形式语言、自动机和语法分析

邹海明 周 新 编著

华中工学院出版社

形式语言、自动机和语法分析

华中工学院　邹海明
复旦大学　周　新
编著

华中工学院出版社

内 容 简 介

本书系论述形式语言、自动机以及语法分析等方面基础理论的著作。由于Chomsky的短语结构文法体系与各类自动机有着极为密切的关系，因此本书将它们组织在一起加以讨论。这样既可使这两者之间的关系更加清晰明晰，而且也大大缩短了篇幅。

语法分析是计算机编译程序的理论基础，书中后三章直接以形式语言与自动机理论的成果，给出了语法分析的各种实用方法。

本书可作为高等学校计算机科学技术专业、软件专业高年级学生及硕士研究生的教材或教学参考书，也可作为计算机应用领域内广大科技人员提高理论素质的参考书。

形式语言、自动机和语法分析

邹海明　周新　编著

责任编辑　　代新林

*

华中工学院出版社出版

（武昌喻家山）

新华书店湖北发行所发行

湖北省新华印刷厂排版

华中工学院出版社沔阳印刷厂印刷

*

开本：850×1168　1/32　印张：11　字数：268,000

1985年11月第一版　1985年11月第一次印刷

印　数：1—5,000

统一书号：15255—041　定价：2.70元

序　言

早在本世纪五十年代，在研究如何使“自然语言”符号化，亦即形式化的过程中，产生并发展了“形式语言与自动机”的理论。不久，人们就意识到这种理论与计算机科学中所创立和使用的程序语言具有极为密切的关系。从此以后，形式语言与自动机的理论和方法的研究，受到了越来越多科学家的重视，它的发展日新月异，时至今日，它不仅已成为计算机科学的理论基础，其应用范围已被扩展到生物工程、自动控制系统、以及图象处理与模式识别等许多领域。

随着形式语言与自动机理论应用领域的不断开拓，为了给那些对形式语言与自动机理论感兴趣的读者，以及需要应用这一技术的广大对象提供一本适当的教科书或参考书，我们特编写了本书。

本书的主要内容如下：

第一章为数学预备知识，扼要地罗列了贯穿于全书的最基本的数学工具。以便广大读者在复习本章后能够比较顺利地阅读后续章节。

第二章至第五章构成了形式语言与自动机的基本体系，即Chomsky 的短语结构文法体系。

第六至第八章是以程序的编译为背景引出形式语言与自动机理论的主要应用：翻译与句法分析。

据此，本书具有下列特点：

（1）系统性：为了适应当今各门学科相互渗透的形势，有必要给读者以完整的理论体系。因此，本书除重点引入具有实用意义的有限状态文法与有限自动机、上下文无关文法与下推自动机外，还扼要地介绍了具有理论价值的其他几类文法和图灵机。此

外，对全书的主要定理均给出了较为详细的证明。

（2）实用性：形式语言与自动机理论除了自身理论、方法的不断完善与扩充外，它又是一门应用性越来越强的学科。因此，在编写本书的过程中尽量联系实际的例子，特别是最后的三章，主要是围绕着编译理论与方法展开的。

（3）易读性：学习形式语言与自动机理论的最大障碍在于单纯的形式化之后，会使得读者其中特别是初学者感到理解上的困难与学习上的枯燥和乏味。为解决这个问题，本书除了对所有主要的定义或定理均作了直观的解释之外，还列举了大量例题，并依此反复地对主要概念予以澄清。

本书可作为高等院校计算机软件、计算机科学技术、计算机应用等有关专业的学生学习形式语言与自动机或者编译原理等课程的教科书或者教学参考书，也可作为硕士研究生的教学用书以及广大软件工作者提高基础理论的自学参考书。

美国工程院院士，PURDUE 大学的傅京荪(K.S.FU)教授审阅了本书的编写大纲，对编者给予了许多鼓励和指导，书中的某些素材也取自傅教授在PURDUE大学研究生院的讲授材料，对此编者表示衷心感谢。

编　者

一九八四年十月

目　录

第一章　预备知识

为了描述的清晰与严密，我们需要精确且适合于定义的语言，用于表示诸如自动机、语言[①]、解析，以及包含在本书中的其它一些论题。为此，我们在本章引入有关图论、逻辑代数与集合论等的一些基本概念。即使已经具备这些知识的读者，最好还是浏览一下本章的内容，以便将其中出现的记号和定义作为阅读后继章节的依据。

