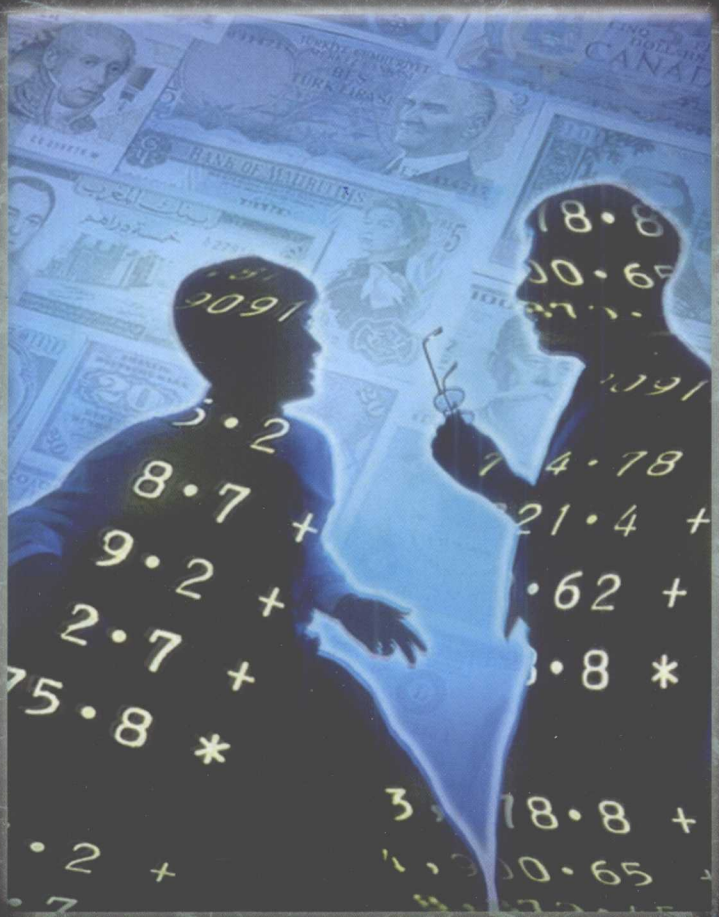
数据学

朱扬勇 熊赟 著

Dataology and Data Science

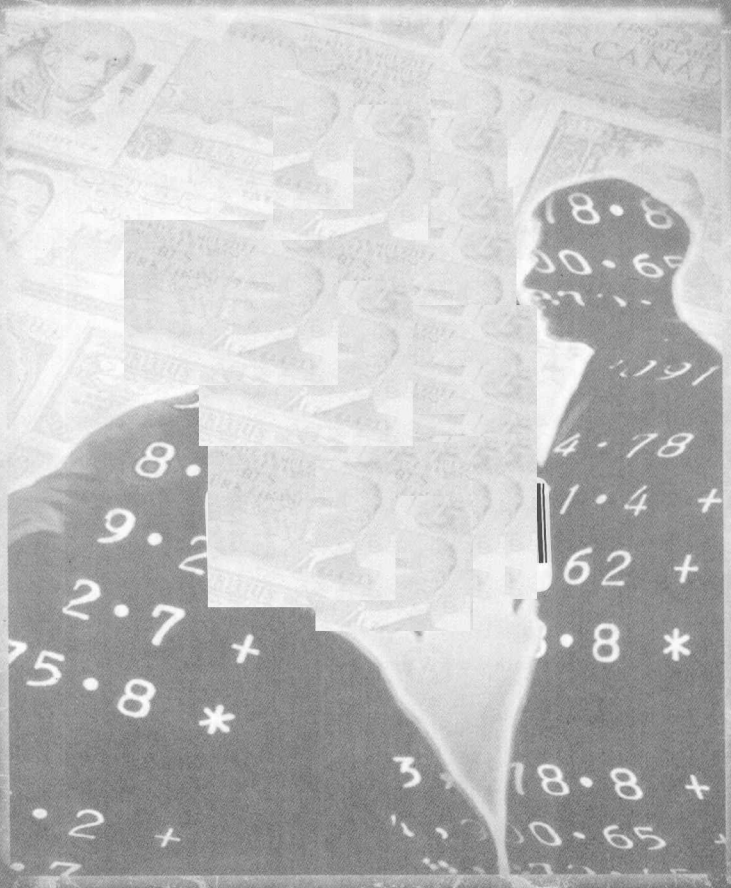


复旦大学出版社
www.fudanpress.com.cn

数据学

朱扬勇 熊赟 著

Dataology and Data Science



复旦大学出版社
www.fudanpress.com.cn

图书在版编目(CIP)数据

数据学/朱扬勇,熊赟著. —上海:复旦大学出版社,2009.12
ISBN 978-7-309-06956-3

Ⅰ. 数… Ⅱ. ①朱…②熊… Ⅲ. 数据管理-基本知识 Ⅳ. TP311.13

中国版本图书馆 CIP 数据核字(2009)第 203090 号

数据学

朱扬勇 熊 赟 著

出版发行 复旦大学出版社 上海市国权路 579 号 邮编 200433
86-21-65642857(门市零售)
86-21-65100562(团体订购) 86-21-65109143(外埠邮购)
fupnet@fudanpress.com http://www.fudanpress.com

责任编辑 范仁梅
出 品 人 贺圣遂

印 刷 上海浦东北联印刷厂
开 本 787×960 1/16
印 张 8.75
字 数 113 千
版 次 2009 年 12 月第一版第一次印刷

书 号 ISBN 978-7-309-06956-3/T·350
定 价 22.00 元

如有印装质量问题,请向复旦大学出版社发行部调换。
版权所有 侵权必究

内 容 提 要

本书介绍了数据学的由来、基本概念和基本原理，包括：数据大爆炸、数据自然界、数据学基础；介绍了数据学的主要方法，包括：数据勘探、数据获取与整合、数据挖掘、数据实验；还介绍了数据学的应用和数据学面临的挑战。

本书主要作为科学研究工作者的参考书，试图向科学工作者展示一种新的科学，并且能够利用这种新的科学为当前的科学研究工作服务。本书基本以较为通俗化的语言来表达数据学的基本原理、方法和技术，希望对各领域的科学家，包括行为科学家和社会科学家能够有所启迪。

Abstract

