大数据应用分析技术与方法

刘汝焯 戴佳筑 何玉洁 编著

清华大学出版社

大数据应用分析技术与方法

刘汝焯 戴佳筑 何玉洁 编著

清华大学出版社
北京

内 容 简 介

本书强调了大数据的宝贵价值，介绍了常用的数据分析技术与方法，论述了大数据分析的思维特征，紧扣大数据的特点演示了可视化分析与可视化挖掘的方法，详细讨论了数据清洗与元数据管理，对大数据的风险予以充分揭示，同时提出了大数据风险管理的对策，对大数据治理作了简介。

本书具有很强的实用性、可操作性和指导性，对于企业管理人员、企业数据分析人员、业务分析人员和市场营销人员，政府监管机构如证监会、银监局、保监会的监管人员，审计师、注册会计师，纪检监察和司法机关执纪执法人员有参考价值，同时可供高等院校相关专业的师生参阅。

图书在版编目(CIP)数据

大数据应用分析技术与方法/刘汝焯，戴佳筑，何玉洁编著. —北京：清华大学出版社，2018
ISBN 978-7-302-48707-4

Ⅰ. ①大…　Ⅱ. ①刘…　②戴…　③何…　Ⅲ. ①数据处理　Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2017)第 265959 号

责任编辑：王　青
封面设计：何凤霞
责任校对：宋玉莲
责任印制：杨　艳

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：三河市铭诚印务有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：13　　**字　　数**：299 千字
版　　次：2018 年 1 月第 1 版　　**印　　次**：2018 年 1 月第 1 次印刷
印　　数：1～3000
定　　价：39.00 元

产品编号：065688-01

前　言

随着大数据的迅猛发展和日益普及，越来越多的与数据分析有关的人员，如企业管理人员、企业数据分析人员、业务分析人员、市场营销人员，政府监管机构如证监会、银监局、保监会的监管人员，审计师、注册会计师，纪检监察和司法机关执纪执法人员等需要掌握大数据应用分析技术与方法，迫切需要从大数据中挖掘有用的信息，提升工作水平和工作效率。这是信息化发展提出的必然要求，尤其是在业务与信息技术密切融合的形势下，这种需求越来越强劲。适应这种需求，我们在编著这本书时，着力突出实用性、可操作性和指导性。

实用性。贴近大数据发展的现状和趋势，全书共安排了 10 章内容，强调了大数据的宝贵价值，介绍了常用的数据分析技术与方法，紧扣大数据的特点演示了可视化分析与可视化挖掘的方法，详细讨论了数据清洗与元数据管理，对大数据的风险予以充分揭示，同时提出了大数据风险管理的对策，对大数据治理作了简介。

可操作性。在介绍大数据应用分析技术和方法时，由浅入深，逐步引导，屏蔽技术细节，让读者直接进入业务应用的层面，熟练掌握操作。尤其是全书从大数据分析的应用实践中精选了大量案例，进行了生动讲解。这些案例是大数据分析实践中的可贵探索和经验总结。通过案例的操作可以更好地引导读者加深对理论部分的理解，掌握分析技术与方法。

指导性。本书创新性地把大数据应用分析划分为器、技、道和美四个层面。器，指大数据分析的硬件和软件；技，指大数据分析的技术和方法；道，指大数据分析的思维方式；美，指审美体验、感觉和想象力。在开展大数据分析时需要硬件和软件、需要技术方法，这是毋庸讳言的。但大数据数量巨大、类型繁多、来源复杂，而且很多过去从来没有遇到过，单靠工具和技术方法是不能胜任大数据分析的多变情况的，清晰的分析思路、科学的思维方式显得更为重要，具有更强的更普遍的指导性。本书详细介绍了特征发现的思维方式，通过案例介绍了特征枚举、特征捕捉、特征分析的实际应用，同时对大数据分析中如何结合审美体验，张开想象的翅膀，激发分析的灵感，打开分析的思路给予了必要强调。

为天天和大数据打交道的人尽快掌握大数据分析的实用技能助一臂之力，为天天使用大数据的人通过最简单的路径掌握大数据分析的技能提供支持和帮助，这是我们的初衷。为了这个初衷我们确实努力了。大数据在发展中，加之编著者的水平和经验有限，书中有些问题的研究还不透彻，有些内容还有待于在实践中检验和完善，还有些可能本身就是存在问题的，这也是在所难免的。真诚希望广大读者批评指正！

