CCF 大数据教材系列丛书
CCF 大数据专家委员会 组编

主编 程学旗

大数据分析

高等教育出版社

数据教材系列丛书
数据专家委员会 组编

程学旗

大数据分析

高等教育出版社·北京

大数据分析

Dashuju Fenxi

图书在版编目（CIP）数据

大数据分析 / 程学旗主编. -- 北京 : 高等教育出版社, 2019.4
ISBN 978-7-04-051632-6

Ⅰ. ①大… Ⅱ. ①程… Ⅲ. ①数据处理-高等学校-教材 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2019)第050472号

策划编辑 张江漫
责任编辑 张江漫
书籍设计 张申申
插图绘制 于 博
责任校对 刁丽丽
责任印制 赵义民

出版发行 高等教育出版社
社址 北京市西城区德外大街4号
邮政编码 100120
购书热线 010-58581118
咨询电话 400-810-0598
网址
http://www.hep.edu.cn
http://www.hep.com.cn
网上订购
http://www.hepmall.com.cn
http://www.hepmall.com
http://www.hepmall.cn
印刷 北京中科印刷有限公司
开本 787mm×1092mm 1/16
印张 22
字数 370千字
版次 2019年4月第1版
印次 2019年4月第1次印刷
定价 48.00元

物料号 51632-00

内容简介

本书系统地介绍了大数据分析的理论、算法及应用，全书共10章，包括基本概念和基本知识、大数据统计分析方法、大数据机器学习方法、机器学习理论、大数据算法理论、文本大数据分析、知识计算、网络数据挖掘、社会媒体分析、大数据分析系统架构等内容。除章节内容外，每章还提供“小节及进一步阅读”“参考文献”“习题”等内容帮助读者学习。

本书既可作为普通高等学校大数据相关专业的教材使用，也可供大数据分析领域的专业技术人员参考。

编者按

随着大数据的蓬勃发展，大数据领域人才的需求越来越大，大数据人才培养受到了各界的广泛关注。2016年，教育部开始批准设立“数据科学与大数据技术”本科专业，越来越多的高校申请开设“数据科学与大数据技术”专业或开设大数据方向的相关课程，截至2018年3月，已有近三百所高校获批建设“数据科学与大数据技术”专业。虽然大数据专业和大数据方向的课程不断开设，但是，当前我国高校的大数据教学尚处在摸索阶段，尤其缺乏成熟的、系统性和规范性的大数据教学体系和教材。

在此背景下，中国计算机学会大数据专家委员会成立了大数据教材系列丛书编委会，着手编著系列化、规范化的大数据教材。自2017年6月，经编委会多次研讨，形成丛书框架，作者们随即开始紧张的编写工作。编委会和作者间也有多轮的初稿审阅和研讨交流。数易其稿，终于付梓。

大数据教材系列丛书采用“1+3+X”的体系，即以1本《大数据导论》为基础，设置《大数据管理》《大数据处理》《大数据分析》3本关键技术教材，以及针对行业领域的X本应用教材。本套教材系列丛书既适合高校大数据专业的专科生、本科生以及研究生系统地学习大数据相关知识与技术，也适合从事大数据相关技术的企事业单位研究人员、工程师作为参考用书。

《大数据导论》是一本全面介绍大数据相关知识的专业通识教材，其系统地介绍大数据涵盖的内容，包括数据与大数据、大数据获取与感知、大数据存储与管理、大数据分析、大数据处理、大数据治理、大数据安全与隐私等，同时还介绍了部分行业中大数据的典型应用案例，反映了大数据在社会经济生活中的重要价值。

《大数据管理》首先综述数据管理系统的发展，指出发展大数据管理系统是历史的必然，并沿着数据模型和系统构件两个维度上展开。在数据模型的维度上，主要介绍关系、键值对、图和文档数据模型及其语言；在系统维度上，介绍系统结构、存储与组织、查询处理、事务管理、故障恢复等话题。

《大数据处理》包括大数据处理基础技术、大数据处理编程与典型应用处理、大数据处理系统与优化三个方面。本教材以大数据处理编程为核心，从基础、编程到优化等多个方面对大数据处理技术进行系统介绍，

使得读者能够快速入门，同时体会大数据处理系统的设计理念与优化方法本质。

《大数据分析》包括大数据分析方法和理论、典型大数据分析任务以及大数据分析系统与应用。本教材特色是理论联系实际，本书从基础理论、典型任务以及系统应用多个方面对大数据分析相关知识进行了系统而详细的介绍，使得读者能够快速入门，体会大数据分析技术的本质特征，领略大数据技术带来的创新理念。

