第 1 章 绪 论

计算机科学与技术学科**系统地研究信息描述及其变换算法，包括它们的理论、分析、效率、实现和应用**。学科的根本问题是：**什么能且如何被（有效地）自动计算**。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）。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。另外，深入分析不难发现，即使像形式语言与自动机理论这样的内容，也同时具有抽象、理论、设计 3 个形态的内容。对不同类型学生的教学，可以通过强调不同形态的内容来达到教学目的。

众所周知，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“离散变换特征”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、有向图和树。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相关基本概念,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。讲授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。注意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。

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

(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。

如果集合 A 是集合 B 的子集 $A\subseteq B$,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 $A=B$。

2. 注意事项

(1) 对于集合的如下结论,可以在回忆上述基本概念时穿插在其中考虑,不用特意去证明,在提到这些结论时,可以以思考题的方式引导学生去思考,并鼓励这方面知识不扎实的同学在课后自行努力完成几个证明。

对任意集合 A,B,C:

① $A=B$ iff $A\subseteq B$ 且 $B\subseteq A$。

② 如果 $A\subseteq B$,则 $|A|\leqslant|B|$。

③ 如果 $A\subset B$,则 $|A|\leqslant|B|$。

④ 如果 A 是有穷集,且 $A\subset B$,则 $|B|>|A|$。

⑤ 如果 $A\subseteq B$,则对 $\forall x\in A$,有 $x\in B$。

⑥ 如果 $A\subset B$,则对 $\forall x\in A$,有 $x\in B$ 并且 $\exists x\in B$,但 $x\notin A$。

⑦ 如果 $A\subseteq B$ 且 $B\subseteq C$,则 $A\subseteq C$。

⑧ 如果 $A\subseteq B$ 且 $B\subset C$,或者 $A\subset B$ 且 $B\subset C$,或者 $A\subset B$ 且 $B\subseteq C$,则 $A\subset C$。

⑨ 如果 $A=B$,则 $|A|=|B|$。

(2) 通过充要条件、存在量词、全称量词的使用,告诉学生尽量用符号、式子去表达和叙述问题,培养学生形式化表达问题的能力。

1.1.3 集合的运算

1. 知识点

(1) 并

将集合 A 的元素和 B 的元素放在一起构成的集合称为 A 与 B 的**并**(union),记作 $A\cup B$。$A\cup B=\{a\mid a\in A$ 或者 $a\in B\}$。

"$\cup$"为并运算符,$A\cup B$ 读作 A 并 B。

设 $A_1,A_2,\cdots,A_n$ 是 n 个集合,则它们的并 $A_1\cup A_2\cup\cdots\cup A_n=\{a\mid\exists i,1\leqslant i\leqslant n$,使得 $a\in A_i\}$;

设 $A_1,A_2,\cdots,A_n,\cdots$ 是一个集合的无穷序列,则它们的并 $A_1\cup A_2\cup\cdots\cup A_n\cup\cdots=\{a\mid\exists i,i\in N$,使得 $a\in A_i\}$,也可记为 $\bigcup_{i=1}^{\infty}A_i$。

当一个集合的元素都是集合时,可以称为集族。设 S 是一个集族,则 S 中的所有元素的并为:

$\bigcup_{A\in S} A = \{a \mid \exists A \in S, a \in A\}$

(2) 交

集合 A 和 B 中都有的所有元素放在一起构成的集合称为 A 与 B 的**交**(intersection),记作 $A\cap B$。$A\cap B=\{a|a\in A$ 且 $a\in B\}$。

"$\cap$"为交运算符。$A\cap B$ 读作 A 交 B。

如果 $A\cap B=\varnothing$,则称 A 与 B 不相交。

(3) 差

由属于 A,但不属于 B 的所有元素组成的集合称为 A 与 B 的**差**(difference),记作 $A-B$。$A-B=\{a|a\in A$ 且 $a\notin B\}$。

"$-$"为减(差)运算符,$A-B$ 读作 A 减 B。

(4) 对称差

由属于 A 但不属于 B,以及属于 B 但不属于 A 的所有元素组成的集合称为 A 与 B 的**对称差**(symmetric difference),记作 $A\oplus B$。$A\oplus B=\{a|a\in A$ 且 $a\notin B$ 或者 $a\notin A$ 且 $a\in B\}$。

"$\oplus$"为对称差运算符。$A\oplus B$ 读作 A 对称减 B。

$A\oplus B=(A\cup B)-(A\cap B)=(A-B)\cup(B-A)$。

(5) 笛卡儿积

A 与 B 的**笛卡儿积**(cartesian product)是一个集合,该集合是由所有这样的有序对 (a,b) 组成的:其中,$a\in A$,$b\in B$,记作 $A\times B$。$A\times B=\{(a,b)|a\in A$ 且 $b\in B\}$。

"$\times$"为集合的笛卡儿乘运算符。$A\times B$ 读作 A 叉乘 B。

(6) 幂集

A 的**幂集**(power set)$2^A=\{B|B\subseteq A\}$。

(7) 补集

补集又称为余集,它是基于某个论域而言的。论域 U 中的、不在 A 中的所有元素组成的集合称为 A 关于论域 U 的**补集**(complementary set),简称为 A 的补集,记作 $\overline{A}$。$\overline{A}=U-A$。