参加本书撰写的有刘汝焯、戴佳筑、何玉洁。刘汝焯设计了全书的章节，撰写了其中第2、4、7章，对全书进行了统稿。戴佳筑撰写了第1、3、6、10章和附录B。何玉洁撰写了第5、8、9章和附录A，对全书的书稿进行了统一修订。

刘汝焯

2017年6月11日于北京

目　录

第 1 章　大数据是信息社会的宝贵资源 …… 1
1.1　大数据产生的背景和概念 …… 1
1.2　大数据的特征 …… 3
1.3　大数据与传统数据的区别 …… 4
1.4　大数据的价值和开发应用 …… 5
1.5　大数据时代的新机遇和新挑战 …… 8
1.5.1　依据大数据进行决策成为一种新的决策方式 …… 8
1.5.2　大数据与各行业深度融合带来层出不穷的新应用 …… 8
1.5.3　大数据推动新技术的不断涌现 …… 9
1.6　本书的特定视野 …… 10
参考文献 …… 11

第 2 章　大数据应用分析 …… 12
2.1　大数据的处理流程 …… 12
2.2　大数据分析的概念 …… 14
2.3　大数据分析的关键技术 …… 15
2.3.1　云计算 …… 15
2.3.2　数据分析方法 …… 16
2.3.3　数据可视化 …… 17
2.4　大数据分析工具介绍 …… 17
2.4.1　Hadoop …… 18
2.4.2　R …… 19
2.4.3　Python …… 19
2.4.4　RapidMiner …… 20
2.4.5　Tableau …… 20
2.5　大数据分析示例——查处虚假出口贸易 …… 22
2.5.1　案例概述 …… 22
2.5.2　查询分析 …… 23
2.5.3　可视化分析 …… 25
2.5.4　分析小结 …… 27
参考文献 …… 30

第 3 章　常用数据分析与预测方法 …… 31
3.1　方差分析 …… 31
3.1.1　分析方法 …… 31
3.1.2　示例介绍 …… 31
3.1.3　示例分析 …… 33
3.1.4　结果分析与总结 …… 35
3.2　相关分析 …… 35
3.2.1　分析方法 …… 35
3.2.2　示例介绍 …… 36
3.2.3　示例分析 …… 37
3.2.4　结果分析与总结 …… 40
3.3　回归分析 …… 40
3.3.1　分析方法 …… 40
3.3.2　示例介绍 …… 41
3.3.3　示例分析 …… 41
3.3.4　结果分析与总结 …… 42
3.4　时间序列分析 …… 44
3.4.1　平稳性检验 …… 44
3.4.2　纯随机性检验 …… 44
3.4.3　适用性检测 …… 44
3.5　聚类分析 …… 45
3.6　可视化数据分析 …… 46
3.6.1　常用的可视化数据展示方法 …… 47
3.6.2　可视化分析示例 …… 51
3.7　环境准备 …… 61
参考文献 …… 62

第 4 章　大数据分析的思维特征 …… 63
4.1　大数据应用分析的实务框架 …… 63
4.1.1　大数据应用分析的四个层面 …… 63
4.1.2　四个层面的关系 …… 65
4.2　大数据分析的特征发现 …… 65
4.2.1　特征发现的案例 …… 66
4.2.2　特征发现的概念 …… 73
4.3　对数据的分类 …… 73
4.4　特征发现的一般过程 …… 79
参考文献 …… 81

第 5 章　大数据的可视化分析 …… 82
5.1　不良贷款分析 …… 82
5.1.1　数据准备 …… 82
5.1.2　各银行的不良贷款情况分析 …… 86
5.1.3　各经济类型的企业的不良贷款情况分析 …… 95
5.1.4　各类贷款的不良贷款情况分析 …… 99
5.2　保险公司客户索赔分析 …… 103
5.2.1　数据准备 …… 103
5.2.2　数据分析 …… 104
参考文献 …… 119

第 6 章　可视化挖掘分析 …… 120
6.1　挖掘分析在审计线索特征发现中的应用 …… 120
6.1.1　案例背景 …… 120
6.1.2　数据准备 …… 120
6.1.3　聚类分析 …… 122
6.2　挖掘分析在推荐系统中的应用 …… 131
6.2.1　案例背景 …… 131
6.2.2　数据准备 …… 131
6.2.3　构建推荐系统 …… 132

