陈 明◎编著

大数据基础与应用

◆ 大数据技术与教育**扛鼎之作**

◆ 全面系统深入讲述**大数据**

北京师范大学出版集团
BEIJING NORMAL UNIVERSITY PUBLISHING GROUP
北京师范大学出版社

陈 明◎编著

大数据基础与应用

北京师范大学出版集团
BEIJING NORMAL UNIVERSITY PUBLISHING GROUP
北京师范大学出版社

图书在版编目（CIP）数据

大数据基础与应用/陈明编著. —北京：北京师范大学出版社，2016
ISBN 978-7-303-20018-4

I. ①大… II. ①陈… III. ①数据处理 IV. ①TP274

中国版本图书馆 CIP 数据核字（2016）第 005840 号

营 销 中 心 电 话 010-62978190 62979006
北师大出版社科技与经管分社网 http://jsws.bnupg.com
电 子 信 箱 kjjg@bnupg.com

出版发行：北京师范大学出版社 www.bnupg.com
北京新街口外大街 19 号
邮政编码：100875
印 刷：北京京师印务有限公司
经 销：全国新华书店
开 本：184 mm×260 mm 1/16
印 张：21.25
字 数：504 千字
版 次：2016 年 3 月第 1 版
印 次：2016 年 3 月第 1 次印刷
定 价：42.00 元

策划编辑：李 丹 责任编辑：李 丹
美术编辑：刘 超 装帧设计：刘 超
责任校对：马军令 责任印制：曲利华

前　言

需求是科学技术发展的原动力。大数据问题的出现与研究已经成为计算机科学与技术研究的新热点，并显示出日益强大的吸引力，科学大数据的出现催生了数据密集型知识发现的第四科学研究范式的出现。对于信息领域，大数据带来的不仅是机遇，还有一系列的困难和挑战。大数据技术与应用展现出锐不可当的强大生命力，科学界与企业界对其寄予无比的厚望。大数据成为继 20 世纪末、21 世纪初互联网蓬勃发展以来的又一轮 IT 工业革命。数据本身是无意义的，而通过统计、分类、萃取、特征抽取等一系列技术手段，可以从数据中提取信息与知识。所以说，数据是重要的战略资源，隐含巨大的经济价值，已经引起科技界和企业界的高度重视。有效地组织和使用数据，将对经济发展产生巨大的推动作用。大数据是以大样本或全样本代替抽样，以近似代替准确，以联系代替因果，因此，大数据是对传统的 IT 各领域的挑战，研究大数据意义深远。大数据孕育着前所未有的机遇。对大数据的交换、整合和分析，可以发现新的知识，创造新的价值，带来大知识、大科技、大利润和大发展。

本书概括性地介绍了大数据的主要内容，是大数据技术入门的参考书。全书分为 17 章，其中第 1 章是对大数据的简单概述，第 2 章介绍大数据研究的方法论，第 3 章、第 8 章、第 9 章和第 14 章介绍大数据的生态环境，第 4 章、第 5 章、第 6 章、第 7 章、第 10 章、第 11 章、第 12 章、第 13 章、第 15 章和第 16 章介绍大数据技术及应用方法，第 17 章简单介绍数据科学的内容。

大数据的规模、活性及解释运用的能力将成为综合国力的重要组成部分，对数据的占有和控制将成为陆权、海权、空权之外的另一种国家核心资产。联合国在 2012 年发布了《大数据政务白皮书》，指出大数据对于联合国和各国政府是一个历史性的机遇，通过使用极为丰富的数据资源，对社会经济进行前所未有的实时分析，帮助政府更好地调整社会和经济运行。数据为王的大数据时代已经到来，对数据的占有和控制也将成为新的争夺点。大数据技术的专业人才，特别是数据分析复合型人才的稀缺将会影响市场的发展。

