第 3 章　正规表达式与正规语言

除了上下文无关语言，本书重点讨论的另一类语言是正规语言。第 2 章介绍过，正规文法对应的语言称为正规语言。除正规文法外，重要的正规语言计算模型还包括有限状态自动机与正规表达式，如图 3.1 所示。本章介绍正规表达式的内容，第 4 章将介绍三类有限状态自动机模型：确定有限自动机、非确定有限自动机，以及带 ε-转移的非确定有限自动机。

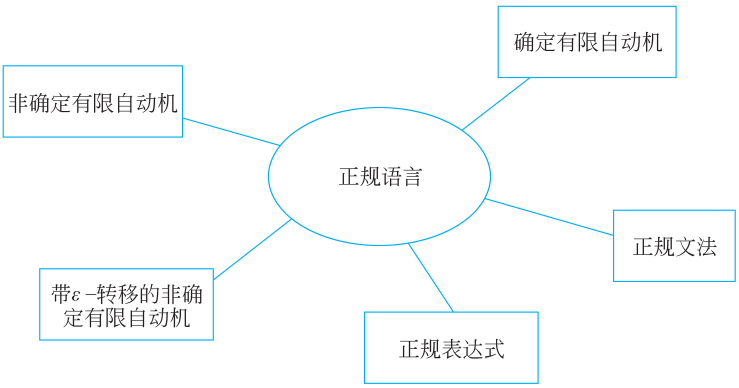

图 3.1　正规语言的不同计算模型

本章首先介绍正规表达式的基本内容，包括其语法、语义及其设计例子。语法是指它的表示形式，语义是指它所代表的语言。

接着，独立于正规表达式，引出正规语言的概念。因正规表达式的语言与正规语言的定义完全对应，本章也借助正规表达式的语言定义正规语言。进一步，图 3.1 中的正规表达式、正规文法，以及各种有限自动机模型的计算能力是相互等价的。因此，无论是正规文法和正规表达式，还是有限状态自动机，都可用于定义正规语言。

关于正规表达式与正规文法之间等价性的讨论，参见 3.4 节。

最后，我们讨论正规表达式的一些基本的代数性质；讨论代数定律的具体化，及其如何用于发现和测试有关正规表达式的定律。

3.1　正规表达式

正规表达式是用来表示正规语言（见随后的定义）的一种代数表达式。从表示形式（即语法）上看，最基本的正规表达式有三个运算符，分别对应正规语言上的三种运算：并、连接、闭包。这三种运算的定义已在第 1 章给出，为方便，现将其重新叙述如下。

设 Σ 为字母表，L 和 M 是 Σ 上的两个语言，则

- L 和 M 的并，$L \cup M = \{w \mid w \in L \vee w \in M\}$。
- L 和 M 的连接，$LM = \{w_1 w_2 \mid w_1 \in L \wedge w_2 \in M\}$。

- L 的(星)闭包(或称闭包),$L^* = L^0 \cup L^1 \cup L^2 \cup \cdots = \cup_{i \geqslant 0} L^i$,其中 $L^0 = \{\varepsilon\}$,$L^1 = L$,$L^2 = LL$,…,$L^n = L^{n-1}L$。

为方便,随后会引入一些扩展的其他助记运算符,这些助记运算符均可用三个基本运算符表示。在实际应用中,比如一类著名的词法分析器生成工具 LEX 中的正规表达式,会用到非常多的扩展运算符。

下面先给出本书中基本正规表达式的定义。

设 Σ 为字母表,则 Σ 上的**正规表达式**(regular expression)集合$\mathfrak{R}$归纳定义如下:

(1) 基础。

- $\varepsilon \in \mathfrak{R}$;
- $\phi \in \mathfrak{R}$;
- 若 $a \in \Sigma$,则 $a \in \mathfrak{R}$;
- 任一变量 $L \in \mathfrak{R}$。

(2) 归纳。

- 若 $E \in \mathfrak{R}$和 $F \in \mathfrak{R}$,则 $E+F \in \mathfrak{R}$;
- 若 $E \in \mathfrak{R}$和 $F \in \mathfrak{R}$,则 $EF \in \mathfrak{R}$;
- 若 $E \in \mathfrak{R}$,则 $E^* \in \mathfrak{R}$;
- 若 $E \in \mathfrak{R}$,则$(E) \in \mathfrak{R}$。

以上定义给出了正规表达式的表示形式,即它的语法。下面给出正规表达式的语义,即它所代表的语言。

设$\mathfrak{R}$为字母表 Σ 上的正规表达式集合。对于每个不含变量的 $E \in \mathfrak{R}$,我们用 $L(E)$表示 E 所代表的语言。$L(E)$可以归纳定义如下:

(1) 基础。

- $L(\varepsilon) = \{\varepsilon\}$,这里,第一个 ε 表示正规表达式ε,第二个 ε 表示空串;
- $L(\phi) = \varnothing$,这里,ϕ 表示正规表达式 ϕ,$\varnothing$表示空语言;
- 若 $a \in \Sigma$,则 $L(a) = \{a\}$,这里,$L(a)$中的 a 表示正规表达式 a,$\{a\}$中的 a 表示仅含一个字符的字符串 a。

(2) 归纳。

- 若 $E \in \mathfrak{R}$和 $F \in \mathfrak{R}$,则 $L(E+F) = L(E) \cup L(F)$;
- 若 $E \in \mathfrak{R}$和 $F \in \mathfrak{R}$,则 $L(EF) = L(E)L(F)$;
- 若 $E \in \mathfrak{R}$,则 $L(E^*) = (L(E))^*$;
- 若 $E \in \mathfrak{R}$,则 $L((E)) = L(E)$。

如果 $E \in \mathfrak{R}$包含变量 L,则 L 可以解释为任何正规语言。

为减少括号的数目,通常规定正规表达式中三种运算的优先级从高到低依次为:闭包、连接、并。

设 L 是正规表达式。为方便,下面引入正规表达式的几个派生运算符:

- 正闭包运算:$L^+ = LL^* = L^*L$
- 任选运算:$L? = \varepsilon + L$
- 幂运算:$L^n = LL^{n-1}(n > 0)$,$L^0 = \varepsilon$

