21世纪计算机科学与技术系列教材

形式语言与自动机

XINGSHIYUYAN YU ZIDONGJI

王柏 杨娟 编著

北京邮电大学出版社
www.buptpress.com

形式语言与自动机

王 柏　杨 娟　编著

北 京 邮 电 大 学 出 版 社
·北京·

内 容 简 介

本书扼要地介绍了形式语言与自动机的基本体系，是学习理论计算机科学基础的教材和参考书。书中主要介绍了形式语言的基本概念、自动机的模型以及形式语言与自动机的等价性，包括右线性文法与有限自动机、上下文无关文法与下推自动机、图灵机以及无限制文法等。同时介绍了自动机在通信领域的某些应用。

本书不追求过多形式化讨论，强调基本概念的直观背景和主要定理证明的思路分析。书中配有较多的例题和习题，可作为工科计算机专业本科生的教材和研究人员的参考书。

图书在版编目(CIP)数据

形式语言与自动机/王柏，杨娟编著．—北京：北京邮电大学出版社，2003
ISBN 7-5635-0669-1

Ⅰ．形…　Ⅱ．①王…②杨…　Ⅲ．①形式语言—高等学校—教材②自动机理论—高等学校—教材　Ⅳ．TP301

中国版本图书馆 CIP 数据核字(2003)第 000544 号

出 版 者：北京邮电大学出版社(北京市海淀区西土城路 10 号)邮编：100876
发行部电话：(010)62282185　62283578(传真)

经　　销：各地新华书店

印　　刷：北京忠信诚胶印厂印刷

开　　本：850 mm × 1 168 mm　1/32　　印　　张：9.25

字　　数：240 千字　　印　　数：1—3 000 册

版　　次：2003 年 2 月第 1 版　2003 年 2 月第 1 次印刷

ISBN 7-5635-0669-1/TP·95　　定价：16.00 元

前　言

任何一门科学都有它自身的理论基础，计算机科学也不例外。计算机技术变化很快，专门的技术知识今天有用，但常常在几年内就变成过时的东西。我们更应该培养思考的能力、清楚而准确地表达自己的能力、解决问题的能力以及知道问题什么时候还没有解决的能力。这些能力具有持久的价值。学习理论能够拓展人们的思维，并能使人们在这些方面得到训练。

形式语言与自动机是将数学系统应用于计算的模型。本课程重点介绍了形式语言及与之相对应的自动机体系。形式语言给出了对语言的语法规则进行描述和分类的形式化方法；而自动机则描述了能够识别语言的自动装置。这种形式化描述方法及自动机的工作原理是课程的核心内容，在编译理论、人工智能、电路设计、现代密码协议及通信等领域都有着极为广泛的应用，是每个对计算机科学感兴趣的人应该熟悉的内容。

本书共分 7 章。第 1 章是预备知识，第 2 章至第 5 章讨论了 Chomsky 四类文法所产生的语言和这些语言的识别装置。在这几章中，较详尽地论述了在计算机科学中有重大意义的 3 型文法与有限自动机、2 型文法与下推自动机，并简要地介绍了 0 型文法与图灵机，形成了较完整的理论体系。第 6 章简单介绍了前几章的理论与方法在编译方面的应用。第 7 章讨论了自动机理论在通信领域的某些应用。附录中简要介绍了计算复杂性理论。为避免过分形式化，本书强调基本概念的直观背景，主要定理证明的思路分析和各部分内容的内在联系，并配有较多的例题和习题，使比较抽

象的概念、理论易于理解与接受。

本书是在北京邮电大学陈崇昕教授1988年编写的《形式语言与自动机》这一教材的基础上结合编者的教学实践扩充、修改而成，在此谨向陈教授表示衷心感谢。同时，北京邮电大学出版社为本教材的顺利出版付出了很大努力，谨此一并致以诚挚的谢意。

由于编者水平所限，书中的不足之处在所难免，请广大专家和读者给予批评指正。

编　者

2002年12月

目　　录

第 1 章　基础知识

作为阅读本书的一些基础知识，在本章内引入有关集合、图、逻辑和证明技术等方面的基本概念。熟悉这些内容的读者可略去本章，直接学习第 2 章。

1.1　集合与关系

1. 集合

当研究某一类对象时，可把这类对象的整体称为**集合**，组成一个集合的对象称为该集合的**元素**。

设 A 是一个集合，a 是集合 A 中的元素，可表示为 $a \in A$，读作 a 属于 A。如果 a 不是集合 A 的元素，则表示为 $a \notin A$，读作 a 不属于 A。

