

第 1 章 数据、数学模型和算法

1.1 数据时代

大数据 (big data) 是当今学术界和产业界最炙手可热的名词之一，其重要性和价值已经得到广泛的认同。数据科学，继实验科学、理论科学、计算机仿真之后，被称为科学研究的第四范式。为什么数据处理技术会得到如此普遍的重视？为什么人类会对数据中的价值寄予如此巨大的期望？又为什么人类社会发展到今天，数据的重要性会特别地凸显出来呢？我们从什么是数据谈起。

1.1.1 什么是数据

“数”是人们用来表示事物的量的基本数学概念。在人类发展的历史上，这种抽象的“数”的概念是从具体事物中逐步获得和建立起来的。例如“一个苹果”“二个橘子”“三个香蕉”描述的是具体的事物，而“一”“二”“三”则是与具体事物无关的抽象的“数”。另一个相关的概念是“数字”，数字是人们用来计数的符号，如现在人们常用的阿拉伯数字“1”“2”“3”，又如中文的数字“一”“二”“三”和罗马数字“I”“II”“III”。而我们在这里要讨论的“数据”，则是一个范围大得多的概念。

数据是客观事物的符号表示，往往是通过对外观事物的观察得到的未经加工的原始素材，是包含知识和信息的原始材料。在今天的信息社会中，数据可以说无处不在，其表现形式也是多种多样，例如：

文字和符号：不仅普遍存在于书籍、报纸等传统的纸质媒介上，也广泛存在于计算机、手机、平板电脑等电子设备上；既包括今天人们使用的各种文字符号，也包括从远古时代遗存下来的象形文字和甲骨文等。

多媒体数据：计算机的图形界面、广播电视电影、数码相机 (DC) 和数码摄像机 (DV)，使得我们身处于丰富多彩的多媒体时代。多媒体数据的采集、保存和播放已经非常方便；图像、音频、视频等各种媒体数据在我们的日常生活中随处可见。

通信信号：电信号和电磁波已经成为人类社会信息最方便快捷的传输方式，这些用于通信和控制的电话信号、导航信号、手机信号、广播信号，无论是在发送端还是在接收端都是数据。

传感器采集的数据：通过各种各样的温度传感器、压力传感器，以及 CT、B 超、声呐等，人们可以采集到各种各样能够描述客观事物的数据。

社会性数据：人类社会生活的方方面面同样需要大量数据来描述，如社会普查数据、

人口统计数据 and 民意调查数据等，著名的如美国总统大选期间盖洛普所做的候选人支持率的民意测验；也包括紧密联系我们日常生活的经济运行数据，如物价、收入等。随着社交网络的发展和普及，人们之间通过互联网和移动互联网的交互行为也成为重要的海量数据来源。

可以很清楚地看到对数据的掌握和处理是当今社会的一个基本问题，在科研活动、经济活动、文化活动和政治活动中，我们随时都会面对各种各样的数据。数据和对数据的处理与我们每个人都息息相关。

我们在这里讨论的数据，进一步被特指为能够输入到计算机并被计算机处理的。

1.1.2 大数据时代

数据处理技术包括了数据的获取、数据的存储、数据的传输，以及针对数据的计算等。

数据是客观事物的表示和描述，人具有很强的获取数据的能力，如人对客观事物的观察，社会普查等；数据获取也可以通过多种多样的设备，如温度和压力等各种传感器，万用表和光谱仪等各种测量仪器，照相机和摄像机等图像视频采集设备，麦克风和录音机等声音采集设备，雷达接收机和卫星接收机等信号接收设备等。

传统的数据存储主要依靠纸质媒介，如书籍、报表和纸质文件等，典型的模拟存储介质有胶片和磁带。随着数字技术的发展，数字存储介质已经成为主流。从大型的磁盘存储系统，到容量越来越大的计算机硬盘，再到便携的移动硬盘、U 盘、光盘和闪存卡，存储容量不断增大，而且价格越来越便宜。

语言交流和书信曾是人类历史上数据传输和信息交互的主要手段。电磁波和电信号的发现和利用，造就了电话、电报等快捷的数据传输方式。互联网、移动通信，以及 USB 和 IEEE 1394 等高速率数据传输技术的发展，使数据传输的快速、高效和方便达到了前所未有的程度。