The large scale of data is rapidly generated and stored in computer systems, which is called data explosion. Data explosion forms data nature in computer systems. Dataology (also called data science or science of data) is an umbrella of theories, methods and technologies for studying data nature.

This book is the first book to introduce the concepts, theories and principles of dataology systematically.

More detail please visit 'http://www.dataology.fudan.edu.cn'.

前　言

科学研究所用的方法是逻辑推理和实验，逻辑推理依靠数学，实验依靠观测。在计算机出现后，科学研究开始使用计算机技术，使得科学研究增加了计算的方法。在自然科学的研究过程中，遇到了大量的计算问题，这些计算是手工无法完成的，因此计算机的大规模计算能力在科学计算方面获得了很好的应用，并逐渐成为一种科学研究的新方法，相继出现了计算数学、计算物理、计算化学、计算生物学，等等，并且计算生物学已经成为现代生物学研究的核心方法之一。后来，科学研究的对象也信息化了，变成了计算机中的数据，最典型的是生命科学领域中基因的信息化，形成由 ACGT 这 4 个字母组成的 DNA 序列数据，研究对象变成了 DNA 序列数据，出现了生物信息学。同样的状况也出现在其他研究领域，于是有了脑信息学、地理信息学、行为信息学、社会信息学、经济信息学、历史信息学，等等。

随着国民经济和社会的信息化进程，自然界中的事物以数据的形式存储到计算机系统中，即信息化是一个生产数据的过程。这些数据是自然和生命的一种表示形式，这记录了人类的行为，包括工作、生活和社会发展。数据被快速大量地生产并存储到计算机系统中，这种现象称为数据爆炸，数据爆炸在计算机系统中形成数据自然界。目前，数据爆炸还在进行中，人类还不能清晰地描述数据自然界，但已经在其中工作和生活

