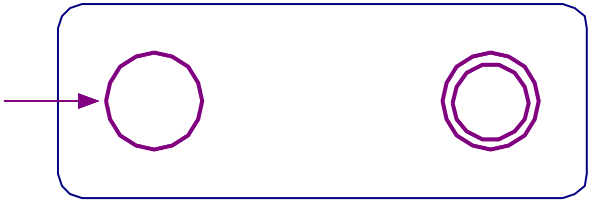
从正则表达式构造等价的ε-NFA
(归纳构造过程)

基础:

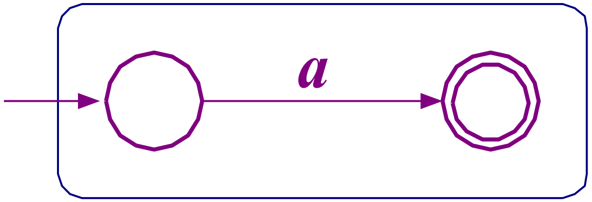
1 对于 ε，构造为



2 对于 ϕ，构造为



3 对于 a，构造为

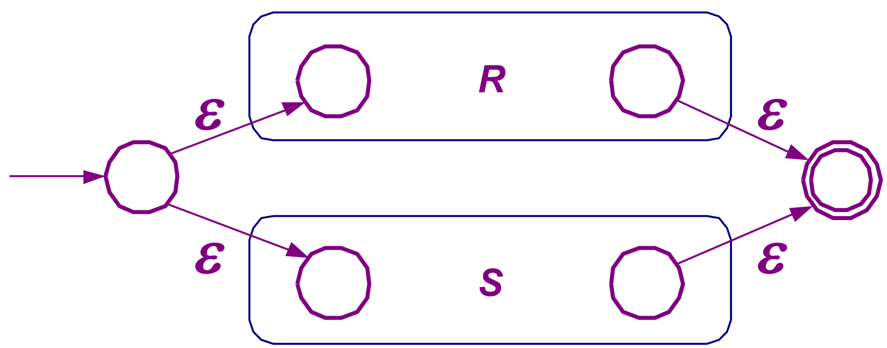


从正则表达式构造等价的ε-NFA

（归纳构造过程）

归纳：

1 对于 R+S，构造为

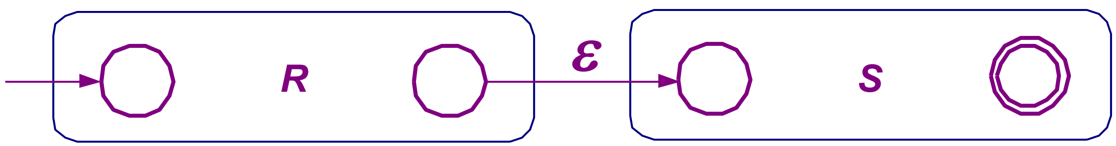


从正则表达式构造等价的ε - NFA

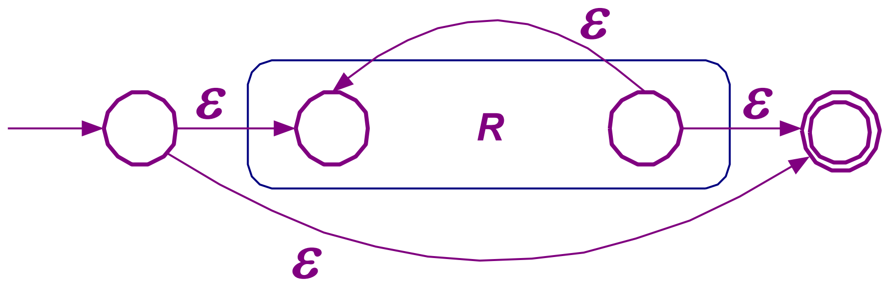
（归纳构造过程）

归纳：

2 对于 *RS*，构造为



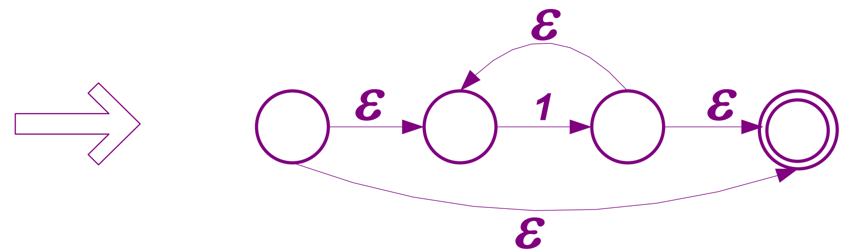
3 对于 *R**，构造为



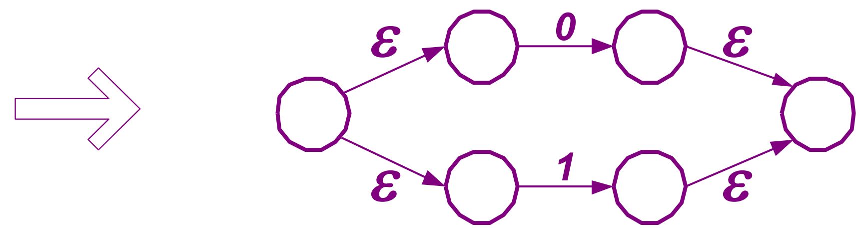
从正则表达式构造等价的ε - NFA

举例：设正则表达式 1*0(0+1)*，构造等价的ε - NFA.

