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

回顾

- 词法分析器的功能
- 词法分析器的设计
  - 状态转换图
  - 状态转换图的实现

是否有自动的方法产生词法分析程序？

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

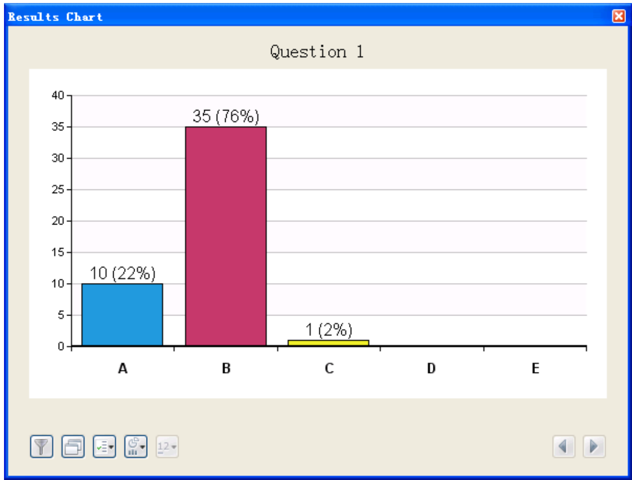
调查：词法分析程序

在操作系统的“ shell 命令解释器”实验中，你是如何设计和实现命令的单词识别程序的 (   )

A. 全部自己实现

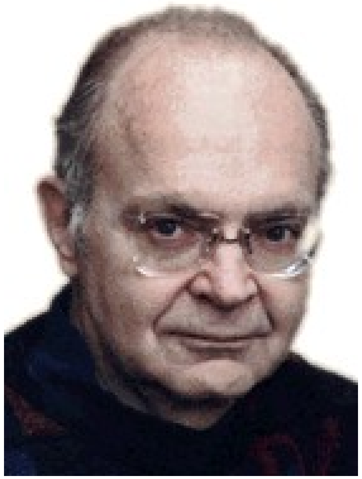
B. 使用 LEX(FLEX) 工具实现

C. 使用其它词法分析程序开发工具实现



iclicker

Knuth on *Theory and Practice*



Donald Ervin Knuth

Theory and practice are not mutually exclusive; they are intimately connected. They live together and support each other.

3.3 正规表达式与有限自动机

- 几个概念
  - 考虑一个有穷 字母表∑ 字符集
  - 其中每一个元素称为一个字符
  - ∑ 上的字（也叫字符串） 是指由∑中的字符所构成的一个有穷序列
  - 不包含任何字符的序列称为空字，记为 ε
  - 用∑ * 表示∑上的所有字的全体，包含空字 ε
  - 例如：设 ∑ ={a， b}，则

    $\Sigma^* = \{\varepsilon, a, b, aa, ab, ba, bb, aaa, ...\}$

- Σ^* 的子集 U 和 V 的连接（积）定义为

$$UV = \{ \alpha\beta \mid \alpha \in U \ \& \ \beta \in V \}$$

- V 自身的 n 次积记为

$$V^n = V\,V \cdots V$$

- 规定 $V^0 = \{\varepsilon\}$
- 令

$$V^* = V^0 \cup V^1 \cup V^2 \cup V^3 \cup \cdots$$

称 V^* 是 V 的闭包

- 记 $V^+ = V\,V^*$，称 V^+ 是 V 的正规闭包

3.3.1 正规式和正规集

- 正规集可以用正规表达式（简称正规式）表示
- 正规表达式是表示正规集一种方法
- 一个字集合是正规集当且仅当它能用正规式表示

冯 - 诺伊曼构造自然数的方案

- $\varnothing$ 0
- $\{\varnothing\}$ 1
- $\{\varnothing, \{\varnothing\}\}$ 2
- $\{\varnothing, \{\varnothing\}, \{\varnothing, \{\varnothing\}\}\}$ 3

