

## 第三章 词法分析

# 正则表达式

哈尔滨工业大学 陈鄞



# 正则表达式

语言  $L = \{a\}\{a,b\}^*({\varepsilon} \cup (\{.,\_ \}\{a,b\}\{a,b\}^*))$

➤ **正则表达式** (Regular Expression, RE) 是一种用来描述正则语言的更紧凑的表示方法

➤ 例:  $r = a(a/b)^*(\varepsilon \mid (./\_)(a/b)(a/b)^*)$

➤ 正则表达式可以由较小的正则表达式按照特定规则递归地构建。每个正则表达式  $r$  定义（表示）一个语言，记为  $L(r)$ 。这个语言也是根据  $r$  的子表达式所表示的语言递归定义的

# 正则表达式的定义

- $\varepsilon$  是一个  $RE$ ,  $L(\varepsilon) = \{\varepsilon\}$
- 如果  $a \in \Sigma$ , 则  $a$  是一个  $RE$ ,  $L(a) = \{a\}$
- 假设  $r$  和  $s$  都是  $RE$ , 表示的语言分别是  $L(r)$  和  $L(s)$ , 则
  - $r|s$  是一个  $RE$ ,  $L(r|s) = L(r) \cup L(s)$
  - $rs$  是一个  $RE$ ,  $L(rs) = L(r) L(s)$
  - $r^*$  是一个  $RE$ ,  $L(r^*) = (L(r))^*$  克林闭包
  - $(r)$  是一个  $RE$ ,  $L((r)) = L(r)$

运算的优先级:  $*$ 、连接、 $|$

## 例

➤ 令  $\Sigma = \{a, b\}$ , 则

$$\text{➤ } L(a|b) = L(a) \cup L(b) = \{a\} \cup \{b\} = \{a, b\}$$

$$\text{➤ } L((a|b)(a|b)) = L(a|b) L(a|b) = \{a, b\} \{a, b\} = \{aa, ab, ba, bb\}$$

$$\text{➤ } L(a^*) = (L(a))^* = \{a\}^* = \{\varepsilon, a, aa, aaa, \dots\} \text{ 任意长度的 } a$$

$$\text{➤ } L((a|b)^*) = (L(a|b))^* = \{a, b\}^* = \{\varepsilon, a, b, aa, ab, ba, bb, aaa, \dots\}$$

任意长度的ab串

$$\text{➤ } L(a|a^*b) = \{a, b, ab, aab, aaab, \dots\}$$

a 或 任意长度a, 结尾+b

## 例：C语言无符号整数的 $RE$

➤ 十进制整数的 $RE$

➤  $(1|...|9)(0|...|9)^*|0$

➤ 八进制整数的 $RE$

➤  $0(0|1|2|3|4|5|6|7)(0|1|2|3|4|5|6|7)^*$

➤ 十六进制整数的 $RE$

➤  $0x(0|1|...|9|a|...|f|A|...|F)(0|...|9|a|...|f|A|...|F)^*$

# 正则语言

➤ 可以用 $RE$ 定义的语言叫做

**正则语言**(*regular language*)或**正则集合**(*regular set*)

# RE的代数定律

定律	描述
$r \mid s = s \mid r$	$\mid$ 是可以交换的
$r \mid (s \mid t) = (r \mid s) \mid t$	$\mid$ 是可结合的
$r (s t) = (r s) t$	连接是可结合的
$r (s \mid t) = r s \mid r t ;$ $(s \mid t) r = s r \mid t r$	连接对 $\mid$ 是可分配的
$\varepsilon r = r \varepsilon = r$	$\varepsilon$ 是连接的单位元
$r^* = (r \mid \varepsilon)^*$	闭包中一定包含 $\varepsilon$
$r^{**} = r^*$	$*$ 具有幂等性

# 正则文法与正则表达式等价

- 对任何正则文法  $G$ ，存在定义同一语言的正则表达式  $r$
- 对任何正则表达式  $r$ ，存在生成同一语言的正则文法  $G$



## 3型文法 (Type-3 Grammar)

$$\alpha \rightarrow \beta$$

### ➤ 正则文法 (Regular Grammar, RG)

2种分类

➤ 右线性 (Right Linear) 文法:  $A \rightarrow wB$  或  $A \rightarrow w$

右侧最多1个非终结符, 且  
在同一侧

➤ 左线性 (Left Linear) 文法:  $A \rightarrow Bw$  或  $A \rightarrow w$

➤ 左线性文法和右线性文法都称为正则文法

例 (右线性文法)

字母打头的标识符

①  $S \rightarrow a \mid b \mid c \mid d$

②  $S \rightarrow aT \mid bT \mid cT \mid dT$

③  $T \rightarrow a \mid b \mid c \mid d \mid 0 \mid 1 \mid 2 \mid 3 \mid 4 \mid 5$

④  $T \rightarrow aT \mid bT \mid cT \mid dT \mid 0T \mid 1T \mid 2T \mid 3T \mid 4T \mid 5T$


文法  $G$  (上下文无关文法)

①  $S \rightarrow L \mid LT$

②  $T \rightarrow L \mid D \mid TL \mid TD$

③  $L \rightarrow a \mid b \mid c \mid d$

④  $D \rightarrow 0 \mid 1 \mid 2 \mid 3 \mid 4 \mid 5$



## 第三章 词法分析

# 正则表达式

哈尔滨工业大学 陈鄞





### 第三章 词法分析

# 正则定义

哈尔滨工业大学 陈鄞



# 正则定义 (Regular Definition)

➤ 正则定义是具有如下形式的定义序列：

$$d_1 \rightarrow r_1$$

$$d_2 \rightarrow r_2$$

...

$$d_n \rightarrow r_n$$

给一些RE命名，并在之后的RE中像使用字母表中的符号一样使用这些名字

其中：

➤ 每个 $d_i$ 都是一个新符号，它们都不在字母表 $\Sigma$ 中，而且各不相同

➤ 每个 $r_i$ 是字母表 $\Sigma \cup \{d_1, d_2, \dots, d_{i-1}\}$ 上的正则表达式

## 例1

➤ C语言中标识符的正则定义

➤  $digit \rightarrow 0|1|2|\dots|9$

➤  $letter\_ \rightarrow A|B|\dots|Z|a|b|\dots|z|_$

➤  $id \rightarrow letter\_ (letter\_ | digit)^*$

表示字母打头的字母数字串

## 例2

➤ (整型或浮点型) 无符号数的正则定义

➤  $digit \rightarrow 0|1|2|\dots|9$

➤  $digits \rightarrow digit\ digit^*$

➤  $optionalFraction \rightarrow .digits|\epsilon$  可选小数部分

➤  $optionalExponent \rightarrow ( E(+|-|\epsilon)digits )|\epsilon$  可选指数部分

➤  $number \rightarrow digits\ optionalFraction\ optionalExponent$

2

2.15

2.15E+3

2.15E-3

2.15E3

2E-3



## 第三章 词法分析

# 正则定义

哈尔滨工业大学 陈鄞







# 提纲

## 3.3 单词的识别

- 有穷自动机 (*Finite Automata*)
- 从正则表达式到自动机





# 第三章 词法分析

## 有穷自动机

哈尔滨工业大学 陈鄞



# 有穷自动机

- 有穷自动机 ( *Finite Automata*, *FA* ) 由两位神经物理学家 *McCulloch* 和 *Pitts* 于 1948 年首先提出，是对一类处理系统建立的数学模型
- 这类系统具有一系列离散的输入输出信息和有穷数目的内部状态（状态：概括了对过去输入信息处理的状况）
- 系统只需要根据当前所处的状态和当前面临的输入信息就可以决定系统的后继行为。每当系统处理了当前的输入后，系统的内部状态也将发生改变

