



21 世纪大学本科
计算机专业系列教材

蒋宗礼 姜守旭 编著

形式语言与自动机理论

<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

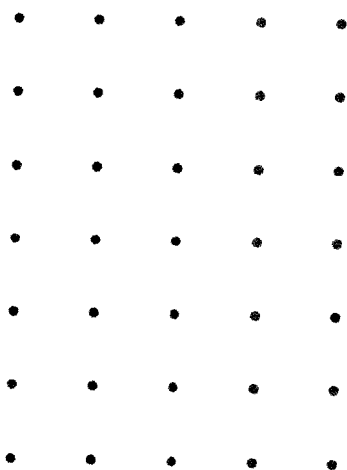


清华大学出版社

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

蒋宗礼 姜守旭 编著



清华大学出版社

内 容 简 介

形式语言与自动机理论是计算机科学与技术专业本科生的一门重要课程。本书是作者结合其10余年来在大学讲授该门课程的经验 and 体会,选择和组织有关内容撰写而成。本书不仅含有有关正则语言、上下文无关语言的文法、识别模型及其性质、图灵机的基本知识,更涉及到本学科方法论中所包含的三个学科形态。其内容特点是抽象和形式化,既有严格的理论证明,又具有很强的构造性,从而培养学生的形式化描述和抽象思维能力,使学生了解和初步掌握“问题、形式化、自动化(计算机化)”的解题思路。为了便于学生对内容的掌握,附录A还给出了建议的教学设计。

本书适合于计算机科学与技术专业的高年级本科生、研究生使用,也可供相关专业的学生、教师和科研人员参考。

版权所有,翻印必究。

本书封面贴有清华大学出版社激光防伪标签,无标签者不得销售。

图书在版编目(CIP)数据

形式语言与自动机理论/蒋宗礼,姜守旭编著. —北京:清华大学出版社,2003

21世纪大学本科计算机专业系列教材

ISBN 7-302-06103-3

I. 形… II. ①蒋… ②姜… 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(2002)第092839号

出版者: 清华大学出版社(北京清华大学学研大厦,邮编100084)

<http://www.tup.com.cn>

责任编辑: 马珂

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:** 22.75 **字数:** 439千字

版次: 2003年1月第1版 2003年1月第1次印刷

书号: ISBN 7-302-06103-3/TP·3647

印数: 0000~5000

定价: 28.00元

21 世纪大学本科计算机专业系列教材编委会

名誉主任：陈火旺

主任：李晓明

副主任：钱德沛 焦金生

委员：(按姓氏笔画为序)

马殿富 王志英 王晓东 宁洪 刘辰

孙茂松 李大友 李仲麟 吴朝晖 何炎祥

宋方敏 张大方 张长海 周兴社 侯文永

袁开榜 钱乐秋 黄国兴 蒋宗礼 曾明

廖明宏 樊孝忠

秘书：张瑞庆

本书责任编辑：宋方敏

序 言

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。

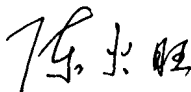
为了提高教学质量,需要不断改革教学方法与手段,倡导因材施教,强调知识的总结、梳理、推演和挖掘,通过加快教案的不断更新,使学生掌握教材中未及时反映的学科发展新动向,进一步拓广视野。教学与科研相结合是培养学生实践能力的有效途径。高水平的科研可以为教学提供最先进的高新技术平台和创造性的工作环境,使学生得以接触最先进的计算机理论、技术和环境。高水平的科研还可以为高水平人才的素质教育提供良好的物质基础。学生在课题研究中不但能了解科学研究的艰辛和科研工作者的奉献精神,而且能熏陶和培养良好的科研作风,锻炼和培养攻关能力和协作精神。

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

当我们用计算机进行问题的求解时,首先需要用适当的数据进行问题表示,然后再用适当的算法通过对这些数据进行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。这些模型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,也向人们展示了一种典型的方法和思想。另外,它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计算机科学与技术工作者是非常重要的,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。CC2001-CS 和 CCC2002 给出了明确的要求,里面不仅含有本学科最基本的知识内容,更涉及到本学科方法论中所包含的三个学科形态。它们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,从关键点上以“计算机”的方式解决问题。难怪作者在 1989 年到美国进修时被首先问到的两个问题之一就是“是否学过形式语言与自动机理论?”(另一个问题是“是否学习过算法设计与分析?”)。据统计,在每年 GRE 的考题中,大约有 8~15 道题是关于本课程内容的。