正规式和正规集的递归定义

- 对给定的字母表Σ

  1)ε 和∅都是Σ上的正规式，它们所表示的正规集为 {ε} 和∅；

  2) 任何 a∈Σ ， a 是Σ上的正规式，它所表示的正规集为 {a}；

正规式和正规集的递归定义（续）

3）假定 e_1 和 e_2 都是Σ上的正规式，它们所表示的正规集为 $L(e_1)$ 和 $L(e_2)$，则

i）$(e_1 | e_2)$ 为正规式，它所表示的正规集为 $L(e_1) \cup L(e_2)$

ii）$(e_1.e_2)$ 为正规式，它所表示的正规集为 $L(e_1)L(e_2)$

iii）$(e_1)^*$ 为正规式，它所表示的正规集为 $(L(e_1))^*$

仅由有限次使用上述三步骤而定义的表达式才是Σ上的正规式，仅由这些正规式表示的字集才是Σ上的正规集。

- 所有词法结构一般都可以用正规式描述
- 若两个正规式所表示的正规集相同，则称这两个正规式等价。如

$$b(ab)^*=(ba)^*b$$

$L(b(ab)^*)$
$= L(b)L((ab)^*)$
$= L(b)\ (L(ab))^*$
$= L(b)\ (L(a)L(b))^*$
$=\{b\}\ \{ab\}^*$
$= \{b\}\ \{\varepsilon, ab, abab, ababab, \ldots\}$
$= \{b, bab, babab, bababab, \ldots\}$

请证明：

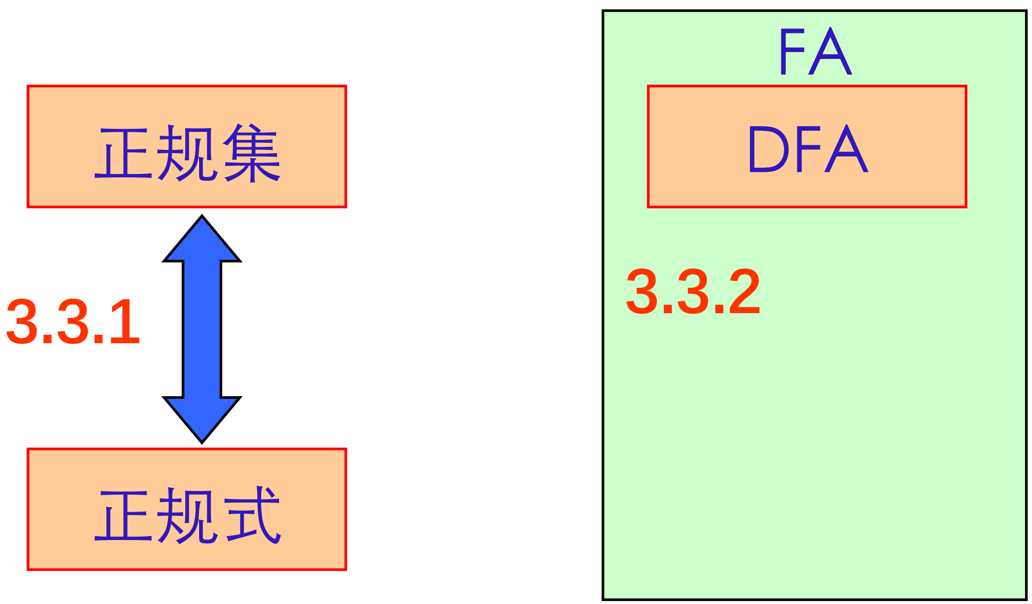
$$(a^*b^*)^*=(a|b)^*$$

$L(\ (ba)^*b)$
(b)
b)
* L(b)
$= \{ba\}^*\ \{b\}$
$= \{\varepsilon, ba, baba, bababa, \ldots\}\ \{b\}$
$= \{b, bab, babab, bababab, \ldots\}$