第 7 章　大数据资源的元数据管理 …… 140
7.1　元数据简介 …… 140
7.1.1　元数据和对象数据 …… 140
7.1.2　应用元数据管理技术的意义 …… 140
7.2　著录对象分析 …… 142
7.2.1　审计中间表 …… 142
7.2.2　审计分析模型 …… 142
7.2.3　审计专家经验 …… 143
7.2.4　审计情景案例 …… 144
7.2.5　被审计单位资料 …… 144
7.3　元数据结构设计 …… 145
7.3.1　审计中间表的元数据结构 …… 145
7.3.2　审计分析模型的元数据结构 …… 146
7.3.3　审计专家经验的元数据结构 …… 147
7.3.4　审计情景案例的元数据结构 …… 149
7.3.5　被审计单位资料的元数据结构 …… 150
7.4　应用大数据审计分析数字信息元数据规范的扩展规则 …… 151

参考文献…… 152

第 8 章　大数据分析的数据清洗…… 153
8.1　大数据清洗的基本概念 …… 153
8.1.1　大数据清洗的基本架构…… 153
8.1.2　数据清洗的基本步骤…… 154
8.2　数据清洗 …… 157
8.2.1　数据清洗的一些注意事项…… 157
8.2.2　常见的数据清洗…… 158
参考文献…… 163

第 9 章　大数据分析的风险与对策…… 164
9.1　大数据分析的风险及产生原因 …… 164
9.2　大数据采集的风险 …… 165
9.3　大数据处理与集成的风险 …… 167
9.4　大数据分析的风险 …… 168
9.5　大数据解释的风险 …… 168
9.6　大数据的隐私和安全风险及其对策 …… 169
9.6.1　大数据处理流程的隐私风险…… 170
9.6.2　大数据处理平台带来的安全和隐私风险…… 172
9.6.3　保护大数据隐私和安全的对策…… 173
参考文献…… 175

第 10 章　大数据治理简介 …… 177
10.1　大数据治理的必要性…… 177
10.2　大数据治理的概念…… 178
10.3　大数据治理的核心内容…… 180
10.4　案例…… 181
10.4.1　工作思路 …… 182
10.4.2　数据真实性的验证方法 …… 182
10.4.3　数据完整性的验证 …… 186
参考文献…… 187

附录 A　Tableau 10.0 简介 …… 188
A.1　Tableau 工作区 …… 188
A.1.1　工作表工作区 …… 189
A.1.2　仪表板工作区 …… 190
A.1.3　故事工作区 …… 191

A.2 Tableau 的文件管理 …… 192

附录 B RapidMiner 使用方法简介 …… 194

B.1 RapidMiner 的主界面 …… 194

B.2 使用 RapidMiner 分析数据的方法 …… 195

第 1 章　大数据是信息社会的宝贵资源

1.1　大数据产生的背景和概念

大数据是随着信息数据快速增长和网络计算技术迅猛发展而兴起的一个新概念。大数据通过对海量数据的收集、处理和展示，揭示规律，预测未来。大数据能够帮助企业从海量数据中挖掘用户的需求，从而使数据真正产生价值。随着大数据的发展，其应用已经渗透到农业、工业、商业、服务业和医疗领域等各个方面。

随着计算机信息技术的发展和网络的普及，以博客、社交网络、基于位置的服务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、移动互联网等技术的兴起和普及，数据正以前所未有的速度在不断地增长和累积，特别是进入 DT(数据技术)时代，在线数据存储和计算量以及人类在日常学习、生活、工作中产生的数据量正以指数形式增长，呈现“爆炸”状态。国际数据公司(IDC)的研究结果表明，2008 年全球产生的数据量为 0.49ZB(1024GB=1TB，1024TB=1PB，1024PB=1EB，1024EB=1ZB)，2009 年的数据量为 0.8ZB，2010 年增长为 1.2ZB，2011 年的数量更是高达 1.82ZB，相当于全球每人每年产生 200GB 以上的数据。而到 2012 年为止，人类生产的所有印刷材料的数据量是 200PB。2014 年，全球产生的数据量估计已经达到了 3.6ZB。

全球信息数据量的飞速膨胀成为大数据产业存在并发展的基础。国际数据公司(IDC)预计，未来全球数据总量增长率将维持在 50%左右，到 2020 年全球数据总量将达到 40ZB，其中，我国将达到 8.6ZB，占全球的 21%。中国信息产业研究院的数据显示，2014 年我国大数据市场规模约为 116 亿元，同比增长 38%。预计未来几年，随着应用效果的逐步显现，我国大数据市场规模还将维持 40%左右的高速增长。