本书包括了 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,本书主要突出如下特点:(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进一步的尝试。(3)尽量照顾国内读者的特点,并且按照国内的教学风格讨论问题。

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,掌握逻辑思维方法。我们称这种能力为“计算思维”能力,或者叫“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,它需要一系列的课程,并且通过长期的修养来完成。本课程是这个系列课程中的一门,关于这个系列课程的具体讨论我们将放到 1.4 节进行。本书内容的主要特点是抽象和形式化,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、模型的建立与性质等。通过对本课程的学习,除了使学生掌握有关正则语言、上下文无关语言的文法、识别模型及其基本性质、图灵机的基本知识外,更重要的是还能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“什么能被有效地自动化”这一计算学科的主题。

哈尔滨工业大学从 1977 级本科生开始,一直坚持在本科教学计划中设置此课程。为了给没有学过此课程的研究生提供机会,还从 1982 级工学硕士研究生开始,在其计算机科学与技术学科的硕士研究生的培养方案中安排了此门课程。与其他课程相配合,在对学生进行计算思维能力的培养上,取得了良好的效果。本书是作者根据其在该校进行 10 余年的形式语言与自动机理论课程教学的教案,并参考有关教材撰写而成的。促使作者将教案变成教材的另一个原因是,在国内的教材市场上,这类教材少之又少,根本无法与它在计算机学科的人才培养中的地位相匹配。另外,我们也希望将自己积累的经验和体会提供给大家参考。在本书中,我们希望通过对一些思想和方法的介绍,使读者在这门课程中享受其高度抽象和形式化所带来的美和乐趣。希望通过这些努力,能使这些看似抽象枯燥的内容活起来。许多都是我们自己的体会,其中也难免存在不完善的地方。为了帮助读者更好地学习,附录 A 提供了包括内容取舍、讲授要点等在内的教学设计。在每章的后面,我们都附有一定量的习题。这些习题用来深化对课程知识的理解,并为读者提供应用所学知识解决问题的机会,使读者亲身体验用相关方法和思想进行探索的乐趣。特别难的习题我们没有列出来,请感兴趣的读者查阅本书后面给出的参考文献。

虽然目前国内计算机科学与技术学科本科生的课程计划中,除了一些重点院校外,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难以见到此课程。但是,我们相信,随着我国计算机学科教学的不断发展,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 10 章。第 1 章绪论,带领读者回顾在离散数学中学过的本书将要用到的一些基础知识,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;另外,介绍形式语言及其相关的基本概念,为后续的章节做准备。第 2 章介绍文法,包括文法的直观意义与形式定义,推导,归约,文法产生的语言、句子、句型,文法的构造,乔姆斯基体系,左线性文法,右线性文法,空语句。第 3 章讨论有穷状态自动机,包括 DFA 作为对

实际问题的抽象,直观物理模型,形式定义,DFA 接受的句子、语言,状态转移图,构造方法,NFA 与 DFA 的等价性,带空移动的 NFA 与 NFA 的等价性,正则文法与 FA 的等价性及其相互转换方法,基本问题的判定。第 4 章研究正则表达式,包括正则表达式的定义及其与 FA 的等价性证明。第 5 章讨论正则语言的性质,包括正则语言的泵引理的证明及其应用,正则语言的封闭性,Myhill-Nerode 定理与 FA 的极小化,判定算法。第 6 章讲述上下文无关语言,包括文法二义性,派生与派生树,上下文无关文法的化简,乔姆斯基范式,格雷巴赫范式。第 7 章叙述下推自动机,包括下推自动机的基本定义,下推自动机用终态接受的语言和用空栈接受的语言,构造方法,下推自动机与上下文无关文法的等价性。第 8 章研究上下文无关语言的性质,包括上下文无关语言的泵引理、Ogden 引理及其应用,上下文无关语言的封闭性,判定算法。第 9 章介绍图灵机,包括图灵机作为一个计算模型的基本定义,图灵机接受的语言,构造技术,通用图灵机,丘奇-图灵论题,图灵机的变形,可计算语言,不可判定性,P-NP 问题。第 10 章介绍上下文有关语言,包括图灵机与短语结构文法的等价性,线性有界自动机的定义及其与上下文有关语言的关系。

