ITAT 教育部实用型信息技术人才培养系列教材

大数据概论

陈 明 编著

科学出版社

教育部实用型信息技术人才培养系列教材

大数据概论

陈　明　编著

科学出版社

北　京

内 容 简 介

本书主要介绍大数据概论，内容包括大数据概述、科学研究第四范式、分布系统设计的 CAP 理论、NoSQL 数据库、复杂网络、MapReduce 分布编程模型、大数据存储、大数据分析、大数据挖掘、大数据可视化、大数据安全、大数据机器学习、大数据推荐技术，以及数据科学与数据思维。全书对上述内容概念性地介绍，语言精练、内容全面。

本书可作为高等院校的大数据入门教材，也可以作为学习大数据的科学技术人员的参考书。

图书在版编目(CIP)数据

大数据概论 / 陈明编著. —北京：科学出版社，2015.1
教育部实用型信息技术人才培养系列教材
ISBN 978-7-03-042467-9

Ⅰ. ①大… Ⅱ. ①陈… Ⅲ. ①数据处理—技术培训—教材 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2014)第 268307 号

责任编辑：于海云 张丽花 / 责任校对：郭瑞芝
责任印制：霍 兵 / 封面设计：迷底书装

科 学 出 版 社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

三河市骏杰印刷有限公司 印刷

科学出版社发行 各地新华书店经销

*

2015 年 1 月第 一 版 开本：787×1092 1/16
2015 年 1 月第一次印刷 印张：17 1/2
字数：414 000

定价：48.00 元

(如有印装质量问题，我社负责调换)

前　言

需求是科学技术发展的原动力。目前，大数据问题的出现与研究已经成为了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科学研究第四范式。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。目前，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(信息技术)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，可以从数据中产生信息与知识。所以说，数据是重要的战略资源，隐含着巨大的经济价值，因此已经引起人类社会的广泛关注和高度重视。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。大数据的出现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。

本书概括性地介绍大数据的主要内容，是大数据技术入门的参考书。全书分为14章，内容包括大数据概述、科学研究第四范式、分布系统设计的CAP理论、NoSQL数据库、复杂网络、MapReduce分布编程模型、大数据存储、大数据分析、大数据挖掘、大数据可视化、大数据安全、大数据机器学习、大数据推荐技术，以及数据科学与数据思维。为了便于学习，在许多章节首先介绍传统的结构化数据的方法，然后再介绍非结构化数据的方法。

拥有数据的规模、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，帮助政府更好地响应社会和经济运行。以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国家间和企业间新的争夺点。大数据技术的专业人才，特别是数据分析专业复合型人才的稀缺将会影响该市场的发展。

在技术层面上，大数据、海量数据与超大规模数据无本质的区别，它们都是指用传统处理方法无法处理的大量数据。通过对大数据的高速有效处理，可以发现数据中蕴藏的规律与规则，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是小数据处理技术与工具，一些海量数据处理方法与工具是一种过渡性的方法与工具，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。大数据技术是一门崭新的技术，大数据的出现，对IT各领域的传统处理方法提出了新的挑战。

2012年作者接受了教育部ITAT教育工程的大数据处理工程师培训《大数据概论》的编写任务，通过对大数据的论文、著作的学习与实践，编写了本书，在这里，向所有大数据的研究者表示谢意。本书在结构上为积木状，各章内容是独立的概念性论述。