此为试读，需要完整PDF请访问：www.ertongbook.com

了。在数据自然界中，人类将面临许多新的问题，例如：人类不知道从互联网上获得的数据是否是正确的和真实的；也许网络中某个数据库早已显示人类将面临能源危机，但人们却无法得到这样的知识；等等。

面对数据自然界，面对科学研究方法和研究对象的变革，需要有新的科学技术，称其为数据学(dataology)或数据科学(data science)。数据学定义为探索数据自然界奥秘的理论、方法和技术。数据学将为许多(也许是所有)学科和领域的科学研究提供基础理论和方法，包括：数据勘探、数据实验、数据获取与整合、数据挖掘，等等。数据学方法和技术将会被应用于许多领域，开发出专门的理论、技术和方法，从而形成专门领域的数据学。

本书介绍了数据学的由来、基本概念和基本原理，包括：数据大爆炸、数据自然界、数据学基础；介绍了数据学的主要方法，包括：数据勘探、数据获取与整合、数据挖掘、数据实验；还介绍了数据学的应用和数据学面临的挑战。由于本书介绍的数据学还没有形成严谨的科学体系，因此还存在很多问题。我们将数据学介绍给读者，是希望能尽早引起科学界的关注，希望在批评、争论、质疑的过程中发展数据学。

感谢能倾听我们叙述并认真提出建议的科学家们，他们是李亦学教授、钟扬教授、石勇教授、钟宁教授、操龙兵教授等。尤其感谢钟宁教授的合作，使我们完成了第一篇数据学论文。感谢他们的鼓励，使我们有勇气来写这本书。也感谢复旦大学出版社，尤其感谢范仁梅女士，在她的理解和支持下本书得以出版。

恳切希望读者批评指正。

朱扬勇 熊 赟

2009年10月11日于复旦

目　录

此为试读，需要完整PDF请访问：www.ertongbook.com

第1章 绪　论

信息化的本质是将现实世界中的事物以数据的形式存储到计算机系统中，即信息化是一个生产数据的过程。这些数据是自然和生命的一种表示形式，这些数据还记录了人类的行为，包括工作、生活和社会发展。今天，数据被快速大量地生产并存储在计算机系统中，这种现象称为数据爆炸(data explosion)。数据爆炸在计算机系统中形成数据自然界(data nature)。研究数据自然界是研究自然界(real nature)的一种有效方法，例如：可以通过研究数据来研究生命(生物信息学)、研究人类行为(行为信息学)。数据学(dataology)或数据科学(data science)是探索数据自然界奥秘的理论、方法和技术。

本章介绍了数据爆炸、数据自然界和数据学的基本概念，并给出了数据学[ZZX09]的定义及其基本框架。

1.1　数据

计算机系统中存放的是数据，“数据”的含义很广，不仅指 1011、8084 这样一些数字，还指“dataology”、“小舟扬帆出海”、“11/11/11”等符号、字符、日期形式的数据。确切地说，本书讨论的数据是指能够输入到计算机中的任何东西，如：数字、字符、声音、图像、照片，等等，并且处理数据的计算机程序本身也是“数据”。

1.1.1 数据的概念

数据在物理上以字节(Byte)作为其大小的计量单位,一个字节为一个数据单位,数据物理存在于计算机系统中。

数据原子(data atomic) 是不可再分割的最小数据单位,是计算机系统所使用的基本字符集。数据原子一般为单个字节字符,例如 ASCII 码字符表里的字符;也有一些是双字节字符,例如 CJK 字库中的字符。

数据对象(data object) 是识别数据的基本单位,是可命名的,具有独立含义。一个数据对象由有限个数据项组成,必须要有一个对象标识,其他为对象内容。数据项(data item)是一个数据原子的有限集,用于描述数据对象的特性,也是可命名的,并且可以定义其数据类型,但没有独立含义,即脱离数据对象单独讨论数据项是没有意义的。

