BIG DATA

大数据技术与应用基础

DASHUJU JISHU YU YINGYONG JICHU

刘红英 刘 博 李韵琴 编著

海洋出版社

大数据技术与应用基础

刘红英　刘　博　李韵琴 编著

海洋出版社
2016 年 · 北京

内 容 简 介

本书全面讲述了大数据技术与应用相关的基础概念和知识，着重介绍了大数据的国内外发展状况、技术架构以及大数据分析的基本知识、数据分析、挖掘的流程、方法、工具等，并选择了一个数据分析、挖掘软件作为实例进行了重点说明。

全书分四个部分共18章，第一部分包括第1章至第4章，介绍大数据基础知识；第二部分包括第5章至第17章，介绍大数据初级应用基础知识；第三部分包括第8章至第16章，介绍数据分析挖掘模型和方法；第四部分包括第17章和第18章，介绍大数据关联工作说明。

本书既适合于开展国家职业技能培训的各类培训机构作为大数据基础分析师，以及大数据技术与应用基础的教材使用，也适合于已经掌握大数据基础知识，但希望使用数据挖掘软件进行数据分析实践的读者自学参考。

图书在版编目（CIP）数据

大数据技术与应用基础/刘红英，刘博，李韵琴编著．—北京：海洋出版社，2016.6

ISBN 978-7-5027-8927-5

Ⅰ.①大… Ⅱ.①刘… ②刘… ③李… Ⅲ.①数据-技术-应用 Ⅳ.①P23.5

中国版本图书馆CIP数据核字（2014）第057500号

责任编辑：苏 勤 黄新峰

责任印制：赵麟苏

海洋出版社 出版发行

http://www.oceanpress.com.cn

北京市海淀区大慧寺路8号 邮编：100081

北京华正印刷有限公司印刷 新华书店北京发行所经销

2016年6月第1版 2016年6月第1次印刷

开本：787mm×1092mm 1/16 印张：14

字数：260千字 定价：45.00元

发行部：62132549 邮购部：68038093 总编室：62114335

《大数据技术与应用基础》编委会

前　言

随着以博客、社交网络、基于位置的服务LSB为代表的新型信息发布方式的不断涌现，以及云计算、物联网等技术的兴起，数据正以前所未有的速度在不断地增长和累积，大数据时代已经到来。学术界、工业界乃至政府机构都已经开始密切关注大数据问题，大数据（Big Data）逐渐成为对于ICT产业具有深远影响的技术变革。其技术的发展与应用，将对社会的组织结构、国家的治理模式、企业的决策架构、商业的业务策略以及个人的生活方式产生深刻影响。可以说，大数据是一场革命，大数据将改变我们的生活、工作和思维方式。

尽管很多人对于大数据这个概念已经“耳熟能详”了，但是对于其具体含义及其相关内涵却并不是很清楚。本书着重从大数据的技术基础和初步应用出发，阐述如何将数据从简单的处理对象转变为一种基础性资源，如何更好地管理和利用大数据，如何分析和挖掘数据价值等问题。对大数据的基本概念进行剖析，并对大数据的主要应用作简单说明。在此基础上，阐述大数据处理的基本框架，并就云计算技术对于大数据时代数据管理所产生的作用进行分析。最后归纳总结大数据时代所面临的新挑战。

书中第1章简明扼要地介绍大数据的基本概念；第2章讲述了大数据的发展状况；第3章介绍了大数据的技术体系；第4章介绍了大数据的标准化知识；第5章至第16章每章介绍一种大数据分析的概念或技术；第17章和第18章介绍了大数据的关联工作。

本书由中关村软件园孵化器服务有限公司刘红英高级工程师、明博智创（北京）软件技术有限责任公司的刘博先生以及李韵琴高

级工程师合作编著。同时，中关村软件园孵化器服务有限公司的张征先生、何彦普先生、付延强先生、孙齐先生等参与了编写；明博智创（北京）软件技术有限责任公司赵晋平女士以及王野平先生等一起参与了编写。在编写过程中，还得到大数据标准化领域、国家职业技能培训领域各位专家的大力帮助、支持，在此深表谢意！

本书既适合于开展国家职业技能培训的各类培训机构作为大数据基础分析师，大数技术与应用基础的教材使用，也适合于已经掌握大数据基础知识，但希望使用数据挖掘软件进行数据分析实践的读者自学参考。