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。

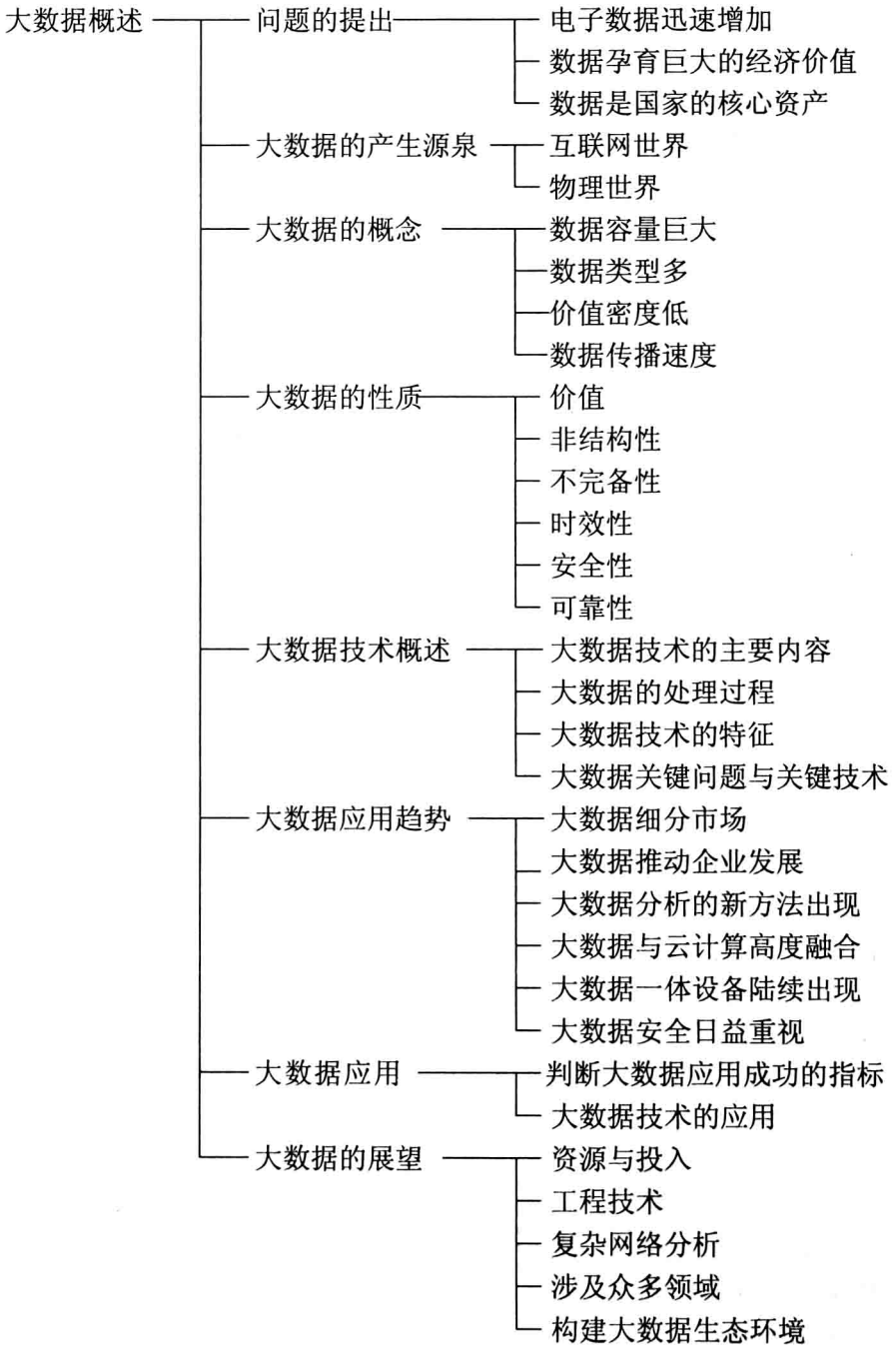
陈明

2014年2月

目　录

第 1 章　大数据概述

本章主要内容



需求是科学技术发展的原动力。目前，大数据问题的出现与研究已经成为了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科学研究第四范式的出现。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。目前，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。

1.1　问题的提出

在全世界范围内，以电子方式存储的数据(又简称为电子数据)总量空前巨大。在 2011 年电子数据总量已达到 1.8ZB(1ZB=1024PB)，较 2010 年同期提高超过 1ZB，统计结果表明，每经过 2 年就可以增加 1 倍，预计到 2020 年可达到 35ZB，如图 1-1 所示。面对数据增长的速度迅猛提升，数据量的飞速增加，对大量电子数据的高效存储、高效传输与快速的处理是必须面对的研究问题。

