形式语言和自动机初步

第10章 形式语言和自动机初步

- 10.1 形式语言和形式文法
- 10.2 有穷自动机
- 10.3 有穷自动机和正则文法的等价性
- 10.4 图灵机

10.1 形式语言与形式文法

- 字符串和形式语言
- 形式文法
- 形式文法的分类
  - 0型文法
  - 1型文法 或上下文有关文法
  - 2型文法 或上下文无关文法
  - 3型文法 或正则文法

语言的基本要素

汉语

字符:汉字和标点符号

字符集:合法字符的全体

句子:一串汉字和标点符号

语法:形成句子的规则

形式语言

字符

字母表

字符串

形式文法

字符串

字母表Σ: 非空的有穷集合

字符串: Σ中符号的有穷序列

如 $\Sigma=\{a,b\}$

$a, b, aab, babb$

字符串ω的长度$|\omega|$: ω中的字符个数

如 $|a|=1$, $|aab|=3$

空字符串ε: 长度为0, 即不含任何符号的字符串

a^n : n个a组成的字符串

Σ^*: Σ上字符串的全体

子字符串(子串):

字符串中若干连续符号组成的字符串

前缀: 最左端的子串

后缀: 最右端的子串

例如 $\omega=abbaab$

a,ab,abb是ω的前缀

aab,ab,b是ω的后缀

ba是ω的子串, 但既不是前缀, 也不是后缀

ω本身也是ω的子串, 且既是前缀, 也是后缀

ε也是ω的子串, 且既是前缀, 也是后缀

字符串的连接运算

设$\alpha = a_1a_2 \ldots a_n$, $\beta = b_1b_2 \ldots b_m$,

$\alpha\beta = a_1a_2 \ldots a_nb_1b_2 \ldots b_m$称作$\alpha\beta$与$\beta$作的连接

如 $\alpha = ab$, $\beta = baa$, $\alpha\beta = abbaa$, $\beta\alpha = baaab$

对任意的字符串$\alpha\beta$, β, γ

(1) $(\alpha\beta)\gamma = \alpha(\beta\gamma)$

即, 连接运算满足结合律

(2) $\varepsilon\alpha = \alpha\varepsilon = \alpha$

即, 空串ε是连接运算的单位元

n个α的连接记作α^n

如 $(ab)^3 = ababab$, $\alpha^0 = \varepsilon$

形式语言

定义: Σ^*的子集称作字母表Σ上的形式语言,
简称 语言

例如 $\Sigma=\{a,b\}$

$A=\{a,b,aa,bb\}$

$B=\{a^n \mid n\in N\}$

$C=\{a^n b^m \mid n,m\geq 1\}$

$D=\{\varepsilon\}$

空语言$\varnothing$

形式文法

一个例子——标识符

<标识符> : <字母> | <下划线> | <标识符> <字母> | <标识符> <下划线>| <标识符><数字>

<字母> : a | b | … | z | A | B | … | Z

<下划线> : _

<数字> : 0 | 1 | … | 9

形式文法的定义

定义 形式文法是一个有序4元组$G=<V,T,S,P>$,

其中

(1) V是非空有穷集合, V的元素称作变元或非终极符

(2) T是非空有穷集合且$V\cap T=\varnothing$, T的元素称作终极符

(3) $S\in V$称作起始符

(4) P是非空有穷集合, P的元素称作产生式或改写规则, 形如$\alpha\rightarrow\beta$, 其中$\alpha,\beta\in(V\cup T)^*$且$\alpha\neq\varepsilon$.

文法生成的语言

设文法 $G=\langle V,T,S,P\rangle$, $\omega,\lambda\in(V\cup T)^*$,

$\omega\Rightarrow\lambda$: 存在$\alpha\to\beta\in P$和$\xi,\eta\in(V\cup T)^*$, 使得

$$\omega=\xi\alpha\eta,\quad \lambda=\xi\beta\eta$$

称ω直接派生出λ.

