大数据技术及其应用

陈　燕　张金松　著

大连海事大学出版社

大数据技术及其应用

陈　燕　张金松　著

大连海事大学出版社

图书在版编目(CIP)数据

大数据技术及其应用 / 陈燕,张金松著. — 大连 : 大连海事大学出版社, 2015.12
ISBN 978-7-5632-3262-8

Ⅰ.①大…　Ⅱ.①陈…②张…　Ⅲ.①数据处理　Ⅳ.①TP274

中国版本图书馆 CIP 数据核字(2015)第 302562 号

大连海事大学出版社出版
地址:大连市凌海路1号　邮编:116026　电话:0411-84728394　传真:0411-84727996
http://www.dmupress.com　E-mail:cbs@dmupress.com
大连住友彩色印刷有限公司印装　大连海事大学出版社发行
2015 年 12 月第 1 版　2015 年 12 月第 1 次印刷
幅面尺寸:185 mm × 260 mm　印张:14.75
字数:295 千　印数:1 ~ 600 册
出版人:徐华东
责任编辑:苏炳魁　责任校对:宋彩霞　张　冰
封面设计:王　艳　版式设计:解瑶瑶
ISBN 978-7-5632-3262-8　定价:39.00 元

本书由

辽宁省交通厅项目(201401)、国家自然科学基金项目(71271034)资助出版

The published book is sponsored by

the Funds of Department of Communications of Liaoning Province

(201401) and the National Natural Science Foundation of China(71271034)

作者简介

陈燕，博士，大连海事大学交通运输管理学院教授/博士生导师，管理科学与工程学科一级学科带头人、省重点学科负责人，担任辽宁省物流航运管理系统工程重点实验室主任、辽宁省创新团队负责人。曾撰写《数据挖掘技术与应用》、《数据仓库与数据挖掘》、《数据仓库技术及其应用》、《管理信息系统开发教程》、《信息经济学》、《信息系统集成技术与方法教程》等学术专著与教材。主持并完成多项国家自然科学基金、国家科技计划项目及多项省部级、市级项目，获得省部级奖励10余项，发表相关学术论文200余篇。

张金松，博士，大连海事大学交通运输管理学院讲师，曾在美国印第安纳大学布鲁明顿分校进行过为期一年的交流学习，在JASIST期刊、CIKM、JCDL、AIRS等知名国际期刊、会议发表过学术论文，主要研究方向为信息检索、文本挖掘、数据挖掘等。

内容提要

本书系统详细地阐述了与大数据相关的概念、理论、技术及应用。主要内容包括：大数据相关概念，云计算与大数据，面向大数据的数据分析工具，数据库、数据仓库与数据挖掘，数据集成、中间件与数据压缩，非关系型数据库技术及应用，信息检索与主题模型相关技术及应用，可视化技术及应用等内容。

本书可作为信息管理与信息系统、电子商务、计算机应用、软件工程等高年级本科相关专业的教科书，同时也可以作为管理科学与工程、计算机应用及软件工程、工业工程等相关学科研究生的教科书或参考资料。

前　言

随着计算机硬件技术的飞速发展，传感器、微处理器和大容量存储设备的广泛应用，加之互联网、物联网、云计算等新一代信息技术的产生，网络中的数据量呈现暴增的趋势，可以说，信息技术已经逐步走向大数据时代。大数据不仅是信息科学技术领域的研究问题，其影响力更是已经覆盖了社会的各个领域。事实上，大数据并非一门全新的技术，而是信息管理科学、计算机应用科学发展的必然产物。因此，对大数据相关问题的研究，离不开数据库、数据仓库、数据挖掘、数据整合、数据压缩、中间件等基础理论的支撑，同时，非关系型数据库、信息检索、主题模型、可视化等技术又为大数据分析提供了有效的手段。

作者长期从事数据仓库、数据挖掘、数据分析等与大数据相关的理论研究，通过企、事业单位委托开发的项目，在数据采集、数据存储、数据管理、数据分析等方面积累了丰富的经验，并获得多项科研成果。对大数据相关问题的研究是在作者多年对数据挖掘等领域的研究基础之上展开的。尽管书中阐述的各种方法由于数据等诸多方面的限制，没能在大数据平台中运行实施，但相关理论与技术的阐述都是作者多年承担各项研究课题中真实案例所运用的，其与大数据的研究息息相关。