# FA的典型例子

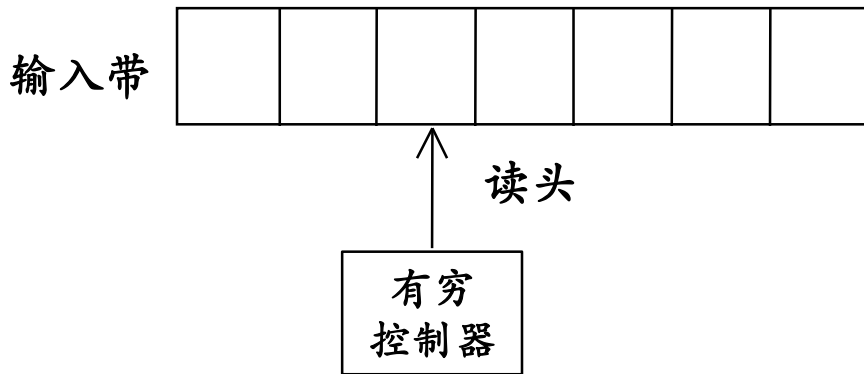
## ➤ 电梯控制装置

➤ 输入：顾客的乘梯需求（所要到达的层号）

➤ 状态：电梯所处的层数+运动方向

➤ 电梯控制装置并不需要记住先前全部的服务要求，只需要知道电梯当前所处的状态以及还没有满足的所有服务请求

# FA模型



- **输入带** (*input tape*): 用来存放输入符号串
- **读头** (*head*): 从左向右逐个读取输入符号, 不能修改 (只读)、不能往返移动
- **有穷控制器** (*finite control*): 具有有穷个状态数, 根据当前的状态和当前输入符号控制转入下一状态

# FA的表示

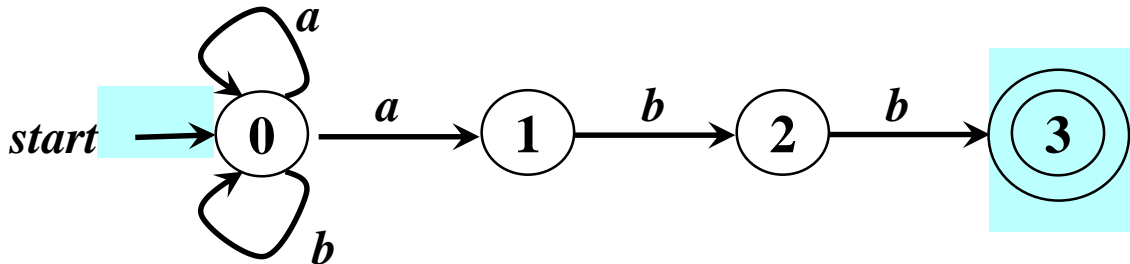
## ➤ 转换图 (Transition Graph)

### ➤ 结点：FA的状态

➤ 初始状态（开始状态）：只有一个，由start箭头指向

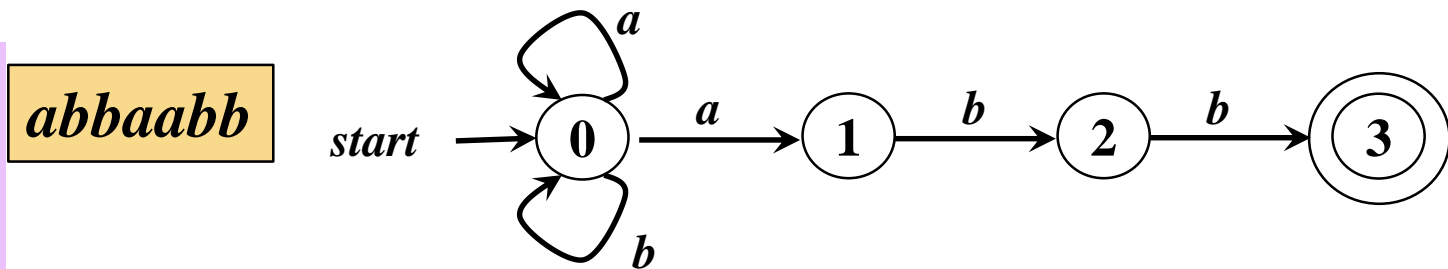
➤ 终止状态（接收状态）：可以有多个，用双圈表示

➤ 带标记的有向边：如果对于输入 $a$ ，存在一个从状态 $p$ 到状态 $q$ 的转换，就在 $p$ 、 $q$ 之间画一条有向边，并标记上 $a$



## FA定义(接收)的语言

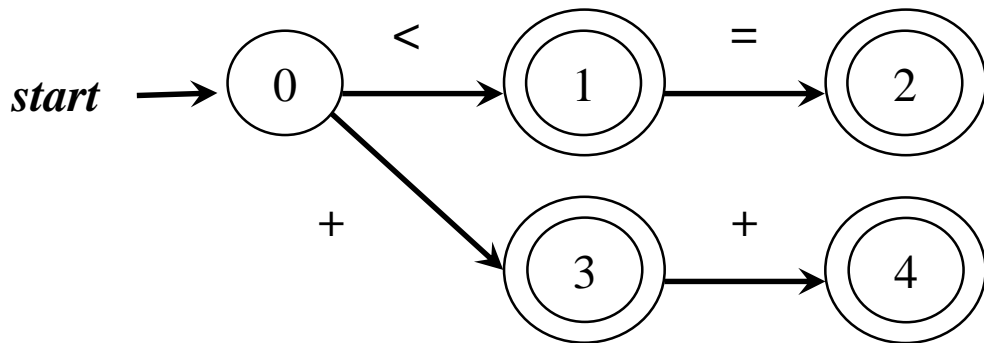
- 给定输入串 $x$ ，如果存在一个对应于串 $x$ 的从初始状态到某个终止状态的转换序列，则称串 $x$ 被该FA接收
- 由一个有穷自动机 $M$ 接收的所有串构成的集合称为是该FA定义(或接收)的语言，记为 $L(M)$



$L(M)$  = 所有以 $abb$ 结尾的字母表 $\{a, b\}$ 上的串的集合

## 最长子串匹配原则 (Longest String Matching Principle)

- 当输入串的多个前缀与一个或多个模式匹配时，总是选择最长的前缀进行匹配



- 在到达某个终态之后，只要输入带上还有符号，DFA就继续前进，以便寻找尽可能长的匹配



# 第三章 词法分析

## 有穷自动机

哈尔滨工业大学 陈鄞







## 第三章 词法分析

# 有穷自动机的分类

哈尔滨工业大学 陈鄞





## ***FA*的分类**

➤ 确定的*FA* (*Deterministic finite automata, DFA*)

➤ 非确定的*FA* (*Nondeterministic finite automata, NFA*)