$\because L(b(ab)^*)= L(\ (ba)^*b) \quad \therefore b(ab)^*=(ba)^*b$

- 对正规式，下列等价成立：
  - $e_1 | e_2 = e_2 | e_1$　　交换律
  - $e_1 | (e_2 | e_3) = (e_1 | e_2) | e_3$　　结合律
  - $e_1 (e_2 e_3) = (e_1 e_2) e_3$　　结合律
  - $e_1 (e_2 | e_3) = e_1 e_2 | e_1 e_3$　　分配律
  - $(e_2 | e_3) e_1 = e_2 e_1 | e_3 e_1$　　分配律
  - $e\varepsilon = \varepsilon e = e$　　$e_1 e_2 <> e_2 e_1$

$$
\begin{aligned}
&\mathbf{L(e_1|e_2)} \\
&= \mathbf{L(e_1) \cup L(e_2)} \\
&= \mathbf{L(e_2) \cup L(e_1)} \\
&= \mathbf{L(e_2|e_1)}
\end{aligned}
$$


正规集
3.3.1
正规式
FA
DFA
3.3.2

Deterministic Finite Automata

3.3.2 确定有限自动机 (DFA)

- 对状态图进行形式化，则可以下定义：

确定有限自动机 (DFA)M 是一个五元式 $M=(S, \Sigma, f, S_0, F)$，其中：

1. S: 有穷状态集
2. Σ : 输入字母表 (有穷)
3. f: 状态转换函数，为 $S\times\Sigma\rightarrow S$ 的单值部分映射，f(s，a)=s' 表示：当现行状态为 s，输入字符为 a 时，将状态转换到下一状态 s'，s' 称为 s 的一个后继状态
4. $S_0\in S$ 是唯一的一个初态
5. $F\subseteq S$ ：终态集 (可空)