1·1 集合论基础

本节扼要地回顾集合论中某些最基本的概念，例如，集合、关系、函数、有序性以及集合的运算等。

1·1·1 集合

存在某个称为原子(atom)的客体。原子这个述语顾名思义是指最基础的成分。选择什么对象当作原子，这依赖于所讨论的范畴。通常，我们用整数或小写的拉丁字母来表示原子。

除原子外，还有一种抽象的成员(membership)隶属的表示方法，即若 a 是 A 中的成员，则可写成 $a \in A$；反之，则写成 $a \notin A$。假设 a 是原子，则它不再具有任何成员。

我们还使用某个不是原子的客体，称为集合(sets)。假设 A 为集合，则集合中的成员可以是原子，也可以是另外的集合。对于集合，我们假设其中的每个成员恰好出现一次。如果 A 具有有穷数目的成员 $a_1, a_2, \cdots, a_n$，且若 $i \neq j$，则 $a_i \neq a_j$，于是，称 A 为有

① 这里的语言系指形式系统的语言。

穷集(finite sets)，通常写成$A=\{a_1, a_2, \cdots, a_n\}$。注意，$A$内各成员间的先后次序是无关紧要的。例如，上面的集合也可以写成$A=\{a_n, \cdots, a_1\}$。一个特殊的集合——空集合(empty sets)是指没有任何成员的集合，常用符号ϕ表示。读者一定注意到了原子也没有任何成员。但是，应该强调：ϕ不是原子；也没有一个原子是ϕ。此外，我们使用语句$\#A=n$表示集合A具有n个成员。

例1·1 令正整数是原子，于是，$A=\{1,\{2,3\},4\}$是集合，其中，A的成员是1，$\{2,3\}$，及4。注意：作为A的成员之一的$\{2,3\}$本身也是集合，其成员是2和3；然而，原子2和3却不是A的成员。显然，我们还能够将集合A等价地写成：$A=\{4,1,\{2,3\}\}$。此外，也可将其表示为$\#A=3$。

另一种定义集合的方法是借助于谓词(predicate)。谓词是指含有一个或多个未知量的语句，且其中未知量一定处于两个量值中的一个：真(T)或假(F)。用谓词定义的集合应恰好由那些使得谓词的值为真的成分组成。因而，一定要细心的挑选用于定义集合的谓词，不然，象下面的例1·2那样，可能定义了一个不可能存在的集合。

例1·2 一个有名的现象称为"Russell 悖论"：令$P(X)$为谓词"X不是自身的成员"；即$X\notin X$。于是可以认为：能够定义集合Y，其中的成员X使得$P(X)$为真，也就是说，Y是仅仅由那些并非自身成员的量组成的集合。

假使Y存在，我们应该有可能回答这样的问题："Y是自身的成员吗？"然而，这将导致一个不可能存在的情况。假使$Y\in Y$，于是，$P(Y)$为假。由Y的定义得出Y不是自身的成员，因此，$Y\notin Y$是不可能的。反之，假设$Y\notin Y$，于是，$P(Y)$为真，由Y的定义得出$Y\in Y$。至此，我们看到了，若$Y\in Y$，则蕴含着$Y\notin Y$；若$Y\notin Y$，则蕴含着$Y\in Y$。显然，上列这些情况都是不成立的。于是，出路只有一条：集合Y不存在。

避免"Russell悖论"的常规方法是定义"X既在A内又在$P_1(X)$内"的谓词$P(X)$的集合，其中A是已知的集合，而P_1是任意的谓词。假使A是明确无误的，则谓词"X既在A内又在$P_1(X)$内"可以简化为"X在$P_1(X)$内"。假使$P(X)$是谓词，于是，我们用$\{X|P(X)\}$来表示客体为X且$P(X)$的值为真的集合。

例 1·3 令$P(X)$是谓词"X是非负的偶整数"，即$P(X)$为"X既在正整数集又在$P_1(X)$内"，其中$P_1(X)$是谓词"X是偶数"。于是$A=\{X|P(X)\}$就是通常的集合$\{0,2,4,\cdots,2n,\cdots\}$。如果非负整数的集合是已知且明确的，则集合可以直接写成：$A=\{X|X$为偶数$\}$。