例如，26 个小写英文字母可组成一个字母集合 A，每个小写字母皆属于 A，可写为 $a \in A, b \in A, \cdots, z \in A$。所有阿拉伯数字都不属于 A，则有 $2 \notin A, 8 \notin A$ 等。

注意： 为书写方便，今后对 $a \in A, b \in A, \cdots, z \in A$，可改写为 $a, b, \cdots, z \in A$。

有限个元素 $x_1, x_2, \cdots, x_n$ 组成的集合，称为**有限集合**。无限个元素组成的集合，称为**无限集合**。例如，整数构成的集合是一个无限集合。

不含元素的集合，称为**空集**，记为 $\varnothing$。

集合的表示法，有列举法和描述法。

列举法　列举法是把集合的元素一一列举出来。例如 26 个小写英文字母组成的集合 A，可写成 $A=\{a,b,c,\cdots,z\}$；阿拉伯数字的集合 $D=\{0,1,2,\cdots,9\}$，以及集合 $C=\{a^1,a^2,a^3,\cdots\}$等。

描述法　列举法是描述出集合中元素所符合的规则。

例如，$\mathbf{N}=\{n\mid n$ 是自然数$\}$，表明是自然数集合$\mathbf{N}$。

$A=\{x\mid x\in\mathbf{Z}$且 $0\leqslant x\leqslant5\}$，其中$\mathbf{Z}$是整数，则$A=\{0,1,2,3,4,5\}$。

2. 集合之间的关系

(1) 设两个集合 A、B 包含的元素完全相同，则称集合 A 和 B 相等，表示为 $A=B$。

例如，集合 $A=\{a,b,c\}$，集合 $B=\{b,a,c\}$，则有 $A=B$。

应指出，一个集合中元素排列的顺序是无关紧要的。

有限集合 A 中不同元素的个数称为**集合的基数**，表示为 $\# A$ 或$|A|$。

例如，$B=\{a,b,c,4,8\}$，其基数 $\# B=5$。

(2) 设两个集合 A、B，当 A 的元素都是 B 的元素，则称 A 包含于 B，或称 A 是 B 的子集，表示为 $A\subseteq B$。当 $A\subseteq B$ 且 $A\neq B$，称 A 是 B 的真子集，表示为 $A\subset B$。

如果所研究的集合皆为某个集合的子集时，称该集合为**全集**，记为 E。

(3) 根据(1)和(2)，对于任意两个集合 A、B，$A=B$ 的充要条件是$A\subseteq B$ 且 $B\subseteq A$。

3. 幂集

设 A 是集合，A 的所有子集组成的集合称为A 的**幂集**，表示为 2^A 或$\rho(A)$。

例如，$A=\{a,b,c\}$，则有

$$\rho(A)=\{\varnothing,\{a\},\{b\},\{c\},\{a,b\},\{b,c\},\{a,c\},\{a,b,c\}\}$$

当 A 是有限集，$\# A = n$，则 $\rho(A)$ 的元素数为

$$C_n^0 + C_n^1 + \cdots + C_n^n = 2^n$$

应指出，空集 $\varnothing$ 是任何集合的一个子集。

4. 集合的运算

(1) 设两个集合 A、B，由 A 和 B 的所有共同元素构成的集合，称为 A 和 B 的**交集**，表示为 $A \cap B$。

例如，$A = \{a, b, c\}$，$B = \{c, d, e, f\}$，则 $A \cap B = \{c\}$。

(2) 设两个集合 A、B，所有属于 A 或属于 B 的元素组成的集合，称为 A 和 B 的**并集**，表示为 $A \cup B$。

例如，$A = \{a, b\}$，$B = \{7, 8\}$，则 $A \cup B = \{a, b, 7, 8\}$。

(3) 设两个集合 A、B，所有属于 A 而不属于 B 的一切元素组成的集合，称为 B 对 A 的**补集**，表示为 $A - B$。

例如，$A = \{a, b, c, d\}$，$B = \{c, d, e\}$，则 $A - B = \{a, b\}$，$B - A = \{e\}$。

设全集 E 和集合 A，则称 $E - A$ 是集合 A 的补集，表示为 $\overline{A}$。

(4) 设两个集合 A、B，所有序偶 (a, b) 组成的集合，称为 A、B 的**笛卡尔乘积**，表示为 $A \times B$。

$$A \times B = \{(a, b) \mid a \in A \text{ 且 } b \in B\}$$

例如，$A = \{a, b, c\}$，$B = \{0, 1\}$，则 $A \times B = \{(a, 0), (a, 1), (b, 0), (b, 1), (c, 0), (c, 1)\}$。