# 确定的有穷自动机 ( $DFA$ )

$$M = ( S, \Sigma, \delta, s_0, F )$$

➤  $S$ : 有穷状态集

➤  $\Sigma$ : 输入字母表, 即输入符号集合。假设  $\varepsilon$  不是  $\Sigma$  中的元素

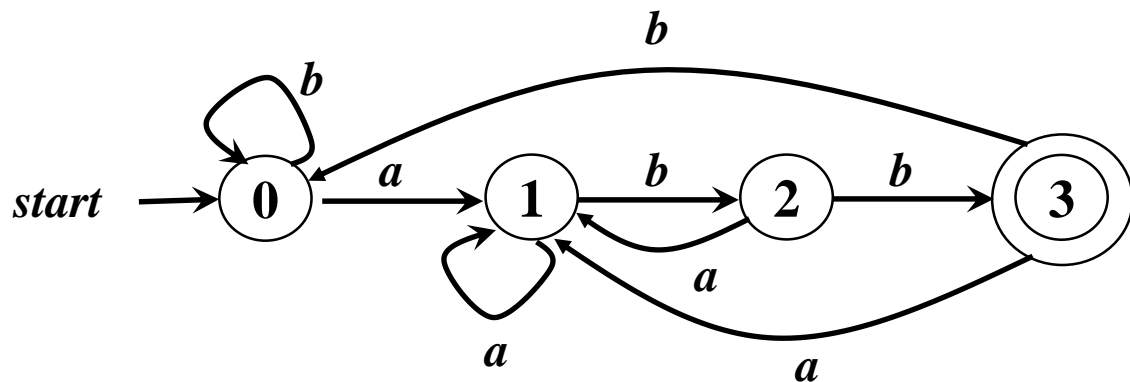
➤  $\delta$ : 将  $S \times \Sigma$  映射到  $S$  的转换函数。  $\forall s \in S, a \in \Sigma, \delta(s, a)$  表示从状态  $s$  出发, 沿着标记为  $a$  的边所能到达的状态。

➤  $s_0$ : 开始状态 (或初始状态),  $s_0 \in S$

➤  $F$ : 接收状态 (或终止状态) 集合,  $F \subseteq S$

## 例：一个DFA

$$M = (S, \Sigma, \delta, s_0, F)$$



转换表

状态 \ 输入	a	b
0	1	0
1	1	2
2	1	3
3	1	0

DFA可以用转换图或者转换表来表示，二者是等价的

可以用转换表表示DFA

表示终止状态

# 非确定的有穷自动机(NFA)

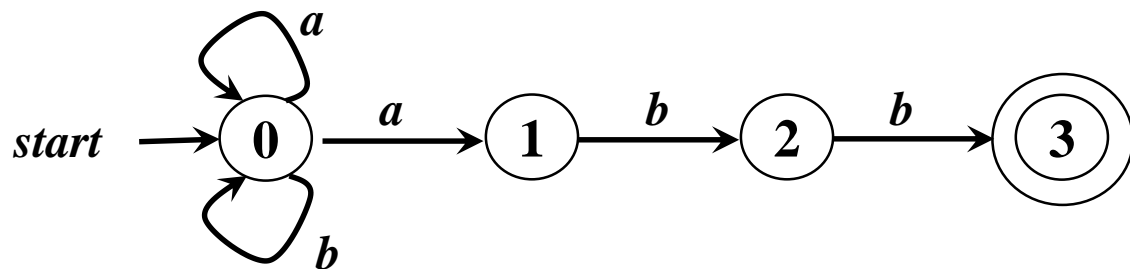
$$M = ( S, \Sigma, \delta, s_0, F )$$

- $S$ : 有穷状态集
- $\Sigma$ : 输入符号集合, 即输入字母表。假设  $\varepsilon$  不是  $\Sigma$  中的元素
- $\delta$ : 将  $S \times \Sigma$  映射到  $2^S$  的转换函数。  $\forall s \in S, a \in \Sigma, \delta(s, a)$  表示从状态  $s$  出发, 沿着标记为  $a$  的边所能到达的状态集合
- $s_0$ : 开始状态 (或初始状态),  $s_0 \in S$
- $F$ : 接收状态 (或终止状态) 集合,  $F \subseteq S$

状态不是唯一确定的

## 例：一个NFA

$$M = (S, \Sigma, \delta, s_0, F)$$



转换表

状态 \ 输入	a	b
0	{0,1}	{0}
1	$\emptyset$	{2}
2	$\emptyset$	{3}
3 •	$\emptyset$	$\emptyset$

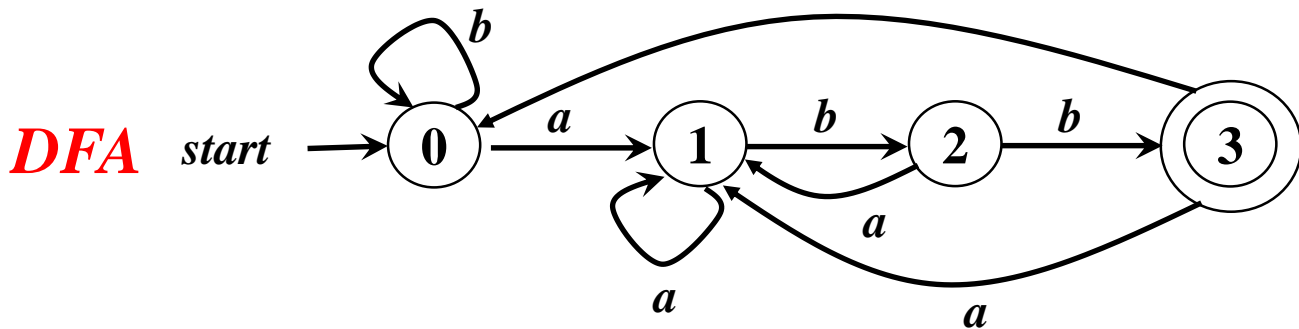
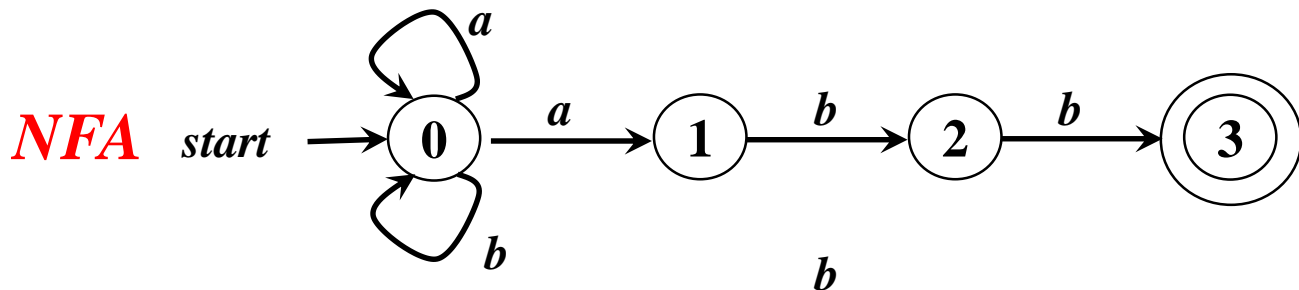
如果转换函数没有给出对应于某个状态-输入对的信息，就把 $\emptyset$ 放入相应的表项中

# DFA和NFA的等价性

- 对任何非确定的有穷自动机 $N$ ，存在定义同一语言的确定的有穷自动机 $D$
- 对任何确定的有穷自动机 $D$ ，存在定义同一语言的非确定的有穷自动机 $N$

# DFA和NFA的等价性

➤ DFA和NFA可以识别相同的语言



状态1: 串以 $a$ 结尾

状态2: 串以 $ab$ 结尾

状态3: 串以 $abb$ 结尾

$$r = (a|b)^*abb$$

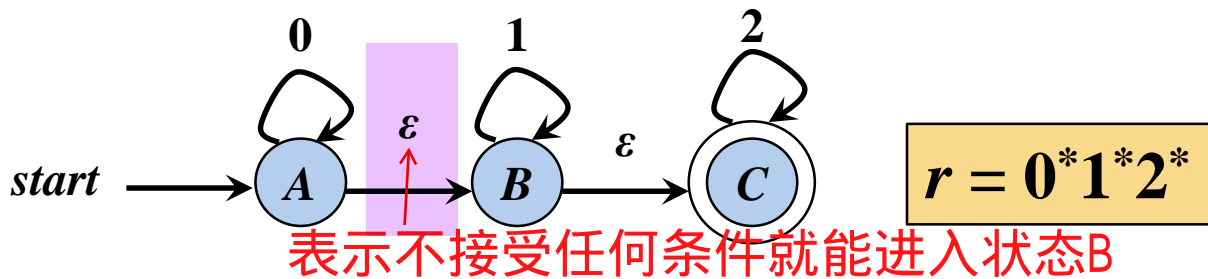
正则文法  $\Leftrightarrow$  正则表达式  $\Leftrightarrow$  FA