$\omega\overset{*}{\Rightarrow}\lambda$: 存在$\omega_1, \omega_2, \ldots, \omega_m$, 使得

$$\omega=\omega_1\Rightarrow\omega_2\Rightarrow\ldots\Rightarrow\omega_m=\lambda$$

称ω 派生出λ.

恒有 $\omega\overset{*}{\Rightarrow}\omega$ (当m=1时)

$\overset{*}{\Rightarrow}$是 $\Rightarrow$ 的自反传递闭包

文法生成的语言

定义 设文法$G=<V,T,S,P>$, G生成的语言

$$L(G)=\{\omega\in T^* \mid S \stackrel{*}{\Rightarrow}\omega\}$$

$L(G)$由所有满足下述条件的字符串组成:

(1) 仅含终结符;

(2) 可由起始符派生出来.

定义 如果$L(G_1)=L(G_2)$, 则称文法G_1与G_2等价.

举例

例1 文法$G_1 = <V,T,S,P>$,其中$V=\{S\}$, $T=\{a,b\}$,

P: $S \rightarrow aSb \mid ab$

$L(G_1)=\{a^n b^n \mid n>0\}$

例2 文法$G_2 = <V,T,S,P>$,其中$V=\{A,B,S\}$, $T=\{0,1\}$,

P: $S \rightarrow 1A$, $A \rightarrow 0A \mid 1A \mid 0B$, $B \rightarrow 0$

$L(G_2)=\{1x00 \mid x \in \{0, 1\}^*\}$

例3 文法$G_3 = <V,T,S,P>$,其中$V=\{A,B,S\}$, $T=\{0,1\}$,

P: $S \rightarrow B0$, $B \rightarrow A0$, $A \rightarrow A1 \mid A0$, $A \rightarrow 1$

$L(G_3)= L(G_2)$,　G_3与G_2等价

举例 (续)

例4 $G=<V,T,S,P>$,其中$V=\{S,A,B,C,D,E\}$, $T=\{a\}$,

P: (1) $S\to ACaB$ (2) $Ca\to aaC$ (3) $CB\to DB$

(4) $CB\to E$ (5) $aD\to Da$ (6) $AD\to AC$

(7) $aE\to Ea$ (8) $AE\to \varepsilon$

试证明: $\forall i\geq 1, S \stackrel{*}{\Rightarrow} a^{2^i}$

证: a^2 和 a^4 的派生过程

$S \Rightarrow ACaB$ (1)

$\Rightarrow AaaCB$ (2)

$\Rightarrow AaaE$ (4)

$\stackrel{*}{\Rightarrow} AEaa$ 2次 (7)

$\Rightarrow a^2$ (8)

例4 (续)

$$
\begin{aligned}
S &\overset{*}{\Rightarrow} AaaCB & & \\
&\Rightarrow AaaDB & & (3) \\
&\overset{*}{\Rightarrow} ADaaB & & 2\text{次}(5) \\
&\Rightarrow ACaaB & & (6) \\
&\overset{*}{\Rightarrow} AaaaaCB & & 2\text{次}(2) \\
&\Rightarrow AaaaaE & & (4) \\
&\overset{*}{\Rightarrow} AEaaaa & & 4\text{次}(7) \\
&\Rightarrow a^4 & & (8)
\end{aligned}
$$

例4(续)

先用归纳法证明 $\forall i \geq 1,\ S \overset{*}{\Rightarrow} Aa^{2^i}CB$
当 i=1 时结论成立, 假设对i 结论成立,

$$
\begin{aligned}
S &\overset{*}{\Rightarrow} Aa^{2^i}CB \\
&\Rightarrow Aa^{2^i}DB && (3) \\
&\overset{*}{\Rightarrow} ADa^{2^i}B && 2^i\text{ 次}(5) \\
&\Rightarrow ACa^{2^i}B && (6) \\
&\overset{*}{\Rightarrow} Aa^{2^{i+1}}CB && 2^i\text{ 次}(2)
\end{aligned}
$$

得证对 i +1 结论成立, 故对所有的 i 成立.

