第 2 章　形式语言基础

【前 言】

计算机处理语言，首先应考虑语言的形式化、规范化，使其具有可计算性和可操作性；这就是形式语言理论研究的问题。

形式语言诞生于1956年，由chomsky创立。通常，语言研究至少涉及三个方面：语法、语义和语用；

这里仅侧重于语法的研究。

形式语言的基本观点是 ：

语言是符号串之集合！

形式语言理论研究的基本问题是：

研究符号串集合的表示方法、结构特性以及运算规律。

第 2 章　形式语言基础

【内容提要】

2.1　形式语言是符号串集合
2.2　形式语言是由文法定义的
2.3　主要语法成分的定义
2.4　两类特性文法
2.5　文法变换方法
2.6　关于形式语言的分类问题

2.1 形式语言是符号串集合

【形式语言】是字母表上的符号，按一定的规则组成的所有符号串集合；其中的每个符号串称为句子。

【名词解释】：

- 字母表 -- 元素(符号)的非空有限集合；
- 符号串 -- 符号的有限序列；
- 符号串集合 -- 有限个或者无限个符号串组成的集合；
- 规　则 -- 以某种形式表达的在一定范围内共同遵守的章程和制度；这里，指符号串的组成规则。

➢ 形式语言概念示例：

两个语言！

【例2. 1】 $L_1=\{\ 00, 01, 10, 11\ \}$；

字母表$\Sigma_1=\ \{0, 1\}$，

句子有：00, 01, 10, 11

【例2. 2】 $L_2=\{\ ab^mc, b^n \mid m>0, n\geqslant 0\ \}$

字母表$\Sigma_2=\ \{a, b, c\}$，

- 句型1： ab^mc ，

  有句子： abc， abbc， abbbc, …

- 句型2： b^n ；

  有句子： ε， b， bb， bbb, …

【注】 (1) $b^0=\varepsilon$（空符号串），$b^1=b, b^2=bb, b^3=bbb, \cdots$

(2) L_1 为有限语言；L_2 为无限语言。

2.1.1 符号串(集合)的运算

Ⅰ. 符号串的运算

设 α ， β 为两个符号串，则：

1. 连接：$\alpha.\beta = \alpha\beta$　　如　a. b=ab

2. 或：$\alpha|\beta = \alpha$（或者 β）

3. 方幂：$\alpha^n = \underbrace{\alpha\alpha \ldots \alpha}_{n个} = \alpha\alpha^{n-1} = \alpha^{n-1}\alpha$

※ $\alpha^0 = \varepsilon$（空符号串）

↳ 什么符号也没有的符号串 ！

$\alpha^1 = \alpha$ ； $\alpha^2 = \alpha\alpha$ ； …

4. 闭包：

α 的正闭包：$\alpha^+ = \alpha^1|\alpha^2|\ldots|\alpha^n|\ldots$

α 的星闭包：$\alpha^* = \alpha^0|\alpha^1|\alpha^2|\ldots|\alpha^n|\ldots$

2.1.1 符号串(集合)的运算(续1)

※ 符号串运算示例

【例】：

(1) abc. de= abcde

(2) ab|cd = ab（或者 cd ）

(3) $(a|b)^1 = (a|b) = a|b$

$(a|b)^2 = (a|b)(a|b) = aa|ab|ba|bb$

…

$(a|b)^* = (a|b)^0 | (a|b)^1 | (a|b)^2 | \ldots$

即：$(a|b)^* = (a|b)^n$ ， $n \geqslant 0$

同理：$(a|b)^+ = (a|b)^n$ ， $n > 0$

2.1.1 符号串(集合)的运算(续2)