本书中,正规表达式也简称**正规式**。

此外,在一些文献中,正规表达式被称作**正则表达式**。

3.2 正规语言

字母表 Σ 上的**正规语言**(regular language)归纳定义如下：

(1) 基础。

- $\{\varepsilon\}$是正规语言；
- $\varnothing$是正规语言，这里，$\varnothing$表示空语言；
- 若 $a\in\Sigma$，则$\{a\}$是正规语言。

(2) 归纳。

- 若 L 和 R 是正规语言，则 $L\cup R$ 也是正规语言；
- 若 L 和 R 是正规语言，则 LR 也是正规语言；
- 若 L 是正规语言，则 L^* 也是正规语言。

我们还可以借助正规表达式的概念定义正规语言。

对于字母表 Σ 上的语言 R，若存在 Σ 上的正规表达式 E(不含变量)，满足 $L(E)=R$，则 R 是**正规语言**。

由 3.1 节中关于正规表达式的语义解释，不难看出上述两种定义是等价的。

此外，3.4 节将介绍正规表达式与正规文法的等价性，因此也可以采用正规文法定义正规语言。如前所述，也可以基于有限状态自动机(参见第 4 章)给出正规语言的定义。

3.3 正规表达式的设计

正规表达式在实践中用途很广，设计正规表达式的需求十分普遍。类似上下文无关文法的设计，正规表达式的设计通常也需要从语言的字符串构成方式入手发现其生成规律，有时还需要将复杂问题分解为易解决的子问题，对于有难度的问题，也需发挥一定的想象力。

当然，有些时候遇到的语言本身不是正规语言。如在第 6 章，我们会介绍正规语言的一个必要条件，从而证明某些语言不是正规语言，不可能设计出相应的正规表达式。

例 3.1 设计表示如下语言 L 的正规表达式：

$$L=\{w\mid w\in\{0,1\}^*,\text{且 } w \text{ 由交替的 0 和 1 构成}\}$$

解 表示语言 L 的一个正规表达式为

$$(01)^*+(10)^*+0(10)^*+1(01)^*$$

我们简单分析一下这样设计的思路。由＋分开的 4 个子表达式代表的子语言分别是：$(01)^*$ 表示 0 开头 1 结尾的字符串集合，并包含空串 ε；$(10)^*$ 表示 1 开头 0 结尾的字符串集合，并包含空串 ε；$0(10)^*$ 表示 0 开头 0 结尾的字符串集合，并包含串 0；$1(01)^*$ 表示 1 开头 1 结尾的字符串集合，并包含串 1。

一个语言的正规表达式并非唯一的，不同的设计思路会得出看似不同形式的表达式。例如，例 3.1 中 L 的另外两个正规表达式为

(1) $(\varepsilon+1)(01)^*(\varepsilon+0)$

(2) $(\varepsilon+0)(10)^*(\varepsilon+1)$

读者可以想一想这两个正规表达式的设计思路。

例 3.2 设计表示如下语言的正规表达式：

$L_1=\{w \mid w\in\{0,1\}^*, |w|\geqslant 5$，且从右端数第 5 个位置是“1”$\}$，

$L_2=\{w \mid w\in\{0,1\}^*, |w|>0$，且 w 的后 5 位中至少有 1 个“1”$\}$，

$L_3=\{w \mid w\in\{0,1\}^*, |w|>0$，且 w 的前 5 位中至少有 1 个“1”$\}$。

解 语言 L_1 的一个正规表达式为

$$(0+1)^*1(0+1)(0+1)(0+1)(0+1)$$

显然，该正规表达式定义的语言中，每个字符串长度至少为 5，且倒数第 5 个位置是“1”。

语言 L_2 的情况相比 L_1 略微复杂一些。上述表达式中，$(0+1)(0+1)(0+1)(0+1)$代表长度为 4 的 0、1 串，可确保倒数第 5 个位置一定是“1”。试想，对于倒数的 4 个位置，如果可以选择每一位置上的 0、1 字符出现或不出现，即将$(0+1)$改为$(\varepsilon+0+1)$，那么这些位置所组成的字符串的长度就是 0～4。也就是说，原来倒数第 5 个位置的“1”，将可能位于倒数第 1、2、3、4 或 5 的位置，或者说后 5 位至少有一个“1”，且整个字符串的长度一定大于 0。根据这一思路，L_2 的正规表达式可设计为

$$(0+1)^*1(\varepsilon+0+1)(\varepsilon+0+1)(\varepsilon+0+1)(\varepsilon+0+1)$$

再进一步分析，后 5 位中至少有 1 个“1”，实际上相当于从最后一位往前数，首次出现“1”的位置可以位于倒数第 1、2、3、4 或 5 的位置。因此，L_2 的正规表达式也可设计为

$$(0+1)^*1(\varepsilon+0)(\varepsilon+0)(\varepsilon+0)(\varepsilon+0)$$

借鉴 L_2 正规表达式的设计思路，语言 L_3 的正规表达式可以设计为

$$(\varepsilon+0+1)(\varepsilon+0+1)(\varepsilon+0+1)(\varepsilon+0+1)1(0+1)^*$$

同样，语言 L_3 的正规表达式设计还可以有另一种解法：

$$(\varepsilon+0)(\varepsilon+0)(\varepsilon+0)(\varepsilon+0)1(0+1)^*$$

若使用幂运算表示，可以将上述结果重写为：L_1 的正规表达式$(0+1)^*1(0+1)^4$，L_2 的正规表达式$(0+1)^*1(\varepsilon+0+1)^4$ 或$(0+1)^*1(\varepsilon+0)^4$，以及 L_3 的正规表达式$(\varepsilon+0+1)^4 1(0+1)^*$ 或$(\varepsilon+0)^4 1(0+1)^*$。

3.4 正规表达式与正规文法*

3.4.1 从正规表达式到正规文法*

定理 3.1 任何正规表达式均存在与之等价的正规文法。

证明 设 Σ 为字母表，E 为 Σ 上的正规表达式。通过下列步骤，定义文法 $G=(V,\Sigma,P,S)$：

(1) 置 $V=\varnothing$；

(2) 归纳于 E 的结构，定义相应子文法的产生式集合 $P(E)$和开始符号 $S(E)$如下：

① 基础

- 若 E 为ε，选 $S'\notin V$，令 $S(E)=S'$，$P(E)=\{S'\to\varepsilon\}$，$V=V\cup\{S'\}$；
- 若 E 为$\varnothing$，选 $S'\notin V$，令 $S(E)=S'$，$P(E)=\varnothing$，$V=V\cup\{S'\}$；
- 若 E 为$a(a\in\Sigma)$，选 $S'\notin V$，令 $S(E)=S'$，$P(E)=\{S'\to a\}$，$V=V\cup\{S'\}$；