对补运算,De Morgan 公式成立:$\overline{A\cap B}=\overline{A}\cup\overline{B}$,$\overline{A\cup B}=\overline{A}\cap\overline{B}$。

2. 注意事项

(1) 应强调有穷集合的并和无穷集合的并的异同,尤其注意使学生较好地掌握无穷集合的运算问题和 $\bigcup_{A\in S} A = \{a \mid \exists A \in S, a \in A\}$ 的用法。

(2) 与 1.1.2 节提到的注意事项类似,本节的一些结论,也应该进行同样的处理。

1.2 关　　系

1.2.1 二元关系

1. 知识点

(1) 二元关系

任意的 $R \subseteq A \times B$,R 是 A 到 B 的**二元关系**(binary relation)。

① $(a,b) \in R$,表示 a 与 b 满足关系 R,按照中缀形式,也可表示为 aRb。其中,A 称为**定义域**(domain),B 称为**值域**(range)。当 $A=B$ 时,则称 R 是 A 上的二元关系。

② 二元关系的性质:自反(reflexive)性、反自反(irreflexive)性、对称(symmetric)性、反对称(asymmetric)性、传递(transitive)性。

③ 自反性、对称性、传递性合在一起称为关系的三歧性。

(2) 等价关系

具有三歧性的二元关系称为**等价关系**(equivalence relation)。

(3) 等价类

设 R 是集合 S 上的等价关系,则 S 的满足如下要求的划分 $S_1, S_2, S_3, \cdots, S_n, \cdots$ 称为 S 关于 R 的等价划分,S_i 称为**等价类**(equivalence class)。

① $S = S_1 \cup S_2 \cup S_3 \cup \cdots \cup S_n \cup \cdots$。

② 如果 $i \neq j$,则 $S_i \cap S_j = \varnothing$。

③ 对任意的 i,S_i 中的任意两个元素 a,b,aRb 恒成立。

④ 对任意的 i,j,$i \neq j$,S_i 中的任意元素 a 和 S_j 中的任意元素 b,aRb 恒不成立。

R 将 S 分成的等价类的个数称为 R 在 S 上的**指数**(index)。如果 R 将 S 分成有穷多个等价类,则称 R 具有有穷指数;如果 R 将 S 分成无穷多个等价类,则称 R 具有无穷指数。

(4) 关系的合成

设 $R_1 \subseteq A \times B$ 是 A 到 B 的关系,$R_2 \subseteq B \times C$ 是 B 到 C 的关系,R_1 与 R_2 的**合成**(composition) R_1R_2 是 A 到 C 的关系:$R_1R_2 = \{(a,c) \mid \exists (a,b) \in R_1 \text{且} (b,c) \in R_2\}$。

2. 注意事项

(1) 关系用来反映对象——集合元素之间的联系和性质。二元关系则是反映两个元素之间的关系,包括某个元素的某种属性。读者需要建立这种观念,以促进对后续内容的理解。

(2) 注意全称量词是对什么样的范围而言的。

(3) 需要通过等价关系强化等价分类的概念，为后面章节中的正则语言的描述(RG,FA)和对相应的 R_M,R_L 的理解做好准备。

(4) 关系合成的几个结论无需仔细证明。

1.2.2 递归定义与归纳证明

1. 知识点

递归定义(recursive definition)又称为**归纳定义**(inductive definition)，可用来定义一个集合。一般地，一个集合的递归定义由以下 3 个部分组成。

(1) 基础(basis)：用来定义该集合最基本的元素。

(2) 归纳(induction)：它指出用集合中的元素来构造集合的新元素的规则。其一般形式为：如果 $a,b,c,\cdots,d$ 是被定义集合的元素，则用某种运算、函数或者组合方法对 $a,b,c,\cdots,d$ 进行处理后所得的结果也是集合中的元素。

(3) 极小性限定：指出一个对象是所定义集合中的元素的充要条件是可以通过有限次的使用基础和归纳条款中所给的规定构造出来。

与递归定义相对应，归纳证明方法包括以下 3 个步骤。

(1) 基础(basis)：证明该集合的最基本元素具有给定性质。

(2) 归纳(induction)：证明如果某些元素具有相应性质，则根据所规定的方法得到的新元素也具有相应的性质。它的形式一般为：如果 $a,b,c,\cdots,d$ 具有相应的性质，则用规定的方法根据 $a,b,c,\cdots,d$ 构造出来的新元素也具有相应的性质。

(3) 根据归纳法原理，所有的元素具有给定的性质。

2. 注意事项

(1) 递归定义给出的概念有利于归纳证明。在计算机科学与技术学科中，有许多问题可以用递归定义描述或者用归纳方法进行证明，而且在许多时候，这样做会带来很多方便，特别是有利于给出无穷对象的有穷描述。因此，读者应该逐步掌握这种对问题的描述和证明方法。

(2) 在课堂上，最多选择一个典型例子仔细讲解一遍即可。主要是让学生掌握这种方法的叙述格式。$|2^A|=2^{|A|}$(主教材例 1-19)，表达式的前缀形式与中缀形式对应(主教材例 1-20)是值得考虑使用的例子。

