编译原理

第三章 词法分析

第三章 词法分析

- 对于词法分析器的要求
- 词法分析器的设计
- 正规表达式与有限自动机
- 词法分析器的自动产生 --LEX

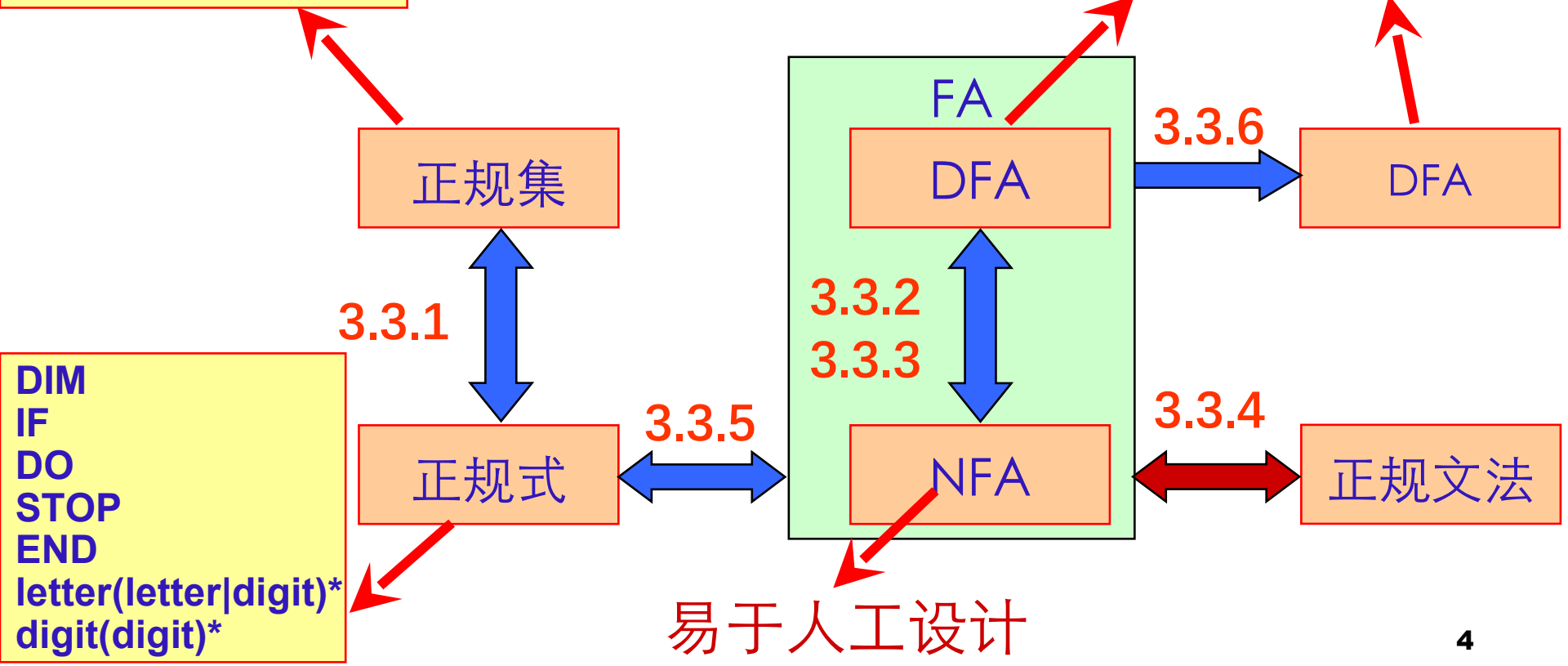
第三章 词法分析

- 对于词法分析器的要求
- 词法分析器的设计
- 正规表达式与有限自动机
- 词法分析器的自动产生 --LEX

关系图

```
curState = 初态
GetChar();
while( stateTrans[curState][ch] 有定义 ){
    // 存在后继状态，读入、拼接
    Concat();
    // 转换入下一状态，读入下一字符
   curState= stateTrans[curState][ch];
    if cur_state 是终态 then 返回 strToken 中的单
    GetChar( );
}
```