1*

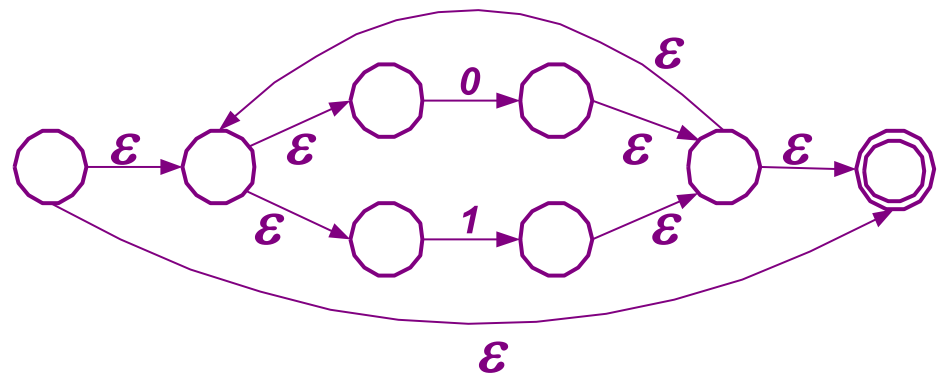


0+1

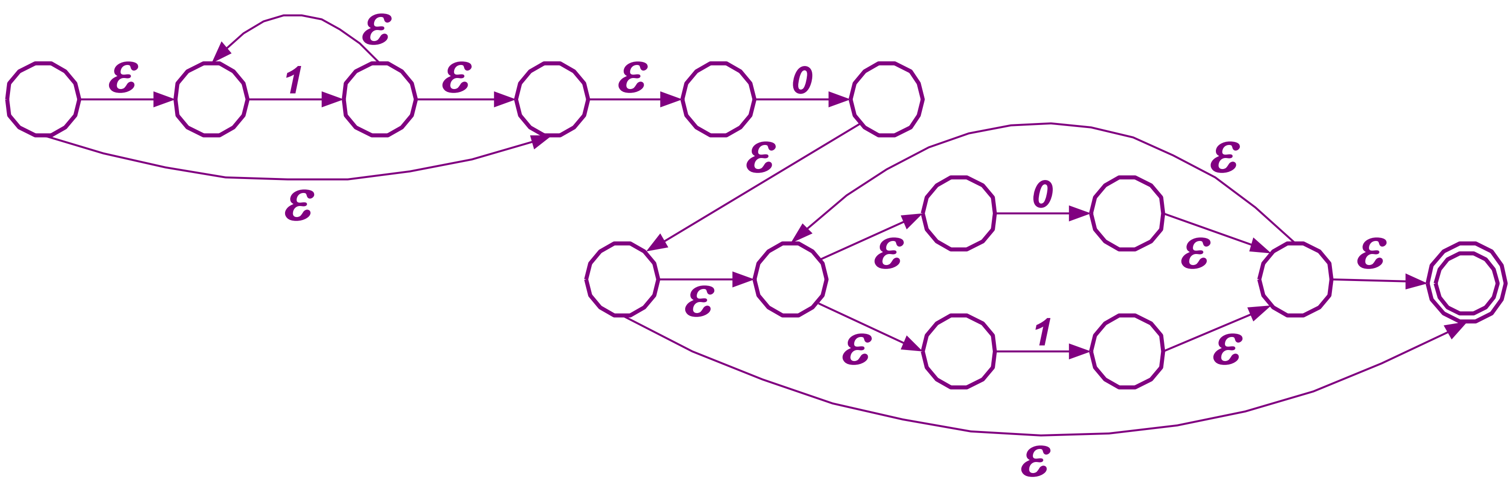


从正则表达式构造等价的ε - NFA

*(0+1)** ⇒



*1*0(0+1)** ⇒



3.8 右线性语言与有限自动机

至此，我们已学到正则集有三种定义方式，且这三种方式等价：

1. 正则集是含有{ε}，φ，{a} 以及在并、连接和 * 运算下封闭的语言
2. 由正规表达式定义的集合是正则集。
3. 由右线性文法生成的语言是正则集。

此外，还有第四种方式：

将正则集作为由有限自动机定义的集合。

即 正则集(右线性语言) <＝> 有限自动机

右线性文法 => 有限自动机

定理3.8.1：由任意右线性文法G定义的语言必然能被一个NFA M所接受。即L（G）＝L（M）

证明思路（构造证明）：

设右线性文法G＝（N，T，P，S），构造一个与G等价的有限自动机NFA M＝（Q，T，δ，q_0，F），其中：

Q＝N U {H}，H为一个新增加的状态，$H \notin N$， q_0＝S

$$F=\begin{cases}\{H, S\} & \text{当}S\to\varepsilon\text{属于}P\text{。}\\ \{H\} & \text{否则}\end{cases}$$

