普通高等教育“十三五”软件工程专业规划教材

大数据导论

宁兆龙　孔祥杰
杨　卓　夏　锋　编著

科学出版社

普通高等教育“十三五”软件工程专业规划教材

大数据导论

宁兆龙　孔祥杰　杨　卓　夏　锋　编著

科学出版社

北　京

内 容 简 介

本书是编者在多年从事大数据相关领域教学和科研的基础上编写而成的。全书系统地对大数据采集、存储、计算、处理、分析、挖掘和可视化等相关内容进行介绍，并结合大数据在社交、交通、医疗、金融、教育等方面的应用进行剖析阐述。

该书既可以作为计算机和软件工程专业的研究生和本科生教材，也可供从事信息技术领域的工程技术人员进行学习、使用和参考。本书相关内容基本覆盖了近些年大数据领域的最新技术和相关研究进展。

图书在版编目(CIP)数据

大数据导论/宁兆龙等编著. —北京：科学出版社，2017.5
普通高等教育“十三五”软件工程专业规划教材

ISBN 978-7-03-052662-5

Ⅰ.①大… Ⅱ.①宁… 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（2017）第 091568 号

责任编辑：张 帆 / 责任校对：郭瑞芝
责任印制：吴兆东 / 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

北京九州逐驰传媒文化公司印刷
科学出版社发行 各地新华书店经销
*
2017 年 5 月第 一 版 开本：787×1092 1/16
2017 年 5 月第一次印刷 印张：18 1/4
字数：467 000

定价：58.00 元

（如有印装质量问题，我社负责调换）

前　言

随着物联网和云计算技术的兴起，“大数据”已成为当今炙手可热的明星词汇。党中央在“十三五”规划建议中提出：“实施国家大数据战略，推进数据资源开放共享”。全球知名咨询公司麦肯锡称：“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。人们对于海量数据的挖掘和运用，预示着新一波生产率增长和消费者盈余浪潮的到来。”“大数据”因为近年来互联网和信息行业的发展而得到人们的广泛关注。

本书是编者在多年从事大数据相关领域教学和科研的基础上编写而成的。本书既可作为计算机和软件工程专业的本科生和研究生教材，也可供从事信息技术领域的工程技术人员进行学习和参考。本书全面系统地介绍了大数据理论和技术，相关内容基本涵盖了近些年大数据领域的最新技术和相关研究进展。

本书共 13 章，涉及的内容包括三大部分。

第一部分为大数据相关概述，由第 1 章组成，对大数据定义、结构类型、大数据技术的发展和大数据相关应用及面临的挑战进行阐述。

第二部分为大数据相关技术，由第 2～8 章组成。第 2 章介绍了大数据采集技术，详细介绍了大数据的来源、采集设备、采集方法和预处理技术。第 3 章对大数据相关存储技术进行了阐述，并对传统存储、云存储、大数据存储、数据中心和数据仓库分别进行了介绍。第 4 章对大数据计算平台进行介绍，着重对云计算平台、MapReduce 计算平台、Hadoop 计算平台、Spark 计算平台进行了说明。第 5 章为大数据分析技术，分别对传统数据分析方法、大数据分析方法、大数据分析架构和大数据分析应用进行了介绍。第 6 章介绍大数据挖掘相关知识，分别从大数据挖掘算法、挖掘工具、挖掘平台和挖掘应用展开论述。第 7 章对大数据下的机器学习算法进行阐述，首先对大数据的特征选择进行了简要介绍，接着对大数据的分类、聚类和关联分析进行着重讲解，最后对大数据的并行算法进行介绍。第 8 章介绍大数据可视化的相关内容，具体内容包括：大数据可视化技术、可视化工具、可视化案例以及可视化的未来。

第三部分为大数据相关应用，由第 9～13 章组成。第 9 章对社交大数据的相关应用进行阐述，首先对社交大数据的来源进行说明，接着对社交大数据在国内外社交网络中的应用进行实例分析。第 10 章为交通大数据，相关内容包括：交通数据分类及其相关分析、交通情况监测和基于交通大数据的人类移动行为预测。第 11 章为医疗大数据，首先对医疗大数据进行简介，接着对基于大数据的临床决策分析、基于大数据的医疗数据系统分析和基于大数据的远程病人监控进行着重阐述。第 12 章为金融大数据，分别从摩根大通信贷市场分析、瑞士银行集合风险分析、民生银行新核心业务平台分析和阿里信贷金融模式分析的实际案例对金融大数据相关应用进行展开说明。第 13 章为教育大数据，着重从微课教育、慕课教育和云平台教育对大数据教育的相关应用进行系统阐述。

