



第2讲

词法分析-分析器自动生成

李 诚

国家高性能计算中心(合肥)、信息与计算机国家级实验教学示范中心

计算机科学与技术学院

2023年09月06日

三个问题!

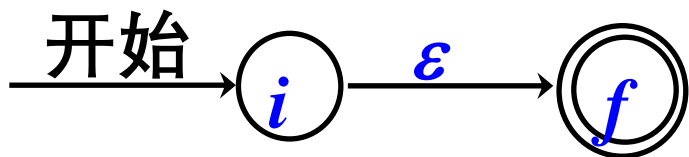


- ❑ 如何进行正则表达式→**NFA**的转换
- ❑ 如何进行**NFA** → **DFA**的转换
- ❑ 如何进行**DFA** →化简**DFA**的转换

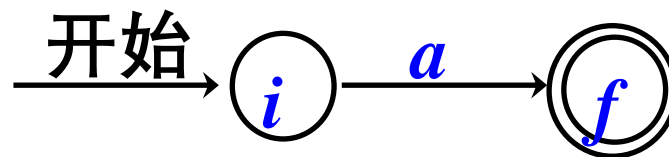


- 首先构造识别 ϵ 和字母表中一个符号 a 的NFA

❖ 重要特点：仅一个接受状态，它没有向外的转换



识别正则表达式 ϵ 的
NFA



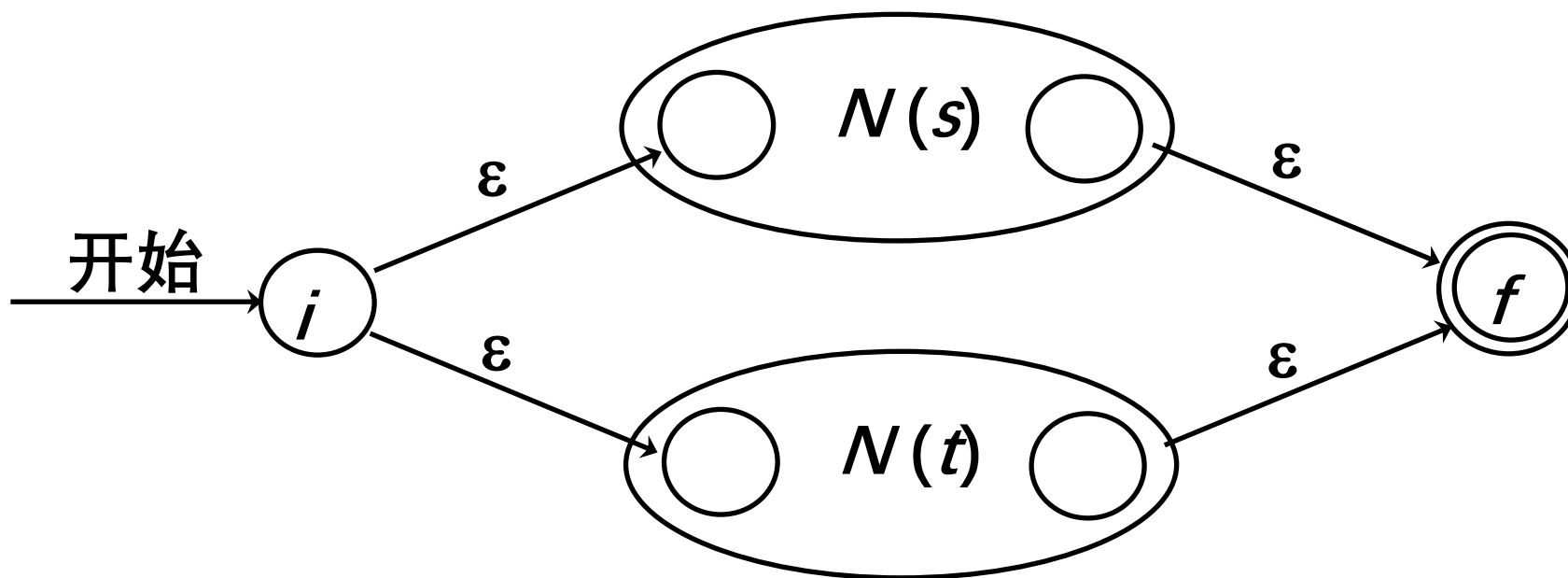
识别正则表达式 a 的
NFA

- 对于加括号的正则表达式(s)，其NFA可用 s 的NFA（用 $N(s)$ 表示）代替



- 构造识别主算符为选择的正则表达式的NFA

❖ 重要特点：仅一个接受状态，它没有向外的转换

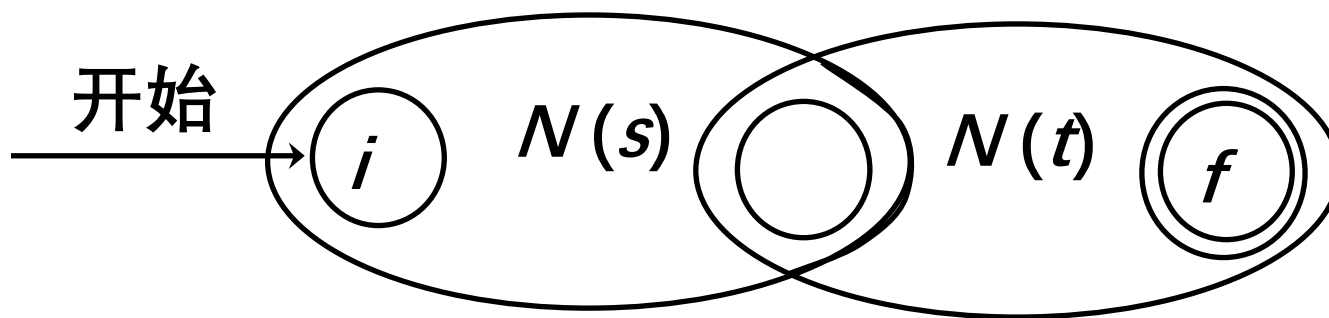


识别正则表达式 $s \mid t$ 的NFA



- 构造识别主算符为连接的正则表达式的NFA

❖ 重要特点：仅一个接受状态，它没有向外的转换

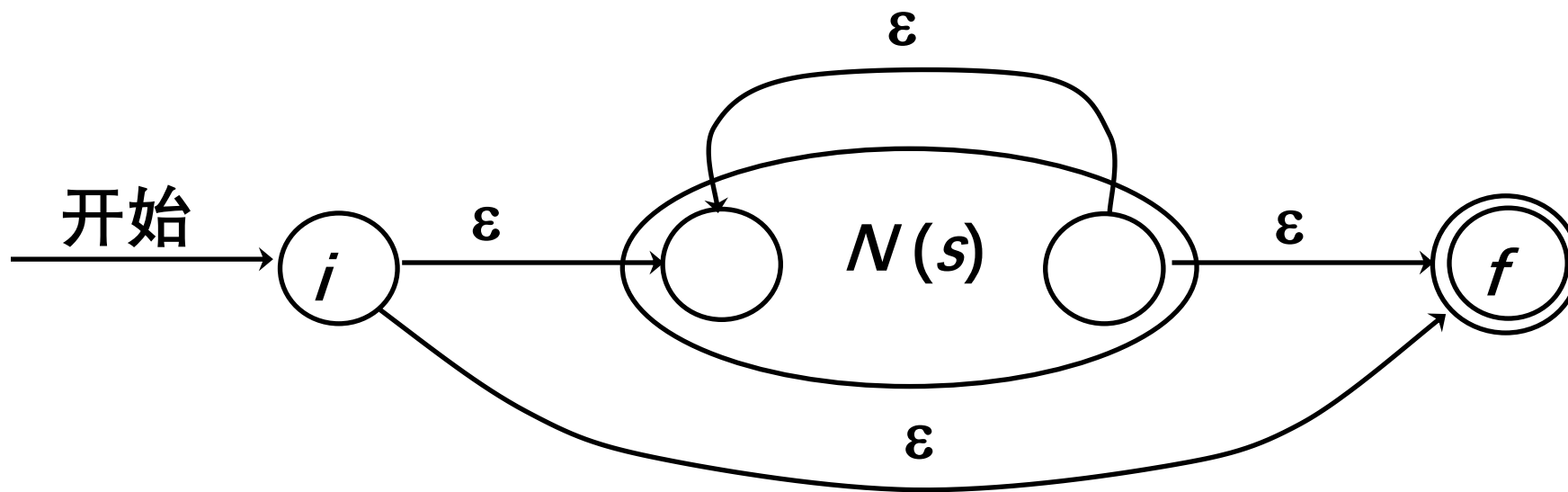


识别正则表达式 st 的NFA



- 构造识别主算符为闭包的正则表达式的NFA

❖ 重要特点：仅一个接受状态，它没有向外的转换

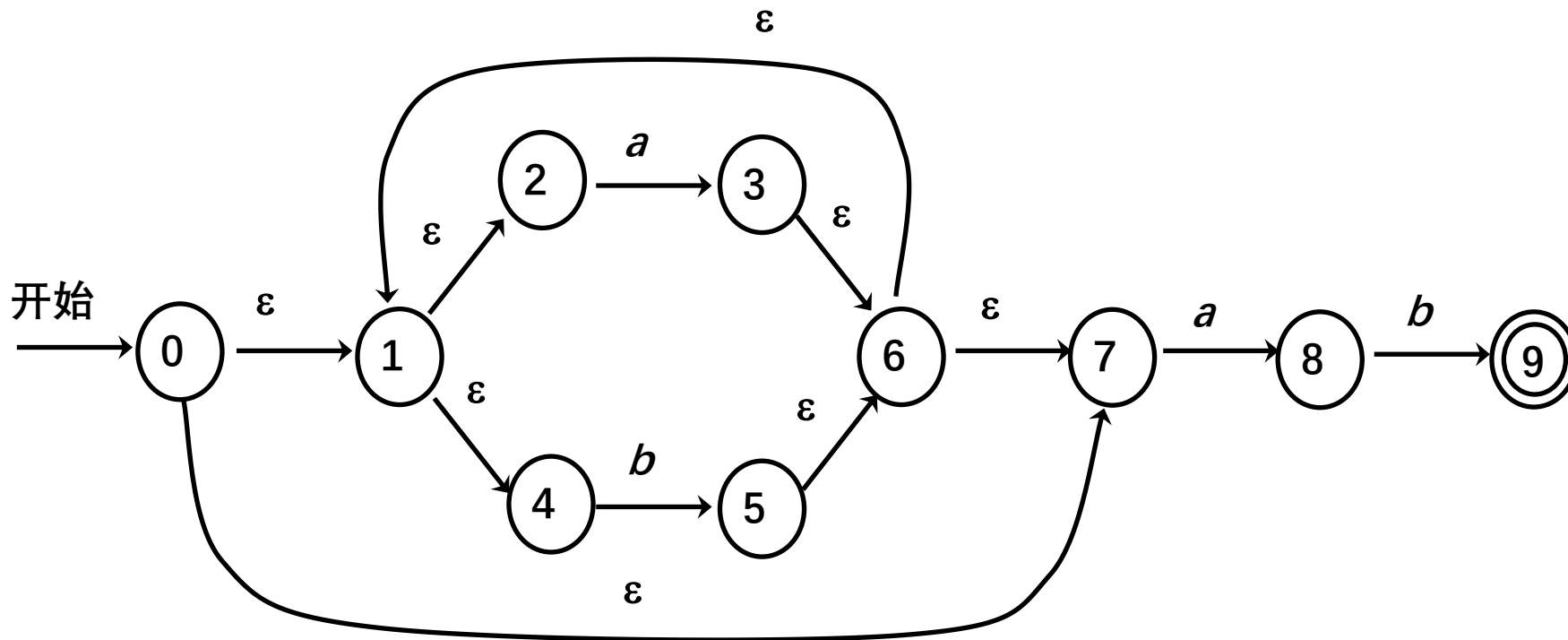


识别正则表达式 s^* 的NFA



- 由本方法产生的NFA具有下列性质:

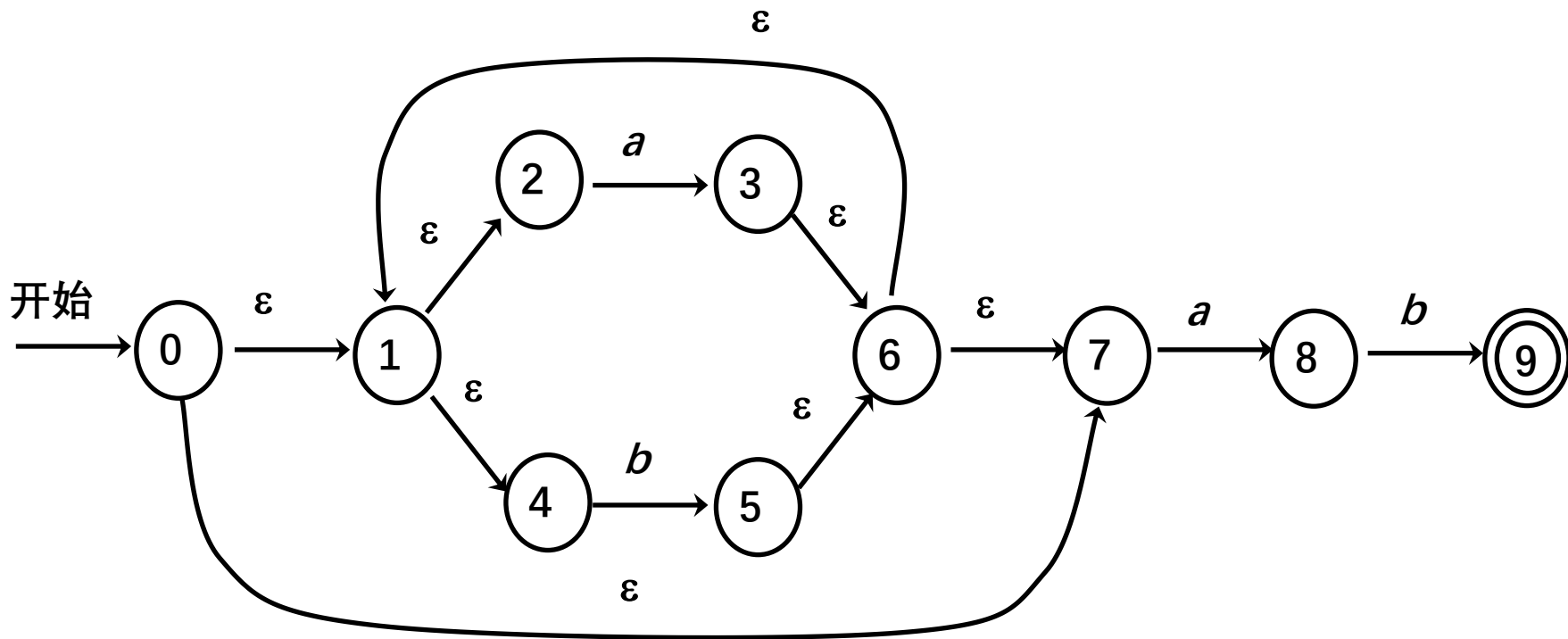
- ❖ $N(r)$ 的状态数最多是 r 中符号和算符总数的两倍
- ❖ $N(r)$ 只有一个接受状态, 接受状态没有向外的转换

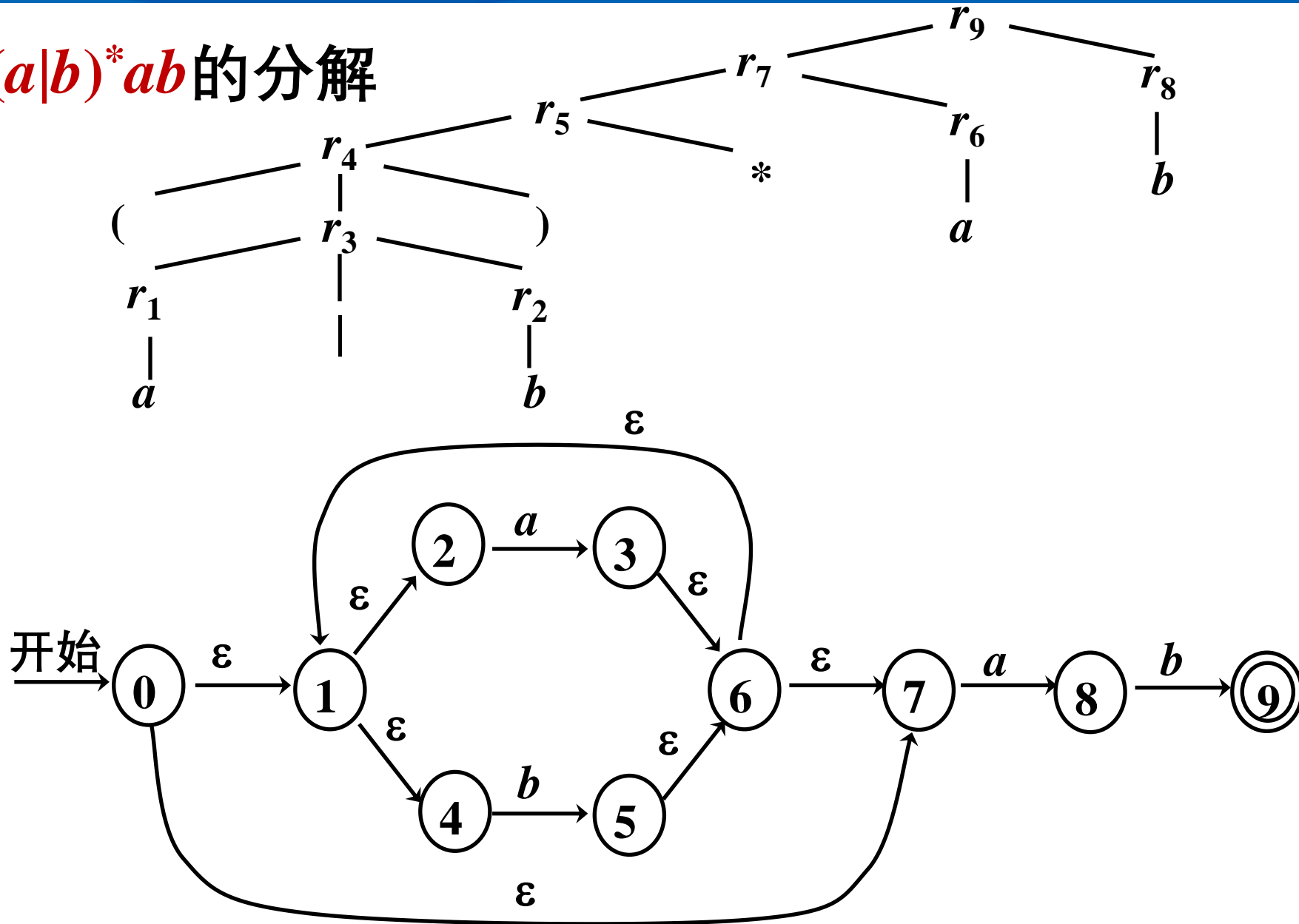




- 由本方法产生的NFA具有下列性质:

- ❖ $N(r)$ 的每个状态有 (1) 一个其标号为 Σ 中符号的指向其它状态的转换, 或者 (2) 最多两个指向其它状态的 ϵ 转换



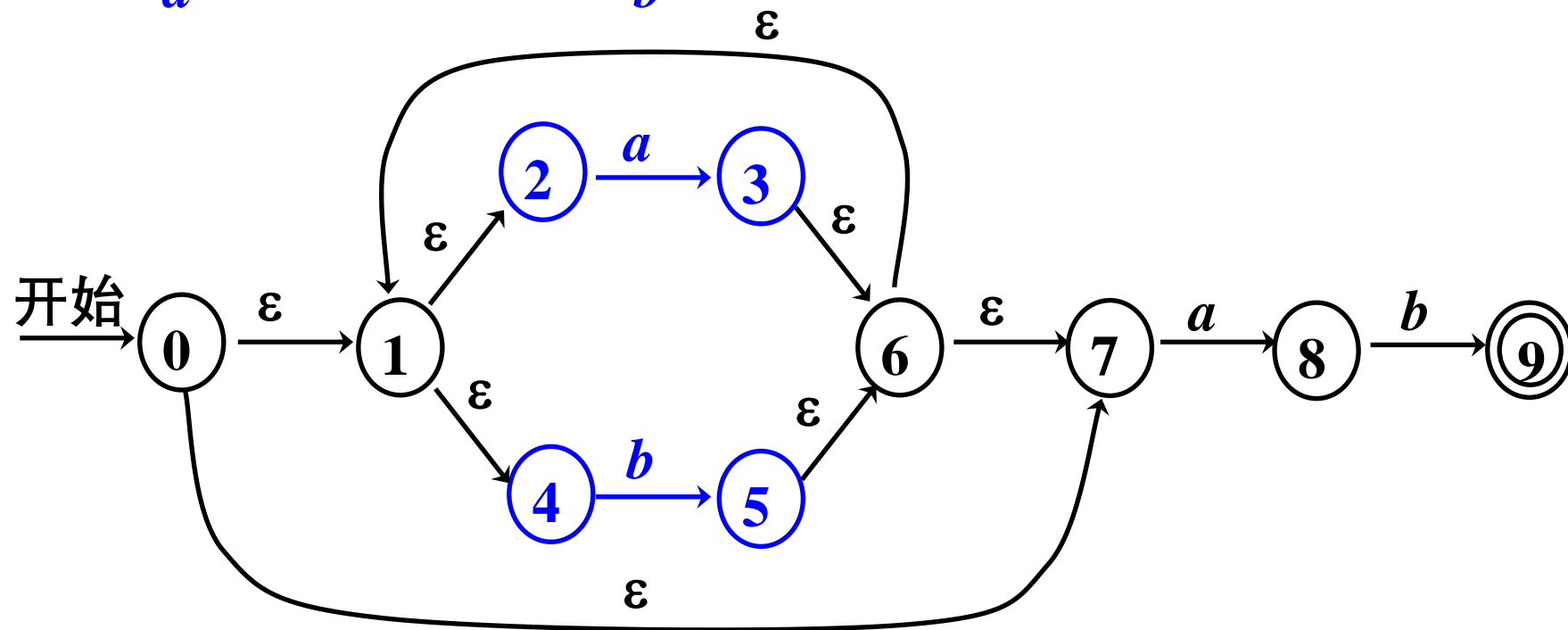
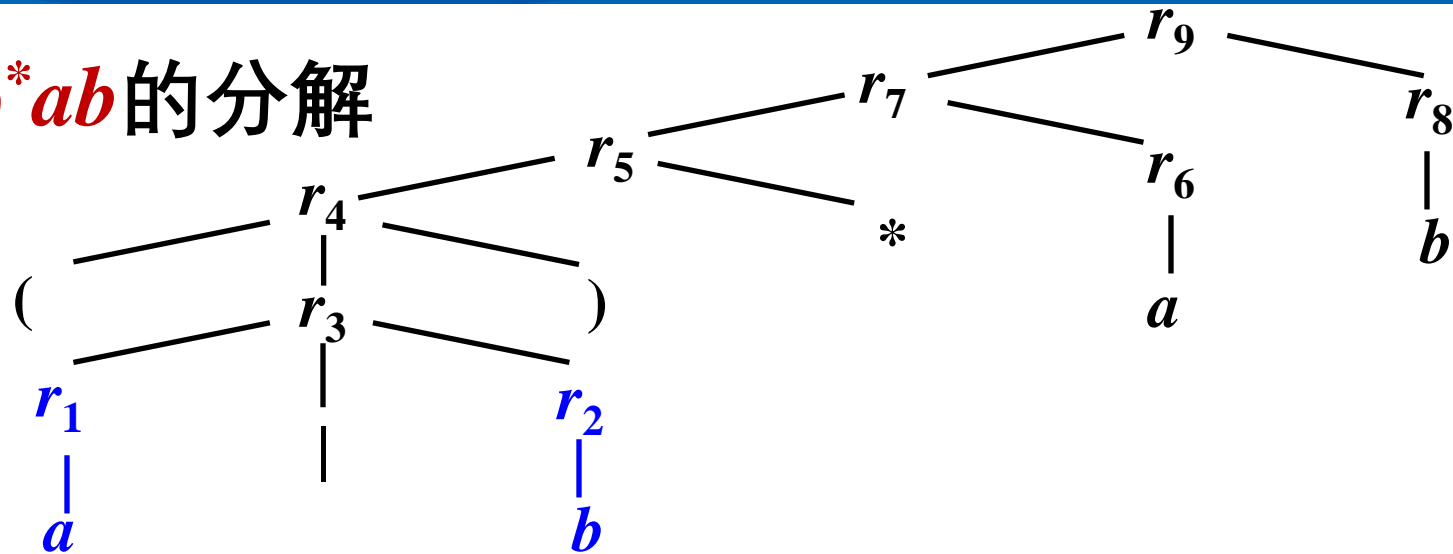