除了迅速增长的数据洪流，数据的结构越来越趋于复杂化，除了传统数据库中的数据，还有文档、网页、图像、音频和视频等，而且后者所占的比例也越来越大。这些数据的量变到底有多大呢？2014 年产生了大约 5ZB(Zettabyte)字节的非结构化数据，到 2020 年预计将增加到大约 40ZB 字节的非结构化数据。如图 1-1 所示为非结构化数据 2005—2020 年的实际和预期增长对比，该图片引自 Evangelos Simoudis 的“认知应用：大数据的下个转折点”一文。

这些数量巨大、种类繁多、结构复杂的数据早已远远超越了传统技术所能处理的范畴，如何合理、高效、充分地管理和使用这些数据，使之能够给人们的生活和工作带来更大的效益和价值，逐渐成为人们的共识，在这种背景下，大数据应运而生。

什么是大数据呢？大数据一词源于英文的“Big Data”，以前也有类似的词语，如“海量数据”“信息爆炸”等，但似乎都很难准确描述这个词的具体内涵。目前国内外对大数据

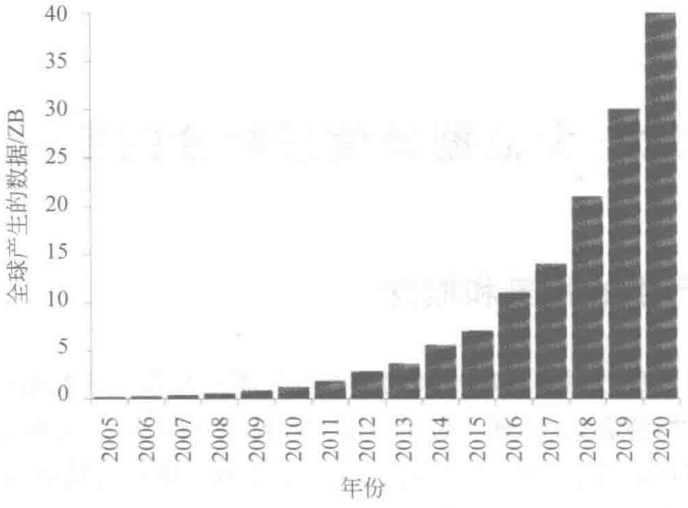

图 1-1　非结构化数据 2005—2020 年的实际和预期增长对比

没有一个统一的定义，国内外政府机构、企业和专家从不同角度给出了大数据的定义。维基百科对大数据的定义是“大数据是数据规模巨大，通过目前主流软件工具无法在合理时间内捕获、管理、处理并整理成为帮助经营决策的数据集”；美国国家标准和技术研究院(NIST)则认为“大数据是指由于数据的容量、数据的获取速度或者数据的表示限制了使用传统关系方法对数据的分析处理能力，需要使用扩展的机制以提高数据处理效率的技术”；著名的管理咨询公司麦肯锡公司的研究报告中将大数据定义为“超过了传统数据库软件工具捕获、存储、管理和分析能力的数据集”；国际数据公司(IDC)是研究大数据及其影响的先驱，在其 2011 年的报告中指出“大数据技术描述了一个技术和体系的新时代，被设计用于从大规模、多样化的数据中通过高速捕获、发现和分析技术提取数据的价值”。著名的大数据专家维克托·迈尔—舍恩伯格在其经典著作《大数据时代》中，指出大数据“是当今社会所独有的一种新型能力，以一种前所未有的方式，通过对海量数据进行分析，获得有巨大价值的产品和服务，或深刻的洞见。”

大数据中的海量数据有三个主要来源，首先是海量交易数据。随着信息技术的广泛应用，越来越多的企业和机构比以往任何时候都依赖信息系统，如超市的销售记录系统、火车售票系统、银行的交易记录系统、医院病人的医疗记录等，由此产生了大量的交易数据。其次是海量的网络信息。互联网的诞生促使人类社会数据量出现一次巨大的飞跃，但是真正的数据爆发产生于移动互联网时代特别是社交媒体的兴起，这类数据近几年一直呈现爆炸性的增长，涵盖了海量的聊天记录、Web 网页、电子邮件、图片、视频、音频等。最后是海量的感知数据。物联网(The Internet of Things)是新一代信息技术的重要组成部分，通过传感器和网络技术实现了物与物、人与物、人与人之间的互联。物联网时代，除了智能手机、平板电脑等常见的客户终端之外，更多更先进的传感设备和智能设备，如智能手表、智能眼镜、智能汽车、智能电视、工业设备和手持设备等都将接入网络，由此产生的海量感知数据量及其增长速度比以往任何时期都要多。

