形式语言与自动机

陈崇昕 编著

北京邮电学院出版社

形式语言与自动机

陈崇昕　编著

北京邮电学院出版社

内 容 提 要

本书是讨论形式语言、自动机及语法分析等方面的内容，是属专业基础性教材。

书的内容，除第一章为基础知识外，分两方面：一是形式语言与自动机的基本体系，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。二是形式语言与自动机的重要应用，即语法分析，包括翻译和解析方法。

本书可作为高等学校计算机有关专业高年级学生和硕士研究生学习本课程的教材或教学参考书，也可作为计算机应用领域中广大科技工作者的参考书。

形式语言与自动机

陈崇昕 编著

责任编辑 王履锦

北京邮电学院出版社出版

新华书店北京发行所发行 各地新华书店经售

北京市怀柔黄坎印刷厂印刷

*

850×1168毫米 1/32 印张9.25 字数231.28千字

1988年11月第一版 1988年11月第一次印刷

印数：1- 3000 册

ISBN 7-5635-0021-9/TP·5 定价：2.25 元

序　言

本世紀中期，人們在研究使“自然语言”形式化的过程中，提出和发展了形式语言与自动机理论。由于它与計算机科学中的程序语言有着密切的关系，所以随着計算机科学的飞速发展，形式语言与自动机理论和方法的研究也越来越受到人們的重視，当前已成为計算机科学的理论基础。应用領域已扩展到图象处理、模式识别、自动控制以及生物工程等方面。

本书共分七章，第一章是基础知识，涉及到本书所使用的最基本的数学知识。第二、三、四、五章在介绍语言和文法分类后，重点討论右綫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，这些构成了形式語言与自动机的基本体系。第六、七章是形式语言与自动机理论的重要应用部分，主要討论编译程序的理论基础、语法分析以及有关实用方法等。

为了保持理论的系统性，給读者一个較完整的理论体系，书中討论的重点放在有实用意义的右綫性文法与有限自动机、上下文无关文法与下推自动机，而对图灵机和无限制文法等仅作扼要地介绍。由于形式语言与自动机是一門应用性較強的学科，本书以第六、七章重点联系在编译方面的应用。

本书可作为高等学校計算机有关专业本科学生和硕士研究生学习形式语言与自动机理论的教材或教学参考书，亦可作为广大計算机科技工作者的参考书。

编　者

1988年6月

目　录

第四章　上下文无关文法与下推自动机

第五章　图灵机

第六章　翻译

第七章　解析方法

第一章　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、图和逻辑等方面的基本概念。

§1.1　集合与关系

集合

当我们研究某一类对象时，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$a\in A$，$b\in A$，…，$z\in A$。所有阿拉伯数字都不属于A，则有$2\in A$，$8\in A$等。

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

有限个元素x_1，x_2，…，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。

我们把不含元素的集合，称为**空集**，记为ϕ。

集合的表示法，有列举法和描述法。

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：是描述出集合中元素所符合的规则。

例如，$N=\{n\mid n$是自然数$\}$，表明N是自然数集合。

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

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。

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

例如，$B=\{a,b,c,c,4,8\}$，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$B\subseteq A$。

幂集

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

例如，$A=\{a,b,c\}$，则有

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

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ϕ是任何集合的一个子集。

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

$$A-B=\{a,b\},\quad B-A=\{e\}。$$

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，称是A，B的**笛卡儿乘积**，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) $A\cup\overline{A}=E$, $A\cap\overline{A}=\phi$;

(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) $A\cup\phi=A$, $A\cap\phi=\phi$。

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N中的大于关系，可表示为

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

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Z是整数集合，R是Z上模3同余关系，也是一个等价关系，即

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

由于R是等价关系，则存在三个等价类为

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

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

$$[2]_R=\{\cdots,-4,-1.2,5,8,\cdots\}$$

其中$[0]_R$，$[1]_R$，$[2]_R$是表示等价类的符号。

逆关系

设R是集合A上的一个关系，则

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

则称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。

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

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

在$\rho(A)$上的包含关系可用图的方法表示，如图1.1.1所示。

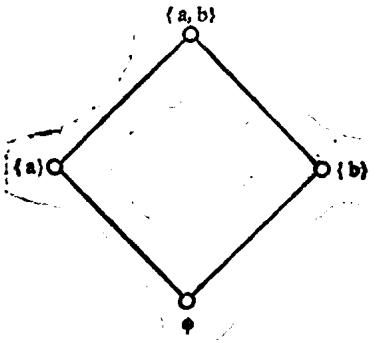

图 1.1.1

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结点$\{a\}$到结点$\{a,b\}$的边，这条边是自下而上的。又因$\phi\subseteq\{a\}$，故从结点ϕ到结点$\{a\}$也存在一条自下而上的边。而对于$\phi\subseteq\{a,b\}$，由于以上两条边的存在，靠偏序关系的传递性，所以从结点ϕ到结点$\{a,b\}$之间的边是不必要的。

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）$t(R) = R \cup R^2 \cup \cdots \cup R^n$，其中$A = n$。

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$tr(R)$

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

今后用R^+表示R的传递闭包，用R^*表示R的自反传递闭包。

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$a \in A$，有唯一的$b \in B$，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作用下的象点，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$a\in A$只能对应于唯一的一个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：

（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象点，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：

$$f(a)=x \qquad f(b)=x$$

$$f(c)=y \qquad f(d)=z$$

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。

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$\pi_i\subseteq A$，$\pi_i\neq\phi(i=1,2,\cdots,n)$，如果有$\bigcup_{i=1}^{n}\pi_i=A$且$\pi_i\cap\pi_j=\phi(i\neq j)$，则称$\Pi$是$A$的划分。其中$\pi_i$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。

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，我们可以使用一一对应（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。对于无限集可以看下面的例子。

例7 设偶数集合$Ne=\{2,4,6,\cdots\}$，定义函数$f:N\to Ne$，

N为自然数集合。如果对每个$n\in N$，有$f(n)=2n$，显然f是从N到Ne的双射，所以存在$N\sim Ne$。

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Ne和自然数集合都是无限集。

通常，我们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。我们把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R是不可数集；

集合 $\{x\mid x\in R,\ 0<x<1\}$ 是不可数集，

其中R是实数。

§1.2 逻　辑

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)。

连接词是用于把命题构成复合的命题，连接词包括“非”、“与”、“或”和“蕴含”。通常用符号“$\neg$”表示“非”，符号“$\wedge$”表示“与”、符号“$\vee$”表示“或”和符号“$\rightarrow$”表示“蕴含”。下面用真值表的方法，给出这些连接词的定义，见表1.2.1。