例4 (续)

于是, $\forall i \geq 1$,

$$
\begin{aligned}
S &\overset{*}{\Rightarrow} Aa^{2^i}CB \\
&\Rightarrow Aa^{2^i}E && (4) \\
&\overset{*}{\Rightarrow} AEa^{2^i} && 2^i\text{ 次}(7) \\
&\Rightarrow a^{2^i} && (8)
\end{aligned}
$$

可以证明:

$$L(G) = \{\ a^{2^i} \mid i \geq 1\}$$

形式文法的分类 —Chomsky谱系

0型文法(短语结构文法,无限制文法)

1型文法(上下文有关文法):

所有产生式$\alpha \to \beta$, 满足 $|\alpha| \le |\beta|$

另一个等价的定义: 所有的产生式形如

$$\xi A \eta \to \xi \alpha \eta$$

其中$A \in V$, $\xi, \eta, \alpha \in (V \cup T)^*$,且$\alpha \neq \varepsilon$

2型文法(上下文无关文法):

所有的产生式形如 $A \to \alpha$

其中$A \in V, \alpha \in (V \cup T)^*$,

形式文法的分类 (续)

3型文法(正则文法): 右线性文法和左线性文法的统称

右线性文法: 所有的产生式形如

$$A \to \alpha B \quad 或 \quad A \to \alpha$$

左线性文法: 所有的产生式形如

$$A \to B\alpha \quad 或 \quad A \to \alpha$$

其中$A,B \in V, \alpha \in T^*$

例1是上下文无关文法

例2是右线性文法,例3是左线性文法,都是正则文法

例4是0型文法

Chomsky谱系

0型语言: 0 型文法生成的语言

1型语言(上下文有关语言): 如果L-{ε}可由1型文法生成, 则称 L 是1型语言

2型语言(上下文无关语言) : 2 型文法生成的语言

3型语言(正则语言): 3 型文法生成的语言

如 $\{1x00 \mid x\in\{0, 1\}^*\}$ 是正则语言 (例1)

$\{a^n b^n \mid n>0\}$ 是上下文无关语言 (例2,3)

$\{ a^{2^i} \mid i \geq 1\}$ 是 0 型语言 (例4)

定理 0型语言⊃1型语言⊃2型语言⊃3型语言

描述算术表达式的文法

$G=\{\{E,T,F\},\{a,+.-.*,/,(,)\},E,P\}$

其中E:算术表达式, T:项,

F:因子, a:数或变量

$$P:\ E \rightarrow E+T \mid E-T \mid T$$

$$T \rightarrow T*F \mid T/F \mid F$$

$$F \rightarrow (E) \mid a$$

这是上下文无关文法

左、右线性文法的等价性

定理 设G是右(左)线性文法,则存在左(右)线性文法 G'使得$L(G')=L(G)$.

证明: G'用模拟G

$G=<V,T,S,P>$

P: $A\to\alpha B$

$A\to\alpha$

$G'=<V\cup\{S'\},T,S',P'>$

P': $B\to A\alpha$

$S'\to A\alpha$

$S\to\varepsilon$

一个实例——模拟例2中的G_2

$G_2 = \langle V,T,S,P \rangle$

$V=\{A,B,S\}$

$T=\{0,1\}$

P: $S \rightarrow 1A$

$A \rightarrow 0A$

$A \rightarrow 1A$

$A \rightarrow 0B$

$B \rightarrow 0$

$G_2' = \langle V',T',S',P' \rangle$

$V'=\{A,B,S,S'\}$

$T'=\{0,1\}$

P': $A \rightarrow S1$

$A \rightarrow A0$

$A \rightarrow A1$

$B \rightarrow A0$

$S' \rightarrow B0$

$S \rightarrow \varepsilon$

可删去G_2'中的S, 这实际上就是G_3