序偶的元素排列是有顺序的，不能任意颠倒，(a, b) 和 (b, a) 是不相同的两个序偶，因此两个序偶相等，应该是对应元素相同，例如，$(a, b) = (c, d)$，应有 $a = c$ 和 $b = d$。

对任意集合 A、B、C 有如下运算律：

(1) $A \cup A = A$，$A \cap A = A$；

(2) $A \cup B = B \cup A$，$A \cap B = B \cap A$；

(3) $(A \cup B) \cup C = A \cup (B \cup C)$，

$(A \cap B) \cap C = A \cap (B \cap C)$；

(4) $A\cup(B\cap C)=(A\cup B)\cap(A\cup C)$,

$A\cap(B\cup C)=(A\cap B)\cup(A\cap C)$;

(5) $A\cup(A\cap B)=A, A\cap(A\cup B)=A$;

(6) $A\cup\overline{A}=E, A\cap\overline{A}=\varnothing$;

(7) $\overline{A\cup B}=\overline{A}\cap\overline{B}, \overline{A\cap B}=\overline{A}\cup\overline{B}$;

(8) $E\cup A=E, E\cap A=A$;

(9) $A\cup\varnothing=A, A\cap\varnothing=\varnothing$。

5. 关系

关系的概念在数学中是常用的,诸如大于、小于、等于、包含等都属于关系。下面给出关系的形式定义。

定义 1.1.1 设 A 是一个集合,$A\times A$ 的一个子集 R,称为是集合 A 上的一个二元关系,简称**关系**。

对于 $a\in A, b\in A$,如果$(a,b)\in R$,称 a 和 b 存在关系 R,表示为 aRb;如果$(a,b)\notin R$,称 a 和 b 不存在关系 R,表示为 $a\not R b$。

例如,自然数集合$\mathbf{N}$中的大于关系,可表示为

$$>=\{(a,b)\mid a,b\in\mathbf{N}\text{且 } a>b\}$$

当有两个集合 A、B,则从 A 到 B 的关系是 $A\times B$ 的一个子集。

例 1 设 $A=\{x,y,z\}, B=\{0,1\}$

$$A\times B=\{(x,0),(x,1),(y,0),(y,1),(z,0),(z,1)\}$$

则下列子集均为从 A 到 B 的关系:

$$R_1=\{(x,0),(y,0)\}$$

$$R_2=\{(x,1),(y,1),(z,0)\}$$

$$R_3=\{(y,0)\}$$

定义 1.1.2 设集合 A,R 是 A 上的关系:

- 对每个 $a\in A$,如果有 aRa,称 R 是自反的;
- 对于 $a,b\in A$,如果有 aRb,又有 bRa,称 R 是对称的;
- 对于 $a,b\in A$,如果有 aRb 和 bRa,则必有 $a=b$,称 R 是反对称的;
- 对于 $a,b,c\in A$,如果有 aRb 和 bRc,则有 aRc,称 R 是传