面向数据的计算涵盖了对数据的分析、管理和利用。其中既包括了以处理器性能为代表的计算能力，又包括了对数据进行处理以实现信息抽取和知识发现的技术方法。

随着信息技术的飞速发展，人类在数据采集、数据存储和数据传输方面的能力得到了长足的发展。我们都知道，二进制是数字计算机的基础，计算机存储容量的基本单位是字节 (Byte)，每个字节包含 8 个二进制位。为了描述不同规模的数据，人们定义了一系列的数据计量单位：

Bytes $\rightarrow$ Kilobyte(2^{10} Bytes) $\rightarrow$ Megabyte(2^{20} Bytes) $\rightarrow$ Gigabyte(2^{30} Bytes) $\rightarrow$ Terabyte(2^{40} Bytes) $\rightarrow$ Petabyte(2^{50} Bytes) $\rightarrow$ Exabyte(2^{60} Bytes) $\rightarrow$ Zettabyte(2^{70} Bytes) $\rightarrow$ Yottabyte(2^{80} Bytes)

其中我们比较熟悉的有千字节 (KB)、兆字节 (MB) 和吉字节 (GB)。我们甚至难以想象更大的数据量单位意味着什么？美国国会图书馆所有藏书的数据约为 10TB。按照 2001 年的数据估算，美国国家航空航天局地球观测系统 (Earth Observing System) 三年的数据总和约为 1PB^[1]。据称 1 个 ZB 大概相当于全世界所有海滩上的沙子总和，而 1 个 YB 大概相当于 7000 人体内的原子数总和^[2]。如果以每分钟 1MB 的速度不间断播放 MP3 格式的歌曲，1ZB 存储的歌曲可以让人听上 19 亿年。

根据 IDC 的统计和预测, 2007 年全球数据量约为 161EB; 2008 年激增到 487EB; 金融危机的 2009 年, 全球数据量达到 0.8ZB, 增长 62%; 2010 年进一步增长到 1.2ZB, 约为 2007 年的 8 倍; 而到 2020 年, 这一数字将达到 35ZB。人类所拥有的数据量还在以更快的速度增长, 2010 年 3 月, 视频网站 YouTube 宣布每分钟就会有 24 小时的视频被上传, 而到了 2010 年 11 月, 每分钟上传至 YouTube 的视频长度已达 35 小时。根据 YouTube 产品管理负责人的计算: “如果美国三大电视网每天播放 24 小时, 一周 7 天, 一年 365 天不间断播放 60 年, 那么这些视频内容才与 YouTube 每 30 天增加的内容一样多。” 而到了 2012 年 5 月, 每分钟上传的视频长度已经超过 72 小时, YouTube 上已经有超过一万亿个视频。

2012 年初, Royal Pingdom 网站给出 2011 年与互联网相关的一些统计数据:

- ◇ 全世界有 31.46 亿个电子邮件账户;
- ◇ 全世界有 5.55 亿个网站, 其中有 3 亿个是在 2011 年创建的;
- ◇ 全世界有 21 亿互联网用户;
- ◇ 3.5 亿用户使用移动设备登录互联网;
- ◇ 全世界有超过 24 亿个社交媒体账户;
- ◇ 全球有 26 亿个即时通信账户;
- ◇ 截至 2011 年 10 月, 互联网用户每月在线浏览视频量达到 2014 亿个;
- ◇ 截至 2011 年中期, Facebook 上有 1000 亿张照片;
- ◇ Flickr 上有 5100 万注册用户, 这些用户每天上传 450 万张照片。Flickr 上一共有 60 亿张照片。

很显然, 人类获取和生产数据的能力已经十分惊人, 当今的时代已经是一个“数据爆炸”的时代。为了应对数据爆炸性的增长, 最近二十年以来, 人类在数据存储能力上的进步极为迅速。二十年前, 我们使用的个人计算机往往只有 40MB 的硬盘, 数据交换依靠 720KB 的 5 英寸软盘和 1.44MB 的 3.5 英寸软盘。对于今天的个人计算机而言, 500GB 硬盘几乎成了标准配置, 用于数据交换的移动存储设备多为 250GB 以上移动硬盘和 2GB 以上的 U 盘。个人数据存储产品的容量在二十年间增大了成千上万倍。在二十年间, 数据中心更是从萌芽走向成熟, 当今的数据中心的存储规模往往能达到 PB 量级, 并且在能效、安全、接入和管理等方面有了越来越完善的考虑和设计。

