大数据科学与应用丛书

大数据技术
基础与应用导论

杨毅 王格芳 王胜开 等编著

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

大数据科学与应用丛书

大数据技术
基础与应用导论

杨毅 王格芳 王胜开 陈国顺 孙甲松 编著

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

本书从大数据的前身——数据挖掘技术入手，首先介绍了数据挖掘技术及在大数据中常用的采集、存储和分析方法；然后以连续语音识别和多语言语音识别为例，对大数据信息处理技术的关键应用给出了详细的说明；接着给出了大数据场景分析，详细介绍了基于场景分析的大数据信息处理应用，如MOOC 大数据教学分析系统、社交网络大数据关系推荐系统、金融服务大数据风险预警系统等；随后介绍了互联网+大数据的应用，对电子商务、互联网金融、能源大数据等具有差异性的行业应用进行了简要介绍；最后对大数据的应用进行了展望。

本书包括大数据、数据挖掘和场景感知等基本内容及其应用，可作为相关专业本科生及研究生学习大数据应用的入门用书，对工程人员来说，本书也是一本综合性较强的参考图书。

图书在版编目（CIP）数据

大数据技术基础与应用导论 / 杨毅等编著. — 北京：电子工业出版社，2018.6
（大数据科学与应用丛书）
ISBN 978-7-121-34336-0

I. ①大… II. ①杨… III. ①数据处理 IV. ①TP274

中国版本图书馆 CIP 数据核字（2018）第 117582 号

责任编辑：田宏峰
印　　刷：三河市双峰印刷装订有限公司
装　　订：三河市双峰印刷装订有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　　邮编：100036
开　　本：787×980　1/16　印张：12　　字数：268 千字
版　　次：2018 年 6 月第 1 版
印　　次：2018 年 6 月第 1 次印刷
定　　价：68.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：tianhf@phei.com.cn。

前　言

“大数据”这个词汇已经与“移动互联网”“云计算”“人工智能”等一起成为科技从业人员中，甚至是街头巷尾的流行词汇之一。中国工程院邬贺铨院士在 2013 年撰写的大数据时代的机遇与挑战至今已被引用 200 多次；同年出版的维克托·迈尔·舍恩伯的专著《大数据时代》则一直在亚马逊的热销图书商品排名中，其热度可见一斑。从 2016 年美国总统选举到相亲网站用户匹配，大数据的身影无处不在，每个人的工作和日常生活，都自觉或不自觉地受到大数据的影响和支配。但什么是大数据，每个人、每个机构，甚至每个国家，都对此有不同的答案。我们需要给大数据一个清晰的、统一的、完整的定义。幸运的是，麦肯锡全球研究所给出了一个标准答案：大小超出了传统数据库软件工具的抓取、存储、管理、分析能力的数据群被称为大数据。

虽然大数据如此之热，但是在具体深入研究下去后就会发现，大数据技术的研究和应用的主要领域仍然集中在与 IT 产业密切相关的互联网产业界，在电子商务、搜索推荐、可穿戴设备、无人车/机等方向上，各种规模的创新、创业公司层出不穷，各类应用更是五花八门、纷繁复杂，而大数据相关的国内外文献也是种类繁多、涉及广泛。

大数据分析应用于科学、医药、商业等各个领域，用途差异巨大，但其目标可以归纳为如下几类。第一，获得知识与推测趋势。大数据包含大量原始的、真实的信息，大数据分析能够有效摒弃个体差异，帮助人们可以透过现象更准确地把握事物背后的规律。第二，分析掌握个性化特征。企业通过长时间、多维度的数据积累，可以分析用户的行为规律，更准确地描绘个体轮廓，为用户提供更好的个性化产品和服务，以及更准确的广告推荐等。第三，通过分析辨识真相。由于网络中的信息传播更加便利，所以网络虚假信息造成的危害也更大。由于大数据的来源广泛且具有多样性，因此在一定程度上可以帮助实现信息的去伪存真。目前，人们开始尝试利用大数据进行虚假信息的识别。