在技术上，大数据、海量数据与超大规模数据都是指用传统处理方法无法处理的大量数据。通过对大数据进行高速有效的处理，可以发现数据中蕴藏的规律，进而为各种关键决策提供依据与指导。正确的预测与决策将导致巨大财富的产生。技术与工具密不可分，目前常用的数据处理技术与工具适用于小数据处理，一些海量数据处理方法与工具是过渡性的，大数据处理技术与工具的研究是一项有理论意义和实际价值的工作。大数据的出现，对 IT 各领域的传统处理方法提出了新的冲击与挑战。大数据技术是一门实践性较强的技术，需要重视工具与应用方法的选择与研究。

高校是培养人才的重要基地，需要分析和定位大数据带来的影响。大数据推动学科发展和学科建设、利用大数据技术整合现有的教学资源为已经到来的“数据革命”培养行业紧缺人才都是教育工作面临和需要解决的问题。

本书在结构上呈积木状，对各章内容进行独立的概念性论述。由于作者水平有限，书中不足之处在所难免，敬请读者批评指正。

陈明

2015 年 10 月

目　录

第 1 章　走进大数据时代

本章主要内容

需求是科学技术发展的原动力，大数据中蕴含着巨大而重要的价值，大数据的研究已经成为计算机科学与技术研究的新热点，并显示出日益强大的生命力和吸引力。科学大数据催生了数据密集型知识发现的第四科学研究范式的出现。大数据的出现，不仅带来了机遇，也带来了困难和挑战。无论是科学界，还是企业界，都对大数据所带来的巨大冲击寄予厚望，大数据是继 20 世纪末、21 世纪初互联网蓬勃发展以来的又一轮新的 IT 工业革命。

1.1 应对大数据

在全球范围内，以电子方式存储的数据（简称为电子数据）总量空前巨大。2011 年，电子数据总量达到 1.8ZB，比 2010 年同期提高了 1ZB，统计结果表明，每经过两年就可以增加一倍，预计到 2020 年可达到 35ZB，其预测曲线如图 1-1 所示。面对数据增长速度飞快地提升，数据量狂增，对大量电子数据的高效存储、高效传输与快速处理成为必须解决的问题。

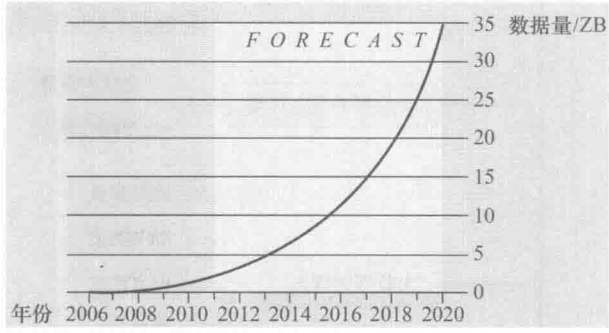

图 1-1 全球数据总量的预测示意图

1.1.1 电子数据迅速增加

云计算、物联网、移动互联网、手机、平板电脑和 PC 机中的数据，各种公开的数据、大型电子商务数据、遍布全球的各种传感器数据、工业蓬勃发展产生的工业数据、大型科学研究设备产生的数据，以及社交媒体的快速发展，构成了大数据持续产生的基本生态环境。尤其是近年来，随着互联网技术的发展，来自人们的日常生活，特别是来自互联网+X 服务的数据迅猛增加。据不完全统计，互联网当前包含 93 亿多个页面，80%～85%的数据存储在数据库的文本中。互联网一天产生的全部内容可以刻满 1.68 亿张 DVD，发出的电子邮件有 2940 亿封之多，发出的社区帖子达 200 万个，相当于《时代》杂志 770 年的文字量。从统计角度来看，由于电子数据量迅速增加，预计中国数据技术和服务市场未来 5 年的复合增长率将达 51.4%，其中增长率最高的是存储市场，将达 60.8%，服务器市场的增长率则是 38.3%，远远高于其他产品相关的市场。