数据传输技术的发展同样迅猛。一方面是依赖于移动存储介质的数据交换, 除了存储量的增大以外, 传输速率也飞速增长。传统的 1.44MB 软盘的传输速率为 62.5KB/s, 计算机串口的传输速率为 14.4KB/s。CD 光盘的读取速度为 7.5MB/s, DVD 光盘的读取速度为 16.6MB/s。现在得到广泛应用的 USB 2.0 理论传输速率为 60MB/s, 实际传输速率能达到 20~30MB/s; 2008 年底发布的 USB 3.0 标准理论传输速率已经达到了 600MB/s。因此基于移动存储介质的传输速率在十多年间也得到数百倍乃至数千倍的提升。互联网的发展使得数据传输不再受到地理位置的约束。早期的 Modem 拨号上网的速率为 7KB/s; 现在 ADSL 上网的下行速率可以达到 1MB/s, 目前家庭常用的速率为 512KB/s~2MB/s。而局域网的传输速率可以达到 10MB/s 甚至 100MB/s。而基于无线传输的移动互联网也可以提供 50~150KB/s 的下行速率。随着互联网, 特别是移动互联网的发展, 人们将继续向随时随地快速传输数据的目标前进。

数据的计算需要强大的处理能力，其中处理器和随机存储器起着至关重要的作用。二十年前的个人计算机，Intel 80386 的典型配置是 33MHz 主频和 1MB 内存；而今天的 Intel Core2 的典型配置是主频 3GHz、64KB 的一级缓存 (L1 cache) 和 6MB 的二级缓存 (L2 cache)；而 Intel Core-i 系列进一步引入了三级缓存，并实现了 CPU 与图形处理单元 GPU 的整合封装。因此今天的处理器，其计算能力已经不可同日而语。然而单处理器计算能力的提高仍然远远不能满足数据处理的需要，因此各种并行计算技术风起云涌，从多核处理器、图形加速器 GPU，并程序程序设计技术如 OpenMP、MPI，到分布式计算、网格计算和今天声名显赫的云计算，给数据计算提供了前所未有的强大能力。

然而数据的计算除了计算能力之外，同样甚至更为重要的是计算方法，因此近年来以机器学习、数据挖掘为代表的海量数据处理技术都得到了普遍的重视和迅速的发展。

数据的重要性导致了数据采集能力和生产速度不断提高，爆炸性的数据增长推动了数据存储、数据传输和数据计算能力和计算方法的飞速发展，而数据处理技术的发展进一步提升的数据的可用性和重要性，这就形成了一个正反馈，从而促使数据和数据处理相关的领域成为当今社会最有活力的发展方向。同时应该看到，相比于数据的采集、传输和存储，数据的计算能使人们更充分更有效地发挥数据的价值，因此我们有理由期待数据的计算有着更为广阔的发展空间。

1.1.3 数据的重要性

我们在上一节中把数据的重要性作为论述的基础，那么在这一节我们试图去回答数据为什么是重要的。数据是客观事物的符号表示，人类通过观察获得数据，通过数据积累和分析去获取知识，人类发展的历史同时也是一个数据积累和知识增长的过程。

远古时代，我们的祖先就为了生存去观察和适应环境，但在很长时间里人们缺乏描述客观事物的有效手段，观察的积累主要依靠个体进行。人们的知识主要来自直接的生活经验，信息的交流和保存非常困难，知识的积累缓慢而艰难。随着语言、文字的出现，以及保存文字的介质，如泥版和纸的发明，人们对客观事物的观察和认识能够以数据的形式保存、积累和交流，人类文明也进入了新的发展阶段。然而在很长的时期里，学习、掌握、传承和发展新的知识仍然主要依靠人类社会中的某些特殊群体，社会整体仍然处于愚昧落后的状态。17 世纪义务教育开始出现，18 世纪印刷术开始普及，书籍和报纸逐渐变得普遍起来，19 世纪电报和电话的诞生，20 世纪计算机和互联网的崛起，这些进步对于人类文明具有重要意义，它们推动了数据的产生、存储和传输，直接促进了信息和知识的分享。人们越来越多的知识来自于对数据的处理，而不是直接的生活经验。一个非常典型的例子是：16 世纪的丹麦天文学家第谷穷尽毕生精力，积累了大量准确细致的天文观测数据，就是在这些数据的基础上，他的学生，德国天体物理学家开普勒提出了著名的行星运动三大定律。可以说，正是第谷长期积累的精确数据加上开普勒创新的思想，对原始观测数据的尊重和有效利用，共同铸就了这一辉煌成就。