定义 假使集合A中的每个成员也是集合B中的成员，我们就说集合A被包含在集合B中，写成$A\subseteq B$。

有时我们也将$A\subseteq B$说成B包含A，写成$B\supseteq A$。在上述这两种情况下，A称为是B的子集(subset)，而B称为是A的超集(Superset)。假使集合B除包含A外，还含有不在A中的其它成员，于是，我们称A真正被包含在B内，写成$A\subset B$(或者说，B真正包含着A，写成$B\supset A$)。在这种情况下，A称为是B的真子集(proper subsets)，或者说B是A的真超集(proper superset)。两个集合A和B是相等的，当且仅当$A\subseteq B$，且$B\supseteq A$。通常使用文氏图表示集合的成员及其蕴含关系。图1·1是$A\subseteq B$关系的文氏(Venn)图。

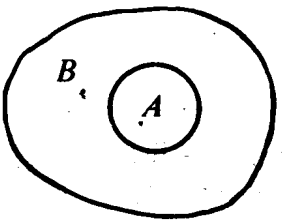


图1·1 蕴含$A\subseteq B$的文氏图

1·1·2 集合的运算

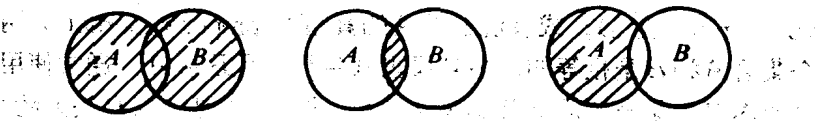
存在若干有关集合的基本运算，通过这些集合的运算能够衍生出新的集合。

定义 设A和B是集合，A和B的并(union)，写成$A\cup B$，是一个新的集合，它含有A中所有的成员加上B中所有的成员。其形式的表示是：$A\cup B=\{x|x\in A或x\in B\}$。

A和B的交(intersection)，写成$A\cap B$，是一个新的集合，它含有既在A内又在B内的所有成员。其形式的表示是：$A\cap B=\{x|x\in A且x\in B\}$。

A和B的差(difference)，写成$A-B$，是一个新的集合，它含有所有不在B中的A的成员。假使$A=U$(U含有定义域中所有的成员，故称U为通集)，于是，通常将$U-B$写成$\overline{B}$，并称为B的补(complement)。

对于上述集合运算的文氏图如图1·2所示。



(a) $A\cup B$　　(b) $A\cap B$　　(c) $A-B$

图1·2 集合运算的文氏图

假使$A\cap B=\phi$，于是，A和B称为互异(disjoint)。

定义 假使I是某些标号的集合，使得对于I内的每个i，A_i是一个已知的集合，于是，我们将$\{x|$存在$i\in I$，使得$x\in A_i\}$写成$\bigcup_{i\in I}A_i$。由于I可能是无穷的，所以上述定义可以看成是两个集

合 并的推广

也可由谓词$P(i)$定义I。这时我们将$\bigcup_{i\in I}A_i$改写成$\bigcup_{P(i)}A_i$。例如，$\bigcup_{i>2}A_i$意味着$A_3\cup A_4\cup A_5\cup\cdots$。

定义 设A为集合，A的幂集(power set)，写成$\rho(A)$或有时写成2^A，为A的所有子集的集合，即$\rho(A)=\{B\mid B\subseteq A\}$。

例1·4 令$A=\{1,2\}$，于是$\rho(A)=\{\phi,\{1\},\{2\},\{1,2\}\}$。作为幂集的另一个例子是$\rho(\phi)=\{\phi\}$。

假使A是具有m个成员的有穷集，则$\rho(A)$具有2^m个成员。注意，空集ϕ也是$\rho(A)$的成员之一。

通常我们考虑集合是无序的，然而，为了满足某种需要，定义有序对偶的集合是方便的。

定义 设a和b是一对客体，则(a,b)表示由a和b在上述顺序下组 成的有序对偶。当且仅当$a=c$和$b=d$时，式$(a,b)=(c,d)$才能成立。

定义 设A和B为集合，$A\times B$的笛卡尔乘积 (Cartesian product) 为$\{(a,b)\mid a\in A,\ 且\ b\in B\}$。

例1·5 设$A=\{1,2\}$，$B=\{2,3,4\}$，于是，$A\times B=\{(1,2),(1,3),(1,4),(2,2),(2,3),(2,4)\}$。