“X”系列教材包含面向各行各业的大数据应用的知识与技术，既可面向工程实践，又可面向职业培训，且将随着产业界大数据应用的发展进行更新迭代。

大数据已成为学术界、产业界和政府共同关注的热点，正在开启信息化的新阶段。大数据人才培养也刚刚起步，还需要付出更多努力去探索。通过汇集中国计算机学会大数据专家委员会的智力资源，丛书编委会希望本系列教材能够为我国大数据人才培养尽到绵薄之力，助力我国大数据事业的蓬勃发展。尽管编委会和作者花费了很大精力规划和编写本系列教材，但是囿于对大数据的认识局限和自身能力限制，难免存在疏漏和错误，欢迎读者批评指正，以待再版时修正完善。

大数据教材系列丛书编委会

2018年7月

前言

大数据开启了信息化发展的第三阶段，大数据分析技术是推动信息化发展新阶段社会进步与经济发展的关键技术。大数据分析涉及统计、机器学习、计算机算法、计算架构、引擎系统等多个方面的技术，是一个综合性技术栈。同时，应用的多样性带来的数据来源、数据类型、数据形态等方面的多样化，使得大数据分析和具体应用紧密关联。撰写大数据分析方面的书籍需要综合考虑大数据分析的理论、算法及应用等多个层面的进展，本书作者从向量化、序列化、网络化三类典型数据出发，结合大数据分析的基础理论和方法，撰写本书。

《大数据分析》一书是大数据系列教材的成员之一，致力于从大数据分析的理论、算法及应用方面为大数据相关专业的本科生、研究生及科研人员提供一本较为全面介绍大数据分析相关技术的专业教材。本书系统地介绍大数据分析方法，包括大数据统计分析方法、大数据机器学习方法、机器学习理论、大数据算法理论、文本大数据分析、知识计算、网络数据挖掘、社会媒体分析、大数据分析系统架构等。除了介绍大数据分析方法之外，本书还介绍了部分行业中大数据的应用案例。本书既可作为普通高等学校大数据相关专业的教材使用，也可供专业技术人员参考。

本书共分10章。第1章简要介绍大数据分析的一些基本概念和基本知识，为读者提供一个使用本书的背景知识，并对本书的架构进行简明扼要的介绍。第2章介绍大数据统计分析方法，包括相关性分析、因果推断和大数据采样分析方法。第3章介绍大数据机器学习方法，由浅入深分别介绍了描述性分析方法、预测性分析方法、深度学习方法和强化学习方法。第4章介绍机器学习理论，从机器学习基础、模型选择、偏差方差分解、PAC学习理论和非独立同分布学习等几个方面展开。第5章介绍大数据算法理论，包括组合优化算法、在线算法、流式算法、参数算法等。第6章着眼于文本大数据分析，包括文本表达、文本匹配、文本生成。第7章围绕知识图谱介绍知识计算，包括知识抽取、知识融合、知识推理等。第8章着眼于网络数据，从网络排序、网络聚类、网络表示学习三个方面介绍了网络数据挖掘的方法和研究进展，包括中心度度量、网络划分、社区发现、网络嵌入等。第9章以社会媒体计算为例介绍了大数据分析的应用，包括社交网络分析、大图异常检测和其他一些社会媒体分析的新应用。第10章介绍大数据分析

系统架构，包括数据与计算的演变历程、大数据分布式计算模型、大数据计算系统等。

参与本书撰写的主要人员包括：程学旗、徐君、沈华伟、郭嘉丰、靳小龙、兰艳艳、刘盛华、张家琳、张志斌、庞亮。程学旗牵头确定了全书的知识架构，第1章“基本概念与基本知识”由程学旗起草；第2章“大数据统计分析方法”由徐君起草；第3章“大数据机器学习方法”由郭嘉丰起草；第4章“机器学习理论”由兰艳艳起草；第5章“大数据算法理论”由张家琳起草；第6章“文本大数据分析”由庞亮起草；第7章“知识计算”由靳小龙起草；第8章“网络数据挖掘”由沈华伟起草；第9章“社会媒体分析”由刘盛华、沈华伟起草。第10章“大数据分析系统架构”由张志斌起草。徐君和沈华伟协助我完成了全书统稿。本书由清华大学王建民教授审稿，在此致以衷心的感谢，同时感谢在本书撰写过程中提供素材或建议的老师和同学。

感谢中国计算机学会大数据专家委员会给予的支持和指导，感谢大数据教材系列丛书编委会各位同仁的努力付出。

作者认识和水平所限，参与本书编写人员众多，时间紧，统稿难度大，书中必然存在不少不一致的认知和叙述，欢迎读者批评指正，并积极反馈意见和建议，联系邮箱：bigdata@ccf.org.cn，我们将在再版时吸纳，使本书逐步趋于完善。

