Chapter 1

简介

1.1 形式语言与自动机理论

计算机科学是关于计算知识的有系统的整体, 其始源可回溯到欧几里德关于一些算法的设计和巴比伦人关于渐进复杂性和归约性的使用. 然而, 现今的计算机学科的发展, 起源于两个重要的事件: 现代数字计算机的出现和算法概念的形式化.

计算机科学有两个主要的部分: 第一, 构成计算系统基础的一些基本概念和模型; 第二, 设计计算系统 (软件和硬件) 的工程技术. 形式语言与自动机理论, 就是作为第一部分, 即构成计算基础的基本概念的引论.

1.2 基础知识

1.2.1 基本概念

(1) 字母表 (*Alphabet*): 有穷非空符号集. 例如, $\Sigma = \{0, 1\}$, $\Sigma = \{a, b, \dots, z\}$. (“符号”是一个抽象的实体, 我们不再去形式的定义它, 如同几何学中对“点”和“线”的概念不加定义一样.)

(2) 字符串 (*Strings*): 某个字母表中符号的有穷序列, 也称字 (*words*). 例如, 若 $\Sigma = \{0, 1\}$, 则 000, 111, 0101, 10101 为 Σ 上的字符串.

(3) 空串 (*Empty string*): 长度为 0 的串. 一般表示为 ε.

(4) 串的长度: 串中符号的个数. 更准确的说, 是串中符号所占的位置数. 若串为 w, 那么长度记为 $|w|$. 例如, $|011| = 3$, $|\varepsilon| = 0$.

(5) 串的连接 (*Concatenation*): 如果 x 和 y 是串, 那么 xy 标识将 x 和 y 进行连接后得到的串. 例如 $x = 01101$ 和 $y = 110$, 那么 $xy = 01101110$.

对任意串 w, 都有 $\varepsilon w = w\varepsilon = w$, 所以 ε 也称为连接运算的单位元.

对任意串 w, 记 $w^0 = \varepsilon$ 和 $w^n = w^{n-1}w$.

例如 $a^3b^2 = aaabb$, $0^n1^n = 00\cdots011\cdots1$(0 和 1 都是 n 个).

(6) 串的逆序 (*Reverse*): 若 $w = a_1a_2\dots a_n$, 则记 w 的逆序为 $w^R = a_na_{n-1}\dots a_1$.

(7) 集合的连接 (或乘积): $AB = \{ab \mid a \in A, b \in B\}$.

例如 $\{0, 1\}\{0, 1\} = \{00, 01, 10, 11\}$; $\{0, 1\}\{a, b, c\} = \{0a, 0b, 0c, 1a, 1b, 1c\}$.

(8) 集合的幂: 递归定义 (1) $\Sigma^0 = \{\varepsilon\}$; (2) $\Sigma^n = \Sigma^{n-1}\Sigma$ $(n \geq 1)$.

那么, 若 Σ 是字母表, 那么定义 Σ^k 就是, 长度为 k 的串的集合. 而且 $\Sigma^0 = \{\varepsilon\}$ 对任何 Σ 都成立.

例如, 若 $\Sigma = \{0, 1\}$, 那么 $\Sigma^2 = \{00, 01, 10, 11\}$, $\Sigma^3 = \{000, 001, 010, 011, 100, 101, 110, 111\}$.

(9) 正闭包 (*Positive closure*): $\Sigma^+ = \Sigma \cup \Sigma^2 \cup \Sigma^3 \cup \Sigma^4 \cup \cdots$

(10) 克林闭包 (*Kleene closure*): $\Sigma^* = \{\varepsilon\} \cup \Sigma \cup \Sigma^2 \cup \Sigma^3 \cup \cdots$, 显然 $\Sigma^* = \Sigma^+ \cup \{\varepsilon\}$.

Σ^* 就是字母表 Σ 上的所有的串的集合. Σ^+ 就是字母表 Σ 上全部非空串的集合.

例如 $\{0, 1\}^* = \{\varepsilon, 0, 1, 00, 01, 10, 11, 000, 001, 010, 011, \cdots\}$

(11) 串的前缀 (*prefix*)、后缀 (*suffix*)、真前缀 (*proper prefix*)、真后缀 (*proper suffix*): 略

1.2.2 语言

语言 (*Languages*): 若 Σ 是字母表, 那么 $\forall L \subseteq \Sigma^*$, L 称为字母表 Σ 上的一个语言.

例如

(1) 自然语言, C 语言等

(2) The language of all strings consisting of n 0's followed by n 1's, for some $n \geq 0$(对某个 $n \geq 0$, 形如 n 个 0 后面跟着 n 个 1 的所有串, 所构成的语言):

$$\{\varepsilon, 01, 0011, 000111, \ldots\}$$

(3) The set of strings of 0's and 1's with an equal number of each:

$$\{\varepsilon, 01, 10, 0011, 0101, 1100, \ldots\}$$

(4) Σ^* 是任意字母表 Σ 上的语言

(5) $\emptyset$ 是任意字母表 Σ 上的语言, 空语言

(6) $\{\varepsilon\}$ 是任意字母表 Σ 上的语言, 仅有一个空串的语言, 但 $\emptyset \neq \{\varepsilon\}$.

语言以集合的方式来描述: $\{w \mid \text{something about } w\}$, 例如

(1) $\{w \mid w \text{ consists of an equal number of 0's and 1's}\}$

(2) $\{w \mid w \text{ 是素数的二进制表示}\}$

(3) $\{0^n 1^n \mid n \geq 1\}$

(4) $\{0^i 1^j \mid 0 \leq i \leq j\}$