相应地，大数据技术也面临巨大的挑战，主要包括：

（1）当前的数据量正以指数方式增长，而大数据处理和分析的能力远远跟不上数据量增长的速度。高效率和低成本的存储技术、非结构化和半结构化数据的高效处理技术、大数据去冗降噪技术、数据挖掘和基于大数据的预测分析技术等都有待发展和完善。

（2）大数据包含丰富的个人信息，通过整合分析，可以精准判断个人的喜好乃至性格，揭示行为规律，使个人的隐私信息更加容易暴露。如何在加强数据获取能力的同时更好地保护个人隐私，是未来大数据研究的一个重大挑战。

（3）大数据使人类对信息掌控的程度相对过去有了质的提升，从这个意义来看，从信息时代进入大数据时代超越了从机械计算时代进入电子计算时代，对于大数据的观念、态度必须要能够适应新时代的要求。

本书尝试从大数据的前身——数据挖掘技术入手，首先介绍在大数据这个词汇发明之前，数据挖掘技术是如何用于金融投资、识别欺诈并保障网络安全的；随后对大数据技术中使用的采集、存储及分析方法，如目前流行的 HDFS 及 MapReduce 进行详细阐述，以便使入门者快速掌握相关的技术；随后以语音识别中的连续语音识别和多语言语音识别为例，介绍大数据信息处理技术在 IT 行业中的关键应用；大数据分析与场景密切相关，因此提供了一系列基于场景分析基础上的大数据信息处理应用，如 MOOC 大数据教学分析系统、社交网络大数据关系推荐系统和金融服务大数据风险预警系统等；以互联网+大数据为特色的应用非常广泛，仅选取了电子商务、互联网金融、城市可持续发展、能源大数据、智能电网大数据等差异性较大的行业应用进行了简单介绍；进一步的大数据信息处理应用则涉及场景感知这一更加复杂的课题，场景感知更近似于人类对场景的观察、判断、分析与响应，相比于场景分析具有更强的灵活性、实时性、准确性，无人驾驶汽车操控系统就是场景感知的典型综合应用案例。

本书包括大数据、数据挖掘和场景感知等基本内容及其应用，可作为 IT 相关专业本科及研究生学习大数据理论、技术与应用的入门用书，对工程人员来说也是一本综合性较强的参考手册。同时，本书引用了大量国内外最新技术实例及作者的国家基金项目研究成果，对互联网领域的技术研究人员也有一定的参考价值。

本书在编写过程中，北京交通大学袁保宗教授、中国科学院声学研究所颜永红教授、北京理工大学谢湘副教授等专家给予了大力指导和支持，并得到国家自然科学基金重大项目（NSFC：11590770）的支持，在此表示衷心的感谢！

由于编著者水平和经验有限，书中错误之处在所难免，敬请读者指正。

编著者

2018 年 5 月

目　　录

第1章

绪论

1.0 引　　言

随着计算机网络用户数量的增长，每天都产生上万亿比特的数据，大数据（Big Data）时代已经到来，这是过去几十年计算机领域没有预见的，这给计算机信息处理技术带来了新的挑战，必须利用新的思路和理念来处理与日俱增的数据。

对于越来越多的海量数据，用以往的方法已经很难进行有效的处理，因此人们开始关注和研究海量数据的处理方法。2011 年 6 月，麦肯锡全球研究所发布了《大数据：创新、竞争和生产力的下一个前沿》的报告，对“大数据”的概念进行了清晰的阐释。报告将“大小超出了传统数据库软件工具的抓取、存储、管理、分析能力的数据群”称为大数据。2012 年 1 月，在瑞士达沃斯召开的世界经济论坛上，大数据是主题之一，会议报告宣称，数据已经成为一种新的经济资产类别，就像货币或黄金一样。2012 年 3 月，奥巴马宣布美国政府将投资 2 亿美元启动“大数据研究和发展计划”，用于研究开发科学探索、环境和生物医学、教育和国家安全等重大领域与行业急需的大数据处理技术和工具，这是继 1993 年美国宣布“信息高速公路”计划后的又一次重大科技发展部署。美国政府认为，大数据是“未来的新石油”，并将对大数据的研究上升为国家意志，这必将对未来的科技与经济发展产生深远影响。在这些事件的推动下，大数据逐渐变为全球关注的热门概念，人们甚至将 2012 年称为“大数据元年”。