1.1.2 数据中蕴含的价值

数据本身是无意义的，但是，有效地组织和使用数据，即对大数据进行交换、整合和分析，可以发现新的知识、创造新的价值、推动科技大发展，凸显了数据是重要的战略资源。

据统计，2012 年中国市场规模达到 4.5 亿元，2016 年估计可达到上百亿元的规模，如图 1-2 所示。

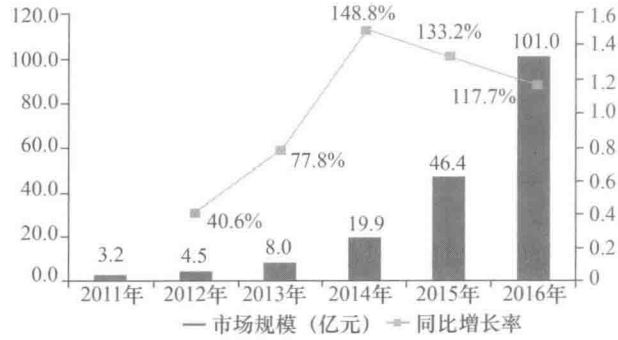

图 1-2　中国大数据应用市场规模与增长的预测

1.1.3　数据是国家的核心资产

一个国家拥有数据的规模及分析获取价值的能力将成为综合国力的重要组成部分，存储与控制的数据将成为国家陆权、海权和空权之外的另一个核心资产。

联合国在 2012 年发布了《大数据政务白皮书》，指出大数据对于联合国和各国政府来说是一个历史性的机遇，通过使用丰富的数据资源，对社会经济进行前所未有的实时分析，能够帮助政府更好地响应社会和经济运行。对数据的占有和控制也将成为国家之间和企业之间的新的争夺点。社会需要大量大数据技术的专业人才，特别是数据分析复合型人才，对这一专业人才的培养，将是一项十分重要的任务。

1.2　大数据的生态环境

大数据是人类活动的产物，它来自人们改造客观世界的过程中，是生产与生活在网络空间的投影。信息爆炸是对信息快速发展的一种逼真的描述，形容信息发展的速度如同爆炸一般席卷整个空间。在 20 世纪 40～50 年代，信息爆炸主要指的是科学文献的快速增长。而经过 50 年的发展，到 20 世纪 90 年代，由于计算机和通信技术的广泛应用，信息爆炸主要指的是所有社会信息快速增长，包括正式交流过程和非正式交流过程所产生的电子式的和非电子式的信息。而到 21 世纪的今天，信息爆炸是由于数据洪流的产生和发展所造成的。在技术方面，新型的硬件与数据中心、分布式计算、云计算、高性能计算、大容量数据存储与处理技术、社会化网络、移动终端设备、多样化的数据采集方式使大数据的产生和记录成为可能。在用户方面，日益人性化的用户界面、信息行为模式等都容易作为数据量化而被记录，用户既可以成为数据的制造者，又可以成为数据的使用者。可以看出，随着云计算、物联网计算和移动计算的发展，世界上所产生的新数据，包括位置、状态、思考、过程和行动等数据都能够汇入数据洪流，互联网的广泛应用，尤其是“互联网+”的出现，促进了数据洪流的发展。

归纳起来，大数据主要来自互联网世界与物理世界。

1.2.1 互联网世界

大数据是计算机和互联网相结合的产物，计算机实现了数据的数字化，互联网实现了数据的网络化，两者结合起来之后，赋予了大数据强大的生命力。随着互联网如同空气、水、电一样无处不在地渗透人们的工作和生活，以及移动互联网、物联网、可穿戴联网设备的普及，新的数据正在以指数级加速产生，目前世界上 90%的数据是互联网出现之后迅速产生的。来自互联网的网络大数据是指“人、机、物”三元世界在网络空间（Cyberspace）中交互、融合所产生并可在互联网上获得的大数据，网络大数据的规模和复杂度的增长超出了硬件能力增长的摩尔定律。