撰写本书的目的在于：将大数据相关的各个理论、技术进行分析，并以实际案例作为佐证，对读者认识大数据、应用大数据提供一定的借鉴作用。

在本书的编写过程中，王勇臻、陈志珍、韩红云、李欧、张鑫、徐慧颖、高鸽、李墨等学生参与完成全书的校对工作；杨明、李桃迎、牟向伟、刘志、蒋卓人、于莹莹、李鹏辉、孙骏雄等同志参与了本书部分章节的案例开发工作。

本书旨在通过应用案例从几个角度对大数据的相关理论与技术进行说明，但由于大数据技术所涵盖的范围非常广泛，因此，还有许多相关技术与方法需要进一步探讨。在编写过程中，笔者查阅了国内外大量文献资料，谨向书中提到的和参考文献中列出的相关人员表示感谢。如果由于我们工作的疏忽，本书中某处内容所参考的文献没有列出，在此向所涉及的作者深表歉意。

由于时间仓促和编者能力有限，书中难免存在一些不当之处，敬请广大读者批评、指正。

作　者

2015 年 10 月

目 录

第1篇 基础知识

第2篇 大数据相关理论及实例

第3篇 大数据相关技术及应用

第 1 篇　基础知识

1　大数据相关概念

1.1　大数据的发展背景

随着互联网、物联网、云计算等新兴技术的发展，以微博、微信等社交网络为代表的新型信息发布方式的不断涌现，加之移动互联网、电子商务应用的快速普及，数据的产生不再局限于特定的时间、地点，全球数据总量正在以前所未有的速度不断累积。也就是说，计算机、互联网与信息技术的普及，在给数字化社会带来管理模式的根本性变化的同时，也为信息化社会带来了种类繁多的数据。事实上，我们已经看到大数据在我们每个人的生活中产生，也就是说，我们每个人都是大数据的制造者，都离不开大数据，比如：我们有车辆定位系统、公交车刷卡系统、单位考勤系统、图像识别与指纹识别系统、生产业务管理系统、一天的工作效果即工作考核指标登记信息系统、商场购物信息管理系统、银行交易系统等相关信息系统所产生的大量数据。

具体来说，当我们在银行办理取款业务时，从申请业务开始到取款结束，整个过程伴随着数据的输入、处理、输出等大量信息和数据的存储，包括办理业务的客户数据、排队等待数据、客户取款产生的相关数据、取款过程中更新的银行数据、取款结束后上传的相关数据等。又比如，当我们在超市购物时同样会产生大量的数据，包括消费者所购产品的价格、货类、货名等相关数据，消费者支付过程中的支付金额、支付方式等相关数据以及消费者的购买记录、个人偏好数据等。另外，当我们出行订购车票时也会产生大量的信息与数据，例如出行所需的车次、航班等时间与价格等数据，到达车站、机场后的各种安全检查与身份认证数据，乘车、登机的确认数据以及安全抵达后的回馈数据等。不仅如此，交通信号灯的管理过程其实也产生了大量的数据，例如专家们将五岔路口信号灯管理的实际问题归结为图的数据结构问题，也就是为了设计一套有效的信号灯使五岔路口通行顺畅，进而运用图的计算机外部表示及其对应的计算机内部的存储方式，通过图的数据结构进行问题的求解，那么，在这一建模、求解过程中也产生了大量的科学管理与决策数据。

总之，在人类日常的生产、生活、出行过程中时刻伴随着数据的产生，这些数据从表面上看可能是杂乱无章的，但是经过分析总结往往能够发现一些潜在的规律，找出事件的深层联系，

甚至通过查看历史数据能够预测事件未来的发展趋势。由此可见，数据的分析不仅关系到科学技术的发展，而且与人类生产、生活息息相关。

通过对数据本身的变化进行分析，可以清楚地看到一个不争的事实，随着网络的应用与发展，计算机接受数据的规模已经从原始的比特（二进制的0、1）发展到字节、字、双字、多字等基本存储单位，同时，其数据存储模式也从单一的业务数据库模式逐步发展成面向对象数据库、主题数据库、多维数据库、结构化数据库、数据仓库、半结构化和非结构化数据库，其计算模式也从基于科学数值计算的结构化计算模式发展为非结构化的大数据与云计算模式。

