第3章 词法分析

Part I 正规文法与正规式

形式文法回顾 1

- 基本概念
  - 文法
  - 推导
  - 句型、句子
  - 语法树
  - 短语、直接短语、句柄

形式文法回顾 2

- 文法与语言的关系
  - 符号→符号串→文法→句子→语言
- 语言表示的两个途经
  - **生成**方式：文法
  - **识别**方式：自动机

文法与自动机的关系

0型文法（短语结构文法）：产生式形式为$\alpha\rightarrow\beta$，其中$\alpha\in(V_N\cup V_T)^+$，$\beta\in(V_N\cup V_T)^*$，可表征任何递归可枚举集，任何0型语言都是递归可枚举的；其表达能力相当于**图灵机**

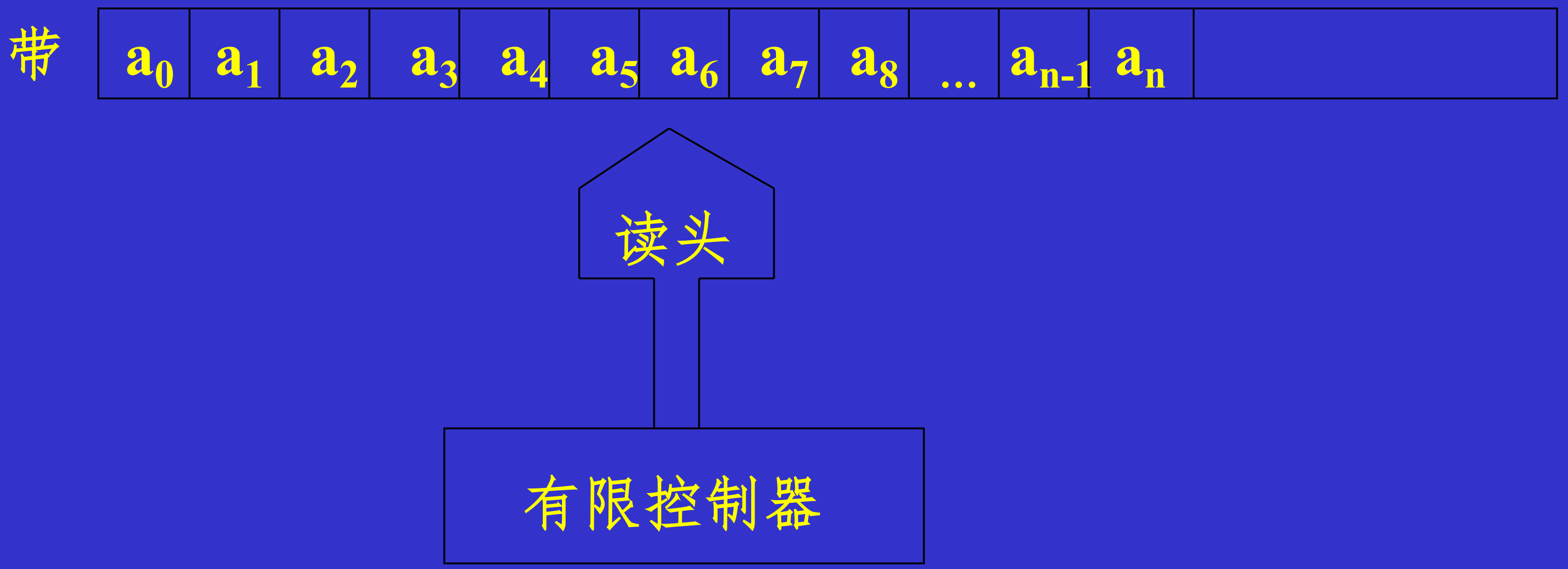
1型文法（上下文有关文法CSG）：产生式形式为$\alpha_1 A\alpha_2\rightarrow\alpha_1\beta\alpha_2$，即只有A出现在$\alpha_1$和$\alpha_2$的上下文中时，才允许$\beta$取代A；其识别系统是**线性有界自动机**

2型文法（上下文无关文法CFG）：产生式形式为A→β，β取代A时与A的上下文无关；其识别系统是不确定的**下推自动机（PDA）**

3型文法（正规文法RG）： 产生式形式为A→aB 或 A→a，产生的语言为正规语言；其识别系统是**有穷自动机（FA）**

任何能用图灵机(**Turing Machine**)描述的计算都能机械地实现，

任何能在计算机上实现的计算都能用图灵机描述



正规文法与自动机

- 正规文法与正规式(Regular Expression)
- 有限自动机(Finite Automata)
  - DFA
  - NFA
  - NFA->DFA
  - ε-FA
  - FA化简
  - FA与RG, RE的等价性

正规文法与正规式

- 单词符号结构的描述方法：
  - 正规文法（3型文法）
  - 正规式（正规表达式）

正规式

正规表达式(regular expression)是说明单词模式(pattern)的一种重要的**表示法**(记号)，是定义正规集的数学工具

正规式也称正则式、正则表达式

在编译中，用以描述单词符号

定义（正规式和它所表示的正规集）：

设字母表为Σ，辅助字母表Σ`={Φ，ε，｜，•，*，(，)}。

1. ε和Φ都是Σ上的正规式，它们所表示的正规集分别为{ε}和{}；

2. 任何$a \in \Sigma$，a是Σ上的一个正规式，它所表示的正规集为$\{a\}$；

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分别为$L(e_1)$和$L(e_2)$，那么，(e_1), $e_1 | e_2$, $e_1 \bullet e_2$, $e_1{}^*$也都是正规式,它们所表示的正规集分别为$L(e_1)$, $L(e_1) \cup L(e_2)$, $L(e_1)L(e_2)$和$(L(e_1))^*$。