② 归纳

- 若 E 为 E_1+E_2，且有 $S(E_1)=S_1$ 以及 $S(E_2)=S_2$，选 $S'\notin V$，令

$$S(E)=S',$$
$$P(E)=\{S'\to\varepsilon\mid S_1\to\varepsilon\in P(E_1)\vee S_2\to\varepsilon\in P(E_2)\}\cup$$
$$\{S'\to\alpha\mid\alpha\neq\varepsilon\wedge(S_1\to\alpha\in P(E_1)\vee S_2\to\alpha\in P(E_2))\}\cup$$
$$\{A\to\alpha\mid A\neq S_1\wedge A\neq S_2\wedge(A\to\alpha\in P(E_1)\cup P(E_2))\},$$
$$V=(V-\{S_1,S_2\})\cup\{S'\};$$

- 若 E 为 E_1E_2，且有 $S(E_1)=S_1$ 以及 $S(E_2)=S_2$，选 $S'\notin V$，令

$$S(E)=S',$$
$$P(E)=\{S'\to\varepsilon\mid S_1\to\varepsilon\in P(E_1)\wedge S_2\to\varepsilon\in P(E_2)\}\cup$$
$$\{S'\to\alpha\mid S_1\to\varepsilon\in P(E_1)\wedge\alpha\neq\varepsilon\wedge S_2\to\alpha\in P(E_2)\}\cup$$
$$\{S'\to aA\mid a\in\Sigma\wedge S_1\to aA\in P(E_1)\}\cup$$
$$\{S'\to aS_2\mid a\in\Sigma\wedge S_1\to a\in P(E_1)\}\cup$$
$$\{A\to aB\mid A\neq S_1\wedge a\in\Sigma\wedge A\to aB\in P(E_1)\}\cup$$
$$\{A\to aS_2\mid A\neq S_1\wedge a\in\Sigma\wedge A\to a\in P(E_1)\}\cup$$
$$\{A\to\alpha\mid\alpha\neq\varepsilon\wedge(A\to\alpha\in P(E_2)\},$$
$$V=(V-\{S_1\})\cup\{S'\};$$

- 若 E 为 E_1^*，且有 $S(E_1)=S_1$，选 $S'\notin V$，令

$$S(E)=S',$$
$$P(E)=\{S'\to\varepsilon\}\cup$$
$$\{S'\to\alpha\mid\alpha\neq\varepsilon\wedge S_1\to\alpha\in P(E_1)\}\cup$$
$$\{S'\to aS_1\mid a\in\Sigma\wedge S_1\to a\in P(E_1)\}\cup$$
$$\{A\to\alpha\mid\alpha\neq\varepsilon\wedge A\to\alpha\in P(E_1)\}\cup$$
$$\{A\to aS_1\mid a\in\Sigma\wedge A\to a\in P(E_1)\}$$
$$V=V\cup\{S'\};$$

- 若 E 为 (E_1)，令 $S(E)=S(E_1)$，$P(E)=P(E_1)$；

(3) 返回 V，$P=P(E)$，以及 $S=S(E)$。

不难看出，步骤(2)中每一规则为相应正规表达式定义的子文法可以识别的字符串集合，与该正规表达式对应的字符串集合是一致的，所以有 $L(G)=L(E)$。

同时，所构造的文法 G 满足正规文法的定义。

证毕。

上述证明中，给出一个从正规表达式 E 至等价的正规文法 G 的构造算法，以下是针对该算法的几点解释：

(1) 动态维护一个非终结符的集合 V。引入新的非终结符时，需选取未在当前 V 中出现的符号。当一些非终结符失去作用时，需从当前 V 中删掉。另一种可能的解决方案是，先不考虑子文法中非终结符的重名问题，待需要进行文法合并时，再设法换名。

(2) 该算法中，新引入的非终结符总是作为新构造文法的开始符号(S')，而被删掉的非

终结符总是相关子文法的开始符号(S_1,S_2)。

(3) 对于基础正规表达式,所构造的子文法仅包含唯一的非终结符,也是开始符号,最多包含一个产生式。对于 ε 和 a,各包含一个产生式。对于 ϕ,则不含任何产生式。

(4) 对于表达式 E_1+E_2,所构造的新文法是 E_1 和 E_2 的两个子文法按照“+”的语义进行合并的,非终结符的重名问题已由构造过程排除。合并的思路是由新的开始符号承接两个子文法开始符号的全部产生式(含 ε-产生式),其余的产生式复制即可。这一轮,V 中新增一个开始符号,删掉原有 E_1 和 E_2 的两个子文法的开始符号。

(5) 对于表达式 E_1E_2,所构造的新文法是 E_1 和 E_2 的两个子文法按照“连接”的语义进行合并的,非终结符的重名问题同样已在构造过程中排除。合并的核心是将 E_1 子文法中形如 $A\to a$ 的产生式替换为 $A\to aS_2$,这里 S_2 为 E_2 子文法的开始符号。然后,再由新的开始符号 S'替代 E_1 子文法的开始符号 S_1,除 ε-产生式外,其余的产生式照搬。如果两个子文法均含 ε-产生式,即存在 $S_1\to\varepsilon$ 和 $S_2\to\varepsilon$,则加入 $S'\to\varepsilon$。

(6) 对于表达式 E_1^*,是按照“(星)闭包”语义由 E_1 子文法构造新的文法。在新文法的构造中,其核心改造是将子文法中形如 $A\to a$ 的产生式替换为 $A\to aS_1$,以体现“闭包”,这里 S_1 为 E_1 子文法的开始符号。然后,由新的符号 S'替代 S_1 成为新的开始符号,而 S_1 则变为普通非终结符。由于 E_1^* 可以识别 ε,因此新构造的文法一定包含 $S'\to\varepsilon$。

(7) 每一轮生成的文法,最多只含一个 ε-产生式:$S'\to\varepsilon$。因此,最终结果中,最多只含一个 ε-产生式:$S\to\varepsilon$。

例 3.3 试给出等价于正规表达式 $a(a+b)^*$ 的一个正规文法。

解 借鉴定理 3.1 证明中的从正规表达式 E 至正规文法 G 的构造过程。首先,置 $V=\varnothing$。然后,归纳于 $a(a+b)^*$ 的结构,构造步骤如下。