本书的参考学时数为 32 学时。在课程学时较少的情况下，可针对具体情况选取部分知识进行学习。本书每章后均有小结和思考题，可使读者更好地回顾每章知识，并检查自己对知识的掌握程度。

本书在编写过程中得到了多位同行专业老师的指导，在此表示感谢。在本书内容编写过程中得到了大连理工大学阿尔法实验室成员的大力支持，他们是杜宏壮、冯玉凡、郭昊尘、郭琳琳、侯杰、侯轲、康文杰、李璐、李梦琳、李世璞、刘嘉莹、刘雷、刘鑫童、马凯、毛梦依、史尉欣、石雅洁、严颖梅、袁宇渊、张凯源。下列人员在本书校稿过程中提供了积极意见，他们是白晓梅、郭腾、彭众、王馨爽、于硕、张君，在此对上述人员进行一并感谢。编者在认真听取同行意见，在潜心研究的基础上细心编写了本书，希望本书对读者对大数据知识的系统学习提供帮助。但由于编写水平有限，书中还难免存在一些缺点和错误，殷切希望广大读者批评指正。

编 者

2017 年 3 月

目　录

第 1 章　大数据概述

当早上被闹铃叫醒，我们可以根据与手机互连的智能手环，从手机 APP 中看到昨晚睡眠的心跳、血压等健康状况信息；我们可以根据手机上即时更新的天气情况添减衣物；我们可以利用导航软件查阅实时交通状况，根据导航软件对用户以往的数据信息分析得出的出行建议进行路线规划；我们还可以利用大数据软件定位寻找附近的餐馆，甚至可以看到餐厅的用餐环境及特色菜品……不可否认，数据应用已渗透到我们生活的方方面面。

互联网带来的数据浪潮给我们的生活带来了极大便利。移动互联、社交网络、电子商务等应用随着互联网的兴起而产生并不断发展，同时大大拓宽了互联网的应用领域，并随之带来了海量的数据。

1.1　大数据定义

1.1.1　初识大数据

20 世纪以来，随着网络及计算机技术的发展，社会各行各业逐步走上了信息化的道路并积累了海量的数据。随着物联网和云计算技术的兴起，数据仍在以前所未有的速度增长和积累，并超越了相应存储仓库和数据处理资源的发展。如何采用新的技术和方法实现 PB 级甚至 ZB 级海量数据的存储和分析是我们当前面临的巨大挑战。爆炸式增长的数据正在引领一场新的时代变革，大数据时代已经来临。

什么是大数据（Big Data）？不同的研究机构基于不同的角度给出了如下定义。

大数据是需要新的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。

——高德纳（Gartner）咨询有限公司

大数据指的是大小超出常规的数据库工具获取、存储、管理和分析能力的数据集。

——麦肯锡[①]

大数据一般会涉及两种或两种以上的数据形式，它需要收集超过 100TB（1TB=2^{40}B）的数据，并且是高速实时数据流；或者是从小数据开始，但数据每年增长速率至少为 60%。

——国际数据公司[②]

总的来说，大数据是指所涉及的数据规模巨大到无法通过人工或计算机，在合理的时间内达到截取、管理、处理并整理成为人们所能解读的形式的信息。

另外，总结以上几种对于大数据的不同定义，我们不难发现大数据概念所具有的两点共性。

（1）大数据的数据量标准是随着计算机软硬件的发展而不断增长的。如 1GB 的数据量在 20 年前可以称为大数据，而今的数据量已上升到了太字节（TB）或拍字节（PB）量级。