近几年，大数据迅速成为科技界和企业界甚至世界各国政府关注的热点，发展的势头

不可阻挡。著名的科技旗舰杂志《自然》和《科学》等相继出版专刊，分别从互联网技术、互联网经济学、超级计算、环境科学、生物医药等多个方面专门探讨大数据带来的机遇和挑战。2011 年 5 月麦肯锡公司在美国拉斯维加斯举办的第 11 届 EMC World 年度大会上称："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"，并于 2012 年 3 月 29 日发布了"大数据研究发展倡议"，正式启动"大数据发展计划"。

我国十分重视大数据的发展。2012 年 8 月，中国科学院启动了"面向感知中国的新一代信息技术研究"战略性先导科技专项，2013 年，科技部正式启动 863 项目"面向大数据的先进存储结构及关键技术"，这些科研项目的任务之一就是研制用于大数据采集、存储、处理、分析和挖掘的未来数据系统。国务院于 2014 年 8 月发布了《国务院关于加快发展大数据、呼叫中心等生产性服务业的指导意见》，从国家层面推动大数据的建设和发展；2015 年 7 月，国务院办公厅印发了《关于运用大数据加强对市场主体服务和监管的若干意见》，要求在政府层面推动大数据的应用；2015 年 9 月国务院颁布了《国务院关于印发促进大数据发展行动纲要》，提出"全面推进我国大数据发展和应用，加快建设数据强国"的方针政策，这标志着大数据建设和应用已经上升为国家发展的长期战略。

1.2　大数据的特征

目前大数据尚未具有统一的描述，不同的定义基本上都是从大数据的特征出发，通过大数据特征的阐述和归纳试图给出其定义。大家都普遍认同大数据具有 4 个基本特征，分别是容量大(Volume)、种类多(Variety)、高速度(Velocity)和价值密度低(Value)，由于这四个特征的英文单词都是以英文字母"V"开头，所以又将其称为大数据的"4V 特征"。

容量大是指大数据的数据量非常巨大。例如，互联网搜索的巨头谷歌现在能够处理的网页数量是在千亿以上，每月处理的数据量超过 400PB(400×10^{15}B)，并且呈继续高速增长的趋势；百度目前数据总量接近 1 000PB($1\ 000\times10^{15}$B)，存储网页数量接近 1 万亿，每天大约要处理 60 亿次搜索请求。

种类多是指大数据的数据种类繁多，结构复杂。在大数据时代，数据来源并非仅仅是计算机产生的信息或者人们在互联网上发布的信息，全世界的工业设备、汽车、电表上有着无数的数码传感器，随时测量和传递有关位置、运动、振动、温度、湿度乃至空气中化学物质的变化等，也产生了海量的数据信息。这些数据既包含传统关系数据库中保存的结构化数据，也包含图像、声音和视频等非结构化数据以及 HTML 网页和 XML 文档等半结构化数据，而且非结构化数据和半结构化数据所占的比例呈现越来越大的趋势。

高速度是指大数据能够更快地满足实时性的需求。目前，对于数据智能化和实时性的要求越来越高，比如开车时会随时通过智能导航仪查询最佳路线，在餐厅吃饭时会查询其他用户对餐厅的评价和推荐的菜肴，见到有趣的事情或可口的食物会拍照发微博等诸如此类的人与人、人与机器之间的信息交流互动，这些都不可避免地带来数据交换，而数

据交换的关键是降低延迟，以近乎实时的方式完成数据交换的任务。

价值密度低是大数据特征里最关键的一点。数据量大并不意味着数据价值的增加，大数据时代数据的价值就像沙里淘金，其应用价值（金子）是隐藏在沙子之中的，数据量越大，里面真正有价值的东西所占的比例就会越少。大数据面临的一个挑战就是从这些TB、PB、EB级的海量数据中，提取有价值的信息，将信息转化为知识，发现规律，最终用知识促成正确的决策和行动。

另外，随着人们对大数据的研究不断深入，有的企业（如IBM公司）认为大数据还应具有第五个特征，即真实性（Veracity），通俗地讲，它是指大数据中数据来源广泛、种类繁多，这些数据具有不可靠或不精确的可能性。当我们试图获得大规模的数据时，必须能够控制这些不可靠或不精确带来的影响，使这些海量数据能够被用来更好地解释和预测客观世界。