尽管各国政府都对大数据技术高度重视，都不遗余力地大力推动大数据的研究。但事实上，大数据技术研究和应用的主要战场，仍然在企业界，特别是在和信息产业密切相关的互联网产业界。如果将大数据的技术版图进行划分，则呈现出以下三大板块，各自有不同的特点。

1. Google 提出并引领大数据技术

大数据概念被关注之前，对于不断增多的数据，人们的应对方法是不断提升服务器的性能、增加服务器集群数量。在海量数据的冲击下，这种模式付出的成本代价越来越大，最终将达到一个无法承受的程度。例如，Oracle 海量数据库系统 Exadata，每个定制集群系统需 2000 千万美元，仅能存储 10 TB 的数据，因此急需研究大数据的索引和查询技术。

在大数据处理技术上具有里程碑意义的事件，是 Google 于 2003 年发表的三篇大数据的技术论文——MapReduce、Google File System、BigTable。这三篇论文描述了采用分布式计算方式来进行大数据处理的全新思路，其主要思想是将任务分解，然后在多台处理能力较弱的计算节点中同时处理，再将结果合并，从而完成大数据处理。这种方式不再采用昂贵的硬件，而是采用廉价的 PC 级服务器集群，实现海量数据的管理。MapReduce 是一种用于大规模数据集并行计算的编程模型，可将一个大作业拆分为多个小作业的框架，进行作业调度和容错管理。Google File System 是一个使用廉价的商用机器构建的大型分布式文件系统，由文件系统来完成容错任务，利用软件方法保证可靠性，使存储成本大幅下降。Big Table 是一个建立在 Google File System 之上的适用性广泛、可扩展、高性能、高可用性的、非关系型分布式结构化数据存储系统，处理的数据通常是分布在数千台普通服务器上的 PB 级的数据。

2. 开源 Hadoop 提供技术基础

Google 的论文给全世界带来了震撼，但由于是私有的技术，无法被其他公司使用。在 Google 思路的启发下，相应的开源项目得到了极大发展，最重要的就是 Apache 基金会下的 Hadoop 项目。Hadoop 项目起源于 2005 年，包含了和 Google 大数据技术相对应的 Google MapReduce、HDFS 和 HBase 等组成部分。Hadoop 可以视为 Google 技术的开源实现，因此具有高可靠性、高扩展性、高效性、高容错、低成本等一系列特点。

Hadoop 技术尽管仍然不能达到 Google 论文中声称的性能，但是它开源的特性使得所有人都可以学习、研究和改进它，同时由于它背后有 Yahoo、Facebook 等 IT 巨头的强力支持，已经完全可以满足当前大数据应用的需求。2011 年以后 Hadoop 的应用越来越多，连 IBM 的智力问答机器人沃森也是基于 MapReduce 数据并行处理的。

3. 各大企业推动大数据应用

在 IT 行业，Yahoo、Facebook、Linkedin 和 eBay 等众多企业纷纷转向 Hadoop 平台，推动和完善 Hadoop 项目，并搭建分布式数据处理平台进行大数据的采集、分析和处理。