δ的定义为：

当$A \rightarrow aB \in P$，则 $B \in \delta$（A，a）

当$A \rightarrow a \in P$，则 $H \in \delta$（A，a）

对于任意输入，δ（H，a）＝φ。

右线性文法 => 有限自动机（例）

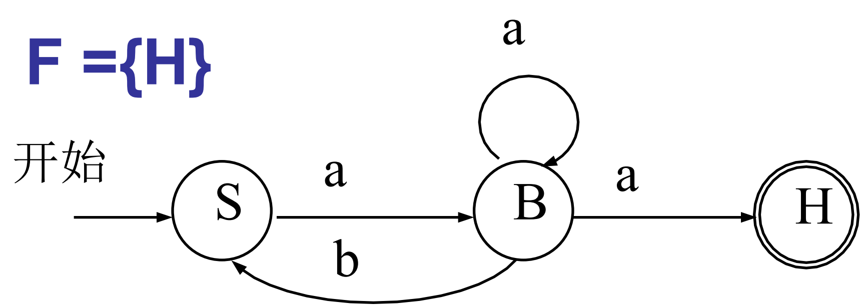
例：设有右线性文法G=({S,B} ,{a,b}, P,S)，其中

P： $S \rightarrow aB$　　$B \rightarrow aB|bS|a$

试构造与G等价的有限自动机M。

解： 设NFA $M=(Q,T, \delta, q_0, F)$

$Q=\{S,B,H\}$　$T=\{a,b\}$　$q_0 = S$　$F =\{H\}$



转换函数δ：

- 对于产生式$S \rightarrow aB$，有$\delta(S,a)=\{B\}$
- 对于产生式$B \rightarrow aB$，有$\delta(B,a)=\{B\}$
- 对于产生式$B \rightarrow bS$，有$\delta(B,b)=\{S\}$
- 对于产生式$B \rightarrow a$，　有$\delta(B,a)=\{H\}$

右线性文法 => 有限自动机（续）

求证 G与NFA M两者定义了同一语言。

证明：

先证（1）文法G产生的语言 L(G) 能够被NFA M所接收；

再证（2）NFA M 接受的语言 L(M) 可由文法G 产生。

右线性文法 => 有限自动机（续）

证明方法：通过两者定义的语言中任意一个字符串来说明。

（1）设 $\omega = a_1a_2\ldots a_n \in L(G)$，且 $n \geq 1$

则有 $S \Rightarrow a_1A_1 \Rightarrow a_1a_2A_2 \Rightarrow \ldots$

$\Rightarrow a_1a_2\ldots a_{n-1}A_{n-1} \Rightarrow a_1a_2 \ldots a_{n-1}a_n$

则由δ的定义，有

$A_1 \in \delta(S, a_1)$，$A_2 \in \delta(A_1, a_2)$，… ，

$A_{n-1} \in \delta(A_{n-2}, a_{n-1})$，$H \in \delta(A_{n-1}, a_n)$，且 $H \in \delta(S, \omega)$

因为 $H \in F$，所以ω被NFA M 所接受。

又若 $\varepsilon \in L（G）$，则表明 $S \rightarrow \varepsilon \in P$，由 NFA M 的定义，

有 $S \in F$，即ε也被NFA M接受。

所以，由文法G派生的任意字符串 $\omega \in L（M）$。 #

右线性文法 => 有限自动机（续）

（2）再证 L(M)可由G产生

设$\omega = a_1a_2\ldots a_n$ 被NFA M接受，即 $\omega \in L$（M），

则必然存在状态序列 S， A_1， A_2 ， … A_{n-1}， H

对M有转换函数为

$$A_1 \in \delta(S, a_1), \ A_2 \in \delta(A_1, a_2), \ \ldots,$$

$$A_{n-1} \in \delta(A_{n-2}, a_{n-1}), \ H \in \delta(A_{n-1}, a_n)$$

则可规定G中含有产生式

$$S => a_1A_1, \quad A_1 => a_2A_2, \ \ldots\ldots, \ A_{n-1} => a_n$$

于是存在推导

$$S => a_1A_1 => a_1a_2A_2 => \ldots => a_1a_2\ldots a_{n-1}A_{n-1} => a_1a_2 \ldots a_{n-1}a_n$$