数据集(data set) 是数据对象的集合。一般情况下,数据集是一个数据对象的有限集合,虽然也有一些无限的数据集需要处理,例如:流数据,但数据任何时刻都是有限的,所以数据学通常是处理有限数据集的。

数据(data) 是数据原子、数据项、数据对象和数据集的统称,可以用一个数据表示一个数据原子、一个数据项、一个数据对象或者一个数据集。数据的大小用数据单位 Byte 来表示,NULL 表示空数据,其大小为 0 Byte。

元数据(mata data) 是描述数据的数据。例如,ASCII 表结构是用来表述数据原子的,NAME 是用来描述数据项的,EMPLOYEE (ID, NAME, RANK)是用来描述职工数据的,DATABASE#1{table1 (a1, a2, …), table2 (b1, b2, …), ….}是用来描述一个数据集的。

对于数据自然界,计算机系统是它的载体,数据是它的唯一存在,为了避免出现"处理数据的数据,或被数据处理的数据"这样的叙述而陷入表达混乱,在需要的时候将计算机程序称为数据工具。

数据工具(data tool) 是计算机系统中存储的能够运行的计算机

程序或软件系统，是一种特殊的数据对象。数据工具通常用于处理数据，但数据工具本身也是数据，可以被其他数据工具处理，例如，杀毒软件是一个数据工具，它用于处理另一个数据工具“病毒程序”，而“病毒程序”还能自己将自己复制传播，即自己处理自己。

1.1.2 数据与物质

数据和自然界中的物质都是存在的，但数据的存在和自然界中物质的存在是非常不同的。主要的不同点表现在可标识性、可共享性和生命周期性等3个方面。

1. 可标识性方面

自然界中的物质，一个是一个，所谓相同的两个东西是同质化的两个东西，例如，面对两杯水，可以说“相同的两杯水”；而对于数据，一个数据的存在和两个相同数据的存在是一样的，“两个相同的数据”的说法意义不大，“两个相同的数据”是表示自然界的一个事物，即是一个数据，一般采用“一个数据的两个复本”的说法。关于数据，讨论数据的相似性比讨论数据的相同性更有意义，相似性由相似性函数定义，可以说“两个相似的数据”。

数据的这种特性说明数据是面向值的，即：如果有两个数据对象有相同的值，则认为它们是一个对象的两个复本。

2. 可共享性方面

共享(share)是指共同分享，在物理世界中主要是指某样东西被多个人分。例如，“共享午餐”是指共享者一起吃午餐，但其实每个共享者吃的东西并不一样，同样的东西不可能被吃进两个人的肚子里。

数据共享的概念有着本质上的不同，数据共享是指同样的数据被多个共享者所拥有，并且每个拥有者具有完全一样的数据量、数据形式和数据内容，即拥有数据的复本。将一个数据复制随意多个复本是轻而易举的事情，因此，数据是可以共享的，并且拥有数据的人常常也愿

意将其拥有的数据拿出来共享。

3. 生命周期性方面

自然界中的物质会老化，有生命周期，而数据不会老化，没有生命周期。数据就其被生产、被存储、被修改、被删除这些过程而言是有生命周期的，但这是该数据在现实中对应的事物的生命周期，不是计算机系统中数据的生命周期。一个数据本身不会随时间的推移而变老变旧，例如，将一张照片数据存放多少年以后，只要载体还存在或者不断替换新载体，这个数据对象本身不会发生变化，数据不会减少，其质量也不会下降。

1.2 数据爆炸

数据爆炸在很多地方也称为信息爆炸(information explosion)，本书采用“数据爆炸”的说法。

数据爆炸随着人类的发展不断前进。试图记住已知的东西是人的天性，从古到今用大脑记住所经历的事物一直是人类生存的最主要手段。由于未知的原因，人并不能做到过目不忘，因此大脑的记忆也不可靠。于是，人类一直寻求辅助的设施来帮助记忆(所有动物都一样)。早先，人在硬物上刻图、刻字来帮助记忆，很快，人就发现被记录在大脑之外的事情可以很容易用来传播和交流，从而更加深化了人记录事情的这种天性。