DIM,IF, DO,STOP,END
number, name, age
125, 2169
…



形式语言鸟瞰

- 2 型 (上下文无关文法，非确定下推自动机)
  - 产生式形如： $A \to \beta$
  - 其中： $A \in V_N$ ； $\beta \in (V_T \cup V_N)^*$
- 3 型 (正规文法，有限自动机)
  - 产生式形如： $A \to \alpha B$ 或 $A \to \alpha$ **右线性文法**
  - 其中： $\alpha \in V_T^*$ ； A ， $B \in V_N$
  - 产生式形如： $A \to B\alpha$ 或 $A \to \alpha$ **左线性文法**
  - 其中： $\alpha \in V_T^*$ ； A ， $B \in V_N$

3.3.4 正规文法与有限自动机的等价性

- 对于正规文法 G 和有限自动机 M，如果 L(G) = L(M)，则称 G 和 M 是等价的
- 关于正规文法和有限自动机的等价性，有以下结论：

1. 对每一个右线性正规文法 G 或左线性正规文法 G，都存在一个有限自动机 (FA) M，使得 L(M) = L(G)。

2. 对每一个 FA M，都存在一个右线性正规文法 G_R 和左线性正规文法 G_L，使得 $L(M) = L(G_R) = L(G_L)$。

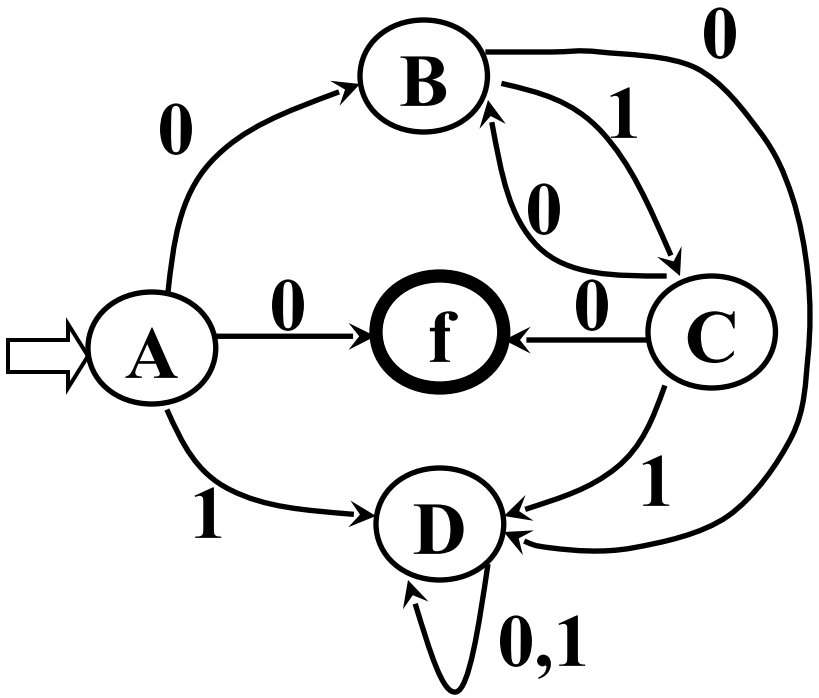
A ⇒ 0B
⇒ 01C
⇒ 010

例：

- $G_R(A)$ ：

  A→0 | 0B | 1D
  B→0D | 1C
  C→0 | 0B | 1D
  D→0D | 1D

- 从 G_R 出发构造 NFA M = <{A, B, C, D, f}, {0, 1}, δ′, A, {f}> ， M 的状态转换图如右图所示。

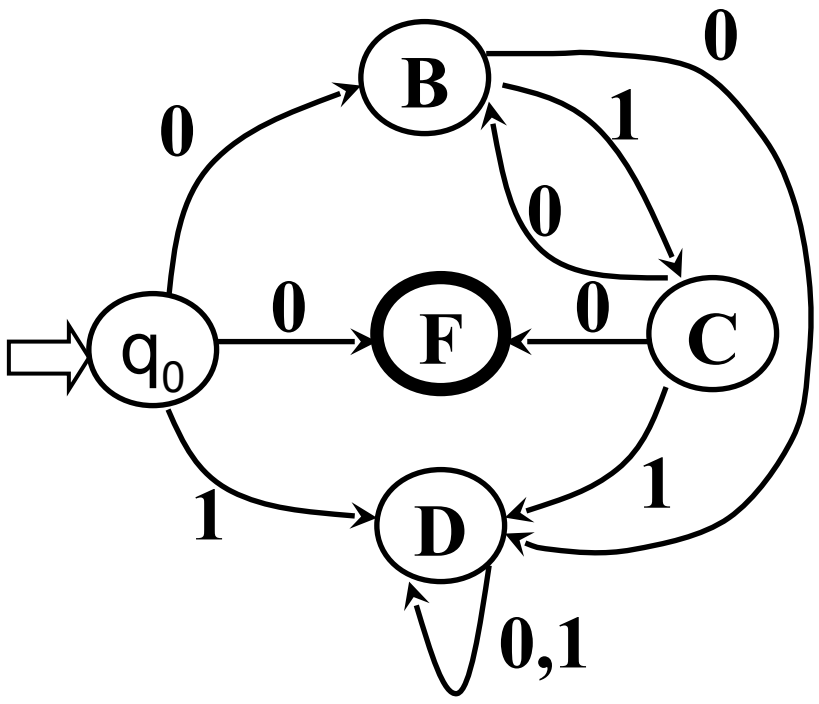


- 显然 L(M) = $L(G_R)$ 。

例：

$F \Rightarrow C0$
$\Rightarrow B10$
$\Rightarrow 010$

- 左线性正规文法 $G_L = <\{0, 1\}, \{B, C, D, F\}, F, P'>$ ， 其中 P' 由下列产生式组成：
  $F \rightarrow 0 \mid C0$
  $C \rightarrow B1$