即$a_1a_2\ldots a_n$ 是文法G的一个句子。

也即 $\omega \in L$（G）。 #

课堂练习：

练习：　设线性文法 G ＝（{S，A，B}，{a，b}，P，S）

P：　$S \to aA \mid a$

$A \to aA \mid aS \mid bB$

$B \to bB \mid b \mid a$

构造相应的 NFA M。

定理3.8.2：设有限自动机 M 接受的语言为L(M)
则存在右线性文法G，它产生的语言L(G)＝L(M)。

证明思路：

构造一个右线性文法G，使它接受由NFA M定义的语言。

构造方法：

设 M＝（Q，T，δ，q_0，F），构造一个右线性文法
G＝（N，T，P，S），其中N＝Q，S＝q0

P定义为：

若δ（A，a）＝B 且 B $\notin$ F，则A → aB 在P中

若δ（A，a）＝B 且 B $\in$ F，则A → a 和A → aB 在P中

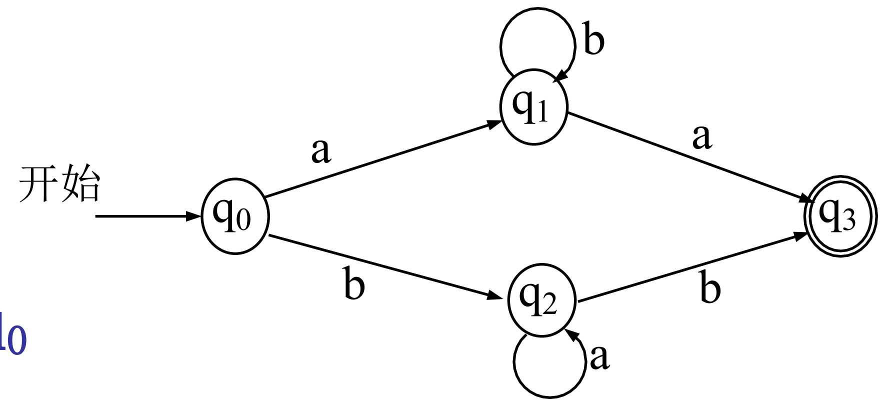
L（M）<=> L（G）的证明见书 P91 （自学）。

有限自动机=>右线性文法（例）

例：设有DFA M =({q0,q1,q2,q3}, {a,b}, δ , q0, {q3})

其中转换函数如图所示，

试构造与之等价的右线性文法G。



解：构造右线性文法G=(N, T, P, S)

$N = \{q_0, q_1, q_2, q_3\}$ $T = \{a, b\}$ $S = q_0$

产生式集合P

$\delta(q_0, a)=q_1,$ $\therefore\ q_0 \to aq_1$

$\delta(q_0, b)=q_2,$ $\therefore\ q_0 \to bq_2$

$\delta(q_1, a)=q_3,\ q_3 \in F,$ $\therefore\ q_1 \to a \mid aq_3$

$\delta(q_1, b)=q_1,$ $\therefore\ q_1 \to bq_1$

$\delta(q_2, a)=q_2,$ $\therefore\ q_2 \to aq_2$

$\delta(q_2, b)=q_3,\ q_3 \in F,$ $\therefore\ q_2 \to b \mid bq_3$

构造的文法G(化简q_3)：

$G=(\{q_0, q_1, q_2\}, \{a, b\}, P, q_0)$

P： $q_0 \to aq_1 \mid bq_2$

$q_1 \to a \mid bq_1$

$q_2 \to aq_2 \mid b$

3.9 右线性语言的性质

主要内容：

- DFA的极小化
- 泵浦引理
- 右线性语言的封闭性

确定有限自动机DFA的化简(极小化)

对DFA　M的极小化是找出一个状态数比M少的DFA M1，使满足 L(M) = L(M1)

1. 等价和可区分的概念

设DFA M = (Q，T，δ，q0，F)

对不同的状态q１，q２∈Q 和每个ω∈T*，如果有

(q１，ω) ⊢* (q，ε) 必有 (q２，ω) ⊢* (q，ε) 且q∈F，

则称q１与q２状态等价. 记为q１≡q２

否则，称q１，q２可区分.

2．不可达状态

如果不存在任何$\omega \in T^*$，使$(q_0, \omega) \vdash^* (q, \varepsilon)$，则称状态$q \in Q$为不可达状态.

3．最小化

若DFA M不存在互为等价状态及不可达状态，则称DFA M是最小化的.

最小化算法

一个DFA M的最小化，是把M的状态集Q构成一个划分。即: 任何两个子集的状态都是可区分的；同一子集中的任何两个状态都是等价的。之后，每个子集用一个状态代表，并取一个状态名.