4. 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，仅由这些正规式所表示的集合才是Σ上的正规集。

正规式中的符号

其中的“|”读为“或”（也有使用“+”代替“|”的）；

“•”读为“连接”；

“*”读为“闭包”（即，任意有限次的自重复连接）。

在不致混淆时，括号可省去，但规定算符的优先顺序为“*”、“•”、“|”。连接符“•”一般可省略不写。“*”、“•”和“|”都是左结合的。

例子

令Σ={a，b}，Σ上的正规式和相应的正规集的例子有：

正规式	正规集
a	{a}
a | b	{a,b}
ab	{ab}
(a | b)(a | b)	{aa,ab,ba,bb}
a*	{ε,a,aa, ……任意个a的串}

正规式	正规集
$(a\|b)^*$	{ε,a,b,aa,ab ……所有由a和b组成的串}
$(a\|b)^*(aa\|bb)(a\|b)^*$	{Σ^*上所有含有两个相继的a或两个相继的b的串}

例

令Σ={l, d} ,

则Σ上的正规式 $r = l(l|d)^*$的正规集为: {l, ll, ld, ldd,……}, 其中l代表字母, d代表数字, 正规式为 字母(字母|数字)* , 它表示的正规集中每个元素的模式是"字母打头的字母数字串", 也是多数程序设计语言允许的**标识符词法规则**

例

Σ={d，•，e，+，-},

则Σ上的正规式$dd^*(\bullet dd^*|\varepsilon)(e(+|-|\varepsilon)dd^*|\varepsilon)$表示无符号数的集合, 其中d为0~9的数字

程序设计语言的单词都能用正规式来定义

正规式等价

若两个正规式e_1和e_2所表示的正规集相同,则说e_1和e_2等价,写作$e_1=e_2$。

例如： $e_1=(a|b)$， $e_2=b|a$

又如： $e_1=b(ab)^*$， $e_2=(ba)^*b$

再如： $e_1=(a|b)^*$， $e_2=(a^*|b^*)^*$

正规式等价变换规则

设r,s,t为正规式， 正规式服从的代数规律有：

1. r | s=s | r　　“或” 的交换律
2. r | (s | t)=(r | s) | t　　“或” 的结合律
3. (rs)t=r(st)　　“连接” 的结合律
4. r(s | t)=rs | rt
   (s | t)r=sr | tr　　分配律

5. $\varepsilon r=r,\ r\varepsilon=r$ 　　　　ε是“连接”的恒等元素

6. $r \mid r=r$

$r^*=\varepsilon \mid r \mid rr \mid \ldots$ 　　　　“或”的抽取律

正规文法和正规式

- 标识符的文法描述
  - G =({ DIGIT, LETTER }, {idn}, P, idn)
  - idn → LETTER
  - idn → idn DIGIT
  - idn → idn LETTER

- 标识符的正则式描述:
  - $LETTER(LETTER|DIGIT)^*$

正规式到正规文法

对∑上的正规式r ,存在一个$RG=(V_N,V_T,P,S)$:

$L(G)=L(r)$

初始，$V_T=\Sigma$, $S \in V_N$，生成正规产生式：$S \to r$

(R.1) 对形如 $A \to r_1 r_2$的正规产生式：$A \to r_1 B$

$B \to r_2$ $B \in V_N$

(R 2) 对形如$A \to r^* r_1$的正规产生式： $A \to rB$

$A \to r_1$

$B \to rB$

$B \to r_1$ $B \in V_N$

(R 3)对形如$A \to r_1 | r_2$的正规产生式: $A \to r_1$

$A \to r_2$

不断应用R做变换，直到每个产生式右端至多有一个V_N

例　$r=a(a|d)^*$

$S \to a(a|d)^*$

$S \to aA$　　　$A \to (a|d)^*$

$A \to (a|d)B$

$A \to \varepsilon$

$B \to (a|d)B$

$B \to \varepsilon$

G[S]:

$S \to aA$

$A \to \varepsilon$　　　$V_T=\{a,d\}$

$A \to aB$　　　$V_N=\{S,A,B\}$

$A \to dB$

$B \to aB$

$B \to dB$

$B \to \varepsilon$

例：标识符定义的转换

- 引入 id

  $id \rightarrow let\ (let \mid dig)^*$

- 引入 rid 消除连接

  $rid \rightarrow (\ let \mid dig\ )^*$

  $rid \rightarrow \varepsilon \mid (let \mid dig)B$

  $B \rightarrow (let \mid dig)B$

  $B \rightarrow \varepsilon$

$id \rightarrow let\ rid$
$rid \rightarrow \varepsilon \mid let\ B \mid dig\ B$
$B \rightarrow let\ B \mid dig\ B$
$B \rightarrow \varepsilon$

正规文法到正规式

对$G=(V_N,V_T,P,S)$,存在一个$\Sigma=V_T$上的正规式r : $L(r)=L(G)$

$A\rightarrow xB,\quad B\rightarrow y$	$\approx$	$A=xy$
$A\rightarrow xA \mid y$	$\approx$	$A=x^*y$
$A\rightarrow x \mid y$	$\approx$	$A=x \mid y$

例子

$G[s]: S \to aA | a$

$A \to aA \mid a \mid dA \mid d$

$A \to (a \mid d)A \mid (a \mid d)$

$A = (a \mid d)^*(a \mid d)$

$S = a(a \mid d)^*(a \mid d) \mid a$

$= a((a \mid d)^*(a \mid d) \mid \varepsilon)$

$= a((a \mid d)^+ \mid \varepsilon)$

$r = a(a \mid d)^*$

请指出下列哪些字符串包含在正规式 ab*c*(a|b)c所对应的正规集合中：

acac

acbbc

abbcac

abc

acc

作 业

- 通读3.1、3.3节

- 练习8