21世纪大学本科
计算机专业系列教材

蒋宗礼 编著

形式语言与自动机理论教学参考书

http://www.tup.com.cn

- 根据教育部高教司主持评审的《中国计算机科学与技术学科教程 2002》组织编写
- 与美国 ACM 和 IEEE/CS《Computing Curricula 2001》同步

清华大学出版社

21世纪大学本科计算机专业系列教材

形式语言与自动机理论教学参考书

蒋宗礼 编著

清华大学出版社
北京

内容简介

本书根据作者对计算机专业教育特点的理解和中国计算机学会“21世纪大学本科计算机专业系列教材”编写的总体要求，作为《形式语言与自动机理论》(主教材)一书的配套教学辅导用书，按照原书的结构编写而成。本书包括有关内容的讲解、学习要点、问题分析、求解思路和方法、注意事项、典型习题的解析等内容，并且按照小节给出知识点和主要内容解读。为读者学习和掌握原书中的知识点和问题求解方法，体会问题求解的核心思想提供帮助，对教师和学生来说，阅读这些内容都是有意义的。

图书在版编目(CIP)数据

形式语言与自动机理论教学参考书/蒋宗礼编著. —北京：清华大学出版社，2003.7
(21世纪大学本科计算机专业系列教材)
ISBN 7-302-06861-5

Ⅰ.形… Ⅱ.蒋… Ⅲ.① 形式语言—高等学校—教学参考资料 ② 自动机理论—高等学校—教学参考书 Ⅳ.TP301

中国版本图书馆CIP数据核字(2003)第053936号

出 版 者：清华大学出版社　　**地　　址**：北京清华大学学研大厦
http://www.tup.com.cn　　**邮　　编**：100084
社 总 机：010-62770175　　**客户服务**：010-62776969
责任编辑：张瑞庆
封面设计：孟繁聪
版式设计：肖　米
印 刷 者：清华大学印刷厂
发 行 者：新华书店总店北京发行所
开　　本：787×960 1/16　**印张**：16.25　**字数**：313千字
版　　次：2003年8月第1版　2003年8月第1次印刷
书　　号：ISBN 7-302-06861-5/TP·5090
印　　数：1～5000
定　　价：22.00元

序　言

PREFACE

21世纪是知识经济的时代，是人才竞争的时代。随着21世纪的到来，人类已步入信息社会，信息产业正成为全球经济的主导产业。计算机科学与技术在信息产业中占据了最重要的地位，这就对培养21世纪高素质创新型计算机专业人才提出了迫切的要求。

为了培养高素质创新型人才，必须建立高水平的教学计划和课程体系。在20多年跟踪分析ACM和IEEE计算机课程体系的基础上，紧跟计算机科学与技术的发展潮流，及时制定并修正教学计划和课程体系是尤其重要的。计算机科学与技术的发展对高水平人才的要求，需要我们从总体上优化课程结构，精炼教学内容，拓宽专业基础，加强教学实践，特别注重综合素质的培养，形成“基础课程精深，专业课程宽新”的格局。

为了适应计算机科学与技术学科发展和计算机教学计划的需要，要采取多种措施鼓励长期从事计算机教学和科技前沿研究的专家教授积极参与计算机专业教材的编著和更新，在教材中及时反映学科前沿的研究成果与发展趋势，以高水平的科研促进教材建设。同时适当引进国外先进的原版教材。

为了提高教学质量，需要不断改革教学方法与手段，倡导因材施教，强调知识的总结、梳理、推演和挖掘，通过加快教案的不断更新，使学生掌握教材中未及时反映的学科发展新动向，进一步拓广视野。教学与科研相结合是培养学生实践能力的有效途径。高水平的科研可以为教学提供最先进的高新技术平台和创造性的工作环境，使学生得以接触最先进的计算机理论、技术和环境。高水平的科研还可以为高水平人才的素质教育提供良好的物质基础。学生在课题研究中不但能了解科学研究的艰辛和科研工作者的奉献精神，而且能熏陶和培养良好的科研作风，锻炼和培养攻关能力和协作精神。

进入21世纪，我国高等教育进入了前所未有的大发展时期，时代的进步与发展对高等教育质量提出了更高、更新的要求。2001年8月，教育部颁发了《关于加强高等学校本科教学工作，提高教学质量的若干意见》。文件指出，本科教育是高等教育的主体