# 带有“ $\varepsilon$ -边”的NFA

$$M = (S, \Sigma, \delta, s_0, F)$$

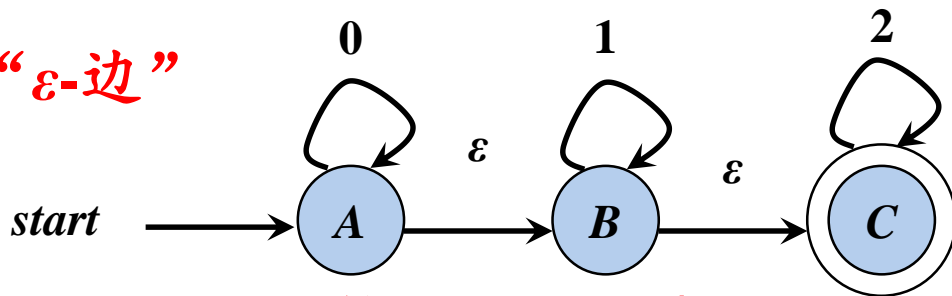
- $S$ : 有穷状态集
- $\Sigma$ : 输入符号集合, 即输入字母表。假设  $\varepsilon$  不是  $\Sigma$  中的元素
- $\delta$ : 将  $S \times (\Sigma \cup \{\varepsilon\})$  映射到  $2^S$  的转换函数。 $\forall s \in S, a \in \Sigma \cup \{\varepsilon\}, \delta(s, a)$  表示从状态  $s$  出发, 沿着标记为  $a$  的边所能到达的状态集合
- $s_0$ : 开始状态 (或初始状态),  $s_0 \in S$
- $F$ : 接收状态 (或终止状态) 集合,  $F \subseteq S$



# 带有和不带有“ $\epsilon$ -边”的NFA的等价性

## ➤ 例

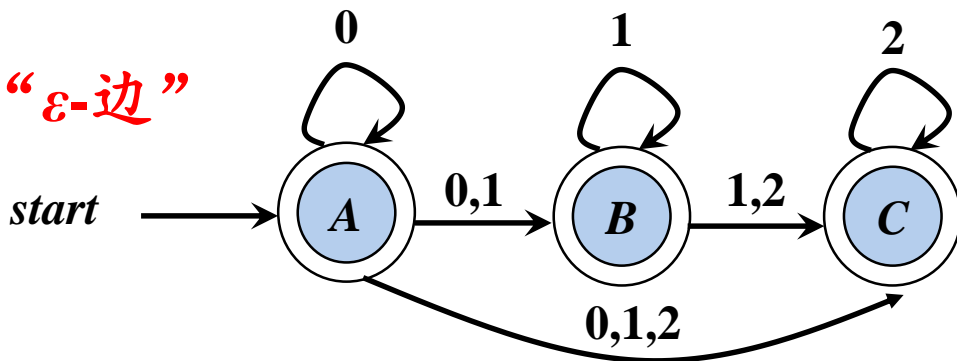
带“ $\epsilon$ -边”



$$r = 0^*1^*2^*$$

上下的AB不是一回事

不带“ $\epsilon$ -边”



状态A:  $0^*$

状态B:  $0^*1^*$

状态C:  $0^*1^*2^*$

# DFA的算法实现

- 输入：以文件结束符eof结尾的字符串 $x$ 。DFA  $D$  的开始状态 $s_0$ ，接收状态集  $F$ ，转换函数 $move$ 。
- 输出：如果  $D$ 接收  $x$ ，则回答 “yes”，否则回答 “no”。
- 方法：将下述算法应用于输入串  $x$ 。

```
 $s = s_0;$   
 $c = nextChar ();$   
while ( $c \neq eof$ ) {  
     $s = move(s, c);$   
     $c = nextChar ();$   
}  
if ( $s$ 在 $F$ 中) return “yes”;  
else return “no”;
```

NFA更直观，DFA更易于实现

- 函数 $nextChar()$ 返回输入串 $x$ 的下一个符号
- 函数 $move(s, c)$ 表示从状态 $s$ 出发，沿着标记为 $c$ 的边所能到达的状态




## 第三章 词法分析

# 有穷自动机的分类

哈尔滨工业大学 陈鄞



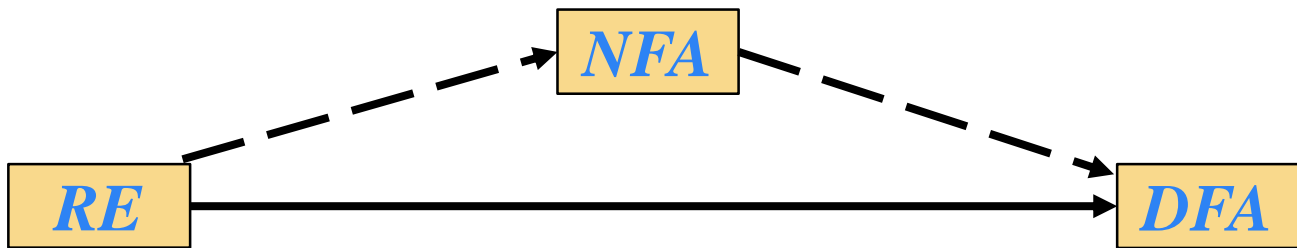
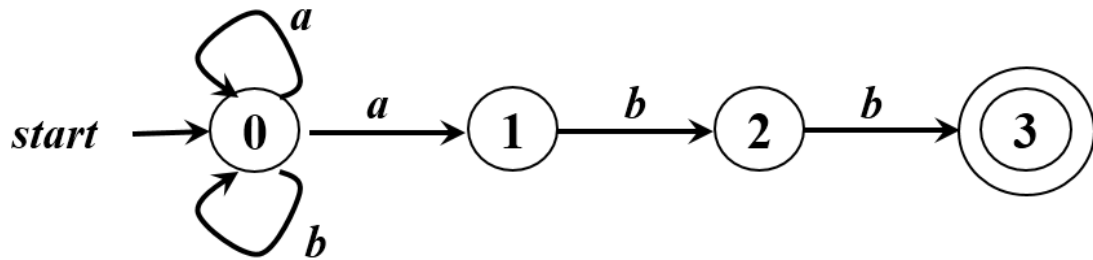