NFA构造过程举例



$(a|b)^*ab$ 的分解

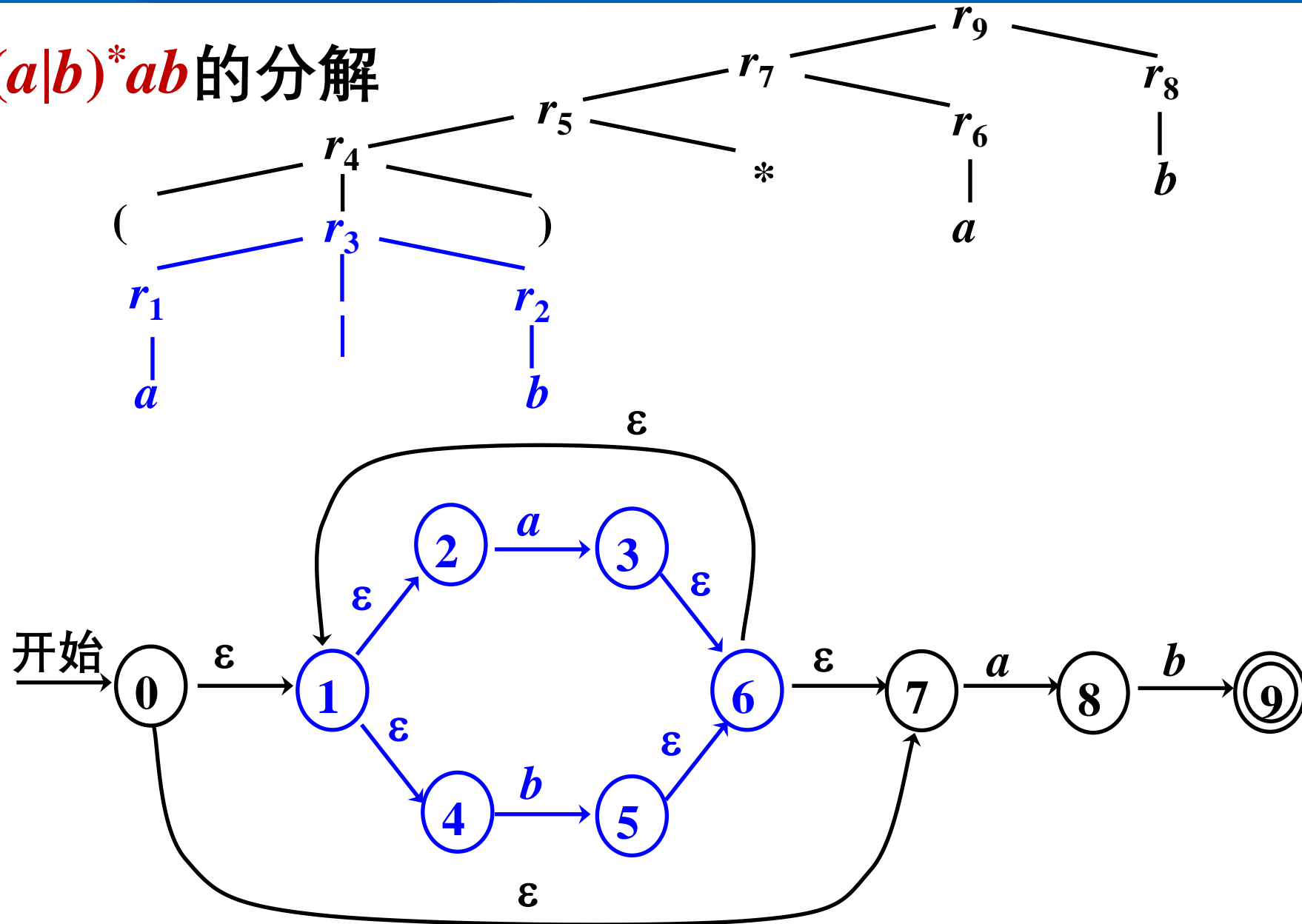




NFA构造过程举例



$(a|b)^*ab$ 的分解

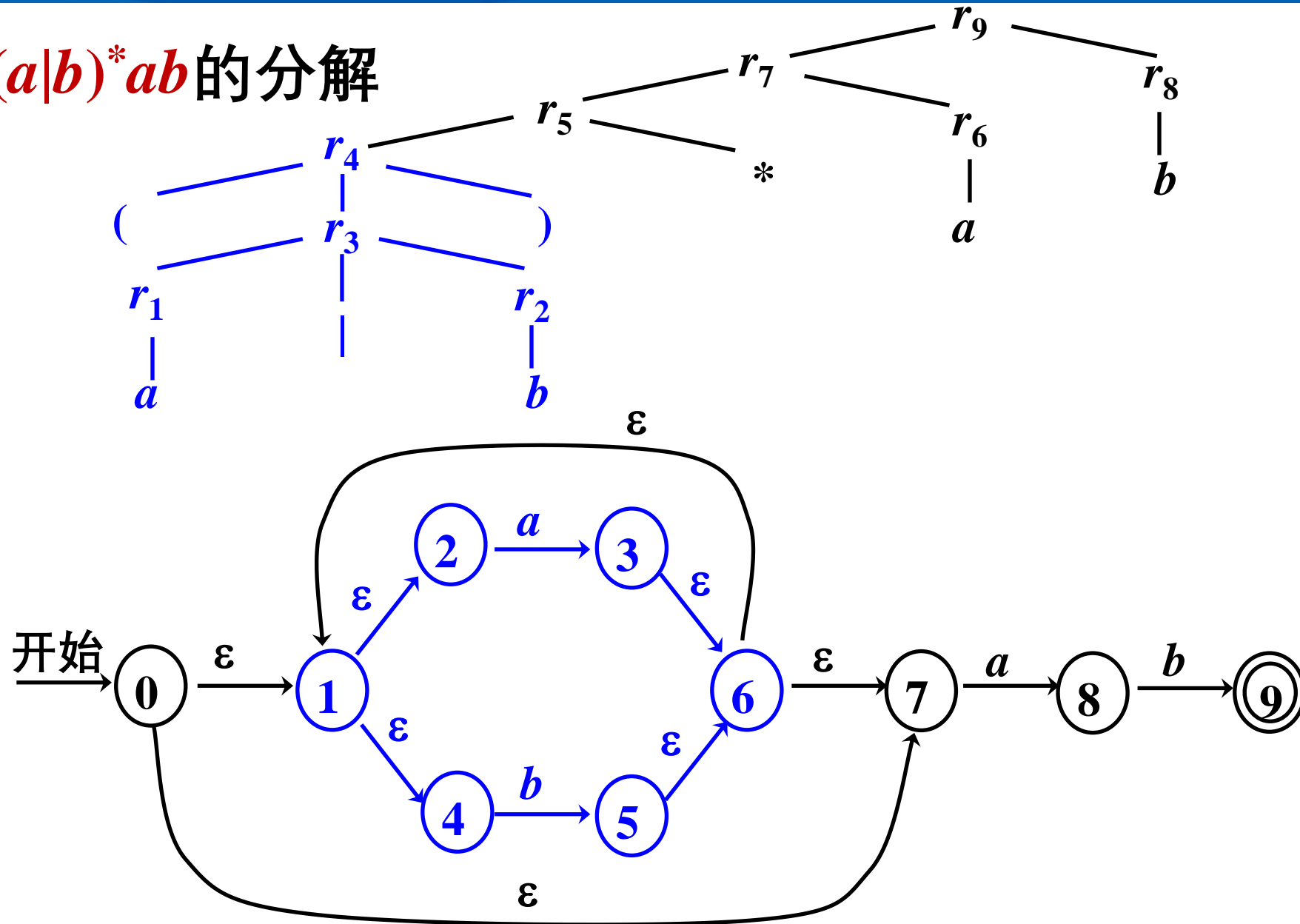




NFA构造过程举例



$(a|b)^*ab$ 的分解

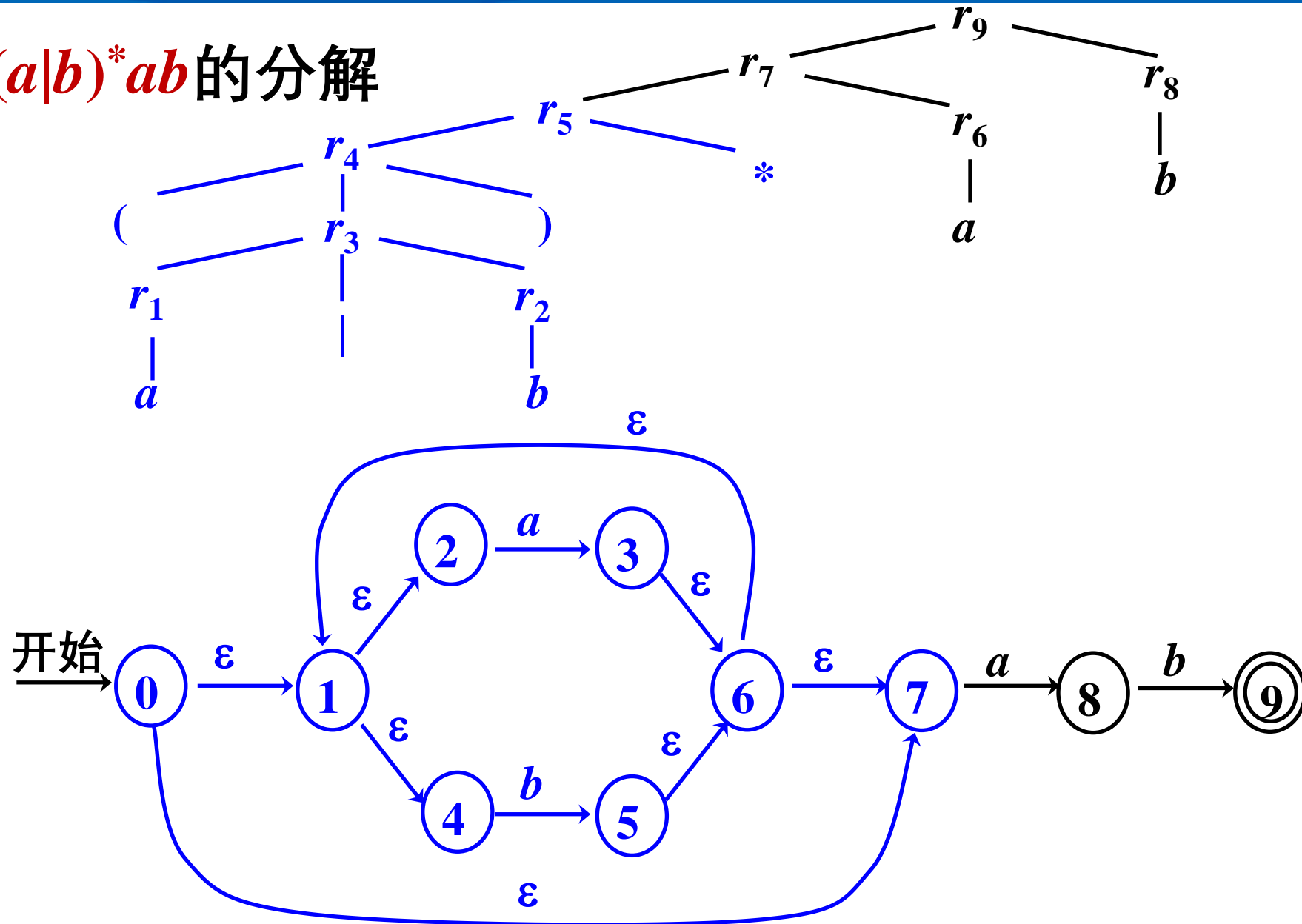




NFA构造过程举例



$(a|b)^*ab$ 的分解

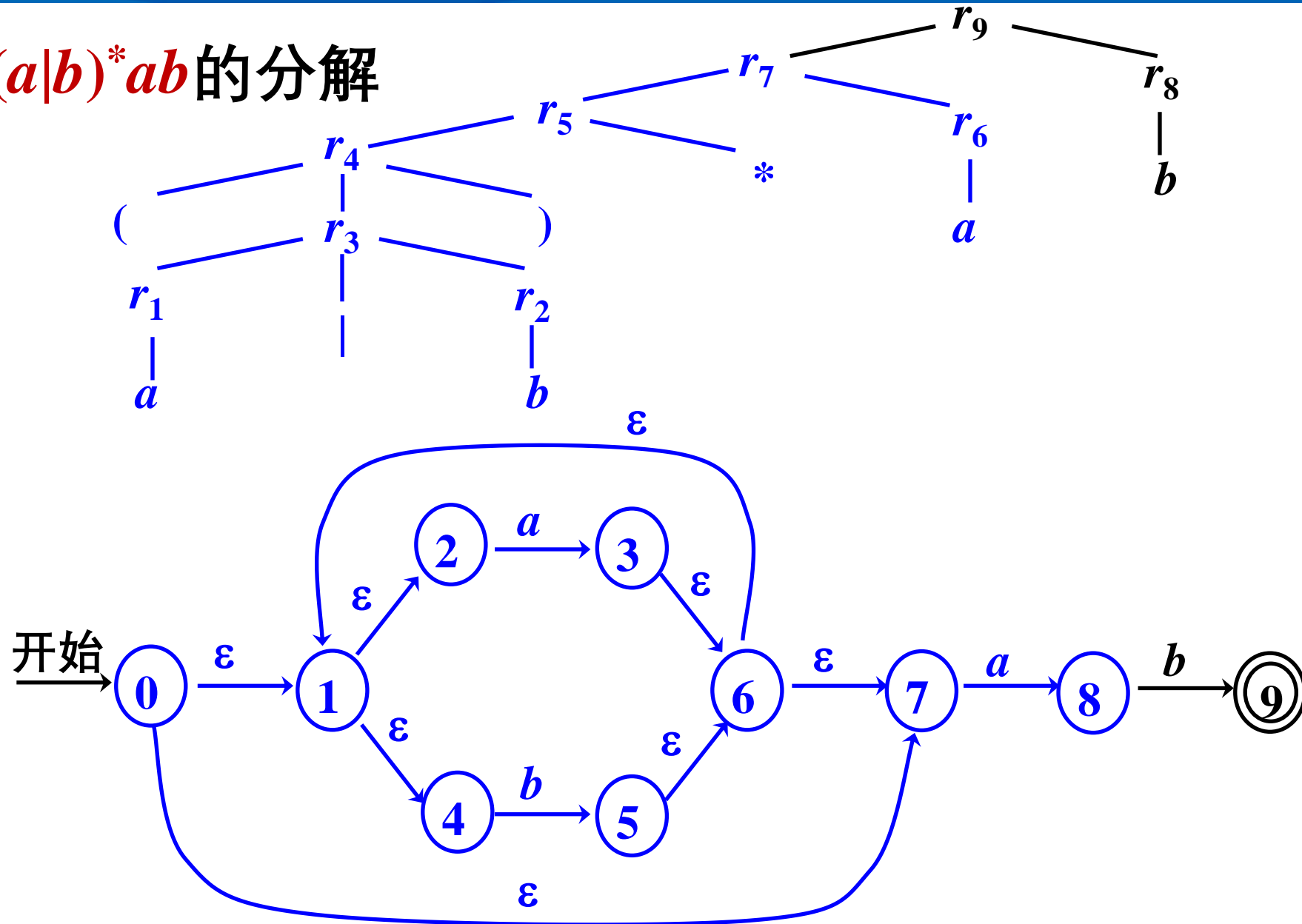




NFA构造过程举例



$(a|b)^*ab$ 的分解

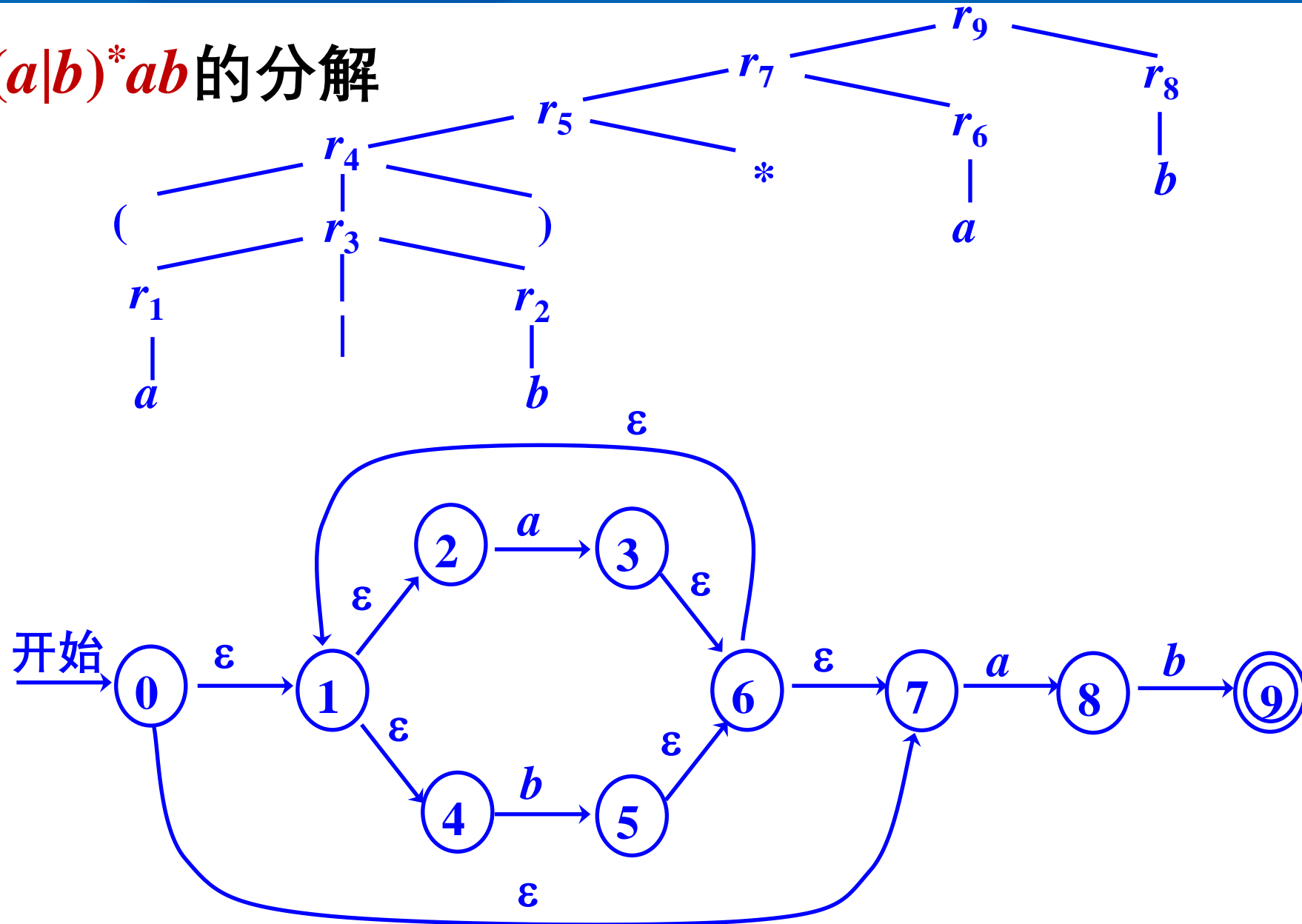




NFA构造过程举例