和基础，抓好本科教学是提高整个高等教育质量的重点和关键。随着高等教育的普及和高等学校的扩招，在校大学本科计算机专业学生的人数将大量上升，对适合21世纪大学本科计算机科学与技术学科课程体系要求的，并且适合中国学生学习的计算机专业教材的需求量也将急剧增加。为此，中国计算机学会和清华大学出版社共同规划了面向全国高等院校计算机专业本科生的**"21世纪大学本科计算机专业系列教材"**。本系列教材借鉴美国ACM和IEEE/CS最新制定的《Computing Curricula 2001》(简称CC2001)课程体系，反映当代计算机科学与技术学科水平和计算机科学技术的新发展、新技术，并且结合中国计算机教育改革成果和中国国情。

中国计算机学会教育专业委员会和全国高等学校计算机教育研究会，在清华大学出版社的大力支持下，跟踪分析CC2001，并结合中国计算机科学与技术学科的发展现状和计算机教育的改革成果，研究出了《中国计算机科学与技术学科教程2002》(China Computing Curricula 2002，简称CCC2002)，该项研究成果对中国高等学校计算机科学与技术学科教育的改革和发展具有重要的参考价值和积极的推动作用。

"21世纪大学本科计算机专业系列教材"正是借鉴美国ACM和IEEE/CS CC2001课程体系，依据CCC2002基本要求组织编写的计算机专业教材。相信通过这套教材的编写和出版，能够在内容和形式上显著地提高我国计算机专业教材的整体水平，继而提高我国大学本科计算机专业的教学质量，培养出符合时代发展要求的具有较强国际竞争力的高素质创新型计算机人才。

中国工程院院士
国防科学技术大学教授
21世纪大学本科计算机专业系列教材编委会名誉主任
2002年7月

前言

FOREWORD

计算机科学与技术学科要求学生具有形式化描述和抽象思维能力，并且能够很好地掌握逻辑思维方法。我们称其为“计算思维”能力，或者叫做“计算机思维”能力。当然，一种能力的培养决不是一、两门孤立的课程可以实现的，尤其是思维能力的培养更是如此。它需要一系列课程，并且通过长期的学习和实践来完成。

形式语言与自动机理论不仅对问题及其求解提供了良好的形式化描述工具，更在通过适当的描述和解析而降低难度之后，成为对本科生和研究生进行“计算思维”能力培养的一门重要技术基础课程。

抽象和形式化是本书所涉及内容的主要特点。在这里，既有严格的理论证明，又具有很强的构造性。包含一些基本模型、模型的建立、性质等。通过对本课程的学习，除了使学生掌握正则语言、下文无关语言的文法、识别模型及其基本性质、图灵机的基本知识以外，主要致力于培养学生的形式化描述和抽象思维能力。同时使学生了解和初步掌握“问题、形式化描述、自动化(计算机化)”的解题思路。这样，我们就扣上了“什么能被有效地自动化”这一计算学科的主题。

当我们用计算机进行问题的求解时，需要实现对“问题”所在系统的状态及其变换的描述，要用适当的数据在计算机中表示该问题，并用适当的算法通过对这些数据的变换来获得问题的求解结果。因此，对问题进行抽象和形式化表示，然后进行处理，是进行计算机问题求解的基本途径。形式语言与自动机理论给出了一类基本问题的基本描述与计算模型——抽象表示。并通过研究这些模型的性质及其变化方法来对这些问题进行研究。它们都是问题的数学模型化的典范，给计算机问题求解提供了一种优美而坚实的基础，而且，它也向人们展示了一种典型的方法和思想。另外，形式语言与自动机理论还是研究算法及其理论的基础。

形式语言与自动机理论对于一个计算机科学与技术工作者来说，是非常重要的。它已经成为国际上计算机科学与技术专业本科生的一门重要课程。ACM 和 IEEE/CS

CC2001 和《中国计算机科学与技术学科教程 2002》(简称为 CCC2002)给出了明确的要求。这里面不仅含有本学科最基本的知识内容,更涉及本学科方法论中所包含的全部三个学科形态。它可以被用来引导学生站在更高的高度去看待问题,去粗存真,直击本质,抓住问题及求解的关键点,以"计算机"的方式解决问题。

《形式语言与自动机理论》一书包括了 CC2001-CS 和 CCC2002 规定的全部相关知识单元的内容,并且完全满足 CC2001 建议的高级课程——自动机理论的教学大纲的要求。它不仅是后续课《编译原理》的理论基础,而且还广泛地用于一些新兴的研究领域。与国外现有的教材比较,《形式语言与自动机理论》一书主要突出了三个特点:(1)充分考虑国内教学计划的容量,进行内容的取舍和组织;(2)在培养读者的计算思维能力上做出进一步的尝试;(3)主要考虑国内读者的特点,并且按照国内的教学风格的要求讨论问题。本书将按照小节清晰地列举出相关知识点,给出主要内容的解读。通过这些,进一步地讨论讲解学习的要点、问题分析、求解思路和方法、注意事项等内容,为读者学习和掌握原书中的知识点和问题求解方法、体会问题求解的核心思想提供帮助。考虑到初学者在解答习题中将会遇到的主要困难,本书选择了一些典型习题,并且给出了解答及分析。