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

2002 年 12 月

目 录

第 1 章 绪论	1
1.1 集合的基础知识	2
1.1.1 集合及其表示.....	2
1.1.2 集合之间的关系.....	5
1.1.3 集合的运算.....	6
1.2 关系.....	12
1.2.1 二元关系	12
1.2.2 等价关系与等价类	13
1.2.3 关系的合成	14
1.2.4 递归定义与归纳证明	15
1.2.5 关系的闭包	18
1.3 图.....	19
1.3.1 无向图	19
1.3.2 有向图	21
1.3.3 树	23
1.4 语言.....	24
1.4.1 什么是语言	24
1.4.2 形式语言与自动机理论的产生与作用	25
1.4.3 基本概念	28
1.5 小结.....	35
习题	35
第 2 章 文法	42
2.1 启示.....	43
2.2 形式定义.....	48

2.3	文法的构造	58
2.4	文法的乔姆斯基体系	68
2.5	空语句	79
2.6	小结	82
	习题	82
第3章	有穷状态自动机	86
3.1	语言的识别	86
3.2	有穷状态自动机	89
3.3	不确定的有穷状态自动机	102
3.3.1	作为对 DFA 的修改	102
3.3.2	不确定的有穷状态自动机的形式定义	104
3.3.3	NFA 与 DFA 等价	106
3.4	带空移动的有穷状态自动机	110
3.5	FA 是正则语言的识别器	115
3.5.1	FA 与右线性文法	115
3.5.2	FA 与左线性文法	120
3.6	FA 的一些变形	122
3.6.1	双向有穷状态自动机	122
3.6.2	带输出的 FA	123
3.7	小结	125
	习题	126
第4章	正则表达式	131
4.1	启示	131
4.2	正则表达式的形式定义	133
4.3	正则表达式与 FA 等价	135
4.3.1	正则表达式到 FA 的等价变换	135
4.3.2	正则语言可以用正则表达式表示	144
4.4	正则语言等价模型的总结	150
4.5	小结	152
	习题	153
第5章	正则语言的性质	156
5.1	正则语言的泵引理	156

5.2	正则语言的封闭性	162
5.3	Myhill-Nerode 定理与 DFA 的极小化	170
5.3.1	Myhill-Nerode 定理	170
5.3.2	DFA 的极小化	180
5.4	关于正则语言的判定算法	189
5.5	小结	190
	习题	191
第 6 章	上下文无关语言	194
6.1	上下文无关语言	195
6.1.1	上下文无关文法的派生树	195
6.1.2	二义性	202
6.1.3	自顶向下的分析和自底向上的分析	205
6.2	上下文无关文法的化简	207
6.2.1	去无用符号	208
6.2.2	去 ϵ -产生式	212
6.2.3	去单一产生式组	216
6.3	乔姆斯基范式	219
6.4	格雷巴赫范式	223
6.5	自嵌套文法	229
6.6	小结	230
	习题	231
第 7 章	下推自动机	235
7.1	基本定义	235
7.2	PDA 与 CFG 等价	242
7.2.1	PDA 用空栈接受和用终止状态接受等价	243
7.2.2	PDA 与 CFG 等价	246
7.3	小结	257
	习题	257
第 8 章	上下文无关语言的性质	260
8.1	上下文无关语言的泵引理	260
8.2	上下文无关语言的封闭性	267
8.3	CFL 的判定算法	273

8.3.1	L 空否的判定	273
8.3.2	L 是否有穷的判定	274
8.3.3	x 是否为 L 的句子的判定	276
8.4	小结	278
	习题	278
第 9 章	图灵机	280
9.1	基本概念	281
9.1.1	基本图灵机	282
9.1.2	图灵机作为非负整函数的计算模型	289
9.1.3	图灵机的构造	293
9.2	图灵机的变形	300
9.2.1	双向无穷带图灵机	300
9.2.2	多带图灵机	304
9.2.3	不确定的图灵机	306
9.2.4	多维图灵机	308
9.2.5	其他图灵机	310
9.3	通用图灵机	313
9.4	几个相关的概念	315
9.4.1	可计算性	315
9.4.2	P 与 NP 相关问题	316
9.5	小结	316
	习题	317
第 10 章	上下文有关语言	320
10.1	图灵机与短语结构文法的等价性	320
10.2	线性有界自动机及其与上下文有关文法的等价性	323
10.3	小结	325
	习题	325
附录 A	教学设计	326
附录 B	缩写符号	337
	词汇索引	339
	参考文献	347