1.2.3 关系的闭包

1. 知识点

(1) 设 R 是 S 上的关系，递归地定义 R 的 n 次幂 R^n：

① $R^0=\{(a,a)\mid a\in S\}$。

② $R^i=R^{i-1}\circ R$，$i=1,2,3,4,5,\cdots$。

(2) 关系 R 的 P **闭包**(closure)是包含 R 并且具有 P 中所有性质的最小关系。

R 的**正闭包**(positive closure) 又称为 R 的**传递闭包**(transitive closure)，用 R^+ 表示，定义为：

① $R\subseteq R^+$。

② 如果$(a,b),(b,c)\in R^+$，则$(a,c)\in R^+$。

③ 除①、②外，R^+不再含有其他任何元素。

$$R^+=R\cup R^2\cup R^3\cup R^4\cup\cdots$$

当 S 为有穷集时，

$$R^+=R\cup R^2\cup R^3\cup\cdots\cup R^{|S|}$$

R 的**克林闭包**(Kleene closure) 又称为 R 的**自反传递闭包**(reflexive and transitive closure)，用 R^* 表示，定义为：

① $R^0\subseteq R^*$，$R\subseteq R^*$。

② 如果$(a,b),(b,c)\in R^*$，则$(a,c)\in R^*$。

③ 除①、②外，R^* 不再含有其他任何元素。

$$R^*=R^0\cup R^+=R^0\cup R\cup R^2\cup R^3\cup R^4\cup\cdots$$

当 S 为有穷集时，

$$R^*=R^0\cup R\cup R^2\cup R^3\cup\cdots\cup R^{|S|}$$

2. 注意事项

(1) 介绍二元关系的传递闭包和自反传递闭包的性质。

(2) 关于闭包运算的几个结论要适当说明，但无需详细证明。

1.3 图

1.3.1 无向图

1. 知识点

(1) 无向图的概念

设 V 是一个非空有穷集合，$E\subseteq V\times V$，$G=(V,E)$称为**无向图**(undirected graph)。其中，V 中的元素称为**顶点**(vertex 或 node)，V 称为顶点集，E 中的元素称为**无向边**(undirected edge)，E 为无向边集。顶点又称为结点。

(2) 图表示

图 $G=(V,E)$的**图表示**是满足下列条件的“图”:其中,V 中称为顶点 v 的元素用标记为 v 的小圈表示,E 中的元素(v_1,v_2)用标记为 v_1,v_2的顶点之间的连线表示。“图”、“图表示”统一简称为图。

(3) 路

如果对于 $0\leqslant i\leqslant k-1,k\geqslant 1$,均有$(v_i,v_{i+1})\in E$,则称 $v_0,v_1,\cdots,v_k$是 $G=(V,E)$的一条长为 k 的**路**(Path),当 $v_0=v_k$时,$v_0,v_1,\cdots,v_k$称为一个**回路或圈**(cycle)。

(4) 顶点的度数

对于 $v\in V$,$|\{v|(v,w)\in E\}|$称为无向图 $G=(V,E)$的顶点 v 的度数,记作 $\deg(v)$。

对于任何一个图,图中所有顶点的度数之和为图中边的两倍:

$$\sum_{v\in V}\deg(v)=2|E|$$

(5) 连通图

如果对于 $\forall v,w\in V,v\neq w$,v 与 w 之间至少有一条路存在,则称 $G=(V,E)$是**连通图**。

2. 注意事项

(1) 尽量用实例说明有关的概念。

(2) 图 G 是连通的充要条件是 G 中存在一条包含图的所有顶点的路。

1.3.2 有向图

1. 知识点

(1) 有向图

V 是一个非空的有穷集合,$E\subseteq V\times V$,$G=(V,E)$称为一个**有向图**(directed grahp)。其中,V 中的元素称为**顶点**(vertex 或 node),V 称为顶点集,$\forall(v_1,v_2)\in E$ 称为从顶点 v_1到顶点 v_2的**有向边**(directed edge),或**弧**(arc),v_1称为**前导**(predecessor),v_2称为**后继**(successor)。E 称为有向边集。顶点又称为结点。

(2) 有向路

如果对于 $0\leqslant i\leqslant k-1,k\geqslant 1$,均有$(v_i,v_{i+1})\in E$,则称 $v_0,v_1,\cdots,v_k$是 $G=(V,E)$的一条长为 k 的**有向路**(directed path),当 $v_0=v_k$时,$v_0,v_1,\cdots,v_k$称为一个**有向回路**或**有向圈**(directed cycle)。

(3) 图表示

$G=(V,E)$的图表示是满足下列条件的“图”:其中,V 中称为顶点 v 的元素用标记

为 v 的小圈表示,E 中的元素 (v_1, v_2) 用从标记为 v_1 的顶点到标记为 v_2 的顶点的弧表示。“图”、“图表示”是图的两种表示形式,统称为图。

(4) 顶点的度

① $\text{ideg}(v) = |\{v|(w,v) \in E\}|$ 称为有向图 $G=(V,E)$ 的顶点 v 的**入度**数,表示到达该顶点的边的个数。