形式语言与自动机理论课程的教学,最大的问题有两个,一是内容非常抽象,这就导致阅读起来比较枯燥,而且它的作用主要是在潜移默化中体现的,难以让学生看出其"用处",似乎让人感到学习这门课程是在"自讨辛苦",而且这种"辛苦"没有太大的意义,不如学习 Java 语言等编程更容易,更实用。其二是这些内容具有较大的难度,难以找到体验感性认识的具体实例,这就导致读者难以发现相关知识点的来龙去脉,以达到深入领会之目的。要想解决这两个问题,必须掌握问题求解的思想和方法,并且通过对它们的研究,来领略这门课程在高度抽象和形式化下的优美和乐趣,使这些看似抽象枯燥的内容活起来。实际上,许多内容都可能是读者自己的体会,哪怕这些体会是不完善的,甚至是过于理想化(理性)的,与历史不十分符合的。但是,它们一定是更理性的"思路"和"想法"。实际上,这正是人们在科学研究中所努力追求的。

虽然目前在国内的计算机科学与技术学科的本科生的课程教学计划中,设置形式语言与自动机理论课程的学校还不是很普遍,甚至在一些学校的研究生的培养方案中也还未开设此课程,但是,随着我国的计算机学科教学的不断发展和条件的逐渐成熟,将会有越来越多的学校开设本课程。

本书共分 12 章。为了便于阅读,从第 1 章到第 10 章完全与原书结构相对应。第 1 章回顾在离散数学中学过的本书将要用到的一些基础知识,为后续的章节做好准备。由于主要是为了复习,所以这里给出的是知识点和应该注意的事项。除最后一节的"形式语言及其相关的基本概念"为新内容外,其他内容都可以由学生自学。第 2 章到第 8

章是教学的重点内容，主要讨论正则语言和上下文无关语言的文法、自动机描述及其性质。第 9 章对计算进行介绍，包括一般计算模型图灵机的概念、构造方法、修改，与计算相关的不可判定性、P-NP 等问题。第 10 章介绍上下文有关语言。在这些章节中，以知识点、主要内容解读、典型习题解析的形式，对讨论的内容进行归纳和解析。在第 11 章中安排了教学设计，从总体上讨论本课程的讲授等问题。第 12 章对本书的内容按类进行全面总结。

由于作者水平有限，书中的错误和不当之处在所难免，敬请读者批评指正。

作者

2003 年 6 月

本书责任编委：宋方敏

目　录

CONTENTS

第 1 章 绪 论

计算机科学与技术学科**系统地研究信息描述和变换算法，主要包括信息描述和变换算法的理论、分析、效率、实现和应用**。学科的根本问题是：**什么能被（有效地）自动化**。经过多年的发展，计算机科学与技术学科已经发展成为计算学科（computing discipline），所以，该学科既研究计算领域中的一些普遍规律，描述计算的基本概念与模型，又研究包括计算机硬件、软件（系统软件和应用软件）在内的计算系统设计与实现的工程技术。理论和实践在该学科占有重要地位，其中的理论扮演着重要基础的角色。这可以从计算学科（计算机科学与技术学科）方法论中找到依据。

建立物理符号系统并对其实施变换是计算机科学与技术学科进行问题的描述和求解的重要手段。“可行性”所要求的“形式化”及其“离散特征”使得数学成为重要工具。尤其是离散数学和计算模型无论从方法还是从工具等方面，更表现出它在计算学科中的直接应用。

虽然形式语言与自动机理论的论述只是用到集合、关系、图等基本概念，但是却不需要对这些基本概念进行过多的解释。因此，从知识的联系的角度来看，集合论和图论不一定要作为本课程的先修课。但是，从理解和掌握本课程的内容来讲，应该是在学习过集合论和图论，具有一定的知识基础和思维能力基础后，再开始本书内容的学习才是比较有利的。考虑到集合论和图论通常都被划入离散数学，所以，在本科生的教学计划中，形式语言与自动机理论被作为离散数学的后续课程。而如果是在研究生阶段学习形式语言与自动机理论，通常也就假定学生具有离散数学的基本知识。为了平稳地过渡，本章首先简要回顾在离散数学中学过的部分基本概念和方法，包括：集合及其表示；集合之间的关系；集合的运算；无穷集合；二元关系及其性质；等价关系与等价类；关系的合成；关系的闭包；无向图、有向图；树。这一部分内容分布在 1.1 节到 1.3 节。建议快速浏览这 3 节内容，以熟悉相应的表达方式。如果要介绍这一部分的内容，需要另外