- 例如： DFA M=({0， 1， 2， 3}，{a，b}，f，0，{3})，其中：f 定义如下：

  f(0，a)=1　　f(0，b)=2

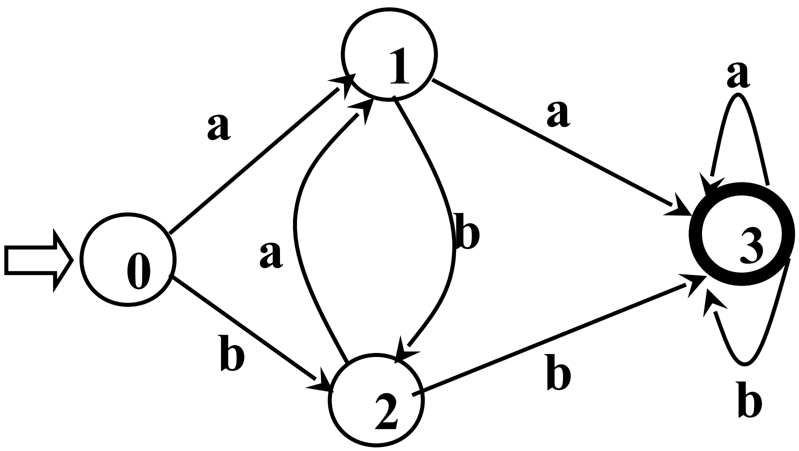
  f(1，a)=3　　f(1，b)=2

  f(2，a)=1　　f(2，b)=3

  f(3，a)=3　　f(3，b)=3

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

状态转换矩阵

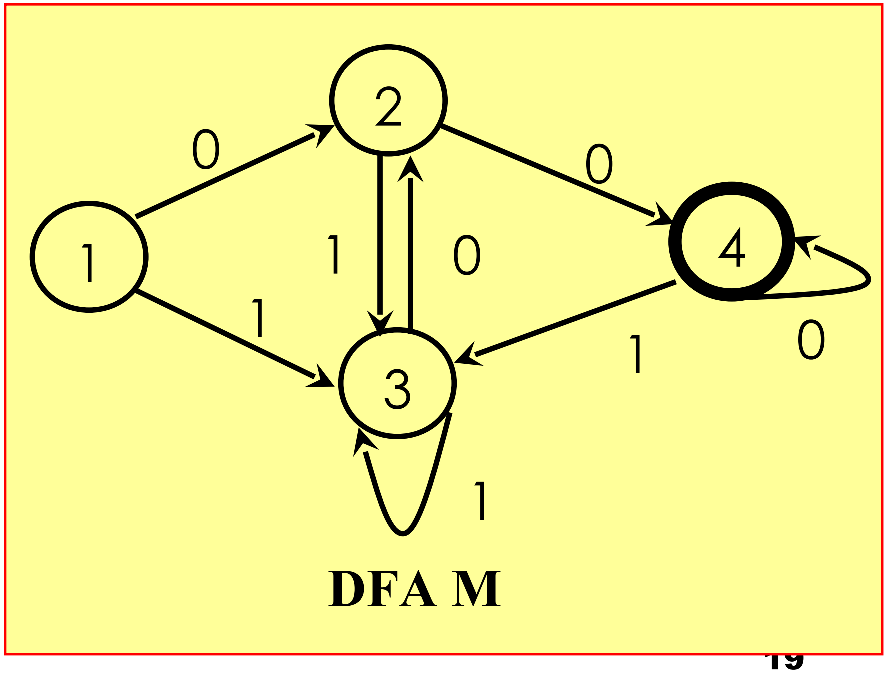


状态转换图

- DFA 可以表示为状态转换图
  - 假定 DFA M 含有 m 个状态和 n 个输入字符
  - 这个图含有 m 个状态结点，每个结点**顶多**含有 n 条箭弧射出，且每条箭弧用 Σ 上的不同的输入字符来作标记

- 对于Σ * 中的任何字α，若存在一条从初态到某一终态的道路，且这条路上所有弧上的标记符连接成的字等于α，则称α为 DFA M 所识别 (接收)
- DFA M 所识别的字的全体记为 L(M)

L(M)={ 以 00 结尾的串 }



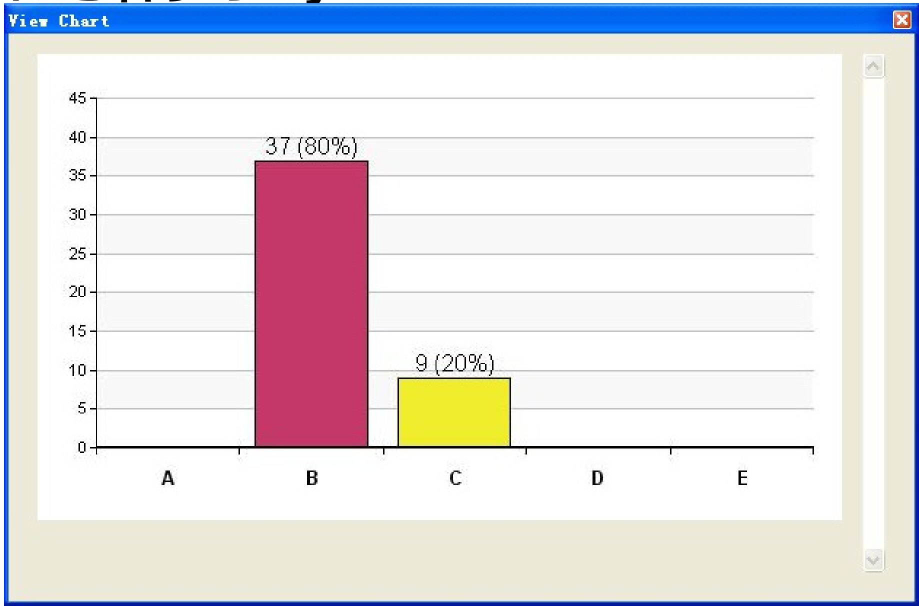
练习

- 图中 DFA M 识别的 L(M) 是什么?