  $B \rightarrow 0 \mid C0$
  $D \rightarrow 1 \mid C1 \mid D0 \mid D1 \mid B0$
- 从 G_L 出发构造 NFA $M = <\{q_0, B, C, D, F\}, \{0, 1\}, \delta, A, \{F\}>$ ， M 的状态转换图如右图所示。
- 显然 $L(M) = L(G_L)$ 。



- 证明：

1. 对每一个右线性正规文法 G 或左线性正规文法 G，都构造一个有限自动机 (FA) M，使得 L(M) = L(G)。

(1) 设右线性正规文法 $G=<V_T, V_N, S, P>$。将 V_N 中的每一非终结符号视为状态符号，并增加一个新的终结状态符号 f，$f \notin V_N$。

令 $M=<V_N \cup \{f\}, V_T, \delta, S, \{f\}>$，其中状态转换函数δ由以下规则定义：

(a) 若对某个 $A \in V_N$ 及 $a \in V_T \cup \{\varepsilon\}$，P 中有产生式 $A \to a$，则令 $\delta(A,a)=f$

(b) 对任意的 $A \in V_N$ 及 $a \in V_T \cup \{\varepsilon\}$，设 P 中左端为 A，右端第一符号为 a 的所有产生式为：

$A \to aA_1 | \cdots | aA_k$ （不包括 $A \to a$），

则令 $\delta(A,a)=\{A_1, \cdots, A_k\}$。

显然，上述 M 是一个 NFA。

对于右线性正规文法 G，在 $S \overset{+}{\Rightarrow} w$ 的最左推导过程中：

- 利用 $A \to aB$ 一次就相当于在 M 中从状态 A 经过标记为 a 的箭弧到达状态 B （包括 $a=\varepsilon$ 的情形）；
- 在推导的最后，利用 $A \to a$ 一次则相当于在 M 中从状态 A 经过标记为 a 的箭弧到达终结状态 f （包括 $a=\varepsilon$ 的情形）。

综上，在正规文法 G 中，$S \overset{+}{\Rightarrow} w$ 的充要条件是：在 M 中，从状态 S 到状态 f 有一条通路，其上所有箭弧的标记符号依次连接起来恰好等于 w，这就是说，$w \in L(G)$ 当且仅当 $w \in L(M)$，故 $L(G) = L(M)$。