递的；

- 对每个 $a \in A$，如果 $a \not R a$，称 R 是反自反的。

例如，数之间的相等关系，具有自反性、对称性和传递性，小于关系和大于关系没有自反性，但有传递性。

定义 1.1.3 设 R 是非空集合 A 上的一个关系，如果 R 有自反性、对称性和传递性，则称 R 是一个等价关系。

由等价关系 R 可以把 A 分为若干子集，每个子集称为一个等价类，同一等价类中的元素互相是等价的。

例 2 设$\mathbf{Z}$是整数集合，R 是$\mathbf{Z}$上模 3 同余关系，也是一个等价关系，即

$$R = \{(x, y) \mid x, y \in \mathbf{Z} \text{且 } x \equiv y (\bmod 3)\}$$

由于 R 是等价关系，则存在 3 个等价类为

$$[0]_R = \{\cdots, -6, -3, 0, 3, 6, \cdots\}$$

$$[1]_R = \{\cdots, -5, -2, 1, 4, 7, \cdots\}$$

$$[2]_R = \{\cdots, -4, -1, 2, 5, 8, \cdots\}$$

其中$[0]_R$，$[1]_R$，$[2]_R$ 是表示等价类的符号。

6. 逆关系

设 R 是集合 A 上的一个关系，则

$$R^{-1} = \{(y, x) \mid x, y \in A \text{ 且有} (x, y) \in R\}$$

称 R^{-1}是关系 R 的**逆关系**。

例如，小于关系的逆关系是大于关系；相等关系的逆关系仍然是相等关系。

7. 偏序关系

定义 1.1.4 设 R 是集合 A 上的一个关系，如果 R 有自反性、反对称性和传递性，则称 R 是**偏序关系**(或部分序关系)。

例 3 设集合 $C = \{2, 3, 6, 8\}$，R 是集合 C 上的整除关系，即

$$R = \{(x, y) \mid x, y \in C \text{ 且 } x \text{ 整除 } y\}$$

可得

$$R=\{(2,2),(3,3),(6,6),(8,8),(2,6),(2,8),(3,6)\}$$

例 4 设集合 $A=\{a,b\}$,幂集 $\rho(A)$上的包含关系$\subseteq$,是一个偏序关系。这里

$$\rho(A)=(\varnothing,\{a\},\{b\},\{a,b\})$$

在 $\rho(A)$上的包含关系可用图的方法表示,如图 1.1.1 所示。

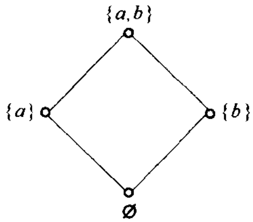

图 1.1.1 $\rho(A)$上的包含关系

描写偏序关系的图称为**哈斯图**。结合本例说明哈斯图的画法:由于 $\rho(A)$中存在$\{a\}\subseteq\{a,b\}$,所以哈斯图中有一条从节点$\{a\}$到节点$\{a,b\}$的边,这条边是自下而上的。又因$\varnothing\subseteq\{a\}$,故从节点$\varnothing$到节点$\{a\}$也存在一条自下而上的边。而对于$\varnothing\subseteq\{a,b\}$,由于以上两条边的存在,靠偏序关系的传递性,从节点$\varnothing$到节点$\{a,b\}$之间的边是不必要的。

8. 关系的闭包

定义 1.1.5 设 R 是集合 A 上的关系,如果另有关系 R' 满足:

(1) R'是传递的(自反的,对称的);

(2) $R'\supseteq R$;

(3) 对任何传递的(自反的、对称的)关系 R'',当有 $R''\supseteq R$,就有 $R''\supseteq R'$,则称关系 R'是 R 的传递(自反、对称)闭包。

R 的自反闭包表示为 $r(R)$,R 的对称闭包表示为 $s(R)$,R 的传递闭包表示为 $t(R)$。

如果给定一个集合 A 上的关系 R,可用以下方法找出传递闭包 $t(R)$,自反闭包 $r(R)$和对称闭包 $s(R)$:

(1) $r(R)=R\cup I_A$,其中 $I_A=\{(x,x)\mid x\in A\}$;

(2) $s(R)=R\cup R^{-1}$;

(3) $t(R)=R\cup R^2\cup\cdots\cup R^n$,其 $\# A=n$。

例 5 设集合 $A=\{a,b,c\}$，A 上的关系 $R=\{(a,b),(b,b),(b,c)\}$，则 R 的传递闭包为

$$t(R)=\{(a,b),(b,b),(b,c),(a,c)\},$$

而 R 的自反传递闭包表示为

$$\mathrm{tr}(R)=\{(a,a),(a,b),(b,b),(b,c),(a,c),(c,c)\}。$$

今后用 R^+ 表示 R 的传递闭包，用 R^* 表示 R 的自反传递闭包。

9. 映射

映射是关系的一个特殊类型，也称函数。

定义 1.1.6 设集合 A 和 B，f 是从 A 到 B 的一个关系，如果对每一个 $a\in A$，有惟一的 $b\in B$，使得 $(a,b)\in f$，称关系 f 是函数，记为 $f:A\to B$。

如果存在 $(a,b)\in f$，则 a 是 f 的自变量，b 是 f 作用下的像点，因此 $(a,b)\in f$ 亦可写成 $f(a)=b$。

由定义 1.1.6 可知，函数有如下特点：

(1) 函数 f 的定义域是 A，不能是 A 的某个真子集。

(2) 一个 $a\in A$ 只能对应于惟一的一个 b，或者说 $f(a)$ 是单值的。f 的值域是 B 的子集，记为 R_f。

例 6 设集合 $A=\{a,b,c\}$，$B=\{x,y\}$

$$f_1=\{(a,x),(b,x),(c,y)\}$$
$$f_2=\{(a,y),(b,y),(c,y)\}$$
$$f_3=\{(a,y),(b,x),(c,x)\}$$
$$f_4=\{(a,x),(a,y),(b,x)\}$$
$$f_5=\{(a,x)\}$$