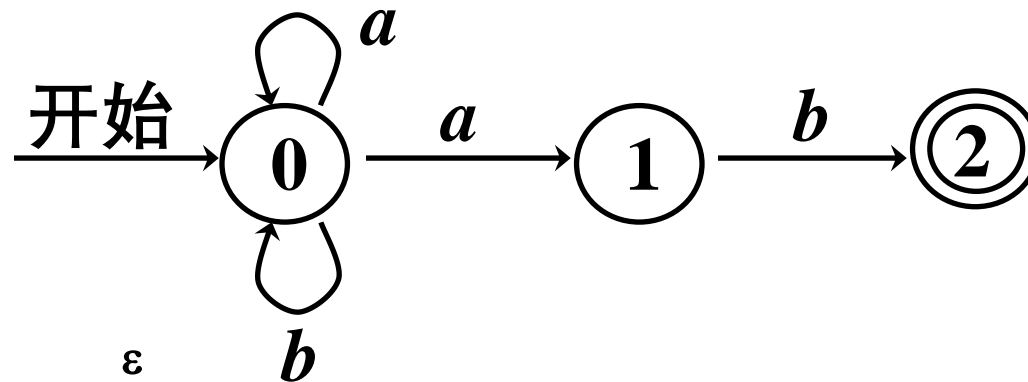
$(a|b)^*ab$ 的分解



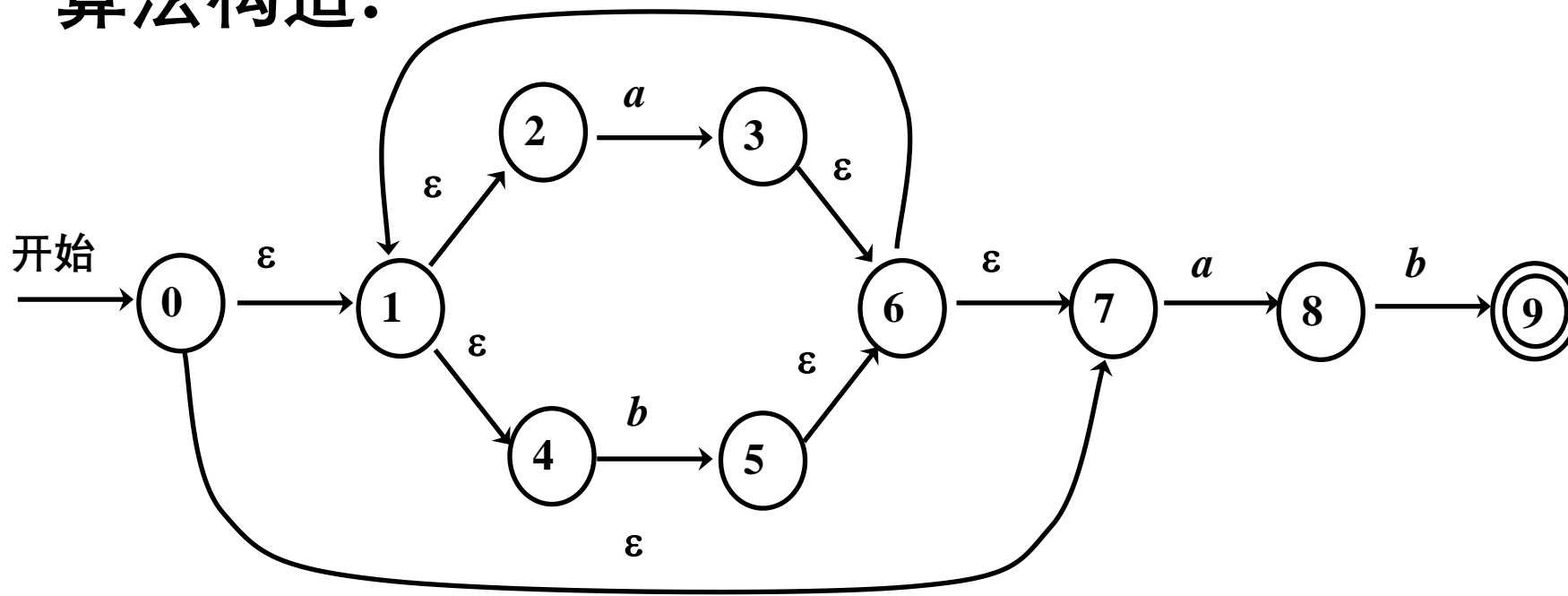


• $(a|b)^*ab$ 的两个NFA的比较

手工构造:



算法构造:



三个问题!



- ❑ 如何进行正则表达式→**NFA**的转换
- ❑ 如何进行**NFA** → **DFA**的转换
- ❑ 如何进行**DFA** →化简**DFA**的转换



• 子集构造法

❖ DFA的一个状态是NFA的一个状态集合

❖ 读了输入 $a_1 a_2 \dots a_n$ 后,

NFA能到达的所有状态: s_1, s_2, \dots, s_k , 则

DFA到达状态 $\{s_1, s_2, \dots, s_k\}$



- 子集构造法(subset construction)