例：

- $G_R(A)$ ：

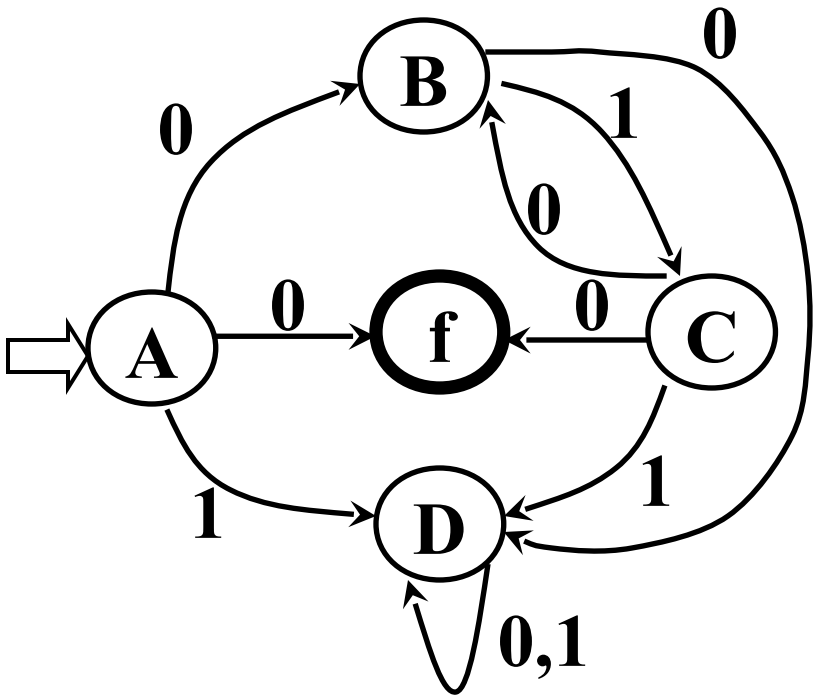
  A→0 | 0B | 1D

  B→0D | 1C

  C→0 | 0B | 1D

  D→0D | 1D

- 从 G_R 出发构造 NFA M = <{A, B, C, D, f}, {0, 1}, δ', A, {f}> ， M 的状态转换图如右图所示。



- 显然 L(M) = $L(G_R)$ 。

3.3.4 正规文法与有限自动机的等价性

- 定理:

1. 对每一个右线性正规文法 G 或左线性正规文法 G，都存在一个有限自动机 (FA) M，使得 L(M) = L(G)。

2. 对每一个 FA M，都存在一个右线性正规文法 G_R 和左线性正规文法 G_L，使得 $L(M) = L(G_R) = L(G_L)$。

(2) 设左线性正规文法 $G=<V_T, V_N, S, P>$。将 V_N 中的每一非终结符号视为状态符号，并增加一个初始状态符号 q_0， $q_0 \notin V_N$。

令 $M=<V_N \cup \{q_0\}, V_T, \delta, q_0, \{S\}>$，其中状态转换函数$\delta$由以下规则定义：

(a) 若对某个 $A \in V_N$ 及 $a \in V_T \cup \{\varepsilon\}$，若 P 中有产生式 $A \to a$，则令$\delta(q_0,a)=A$

(b) 对任意的 $A \in V_N$ 及 $a \in V_T \cup \{\varepsilon\}$，若 P 中所有右端第一符号为 A，第二个符号为 a 的产生式为：

