

形式语言与自动机理论

正则表达式

王春宇

计算机科学与技术学院
哈尔滨工业大学

正则表达式

- 正则表达式
 - 正则表达式的递归定义
 - 正则表达式示例
- 自动机和正则表达式
- 正则表达式的代数定律

正则表达式

- 有穷自动机
 - 通过**机器**装置描述正则语言
 - 用计算机编写相应算法, 易于实现
- 正则表达式
 - 通过**表达式**描述正则语言, 代数表示方法, 使用方便
 - 应用广泛
 - `grep` 工具 (Global Regular Expression and Print)
 - Emacs / Vim 文本编辑器
 - `lex` / `flex` 词法分析器
 - 各种程序设计语言 Python / Perl / Haskell / ...

语言的运算

设 L 和 M 是两个语言, 那么

并 $L \cup M = \{w \mid w \in L \text{ 或 } w \in M\}$

连接 $L \cdot M = \{w \mid w = xy, x \in L \text{ 且 } y \in M\}$

幂

$$L^0 = \{\varepsilon\}$$
$$L^1 = L$$
$$L^n = L^{n-1} \cdot L$$

克林闭包

$$L^* = \bigcup_{i=0}^{\infty} L^i$$

例 1. 若有语言 $L = \{0, 11\}$ 和 $M = \{\varepsilon, 001\}$, 那么

$$L \cup M =$$

$$L^0 =$$

$$LM =$$

$$L^1 =$$

$$ML =$$

$$L^2 =$$

例 2. 对于空语言 \emptyset

$$\cdot \quad \emptyset^0 =$$

$$\forall n \geq 1, \quad \emptyset^n = \emptyset$$

$$\emptyset^* =$$

四则运算表达式的递归定义:

- ① 任何数都是四则运算表达式;
- ② 如果 a 和 b 是四则运算表达式, 那么

$$a + b, a - b, a \times b, a \div b \text{ 和 } (a)$$

都是四则运算表达式.

正则表达式的递归定义

定义

如果 Σ 为字母表, 则 Σ 上的正则表达式递归定义为:

- ① \emptyset 是一个正则表达式, 表示空语言;
- ② ϵ 是一个正则表达式, 表示语言 $\{\epsilon\}$;
- ③ $\forall a \in \Sigma$, a 是一个正则表达式, 表示语言 $\{a\}$;
- ④ 如果正则表达式 r 和 s 分别表示语言 R 和 S , 那么

$$r + s, rs, r^* \text{ 和 } (r)$$

都是正则表达式, 分别表示语言

$$R \cup S, R \cdot S, R^* \text{ 和 } R.$$

运算符的优先级

正则表达式中三种运算以及括号的优先级:

- ① 首先, “括号” 优先级最高;
- ② 其次, “星” 运算: r^* ;
- ③ 然后, “连接” 运算: rs , $r \cdot s$;
- ④ 最后, “加” 最低: $r + s$, $r \cup s$;

例 3.

$$\begin{aligned} 1 + 01^* &= 1 + (0(1^*)) \\ &\neq 1 + (01)^* \\ &\neq (1 + 01)^* \\ &\neq (1 + 0)1^* \end{aligned}$$

正则表达式示例

例 4.

E	$L(E)$
$\mathbf{a + b}$	$\mathbf{L(a) \cup L(b) = \{a\} \cup \{b\} = \{a, b\}}$
\mathbf{bb}	$\mathbf{L(b) \cdot L(b) = \{b\} \cdot \{b\} = \{bb\}}$
$(\mathbf{a + b})(\mathbf{a + b})$	$\{a, b\}\{a, b\} = \{aa, ab, ba, bb\}$
$(\mathbf{a + b})^*(\mathbf{a + bb})$	$\{a, b\}^*\{a, bb\} = \{a, b\}^*\{a\} \cup \{a, b\}^*\{bb\} = \{w \in \{a, b\}^* \mid w \text{ 仅以 } a \text{ 或 } bb \text{ 结尾.}\}$
$\mathbf{1 + (01)^*}$	$\{1, \varepsilon, 01, 0101, 010101, \dots\}$
$(\mathbf{0 + 1})^*\mathbf{01}(0 + 1)^*$	$\{x01y \mid x, y \in \{0, 1\}^*\}$

例 5. 给出正则表达式 $(aa)^*(bb)^*b$ 定义的语言.

$$\begin{aligned} \mathbf{L}((aa)^*(bb)^*b) &= \mathbf{L}((aa)^*) \cdot \mathbf{L}((bb)^*) \cdot \mathbf{L}(b) \\ &= (\{a\}\{a\})^*(\{b\}\{b\})^*\{b\} \\ &= \{a^2\}^*\{b^2\}^*\{b\} \\ &= \{a^{2n}b^{2m+1} \mid n \geq 0, m \geq 0\} \end{aligned}$$

例 6. Design regular expression for $L = \{w \mid w \text{ consists of 0's and 1's, and the third symbol from the right end is 1.}\}$

$$\underline{(0 + 1)^*} \underline{1} \underline{(0 + 1)} \underline{(0 + 1)}$$

例 7. Design regular expression for

$L = \{w \mid w \in \{0, 1\}^* \text{ and } w \text{ has no pair of consecutive 0's.}\}$

$$\underline{1^*(011^*)^*(0 + \varepsilon) \text{ 或 } (1 + 01)^*(0 + \varepsilon)}$$