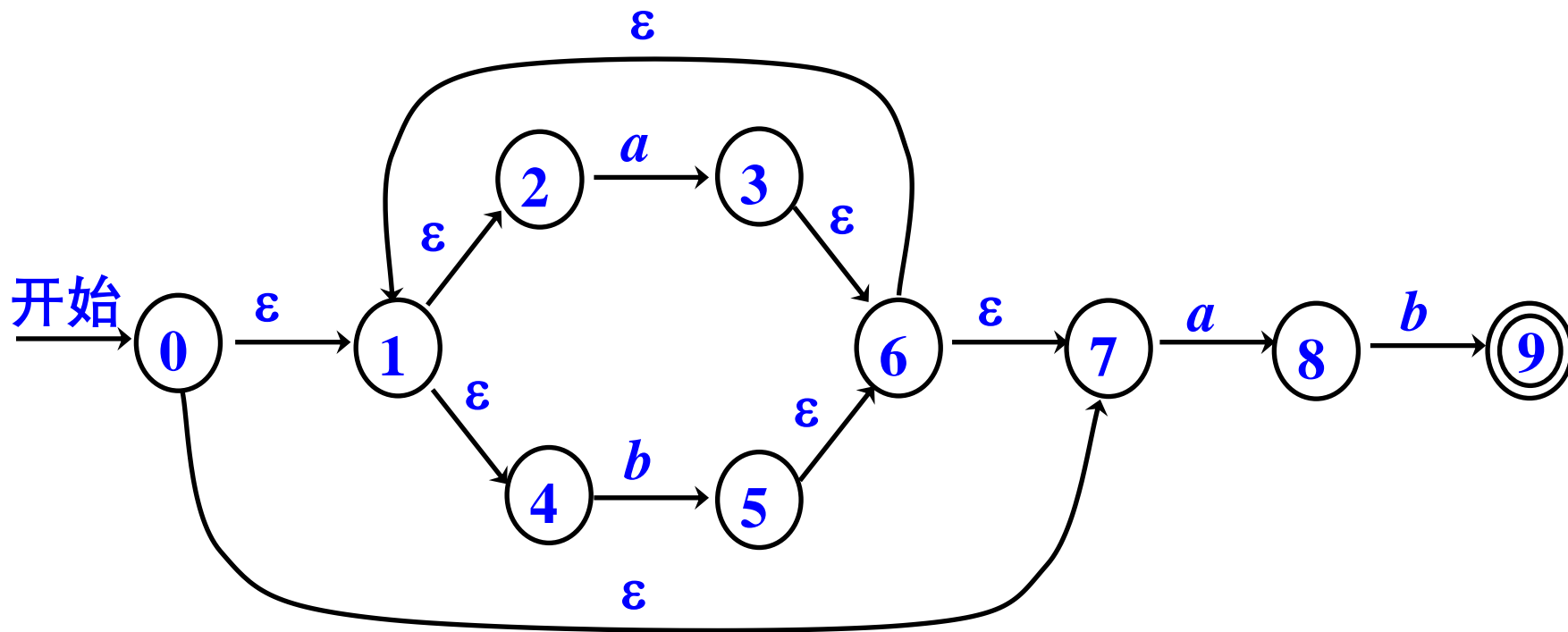
- ❖ ϵ -闭包 (ϵ -closure): 状态 s 的 ϵ -闭包是 s 经 ϵ 转换所能到达的状态集合
- ❖ NFA的初始状态的 ϵ -闭包对应于DFA的初始状态
- ❖ 针对每个DFA状态 – NFA状态子集 A , 求输入每个 a_i 后能到达的NFA状态的 ϵ -闭包并集 (ϵ -closure(move(A, a_i))), 该集合对应于DFA中的一个已有状态, 或者是一个要新加的DFA状态