第 1 章

绪 论

计算机科学与技术学科是以数学和电子学科为基础发展起来的,该学科主要包含两方面的内容:一方面是研究计算机领域中的一些普遍规律,描述计算的基本概念与模型,其重点是描述现象,解释规律。另一方面是包括计算机硬件、软件(系统软件和应用软件)在内的计算系统设计和实现的工程技术。因此,我们称这个学科为计算机科学与技术学科。在科学的“发现自然规律”、“实验”和“计算”三个范型中,计算机科学与技术学科的研究与实践主要涉及实验范型和计算范型,这也表明,有些研究可以是以理论研究为主,有些研究可以是以实践为主。所以,对计算机科学与技术学科来说,理论和实践教学都是非常重要的,而其中的理论是基础。按照计算机科学与技术学科方法论的“抽象”、“理论”、“设计”三个过程来看,实际工作通过理论得到升华,而在理论指导下的设计(实现)才可能是理性的、高水平的。实际系统设计的突破往往等待理论上的突破——新理论的诞生。

简单地说,计算机科学与技术学科通过在计算机上建立模型并模拟物理过程来进行科学调查和研究,它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。

随着计算机科学与技术学科的发展,以及人们对该学科的认识的不断深化,我们认为,该学科的根本问题是:什么能被(有效地)自动化。

问题的计算机求解建立在高度抽象的基础上,问题的符号表示及处理过程的机械化、严格化等固有特性决定了数学是计算机科学与技术学科的重要基础之一,数学及其形式化描述以及严密的表达和计算,是计算机科学与技术学科所用的重要工具。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描述和求解的重要手段。学科所要求的计算机问题求解的“可行性”限定了从问题抽象开始到根据适当理论的指导进行设计(实现)的科学实践过程,“可行性”所需要的“形式化”后呈现出的“离散

特征”，实质上限制了计算机科学与技术学科进行问题求解的重要特征。

本章简要回顾离散数学中学过的部分基本概念和方法，以使读者能够顺利地进行本书主要内容的学习。如果读者对这部分内容比较熟悉，建议快速地浏览 1.1~1.3 节，从而熟悉本书的符号使用方式和对问题的叙述方式。

另外，建议读者能较好地完成本章后面所列的习题，尤其是一些构造性的题目，以及关于语言的所有题目，因为它们对后续内容的学习是十分重要的。

本章的主要内容有两部分：一是有关集合、关系、图、证明方法的基本知识；另外讲述形式语言及其相关的基本概念。

1.1 集合的基础知识

集合论(set theory)是德国数学家康托(Georg Cantor)于 1874 年创立的，至今已经历了两个阶段：1908 年以前称为朴素集合论，又称为康托集合论。在朴素集合论中，存在着严重的集合悖论的问题。在朴素集合论刚出现时，人们认为找到了数学的基础，而集合悖论的发现使得人们无比沮丧，觉得数学失去了重要的基础，甚至有人认为以严密为重要特征的数学是不可靠的。为了避免集合悖论，哲墨罗(E. Zermelo)于 1908 年提出了第一个集合论公理系统，后经富兰科尔(A. A. Fraenkel)和斯库利姆(A. T. Skolem)的改进和补充，形成了 ZF 公理系统。同年，B. 罗素也给出了关于集合的型的层次理论——类型论。

无论如何，人们还是公认集合论在数学中占有非常重要的地位，它的基本概念已经渗透到许多领域。计算机科学与技术学科出现后，也将集合论作为其重要基础之一。计算机科学与技术领域中的大多数基本概念和理论几乎都采用了与集合论有关的术语来描述。

1.1.1 集合及其表示

集合是集合论中最原始的概念。我们只能给出它的非形式描述，以说明它的意义：一定范围内确定的，并且彼此可以区分的对象汇集在一起形成的整体叫作**集合**(set)，简称为**集**。简单地说，集合是具有某种性质的对象的全体。构成集合的每一个对象称为这个集合的一个**成员**，它们可以是具体的东西，也可以是抽象的概念。通常称集合的成员为该集合的**元素**(element)。

例 1-1 集合的实际例子。

- (1) 北京市的所有交通工具汇集在一起构成一个集合。
- (2) 北京市的所有公共汽车是一个集合。

(3) 中国所有高等院校组成一个集合。

(4) 某高校的所有院系构成一个集合。

(5) 全体自然数构成一个集合；全体有理数构成一个集合；全体实数构成一个集合。

(6) 一个学校的所有班级的全体是一个集合；一个班的学生的全体是一个集合；一个学生的所有用品的全体是一个集合。

(7) 1, 2, 3, 5, 8, 13 构成一个集合。

(8) 学生、教师构成一个集合。

显然,对象和集合之间有这样一种关系:该对象要么是该集合的一个元素,要么不是该集合的元素,两者必居其一。一个集合中的元素可能都在另一个集合中,也可能部分在另一个集合中。一个对象可以是某一个集合的元素,它本身也可以是一个集合。