Ⅱ. 符号串集合的运算

设 A 和 B 为两个符号串集合，则：

1. 乘积： $AB=\{xy \mid x\in A \text{ 且 } y\in B\}$

2. 和： $A\cup B=A+B=\{x \mid x\in A \text{ 或 } x\in B\}$

3. 方幂： $A^n = \underbrace{AA\ldots A}_{n\text{个}} = AA^{n-1} = A^{n-1}A$

※ $A^0=\{\varepsilon\}$；

$A^1=A$ ； $A^2=AA$ ； $A^3=AAA$ ； …

4. 闭包：

A 的正闭包： $A^+ = A^1\cup A^2\cup\ldots\cup A^n\cup\ldots$

A 的星闭包： $A^* = A^0\cup A^1\cup A^2\cup\ldots\cup A^n\cup\ldots$

➢ 符号串集合运算示例：

【例2.3】 设 $A=\{a, b\}, B=\{c, d\}$

则 $A+B=\{a, b, c, d\}$

则 $AB=\{xy \mid x\in A, y\in B\}=\{ac, ad, bc, bd\}$

【例2.4】 设 $A=\{a\}$

则 $A^* = A^0 \cup A^1 \cup A^2 \cup \ldots \cup A^n \cup \ldots$

$=\{\varepsilon\}+\{a\}+\{aa\}+\{aaa\}+\ldots$

$=\{\varepsilon, a, aa, aaa, \ldots\}$

$=\{a^n \mid n\geqslant 0\}$

➢ 符号串集合运算示例(续)：

【例2.5】 设 $A=\{a, b\}$， $A^*=?$

$\because A^* = A^0 \cup A^1 \cup A^2 \cup \cdots \cup A^n \cup \cdots$

$A^0 = \{\varepsilon\}$;

$A^1 = A = \{a, b\}$;

$A^2 = A.A = \{a, b\}.\{a, b\} = \{aa, ab, ba, bb\}$;

$A^3 = A.A^2 = \{a, b\}.\{aa, ab, ba, bb\}$

$= \{aaa, aab, aba, abb, baa, bab, bba, bbb\}$;

…

$\therefore A^* = \{ x \mid x=(a|b)^n, n \geqslant 0 \}$

推论：若 A为任一字母表，则 A^* 就是该字母表上的所有符号串(包括空串)的集合。

2.1.2 符号串集合的文法描述

长久以来，探讨符号串集合(即形式语言)的各种描述方法，一直是计算机语言处理的重要任务之一。

【例2.5】 $L = \{ ab^nc \mid n \geqslant 0 \}$，

字母表：$\Sigma = \{a, b, c\}$；

展开：$L = \{ac, abc, abbc, abbbc, \cdots\}$

右图给出的表示方法 --文法规则；

$$S \rightarrow aAc$$
$$A \rightarrow bA \mid \varepsilon$$

(1) S, A — 定义的对象(S 句子, 最大的定义对象，又称为开始符号；A为句型aAc的短语），

(2) a, b, c -- 为字母表Σ中的符号；ε - 空符号串。

(3) -> , | -- 为描述符号(-> 定义为；| 或者是)

➢ 规则应用说明示例：

怎样利用上述文法规则表示语言L?

从开始符号出发，对符号串中的定义对象，采用推导的方法（用其规则右部替换左部）产生新的符号串，如此进行，直到新符号串中不再出现定义的对象为止，则最终的符号串就是一个句子。

【句子产生过程】(=> 推导算符)：

① S => aAc => a ε c = ac

② S => aAc => abAc => ab ε c = abc

③ S => aAc => abAc=> abbAc => abbc

…

∴ $S \overset{+}{\Rightarrow} ab^nc$, $n \geqslant 0$

S -> aAc

A -> bA | ε

2.2 形式语言是由文法定义的

2.2.1 什麽是文法 ？

【定义】 文法(grammar)是规则的有限集，其中的上下文无关文法可定义为四元组：

$G(Z)=(V_N, V_T, Z, P)$

每个元素

V_N ： 非终结符集（定义的对象集，如：语法成分等）；
V_T ： 终结符集（字母表）；
Z ： 开始符号（研究范畴中，最大的定义对象）；
P ： 规则集（又称产生式集）；

每个规则

$A \to \alpha$ 或者 $A \to \alpha \mid \beta$

其中， 描述符号 ： →(定义为)，|(或者是)