# 第三章 词法分析 从正则表达式 到有穷自动机

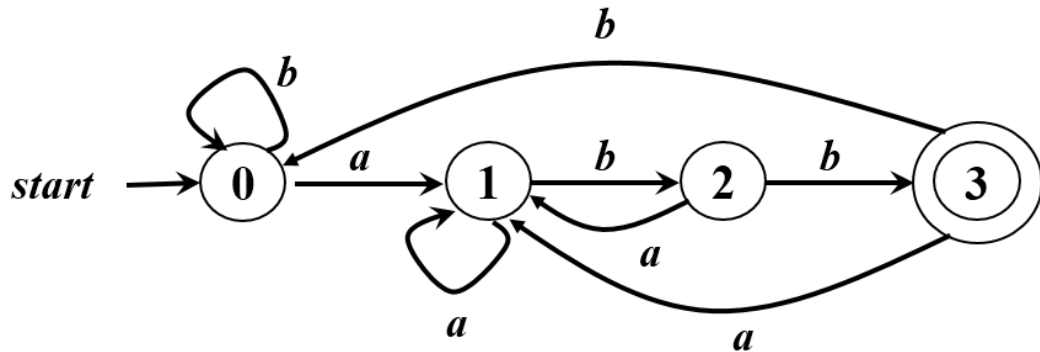
哈尔滨工业大学 陈勤



# 从正则表达式到无穷自动机

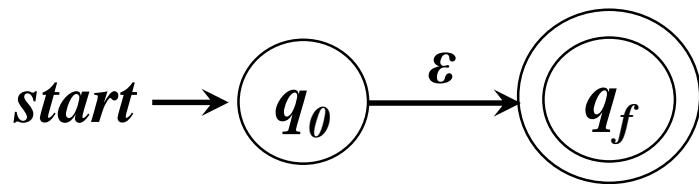


$r = (a/b)^*abb$

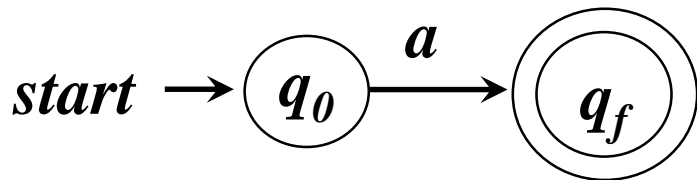


## 根据 $RE$ 构造 $NFA$

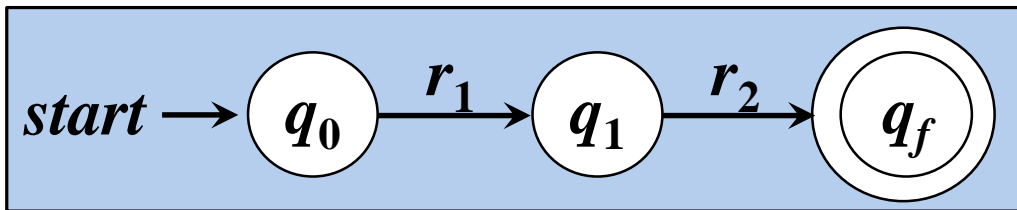
➤  $\epsilon$  对应的  $NFA$



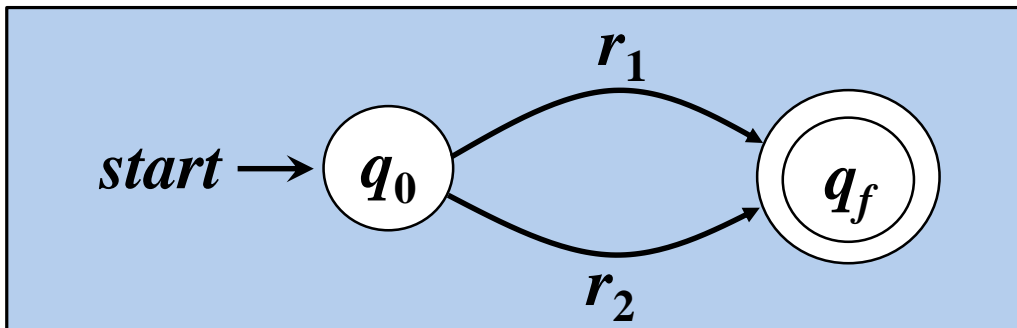
➤ 字母表  $\Sigma$  中符号  $a$  对应的  $NFA$



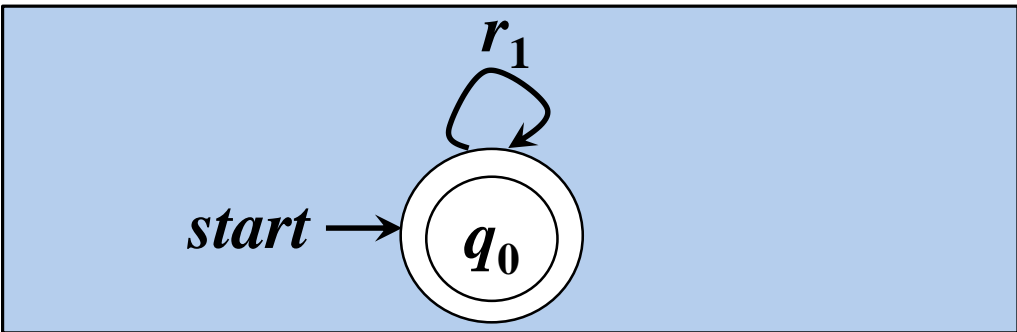
➤  $r = r_1 r_2$  对应的 *NFA*



➤  $r = r_1 / r_2$  对应的 *NFA*



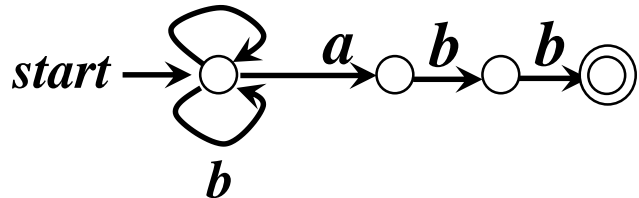
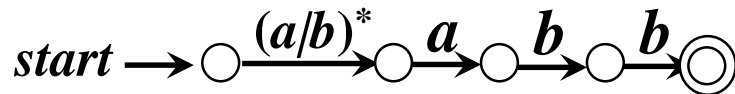
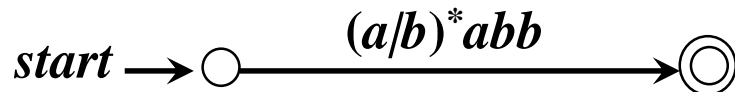
➤  $r = (r_1)^*$  对应的 *NFA*