② $\text{odeg}(v) = |\{v|(v,w) \in E\}|$,称为有向图 $G=(V,E)$ 的顶点 v 的**出度**数,表示离开该顶点的边的个数。

2. 注意事项

(1) 通过强调与无向图中有关定义的异同来解释有向图。

(2) 有向图顶点 v 的出度数和入度数与该图中“经过”v 的长度为 2 的路的条数之间的关系将在后面用到,建议在此处讨论一下。

1.3.3 树

1. 知识点

(1) 树

满足如下条件的有向图 $G=(V,E)$ 称为一棵(有序、有向)**树**(tree):

① $\exists v \in V$,v 没有前导,且 v 到树中其他顶点均有一条有向路,称此顶点为树 G 的**根**(root)。

② 每个非根顶点有且仅有一个前导。

③ 每个顶点的后继按其拓扑关系从左到右排序。

(2) 树的基本概念

① 顶点也可以称为结点。

② 结点的前导为该结点的**父亲**(**父结点** father)。

③ 结点的后继为它的**儿子**(son)。

④ 如果树中有一条从结点 v_1 到结点 v_2 的路,则称 v_1 是 v_2 的**祖先**(ancestor),v_2 是 v_1 的**后代**(descendant)。

⑤ 无儿子的顶点称为**叶子**(leaf)。

⑥ 非叶结点称为**中间结点**(interior)。

(3) 树的层

① 根处在树的第 1 层(level)。

② 如果结点 v 处在第 i 层($i \geqslant 1$),则 v 的儿子处在第 $i+1$ 层。

③ 树的最大层号称为该树的高度(height)。

(4) 二元树

如果对于$\forall v \in V$,v最多只有2个儿子,则称$G=(V,E)$为**二元树**(binary tree)。

对一棵二元树,它的第n层最多有2^{n-1}个结点。一棵n层二元树最多有2^{n-1}个叶子。

2. 注意事项

(1) 适当讨论二元树在"最满"的时候叶子个数的计算,各层结点数的计算,请读者考虑最大路长为n的二元树的叶子的最大个数。

(2) 因为后面将讨论语法树,所以要适当讨论各结点的关系,尤其是祖先与后代的关系。

1.4 语　　言

1.4.1 什么是语言

1. 知识点

(1) 语言的概念

① 关键点:字,组成规则,理解(语义)规则。

② 斯大林强调语言的作用,认为语言是"广大人群所理解的字和组合这些字的方法"。

③ 语言学家韦波斯特(Webster)指出:为相当大的团体的人所懂得并使用的字和组合这些字的方法的统一体。

④ 要想对语言的性质进行研究,用这些定义来建立语言的数学模型是不够精确的。必须有更形式化的定义。

(2) 形式语言

将语言抽象地定义成一个数学系统,其形式性可以使我们能给出语言的严格描述,并能通过此发展出一批知识——理论,而后将这些知识用到适当的模型中,使之能够在科学实践中起到良好的指导作用。

2. 主要内容解读

从自然语言的一个简单句子的构成及其表达意思提取出语言的构成要素,然后给

出语言学家给语言所下的定义。这一部分内容主要围绕着主教材图1-5*理解。

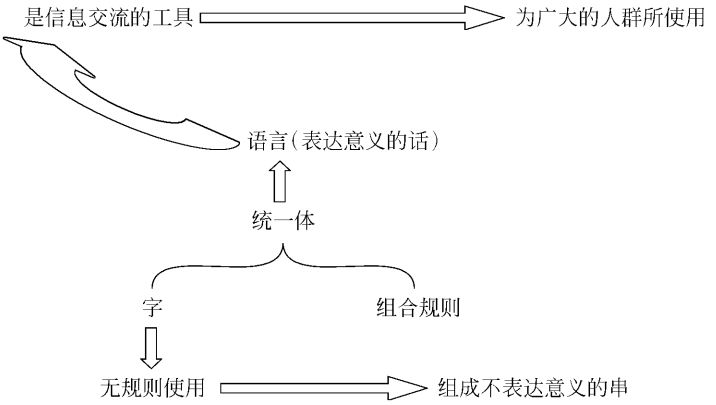

图1-5* 语言是字及其组合规则的统一体

1.4.2 形式语言与自动机理论的产生与作用

1. 知识点

(1) 形式语言的发展历史。

① 语言学家乔姆斯基,毕业于美国宾夕法尼亚大学,最初从产生语言的角度研究语言。1956年,他将语言L定义为一个字母表Σ中的字母组成的一些串的集合:$L\subseteq\Sigma^*$。

② 在字母表上按照一定的规则定义一个文法(grammar),该文法所能产生的所有句子组成的集合就是该文法产生的语言。

③ 1959年,乔姆斯基根据产生语言文法的特性,将语言划分成三大类。

④ 1951年到1956年,克林(Kleene)在研究神经细胞中,建立了识别语言的系统——有穷状态自动机。

⑤ 1959年,乔姆斯基发现文法和自动机分别从生成和识别的角度去表达语言,而且证明了文法与自动机的等价性,这一成果被认为是将形式语言置于了数学的光芒之下,使得形式语言真正诞生了。