Yahoo 投入了大量的资源到 Hadoop 的研究中，目前 Yahoo 在 Hadoop 上的贡献率占了 70%。从 2005 年起，Yahoo 就成立了专门的团队，致力推动 Hadoop 的研发，并将集群从 20 个节点发展到 2011 年的 42000 个节点，初具生产规模。在应用领域，Yahoo 更是积极地将 Hadoop 应用于自己的各种产品中，在搜索排名、内容优化、广告定位、反垃圾邮件、用户兴趣预测等方面得到了充分的应用。Facebook 拥有超过 10 亿的活跃用户，需要存储和处理的数据量巨大。它使用 Hadoop 平台建立日志系统、推荐系统和数据仓库系统等。2012 年，Facebook 甚至宣布放弃自行研发的开源项目 Cassandra，全面采用 HBase 为邮件系统提供数据库支持。Facebook 目前运行着的可能是全球最大规模的基于 Hadoop 的数据搜集平台。另一方面，Facebook 也以自身的强大实力，为 Hadoop 提供强力的支持。2012 年，Facebook 宣布开源 Corona 项目，这是 MapReduce 的改进版本，可以更好地利用集群资源。阿里巴巴同样是 Hadoop 技术的积极响应者，2009 年，阿里推出了以 Hadoop 为基础的分布式数据平台“云梯”。Hadoop 使得大数据的应用已成燎原之势，除了 IT 企业，金融、传媒、零售、能源、制药等传统行业在大数据技术应用方面也积极响应，行业应用如系统研发、服务需求和计算模型研究等都在开展中。

大数据已成为继云计算之后信息技术领域的另一个信息产业增长点。据 Gartner 公司预测，2013 年大数据将带动全球 IT 支出 340 亿美元，到 2016 年全球在大数据方面的总花费将达到 2320 亿美元。Gartner 将大数据技术列入 2012 年对众多公司和组织机构具有战略意义的十大技术与趋势之一。不仅如此，作为国家和社会的主要管理者，各国政府也是大数据技术推广的主要推动者。2009 年 3 月美国政府上线了 data.gov 网站，向公众开放政府所拥有的公共数据。随后，英国、澳大利亚等政府也开始了大数据开放的进程。截至目前，全世界已经有 35 个国家和地区构建了自己的数据开放门户网站。美国政府联合 6 个部门宣布了 2 亿美元的“大数据研究与发展计划”。2012 年，中国通信学会、中国计算机学会等重要学术组织先后成立了大数据专家委员会，为我国大数据应用和发展提供学术咨询。

云计算技术和物联网技术的产生给大数据时代的到来提供了必要条件，是计算机行业又一次重大的革命性变革，并直接影响广大计算机用户、企事业单位和政府机关的活动方式，以及它们之间的交流途径。数据是大数据时代的最重要的核心内容，企业、消费者和网民之间的界限在大数据时代变得模糊，这对企业的运行、经营、管理和发展方向都产生了重要影响，同时也带来各种挑战和机遇。

由于传统计算机硬件的限制，使得计算机网络存在诸多的应用局限，需要将目前的计算机网络转换为云计算机网络，这是大数据时代计算机信息处理技术的发展趋势。事实上，未来计算机网络的发展理念是将计算机硬件和网络数据分开，实现将目前的云计算转

化为云计算机网络。未来的计算机会与信息网络形成大数据网络系统，两者不可分离。

本章将着重介绍大数据相关的背景和基础知识，包括：数据的定义与属性、大数据概念与定义、大数据和小数据、结构化数据和非结构化数据、大数据信息处理技术及其应用等。本章内容为后面的章节做了基础铺垫。

1.1 数据的定义与属性

数据是信息的表现形式和载体，可以是符号、文字、数字、语音、图像、视频等。数据和信息是不可分离的，数据是信息的表达，信息是数据的内涵。数据本身没有意义，数据只有对实体行为产生影响时才成为信息。总的来说，数据是事实或观察的结果，是对客观事物的逻辑归纳，是用于表示客观事物的未经加工的原始素材。数据可以是连续的值，如声音、图像，称为模拟数据；也可以是离散的值，如符号、文字，称为数字数据。

在计算机系统中，各种字母、数字符号的组合、语音、图形、图像等统称为数据，数据经过加工后就成为信息。在计算机科学中，数据是指所有能输入计算机并被计算机程序处理的符号的介质的总称，是用于输入电子计算机进行处理，具有一定意义的数字、字母、符号和模拟量等的通称。