大数据来自人类社会，尤其是互联网的发展为数据的存储、传输与应用创造了基础与环境。依据基于唯象假设的六度分隔理论而建立的社交网络服务（Social Network Service，SNS），以认识朋友的朋友为基础，扩展自己的人脉。基于 Web 2.0 交互网站建立的社交网络，用户既是网站信息的使用者，也是网站信息的制作者。社交网站记录人们之间的交互，搜索引擎记录人们的搜索行为和搜索结果，电子商务网站记录人们购买商品的喜好，微博网站记录人们所产生的即时的想法和意见，图片视频分享网站记录人们的视觉观察，百科全书网站记录人们对抽象概念的认识，幻灯片分享网站记录人们的各种正式和非正式的演讲发言，机构知识库和期刊记录学术研究成果等。归纳起来，来自互联网的数据可以划分为下述几种类型。

1. 视频图像

视频图像是大数据的主要来源之一，电影、电视节目可以产生大量的视频图像，各种室内外的视频摄像头昼夜不停地产生巨量的视频图像。视频图像以每秒几十帧的速度连续记录运动着的物体，一个小时的标准清晰视频经过压缩后，所需的存储空间为 GB 数量级，对于高清晰度视频所需的存储空间就更大了。

2. 图片与照片

图片与照片也是大数据的主要来源之一，截至 2011 年 9 月，用户向脸书（Facebook，美国的一个社会网络服务网站）上传了 1400 亿张以上的照片。如果拍摄者为了保存拍摄时的原始文件，平均每张照片大小为 1MB，则这些照片的总数据量约为 $1.4\times10^{12}\times1\text{MB}=140\text{PB}$，如果单台服务器磁盘容量为 10TB，则存储这些照片需要 14000 台服务器，而且这些上传的照片仅仅是人们拍摄到的照片的很少一部分。此外，许多遥感系统 24 小时不停地拍摄并产生大量照片。

3. 音频

DVD 光盘采用了双声道 16 位采样，采样频率为 44.1kHz，可达到多媒体欣赏水平。如果某音乐剧的时间为 5.5min，计算其占用的存储容量为：

$$\begin{aligned}\text{存储容量}&=(\text{采样频率}\times\text{采样位数}\times\text{声道数}\times\text{时间})/8\\&=(44.1\times1000\times16\times2\times5.5\times60)/8\\&\approx55.5\text{MB}\end{aligned}$$

4. 日志

网络设备、系统及服务程序等，在运作时都会产生 log 的事件记录。每一行日志都记载着日期、时间、使用者及动作等相关操作的描述。Windows 网络操作系统设有各种各样的日志文件，如应用程序日志、安全日志、系统日志、Scheduler 服务日志、FTP 日志、WWW 日志、DNS 服务器日志等，这些根据系统开启的服务的不同而有所不同。用户在系统上进行一些操作时，这些日志文件通常记录了用户操作的一些相关内容，这些内容对系统安全工作人员相当有用。例如，有人对系统进行了 IPC 探测，系统就会在安全日志里迅速地记下探测者探测时所用的 IP、时间、用户名等，用 FTP 探测后，就会在 FTP 日志中记下 IP、时间、探测所用的用户名等。

网站日志记录了用户对网站的访问，电信日志记录了用户拨打和接听电话的信息，假设有 5 亿用户，每个用户每天呼入呼出 10 次，每条日志占用 400B，并且需要保存 5 年，则数据总量为 5×10×365×400×5 Byte≈3.65PB。

5. 网页

网页是构成网站的基本元素，是承载各种网站应用的平台。通俗地说，网站就是由网页组成的，如果只有域名和虚拟主机而没有制作任何网页，客户仍旧无法访问网站。网页要通过网页浏览器来阅读。文字与图片是构成一个网页的两个最基本的元素。可以简单地理解为：文字就是网页的内容，图片就是网页的美观描述。除此之外，网页的元素还包括动画、音乐、程序等。