⑥ 20世纪50年代,巴科斯范式(Backus Nour form或Backus normal form,BNF)实现了对高级语言ALGOL-60的成功描述。这一成功,使得形式语言在20世纪60年代得到了大力的发展。尤其是上下文无关文法被作为计算机程序设计语言的最佳近似描述得到了较为深入的研究。

---

* 注:为了与主教材保持一致,本书的图序号统一采用原主教材的图序号。

⑦ 相应的理论用于其他方面。

(2) 形式语言与自动机理论在计算机科学与技术学科人才计算思维能力的培养中占有极其重要的地位。

(3) 主教材图 1-6 表达的计算学科的主题:“什么能且如何被有效地自动计算”。

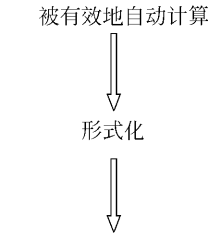

图 1-6　自动计算、形式化与“计算思维”

(4) 计算机科学与技术学科人才基本专业能力构成。

① “计算思维能力”——问题的形式化与模型化描述、抽象思维能力、逻辑思维能力。

② 算法设计与分析能力。

③ 程序设计与实现能力。

④ 计算机系统的认知、分析、开发和应用能力。

(5) 知识的载体属性在能力培养中的体现以及计算思维能力的培养过程见主教材(图 1-7)。

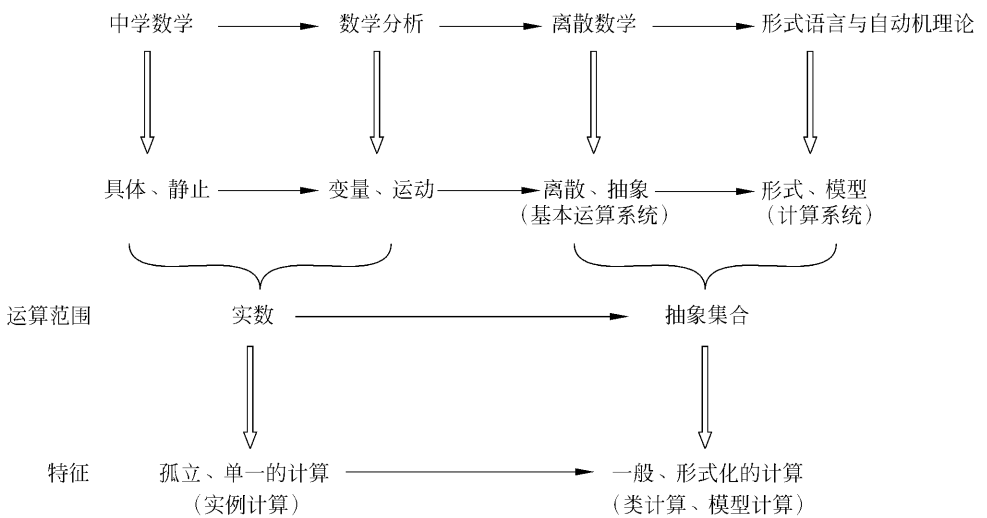

图 1-7　“计算思维能力”梯级训练系统

2. 主要内容解读

(1) 重点论述形式语言与自动机理论在计算机科学与技术学科人才计算思维能力培养中的重要作用,以激发学生的学习热情。这需要通过“什么能且如何被有效地自动计算”这一计算学科的主题,探讨该学科人才的能力构成,如果可能,再适当结合计算学

科方法论的内容进行讨论。这样才能使这段论述更全面、更清楚,才能在今后的教学中获得学生的较好配合。

(2) 尽可能加进教师自己的亲身体会,以增加说服力。

(3) 最后应该能自然地得出结论——形式语言与自动机理论不仅是计算机科学与技术学科重要的基础理论,有着广泛的应用,而且还在计算机科学与技术学科人才的培养中占有十分重要的地位,是一个优秀的计算机科学工作者必修的一门课程。

1.4.3 基本概念

1. 知识点

(1) 对语言研究的3个方面

① 表示(representation)——无穷语言的表示。

② 有穷描述(finite description)——研究的语言要么是有穷的,要么是可数无穷的,这里主要研究可数无穷语言的有穷描述。

③ 结构(structure)——语言的结构特征。

(2) 字母表(alphabet)与字母(letter)

字母表(alphabet) 是一个非空有穷集合,字母表中的元素称为该字母表的一个**字母**(letter)。又叫做**符号**(symbol)或者**字符**(character)。

字母表的非空性和有穷性。

(3) 字符的两个特性

① 整体性(monolith),也叫不可分性。

② 可辨认性(distinguishable),也叫可区分性。

(4) 字母表的乘积(product)

字母表 Σ_1 与 Σ_2 的乘积 $\Sigma_1\Sigma_2=\{ab \mid a\in\Sigma_1 \text{且} b\in\Sigma_2\}$。

(5) 字母表的 n 次幂

Σ 的 n 次**幂**递归地定义为:

① $\Sigma^0=\{\varepsilon\}$。

② $\Sigma^n=\Sigma^{n-1}\Sigma$, $n\geqslant 1$。

(6) 字母表的闭包

字母表 Σ 的**正闭包**:

$$\Sigma^+=\Sigma\cup\Sigma^2\cup\Sigma^3\cup\Sigma^4\cup\cdots$$

Σ 的**克林闭包**:

$$\Sigma^*=\Sigma^0\cup\Sigma^+=\Sigma^0\cup\Sigma\cup\Sigma^2\cup\Sigma^3\cup\cdots$$

(7) 句子

$\forall x\in\Sigma^*$，x 称为 Σ 上的一个**句子**(sentence)。句子还可以称为字(word)、(字符、符号)行(line)、(字符、符号)串(string)。

两个句子被称为相等的，如果它们对应位置上的字符都对应相等。

$x,y\in\Sigma^*$，$a\in\Sigma$，句子 xay 中的 a 称为 a 在该句子中的一个**出现**(apperance)。

$\forall x\in\Sigma^*$，句子 x 中字符出现的总个数称为该**句子的长度**(length)，记作 $|x|$。

长度为 0 的字符串叫空句子，记作 ε。

(8) 句子的并置与幂

① $x,y\in\Sigma^*$，x,y 的**并置**(concatenation)是这样一个串，该串是由串 x 直接连接串 y 所组成的，记作 xy。并置又称为**连接**。

② 对于 $n\geqslant 0$，串 x 的 n 次幂定义为：

- $x^0=\varepsilon$。
- $x^n=x^{n-1}x$。

③ Σ^* 上的并置运算具有如下性质：对 Σ^* 上的任意串 x,y,z，

- 结合律：$(xy)z=x(yz)$。
- 左消去律：如果 $xy=xz$，则 $y=z$。
- 右消去律：如果 $yx=zx$，则 $y=z$。
- 唯一分解性：存在唯一确定的 $a_1,a_2,\cdots,a_n\in\Sigma$，使得 $x=a_1a_2\cdots a_n$。
- 单位元素：$\varepsilon x=x\varepsilon=x$。

(9) 前缀与后缀

设 $x,y,z,w,v\in\Sigma^*$，且 $x=yz,w=yv$，

① y 是 x 的前缀(prefix)。

② 如果 $z\neq\varepsilon$，则 y 是 x 的真前缀(proper prefix)。

③ z 是 x 的后缀(suffix)。

④ 如果 $y\neq\varepsilon$，则 z 是 x 的真后缀(proper suffix)。

⑤ y 是 x 和 w 的公共前缀(common prefix)。

⑥ 如果 x 和 w 的任何公共前缀都是 y 的前缀，则 y 是 x 和 w 的最大公共前缀。

⑦ 如果 $x=zy,w=vy$，则 y 是 x 和 w 的公共后缀(common suffix)。

⑧ 如果 x 和 w 的任何公共后缀都是 y 的后缀，则 y 是 x 和 w 的最大公共后缀。

(10) 子串

设 $t,u,v,w,x,y,z\in\Sigma^*$，

① 如果 $w=xyz$，则称 y 是 w 的子串(substring)。

② 如果 $t=uyv, w=xyz$,则称 y 是 t 和 w 的公共子串(common substring)。

③ 如果 $y_1, y_2, \cdots, y_n$ 是 t 和 w 的公共子串,且 $\max\{|y_1|, |y_2|, \cdots, |y_n|\}=|y_j|$,则称 y_j 是 t 和 w 的**最大公共子串**。

(11) 语言与给定语言的句子

$\forall L\subseteq\Sigma^*$,$L$ 称为字母表 Σ 上的一个语言(language),$\forall x\in L$,x 称为 L 的一个句子。

(12) 语言的乘积

$L_1\subseteq\Sigma_1^*$,$L_2\subseteq\Sigma_2^*$,语言 L_1 与 L_2 的乘积(product)是语言 $L_1L_2=\{xy|x\in L_1, y\in L_2\}$。

(13) 语言的幂

$\forall L\subseteq\Sigma^*$,$L$ 的 n 次幂是一个语言,该语言定义为:

① 当 $n=0$ 时,$L^n=\{\varepsilon\}$。

② 当 $n\geqslant 1$ 时,$L^n=L^{n-1}L$。

(14) 语言的闭包

① L 的正闭包 $L^+=L\cup L^2\cup L^3\cup L^4\cup\cdots$。

② L 的克林闭包 $L^*=L^0\cup L\cup L^2\cup L^3\cup L^4\cup\cdots$。

2. 主要内容解读

(1) 强调字母表为什么是**非空**的、**有穷**的集合。

(2) 字母表中的字母是组成字母表上的语言中的任何句子的最基本元素。

(3) 注意在介绍字母表的闭包时,说明以下两点,这对初学者来讲是很重要的。

$\Sigma^*=\{x|x$ 是 Σ 中的若干个,包括 0 个字符连接而成的一个字符串$\}$

$\Sigma^+=\{x|x$ 是 Σ 中的至少一个字符连接而成的字符串$\}$

(4) 对 ε 强调以下两点:

① ε 是一个句子。

② $\{\varepsilon\}\neq\varnothing$,这是因为$\{\varepsilon\}$是含有一个空句子 ε 的集合,不是一个空集。$|\{\varepsilon\}|=1$,而$|\varnothing|=0$。