文法符号 ： $Z, A \in V_N$， $\alpha, \beta \in (V_N+V_T)^*$

2.2 形式语言是由文法定义的（续3）

【注意】 提供了规则集，就相当给出了一个文法：

G(S):

S -> aAc
A -> bA | ε

$G(Z)=(V_N,\ V_T,\ Z,\ P)$

$V_N=\{ S, A \}$；$V_T=\{ a, b, c \}$；

Z = S ；　　P ：

2.2.2 文法是怎样定义语言的？

设 有文法 G(Z)， L(G)为G所定义的语言；

则 $L(G)=\{ x \mid Z \overset{+}{\Rightarrow} x, x \in V_T^* \}$ 〖2.1〗

利用 => 进行连续推导之意！

即：一个文法所定义的语言，就是由该文法开始符号推导出的所有仅含终结符的符号串之集合。其中的每个符号串皆称为句子。

【例2.6】标识符的文法

【标识符】 指字母开头的字母、数字序列。

令 $G(Z)=(V_N, V_T, Z, P)$

则 $V_N=\{I(标识符), A(标识符尾)\}$;

$V_T=\{\ell(字母), d(数字)\}$;

$Z = I$;

P :

$I \rightarrow \ell A \mid \ell$

$A \rightarrow \ell A \mid d A \mid \varepsilon$

同理，【无符号整数】文法 可写成：

G(N)：

$N \rightarrow d N \mid d$

※其四元组也可写成：G(N)=({N}, {d}, N, P)

※标识符文法所定义的语言求解：

上面构造的标识符文法属于正规文法（定义在后）类，正确性检验较容易；下面给出一个算法：

$$I \rightarrow \ell A \mid \ell$$
$$A \rightarrow \ell A \mid d A \mid \varepsilon$$

令：$I = \ell A \mid \ell$

$A = \ell A \mid d A \mid \varepsilon$

} 正规方程式

※ 求解 I 值步骤：

① 先求解 A：$A=(\ell|d)A$，$A=(\ell|d)^2 A$，…，$A=(\ell|d)^n A$

代入 $A=\varepsilon$ 得：$A=(\ell|d)^n$，$n\geq 0$

② ∵ $I=\ell A \mid \ell$ 代入 $A=(\ell|d)^n$，$n\geq 0$

得：$I=\ell(\ell|d)^n$，$n\geq 0$

《标识符》：字母开头的字母、数字序列；

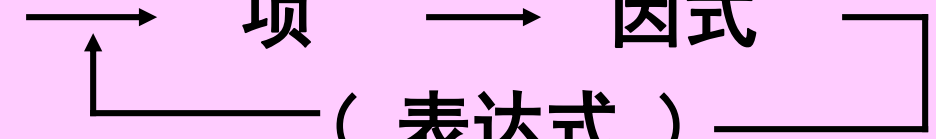
【例2.7】简单算术表达式文法

令　G(Z)=（V_N，V_T，Z，P）

则　V_N ={ E(算术表达式),T(项)，F(因式)};
　　V_T ={ i(变量或常数),+,-,*,/,(,) };
　　Z = E ;
　　P :

E -> T | E + T | E - T
T -> F | T * F | T / F
F -> i | (E)

【注】此文法定义了算术表达式的层次嵌套结构:



※ 算术表达式文法应用示例：

证明 i*(i+i−i)

是文法G(E)的一个句子

(即 合法的算术表达式):

G(E)： E->T |E+T |E−T

T->F |T* F |T/F

F->i |(E)

根据 语言定义式〖2.1〗,合法的算术表达式是指：

$E \overset{+}{\Rightarrow} i*(i+i-i)$ 成立吗？

∵ E =>T =>T*F =>T*(E) =>T*(E−T)

=>T*(E+T−T)=>F*(E+T−T)=>i*(E+T−T)

=>... =>i*(i+i−i)

∴ $E \overset{+}{\Rightarrow} i*(i+i-i)$

观察推导过程，可以看到：一旦产生式选择错了，会导致失败！