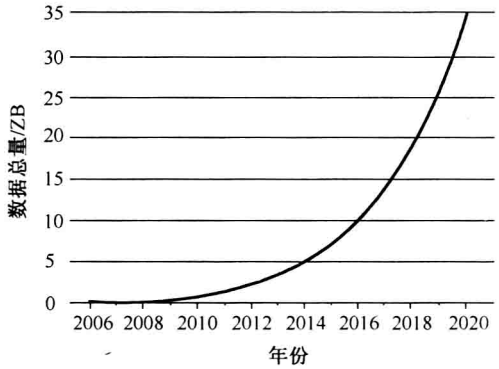


图 1-1　全球数据创建及复制的数据总量预测

1.1.1　电子数据迅速增加

物联网、云计算、移动互联网、车联网、手机、平板电脑、个人计算机(PC)、气候信息、公开的信息，如杂志、报纸和文章、交易记录、网络日志、病历、军事监控、视频和图像、档案及大型电子商务，以及遍布地球各个角落的各种各样的传感器是数据来源或者承载的方式不断更新与发展、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，特别是来自互联网服务而产生的大量数据迅猛增加。据不完全统计，互联网当前包含 93 亿多个页面，80%～85%的数据是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)，卖出的手机为 37.8 万台，高于全球每天出生的婴儿数量 37.1 万……从数据统计角度来看，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　数据孕育巨大的经济价值

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，可以从数据中产生信息与知识。数据是重要的战略资源，隐含巨大的经济价值，因此已经引起科学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。大数据出现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。

越来越多的企业等机构意识到数据正在成为最重要的资产，数据分析能力正在成为核心竞争力。经过了由 PC 成功转向了软件和服务，而这次将远离服务与咨询，更多地专注于因数据分析而带来的全新业务增长点。数据将成为各行业中决定胜负的根本因素，最终数据将成为人类至关重要的自然资源。各著名的大型公司已经致力于开发自己的大数据处理和存储系统，目前已经到了数据化运营的黄金时期，如何整合这些数据成为未来的关键任务。

在互联网、电信、金融等行业，几乎已经到了数据就是业务本身的地步。物联网、社交网络等新的互联网技术在为人们带来便利的同时，也产生了大量的数据。如何有效地存储和查询这些数据，如何通过数据挖掘，从数据中获得有用的信息，为用户提供好的用户体验，增强企业的竞争力，是一个挑战。研究表明，数字领域存在着 1.8 万亿 GB 的数据，企业数据正在以 55%的速度逐年增长。目前，两天就能创造出自人类文明诞生以来到 2003 年所产生的数据总量。大数据已经成为重要的时代特征，充分利用大数据可帮助全球个人定位服务提供商增加 1000 亿美元的收入，帮助欧洲公共部门的管理每年提升 2500 亿美元产值，帮助美国医疗保健行业每年提升 3000 亿美元产值，并可帮助美国零售业获得 60%以上的净利润增长率。由此可见，充分使用大数据和挖掘大数据商业价值将为行业企业带来强大经济效益与竞争力。

大数据既是对信息技术发展的高度抽象和概括，同时也体现了信息技术服务于数据蕴藏的巨大价值。大数据给数据的采集、存储、维护、共享带来了具有研究意义的现象和挑战，但更多的意义是可以处理、分析并使用大量数据，通过这些数据的处理、整合和分析，可以发现新知识、创造新价值，带来大知识、大科学和大发展，逐渐走向创新社会化的新信息时代。

大数据全生命周期可以划分为“数据产生—数据采集—数据传输—数据存储—数据处理—数据分析—数据发布、展示和应用—产生新数据”等阶段。已经形成了大数据的“生产与集聚层—组织与管理层—分析与发现层—应用与服务层”的产业链，而 IT 基础设施为这各环节提供基础支撑。

据统计，2012 年市场规模达到 4.5 亿元，2016 年估计可达到百亿规模，如图 1-2 所示。

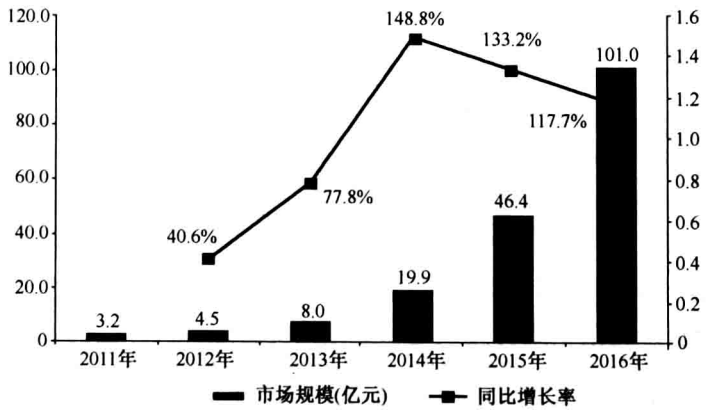


图 1-2　中国大数据应用市场规模与增长