---

① 麦肯锡公司，全球最著名的管理公司。

② 国际数据公司（International Data Group，IDG）。

（2）大数据不仅体现在数据规模上，还包含了不同于传统数据库软件获取、存储、分析和管理能力的提升。

1.1.2 大数据的特征

现在我们普遍以 5V 特征来具体描述大数据，其反映了大数据在 5 个层面上的特点，如图 1-1 所示。

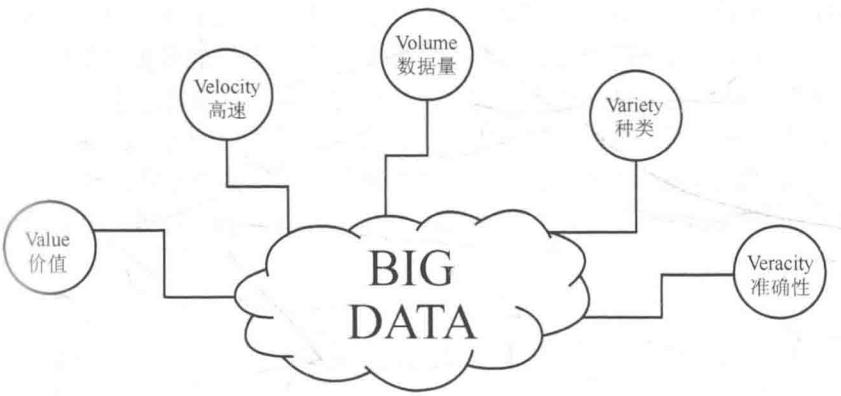

图 1-1 大数据的 5V 特征

（1）Volume：数据量巨大。数据体积大是大数据的显著特征，其数据量由传统 TB 级的基于关系的数据库处理数据量增长为 PB 级及以上的数据量，且不可避免的向泽字节（ZB）发展。

（2）Velocity：数据具有高速性。该特性包括大数据传输方式和处理方式。传输方式包括批处理传输、实时传输、近似实时传输和流传输等方式。数据处理方式包括数据处理时间和相应的时延。在具有时延的情况下，数据依旧需要以较高的速率被分析、处理、存储和管理，并遵循一秒定律[①]。

（3）Variety：数据类型多样。大数据不仅包括结构化数据，如传统文本类和数据库数据，还包括各种非结构化、半结构化以及复杂结构的数据，如网页、Web 日志文件、博客、微博、图片、音频、视频、地理位置信息等。

（4）Value：数据具有潜在价值。该特性是指大数据用户从中获得的价值。大数据的这一特性在商业领域较为关键。大数据中数据的价值密度与数据总量成反比，具有价值密度低的特点，如在视频数据中，一小时的视频中有用数据可能只占几秒。

一般而言，数据容量越大，种类越多，用户得到的信息量越大。获得的知识越多，数据能够发挥的潜在价值越大。但在实际情况中，大数据价值密度低这一特点使其数据价值往往依赖于较好的数据处理方式和工具。因此尽量减少由于数据垃圾和信息过剩造成的数据价值丢失，力求从数据中获得更高的价值回报至关重要。

（5）Veracity：数据准确性。该特性体现了大数据的数据质量。较为典型的应用是垃圾邮件，它们给社交网络带来了严重的困扰。据统计数据显示，网络垃圾占万维网所有内容的 20%以上。

从传统数据到大数据，形象地说类似于从“池塘捕鱼”发展到“大海捕鱼”的过程，而其中的鱼则为待处理的数据。两者的区别见表 1-1。

① 一秒定律是指对处理速度有要求，一般要在秒级时间范围内提出分析结果。

表 1-1　大数据与传统数据对比

比较项目	传统数据	大数据
数据规模	规模小，以 MB、GB 为处理单位	规模大，以 TB、PB 为处理单位
数据生成速率	每小时，每天	更加迅速
数据结构类型	单一的结构化数据	多样化
数据源	集中的数据源	分散的数据源
数据存储	关系数据库管理系统（RDBMS）	分布式文件系统（HDFS）、非关系型数据库（NoSQL）
模式和数据的关系	先有模式后有数据	先有数据后有模式，且模式随数据变化而不断演变
处理对象	数据仅作为被处理对象	作为被处理对象或辅助资源来解决其他领域问题
处理工具	一种或少数几种处理工具	不存在单一的全处理工具