通常我们用大写的英文字母 A, B, C, \dots 和大写的希腊字母 $\Gamma, \Sigma, \emptyset, \dots$ 表示集合,用小写字母 a, b, c, \dots 表示集合的元素。例如,一般地,

\mathbf{N} ——表示全体自然数集合

\mathbf{Q} ——表示全体有理数集合

\mathbf{R} ——表示全体实数集合

Σ ——表示字母的集合

关于集合和元素,我们用如下的记法:

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

例 1-2 集合与元素。

(1) $6 \in \mathbf{N}, 1 \in \mathbf{N}, 1.5 \notin \mathbf{N}, 0.81 \notin \mathbf{N}$ 。

(2) $4 \in \mathbf{N}, 1.5 \notin \mathbf{N}, 4 \in \mathbf{Q}, 1.5 \in \mathbf{Q}, 4 \in \mathbf{R}, 1.5 \in \mathbf{R}$ 。

(3) 设 U 是中国所有高等院校组成的集合, 则有

清华大学 $\in U$, 北京大学 $\in U$, 北京工业大学 $\in U$

(4) 设北京工业大学表示该校所有院系组成的集合, 则有

计算机学院 \in 北京工业大学

集合可以用两种形式加以描述:

第一种形式称为**列举法**(listing): 将所有的元素逐一地列举在大括号 $\{ \}$ 中, 读者能立即看出规律时, 某些元素可用省略号表示。

例 1-3 集合的列举表示。

(1) $\{1, 3, 6, 9, 10\}$ 。

(2) $\{a, b, c, \dots, z\}$ 。

(3) {本科生, 硕士研究生, 博士研究生, 进修生}。

(4) {1, 2, 3, 4, 5, ...}。

(5) {0, 5, 10, 15, ..., 200}。

值得注意的是, 在使用列举法时, 集合中元素出现的先后顺序是没有意义的。例如, {1, 3, 6, 9, 10} 与 {6, 3, 10, 9, 1} 表示的是同一个集合。

集合的第二种表示形式称为**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 。

例 1-4 集合的命题表示。

(1) $\{x \mid 0 \leq x \leq 200 \text{ 且 } (\exists n \in N (n \cdot 5 = x))\}$ 。

(2) $\{x \mid 3x^2 + 8x + 4 = 0\}$ 。

(3) $\{x \mid x \in [0, 1]\}$ 。

(4) $\{x \mid x \in \{\text{本科生, 硕士研究生, 博士研究生, 进修生}\}\}$ 。

在有的集合中, 一个元素可以重复出现, 这种集合称为**多重集合**。本书不考虑多重集合的问题, 所提到的集合均不允许一个元素重复出现。

由有限个元素构成的集合叫作**有限集**(finite set), 又称为**有穷集**。由无穷多个元素构成的集合叫作**无穷集**(infinite set)。这是一个直观的描述, 作为一个思考题, 读者可以根据定义 1-1 分别给有穷集和无穷集一个比较严格的定义。

定义 1-1 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), 通常用 $|A|$ 表示集合 A 的基数。

集合的基数又叫作**集合的势**。

对有穷集来说, 它的基数就是它所包含的元素的个数。

例 1-5 有穷集合的基数。

(1) $|\{x \mid 0 \leq x \leq 200 \text{ 且 } (\exists n \in N (n \cdot 5 = x))\}| = 41$ 。

(2) $|\{x \mid 3x^2 + 8x + 4 = 0\}| = 2$ 。

(3) $|\{a, b, c, \dots, z\}| = 26$ 。

(4) $|\{\text{本科生, 硕士研究生, 博士研究生, 进修生}\}| = 4$ 。

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

无穷集可以分成**可数集**(countable infinite set 或 countable set) 和**不可数集**(uncountable set)。

设 S 是一个无穷集, 如果集合 S 与自然数集 $N(\{1, 2, 3, 4, \dots\})$ 具有相同的基数, 则称 S 是可数无穷的集合, 简称 S 是可数的; 否则, 称 S 是不可数集。

例如, 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