A. L(M)={ 以 aa 或 bb 开头的字 }

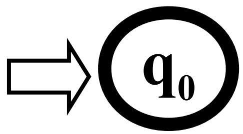
B. L(M)={ 含 aa 或 bb 的字 }

C. L(M)={ 以 aa 或 bb 结尾的字 }

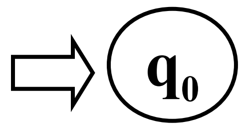


练习

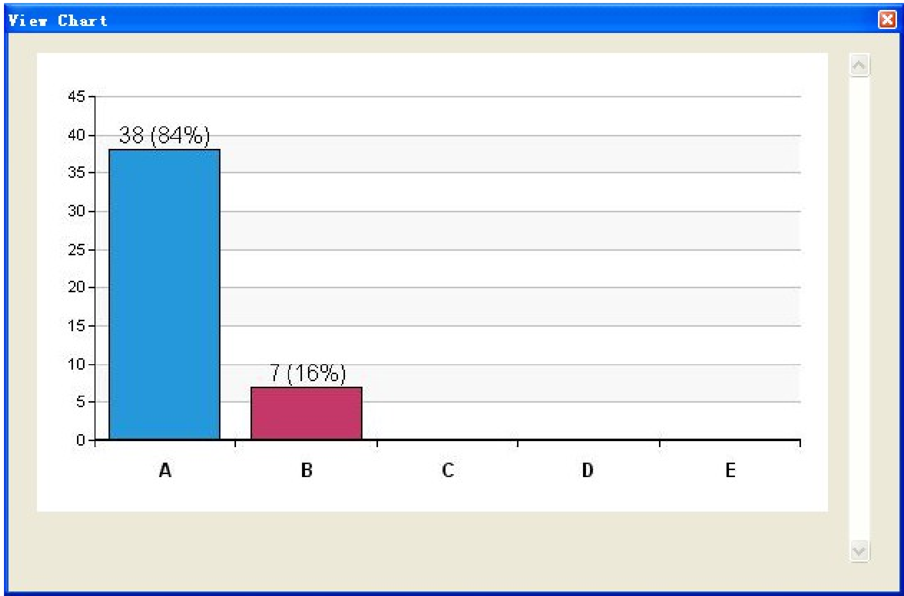
- 哪个 DFA 识别 {ε} ？



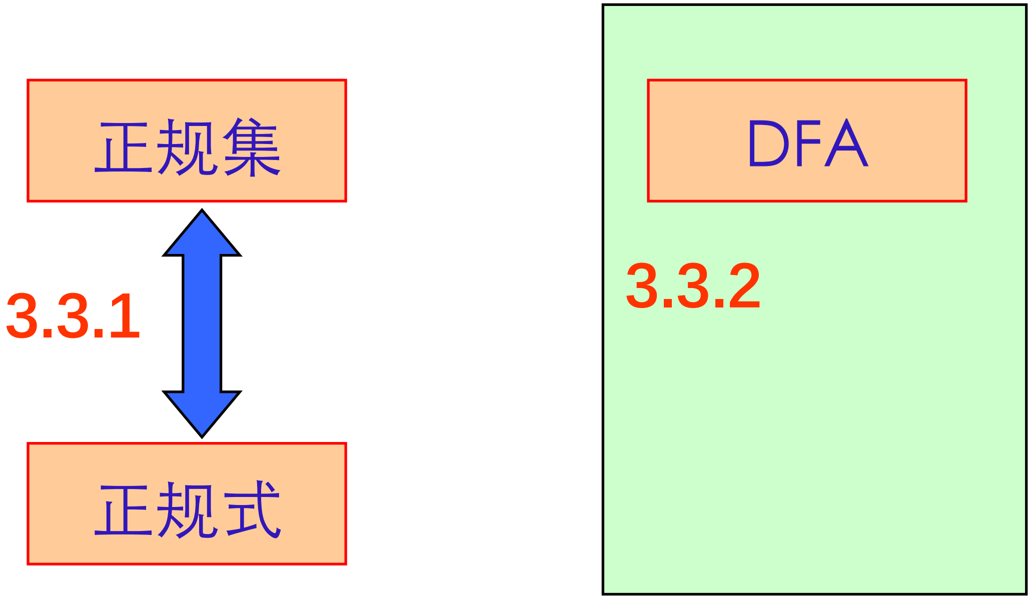
A



B



关系图



- 将证明：Σ上的字集 $V \subseteq \Sigma^*$ 是正规集，当且仅当存在Σ上的 DFA M，使得 $V = L(M)$