(1) 为最左边的子表达式 a,引入非终结符 A,定义产生式 $A\to a$,开始符号为 A。此时,$V=\{A\}$。

(2) 对于子表达式$(a+b)$中的 a 和 b,分别引入非终结符 B 和 C,分别定义产生式 $B\to a$ 和 $C\to b$,开始符号分别为 B 和 C。此时,$V=\{A,B,C\}$。

(3) 为子表达式$(a+b)$引入非终结符 D,则对应的子文法包含产生式 $D\to a$ 和 $D\to b$,开始符号为 D。此时,$V=\{A,D\}$。

(4) 为子表达式$(a+b)^*$引入非终结符 F,定义其子文法包含产生式 $F\to\varepsilon$,$F\to a$,$F\to b$,$F\to aD$,$F\to bD$,$D\to aD$,$D\to bD$,$D\to a$,$D\to b$,开始符号为 F。此时,$V=\{A,D,F\}$。

(5) 对于 $a(a+b)^*$,引入非终结符 S,定义子文法的产生式:$S\to a$,$S\to aF$,$F\to a$,$F\to b$,$F\to aD$,$F\to bD$,$D\to aD$,$D\to bD$,$D\to a$,$D\to b$,开始符号为 S。此时,$V=\{S,D,F\}$。

(6) 得到等价于正规表达式 $a(a+b)^*$ 的一个正规文法 $G=(\{S,D,F\},\{a,b\},P,S)$,其中产生式集合 P 为

$$
\begin{aligned}
S &\to a\mid aF\\
F &\to a\mid b\mid aD\mid bD\\
D &\to a\mid b\mid aD\mid bD
\end{aligned}
$$

3.4.2 从正规文法到正规表达式*

定理 3.2 任何正规文法均存在与之等价的正规表达式。

证明 设正规文法 $G=(V,T,P,S)$。首先介绍从文法 G 构造与之等价的正规表达式 E 的一个算法。为此，引入一种扩展的正规文法，其产生式右端可以包含正规表达式，产生式形如 $A\to r$ 和 $A\to rB$，其中 $A,B\in V$，r 为 T 上的正规表达式。对于这种扩展正规文法，其语言可以理解为从开始符号能推导出的所有正规表达式的语言的并集。借助正规表达式的语义，可以严格定义这种扩展正规文法的语言(限于篇幅，这里忽略)。

初始时，将 P 中的所有产生式通过下列规则进行合并，得到新型产生式的集合 P'：

(1) 将 P 中左边非终结符相同的所有产生式 $A\to a_1$，$A\to a_2$，…，$A\to a_n$ 合并为

$$A\to(a_1+a_2+\cdots+a_n)$$

其中，$a_i\in T$ 或 $a_i=\varepsilon$，$1\leqslant i\leqslant n$。当 $n=1$ 时，可不加括号。

(2) 将 P 中左边和右边非终结符均相同的所有产生式 $A\to a_1B$，$A\to a_2B$，…，$A\to a_nB$ 合并为

$$A\to(a_1+a_2+\cdots+a_n)B$$

其中，$a_i\in T$，$1\leqslant i\leqslant n$。当 $n=1$ 时，可不加括号。

完成上述变换后，我们用 $G'=(V,T,P',S)$ 表示这一新的扩展的正规文法。根据扩展正规文法的语言含义，显然有 $L(G)=L(G')$。证明留作练习。

接下来，针对 P' 迭代进行下列变换：

(1) 取当前 P' 中除 S 外的任一非终结符 A；若当前只有非终结符 S，则转(5)；

(2) 若当前 P' 中存在 $A\to xA$ 和 $A\to y$(x 和 y 均为正规表达式)，则在 P' 中增加产生式 $A\to x^*y$，同时去掉 P' 中的 $A\to xA$ 和 $A\to y$；

(3) 若当前 P' 中存在 $A\to xB$，或者 $A\to y$，其中 $B\neq A$，x 和 y 均为正规表达式，则将 P' 中所有形如 $A'\to x'A$(x' 为正规表达式)的产生式替换为 $A'\to x'y$ 和 $A'\to x'xB$，同时在 P' 中去掉 $A\to xB$ 和 $A\to y$；

(4) 若当前 P' 中存在 $A\to x$ 和 $A\to y$(x 和 y 均为正规表达式)，则在 P' 中增加产生式 $A\to(x+y)$，同时去掉 P' 中的 $A\to x$ 和 $A\to y$；转(1)；

(5) 若当前 P' 中同时存在 $S\to xS$ 和 $S\to y$(x 和 y 均为正规表达式)，则在 P' 中增加产生式 $S\to x^*y$，同时去掉 P' 中的 $S\to xS$ 和 $S\to y$。

完成上述变换后，若存在 $S\to x\in P'$(x 为正规表达式)，则 x 即正规文法 G 的一个等价的正规表达式 E；若不存在这样的 $S\to x\in P'$，则令 E 为 ϕ。

欲证明上述针对 P' 的变换的正确性，需要证明每一步变换都能保证变换前后的扩展正规文法是等价的，即其语言是相等的。核心变换规则为(2)～(4)。规则(2)和(4)仅关系到当前非终结符的产生式，根据上下文无关性(局部性)，证明较为直接。规则(3)关系到整个文法的产生式，证明变换前后的等价性需要用到归纳法(留作练习)。

证毕。

上述证明中，给出一个从正规文法至正规表达式的等价构造算法。该算法的核心思想是引入含正规表达式扩展的产生式，初始化后基于步骤(2)～(4)所述的三条规则，依次逐个消去除开始符号 S 外的非终结符及其相关联的扩展产生式，直至仅剩余 S 的最多两个产生式为止，最后经由步骤(5)可求得与原文法等价的一个正规表达式。

选择消去的非终结符次序不同，得到的正规表达式可能不同，但这些正规表达式一定是相互等价的。

例 3.4 试给出等价于下列正规文法 $G[S]$ 的一个正规表达式：

$$S \to a \mid aF$$
$$F \to a \mid b \mid aD \mid bD$$
$$D \to a \mid b \mid aD \mid bD$$