构成划分的步骤:

1. 构成基本划分 $\prod=\{\prod', \prod''\}$，($\prod'$为终态集，$\prod''$为非终态集)

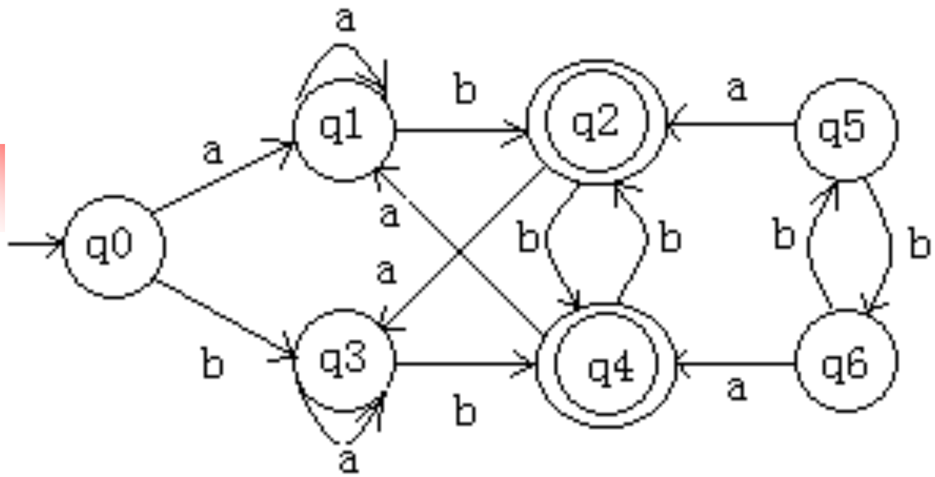
2. 细分 $\prod=\{\prod^1, \prod^2, \ldots, \prod^n\}$,

$\prod^i=\{q_1, q_2, \ldots, q_m\}$

当输入任意字符a时，若$\prod^i$中的状态经标a的边可到达的状态集的元素分属于两个不同的子集中，则将$\prod^i$细分为两个子集.

重复步骤(2)，直至不可再细分，得到M1.

若M1中有不可达状态，将其删除，M1便是最小化的.



例

(1) q 5， q 6 为不可达状态， 删除之.

(2) Q = { q 0， q 1， q 2， q 3， q 4}，

∏= {{q 2， q 4}，{q 0， q 1， q 3}}

构成基本划分 ∏={∏’，∏”}

(a) 对于∏’ ={ q 2， q 4 }，

对字符a，有δ(q 2， a)= q 3， δ(q 4， a)= q 1 q 1， q 3 ∈ 同一子集.

对字符b，有δ(q 2， b)= q 4， δ(q 4， b)= q 2 q 4， q 2 ∈ 同一子集.

∴ ∏’ = { q 2， q 4 } 不能再细分. 可用q 2 表示∏’ 状态.

(b) 对于∏” = { q 0， q 1， q 3 }

对a，δ(q 0 ,a)= q 1， δ(q 1 ,a)= q 1， δ(q 3 ,a)= q 3 q 1 , q 3 ∈同一子集

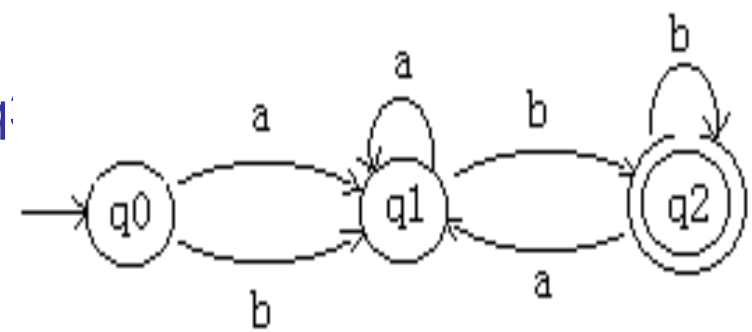
对b，δ(q 0 ,b)= q 3， δ(q 1 ,b)= q 2， δ(q 3 ,b)= q 4

q 3， q 2， q 4 ∉ 同一子集.