在大数据定义过程中，需要注意的是其数据量不一定要满足 TB 级。在实际情况中，我们可以根据具体的数据特征来进行判断，如只有几百 GB 的数据在一定情况下也可以成为大数据。此时需要考虑其他判断标准，即数据处理速度或处理数据的时间维度，如几百 GB 的数据可以在一秒或几秒内被全部处理，而传统数据处理方式可能需要半小时甚至几小时，那么这种处理能力的高速提升极大地增加了数据价值。因此，所谓的大数据技术可以只满足以上部分判断特征。

同时，我们应注意区分“大数据”“大规模数据”和“海量数据”这几个概念。可以从以下两方面加以区分。

（1）从目标性来看，以上三者都具有数据容量大的特点。但大数据的目标是从大量数据中提取相关的价值信息，所以大数据并非只是大量数据无意义的堆积，其数据之间具有一定的直接或者间接联系。因此数据之间是否具有结构性和关联性是大数据和“海量数据”“大规模数据”的重要差别。

（2）就技术方面而言，大数据能够快速、高效地对多种类型的数据进行处理和整合从而获得有价值的信息，这也是大数据不同于“海量数据”和“大规模数据”的最主要特征。在数据处理过程中，大数据处理技术运用了如数据挖掘、分布式处理、聚类分析等多种方法，并对相关的硬件发展和软硬件的集成技术提出了较高要求。

数据量的剧增伴随着数据处理要求的不断提高。因此，大数据的处理技术也得到相应发展。

1.1.3　大数据技术

大数据技术是新兴的，能够高速捕获、分析、处理大容量多种类数据，并从中得到相应价值的技术和架构。大数据处理的关键技术主要包括：数据采集和预处理、数据存储、基础架构、数据分析和挖掘以及大数据应用。利用大数据技术对数据处理流程如图 1-2 所示。

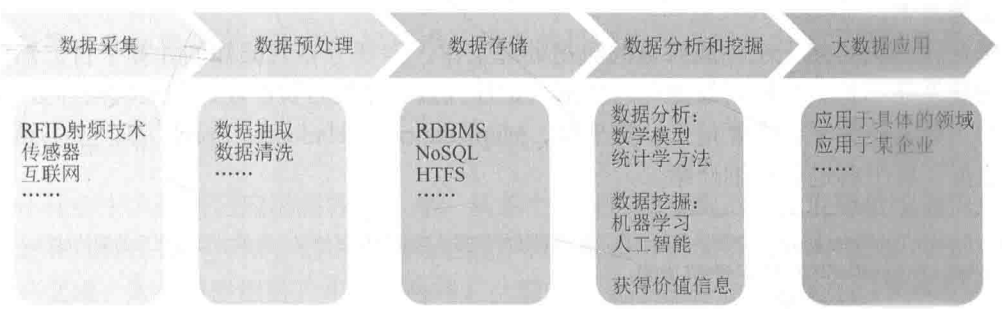

图 1-2　大数据处理流程

1. 数据采集

数据是通过射频识别技术、传感器、交互型社交网络以及移动互联网获得的多类型海量数据，这些数据是大数据知识服务模型的根本。

大数据采集一般分为大数据智能感知层和基础支撑层。智能感知层主要包括数据传感体系、网络通信体系、传感适配体系、智能识别体系以及软硬件资源接入系统，可以实现对结构化、半结构化、非结构化海量数据的智能化识别、定位、跟踪、介入、传输、信号转换、监控、初步处理和管理等。基础支撑层主要提供大数据服务平台所需的虚拟服务器，结构化、半结构化及非结构化数据的数据库及物联网资源等基础支撑环境。本书第 2 章将详细介绍这些内容。

2. 数据预处理

数据预处理是数据分析和挖掘的基础，是将接收数据进行抽取、清洗、集成、转换、归约等并最终加载到数据仓库的过程。

（1）数据清洗：现实世界中接收到的数据一般是不完整、有噪声且不一致的。数据清洗过程试图填充空缺值，光滑噪声并识别离群点，纠正数据中的不一致。因此，为了提高数据挖掘结果的准确性，数据预处理是不可或缺的一步。数据清洗过程主要包括数据的默认值处理、噪声数据处理、数据不一致处理，常见的数据清洗工具有 ETL[①]和 Potter's Wheel[②]。