解 借鉴定理 3.2 证明中的从正规文法 G 至正规表达式的构造过程。首先，经过合并 $G[S]$中的 10 个产生式初始化含正规表达式扩展的产生式集合 P'如下：

$$S \to a \mid aF$$
$$F \to (a+b) \mid (a+b)D$$
$$D \to (a+b) \mid (a+b)D$$

若选择先消去 D，由规则(2)，$D \to (a+b) \mid (a+b)D$ 被替换为 $D \to (a+b)^*(a+b)$，而 $(a+b)^*(a+b)=(a+b)^+$，这样 P'变换为

$$S \to a \mid aF$$
$$F \to (a+b) \mid (a+b)D$$
$$D \to (a+b)^+$$

接着，由规则(3)，当前 P'变换为

$$S \to a \mid aF$$
$$F \to (a+b) \mid (a+b)(a+b)^+$$

然后，由规则(4)，$F \to (a+b) \mid (a+b)(a+b)^+$ 被替换为 $F \to (a+b)+(a+b)(a+b)^+$，而该产生式右边等价于$(a+b)(a+b)^*$，故 P'变换为

$$S \to a \mid aF$$
$$F \to (a+b)(a+b)^*$$

开始新一轮消去，只剩非终结符 F，经规则(3)变换后，P'变换为

$$S \to a \mid a(a+b)(a+b)^*$$

经由规则(4)，得到最终的 P'为

$$S \to a+a(a+b)(a+b)^*$$

这样便求得等价于 $G[S]$的一个正规表达式 $a+a(a+b)(a+b)^*$，为方便起见，将其等价变换为

$$a(a+b)^*$$

若交换一下消去次序，先 F 后 D。从初始的 P'，通过规则(3)消去 F 后，P'变换为

$$S \to a \mid a(a+b) \mid a(a+b)D$$
$$D \to (a+b) \mid (a+b)D$$

由规则(4)，合并 $S \to a \mid a(a+b)$并简化，得到 P'为

$$S \to a(a+b)^* \mid a(a+b)D$$
$$D \to (a+b) \mid (a+b)D$$

再消去 D。通过规则(2)以及记$(a+b)^*(a+b)$为$(a+b)^+$，P'可变换为

$$S \to a(a+b)^* \mid a(a+b)D$$
$$D \to (a+b)^+$$

再经规则(3)以及规则(4)，P'可变换为

$$S \to a(a+b)^* + a(a+b)(a+b)^+$$

所以，可求得等价于$G[S]$的另一个正规表达式$a(a+b)^*+a(a+b)(a+b)^+$，经等价变换后，该表达式同样可简化为$a(a+b)^*$。

例3.4中多次提到正规表达式的等价变换，在5.2节也涉及类似的有关正规表达式的简化问题，相关话题可进一步参考后续章节的内容。

3.5 正规表达式的代数定律

与算术表达式、集合表达式、逻辑表达式类似，正规表达式也满足一些基本的运算性质，或代数定律。本节仅列举关于正规表达式的几组基本性质或定律，如交换律和结合律、幺元和零元、分配律、幂等律，以及与闭包相关的定律，等等。然后介绍正规表达式的一个特殊性质，讨论含变量的一般正规式与具体化（或实例化）的正规式之间的关系，以及如何将代数定律的具体化用于发现和测试正规表达式的定律。

3.5.1 正规表达式的几组代数定律

1. 交换律（commutativity）和结合律（associativity）

$$L+M=M+L$$
$$(L+M)+N=L+(M+N)$$
$$(LM)N=L(MN)$$

上述第一个等式表示运算“+”满足交换律，第二个等式表示运算“+”满足结合律，第三个等式表示运算“连接”满足结合律。

这些定律的证明，均可以转换为证明其对应语言的集合运算性质。比如，$L+M=M+L$的证明，可将L和M看作统一字母表上的字符串集合，然后证明$L\cup M=M\cup L$，相当于普通集合定律的证明。本节后续的定律也是这样，这里不赘述。

2. 幺元（identities）和零元（annihilators）

$$\phi+L=L+\phi=L$$
$$\varepsilon L=L\varepsilon=L$$
$$\phi L=L\phi=\phi$$

幺元有时也称单位元，表示与其他任何元素进行二元运算后结果仍为其他元素。上述第一行的等式表示ϕ为运算“+”的幺元。$\phi+L=L$表明ϕ为“+”的**左幺元**，而$L+\phi=L$表明ϕ为“+”的**右幺元**。因为ϕ同为左幺元和右幺元，所以是“+”的**幺元**。

类似地，上述第二行的等式表示ε为“连接”运算的左幺元、右幺元和幺元。

第三行的等式，表明ϕ为“连接”运算的**左零元**、**右零元**和**零元**。零元与其他任何元素进行运算的结果仍为零元自身。

3. 分配律（distributive law）

$$L(M+N)=LM+LN$$
$$(M+N)L=ML+NL$$

“连接”运算对于“+”运算是**可分配的**。上述第一个等式表示“连接”运算对“+”运算是**左可分配的**，第二个等式指“连接”运算对“+”运算是**右可分配的**。

4. 幂等律（idempotent law）

$$L+L=L$$

运算“+”满足幂等律。

5. 与闭包运算相关的定律

$$(L^*)^* = L^*$$
$$\phi^* = \varepsilon$$
$$\varepsilon^* = \varepsilon$$
$$L^+ = LL^* = L^*L \quad (L^+ \text{的定义})$$
$$L^* = L^+ + \varepsilon$$

这几个等式刻画了两种闭包运算所满足的几条定律。这些定律的直观意义是容易理解的。由于闭包运算的语义涉及无穷个集合的“并”运算，因而相关定律的证明一般需要用到归纳法。

6. 任选运算相关的定律

$$L? = \varepsilon + L$$

该等式即任选运算“?”的定义。

3.5.2 代数定律的具体化

1. 正规表达式的具体化

将正规表达式中的每个变量用单个符号替换，得到一个**具体的**(concrete)正规表达式，称为正规表达式的**具体化**。

反之，将具体的正规表达式中的单个符号用变量表示，得到该正规表达式的一般(general)表达式，称为具体正规表达式的**一般化**(或泛化)。

例如，有一般的正规表达式 $R(M+N)$，含三个变量 R、M 和 N。我们将 R、M 和 N 分别替换为某字母表中的单个符号 a、b 和 c，得到一个具体的正规表达式 $a(b+c)$。反之，可以称 $R(M+N)$ 为正规表达式 $a(b+c)$ 的一个一般表达式。