可以说，数据规模的变化促进了各类数据中心的产生，形成了新型的数据管理与数据处理方式，带动了网络数据营销与电子商务营销模式的发展，从而为新型商业（企业）带来了新的商机和巨大的经济效益，出现了以出售书为主题的当当网、以旅行为主题的携程旅行网、以客户为中心理念的亚马逊。在成功运营的网络新型商业模式中，还有众所周知的淘宝购物网站。

可以发现，随着数据量的爆炸式增长和数据结构的愈加复杂，一般的数值计算和结构化的数据挖掘技术已经不能满足大型企业的需求，这就使得企业在采集数据之后，也开始有意识地寻找新的技术和方法来解决大量数据的存储和处理分析的问题，由此，产生了各个领域的大数据。阿里巴巴大数据团队自研的实时数据计算平台 Galaxy，每秒可运算数据超过 500 万条，2013 年“双 11”当天每秒运算量超过 1 000 万条，日处理消息数据超过 1 万亿条。实践已经证明：凡是成功的数据处理企业，都将处理大数据的重点业务作为本行业（企业）发展的重要举措。因此，针对大数据几乎是无所不在的今天，研究大数据的存储、操作及分析已经成为目前各行各业最重要的任务。

目前，大数据按照行业领域可以分为：金融领域大数据，交通运输领域大数据，综合业务领域大数据，网络营销领域大数据，环境检测及相关业务领域大数据，水利及水文、监控信息及空间信息等业务领域大数据，数字化建筑、数字化城市、数字化地理信息领域等各个领域的大数据。大数据的概念已经渗透到各个领域和各个行业中，因此，运用大数据理念和解决方案来处理大数据带来的问题已经越来越被人们所关注。虽然大数据源于信息行业，但其影响力已经远不止于信息行业，如何更好地管理和处理大数据已经引起学术界、工业界、金融界甚至是政府机构的密切关注。

1.1.1 大数据的产生

（一）企业级应用

随着电子商务业务的拓宽，各行各业信息化应用的普及与深入发展，计算机处理信息的领域不断扩大，其信息处理系统随之产生了大量现行的和历史的重要数据。企业在经营管理过程中，如企业内部业务资源计划系统、业务生产系统、产品市场交易系统、生产资料管理系统、财务系统、办公自动化系统、客户关系管理系统、物流供应链管理系统等都产生了大量的数据，同时也产生了众多文档、交易记录、操作日志、客户反馈等非结构化数据以及传感器数据、图像

视频监控文件等实时多媒体数据。尽管这些企业已经意识到这些复杂格式数据的潜在价值，也通过数据挖掘方法对客户的交易过程、业务处理流程等进行了分析和预测，但是企业所处的信息化环境正在发生着变化，企业应用与互联网、移动互联网的融合越来越快，来自企业外部的非结构化数据在大大增加。

按照传统的数据管理模式和处理方案来解决大数据的管理与分析存在许多弊端，比如：巨量的存储问题、巨量数据的组织与管理模式问题、巨量数据的分析模式问题。这些已经成为制约数据管理与分析的关键要素，并使企业逐步认识到传统数据管理模式在当前大数据环境下的弊端。因此，“大数据”技术开始向传统企业及组织的IT应用领域渗透，一些领先的IT企业和组织开始尝试“大数据”技术实验性部署，这势必会引发企业基础IT架构、数据处理、应用软件的开发和管理模式等发生一些新的变革。为了抢占先机，在大数据环境的竞争中处于有利地位，国内一些硬件厂商首先加入了大数据部署的产业行列，比如：联想公司最早通过与全球知名的存储公司EMC合作，正式进入大数据的企业级应用领域。另外，包括华为公司在内的诸多信息技术产业公司也相继推出了大数据相关产品，如：华为公司的四款面向企业级应用的T系列产品，其在统一存储领域具有显著优势。除此之外，国内的教育界，比如：高校、科研机构也在针对大数据的企业级应用提出各种解决方案，并为未来人才培养提出有效的培训方案。

由此可见，企业级应用的需求拉动了大数据管理与分析技术的发展，而IT企业的积极参与也加速了这一进程的发展步伐，企业级应用的普及成为大数据产生的首要因素。

（二）网络信息与数据的巨量增长

最初形成网络的目的在于提供电子邮件、文件传输协议和网页服务等，而随着互联网的普及，新型互联网应用的不断产生，网络信息与数据不断增长，目前，网络数据量已经占据了大部分的全球数据量。伴随数据量增长而来的是网络中日益增多的数据类型，比如：文本、图片、视频、声音数据等大量半结构化、非结构化的复杂数据类型应运而生。随着社交网络与媒体的发展，各类论坛、博客、社交网站为用户创建、上传和分享数据创造了更为便捷的方式，社交数据开始急速增加。

