字母表与串

Alphabet and String

刘 铎

liuduo@bjtu.edu.cn

字母表与串

- **字母表**（**alphabet**）是指一个有限的非空符号集Σ，Σ 中元素称为**字母**。
- Σ* 为所有由 Σ 中元素生成的有限长度序列全体，Σ* 中元素称为 Σ 上的**词**（**word**）或**串**（**string**）（在不引起混淆时，也可忽略序列各项间的“,”），即串是有限长度的符号序列。
- Σ* 中的空序列称作**空串**（**empty string**），习惯上使用 λ 或 ε 表示，用 Λ 表示集合 $\{\lambda\}$。

字母表与串

- 串 w 所含字母个数（即序列的项数）称作 w 的**长度(length)**，记作 $|w|$ 。

- 可以这样来理解：字母表是有限的符号集，而串是有限长度的符号序列。

- 假设 $w_1=s_1s_2s_3\ldots s_n$ 和 $w_2=t_1t_2t_3\ldots t_m$ 都是字母表 Σ 上的串，则 w_1 和 w_2 的**连接**定义为 $s_1s_2s_3\ldots s_nt_1t_2t_3\ldots t_m$，记作 $w_1\circ w_2$ 或w_1w_2。$\circ$称作 Σ^* 上的**连接运算**。

- 容易看出串在连接运算下形成的代数结构：
  - 假设 Σ 是一个字母表，则 $(\Sigma^*,\circ)$ 构成一个半群

字母表与串

- 例
  - 常用的字母表有：
    - $\Sigma = \{0, 1\}$，二进制字母表；
    - $\Sigma = \{a, b, \ldots, z\}$，所有小写字母的集合；
    - 所有ASCII字符的集合，或者所有可打印的ASCII字符的集合。
- 例
  - **01101是二进制字母表 $\Sigma=\{0, 1\}$ 上的一个串，长度为5**
  - **111是这个字母表上的另一个串，长度为3**
  - **空串 λ 的长度为0**

字母表与串

- 假设 w 是字母表 Σ 上的串，则可以定义 w 的 n 次幂 w^n 为
  - (1) $w^0=\lambda$,
  - (2) $w^n = w^{n-1}\circ w$，$n\geq 1$。
- 例
  - $a^3=aaa$，$(ab)^2a=ababa$
  - $\{0^n1^n \mid n\geq 1\}=\{01,0011,000111, \ldots\}$

字母表与串

- 假设 x, y, z 是字母表 Σ 上的串，且 $x = yz$。
  - 称 y 是 x 的**前缀**（**prefix**）
  - 如果 $z \notin \lambda$，则称 y 是 x 的**真前缀**（**proper prefix**）
  - 称 z 是 x 的**后缀**（**suffix**）
  - 如果 $y \notin \lambda$，则称 z 是 x 的**真后缀**（**proper suffix**）
- 假设 x, y 是字母表 Σ 上的串，且存在字母表 Σ 上的串 z, w 使得 $x = zyw$，则称 y 是 x 的**子串**（**substring**）
  - 简言之，x 的子串就是由 x 删除某个前缀后再删去某个后缀得到的结果。

字母表与串

- 例
  - 串abcde的
    - 前缀是 λ, a, ab, abc, abcd, abcde
    - 真前缀是 λ, a, ab, abc, abcd
    - 后缀是 λ, e, de, cde, bcde, abcde
    - 真后缀是 λ, e, de, cde, bcde
  - 对于任意字符串 x， x 的前缀有 $|x|+1$ 个，真前缀有 $|x|$ 个， x的后缀有 $|x|+1$ 个，真后缀有 $|x|$ 个
- 例
  - 串 abc 的所有子串是λ, a, b, c, ab, bc, abc

字母表与串

- 例

  (a) 对于任意非空字符串 x，λ 是 x 的前缀，且是真前缀；λ 是 x 的后缀，且是真后缀；λ 是 x 的子串

  (b) 对于任何字符串 x，x 是自身的前缀，但不是真前缀；x 是自身的后缀，但不是真后缀；x 是自身的子串

  (c) 对于任何字符串 x，x 的任意前缀 y 有惟一的一个后缀 z 与之对应，使得$x = yz$，反之亦然

End