∴ 将∏’’再分解. ∏’’={ { q 0 }，{ q1，q

q1表示

∴ Q ＝ { {q 0 }， {q 1 }， {q 2 } }



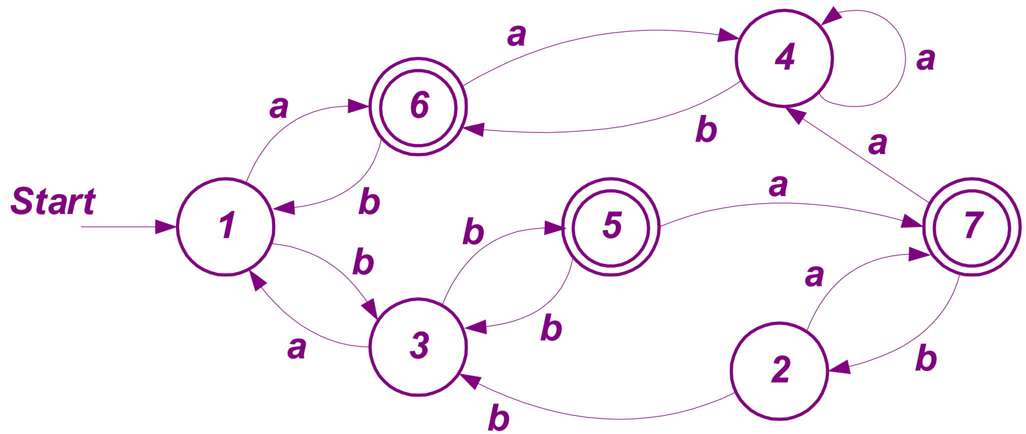
计算状态集划分的算法— 填表法

- 填表算法（*table-filling algorithm*）基于如下递归地标记可区分的状态偶对的过程:
  - 基础 如果 p 为终态，而 q 为非终态，则 p 和 q 标记为可区分的；
  - 归纳 设 p 和 q 已标记为可区分的，如果状态 r 和 s 通过某个输入符号 a 可分别转移到 p 和 q，即 $\delta(r,a)=p$，$\delta(s,a)=q$，则 r 和 s 也标记为可区分的；这是因为：若 p 和 q 可为字符串 w 区分，则 r 和 s 可为字符串 aw 区分.
    （$\because$ $\hat{\delta}(r,aw)=\hat{\delta}(p,w)$，$\hat{\delta}(s,aw)=\hat{\delta}(q,w)$）

计算状态集划分的算法— 填表法

✧ 填表算法举例

2						
3	x	x				
4	x	x	x			
5	x	x	x	x		
6	x	x	x	x	x	
7	x	x	x	x	x	
	1	2	3	4	5	6



(1) 区分所有终态和非终态

(2) 区分(1,3), (1,4), (2,3),
(2,4), (5,6), (5,7)

(3) 区分 (3,4)

(4) 结束． 划分结果：{1,2}, {3}, {4}, {5}, {6,7}

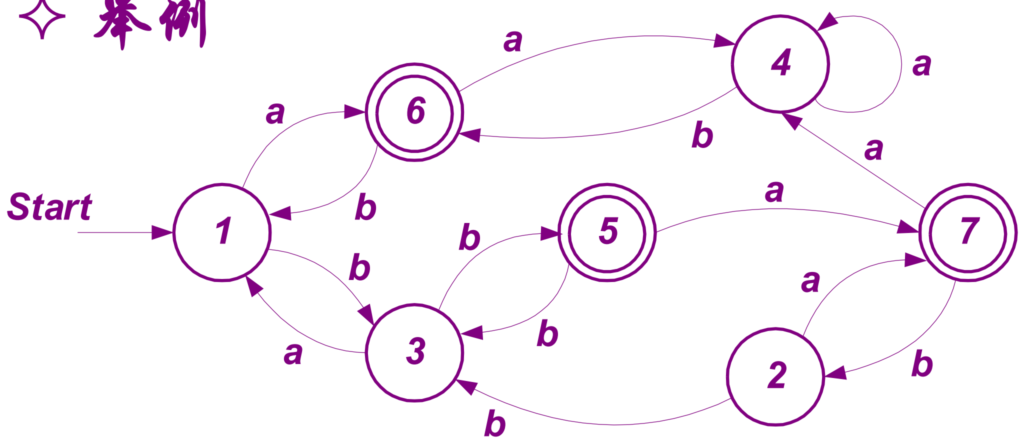
通过合并等价的状态进行 DFA 的优化

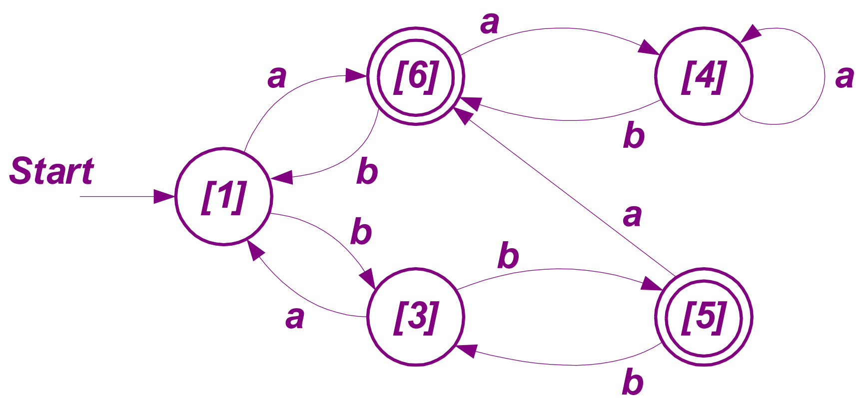
✧ 步骤

1. 删除所有从开始状态不可到达的状态及与其相关的边，设所得到的 DFA 为 $A=(Q, \mathrm{T}, \delta, q_0, F)$；
2. 使用填表算法找出所有等价的状态偶对；
3. 根据 2 的结果计算当前状态集合的划分块，每一划分块中的状态相互之间等价，而不同划分块中的状态之间都是可区分的。包含状态 q 的划分块用 $[q]$ 表示。
4. 构造与 A 等价的 DFA $B=(Q_B, \mathrm{T}, \delta_B, [q_0], F_B)$，其中 $Q_B=\{[q] \mid q\in Q\}$, $F_B=\{[q] \mid q\in F\}$, $\delta_B([q], a)=[\delta(q,a)]$