随着信息技术的发展和对社会的巨大影响，恐怕已经没有人会质疑信息的重要性，然后信息是不能单独存在的，数据是信息的载体。同样，数据尽管非常重要，但需要人们具备从数据中获取信息的能力，才可能有效地利用数据，真正发挥数据的作用。一方面，人们在

信息交互的时候,在传送数据的同时,还需要先验性的知识和约定,才能相互理解;另一方面,即使是同样的数据,不同人从不同的视角也能得到不同的信息和知识。有人说读书实际上是一个二次创作的过程,鲁迅先生在评价古典名著《红楼梦》时写道:“经学家看见易,道学家看见淫,才子看见缠绵,革命家看见排满,流言家看见宫闱秘事……”,恰好说明了这种现象。

人们从对自然和社会的观察中得到数据,这些数据中包含了人们知道和不知道的各种知识和规律,而这些未知的部分就是科学研究试图去发现的。

在以往的数据利用的过程中,人们往往是通过数据的观察产生假设,然后用数据进行验证。但在数据爆炸性增长的今天,很多时候数据中蕴含的规律很难通过观察直接获得,因此从海量数据自动发现其中蕴含规律和知识的数据挖掘技术开始崭露头角。

通过对客观事物的观察获得数据,依靠存储和传输积累数据,针对数据进行分析和思考总结规律,使用更多数据进行验证,这是人类认知的基本途径。因此数据和数据处理极为重要,数据的积累和数据处理能力的提高是人类文明发展的重要阶梯。

1.2 数据的表示

1.2.1 二元关系及其性质

数据的基本单元称为数据元素。数据是从对客观事物的观测中得到的,数据元素并不是孤立存在的,而是存在密切的联系,也因此才能表示和描述客观事物。数据元素之间的联系,归纳起来有三种,即一对一的联系、一对多的联系和多对多的联系。无论是哪种联系,都可以借助于“二元关系”进行描述,因此“二元关系”是描述数据元素关系的基础。

二元关系是一个数学概念,它定义在集合的基本运算——笛卡儿积(Cartesian product)的基础上。因此我们下面将从集合的笛卡儿积的定义出发,来解释二元关系的概念及其性质。

1.2.1.1 笛卡儿积

对于两个集合可以定义一种乘积运算,即集合的笛卡儿积。设有集合 M 和 N , 分别表示为 $M = \{x\}, N = \{y\}$, 则集合 M 和 N 的笛卡儿积, 记作: $M \times N$, 定义为

$$M \times N = \{(x, y) | x \in M \text{ 且 } y \in N\} \quad (1.1)$$

也就是说,两个集合 M 和 N 的笛卡儿积也是一个集合,这个集合 $M \times N$ 中的每个元素都是一个二元组,称为有序对或者序偶。有序对的第一个元素来自第一个集合 M , 有序对的第二个元素来自第二个集合 N 。笛卡儿积 $M \times N$ 将取遍集合 M 和 N 中所有元素的组合,如果集合 M 中的元素个数为 m , 集合 N 中的元素个数为 n , 那么 $M \times N$ 中的元素个数为 $m \times n$ 。

例: $M = (a_1, a_2), N = (0, 1, 2)$ 则

$$M \times N = \{(a_1, 0), (a_1, 1), (a_1, 2), (a_2, 0), (a_2, 1), (a_2, 2)\}$$

笛卡儿积的元素是有序对，因此集合的笛卡儿积是不可交换的，即

$$M \times N \neq N \times M \quad (1.2)$$

1.2.1.2 二元关系

有了集合的笛卡儿积，就可以进一步讨论二元关系。

定义：设有集合 M 、 N ，其笛卡儿积 $M \times N$ 的任意一个子集 $R \subset M \times N$ ，被称为 M 到 N 的一个二元关系。