1.2 大数据概念与定义

近年来，大数据迅速发展成为科技界和企业界甚至世界各国政府关注的热点。《自然》（Nature）和《科学》（Science）等期刊相继出版专刊专门探讨大数据带来的机遇和挑战。对于大数据，研究机构 Gartner 给出了这样的定义：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。麦肯锡全球研究所给出的定义是：一种规模大到在获取、存储、管理、分析方面大大超出了传统数据库软件工具能力范围的数据集合，具有海量的数据规模、快速的数据流转、多样的数据类型和价值密度低四大特征。麦肯锡还认为，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。大数据已成为社会各界关注的新焦点，大数据时代已然来临。

所谓的大数据，顾名思义就是数据量巨大的意思，指的是信息的数据量巨大，以目前的计算机主流软件无法在短时间内实现对其进行获取、处理、存储、传输等管理功能，以

便为客户提供合理的信息技术服务。对于数据量巨大到什么程度，业内目前还没有统一的标准，一般认为数据量在 10 TB～1 PB（1 TB=1024 GB，1 PB=1024 TB）以上。

从宏观世界角度来讲，大数据是融合物理世界（Physical World）、信息空间和人类社会（Human Society）三元世界的纽带，因为物理世界通过互联网、物联网等技术有了在信息空间（Cyberspace）中的大数据反映，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，大数据还是新一代信息技术产业的强劲推动力。所谓新一代信息技术产业，其本质上是构建在第三代平台上的信息产业，主要是指大数据、云计算、移动互联网（社交网络）等。从社会经济角度来讲，大数据是数字经济的核心内涵和关键支撑。

相较于传统的数据，人们将大数据的特征总结为五个 V，即体量大（Volume）、速度快（Velocity）、模态多（Variety）、难辨识（Veracity）和价值大密度低（Value）。

根据来源的不同，大数据大致可分为如下几类。

（1）来自人们在互联网上的活动，以及使用移动互联网过程中产生的各类数据，包括文字、图片、视频等信息；

（2）来自各类计算机信息系统产生的数据，以文件、数据库、多媒体等形式存在，也包括审计、日志等自动生成的信息；

（3）来自各类数字设备所采集的数据，如摄像头产生的数字信号、医疗物联网中产生的人的各项特征值、天文望远镜所产生的大量数据等。

大数据的主要难点并不在于数据量大，因为通过对计算机系统的扩展可以在一定程度上缓解数据量大带来的挑战。其实，大数据真正难以对付的挑战来自数据类型多样（Variety）、要求及时响应（Velocity）和数据的不确定性（Veracity）。数据类型多样使得一个应用往往既要处理结构化数据，同时还要处理视频、语音等非结构化数据，这对现有数据库系统来说是难以应付的；在快速响应方面，在许多应用中时间就是利益；在数据的不确定性方面，数据真伪难辨是大数据应用的最大挑战。追求高数据质量是对大数据的一项重要要求，最好的数据清理方法也难以消除某些数据固有的不可预测性。为了应对大数据带来的上述困难和挑战，以 Google、Facebook、Linkedin、Microsoft 等为代表的互联网企业在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得以迅速发展，并逐渐被广泛应用于不同的行业和领域。

目前大数据分析应用于科学、医药、商业等各个领域，用途差异巨大，但其目标可以归纳为如下几类。

（1）获得知识与推测趋势。人们进行数据分析由来已久，最初且最重要的目的就是获得知识、利用知识。由于大数据包含大量原始、真实信息，大数据分析能够有效地摒弃个体差异，帮助人们透过现象、更准确地把握事物背后的规律。基于挖掘出的知识，可以更准确地对自然或社会现象进行预测。典型的案例是 Google 公司的 Google Flu Trends 网站，它通过统计人们对流感信息的搜索，查询 Google 服务器日志的 IP 地址判定搜索来源，来发布对世界各地流感情况的预测；又如，人们可以根据 Twitter 信息预测股票行情等。

（2）分析掌握个性化特征。个体活动在满足某些群体特征的同时，也具有鲜明的个性化特征，正如“长尾理论”中那条细长的尾巴那样，这些特征可能千差万别。企业通过长时间、多维度的数据积累，可以分析用户行为规律，更准确地描绘其个体轮廓，为用户提供更好的个性化产品和服务，以及更准确的广告推送。例如，Google 通过其大数据产品对用户的习惯和爱好进行分析，帮助广告商评估广告活动效率，预估在未来可能存在高达数千亿美元的市场规模。