NFA到DFA的变换

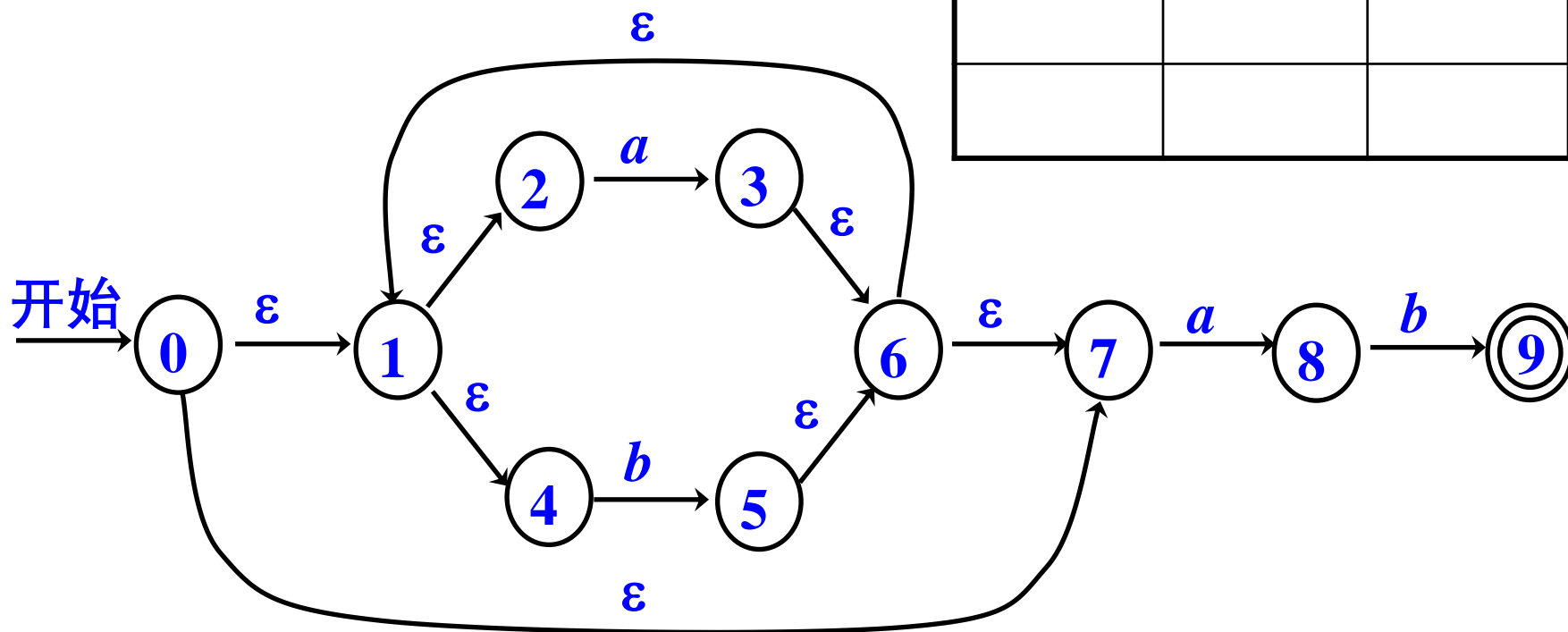


- 例 $(a|b)^*ab$, NFA如下, 把它变换为DFA





NFA到DFA的变换



状态	输入符号	
	a	b

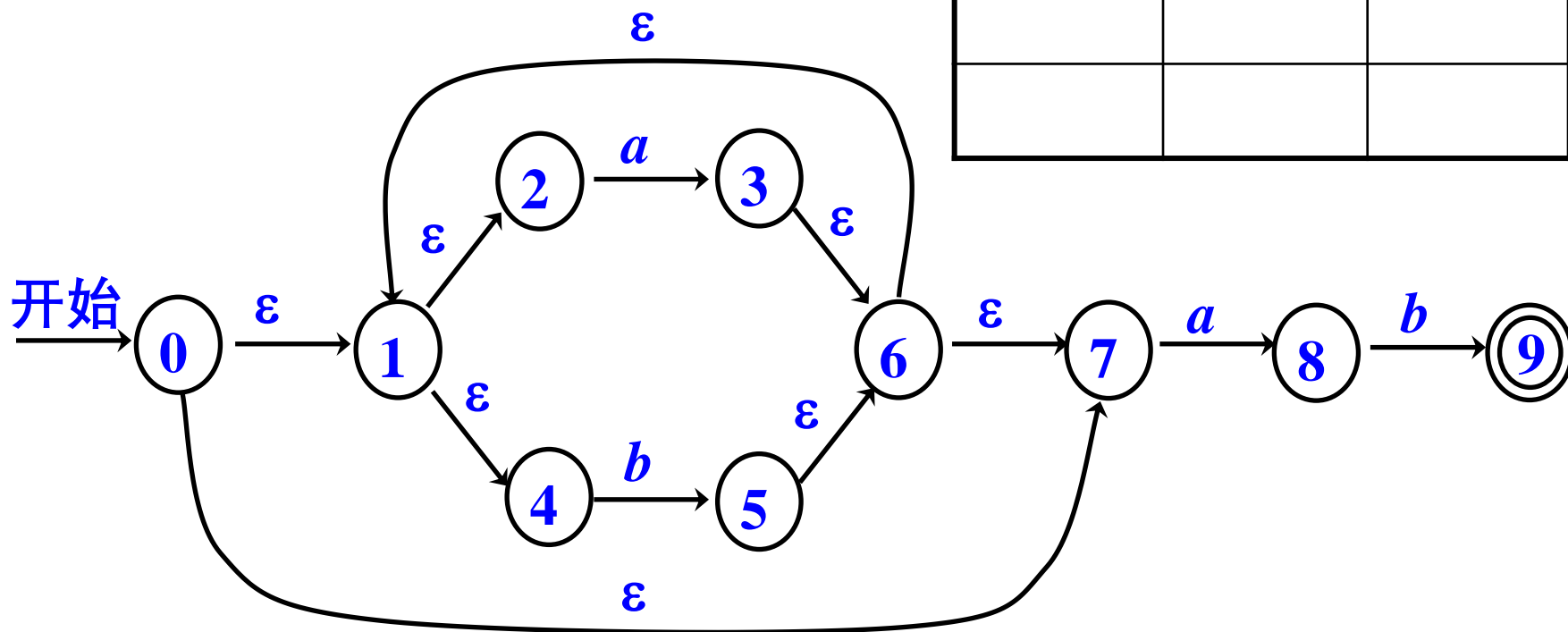


NFA到DFA的变换



$$A = \{0, 1, 2, 4, 7\}$$

状态	输入符号	
	<i>a</i>	<i>b</i>
A		





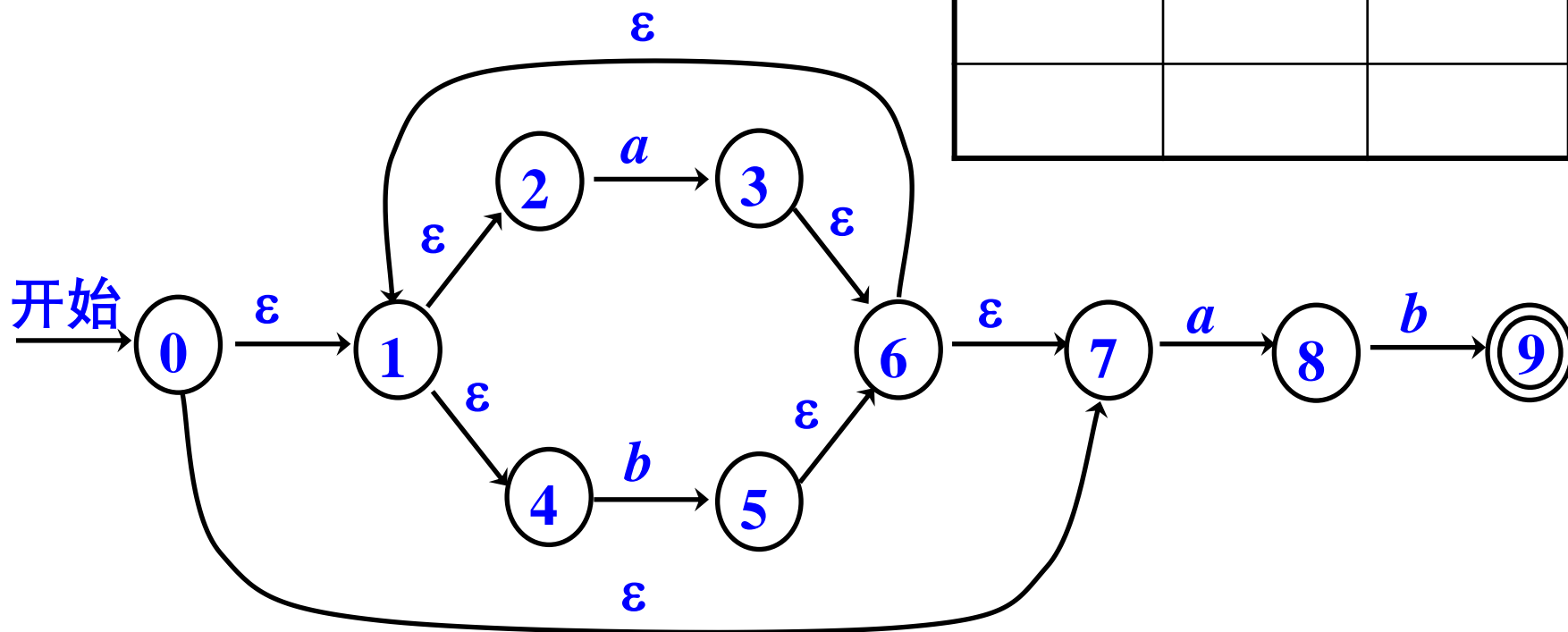
NFA到DFA的变换



$A = \{0, 1, 2, 4, 7\}$

$B = \{1, 2, 3, 4, 6, 7, 8\}$

状态	输入符号	
	a	b
A	B	





NFA到DFA的变换

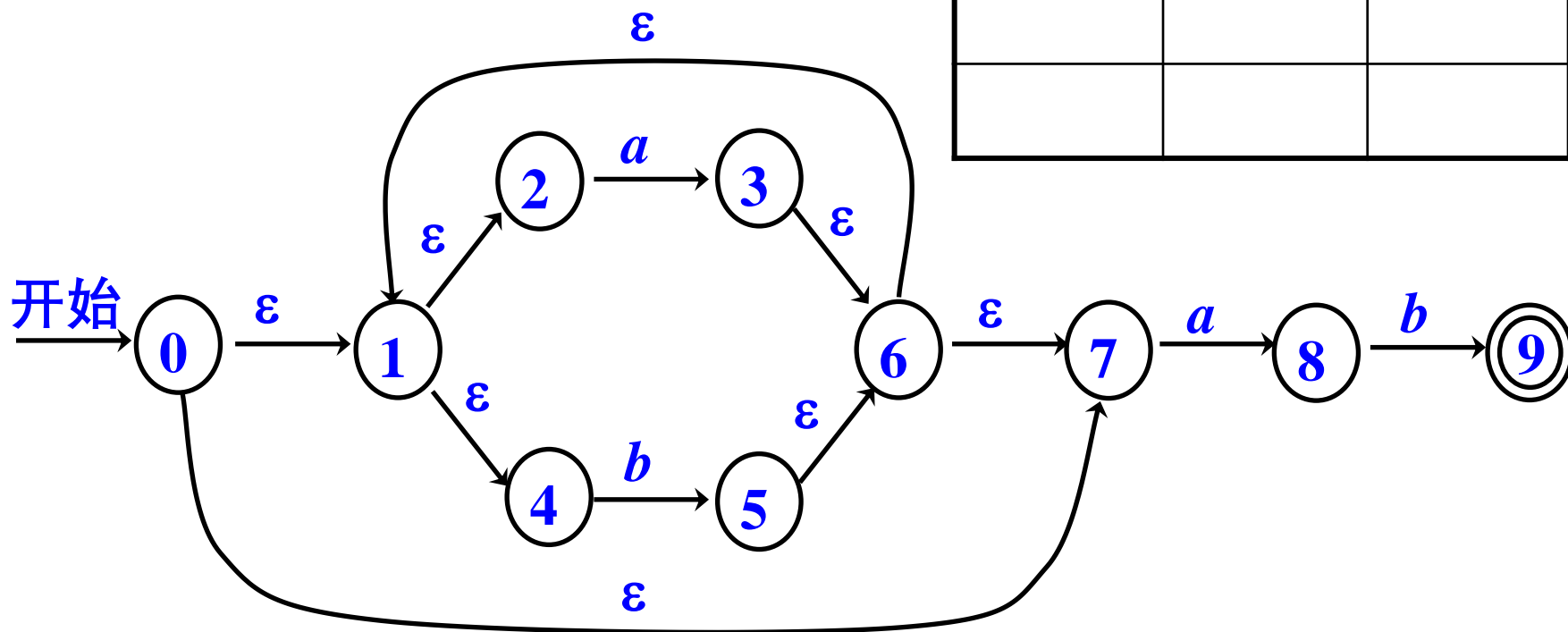


$A = \{0, 1, 2, 4, 7\}$

$B = \{1, 2, 3, 4, 6, 7, 8\}$

$C = \{1, 2, 4, 5, 6, 7\}$

状态	输入符号	
	a	b
A	B	C





NFA到DFA的变换

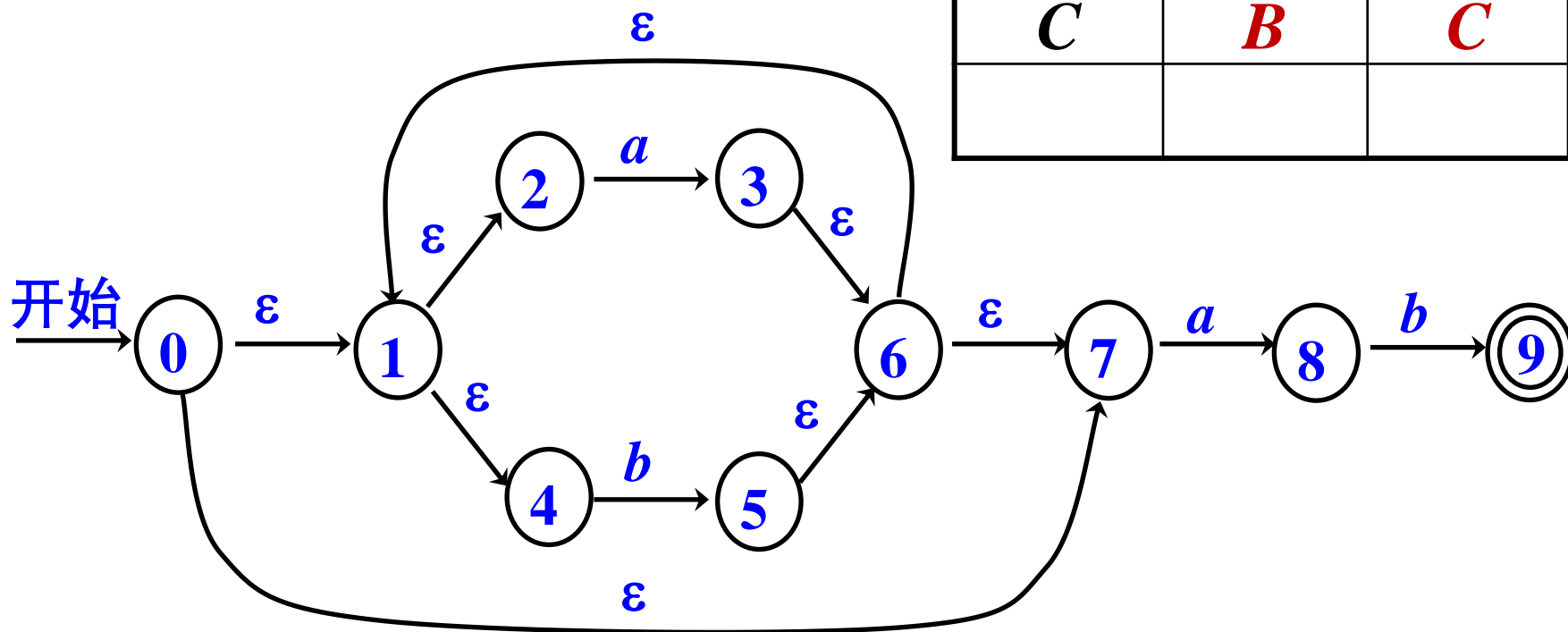


$$A = \{0, 1, 2, 4, 7\}$$

$$B = \{1, 2, 3, 4, 6, 7, 8\}$$

$$C = \{1, 2, 4, 5, 6, 7\}$$

状态	输入符号	
	<i>a</i>	<i>b</i>
<i>A</i>	<i>B</i>	<i>C</i>
<i>B</i>	<i>B</i>	
<i>C</i>	<i>B</i>	<i>C</i>





NFA到DFA的变换



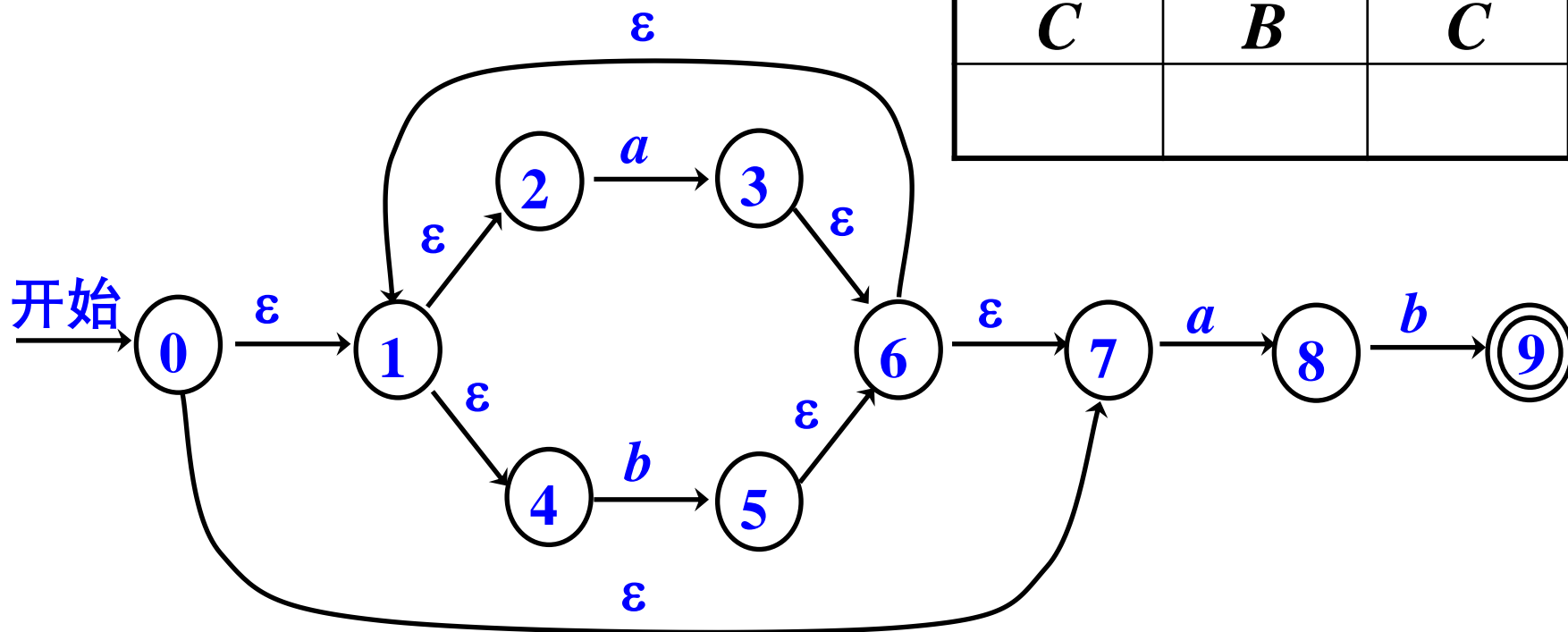
$A = \{0, 1, 2, 4, 7\}$

$B = \{1, 2, 3, 4, 6, 7, 8\}$

$C = \{1, 2, 4, 5, 6, 7\}$

$D = \{1, 2, 4, 5, 6, 7, 9\}$

状态	输入符号	
	a	b
A	B	C
B	B	
C	B	C





NFA到DFA的变换



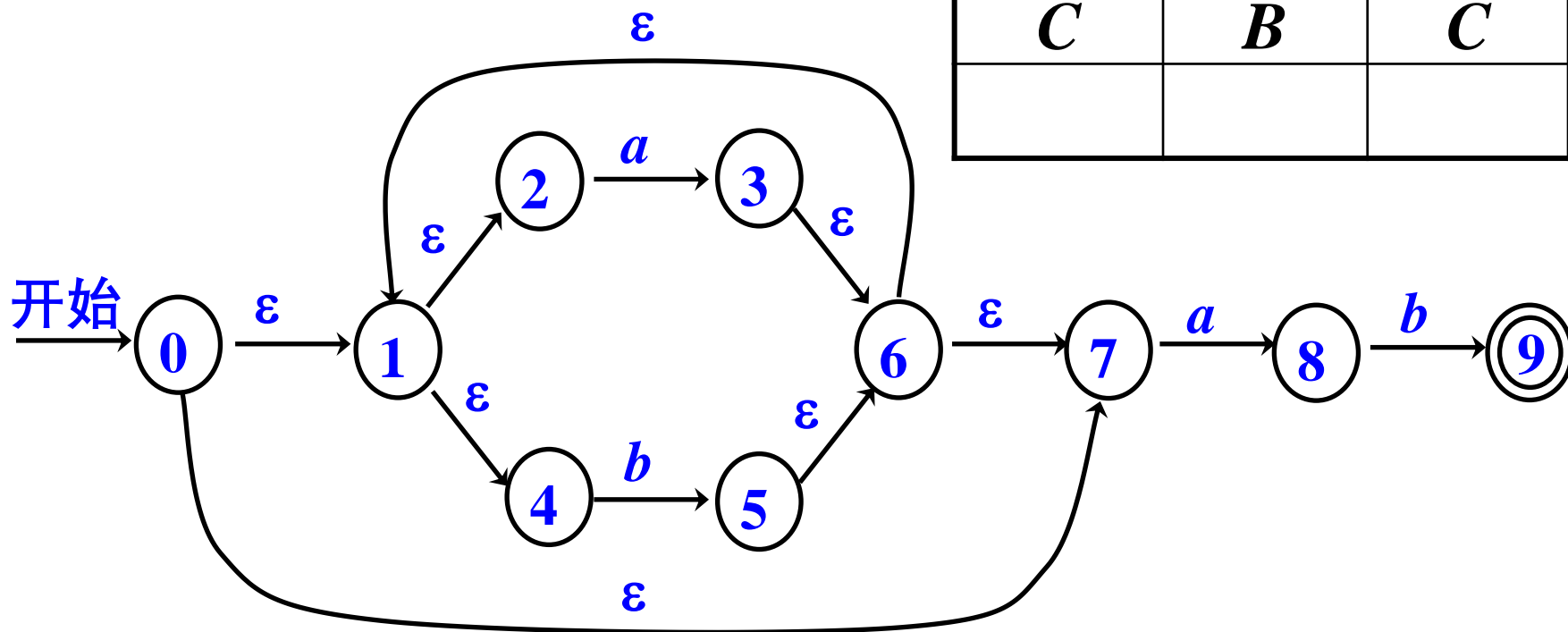
$$A = \{0, 1, 2, 4, 7\}$$

$$B = \{1, 2, 3, 4, 6, 7, 8\}$$

$$C = \{1, 2, 4, 5, 6, 7\}$$

$$D = \{1, 2, 4, 5, 6, 7, 9\}$$

状态	输入符号	
	<i>a</i>	<i>b</i>
<i>A</i>	<i>B</i>	<i>C</i>
<i>B</i>	<i>B</i>	<i>D</i>
<i>C</i>	<i>B</i>	<i>C</i>