(5) 前缀、后缀、真前缀和真后缀,是比较简单的概念,关于它们的讲解,可以和子串的概念一起,考虑在计算机系统处理"语句"的时候是如何处理它们的。

(6) 将字母用法的约定当成是培养良好表达习惯的一种努力。

(7) 关于两个串的最大公共子串并不一定是唯一的这一结论,可以举一个例子说明,也可以作为一道思考题留给学生。

(8) 虽然曾经给过一些乘积的定义,但是,关于语言的相应定义还必须在此给出,

以保证定义的严格性。在这里注意插入一些例子，让学生开始接触语言的结构这一重要问题，否则，到构造文法和自动机时，会因“新”内容过多而造成理解上的困难。

1.5 小　结

本章简要叙述了基础知识，一方面，希望读者通过对本章的阅读，熟悉集合、关系、图、形式语言等相关的基本知识点，为以后各章的学习做适当的准备。另一方面，也使读者熟悉本书中一些符号的意义。

(1) 集合：集合的表示、集合之间的关系、集合的基本运算。

(2) 关系：等价关系、等价分类、关系合成、关系闭包。

(3) 递归定义与归纳证明。

(4) 图：无向图、有向图、树的基本概念。

(5) 语言与形式语言：自然语言的描述，形式语言和自动机理论的出现，形式语言和自动机理论对计算机科学与技术学科人才能力培养的作用。

(6) 基本概念：字母表、字母、句子、字母表上的语言、语言的基本运算。

1.6 典型习题解析*

可以从主教材第1章习题16～32中选择适量的习题要求学生完成，其他的可以作为学生课外自习时参考。

16. 设 L 是 Σ 上的一个语言，Σ^* 上的二元关系 R_L 定义为：对任给的 $x,y\in\Sigma^*$，如果对于 $\forall z\in\Sigma^*$，均有 $xz\in L$ 与 $yz\in L$ 同时成立或者同时不成立，则 xR_Ly。请证明 R_L 是 Σ^* 上的一个等价关系。将 R_L 称为由语言 L 所确定的等价关系。实际上，R_L 还有另外一个性质：如果对任给的 $x,y\in\Sigma^*$，当 xR_Ly 成立时，必有 xzR_Lyz 对 $\forall z\in\Sigma^*$ 都成立。这将被称为 R_L 的“右不变”性。你能证明此性质成立吗？

证明提示：直接参考5.3节中命题5-2的证明。

18. 设 $\{0,1\}^*$ 上的语言 $L=\{0^n1^n\mid n\geqslant 0\}$，请给出 $\{0,1\}^*$ 的关于 L 所确定的等价关系 R_L 的等价分类。

解：根据第16题的定义及其证明，考虑 R_L 对 $\{0,1\}^*$ 的等价分类时，主要需根据语言 L 的结构，分析哪些串按照 L 的要求具有相同的特征。

首先，取 $0^n1^n,0^m1^m\in L,0^n1^n\varepsilon\in L,0^m1^m\varepsilon\in L$ 同时成立，但是，对所有 $x\in\{0,1\}^+$，

---

* 注：本书的习题序号与原主教材习题序号保持一致。

$0^n1^nx\in L$,$0^m1^mx\in L$ 同时不成立,满足第16题中所给定义的要求;所以,L 中的元素是属于同一类的。

第二,分析是否还有其他的元素与 L 中的元素属于同一类。根据 L 的结构,一种类型的串不含子串10,这种串可以表示成 0^k1^h,$k\neq h$;另一种是含有子串10的串(L 的句子不含这种子串),这种串可以表示成 $x10y$。显然,$01\varepsilon\in L$,但是,$0^k1^h\notin L(k\neq h)$,$x10y\notin L(x,y\in\{0,1\}^*)$。根据等价分类的性质,$\{0,1\}^*$ 中的不在 L 中的串与 L 中的串不在同一个等价类中。

第三,考察不含子串10的串。这些串有如下几种形式:

① 0^n,$n\geqslant 1$。

② 1^n,$n\geqslant 1$。

③ 0^m1^n,$m,n\geqslant 1$ 且 $m>n$。

④ 0^m1^n,$m,n\geqslant 1$ 且 $m<n$。

对于 0^n,$n\geqslant 1$,1^n 接在它后面时,构成串 0^n1^n。显然,当 $m\neq n$ 时,$0^n1^n\in L$,但 $0^m1^n\notin L$。所以,0^n 和 0^m 一定不在同一个等价类中。

类似的讨论可知,对于 0^n,$n\geqslant 1$:

0^n 不可能与形如 $0^m1^n(m,n\geqslant 1$ 且 $m<n)$ 的串在同一个等价类中;