正规表达式的一般形式代表的任何正规语言与其对应的具体表达式的语言之间可以建立特定的对应关系。以上述的一般正规表达式 $R(M+N)$ 对应具体正规表达式 $a(b+c)$ 为例。在 $R(M+N)$ 中，R、M 和 N 可代表任何正规语言，任取它们代表的一个具体语言，分别记为 $L(R)$、$L(M)$ 和 $L(N)$，则可将 $R(M+N)$ 的语言表示为 $L(R(M+N))=L(R)(L(M)\cup L(N))=L(R)L(M)\cup L(R)L(N)$。同时，$R(M+N)$ 对应的具体表达式 $a(b+c)$ 的语言为 $L(a(b+c))=\{ab,ac\}$。任取 $w\in L(R(M+N))$，一定存在 $w_1\in L(R)$，$w_2\in L(M)\cup L(N)$，满足 $w=w_1w_2$。根据上述具体化方式，R 被具体化为 a，可将 w_1 对应到 a。若 $w_2\in L(M)$，则将 w_2 对应到 b；若 $w_2\in L(N)$，则将 w_2 对应到 c。总之，w 可以对应到 ab 或 ac，二者均属于 $L(a(b+c))$。

另外，$L(a(b+c))$ 中仅包含 ab 和 ac 两个串。对于 $ab\in L(a(b+c))$，任取 $w_1\in L(R)$，$w_2\in L(M)$，即有 $w_1w_2\in L(R)L(M)$，对应有 $w_1w_2\in L(R(M+N))$。对于 $ac\in L(a(b+c))$，任取 $w_1'\in L(R)$，$w_2'\in L(N)$，即有 $w_1'w_2'\in L(R)L(N)$，对应有 $w_1'w_2'\in L(R(M+N))$。

定理 3.3 设 E 为正规表达式，$L_1,L_2,\cdots,L_m$ 为其中的变量(为简化表述，这里假设 E 中不含非变量符号[①])。将每一 L_i 替换为符号 a_i($1\leqslant i\leqslant m$)，得到对应 E 的一个具体表达式 C，则对于这些变量的任何实例语言 $S_1,S_2,\cdots,S_m$，$L(E)$ 中的任何串 w 可写成 $w=$

① 若包含非变量符号，仅需对定理的表述进行适当扩展，其证明过程也类似。

$w_1w_2\cdots w_k$ 的形式，其中 $w_l(1\leqslant l\leqslant k)$ 是某一语言 $S_{j_l}(1\leqslant j_l\leqslant m)$ 中的串。令 a_{j_l} 为替换 L_{j_l} 的符号，其中 $1\leqslant l\leqslant k$，$1\leqslant j_l\leqslant m$，则必有 $a_{j_1}a_{j_2}\cdots a_{j_k}\in L(C)$。另外，若有串 $a_{r_1}a_{r_2}\cdots a_{r_n}\in L(C)$，$a_{r_p}$ 为替换 L_{r_p} 的符号，其中 $1\leqslant p\leqslant n$，$1\leqslant r_p\leqslant m$，对于 $S_{r_p}(1\leqslant r_p\leqslant m)$ 中的任意串 $w_p(1\leqslant p\leqslant n)$，则一定有 $w_1w_2\cdots w_n\in L(E)$。

注： 默认有一个统一的字母表，包含了所涉及的所有非变量符号。

证明 归纳于正规表达式 E 的结构。

基础：若 E 为 ε、ϕ，显然有 $E=C$，定理显然成立（注：因假设了 E 中不含非变量符号，所以 E 不可能为 a）。若 E 为 L，将唯一的变量 L 替换为符号 c，则其具体表达式为 c；L 的任何一个实例语言中的串 w，对应表达式 c 的语言 $L(c)$ 中的串 c；反之，将 $L(c)$ 中唯一的串 c 替换为 L 实例语言中的任何串 w，自然是属于该语言的。

归纳：假设 $E=E_1E_2$。依题意将每一 L_i 替换为符号 $a_i(1\leqslant i\leqslant m)$，则 E 具体化为 C，E_1 和 E_2 分别具体化为 C_1 和 C_2，并且 $C=C_1C_2$。

任意取定各变量 $L_i(1\leqslant i\leqslant m)$ 的实例语言 $S_1,S_2,\cdots,S_m$。设任何 $w\in L(E)$，则存在 $w_1\in L(E_1)$ 和 $w_2\in L(E_2)$，且满足 $w=w_1w_2$。由归纳假设，w_1 可写成 $s_1s_2\cdots s_k$ 的形式，其中 $s_l(1\leqslant l\leqslant k)$ 是 L_{j_l} 实例语言 $S_{j_l}(1\leqslant j_l\leqslant m)$ 中的串，令 a_{j_l} 为替换 L_{j_l} 的符号，则有 $a_{j_1}a_{j_2}\cdots a_{j_k}\in L(C_1)$；同样，$w_2$ 可写成 $t_1t_2\cdots t_{k'}$ 的形式，其中 $t_l(1\leqslant l\leqslant k')$ 是 $L_{j'_l}$ 实例语言 $S_{j'_l}(1\leqslant j'_l\leqslant m)$ 中的串，令 $a_{j'_l}$ 为替换 $L_{j'_l}$ 的符号，则有 $a_{j'_1}a_{j'_2}\cdots a_{j'_{k'}}\in L(C_2)$。这样，$w$ 可写成 $s_1s_2\cdots s_kt_1t_2\cdots t_{k'}$ 的形式，对应有 $a_{j_1}a_{j_2}\cdots a_{j_k}a_{j'_1}a_{j'_2}\cdots a_{j'_{k'}}\in L(C)$。

另外，任取 $c\in L(C)$，则存在 $c_1\in L(C_1)$ 和 $c_2\in L(C_2)$，且满足 $c=c_1c_2$。设 $c_1=a_{r_1}a_{r_2}\cdots a_{r_n}(1\leqslant p\leqslant n,1\leqslant r_p\leqslant m)$，$c_2=a_{r'_1}a_{r'_2}\cdots a_{r'_{n'}}(1\leqslant p\leqslant n',1\leqslant r'_p\leqslant m)$。由归纳假设，对于 L_{r_p} 的任何实例语言，比如 $S_{r_p}(1\leqslant p\leqslant n,1\leqslant r_p\leqslant m)$，任取 $s_p\in S_{r_p}$，以及对于 $L_{r'_p}$ 的任何实例语言，比如 $S_{r'_p}(1\leqslant p\leqslant n',1\leqslant r'_p\leqslant m)$，任取 $t_p\in S_{r'_p}$，有 $s_1s_2\cdots s_n\in L(E_1)$ 和 $t_1t_2\cdots t_{n'}\in L(E_2)$。因此，存在与 $c\in L(C)$ 对应的 $w=s_1s_2\cdots s_nt_1t_2\cdots t_{n'}\in L(E)$。