网页分为静态网页和动态网页。静态网页的内容是预先确定的，并存储在 Web 服务器或者本地计算机、服务器之上，动态网页取决于用户提供的参数，并根据存储在数据库中的网站上的数据而创建。通俗地讲，静态页是照片，每个人看都是一样的，而动态页则是镜子，不同的人（不同的参数）看都不相同。

网页中的主要元素有感知信息、互动媒体和内部信息等。感知信息主要包括文本、图像、动画、声音、视频、表格、导航栏、交互式表单等。互动媒体主要包括交互式文本、互动插图、按钮、超链接等。内部信息主要包括注释，通过超链接链接到某文件、元数据与语义的元信息，字符集信息，文件类型描述，样式信息和脚本等。

网页内容丰富，数据量巨大，每个网页有 25KB 数据，则一万亿个网页的数据总量约为 25PB。

1.2.2　物理世界

来自物理世界的大数据又被称为科学大数据，科学大数据主要来自大型国际实验：跨实验室、单一实验室或个人观察实验所得到的科学实验数据或传感数据。最早提出大数据概念的学科是天文学和基因学，这两个学科从诞生之日起就依赖于基于海量数据的分析方法。由于科学实验是科技人员设计的，数据采集和数据处理也是事先设计的，所以不管是检索还是模式识别，都有科学规律可循。例如希格斯粒子，又称为“上帝粒子”的寻找，

采用了大型强子对撞机实验。这是一个典型的基于大数据的科学实验，至少要在 1 万亿个事例中才可能找出一个希格斯粒子。从这一实验可以看出，科学实验的大数据处理是整个实验的一个预定步骤，这是一个有规律的设计，发现有价值的信息可在预料之中。大型强子对撞机每秒生成的数据量约为 1PB。建设中的下一代巨型射电望远镜阵每天生成的数据量大约在 1EB。波音发动机上的传感器每小时产生 20TB 左右的数据量。

随着科研人员获取数据方法与手段的变化，科研活动产生的数据量激增，科学研究已成为数据密集型活动。科研数据因其数据规模大、类型复杂多样、分析处理方法复杂等特征，已成为大数据的一个典型代表。大数据所带来的新的科学研究方法反映了未来科学的行为研究方式，数据密集型科学研究将成为科学研究的普遍范式。

利用互联网可以将所有的科学大数据与文献联系在一起，创建一个文献与数据能够交互操作的系统，即在线科学数据系统，如图 1-3 所示。

图 1-3　在线科学数据系统示意图

对于在线科学数据，由于各个领域互相交叉，不可避免地需要使用其他领域的数据。利用互联网能够将所有文献与数据集成在一起，可以实现从文献计算到数据的整合。这样可以提高科技信息的检索速度，进而大幅度地提高生产力。也就是说，在线阅读某人的论文时，可以查看他们的原始数据，甚至可以重新分析，也可以在查看某些数据时查看所有关于这一数据的文献。

1.3　大数据的概念

大数据是指数据规模大，尤其是指因为数据形式多样性、非结构化特征明显，导致数据存储、处理和挖掘异常困难的那类数据集。大数据增长快速，类型繁多，如文本、图像、视频等。大数据处理包含数千万个文档、数百万张照片或者工程设计图的数据集，如何快速访问数据成为核心挑战。无法用常规的软件工具捕捉与处理。

通常将大数据归纳为 5 个“V”：Volume（数据容量）、Variety（数据类型）、Viscosity（价值密度）、Velocity（速度）、Veracity（真实性），如图 1-4 所示。

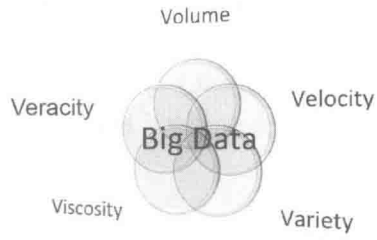

图 1-4　大数据的 5 个“V”

1.3.1　数据容量

Volume 代表大数据的数据量巨大，存储容量单位的定义如表 1-1 所示。