二元关系表示了集合 M 和集合 N 中元素之间的某种相关性。若有序对 $(a, b) \in R$ ，也可以记为 aRb ，则称 a 是 b 的关于 R 的前件，或者说直接前驱； b 是 a 关于 R 的后件，或者说直接后继。

例如某学生学习语文，数学，外语，表示为 $M = \{\text{语文}, \text{数学}, \text{外语}\}$ 。

功课的成绩分为四个等级，记作 $N = \{A, B, C, D\}$ 。

这个学生成绩的全部可能为

$$\begin{aligned} M \times N = \{ & \langle \text{语文}, A \rangle, \langle \text{语文}, B \rangle, \langle \text{语文}, C \rangle, \langle \text{语文}, D \rangle, \\ & \langle \text{数学}, A \rangle, \langle \text{数学}, B \rangle, \langle \text{数学}, C \rangle, \langle \text{数学}, D \rangle, \\ & \langle \text{外语}, A \rangle, \langle \text{外语}, B \rangle, \langle \text{外语}, C \rangle, \langle \text{外语}, D \rangle \} \end{aligned}$$

如果这个学生的实际成绩 $R = \{\langle \text{语文}, B \rangle, \langle \text{数学}, A \rangle, \langle \text{外语}, D \rangle\}$ ，那么可以看到 R 是笛卡儿积 $M \times N$ 的一个子集，因此 R 是 M 到 N 的一个二元关系，它表示了这个学生功课和成绩的对应关系。

集合 M 和 N 的笛卡儿积 $M \times N$ 可以有很多个子集，也就是说从集合 M 到集合 N 可以定义若干种不同的关系，其数目由集合 M 和 N 的元素个数决定。但在实际应用中，我们所取的关系是其中很少的一部分，只有对我们所讨论问题有用的关系，我们才会去关心。

在实际应用的大多数情形下，两个集合是相同的，即 $M = N$ ，则二元关系 $R \subset M \times M$ 被称为 M 上的二元关系。

二元关系是普遍存在的，例如在实数域上相等关系“=”、小于等于关系“ $\leq$ ”、小于关系“ $<$ ”；集合的相等关系“=”、包含关系“ $\subseteq$ ”、真包含关系“ $\subset$ ”；平面上三角形的全等关系、相似关系；生活中的父子关系、同班同学关系等。

1.2.1.3 二元关系的性质

下面介绍二元关系 R 的基本性质。

设 R 是集合 M 上的一个二元关系：

(1) 如果对于集合 M 中的任一个元素 $a \in M$ ，有 $(a, a) \in R$ 成立，则称二元关系 R 是自反的。

(2) 如果对于集合 M 中的任一个元素 $a \in M$ ，都有 $(a, a) \notin R$ 成立，则称二元关系 R 是反自反的。

(3) 如果对于二元关系 R 中的任一个元素 $(a, b) \in R$ ，都有交换前件后件的元素 $(b, a) \in R$ 成立，则称二元关系 R 是对称的。

(4) 反对称性有两种表示方法：其一是：如果二元关系 R 中的任一个元素 $(a, b) \in R$ 且 $(b, a) \in R$ 时, 就有 $a = b$ 成立, 则称二元关系 R 是反对称的; 其二是：如果二元关系 R 中的任一个元素 $(a, b) \in R$, 且交换前件后件的元素 $(b, a) \notin R$, 则称二元关系 R 是反对称的。

(5) 当二元关系 R 中的任两个元素 $(a, b) \in R$ 和 $(b, c) \in R$ 时, 必有元素 $(a, c) \in R$ 成立, 则称二元关系 R 是传递的。

假设有集合 $M = \{1, 2, 3, 4, 5\}$, M 上的一个关系为

$$R = \{(1, 2), (2, 3), (3, 4), (3, 5), (1, 3), (1, 4), (1, 5), (2, 4), (2, 5)\}$$

根据上述性质的定义, 很容易看出该关系 R 是反自反的、反对称的和传递的。

利用二元关系的性质, 我们可以更方便、更准确地分析和描述数据集合中数据之间的关联规律和特点。

1.2.1.4 等价关系和等价类

如果非空集合 M 上二元关系 R 同时满足自反性、对称性和传递性, 则称这个二元关系 R 为集合 M 上的等价关系。如果 R 是 M 上的等价关系, a 和 b 是集合 M 中的任意元素, $a, b \in M$, 如果有 $(a, b) \in R$, 则称 a 等价于 b , 记作 $a \sim b$ 。