NFA到DFA的变换



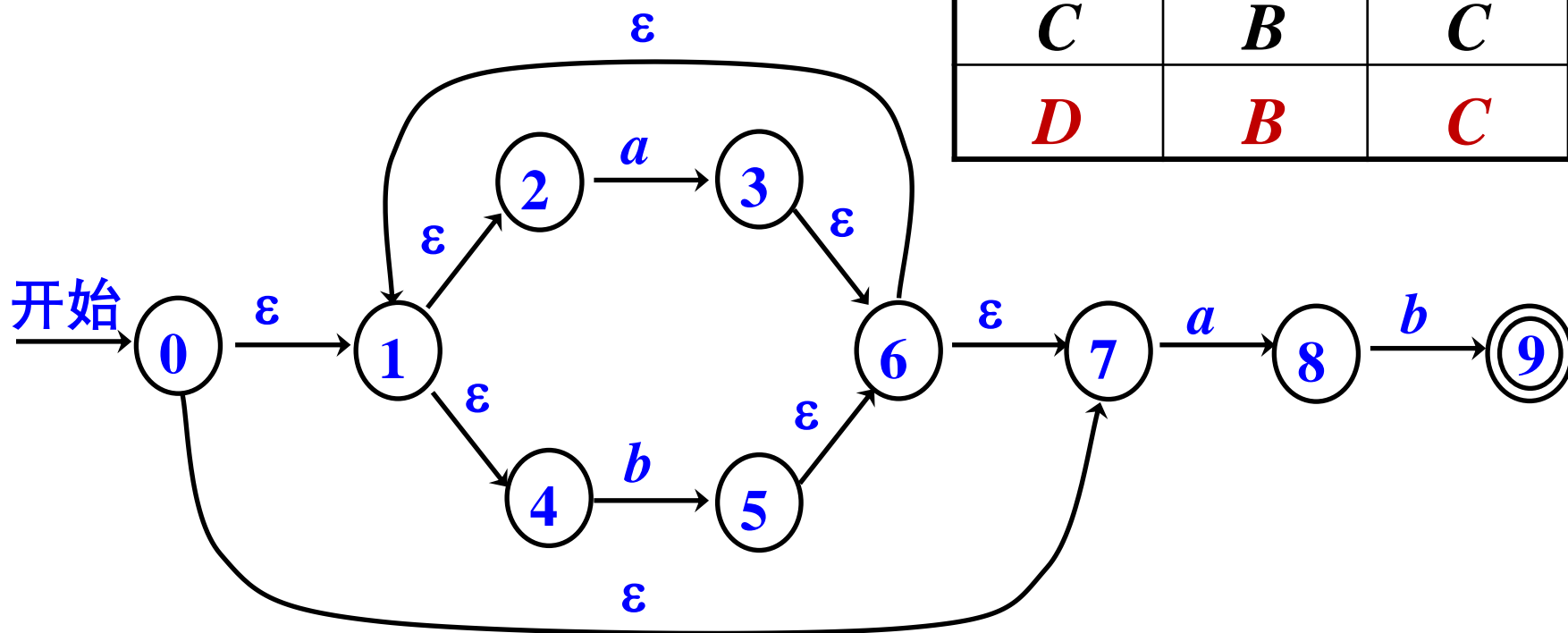
$$A = \{0, 1, 2, 4, 7\}$$

$$B = \{1, 2, 3, 4, 6, 7, 8\}$$

$$C = \{1, 2, 4, 5, 6, 7\}$$

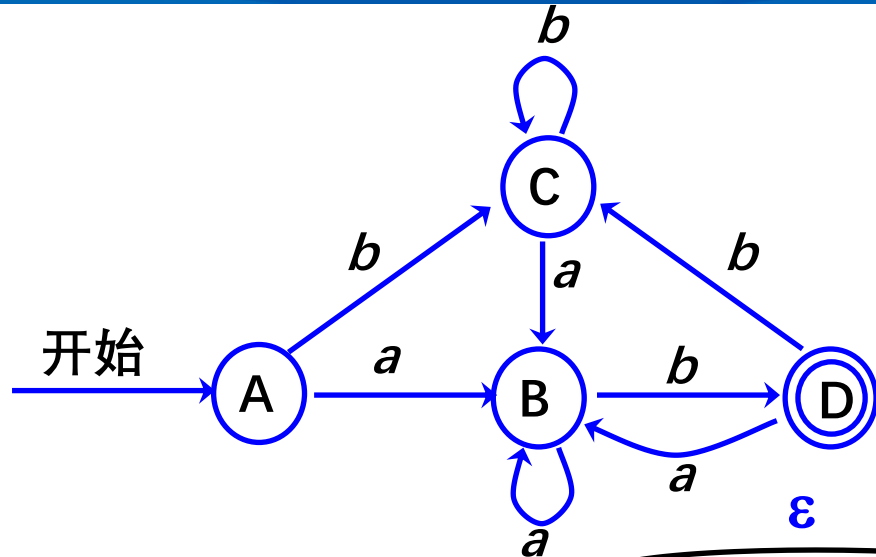
$$D = \{1, 2, 4, 5, 6, 7, 9\}$$

状态	输入符号	
	<i>a</i>	<i>b</i>
<i>A</i>	<i>B</i>	<i>C</i>
<i>B</i>	<i>B</i>	<i>D</i>
<i>C</i>	<i>B</i>	<i>C</i>
<i>D</i>	<i>B</i>	<i>C</i>

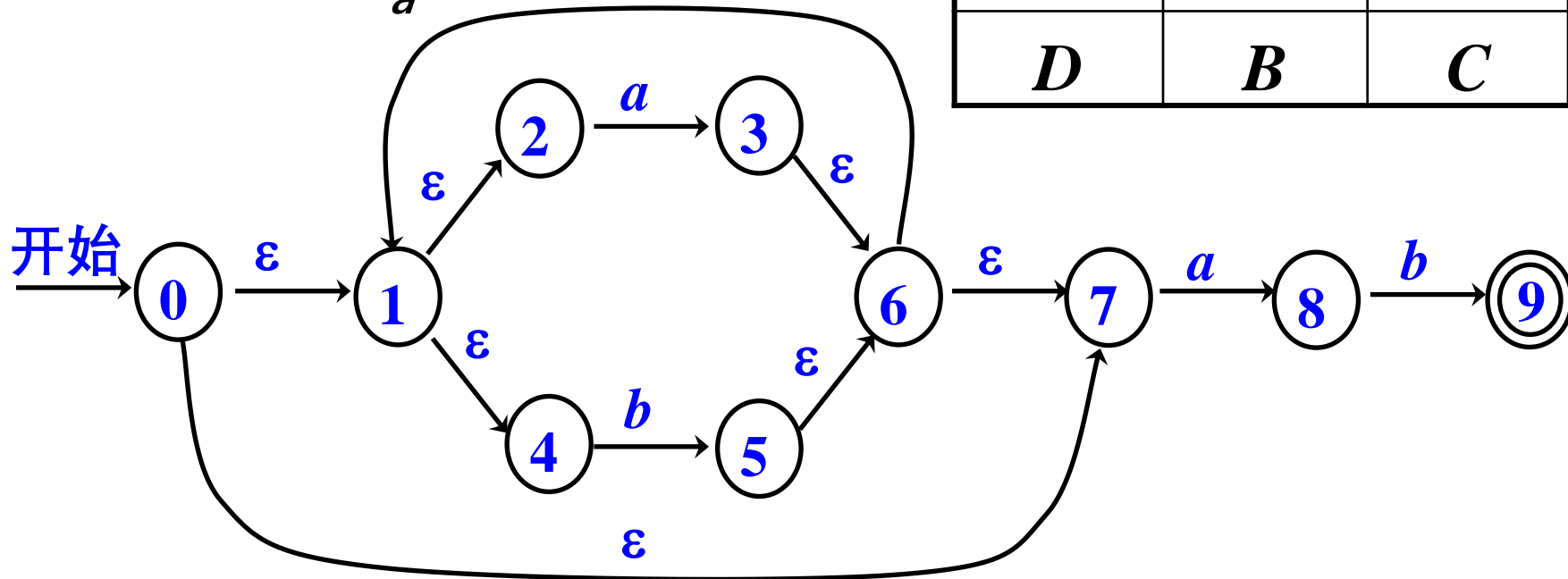




NFA到DFA的变换



状态	输入符号	
	<i>a</i>	<i>b</i>
<i>A</i>	<i>B</i>	<i>C</i>
<i>B</i>	<i>B</i>	<i>D</i>
<i>C</i>	<i>B</i>	<i>C</i>
<i>D</i>	<i>B</i>	<i>C</i>