# 例： $r = (a|b)^*abb$ 对应的 NFA

不带  $\epsilon$ -边的 NFA



不断分解各个部分



# 第三章 词法分析 从正则表达式 到有穷自动机

哈尔滨工业大学 陈勤





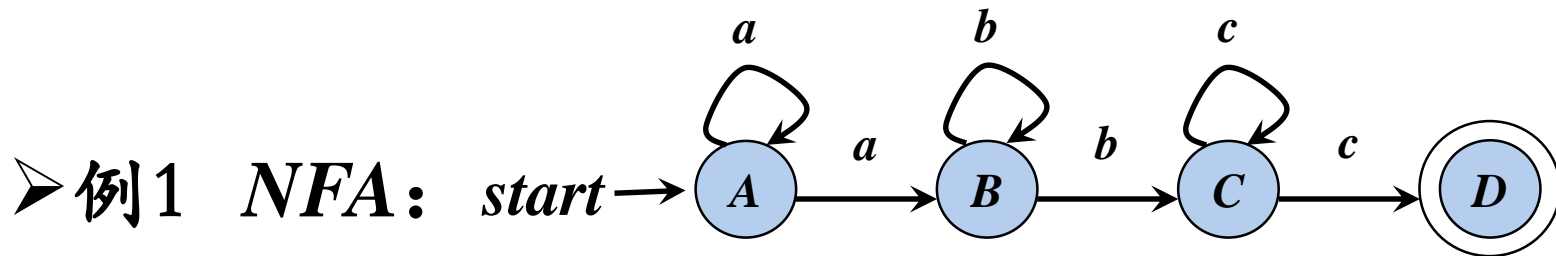
## 第三章 词法分析

# 从 $NFA$ 到 $DFA$ 的转换

哈尔滨工业大学 陈鄞



# 从NFA到DFA的转换

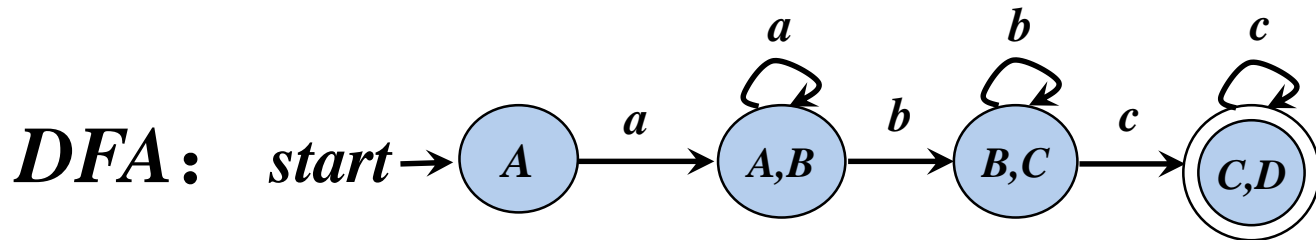


转换表

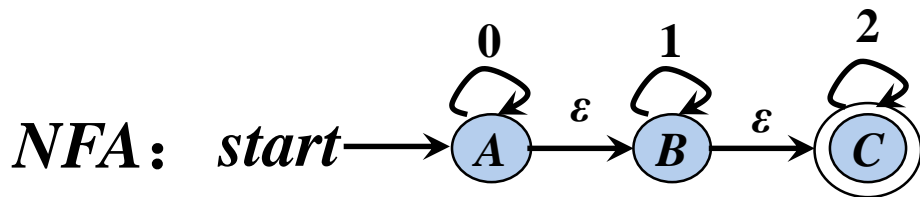
状态 \ 输入	a	b	c
A	{A,B}	$\emptyset$	$\emptyset$
B	$\emptyset$	{B,C}	$\emptyset$
C	$\emptyset$	$\emptyset$	{C,D}
D ●	$\emptyset$	$\emptyset$	$\emptyset$

DFA的每个状态都是一个由NFA中的状态构成的集合,即NFA状态集合的一个子集

$r = aa^*bb^*cc^*$



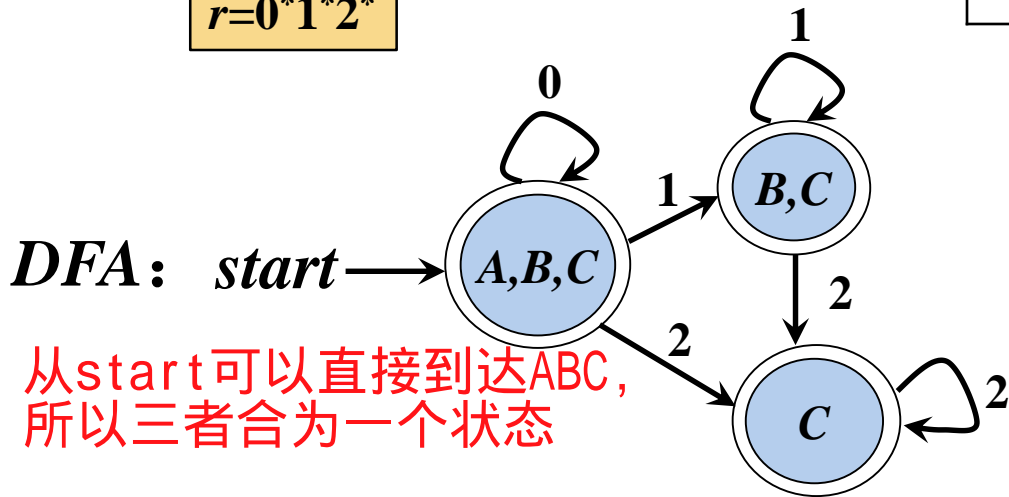
## 例2：从带有 $\varepsilon$ -边的NFA到DFA的转换



$r=0^*1^*2^*$

转换表

状态 \ 输入	0	1	2
A	{A,B,C}	{B,C}	{C}
B	$\emptyset$	{B,C}	{C}
C •	$\emptyset$	$\emptyset$	{C}



# 子集构造法 (subset construction)

- 输入:  $NFA\ N$  都是NFA的一个子集，所以叫子集构造法
- 输出: 接收同样语言的 $DFA\ D$
- 方法: 一开始,  $\varepsilon\text{-closure}(s_0)$  是  $Dstates$  中的唯一状态, 且它未加标记;  
while (在  $Dstates$  中有一个未标记状态  $T$ ) {  
    给  $T$  加上标记;  
    for (每个输入符号  $a$ ) {  
         $U = \varepsilon\text{-closure}(\text{move}(T, a))$ ;  
        if (  $U$  不在  $Dstates$  中 )  
            将  $U$  加入到  $Dstates$  中, 且不加标记;  
         $Dtran[T, a] = U$ ;  
    }  
}

操作	描述
$\varepsilon\text{-closure}(s)$	能够从 $NFA$ 的状态 $s$ 开始只通过 $\varepsilon$ 转换到达的 $NFA$ 状态集合
$\varepsilon\text{-closure}(T)$	能够从 $T$ 中的某个 $NFA$ 状态 $s$ 开始只通过 $\varepsilon$ 转换到达的 $NFA$ 状态集合, 即 $U_{s \in T} \varepsilon\text{-closure}(s)$
$\text{move}(T, a)$	能够从 $T$ 中的某个状态 $s$ 出发通过标号为 $a$ 的转换到达的 $NFA$ 状态的集合

## 计算 $\varepsilon$ -closure ( $T$ )

将 $T$ 的所有状态压入 $stack$ 中；

将 $\varepsilon$ -closure ( $T$ )初始化为  $T$  ；

while ( $stack$ 非空) {

    将栈顶元素  $t$  给弹出栈中；

    for (每个满足如下条件的 $u$ ：从 $t$ 出发有一个标号为 $\varepsilon$ 的转换到达状态 $u$ )