由于作者水平所限，书中难免出现错误和不妥之处，衷心希望各位读者批评指正！

编者

2015 年 6 月

目　次

第一部分　基础知识

第二部分 初级应用

第三部分 数据分析、挖掘模型和方法

第四部分 关联工作说明

第一部分　基础知识

第1章　大数据的基本概念、特征与作用

1.1　背景和概要说明

大数据是一场革命，大数据将改变我们的生活、工作和思维方式。继移动互联网、云计算后，大数据逐渐成为对于ICT产业具有深远影响的技术变革。大数据技术的发展与应用，将对社会的组织结构、国家的治理模式、企业的决策架构、商业的业务策略以及个人的生活方式产生深刻影响。

本书旨在介绍大数据技术与应用方面的常见概念和做法。主要目标读者除了国家职业技能培训的学生之外，还有希望通过大数据解决自身业务问题，但在计算机科学方面却没有相关知识背景的业务专家。尽管大数据融合了应用统计、逻辑、人工智能、机器学习和数据管理系统，但学习本书之前不需要在这些领域具有很强的背景。虽然学过统计学和数据库方面的初级大学课程将会对本书的学习有所帮助，但本书对大数据技术的基本概念和初级应用做了基本描述。

本书中的后续每一章都将介绍一种大数据的概念、技术或应用。本书并不是针对我们将使用的软件工具（明智商业分析系统V3.0、OpenOffice Base和OpenOffice Calc）的说明手册或教材。这些软件包能够进行许多类型的数据分析，本书并未涵盖它们的所有功能，只是介绍了如何使用这些软件工具进行某些类型的大数据分析、挖掘等。此外，本书并非面面俱到，虽然其中包含了众多常见的大数据分析、挖掘技术，但这些工具（尤其是明智商业分析系统V3.0）还能够执行许多本书中未涵盖的大数据分析、挖掘工作。

各章都将遵循相同的格式。首先，各章都将提供一个“背景和概要说明”，将帮助读者了解大数据分析、挖掘可以解决的某些类型的问题，旨在帮助读者思考实际工作中可能面临的各类问题。在“背景和概要说明”之后，是介绍每章主题内容的部分。在这些部分中，常常会给出一些逐步操作示例，读者可以跟随这些示例进行实际的大数据分析、挖掘工作。

1.2 大数据的基本概念和内涵

针对大数据，目前存在多种不同的理解和定义。按照 NIST 研究报告中的定义，大数据是用来描述在我们网络的、数字的、遍布传感器的、信息驱动的世界中呈现出的数据泛滥的常用词语。大量数据资源有可能解决以前不能解决的问题。

按照 Gartner 的定义，大数据是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产[①]。

根据百度百科词条的定义，大数据，或称巨量资料，指的是所涉及的资料量规模巨大到无法通过目前主流软件工具，在合理时间内达到撷取、管理、处理，并整理成为帮助企业经营决策更积极目的的资讯[②]。数据规模超出传统数据库软件采集、存储、管理和分析等能力的范畴，多种数据源，多种数据种类和格式冲破传统的结构化数据范畴，社会向着数据驱动型的预测、发展和决策方向转变，决策、组织、业务等行为日益基于数据和客观分析做出。

除了学术界、科研界的定义外，我国 IT 学术界和企业对大数据是如何理解的呢？通过调研，我们发现“新型的数据和分析”被超过一半的受访者所认同，而“新形式的数据应用”和“更大范围的信息”则分列二、三位，“大量的数据”这一选项仅仅列第四位。由此可见，大量的受访者已经意识到大数据的重点在于“数据”的分析和应用，而“大”不过是信息技术不断发展所产生的海量数据的表象而已（见图 1－1）。

我们认为这显示了大数据从量到质的变化过程；代表着数据作为一种资源在经济与社会实践中扮演着越来越重要的角色，相关的技术、产业、应用、政策等环境会与之互相影响、互为促进。从技术角度来看，这种数据规模质变后带来新的问题，即数据从静态变为动态，从简单的多维度变成巨量的维度，而且其种类日益丰富，超出当前技术与工具控制管理的范畴。这些数据的采集、分析、处理、存储、展现都涉及复杂的多模态高维计算过程，涉及异构媒体的统一语义描述、数据模型、大容量存储建设，涉及多维度数据的特征关联与模拟展现。然而，大数据发展的最终目标还是挖掘其应用价值，没有价值或者没有发现其价值的大数据从某种意义上讲是一种冗余和负担。