对于 $E=E_1+E_2$ 和 $E=E_1^*$ 的情形，可以类似证明。留作练习。

证毕。

为更好地对应定理 3.3 的叙述，下面再举一个简单的例子。

例 3.5 正规表达式 S^*M 对应的一个具体表达式为 a^*b。试借助 S 和 M 的具体实例，$S=\{01,10\}$ 和 $M=L(2^*)$，对定理 3.3 的结论予以相应的描述。

解 从题目可知，变量 S 和 M 分别具体化为符号 a 和 b。任取 S 和 M 的实例：$S=\{01,10\}$，$M=L(2^*)$，则有：

任一 $w\in L(S^*M)=\{01,10\}^*L(2^*)$，可以写成 $w_1w_2\cdots w_k$ 的形式，其中 $w_i(1\leqslant i\leqslant k)$ 是 S 或 M 中的串。显然，在出现 2 之前，所有的 w_i 均属于 S；当出现 2 之后，所有这些 2 构成 M 中的一个串。比如，1001222 可分为 3 个子串 10，01 和 222，对应 $L(a^*b)$ 中的串 aab；同理，串 1001012222，可对应 $L(a^*b)$ 中的串 $aaab$。以此类推，对于 $w_1w_2\cdots w_k\in L(S^*M)$，显然有 $c_1c_2\cdots c_k\in L(a^*b)$，其中，若 $w_i(1\leqslant i\leqslant k)$ 是 S 中的串，则有 $c_i=a$，否则 $c_i=b$。

另外，任取 $c'_1c'_2\cdots c'_k\in L(a^*b)$，必然有 $c'_i(1\leqslant i\leqslant k-1)$ 均为 a，而 c'_k 为 b。任取 $w'_i(1\leqslant i\leqslant k-1)\in S=\{01,10\}$，$w'_k\in M=L(2^*)$，则一定有 $w=w'_1w'_2\cdots w'_k\in L(S^*M)=\{01,10\}^*L(2^*)$。

2. 代数定律的具体化及其应用

将代数定律中的正规表达式进行具体化，即为**代数定律的具体化**。从定理 3.3 可推论出如下关于代数定律具体化的定理。

定理 3.4 设 E、F 为正规表达式，它们具有相同的变量集；采用同样的替换方式，得到对应 E、F 的具体表达式分别为 C、D。无论将 E、F 中的变量实例化为任何正规语言，均满足

$$L(E)=L(F),\quad \text{iff } L(C)=L(D)$$

证明 设 E、F 的变量集为 $L_1,L_2,\cdots,L_m$。分两方面证明。

(⇒)假设 $L(E)=L(F)$ 成立，证明 $L(C)=L(D)$。

设 $c=c_1c_2\cdots c_k\in L(C)$，其中每个 $c_i(1\leqslant i\leqslant k)$ 均为单个符号。任取 $w\in L(E)$，满足 $w=w_1w_2\cdots w_k$，且若有 $w_i\in L_j$，则 E 具体化为 C 时是用 c_i 替换 L_j。

∵ $L(E)=L(F)$，∴ $w\in L(F)$。因而，有 $c\in L(D)$。所以，$L(C)\subseteq L(D)$。

同理可证 $L(D)\subseteq L(C)$。

因此，$L(C)=L(D)$。

(⇐)假设 $L(C)=L(D)$，证明 $L(E)=L(F)$。类似(⇒)，留作练习。

证毕。

根据定理 3.4 的结论，可以将代数定律的具体化应用于发现和测试关于正规表达式的定律。下面通过具体例子予以说明。

例 3.6 基于代数定律的具体化，验证关于正规表达式的定律 $LL^*=L^*L$。

解 将 L 替换为具体符号 a。容易证明 $aa^*=a^*a$ 成立，因二者代表的语言相等，均为由 a 组成的长度至少为 1 的字符串集合。由定理 3.4 可知，$LL^*=L^*L$ 成立。换句话说，从 $aa^*=a^*a$，可以发现定律 $LL^*=L^*L$。

例 3.7 基于代数定律的具体化，测试 $L(M+N)=LM+LN$ 是否为关于正规表达式的定律？

解 将 L、M 和 N 分别替换为具体符号 a、b 和 c。只需测试 $a(b+c)=ab+ac$ 是否成立。显然，$a(b+c)=ab+ac$ 是成立的，因为二者代表的语言均为 $\{ab,ac\}$。

因此，$L(M+N)=LM+LN$ 是关于正规表达式的定律。

例 3.8 基于代数定律的具体化，测试 $L+ML=(L+M)L$ 是否为关于正规表达式的定律？

解 同例 3.7，将 L、M 和 N 分别替换为具体符号 a、b 和 c。欲测试 $L+ML=(L+M)L$ 是否为一条定律，可以转为验证 $a+ba=(a+b)a$ 是否成立。但后者不成立，因为 aa 属于 $(a+b)a$ 代表的语言，而不属于 $a+ba$ 代表的语言。因此，$L+ML=(L+M)L$ 不成立，不是正规表达式的定律。