此外，网络应用产生的数据不仅仅来自于互联网，传统互联网到移动互联网的转变，移动宽带的迅速提升，产生数据的终端由个人计算机转向了包括个人计算机、功能手机等在内的多样化终端，移动互联网也成为网络数据的重要来源。个人智能手机和平板电脑的快速普及，越来越多的人、设备和传感器通过数字网络连接起来产生、分享和访问数据，移动互联网正在逐渐渗透到人们工作和生活的各个领域，移动终端逐渐演变成了一个提供通话管理、游戏娱乐、办公记录、网页浏览、购物理财、视频分享等各类应用在内的运行环境。根据互联网数据中心最新报告，2014年第三季度全球智能手机出货量超3.2亿部，在新兴市场需求量大增的推动下，2014年第三季度智能手机出货量相较2013年同期增长了近25.2%，相比2014年第二季

度增长了8.7%。另外,据2015年7月由中国互联网信息中心(CNNIC)发布的第36次《中国互联网络发展状况统计报告》指出,截至2015年6月,我国互联网普及率已经达到48.8%,其中,网民中使用手机上网的人群比例为88.9%。

由此可见,在移动互联网应用的发展过程中,移动终端使用日益普遍,信息与数据的增长势头仍将持续,特别是文本、视频、声音、图像等半结构化数据或非结构化数据的增长势头不可小觑。

(三)云计算的出现

云计算是信息技术领域继计算机、互联网之后的第三次革新浪潮。2006年"谷歌"在搜索引擎大会首次提出"云计算"的概念,短短数年间,云计算给信息领域带来了巨大的变革。目前,从国家角度来看,各国纷纷制订了云计算发展的国家计划,我国也掀起了兴建云计算基地的热潮。对IT企业而言,国内外的知名信息技术企业更是竞相推出云计算的产品和系统。另外,学术界也对云计算技术积极开展深入的研究。

尽管对云计算的概念还没有统一的定义,但通常认为,云计算是一种基于互联网的相关服务的增加、使用和交付模式,涉及通过互联网来提供动态易扩展且经常是虚拟化的数据资源。在云计算出现之前,数据大多保存在个人计算机或远程服务器中,而云平台能够将海量的网页数据集中存储到云端,用户仅需通过浏览器或专用的应用程序即可访问云端数据。云计算作为一种新型计算模式,体现了网格计算、分布计算、并行计算、效用计算等技术的融合与发展。

随着以云计算为代表的新型信息技术在国民经济、国家安全、科学研究、社会民生等各个领域的不断深入应用,社会生活模式、工作模式和商业模式也在发生着重大转变,以云计算为代表的信息产业通过其技术设施即服务(IAAS)、平台即服务(PAAS)和软件即服务(SAAS)等服务模式正带动着众多产业形态的创新和改革。目前,"谷歌"云计算平台已经达到了一百多万台服务器的规模,亚马逊、国际商用机器公司、微软、雅虎等公司的云平台也都达到了几十万台服务器的规模。

云计算为数据存储与计算提供了一个新型的平台,这也正为大数据的发展提供了动力。从技术上看,大数据与云计算是密不可分的,云计算作为大数据的基础与平台,而大数据则是云计算的重要应用,两者相辅相成,缺一不可。大数据的特点在于对海量数据的挖掘,其必然无法用单台的计算机进行处理,需要依托云计算的分布式处理、分布式数据库、云存储和虚拟化技术。由此可见,云计算的出现是大数据产生必不可缺的前提。

(四)物联网的应用

物联网是互联网的延伸和扩展,通过局部网络或者互联网等通信技术将射频识别、红外感应、全球定位系统、激光扫描器、蓝牙等信息传感设备进行信息交换和通信,实现对物体的智能化识别、定位、跟踪和监控管理,形成人与物、物与物、人与人之间的互联,实现信息化、远程管理控制和智能化的网络,它包括了互联网以及互联网上的所有资源,物联网的用户端延伸和扩