（2）数据集成：数据集成过程是将多个数据源中的数据合并同时存放到一个一致的数据存储（如数据仓库）中，其中数据源可以包含多个数据库、数据立方体或一般文件。数据集成需要考虑诸多问题，如数据集成中对象匹配问题、冗余问题和数据值的冲突检测与处理问题。

（3）数据转换：将原始数据转化为适合于数据挖掘的数据形式。数据转化主要包括数据泛化、数据规范化和新属性构造。

（4）数据归约：数据归约指在尽可能保持数据原貌的前提下，最大限度地精简数据量，该处理过程主要针对较大的数据集。数据归约主要有两个途径：属性选择和数据采样。这两种途径分别针对原始数据集中的属性和记录进行处理。

3. 数据存储

数据存储过程需要将采集到的数据进行存储管理，建立相应的数据库。详解可见本书第 3 章。根据采集数据多样化的特点，数据主要存储在关系数据库、NoSQL、HTFS 等数据库中。

为了保证数据的安全性，数据存储也需要考虑相应的安全技术，主要包括：分布式访问控制、数据审计、透明加解密、数据销毁、推理控制、数据真伪识别和取证、数据持有完整性验证等技术。

单台计算机必然无法完成海量的数据处理工作，需要分布式架构的计算平台。然而高可用性的硬件并不是大数据高效处理的全部决定性因素，合理的软件设计和架构同样必不可少。现有的大数据计算平台主要是云计算平台、MapReduce[③]、Hadoop、Spark[④]等，本书第 4 章将对大数据计算平台进行详细介绍。

---

① ETL（Extract Transfrom Load），常用于数据仓库，用来描述将数据从来源端经过抽取、转换、加载至目的端的过程。[http://baike.so.com/doc/2126217-2249603.html]。

② http://control.cs.berkeley.edu/abc/。

③ MapReduce 是一种编程模型，用于大规模数据集的并行运算。

④ Spark 是一种与 MapReduce 类似的通用并行架构。

4. 数据分析和挖掘

数据分析是指利用相关数学模型以及机器学习算法对数据进行统计、预测和文本分析。数据分析可分为预测性分析、关联分析和可视化分析。数据的主要分析方法有探索性数据分析方法、描述统计法、数据可视化等。关于数据分析的详细内容请查阅本书第 5 章。

预测性分析是通过大数据中某些特点科学地建立模型，并将最新数据应用到已建立的模型中，达到预测未来数据趋势的目的，从而减少对未来事物认知的不确定性。关联分析的目的是寻找数据之间的内在联系。可视化分析是将大型数据集中的数据以图形图像的形式表示，并利用数据分析和开发工具发现其中未知信息的处理过程。对应处理工具主要有动态分析工具和以图形、表格等可视化元素为主的工具。可视化分析可以直观地呈现大数据的特点。

数据挖掘是利用人工智能、机器学习、统计学等多学科方法从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据集中提取隐含在其中的有价值信息或模式的计算过程。数据挖掘技术众多，根据分类方法的不同可以分为多类，具体可见表 1-2。

表 1-2　数据挖掘技术分类表

分类标准	类别
挖掘任务不同	分类或预测模型发现
	数据总结、聚类、关联规则发现
	序列模式发现
	异常和趋势发现
挖掘对象不同	关系数据库
	面向对象数据库
	空间数据库
	时态数据库
	文本数据源
	多媒体数据库
	异质数据库
	遗产数据库
	互联网 Web
挖掘方法不同	机器学习方法（监督、非监督、半监督学习法）
	统计方法（回归分析、判别分析、探索性分析等）
	神经网络方法
	数据库方法

本书第 6 章和第 7 章对数据挖掘技术和相关的机器学习下的数据挖掘算法进行了详细介绍。

5. 大数据应用

现今社会中大数据已应用到各行各业。从各个领域的海量数据中提取有价值的信息进行相关预测和选择决策，可以有力地推动社会进步和发展。目前大数据的典型应用包括社交网络、公共交通、医疗卫生服务、电子商务等。大数据无处不在并与我们的生活紧密相连。