程学旗

2019年1月

目录

第1章 基本概念与基本知识

1

1.1 大数据与大数据分析

近几年，大数据迅速发展成为科技界和企业界乃至世界各国政府关注的热点。《自然》和《科学》等刊物相继出版专刊探讨大数据带来的机遇和挑战。著名管理咨询公司麦肯锡称："数据已经渗透到当今每一个行业和业务职能领域，成为重要的生产因素。人们对于大数据的挖掘和运用，预示着新一波生产力增长和消费盈余浪潮的到来。"美国政府认为大数据是"未来的新石油"，一个国家拥有数据的规模和运用数据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国家间和企业间新的争夺焦点。大数据逐渐成为社会各界关注的新焦点，"大数据时代"已然来临。

什么是大数据，迄今没有公认的定义。从宏观世界角度来讲，大数据是衔接物理世界、信息空间和人类社会三元世界的纽带。物理世界通过互联网、物联网等信息技术有了在信息空间中的大数据投影，而人类社会则借助人机界面、脑机界面、移动互联等手段在信息空间中产生自己的大数据映像。从信息产业角度来讲，大数据是新一代信息技术产业的强劲推动力。新一代信息技术产业本质上是构建在大数据、云计算、移动互联网等第三代信息技术平台上的信息产业。IDC（International Data Coporation）预测，到2020年第三代信息技术平台的市场规模将达到5.3万亿美元，而从2013—2020年，IT产业90%的增长将由第三代信息技术平台驱动。从社会经济角度来讲，大数据是第二经济（second economy）的核心内涵和关键支撑。第二经济的概念是由美国经济学家Auther在2011年提出的。他指出，由处理器、链接器、传感器、执行器以及运行在其上的经济活动形成了人们熟知的物理经济（第一经济）之外的第二经济（不是虚拟经济）。第二经济的本质是为第一经济附着一个

“神经层”，使人们的经济活动智能化，这是100年前电气化以来最大的变化。Auther还估算了第二经济的规模，他认为到2030年，第二经济的规模将逼近第一经济。而第二经济的主要支撑是大数据，因为大数据是永不枯竭并不断丰富的资源产业。借助于大数据，未来第二经济下的竞争将不再是劳动生产率而是知识生产率的竞争。

从非正式的说法来看，大数据是指无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合，是需要更新处理模式才能具有更强的决策力、洞察力和流程优化能力的海量、高增长和多样化的信息资产。国际商用机器公司IBM提出用5V特征来描述大数据，得到了业界的广泛认可，这5V特征分别是：

（1）体量（volume）大：2003年，人类第一次破译人体基因密码时，用了10年才完成了30亿对碱基对的排序；而在10年之后，世界范围内的基因仪15min就可以完成同样的工作量。

（2）种类（variety）多：随着传感器、智能终端以及在线社交协作技术的飞速发展，组织中的数据也变得更加复杂，因为它不仅包含传统的关系型数据，还包含来自网页、互联网日志（包括点击流数据）、搜索索引、社交媒体、电子邮件、文档文件、音视频文件各种传感器数据等原始、半结构化和非结构化数据。

（3）速度（velocity）快：所谓“1秒定律”指的是如果不能在秒级时间范围内给出分析结果，数据就会失去价值。而对于大数据来说，这里的快不仅仅指的是实时进行数据处理和结果展示要求数据处理速度快，也包含数据产生速度快的特点。有的数据是爆发式产生，例如，欧洲核子研究中心的大型强子对撞机在工作状态下每秒产生PB级的数据；有的数据是涓涓细流式产生，但是由于用户众多，短时间内产生的数据量依然非常庞大，例如，互联网点击流、电商平台日志、射频识别数据、GPS（全球定位系统）位置信息。

（4）价值（value）大、密度低：数据价值密度相对较低，却如浪里淘沙又弥足珍贵。随着互联网以及物联网的广泛应用，信息感知无处不在，信息海量，但价值密度较低，如何结合业务逻辑并通过强大的机器算法来挖掘数据价值，是大数据时代最需要解决的问题。

（5）真实（veracity）性难确定：量大导致数据的准确性和可信赖度难以判定，数据质量良莠不齐。

大数据的5V特征使得大数据分析的主要难点并不仅仅在于数据体量大。实际上，通过对计算机系统的扩展可以在一定程度上缓解数据体量大带来的挑战。而大数据分析真正的挑战来自于数据类型多样、数据质量的不确定性和数据分析的实时性要求。数据类型多样使得一个应用往往既要处理结构化数据，又要处理文本、视频、语音等非结构化数据，这对单一的数据库系统来说是难以应付的；而数据质量的不确定性，使得数据真伪难辨成为大数据应用的一大挑战，追求高质量数据是对大数据处理的一项重要要求，最好的数据清洗方法也难以消除某些数据固有的不可预测性；在数据的实时性分析方面，许多应用中时间就是利益，大数据分析的实时响应、在线更新是关键。