$$A_1 \to Aa, \cdots, A_k \to Aa,$$

则令$\delta(A,a)=\{A_1,\cdots,A_k\}$。

与 (1) 类似，可以证明 $L(G) = L(M)$

例：

- 左线性正规文法 G_L = <{0, 1}, {B, C, D, F}, F, P′> ， 其中 P′ 由下列产生式组成：

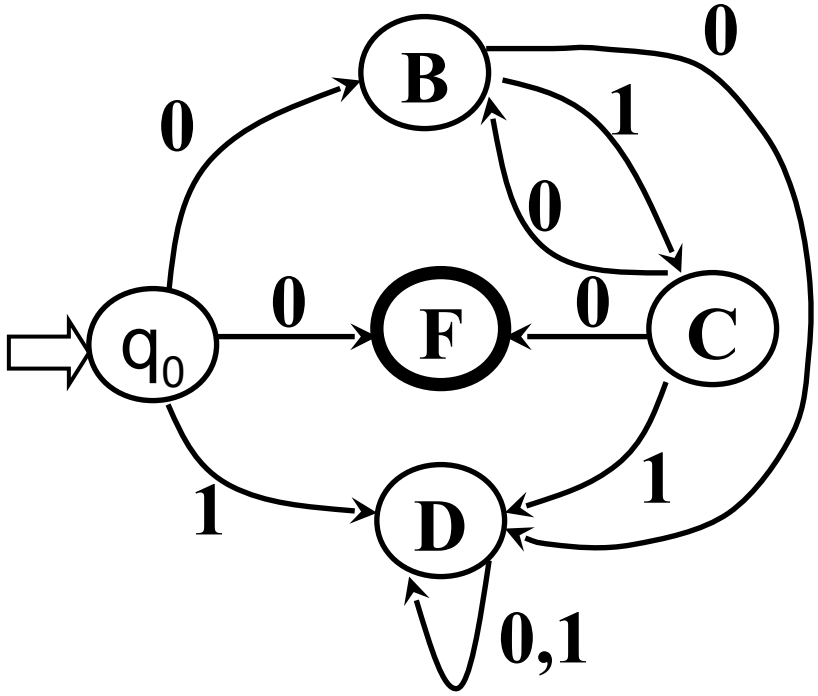
  F→0 | C0

  C→B1

  B→0 | C0

  D→1 | C1 | D0 | D1 | B0

- 从 G_L 出发构造 NFA M = <{q_0, B, C, D, F }, {0, 1}, δ, A, {F}> ， M 的状态转换图如右图所示。



- 显然 L(M) = L(G_L) 。

3.3.4 正规文法与有限自动机的等价性

- 定理:

1. 对每一个右线性正规文法 G 或左线性正规文法 G，都存在一个有限自动机 (FA) M，使得 L(M) = L(G)。

2. 对每一个 FA M，都存在一个右线性正规文法 G_R 和左线性正规文法 G_L，使得 $L(M) = L(G_R) = L(G_L)$。

证明 2：对每一个 DFA M，都存在一个右线性正规文法 G_R 和左线性正规文法 G_L，使得 $L(M) = L(G_R) = L(G_L)$。

设 DFA $M=<S, \Sigma, \delta, s_0, F>$

(1) 若 $s_0 \notin F$，我们令 $G_R=<\Sigma, S, s_0, P>$，其中 P 是由以下规则定义的产生式集合：

对任何 $a \in \Sigma$ 及 $A,B \in S$，若有 $\delta(A,a)=B$，则：

(a) 当 $B \notin F$ 时，令 $A \rightarrow aB$，

(b) 当 $B \in F$ 时，令 $A \rightarrow a|aB$。

对任何 $w\in\Sigma^*$，不妨设 $w=a_1\cdots a_k$，其中 $a_i\in\Sigma$ $(i=1,\cdots k)$。若 $s_0\overset{+}{\Rightarrow} w$，则存在一个最左推导：

$$s_0\Rightarrow a_1A_1\Rightarrow a_1a_2A_2\Rightarrow\cdots\Rightarrow a_1\cdots a_iA_i$$

$$\Rightarrow a_1\cdots a_{i+1}A_{i+1}\Rightarrow\cdots\Rightarrow a_1\cdots a_k$$

因而，在 M 中有一条从 s_0 出发依次经过 A_1，…，A_{k-1} 到达终态的通路，该通路上所有箭弧的标记依次为 $a_1,\cdots,a_k$。反之亦然。所以，$w\in L(G_R)$ 当且仅当 $w\in L(M)$。

- 现在考虑 $s_0 \in F$ 的情形：

因为 $\delta(s_0, \varepsilon)=s_0$，所以 $\varepsilon \in L(M)$。但 ε 不属于上面构造的 G_R 所产生的语言 $L(G_R)$。不难发现，

$L(G_R)=L(M)-\{\varepsilon\}$。

所以，我们在上述 G_R 中添加新的非终结符号 s_0'，$(s_0' \notin S)$ 和产生式 $s_0' \to s_0 | \varepsilon$，并用 s_0' 代替 s_0 作