那么上面介绍的关系中有哪些是等价关系呢? 我们先来看实数域 $\mathbb{R}$ 上实数的相等关系。

设有实数 $a \in \mathbb{R}$, 则必有 $a = a$, 因此实数域上实数的相等关系满足自反性。

设有实数 $a, b \in \mathbb{R}$, 如果 $a = b$, 则必有 $b = a$ 成立, 因此实数域上实数的相等关系满足对称性。

设有实数 $a, b, c \in \mathbb{R}$, 如果满足 $a = b$ 和 $b = c$, 则必有 $a = c$ 成立, 因此实数域上实数的相等关系满足传递性。

由于实数域上实数的相等关系同时满足自反性、对称性和传递性, 所以实数域上实数的相等关系是等价关系。

同样可以分析, 集合的相等关系、平面上三角形的全等和相似关系都是等价关系。

同班同学关系也是一个等价关系, 因为:

- (1) 任何一个人都和自己在同一个班, 即满足自反性;
- (2) 若 a 和 b 是同一个班的, 那么 b 当然也和 a 在同一个班, 即满足对称性;
- (3) 如果 a 和 b 是同一个班的, b 和 c 也是同一个班的, 那么 a 和 c 一定是同一个班的, 即满足传递性。

实际上等价关系可以进一步引出等价类的概念。假设 R 是非空集合 M 上的等价关系, 则 M 中互相等价的元素构成了 M 中若干个互不相交的子集, 这些子集称为 M 的等价类。一个学校或者一个年级的同学就可以由同班同学这个等价关系划分为若干个互不相交的等价类。

下面给出等价类的一般定义:

设 R 是非空集合 M 上的等价关系, 对任意的 $a \in M$, 由 $[a]_R = \{b | b \in M \wedge aRb\}$ 给出的集合 $[a]_R$, 称为 a 关于 R 的等价类, 简称为 a 的等价类, 简记为 $[a]$ 。

在实际应用中,当需要根据某种关系来划分集合中的元素时,划分等价类是常用的方法。被划为等价类的数据元素具有一些共同的性质。

1.2.1.5 偏序和全序

如果二元关系 R 同时满足自反性、反对称性和传递性,则称这个二元关系 R 为集合 M 上的偏序关系 (partial order relation)。我们来看实数域 $\mathbb{R}$ 上实数的小于等于关系“ $\leq$ ”。

设有实数 $x \in \mathbb{R}$, 则必有 $x \leq x$, 因此实数域上实数的小于等于关系“ $\leq$ ”满足自反性。

设有实数 $a, b \in \mathbb{R}$, 如果同时满足 $a \leq b$ 和 $b \leq a$, 则必有 $a = b$ 成立, 因此实数域上实数的小于等于关系“ $\leq$ ”满足反对称性。

设有实数 $a, b, c \in \mathbb{R}$, 如果满足 $a \leq b$ 和 $b \leq c$, 则必有 $a \leq c$ 成立, 因此实数域上实数的小于等于关系“ $\leq$ ”满足传递性。

由于实数域上实数的小于等于关系“ $\leq$ ”同时满足自反性、反对称性和传递性, 所以实数域上实数的小于等于关系“ $\leq$ ”是偏序关系。前面我们提到的子集包含“ $\subseteq$ ”也是一种偏序关系, 读者可以自行验证。

如果二元关系 R 同时满足反自反性、反对称性和传递性, 则称这个二元关系 R 为集合 M 上的拟序关系 (quasi order relation), 也称为严格偏序关系 (strict partial order relation)。我们来看实数域 $\mathbb{R}$ 上实数的小于关系“ $<$ ”。

设有实数 $x \in \mathbb{R}$, 则 $x < x$ 不成立, 因此实数域上实数的小于关系“ $<$ ”满足反自反性。

设有实数 $a, b \in \mathbb{R}$, 如果满足 $a < b$, 则 $b < a$ 不成立, 因此实数域上实数的小于关系“ $<$ ”满足反对称性。