练　习

1. 给出语言的闭包为有限集的所有正规表达式。

2. 设计下列语言的正规表达式：

(1) $\{w\mid w\in\{a,b\}^*,w$ 的长度至少为 2，且 w 的最左字符不同于最右字符$\}$

(2) $\{w\mid w\in\{a,b\}^*,w$ 的长度可被 3 整除$\}$

(3) $\{w\mid w\in\{a,b\}^*,w$ 中 a 的个数能被 5 整除$\}$

(4) $\{a^n\mid \exists i,j.(i,j\geqslant 0\wedge n=3i+5j)\}$

(5) $\{xwx^R \mid x,w\in\{a,b\}^+\}$

(6) $\{w \mid w\in\{0,1\}^*, w$ 包含奇数个 1$\}$

(7) $\{w \mid w\in\{a,b\}^*, |w|\geqslant 2$,且 w 中至少有两个位置的字符是不同的$\}$

(8) $\{a^nb^m \mid n,m\geqslant 0$ 且 $n+m$ 为偶数$\}$

(9) $\{w \mid w\in\{a,b\}^*$,且 w 不在正规表达式 aa^* 的语言中$\}$

(10) $\{a^mb^n \mid m\leqslant 100 \wedge n\geqslant 100\}$

(11) $\{w \mid w\in\{a,b\}^*, |w|\geqslant 1$,且当 w 以 a 结尾时,它的长度为奇数$\}$

(12) $\{w \mid w\in\{0,1\}^*, w$ 至少含有 3 个 1,且倒数第 3 位为 1$\}$

(13) $\{w \mid w\in\{a,b\}^*, w$ 中出现且仅出现一次子串 $aa\}$

(14) $\{w \mid w\in\{0,1\}^*, w$ 至多含有 2 个子串 01$\}$

(15) $\{w \mid w\in\{0,1\}^*, w$ 中包含子串 01 和子串 10$\}$

(16) $\{w \mid w\in\{0,1\}^*, |w|\geqslant 1$,且当 w 以 0 结尾时,它的长度是 3 的倍数;当 w 以 1 结尾时,它的长度是 4 的倍数$\}$

3. 设计下列语言的正规表达式:

(1) $\{w \mid w\in\{a,b\}^*, |w|\geqslant 2$,且正数第 2 位的字符是 a,倒数第 2 位的字符是 $b\}$(注意:从 1 开始数)

(2) $\{w \mid w\in\{a,b\}^* \wedge \exists x,y.(x,y\in\{a,b\}^* \wedge w=xy \wedge |y|=3 \wedge y=y^R)\}$

(3) $\{w \mid w\in\{a,b\}^*, w$ 中既不包含子串 aa,也不包含子串 $bb\}$

(4) $\{w \mid w\in\{a,b,c\}^*, w$ 中 b 和 c 的总数为偶数$\}$

(5) $\{w \mid w\in\{a,b\}^*, w$ 不含子串 $aa\}$

(6) $\{w \mid w\in\{a,b\}^*, |w|\geqslant 2$,且 w 的后 5 位至少有一个子串 $aa\}$

(7) $\{w \mid w\in\{a,b,c\}^*, |w|\geqslant 1$,且 w 前 5 位中(可能不足 5 位)至少有一位不含 $c\}$

(8) $\{w \mid w\in\{a,b\}^*, |w|\geqslant 2$,且 w 的第 2～5 位至少有一个 $a\}$

(9) $\{w \mid w\in\{a,b,c\}^*$,且 w 中每处由连续 b 构成的最长子串的长度一定为奇数$\}$(如 $\varepsilon, abbb, aaa$ 是合法的串,而 $abbabbb$ 不是)

(10) $\{w \mid w\in\{a,b\}^*, w$ 中 a 和 b 的个数相同,且每个前缀中 a 和 b 的个数之差不超过 1$\}$

(11) $\{w \mid w\in\{a,b\}^*, w$ 中子串 ab 和子串 ba 出现的次数相同$\}$

(12) $\{w \mid w\in\{a,b\}^*, w$ 中不含子串 $bab\}$

(13) $\{w \mid w=a^nb^m, m,n\geqslant 0$,且 w 中既不包含子串 $aaabb$,也不包含子串 $aabbbb\}$

4. 试给出等价于正规表达式 $(ab+ba)^*$ 的一个正规文法。

5. 设正规文法 $G=(V,T,P,S)$,若将 P 中所有产生式通过下列规则进行合并,则得到新型产生式的集合 P':

(1) 将 P 中左边非终结符相同的所有产生式 $A\to a_1, A\to a_2, \cdots, A\to a_n$ 合并为

$$A \to (a_1+a_2+\cdots+a_n)$$

其中,$a_i\in T$ 或 $a_i=\varepsilon, 1\leqslant i\leqslant n$。当 $n=1$ 时,可不加括号。

(2) 将 P 中左边和右边非终结符均相同的所有产生式 $A\to a_1B, A\to a_2B, \cdots, A\to a_nB$ 合并为

$$A \to (a_1+a_2+\cdots+a_n)B$$

其中,$a_i\in T, 1\leqslant i\leqslant n$。当 $n=1$ 时,可不加括号。

完成上述变换后，我们用 $G'=(V,T,P',S)$ 表示这一新的扩展的正规文法。

(1) 这种扩展的正规文法，其语言可以理解为从开始符号能推导出的所有正规表达式的语言的并集。试为此类扩展文法的语言给出一种严格定义。

(2) 试证明 $L(G)=L(G')$。

6. 基于题 5 中的扩展正规文法 $G'=(V,T,P',S)$，进一步针对 P' 迭代进行下列变换。

若当前 P' 中存在 $A\to xB$，或者 $A\to y$，其中 $B\neq A$，x 和 y 均为正规表达式，则将 P' 中所有形如 $A'\to x'A$（x' 为正规表达式）的产生式替换为 $A'\to x'y$ 和 $A'\to x'xB$，同时在 P' 中去掉 $A\to xB$ 和 $A\to y$。

设变换后的扩展正规文法为 G''。试证明 $L(G'')=L(G')$。

7. 试给出等价于下列正规文法 $G[S]$ 的一个正规表达式。

$$S \to 0A \mid 1B$$
$$A \to 1S \mid 1$$
$$B \to 0S \mid 0$$

8. 定理 3.3 的证明中，针对 $E=E_1+E_2$ 和 $E=E_1^*$ 情形的归纳证明没有给出，试补充完成。

9. 通过证明所代表的语言相等，验证下列关于正规表达式的定律：

(1) $(LM)N=L(MN)$

(2) $L(M+N)=LM+LN$

(3) $(L^*)^*=L^*$

(4) $(\varepsilon+L)^*=L^*$

10. 基于代数定律的具体化，验证关于正规表达式的定律：

(1) $(RS)T=R(ST)$

(2) $(\varepsilon+R)^*=R^*$

11. 证明或否证下列等式是否为关于正规表达式的定律：

(1) $(RS+R)^*R=R(SR+R)^*$

(2) $(R+S)^*S=(R^*S)^*$

(3) $(R+S)^*=R^*S^*+S^*R^*$

(4) $(\varepsilon+R)^*S^*=R^*S^*$

(5) $(S^*RS)^*R^*=S^*(RSR^*)^*$

(6) $RS^*+SR^*=R^*S+S^*R$