（3）通过分析辨识真相。错误信息不如没有信息，由于网络中信息的传播更加便利，所以网络虚假信息造成的危害也更大。例如，2013 年 4 月 24 日，美联社 Twitter 账号被盗，发布虚假消息称奥巴马总统遭受恐怖袭击受伤，虽然虚假消息在几分钟内被禁止了，但是仍然引发了美国股市短暂跳水。由于大数据来源广泛及其多样性，它在一定程度上可以帮助实现信息的去伪存真，目前人们已开始尝试利用大数据进行虚假信息识别。例如，社交点评类网站 Yelp 利用大数据对虚假评论进行过滤，为用户提供更为真实的评论信息；Yahoo 和 Thinkmail 等利用大数据分析技术来过滤垃圾邮件。

1.3 大数据和小数据

数据技术是一个不断完善的过程，经历了由无数据到小数据、由小数据到大数据的演变。在数据采集、存储、传输、处理、安全等技术环节取得全面突破的前提下，大数据由空想走向理想，由理想走向现实。大数据与小数据判断原则如下。

- 数据的量；
- 数据的种类、格式；
- 数据的处理速度；
- 数据的复杂度。

很多事情在小规模数据的基础上是无法完成的，小数据是对数据价值的全面肯

定，它使人类行为摆脱了对经验的依赖，使人类的决策由主观性开始走向客观性，是人类智慧对蒙昧的一次重要胜利。但是小数据不过是人类的权宜之计，随着数据采集技术、存储技术、传输技术、处理技术和安全技术的全面创新，人类正在告别小数据时代，走向大数据时代。大数据相对于小数据，是一种批判式继承，既继承了小数据的优秀，又创造性地开创了全新的大数据范式。大数据时代只是刚刚开启，数据技术尚需进一步完善。从小数据向大数据进化的路径已经清晰，我们需要的仅仅是耐心的等待，在不完善的大数据中去发现问题，最终实现理想中的大数据。我们应该以“未来大数据”看待“现实大数据”，在这个阶段，“谁拥有大数据”比“怎么使用大数据”更重要。

1.4 结构化数据和非结构化数据

在信息社会，信息可以划分为两大类：一类信息能够用数据或统一的结构加以表示，我们称之为结构化数据，如数字、符号；而另一类信息无法用数字或统一的结构表示，如文本、图像、声音、网页等，我们称之为非结构化数据。结构化数据属于非结构化数据，是非结构化数据的特例。

小数据是以“人力为主、机器为辅”的运行模式，在数据的采集、存储、传输和处理中大量地依赖人力资源。因此，小数据在数据类型上，只能采用人类能够识别的文字、图片、声音、视频等结构化数据。但是并不是所有的社会事物都能够通过结构化语言来进行描述的，还存在着大量的非结构化语言。大数据是以“机器为主、人力为辅”的运行模式，计算机等各类数据设备成为数据采集、存储、传输和处理的主体，人力只在模型设计、参数设置、编辑矫正等环节发挥作用。大数据能够处理的数据来源更加广泛[11]，不仅包括结构化数据，而且包括只有机器方能处理的非结构化数据。例如，Cookie 等非结构化数据，是计算机等智能化设备所能处理的数据类型，它们的出现使人类逐渐摆脱了“语言困境”。

随着网络技术的发展，特别是 Internet 和 Intranet 技术的飞速发展，使得非结构化数据的数量日趋增大。这时，主要用于管理结构化数据的关系型数据库的局限性暴露得越来越明显了。因而，数据库技术相应地进入了“后关系型数据库时代”，发展进入基于网络应用的非结构化数据库时代。所谓非结构化数据库，是指数据库的变长记录由若干不可重复和可重复的字段组成，而每个字段又可由若干不可重复和可重复的子字段组成。简单地说，非结构化数据库就是字段可变的数据库，用它不仅可以处理结构化数据（如数