设有实数 $a, b, c \in \mathbb{R}$, 如果满足 $a < b$ 和 $b < c$, 则必有 $a < c$ 成立, 因此实数域上实数的小于关系“ $<$ ”满足传递性。

由于实数域上实数的小于关系“ $<$ ”同时满足反自反性、反对称性和传递性, 所以实数域上实数的小于关系“ $<$ ”是拟序关系。前面我们提到的子集真包含“ $\subset$ ”也是一种拟序关系, 读者可以自行验证。

R 是集合 M 上的一个二元关系, 那么二元关系 R 是笛卡儿积 $M \times M$ 的一个子集。假设二元关系 R 是 M 上的一个偏序关系, 并不是集合 M 中的任意两个元素 $a, b \in M$ 一定都是可比的, 也就是说存在笛卡儿积 $M \times M$ 的元素 $(a, b) \notin R$ 。

举例来说, 设集合 $M = \{\{a\}, \{a, b\}, \{a, c\}, \{a, b, c\}\}$, 集合 M 中每一个元素都是一个集合, 因此集合 M 是一个集合的集合。集合的包含关系“ $\subseteq$ ”是 M 上的一个偏序关系, 可以看到在集合 M 上, 既有 $\{a\} \subseteq \{a, b\} \subseteq \{a, b, c\}$ 成立, 也有 $\{a\} \subseteq \{a, c\} \subseteq \{a, b, c\}$ 成立, 但是集合 M 的元素 $\{a, b\}$ 和 $\{a, c\}$ 之间不存在子集包含的关系。

假设二元关系 R 是 M 上的一个偏序关系, 如果集合 M 中的任意两个元素 $a, b \in M$ 都可比, 或者有 aRb 成立, 或者有 bRa 成立, 则称 R 是集合 M 上的全序关系 (total order relation)。例如, 实数域 $\mathbb{R}$ 上的小于等于关系“ $\leq$ ”就是全序关系, 因为实数域 $\mathbb{R}$ 中的任意两个元素, 在小于等于关系下都是可比的。

全序关系一定是偏序关系, 但偏序关系不一定是全序关系。定义了全序关系的集合 M , 被称为全序集。

1.2.2 数据的逻辑结构

从二元关系出发就可以定义数据的逻辑结构，在更宏观的尺度上反映数据元素之间在逻辑上的联系。

定义二元组 $B = (D, R)$ ，其中 D 为数据元素的有限集合， R 为 D 上二元关系的有限集合。因此数据的逻辑结构由两个要素组成，一是数据元素的集合，二是集合上的二元关系。由此我们可以得到数据结构的基本形式，如图 1.1 所示，我们用圆圈表示元素，用连线表示关系。

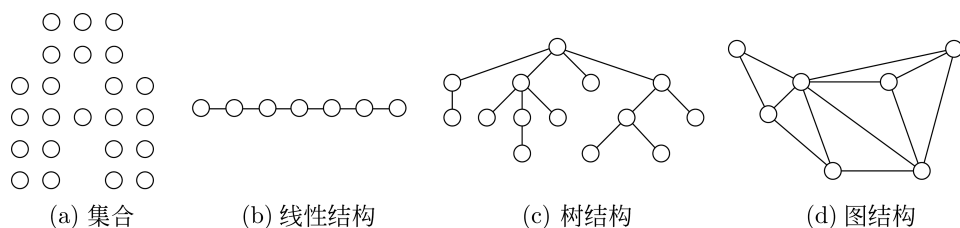


图 1.1 数据的逻辑结构

数据的逻辑结构主要有四种类型，一是集合，集合上没有定义二元关系；二是线性结构，每个元素有唯一的前驱元素和后继元素，这其实就是序列；三是树结构，每个元素可能和多个元素有关系；四是图结构，数据元素之间存在多对多的关系。

在实际应用中，这些逻辑结构既可能反映了数据本身具有的特性，也有可能是人们为了处理的需要而赋予数据的。只要与数据特性和问题需求不相矛盾，这种赋予就是合理的，也是必需的。