为了应对大数据带来的上述困难和挑战，以Google、Facebook、阿里、腾讯、Amazon等为代表的互联网企业近几年推出了各种不同类型的大数据处理系统。借助于新型的处理系统，深度学习、知识计算、可视化等大数据分析技术也得到迅速发展，并逐渐被推广应用于不同的行业和领域。

从价值角度看，大数据可以提供社会科学的新方法和科学研究的新范式，形成高新科技的新领域，推动社会进步的新引擎，而大数据分析技术是实现从数据到价值的关键。按照分析任务的难度以及其产生的价值划分，大数据分析技术可以分为四个层次：描述分析（descriptive analysis）、诊断分析（diagnostic analysis）、预测分析（predictive analysis）、规范分析（prescriptive analysis），如图1-1-1所示。其依次

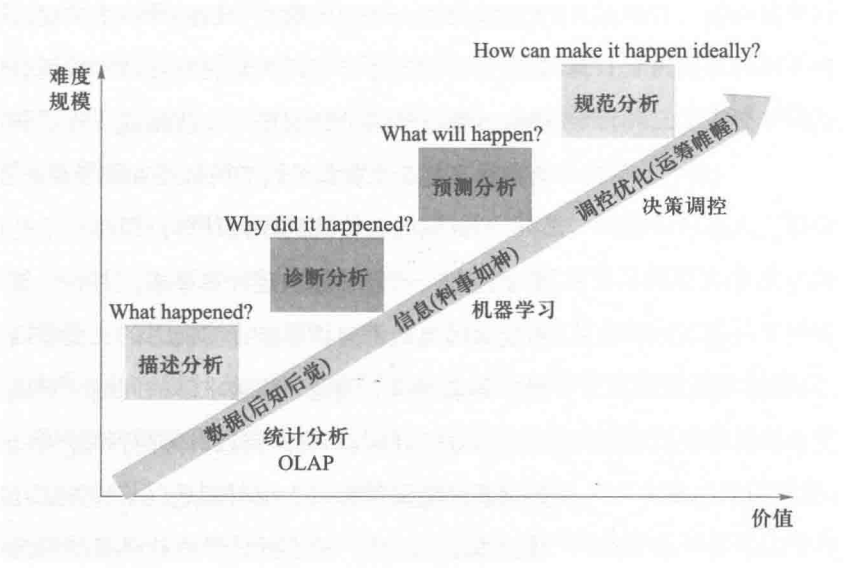

图1-1-1 大数据分析的四个层次

进行描述事情发生结果、分析事情发生原因、预测事情未来趋势以及控制事情发生轨迹。

1.2 重要的问题和概念

融合“人、机、物”三元空间的大数据存在规模大、关联复杂、状态演变等显著特征。部分传统的数据分析方法已经不再适用。在原有机器学习的基础上，设计出适用于大数据分析的方法成为关键问题之一。大数据分析算法模型具有训练样本多、输入数据维度高、输入数据结构复杂、模型参数多的特点。研究者们结合机器学习、认知计算和知识工程，研制新型的大数据分析系统算法，将大大提高大数据的价值利用率。

对于大数据基础分析算法，可以利用数据驱动表达学习与强化学习，提升分类、聚类、查询、检索、匹配、关联分析等基础算法的精准性和适用性。对于大数据融合分析技术，可以利用认知计算和深度学习，提升异构表达、跨媒体特征抽取与内容理解、数据融合分析、异常模式识别等高级算法的效能。大数据预测决策与可视化，主要利用机器学习与知识建模，突破大数据预测、知识推演、可视化分析、辅助决策等大数据分析技术。

随着训练数据集规模以及模型参数规模指数级的增长，如何获取强大的计算能力成为大数据分析的另一个关键问题。研究者利用智能芯片和硬件重构技术，可以大幅提升复杂大数据分析计算的速度。预计将来研究弹性泛流式分析引擎，支持批式、流式和在线分析，通过软硬件集合来提高复杂大数据分析的时效性和规模可扩展性。

与此同时，传统的序列计算，即在一个处理器上按照先后顺序进行指令处理，已经不能满足计算要求。并行计算（parallel computing）技术成为提高数据分析系统计算能力的关键手段。并行计算使用多个处理器协同求解同一问题，将被求解的问题分解成若干部分，各个部分均由一个独立的处理器来并行计算。并行计算可以是专门设计的、含有多个处理器的超级计算机，也可以是以某种方式互联的若干台独立计算机构成的集群。其中，分布式计算将软件系统的组成部分多机协作，提高效