通过合并等价的状态进行 DFA 的优化

◇ 举例





– 等价的状态偶对为：

（1, 2），（6, 7）

– 划分结果：

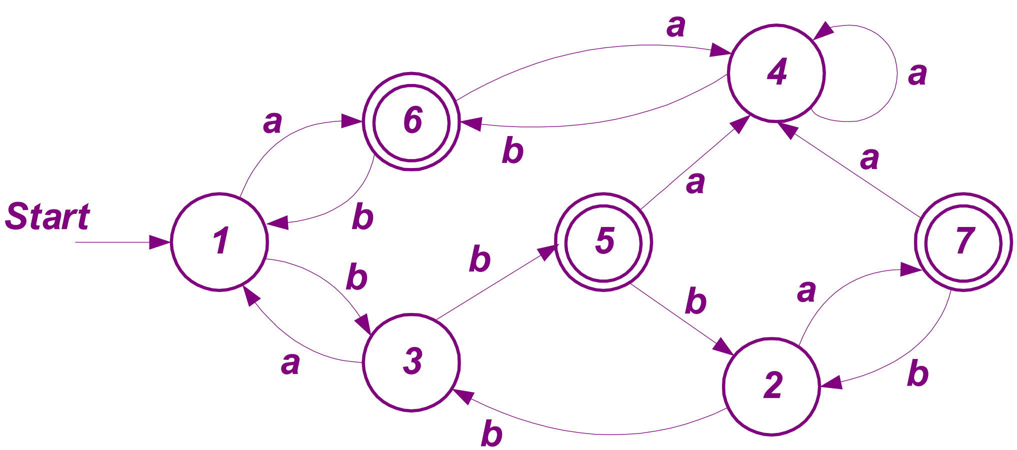
{ 1, 2 }, {3}, {4}, {5}, { 6, 7 }

– 新的状态集合：

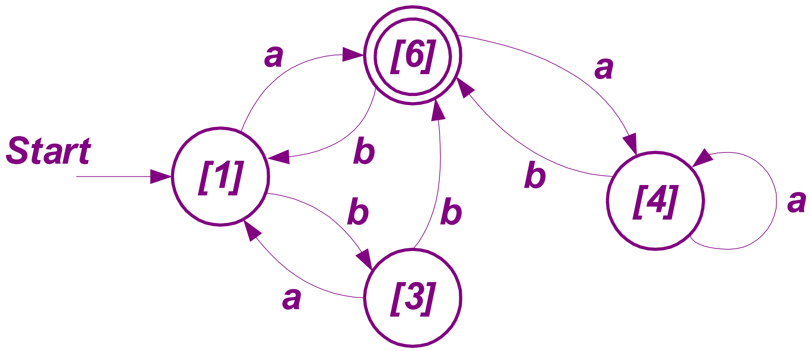
[1], [3], [4], [5], [6]

最小化的 DFA

- 课堂练习 最小化下列 DFA:



- 参考结果



针对正则语言的Pumping引理

- 正则语言应满足的一个必要条件
- 用于判定给定的语言不是正则集。

物理意义：当给定一个正则集和该集合上一个足够长的字符串时，在该字符串中能找到一个非空的子串，并使子串重复，从而组成新的字符串。该新串必在同一个正则集内。

定理：

设L是正则集，存在常数n，对字符串$\omega \in L$且$|\omega| \geq n$，则ω可写成$\omega_1\omega_0\omega_2$，其中$|\omega_1\omega_0| \leq n$，$|\omega_0| > 0$，对所有的$i \geq 0$有$\omega_1\omega_0^i\omega_2 \in L$。

证明 设 L 是 DFA $D = (Q, T, \delta, q_0, F)$ 的语言，取 $n = |Q|$ 即可.

DFA 的 “Pumping”特性

设 DFA $D=(Q, T, \delta, q_0, F)$, $|Q|=n$.