        if ( $u$ 不在 $\varepsilon$ -closure ( $T$ )中) {

            将 $u$ 加入到 $\varepsilon$ -closure ( $T$ )中；

            将 $u$ 压入栈中；

        }

}





## 第三章 词法分析

# 从 $NFA$ 到 $DFA$ 的转换

哈尔滨工业大学 陈鄞







# 第三章 词法分析

## 识别单词的 $DFA$

哈尔滨工业大学 陈鄞



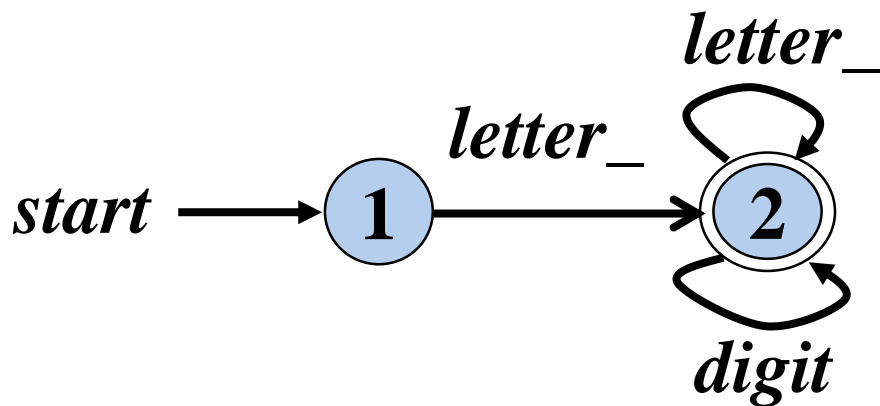
# 识别标识符的DFA

➤ 标识符的正则定义

➤  $digit \rightarrow 0|1|2|\dots|9$

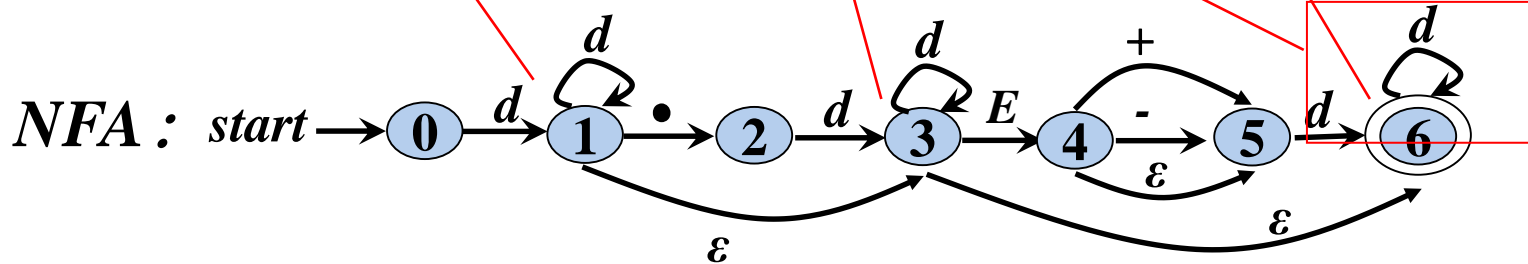
➤  $letter\_ \rightarrow A|B|\dots|Z|a|b|\dots|z|_$

➤  $id \rightarrow letter\_ (letter\_|digit)^*$



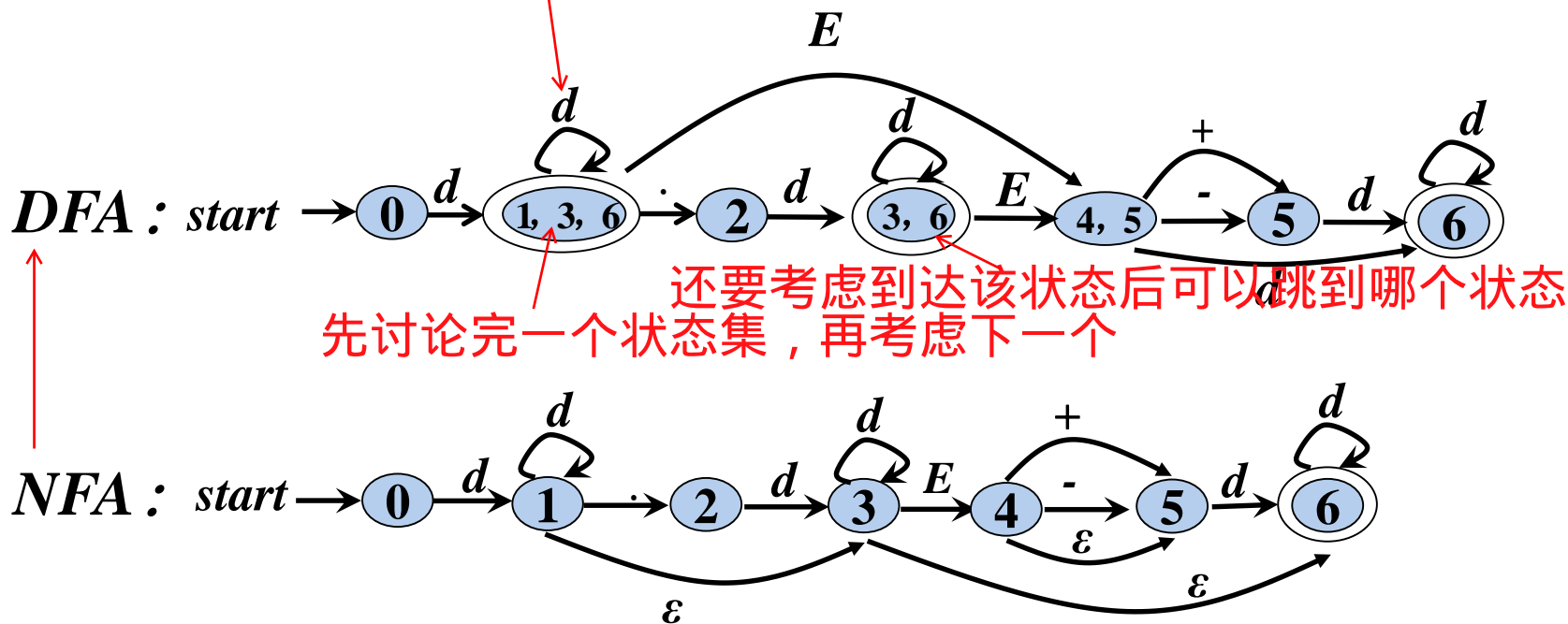
# 识别无符号数的 $DFA$

- $digit \rightarrow 0|1|2|\dots|9$
- $digits \rightarrow digit\ digit^*$
- $optionalFraction \rightarrow .digits|\epsilon$
- $optionalExponent \rightarrow (E(+|-|\epsilon)digits)|\epsilon$
- $number \rightarrow digits\ optionalFraction\ optionalExponent$