0^n 不可能与含有子串10的串在同一个等价类中。

下面再考察形如 0^h 的串和形如 0^m1^n,$m,n\geqslant 1$ 且 $m>n$ 的串是否可能在同一等价类中。注意到当 $m-n=h$ 时,

$$0^h1^h\in L,0^m1^n\,1^h\in L$$

同时成立,但是当 $n\geqslant 1$,$x=01^{h+1}$ 时($x\neq 1^h$),

$$0^hx\in L,0^m1^n\,x\notin L$$

成立。所以,对应 $h>0$,令

$$[h]=\{0^m1^n\mid m-n=h\text{ 且 }n\geqslant 1\}$$

$[h]$ 中的元素在同一个等价类中,而且所有其他的元素都不在这个等价类中。实际上,当 $h=0$ 时,有

$$[0]=L$$

第四,形如 1^m 的串和形如 $0^m1^n(m,n\geqslant 1$ 且 $m<n)$ 的串应该在同一等价类中。事实上,对于 $\{0,1\}^*$ 中的任意字符串 x,

$$1^mx\notin L,0^m1^nx\notin L(m,n\geqslant 1\text{ 且 }m<n)$$

恒成立。所以,这些字符串在同一个等价类中。

第五,所有含子串10的串在同一等价类中。事实上,设 y,z 是含有子串10的串,对于 $\{0,1\}^*$ 中的任意字符串 x,

$yx\notin L, zx\notin L(m,n\geqslant 1$ 且 $m<n)$

恒成立。所以，这些字符串在同一个等价类中。

第六，形如 1^m 的串和含子串 10 的串在同一等价类中。事实上，设 y 是含有子串 10 的串，对于 $\{0,1\}^*$ 中的任意字符串 x，

$1^m x\notin L, yx\notin L(m\geqslant 1)$

恒成立。所以，这些字符串在同一个等价类中。

综上所述，R_L 确定的 $\{0,1\}^*$ 的等价分类为

$[10]=\{x10y\mid x,y\in\{0,1\}^*\}\cup\{0^m1^n\mid n-m\geqslant 1\}$

$[0]=\{0^m1^n\mid m-n=0\}=\{0^n1^n\mid n\geqslant 0\}$

$[1]=\{0^m1^n\mid m-n=1,n\geqslant 1\}$

$[2]=\{0^m1^n\mid m-n=2,n\geqslant 1\}$

$\vdots$

$[h]=\{0^m1^n\mid m-n=h,n\geqslant 1\}$

$\vdots$

$\{0\}$

$\{00\}$

$\vdots$

$\{0^n\}$

$\vdots$

其中，n,m 均为非负整数。

20. 使用归纳法证明下列各题。

(9) 对字母表 Σ 中的任意字符串 x，x 的前缀有 $|x|+1$ 个。

证明：设 $x\in\Sigma^*$，现对 x 的长度施归纳。为了叙述方便起见，用 prefix(x)表示字符串 x 的所有前缀组成的集合。

当 $|x|=0$ 时，有 $x=\varepsilon$，由字符串的前缀定义知道，

prefix(x)$=\{\varepsilon\}$。

ε 就是 x 的唯一前缀。而

$$\begin{aligned}|\text{prefix}(x)| &= |\{\varepsilon\}|\\ &=1\\ &=0+1\\ &=|x|+1\end{aligned}$$

所以，结论对 $|x|=0$ 成立。

设 $|x|=n$ 时结论成立，$n\geqslant 0$。即

$$|\text{prefix}(x)|=|x|+1$$

现在考察$|x|=n+1$的情况。为了叙述方便,不妨设$x=ya$,其中$|y|=n$,并且$a\in\Sigma$。由归纳假设,

$$|\text{prefix}(y)|=|y|+1$$

首先证明y的任何前缀都是x的前缀。事实上,设

$$\text{prefix}(y)=\{u_1,u_2,\cdots,u_n\}$$

对于$\forall u\in\text{prefix}(y)$,根据前缀的定义,存在$v\in\Sigma^*$,使得$uv=y$,注意到$uva=x$,所以,$u$也是$x$的前缀,它对应的$x$的后缀为$va$。

再注意到$x=ya$,所以,一方面,对于$\forall u\in\text{prefix}(y)$,均有

$u\neq x$。

从而

$$x\notin\text{prefix}(y),$$

然而,由

$$x\varepsilon=x$$

可知,x是x的一个前缀。另一方面,由$x=ya$知道,如果u是x的一个前缀,v是u对应的x的后缀,则有如下两种情况:

① $|v|\geqslant 1$,此时必有$u\in\text{prefix}(y)$。

② $|v|=0$,此时必有$v=\varepsilon$并且$u=ya=x$。

由此可见,

$$\begin{aligned}\text{prefix}(x)&=\text{prefix}(y)\cup\{x\}\\&=\{u_1,u_2,\cdots,u_n,x\}\end{aligned}$$

由$x\notin\text{prefix}(y)$可知,

$$\begin{aligned}|\text{prefix}(x)|&=|\text{prefix}(y)\cup\{x\}|\\&=|\text{prefix}(y)|+|\{x\}|\\&=|\text{prefix}(y)|+1\end{aligned}$$

再由归纳假设

$$\begin{aligned}|\text{prefix}(x)|&=|\text{prefix}(y)|+1\\&=|y|+1+1\\&=|x|+1\end{aligned}$$

表明结论对$|x|=n+1$成立。由归纳法原理,结论对于任意$x\in\Sigma^*$成立。

22. 设$\Sigma=\{a,b\}$,求字符串$aaaaabbbba$的所有前缀的集合,后缀的集合,真前缀的集合,真后缀的集合。

解:下面给出结果。