其中，f_1、f_2 和 f_3 均为函数，f_4 和 f_5 不是函数，是关系。

函数的几种特殊类型是：

(1) 对于 $f:A\to B$。如果 f 的值域 $R_f=B$，即 B 的每一个元素都是 A 中一个或多个元素的像点，则称 f 是满射的。

例如，集合 $A=\{a,b,c,d\}$，$B=\{x,y,z\}$，如果 $f:A\to B$ 为：

$$f(a)=x, f(b)=x, f(c)=y, f(d)=z$$

则 f 是满射的。

(2) 对于 $f: A \to B$。如果 A 中没有两个元素有相同的象点，则称 f 是入射的，即对于任意 $a_1, a_2 \in A$：

如果 $a_1 \neq a_2$，则有 $f(a_1) \neq f(a_2)$，或者如果 $f(a_1)=f(a_2)$，则有 $a_1=a_2$。

例如，集合 $A=\{a,b\}$，$B=\{x,y,z\}$，如果 $f: A \to B$ 为：$f(a)=x, f(b)=y$，则称 f 是入射的。

(3) 对于 $f: A \to B$。如果 f 既是满射的，又是入射的，则称 f 是双射的，或称是一一对应的。

例如，集合 $A=\{a,b,c\}$，$B=\{1,2,3\}$，如果 $f: A \to B$ 为

$$f(a)=3, f(b)=1, f(c)=2$$

则称 f 是双射的，或者说是一一对应的。

10. 集合的划分

定义 1.1.7 设非空集合 A，$\Pi=\{\pi_1, \pi_2, \cdots, \pi_n\}$，其中 $\pi_i \subseteq A$，$\pi_i \neq \varnothing (i=1,2,\cdots,n)$，如果有 $\bigcup_{i=1}^{n} \pi_i = A$ 且 $\pi_i \cap \pi_j = \varnothing (i \neq j)$，则称 Π 是 A 的划分。其中 π_i 是一个划分块。

例如，集合 $S=\{a,b,c,d\}$，考虑下列集合：

$$A=\{\{a,b\},\{c,d\}\}$$
$$B=\{\{a\},\{b\},\{c\},\{d\}\}$$
$$C=\{\{a\},\{b,c,d\}\}$$
$$D=\{\{a,b,c,d\}\}$$
$$E=\{\{a,b\},\{b,c,d\}\}$$
$$F=\{\{a,b\},\{c\}\}$$

则 A、B、C 和 D 都是 S 的划分，E 和 F 则不是 S 的划分。

11. 集合的基数(或势)

对于有限集而言，所谓集合的基数，即为集合中不同元素的个

数。但对于无限集来说，集合的基数是什么？两个无限集的基数是否相同呢？在讨论了函数之后，可以使用一一对应（双射）来讨论集合的基数。

定义 1.1.8 设有集合 A、B，如果存在双射函数 $f: A \to B$，则说 A 和 B 有相同的基数，或者说 A 和 B 等势，记为 $A \sim B$。

显然，对于有限集合 A 和 B，称它们有相同的基数，是指它们的元素个数相同。对于无限集，可以看例 7。

例 7 设偶数集合 $N_e = \{2,4,6,\cdots\}$，定义函数 $f: \mathbf{N} \to N_e$，$\mathbf{N}$为自然数集合。如果对每个 $n \in \mathbf{N}$，有 $f(n) = 2n$，显然 f 是从$\mathbf{N}$到N_e的双射，所以存在$\mathbf{N} \sim N_e$。

例 7 说明，一个无限集，存在着它与其自身的一个真子集有相同的基数。这里 N_e 和自然数集合都是无限集。

通常，考虑一个无限集的基数时，总是看它与自然数集合能否建立一一对应。能与自然数集合建立一一对应的无限集称为**可数集**；不能与自然数集合建立一一对应的无限集称为**不可数集**。

例如：整数集合是可数集；集合$\{1,3,5,7,\cdots\}$是可数集；实数集合$\mathbf{R}$是不可数集；集合$\{x \mid x \in \mathbf{R}, 0 < x < 1\}$是不可数集，其中$\mathbf{R}$是实数。

1.2 逻　辑

1. 命题与连接词

命题是一个能判断真假的语句，一般可用一个大写英文字母表示一个命题。例如下列语句皆为命题：

P：3 是奇数

Q：铜是金属

R：1 加 4 是 2

可见，命题P和命题Q的真值均为真（T），命题R的真值为假（F）。