# 识别无符号数的DFA

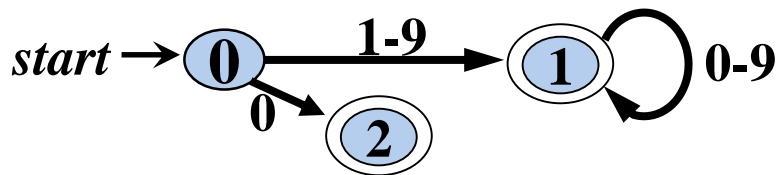
记得分别讨论状态1、3、6在遇到d时的变化，再合集



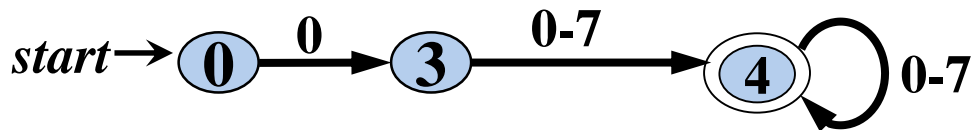
最后可以再验证一下

# 识别各进制无符号整数的DFA

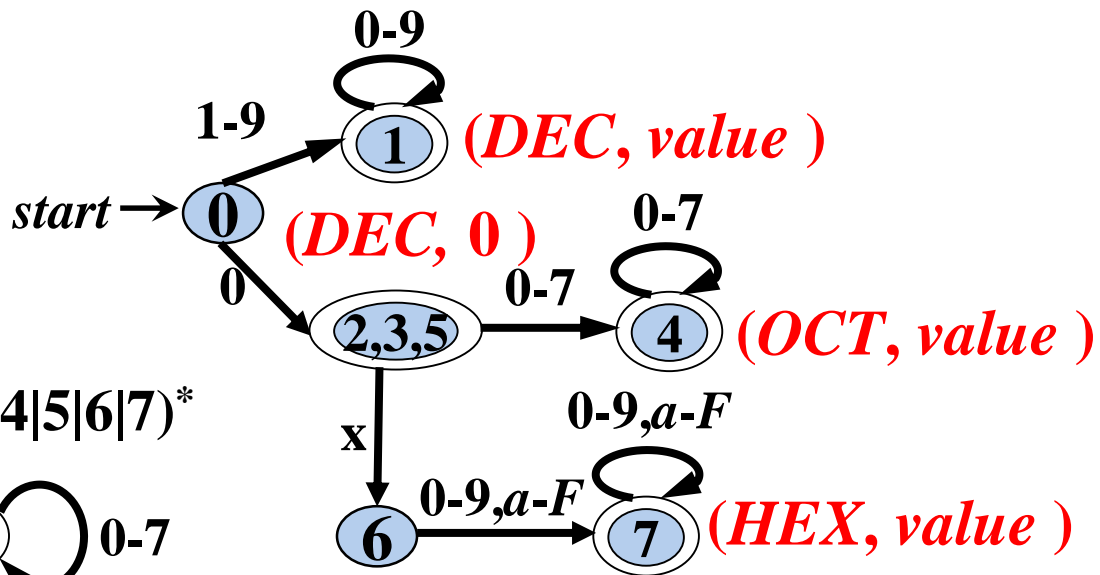
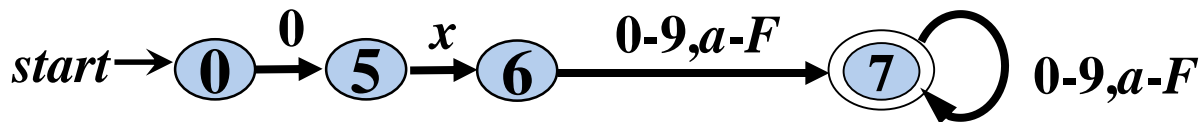
$DEC \rightarrow (1|...|9)(0|...|9)^* | 0$



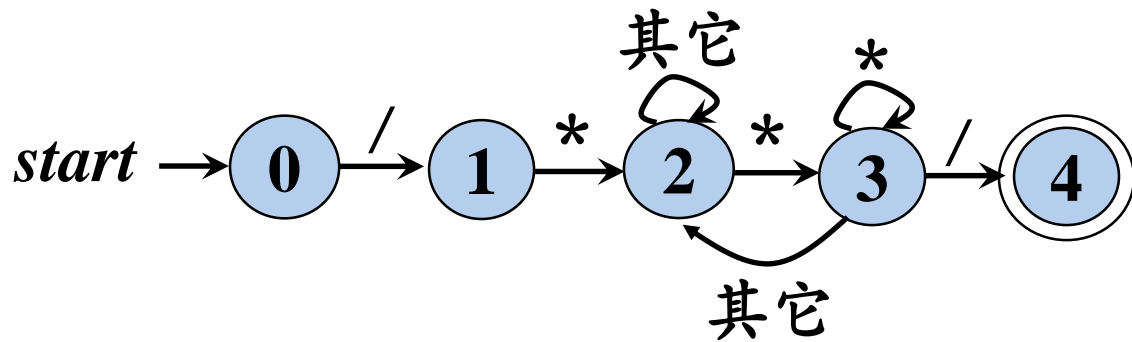
$OCT \rightarrow 0(0|1|2|3|4|5|6|7)(0|1|2|3|4|5|6|7)^*$



$HEX \rightarrow 0x(0|1|...|9|a|...|f|A|...|F)(0|...|9|a|...|f|A|...|F)^*$

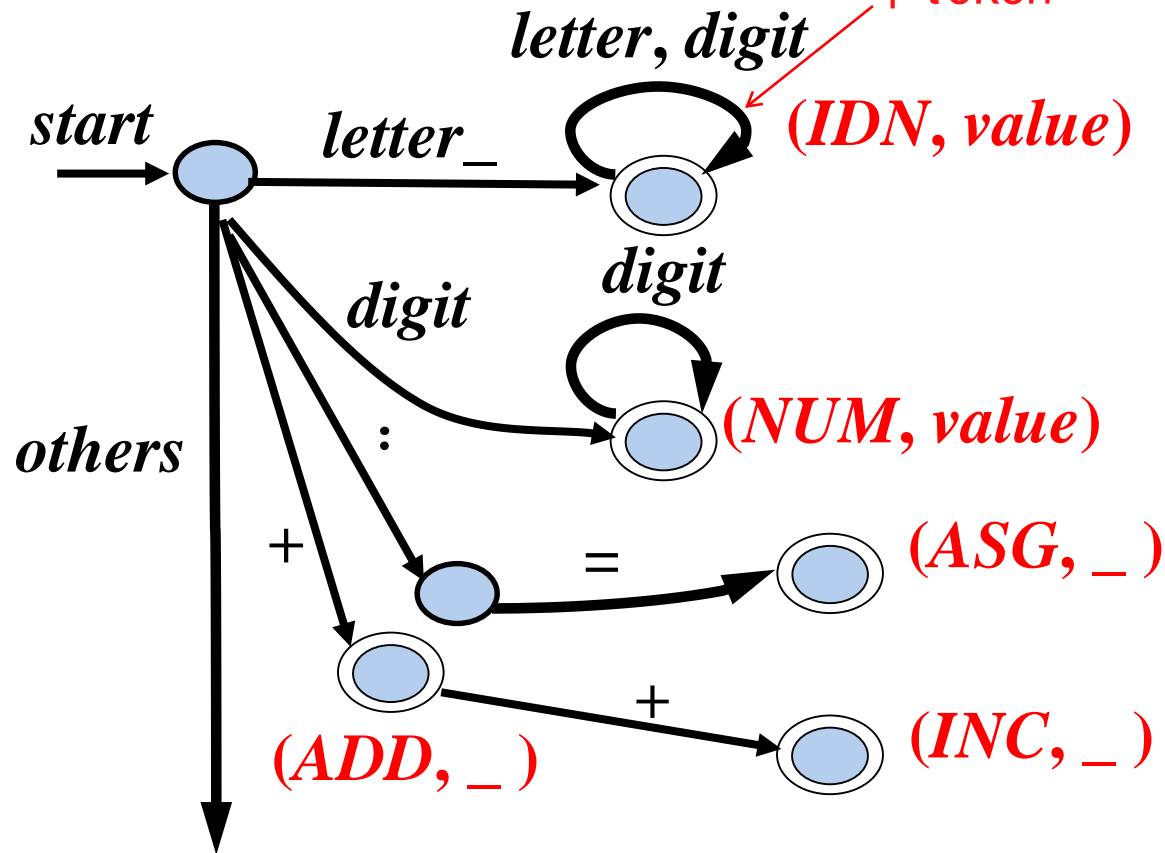


# 识别注释的DFA



# 识别 *Token* 的 *DFA*

形成一个token后对照关键字表，若  
比对成功则作为关键字；否则作为一个  
token



形成token后就返回start继续识别下一个token

# 词法分析阶段的错误处理

## ➤ 词法分析阶段可检测错误的类型

### ➤ 单词拼写错误

➤ 例： `int i = 0x3G; float j = 1.05e;`

### ➤ 非法字符

➤ 例： ~ @

## ➤ 词法错误检测

➤ 如果当前状态与当前输入符号在转换表对应项中的信息为空，而当前状态又不是终止状态，则调用错误处理程序



## 错误处理

- 查找已扫描字符串中最后一个对应于某终态的字符
  - 如果找到了，将该字符与其前面的字符识别成一个单词。  
然后将输入指针退回到该字符，扫描器重新回到初始状态，继续识别下一个单词
  - 如果没找到，则确定出错，采用错误恢复策略

# 错误恢复策略

- 最简单的错误恢复策略：“**恐慌模式** (*panic mode*)” 恢复
  - 从剩余的输入中不断删除字符，直到词法分析器能够在剩余输入的开头发现一个正确的字符为止



# 第三章 词法分析

## 识别单词的 $DFA$

哈尔滨工业大学 陈鄞