2. 对每一个 FA M，都存在一个右线性正规文法 G_R 和左线性正规文法 G_L，使得 $L(M) = L(G_R) = L(G_L)$。

规文法

(a) 当 $A= q_0$ 时，令 $B \to a$，

最后，

结论 2 得

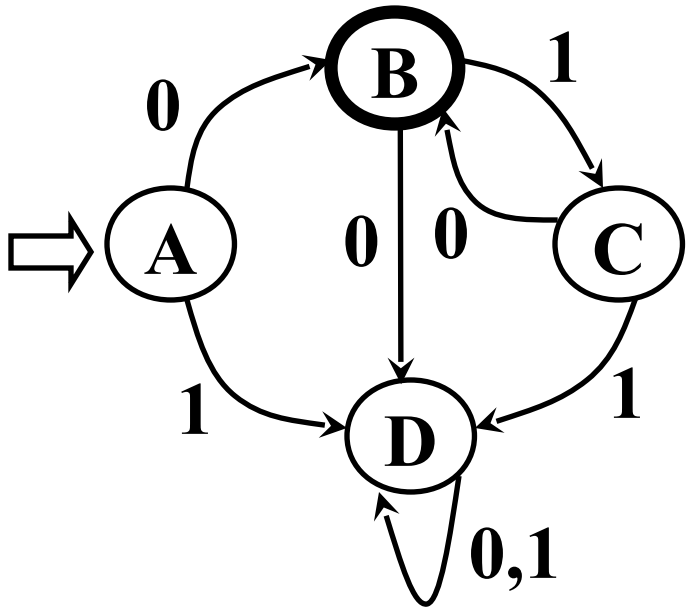
3.3.4 正规文法与有限自动机的等价性

- 定理:

1. 对每一个右线性正规文法 G 或左线性正规文法 G，都存在一个有限自动机 (FA) M，使得 L(M) = L(G)。

2. 对每一个 FA M，都存在一个右线性正规文法 G_R 和左线性正规文法 G_L，使得 $L(M) = L(G_R) = L(G_L)$。

例： 设 DFA M = <{A, B, C, D}, {0, 1}, δ, A, {B}> 。 M 的状态转换图如下图所示。



- L(M) = 0(10)*
- G_R = <{0, 1}, {A, B, C, D}, A, P> ， 其中 P 由下列产生式组成：

  A→0 | 0B | 1D

  B→0D | 1C

  C→0 | 0B | 1D

  D→0D | 1D

$L(G_R) = L(M) = 0(10)^*$

例 设 DFA M = <{A, B, C, D, F }, {0, 1}, δ, A, {F}> 。 M 的状态转换图如下图所示。

- 从 NFA M 出发构造左线性正规文法 G_L = <{0, 1}, {B, C, D, F}, F, P′> ， 其中 P′ 由下列产生式组成:

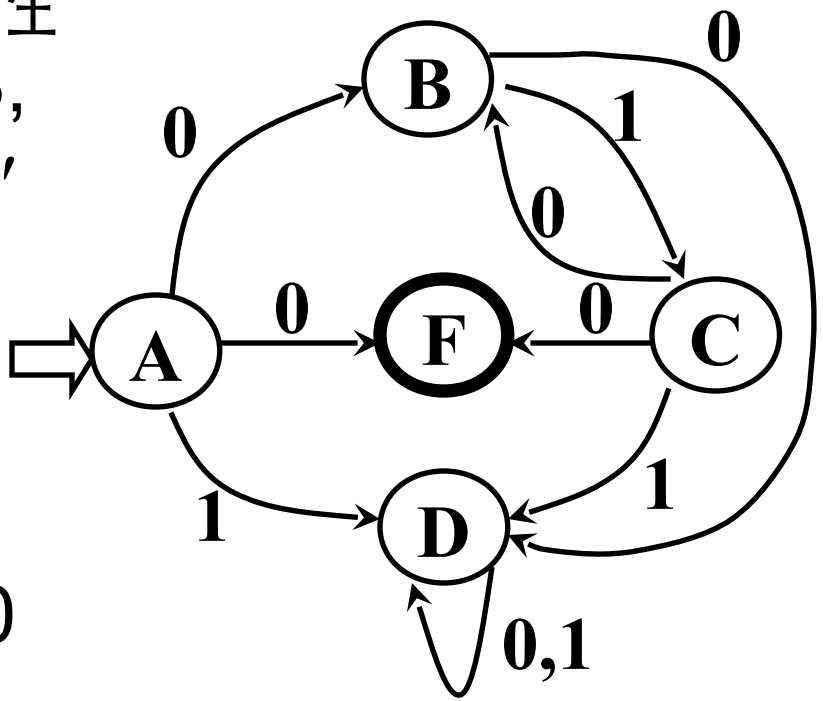
  F→0 | C0

  C→B1

  B→0 | C0

  D→1 | C1 | D0 | D1 | B0

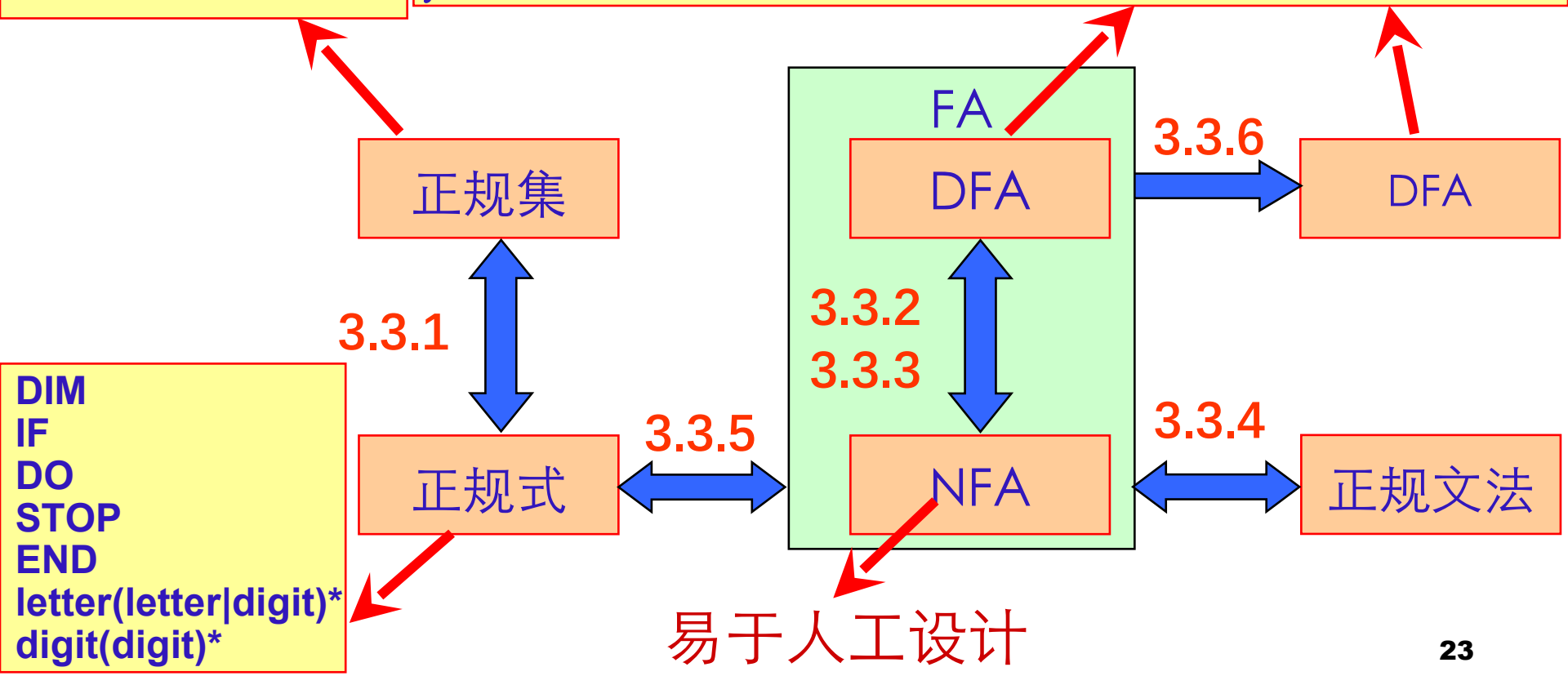
易证 $L(G_L) = L(M)$ 。



小结

```
curState = 初态
GetChar();
while( stateTrans[curState][ch] 有定义 ){
    // 存在后继状态，读入、拼接
    Concat();
    // 转换入下一状态，读入下一字符
   curState= stateTrans[curState][ch];
    if cur_state 是终态 then 返回 strToken 中的单
    GetChar( );
}
```

DIM,IF, DO,STOP,END
number, name, age
125, 2169
…



DIM
IF
DO
STOP
END
letter(letter|digit)*
digit(digit)*

第三章 词法分析

- 对于词法分析器的要求
- 词法分析器的设计
- 正规表达式与有限自动机
- 词法分析器的自动产生 --LEX

作业

- P64-7(选作 2 个小题)， 8(选作 3 个小题)， 12 ， 14