增加 4～6 个学时。

第二部分是关于形式语言及其相关的基本概念，包括字母表、字母及其特性、句子、出现、句子的长度、空语句、句子的前、后缀、语言及其运算。这一部分是本章的重点，属于本课程的正式内容。这一部分的内容需要 2 个学时。

本章后面列举了大量的习题，主要用于使学生对所要求的内容进一步巩固和复习。在这些习题中，希望读者能够完成一些构造性题和证明题。关于语言的题目，应该尽可能多地完成。因为它们都涉及到最基本的训练。

1.1 集合的基础知识

无论是朴素**集合论**(set theory)，还是公理化集合论，都是整个数学的基础。计算机科学与技术领域中的大多数基本概念和理论都采用与集合论有关的术语来描述。

1.1.1 集合及其表示

1. 知识点

(1) 集合：一定范围内的、确定的、并且彼此可以区分的对象汇集在一起形成的整体称为**集合**(set)，简称为集(set)。

(2) 元素：集合的成员为该集合的元素(element)。

(3) a 是集合 A 的一个元素：如果 a 是集合 A 的一个元素，则记为 $a\in A$，且称 a 属于 A，或者 A **含有** a；否则记为 $a\notin A$，且称 a 不属于 A，或者 A 不含 a。$a\in A$ 读作 a 属于 A；$a\notin A$ 读作 a 不属于 A。

(4) 集合描述形式

① **列举法**(listing)：将所有的元素逐一地列举在大括号{ }中，在能使读者立即看出规律时，某些元素可用省略号表示。

② **命题法**(proposition)：其基本形式为$\{x\mid P(x)\}$，其中 P 为谓词，表示此集合包括所有使 P 为真的 x。

(5) 多重集合：一个元素可以在同一个集合里重复出现。

(6) 基数：如果集合 A，B 之间有一个一一对应，则称它们具有相同的**基数**(cardinality)。集合 A 的基数又叫做集合 A 的**势**，一般用$|A|$表示。对有穷集来说，它的基数就是它所包含的元素的个数。

(7) 集合的分类

① 由有限个元素构成的集合叫做**有限集**(finite set)，又称为**有穷集**。由无穷多个

元素组成的集合叫做**无穷集**(infinite set)；

② 如果$|A|=0$,则称 A 为**空集**(null set),一般用 $\varnothing$ 表示。

③ 无穷集可以分成**可数集**(countable infinite set 或 countable set)和**不可数集**(uncountable set)。与自然数集对等的集合称为可数集。

(8) 整数集、有理数集是可数的,实数集是不可数的。实数集的不可数性质可以用著名的**对角线法**(diagonalization)进行证明。

2. 注意事项

本节为回忆集合及其表示的基本内容,不用进一步扩展,而且在回忆中可随时以实际例子加以说明。并注意指出以下表示集合及其元素的习惯。

用大写的英文字母 $A,B,C,\cdots$和大写的希腊字母 $\Gamma,\Sigma,\Phi,\cdots$表示集合,用小写字母 $a,b,c,d,\cdots$表示集合的元素。

N——表示全体自然数集合

Q——表示全体有理数集合

R——表示全体实数集合

Σ——表示字母的集合

1.1.2 集合之间的关系

1. 知识点

(1) P_1是 P_2的充要条件记为 $P_1\Leftrightarrow P_2$,或者 P_1 iff P_2。

(2) 全称量词和存在量词

"$\forall x\cdots$"表示"对(论域中)所有的 $x\cdots$","$\exists x\cdots$"表示"(论域中)存在一个 $x\cdots$"。

(3) 子集

如果集合 A 中的每个元素都是集合 B 的元素,则称集合 A 是集合 B 的**子集**(subset),集合 B 是集合 A 的**包集**(container)。记作 $A\subseteq B$。也可记作 $B\supseteq A$。$A\subseteq B$ 读作集合 A 包含在集合 B 中;$B\supseteq A$ 读作集合 B 包含集合 A。

如果 $A\subseteq B$,且 $\exists x\in B$,但 $x\notin A$,则称 A 是 B 的**真子集**(proper subset),记作 $A\subset B$。

(4) 集合相等

如果集合 A,B 含有的元素完全相同,则称集合 A 与集合 B **相等**(equivalence),记作 $A=B$。