在印刷术和造纸术被发明后，人类以文字和图来记录事情，这时发生了第一次数据爆炸(将印刷在纸上的文字和图看成数据)。在第一次数据爆炸期间，大量自然界的事物(自然现象、人文、社会等)被文字和图表示，然后印刷成书、材料。这些数据(文字和图)被长期保存，并被大量复制，然后被广泛传播。这期间的作者、出版社是生产信息的，图

书、图书馆则是存储和传播信息的。之前一段历史的事物或事情则可能被写到一本书里而被存储和传播，如《圣经》、《史记》，等等。

当计算机及其存储设备被发明后，人类以二进制数位的形式在纸带、磁盘、磁带、光盘、闪存盘上记录事情，这时发生了第二次数据爆炸。它是指人们在使用和应用计算机系统过程中不断向计算机系统存入数据，导致计算机系统中数据成爆炸式增长的过程。这期间，一座图书馆（第一次数据爆炸的主要产物之一）中所有图书所记录的数据便可能被存储在一部个人电脑中，甚至被存储在一块移动硬盘中。在第二次数据爆炸过程中，第一次数据爆炸期间的出版社、报社将逐渐消亡。这也进一步说明，第二次数据爆炸和第一次数据爆炸不同，同时也说明技术的进步使得帮助人类记忆的设施获得了重大发展。

第二次数据爆炸与第一次数据爆炸的一个重要不同是，记录在磁、光、电介质上的数据更容易被共享，能通过计算机网络传播，并且比纸介质传播得更快。

目前，还没有迹象表明有新的设备能取代计算机及其存储设备，也许将来会有某种人造生物，这种人造生物具有“过目不忘”和极强的处理能力，那时人记录的数据将会发生第三次爆炸。

数据爆炸使人们迷失在海量数据中而无所适从，主要表现为以下几点：

（1）数据的正确性和真实性没有保障：人们不知道从计算机系统中获得的数据（如从互联网上看到的网页数据）是否是正确的和真实的。这将导致：虽然数据很多，但不知道哪些数据是可用的。由于人们越来越多地从计算机系统中获得数据，因此如果无法判断数据的真实性，那么获得的数据所表达的信息可能具有误导性，相应地，获得的知识可能是错误的或是过时的。

（2）数据共享越来越困难：虽然提供数据共享是计算机系统的目的之一，但数据共享却越来越困难，在每天产生巨量数据的情况下，人们

不知道要共享什么？也不知道如何使这些数据得以共享，更不知道共享的数据是否会带来副作用。

(3) 数据的一致性越来越困难：难以做到数据的一致性。例如，分别在两个网站对相同的目标进行搜索，得到的结果却不一样。

(4) 数据综合征：人们不停地生产数据，不停地存储数据，不停地从互联网中读取数据，包括网络上瘾、游戏上瘾等，无法从数据中解脱出来。

出现上述这些状况的原因是，数据已经快速膨胀，人类已经无法控制数据，也越来越难认识数据。计算机系统中的海量数据已经形成了一个数据自然界。

1.3 数据自然界

人类社会的进步发展是人类不断探索自然(宇宙和生命)的过程，当人们将探索自然界的成果存储在计算机系统中时，却在不知不觉中创造了一个数据自然界。虽然是人制造了数据，并且人还在不断制造数据的过程中，但当前的数据已经表现出不为人控制，具有未知性、多样性和复杂性等自然界特征。

数据自然界 所有计算机系统中的数据构成了数据自然界，而计算机系统是数据的载体，因此并不将其作为数据自然界的组成部分。

1.3.1 数据不为人类所控制

今天，数据呈爆炸式增长，人们已经无法控制它，除此之外，还有计算机病毒大量出现和传播，垃圾邮件泛滥，网络攻击，数据阻塞信息高速公路，等等，使得人们无法控制数据。

在现在的日常生活中，人们在不断生产数据，不但使用计算机产生

此为试读，需要完整PDF请访问：www.ertongbook.com