① 引自 Gartner 大数据定义。

② 引自百度百科大数据词条。

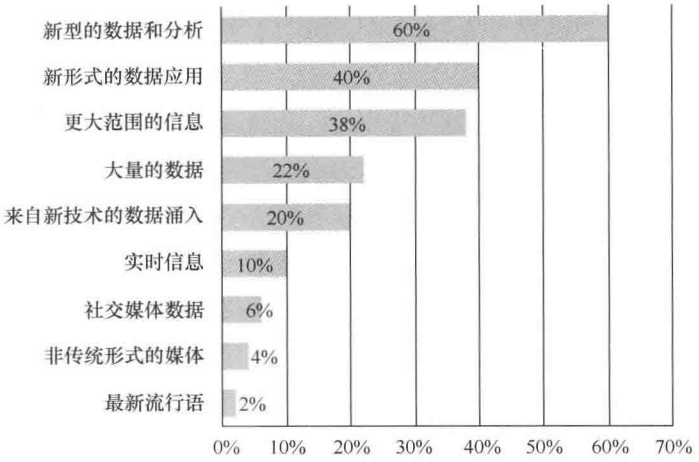

图 1－1　受访者对于大数据的认识

1.3　大数据的特征

目前，业内对于大数据特征的研究主要集中在“3V”、“4V”两种，归纳起来，可以分为规模、变化频度、种类和价值密度等几个维度。研究机构 IDC 定义了大数据的四大特征——海量的数据规模、快速的数据流转和动态的数据体系、多样的数据类型和巨大的数据价值，将“价值”作为第四个“V”。其他一些机构则将真实性作为第四个“V”。还有学者认为应该将供应商（vendor）作为第四个“V”。

我们对于大数据的特征从数量（Volume）、多样性（Variety）、速度（Velocity）、价值（Value）以及真实性（Veracity）几个方面进行认识和理解。在调查过程中，受访者对于大数据特性的关注度如图 1－2 所示，从高到低依次为多样性、价值、真实性、数量、速度。

➢ 多样性：数据形态多样，从生成类型上分为交易数据、交互数据、传感数据；从数据来源上分为社交媒体、传感器数据、系统数据；从数据格式上分为文本、图片、音频、视频、光谱等；从数据关系上分为结构化、非结构化、半结构化数据；从数据所有者分为公司数据、政府数据、社会数据等。

➢ 价值：尽管我们拥有大量数据，但是发挥价值的仅是其中非常小的部分。大数据背后潜藏的价值巨大。美国社交网站 Facebook 有 10 亿用户，网站对这些用户信息进行分析后，广告商可根据结果精准投放广告。对广告商而

言，10 亿用户的数据价值上千亿美元。据资料报道，2012 年，运用大数据的世界贸易额已达 60 亿美元。

➢ 真实性：一方面，对于虚拟网络环境下如此大量的数据需要采取措施确保其真实性、客观性，这是大数据技术与业务发展的迫切需求；另一方面，通过大数据的分析，真实地还原和预测事物的本来面目也是大数据发展未来的趋势。

➢ 数量：聚合在一起供分析的数据规模非常庞大。谷歌执行董事长艾瑞特·施密特曾说，现在全球每两天创造的数据规模等同于从人类文明至 2003 年间产生的数据量总和。“大”是相对而言的概念，对于搜索引擎，EB（1024 × 1024）属于比较大的规模，但是对于各类数据库或数据分析软件而言，其规模量级会有比较大的差别。

➢ 速度：一方面是数据的增长速度快，另一方面是对数据访问、处理、交付等速度的要求快。美国的马丁·希尔伯特说，数字数据储量每 3 年就会翻 1 倍。人类存储信息的速度比世界经济的增长速度快 4 倍。

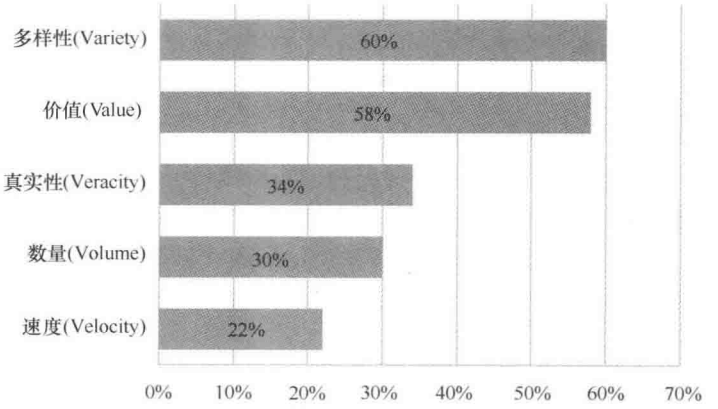

图 1 - 2 受访者对于大数据特征的关注度

从图 1 - 2 中我们不难看出，在大数据的几个特征中，“多样性”和“价值”最被大家所关注。“多样性”之所以被最为关注，在于数据的多样性使得其存储、应用等各个方面都发生了变化，针对于多样化数据的处理需求也成了技术的重点攻关方向。而“价值”则不言而喻，不论是数据本身的价值还是其中蕴含的价值都是企业、部门、政府机关所希望的。因此，如何将如此多样化的数据转化为有价值的存在，是大数据所要解决的重要问题。