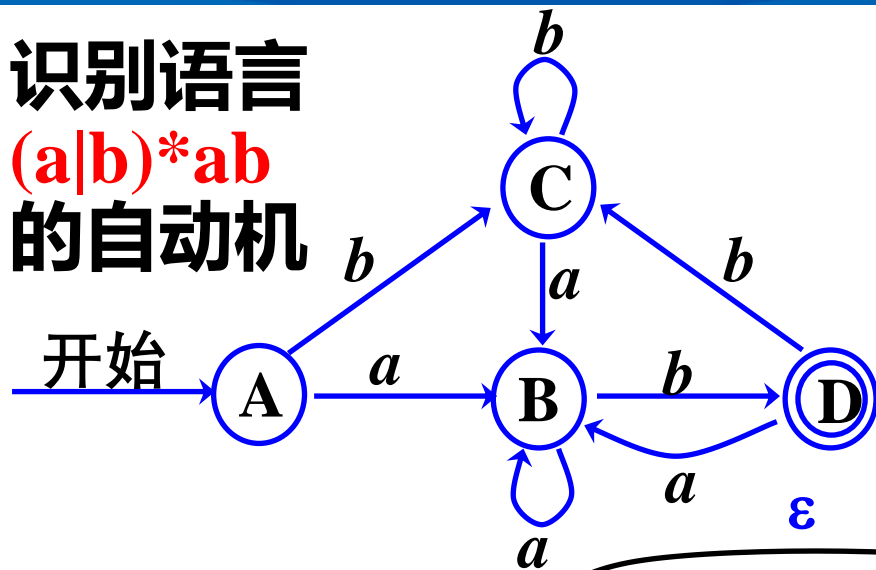




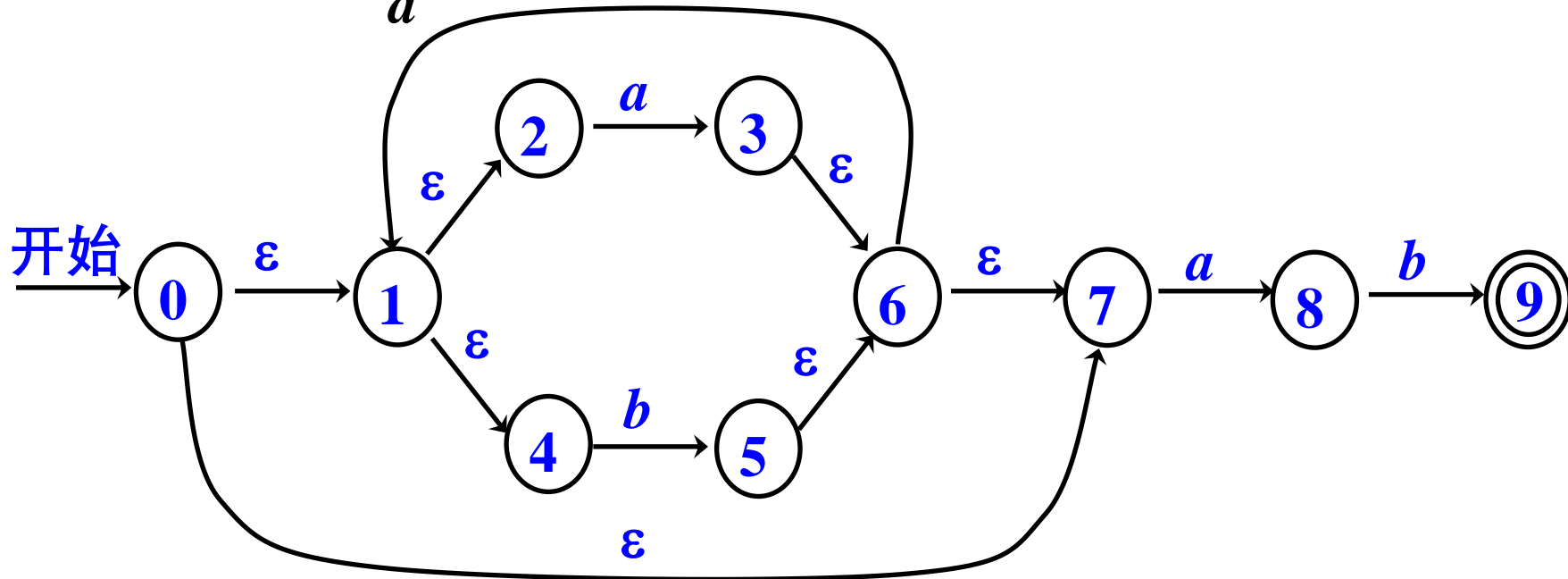
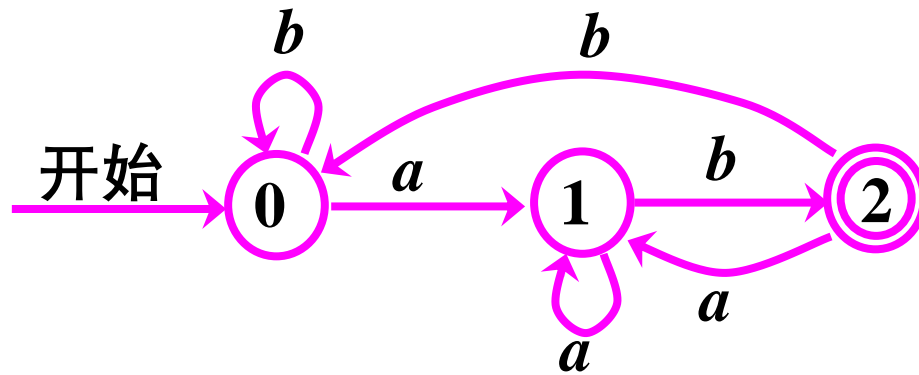
NFA到DFA的变换



识别语言
 $(a|b)^*ab$
的自动机



子集构造法不一定得到最简DFA



三个问题!



- 如何进行正则表达式→**NFA**的转换
- 如何进行**NFA** → **DFA**的转换
- 如何进行**DFA** →化简**DFA**的转换



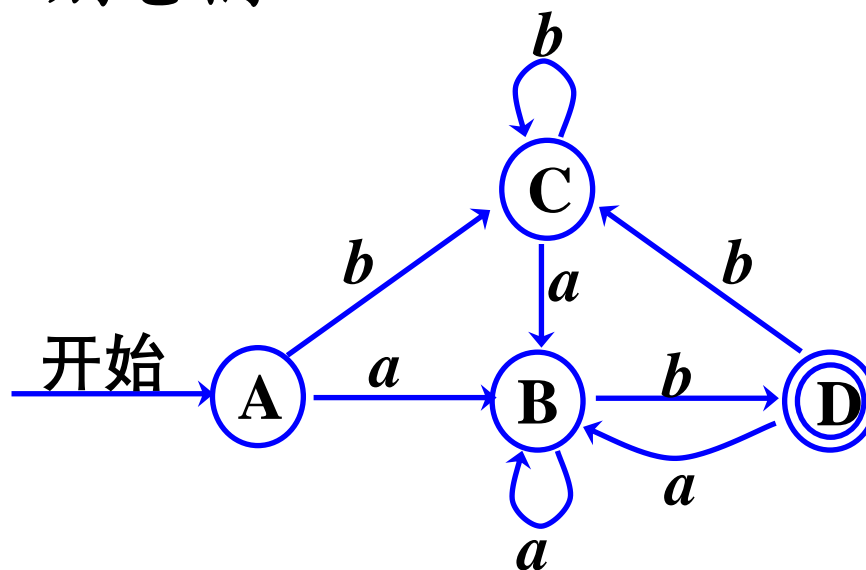
- **A和B是可区别的状态**

- ❖ 从A出发，读过单字符b构成的串，到达非接受状态C，而从B出发，读过串b，到达接受状态D

- **A和C是不可区别的状态**

- ❖ 无任何串可用来像上面这样区别它们

可区别的状态要
分开对待



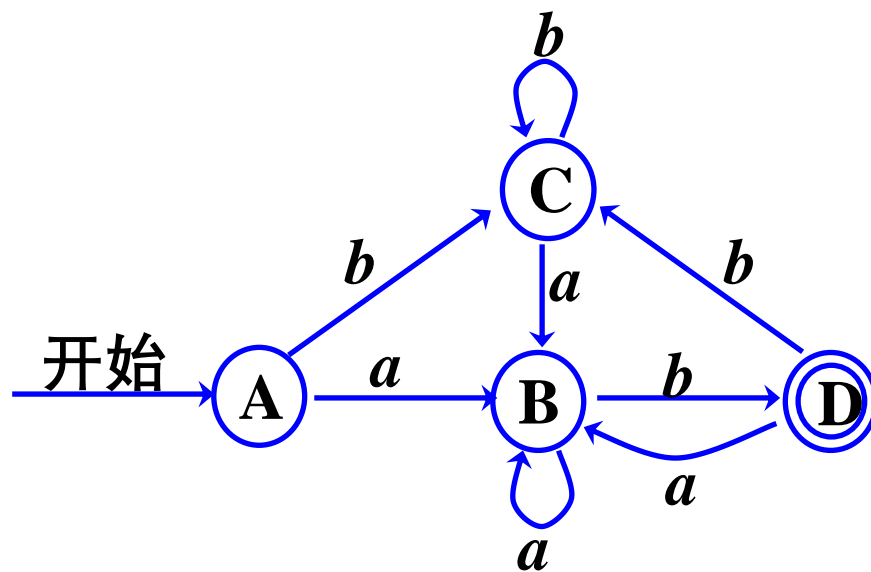


• 1. 按是否是接受状态来区分

$\{A, B, C\}, \{D\}$

$\text{move}(\{A, B, C\}, a) = \{B\}$

$\text{move}(\{A, B, C\}, b) = \{C, D\}$





• 1. 按是否是接受状态来区分

$\{A, B, C\}, \{D\}$

$\text{move}(\{A, B, C\}, a) = \{B\}$

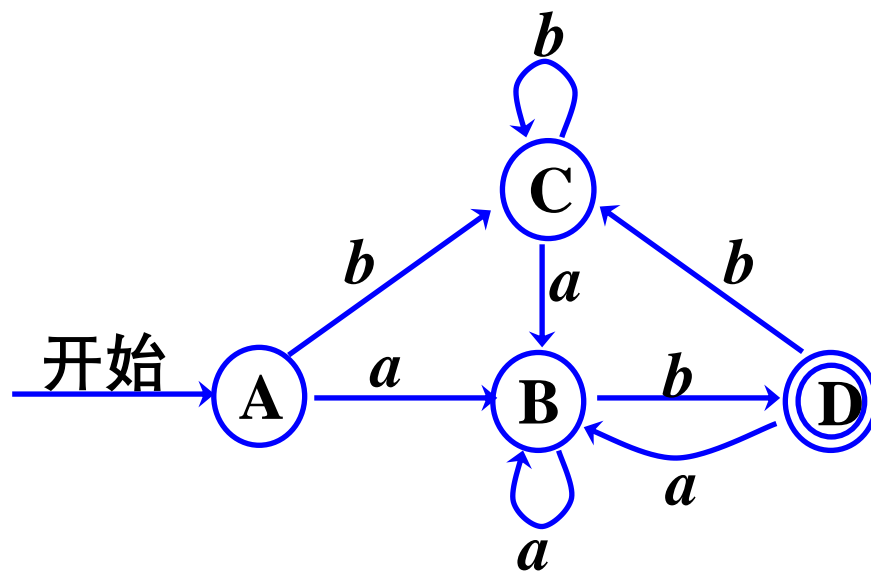
$\text{move}(\{A, B, C\}, b) = \{C, D\}$

• 2. 继续分解

$\{A, C\}, \{B\}, \{D\}$

$\text{move}(\{A, C\}, a) = \{B\}$

$\text{move}(\{A, C\}, b) = \{C\}$



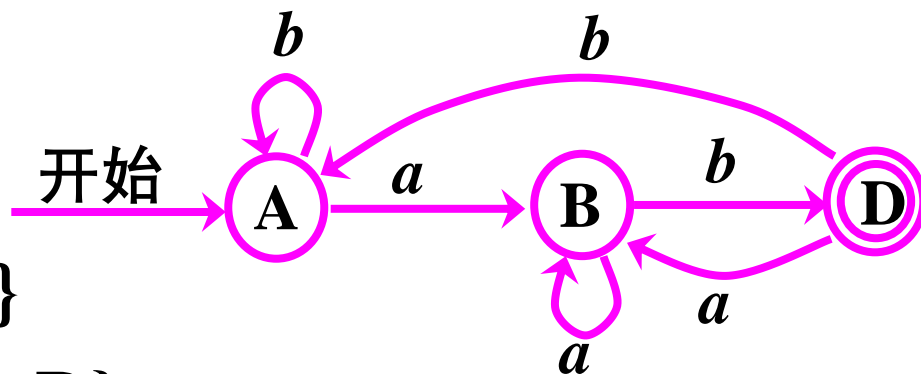


• 1. 按是否是接受状态来区分

$\{A, B, C\}, \{D\}$

$\text{move}(\{A, B, C\}, a) = \{B\}$

$\text{move}(\{A, B, C\}, b) = \{C, D\}$

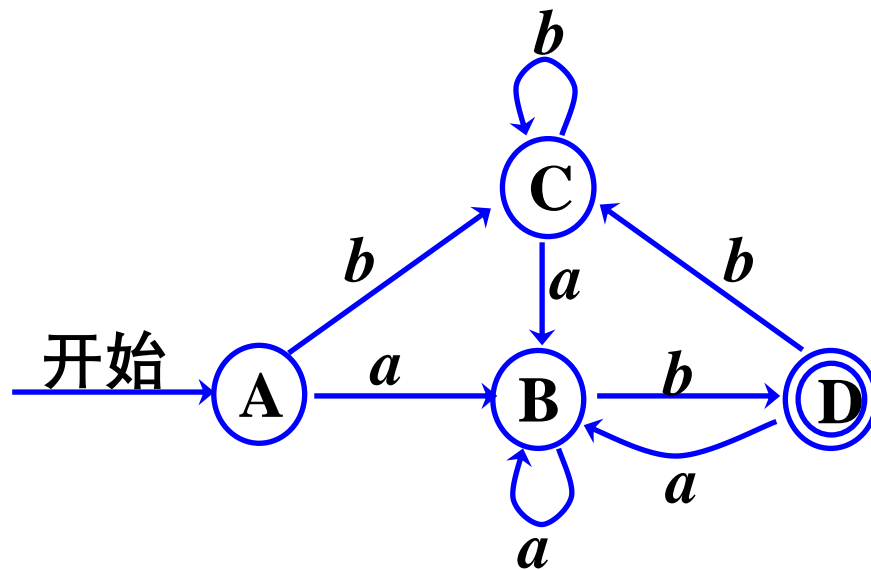


• 2. 继续分解

$\{A, C\}, \{B\}, \{D\}$

$\text{move}(\{A, C\}, a) = \{B\}$

$\text{move}(\{A, C\}, b) = \{C\}$





思考问题



- 正则表达式 $(a|b)^*$ 与 $(a^*|b^*)^*$ 是否等价?
 - ❖ 提示：可利用其最简化DFA的
- 有限自动机如何实现为代码?
 - 请课外阅读[有限自动机的Python实现样例](#)



- **词法分析器的作用和接口，用高级语言编写词法分析器等内容**
- **掌握下面涉及的一些概念，它们之间转换的技巧、方法或算法**
 - ❖ 非形式描述的语言 \leftrightarrow 正则表达式
 - ❖ 正则表达式 \rightarrow NFA
 - ❖ 非形式描述的语言 \leftrightarrow NFA
 - ❖ NFA \rightarrow DFA
 - ❖ DFA \rightarrow 最简 DFA
 - ❖ 非形式描述的语言 \leftrightarrow DFA（或最简 DFA）



一起努力 打造国产基础软硬件体系！

李 诚

国家高性能计算中心(合肥)、信息与计算机国家级实验教学示范中心

计算机科学与技术学院

2023年09月06日