对于任一长度不小于 n 的字符串 $w=a_1a_2\ldots a_m$, 其中 $m\geq n$, $a_k\in T\ (1\leq k\leq m)$, $q\in Q$, 考察如下状态序列

$$\left.\begin{aligned}
&p_0=q\\
&p_1=\delta'(q, a_1)\\
&p_2=\delta'(q, a_1a_2)\\
&\ldots\\
&p_n=\delta'(q, a_1a_2\ldots a_n)
\end{aligned}\right\}$$

$$\begin{aligned}
&p_{n+1}=\delta'(q, a_1a_2\ldots a_{n+1})\\
&\ldots\\
&p_m=\delta'(q, a_1a_2\ldots a_m)
\end{aligned}$$

由 *pigeonhole* 原理, $p_0, p_1, p_2, \ldots, p_n$ 中至少有两个状态是重复的，即存在 i, j, $0\leq i<j\leq n$, $p_i=p_j$.

✧ “*pumping*” 特性：

任一长度不小于状态数目的字符串所标记的路径上，必然出现重复的状态.

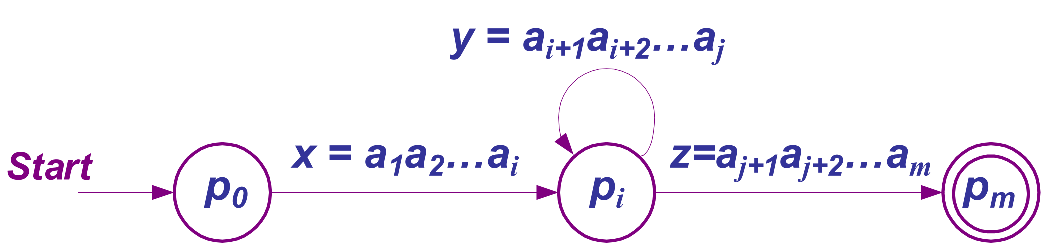
DFA 的 “Pumping”特性

- “pumping”特性：如前，设 DFA $D = (Q, T, \delta, q_0, F)$, $|Q|=n$, $w = a_1a_2\ldots a_m$ $(m\geq n)$, 则存在 i, j, $0\leq i<j\leq n$, $p_i=p_j$, 其中 $p_k=\delta'(p_0, a_1a_2\ldots a_k)$, $0\leq k\leq m$.
- 若假定 $p_0=q_0$, $p_m\in F$, 即 $w\in L(D)$.

令 $w = xyz$, 其中：

$$x = a_1a_2\ldots a_i\ ,\ y = a_{i+1}a_{i+2}\ldots a_j\ ,\ z = a_{j+1}a_{j+2}\ldots a_m$$

则对任何 $k \geq 0$，都有 $xy^kz \in L(D)$. （参考下图）



Pumping 引理的应用

（用于证明某个语言 L 不是正规语言）

✧ 证明步骤

1. 选任意的 *n*.
2. 找到一个满足以下条件的串 $w \in L$ (长度至少为 *n*).
3. 任选满足 $w = xyz \wedge y \neq \varepsilon \wedge |xy| \leq n$ 的 *x,y,z*
4. 找到一个 $k \geq 0$, 使 $xy^kz \notin L$.

✧ 举例　证明 $L=\{ a^nb^n \mid n\geq 1 \}$ 不是正则集.

✧ 证明：由泵浦引理，假设L是正则集，则对足够大的n，a^nb^n 可写成 $\omega_1\omega_0\omega_2$，其中　$0< |\omega_0|\leq n$，$|\omega|= 2n > n$

若 $\omega_0 = a^+$ 或 b^+，设 $|\omega_0|=k\geq 1$，k为常数，

取 $i=0$，有 $\omega_1\omega_0^0\omega_2 = \omega_1\omega_2 = a^{n-k}b^n$ 或 a^nb^{n-k}，此时，a，b字符个数不同，即新组成的串 $\omega_1\omega_2 \notin L$.

若 $\omega_0 = a^+b^+$，可取 $i=2$，有 $\omega_1\omega_0\omega_0\omega_2 = \omega_1a^+b^+a^+b^+\omega_2 \notin L$

∴ 与假设矛盾，故L不是正则集.

Pumping 引理的应用

例　证明$\mathcal{L}=\{a^{k^2}\mid k\geq 1$ 的整数$\}$不是正则集.

证明　假设L是正则集，取足够大的整数n，$\omega=a^{n^2}$.

有$|\omega|=|\omega_1\omega_0\omega_2|=n^2\geq n$，

$0<|\omega_0|\leq n$，　$0<|\omega_1\omega_0|\leq n$

取$i=2$，有$n^2<|\omega_1\omega_0^2\omega_2|\leq n^2+n<(n+1)^2$

∴　$|\omega_1\omega_0^i\omega_2|\notin\mathcal{L}$　(串长不是整数的平方)

与假设矛盾。

∴　L不是正则集。

课后练习

转换下列正则表达式为带ε转移的NFA.

a) 01.*
b) (0+1)01.
c) 00(0+1).*

Chap3 习题 7，9，20