下面我们具体来看数据逻辑结构的一些例子。

1.2.2.1 数据集合

数据集合由数据元素组成，并且其上没有定义二元关系。

瓮模型 (urn model) 是概率论和随机过程中的经典模型。在一个瓮中有 R 个红球和 B 个蓝球： $U = \{Rred, Bblue\}$ ，每次从瓮中拿出一个球，观察其颜色然后放回瓮中，则每次取出红球的概率为 $P(red) = R/(R + B)$ ，取出蓝球的概率为 $P(blue) = B/(R + B)$ ，这是一个伯努利分布 (Bernoulli distribution)，各次取值相互独立。

在瓮模型中，红球和蓝球共同组成了一个集合，且没有定义二元关系，因此是一个数据集合模型的实例。

这是一个很简单的概率模型，但在其上加以变化，还可以定义复杂得多的过程模型。

1.2.2.2 线性结构

线性结构中每个元素最多只会有一个前驱元素，也最多只会有一个后继元素。

中世纪意大利数学家斐波那契在《算盘书》中描述了“兔子问题”。某人买一对小兔子，养殖在完全封闭的围墙内；如果每对兔子每月生一对小兔子，而小兔子出生后，二个月后具备生育能力，问一年后能繁衍到多少对？

按照问题给定的条件进行推导，本月的兔子对数等于上月已有的兔子对数加上本

月新增的兔子对数；考虑到每对兔子每月生一对小兔子，所以本月新增的兔子对数等于本月有生殖能力的兔子对数，也就等于再前一个月的兔子对数；假设每月的兔子对数为 $f(n)$, $n = 1, 2, \dots$ ，因此得到递推式：

$$f(n) = f(n-1) + f(n-2) \quad (1.3)$$

逐月递推，就可以得到表 1.1 中第二行所示结果，这就是兔子的对数随着时间增长的序列，这个序列就是著名的斐波那契数列。

表 1.1 兔子问题的解序列

月 份	0	1	2	3	4	5	6	7	8	9	10	11	12
兔子对数	1	1	2	3	5	8	13	21	34	55	89	144	233

市场上同一类型同一品牌的彩电，无论是早期的 CRT，还是后来的液晶和等离子，都是屏幕尺寸越大，价格越贵。如果我们按照屏幕尺寸从小到大，依次获取彩电的价格，也可以得到一个序列，这也构成一个线性结构。这里定义的二元关系是尺寸上的相邻关系，而不是时间上的相邻关系。

1.2.2.3 树结构

树结构最直观的例子来自于自然界的树，树从树根开始，逐步产生分支，直到叶子；在每个分支处都形成了一对多的关系。

随着生活水平的提高，人们对于食品和健康有了更多的关心。我们从《中国食物成分表 2012 年修正版》中随机选取了一些常见食物及其能量值，如表 1.2 所示，其中食物质量为 100 克，能量为对应的卡路里数。

表 1.2 常见食物的能量表

名称	能量	名称	能量	名称	能量	名称	能量	名称	能量
稻米	346	大白菜	21	草莓	30	香肠	508	蛋糕	347
挂面	344	菠菜	24	橙	47	叉烧肉	279	蛋黄酥	386
花卷	217	萝卜	20	梨	32	酱驴肉	246	凤尾酥	511
烙饼	255	韭菜	26	荔枝	70	牛肉	190	江米条	439
馒头	233	芹菜	14	芒果	32	兔肉	102	麻团	512
糯米	344	生菜	13	苹果	52	羊肉	198	麻花	524
小米	358	西兰花	33	葡萄	43	羊肉串	217	桃酥	481
玉米	336	黄瓜	15	桃	48	猪肉	395	月饼	428
面包	312	青椒	22	香蕉	91	北京烤鸭	436	奶油饼干	429
烧饼	326	番茄	19	西瓜	34	炸鸡	279	巧克力	586

我们可以计算所有这些食物的能量均值和方差。假设上述食物的能量为 $(e_1, e_2, \dots, e_n)$ ，则食物的能量均值为 $\bar{e} = \frac{1}{n} \sum e_i = 217.54$ ；而方差为 $v = \frac{1}{n} \sum (e_i - \bar{e})^2 = 1471386$ ，反映了不同食物之间能量的巨大差异。