1·1·3 关系

许多常用的数学概念，诸如成员的隶属，集合的蕴含，以及算术运算中的小于(<)等都可作为关系(Relation)的例子。本小节将给出这个概念的形式定义。

定义 设A和B为集合，则由A到B的关系是$A\times B$的任何子集若$A=B$，我们称关系于A。若R是由A到B的关系，可写成aRb。每当(a,b)在R内，其中，$a\in A,b\in B$，则称A为R的定义域，B为R的值域。

例1·6 设A为整数集，则关系"<"的完整表示是$\{(a,b)\mid$

a小于b}。也可以简写成$a<b$。

定义 关系$\{(b,a)|(a,b)\in R\}$称为R的逆(inverse)，写成R^{-1}。

关系是一个极普通的概念。对具备某些特殊性质的关系往往有特殊的名称。

定义 假使A为集合，R为A上的关系，则称R为：

（1）自反的(reflexive)，若aRa，对所有的$a\in A$均成立；

（2）对称的(symmetric)，若aRb，则bRa，其中$a,b\in A$；

（3）传递的(transitive)，若aRb和bRc，则aRc，其中，$a,b,c\in A$，且不必是互异的。

例如，在整数集上的关系"$<$"是传递的，因为$a<b$和$b<c$，蕴含着$a<c$。然而，关系"$<$"既不是对称的也不是自反的，因为$a<b$蕴含$b<a$，以及$a<a$都是不成立的。

由于经常遇到同时具有上述性质的关系，从而把同时具备上述三条性质的关系称为等价(equivalence)关系。换句话说，等价关系必定是自反的、对称的、传递的。

等价关系有一个重要的性质，基于A上的等价关系R，可把A分割成互异的若干子集，称为等价类。对A中的某个成员a，定义a的等价类$[a]$为集合$\{b\mid aRb\}$。

例1·7 考虑在非负整数集上模N的同余关系。若存在正整数

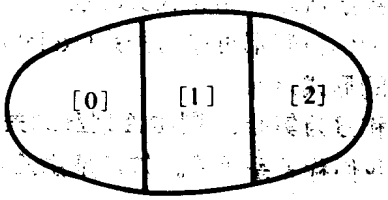


图1·3 模3同余的等价类

k,使得$a-b=kN$,就称以N为模的$a\equiv b$。作为特例取$N=3$,于是,整合$\{0,3,6,\cdots,3n,\cdots\}$组成一个等价类。我们使用记号[0]表示这个类。由于等价类中的任何成员均能用作为代表,所以上述等价类也能使用[3]或[6]或[3n]等记号。

在模3同余关系中的另两个等价类是:

$$\{1\}=\{1,4,7,\cdots,3n+1,\cdots\};$$

$$\{2\}=\{2,5,8,\cdots,3n+2,\cdots\}。$$

显然,[0],[1]与[2]这三个集合的并为所有非负整数集。因此,由模3同余等价关系,可将非负整数集分割成如图1·3所示的三个互异的等价类。

1·1·4 关系闭包

给定关系R,我们往往需要寻找另外的关系R',使得R'不仅包含R且具有某种附加的性质,如传递性等。其次,通常希望R'尽可能"小";也就是说,它是任何包含R的其它关系的子集。

定义 A上关系R的k重积(k - fold product),标记为R^k,定义如下:

(1) aR^1b,当且仅当aRb;

(2) aR^ib,当且仅当A中存在c,使得aRc和$cR^{i-1}b$,其中$i>1$。

上述定义具有递归性。为了验证,假设$i=4$,即aR^4b。应用规则(2),A中存在c_1,使得aRc_1和c_1R^3b。继续应用规则(2),A中存在c_2,使得c_1Rc_2和c_2R^2b。再次应用规则(2),A中存在c_3,使得c_2Rc_3和c_3R^1b。至此,能够应用规则(1)得到c_3Rb。

综上所述,若aR^4b,则A中存在一系列的成分c_1,c_2,及c_3,使得aRc_1,c_1Rc_2,c_2Rc_3,以至c_3Rb。

现在,可以定义集合A上关系R的传递闭包(transitive closure)R^+。也就是说,aR^+b,当且仅当aR^ib,$i\geq 1$。以后将会看到R^+是含有R的最小传递关系。

在集合A上关系R的自反传递闭包(reflexive and transitive