1.3 大数据与传统数据的区别

从传统的数据库到大数据，不仅仅只是一个简单的技术演进，两者既有密切联系又有着本质上的差别。

大数据的出现颠覆了传统的数据管理方式，在数据来源、数据处理方式和数据思维等方面带来革命性的变化。为了说明传统的数据库和大数据的区别，有的专家使用“池塘捕鱼”和“大海捕鱼”的形象比喻。“鱼”是待处理的数据，“池塘捕鱼”代表传统数据库时代的数据管理方式，而“大海捕鱼”则对应着大数据时代的数据管理方式。“捕鱼”环境条件的变化导致了“捕鱼”方式的根本性差异，这些差异主要体现在如下几个方面。

（1）数据规模：“池塘”和“大海”最明显的区别就是规模不一样。“池塘”规模相对较小，“池塘”的处理对象通常以MB为基本单位，而“大海”的规模非常大，则常常以GB，甚至是TB、PB、EB为基本处理单位。

（2）数据类型：“池塘”中的数据种类往往仅仅有几种，这些数据又以结构化数据为主。而在“大海”中数据的种类繁多，这些数据不仅包含结构化数据，还包含半结构化数据以及非结构化的数据，并且半结构化数据和非结构化数据所占份额越来越大。

（3）模式和数据的关系：传统的关系数据库都是先有模式，然后才会产生数据。这就好比是先选好合适的“池塘”，然后才会向其中投放适合在该“池塘”环境生长的“鱼”。而大数据时代很多情况下难以预先确定模式，模式只有在数据出现之后才能确定，且模式随着数据量的增长处于不断的演变之中。这就好比“大海”中鱼的种类和数量都在不断地增长，鱼的变化会使大海的成分和环境处于不断变化之中。

（4）处理对象：在“池塘”中捕鱼，“鱼”仅仅是其捕捞对象。而在“大海”中，“鱼”除了是捕捞对象之外，还可以通过某些“鱼”的存在来判断其他种类的“鱼”是否存在。也就是说，传统数据库中数据仅作为处理对象，而在大数据时代，要将数据作为一种资源来辅助解决其他诸多领域的问题。

（5）处理方法：如果把“渔网”比作数据处理方法的话，捕捞“池塘”中的“鱼”，只需少数几种基本的“渔网”就可以应对，但是在“大海”中，不可能存在少数渔网能够捕获所有的

鱼类。传统意义上的数据处理方式包括数据挖掘、数据仓库、联机分析处理(OLAP)等，而在大数据时代，数据已经不仅仅是需要分析处理的内容，更重要的是人们需要借助专用的思想和手段从大量看似杂乱、繁复的数据中，收集、整理和分析数据，为人们在生产和生活中预测、决策和规划提供强有力的支持。

图灵奖获得者、著名数据库专家吉姆·格雷(Jim Gray)博士观察并总结在人类的科学研究史上，先后经历了实验、理论和计算三种研究方法。而在数据量不断增加和数据结构愈加复杂的今天，这三种方法在一些新的研究领域已经无法很好地发挥作用，所以吉姆·格雷博士提出了科学研究的第四种方法，即“数据探索”，通过大数据的分析和处理来指导科学研究。

(6) 存储方式：“池塘”大都采用关系型数据库保存数据，而“大海”的数据量巨大，关系型数据库已经不能容纳如此巨大的数据，目前只能采用非关系型数据库(如 NoSQL)或分布式文件系统(HDFS)来存储数据。

虽然大数据和传统数据库有本质的差异，但是二者又有密切的联系。首先，大数据不是否定传统的数据库，有些学者认为传统数据库是大数据的一个重要组成部分，大数据只是传统数据库处理能力的拓展和延伸；其次，有些著名的 IT 企业提出传统数据库和大数据是互补的关系，大数据中的结构化数据通过传统数据库能够获得更好的存储和处理；最后，虽然传统的数据库在处理当今海量复杂的数据方面遇到了严峻的挑战，但是它依然是今天主流的数据存储技术，大数据要代替传统数据库成为主流的存储技术尚需时日。

1.4　大数据的价值和开发应用

近几年，大数据迅速发展成为政府、企业界和学术界关注的热点。人们意识到，一个国家和企业拥有数据的规模和运用数据的能力将成为综合国力和企业竞争力的重要组成部分，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。世界 500 强的大公司认为大数据是“重要的生产因素”，而美国政府甚至把大数据称为“未来的新石油”。

毋庸置疑，大数据是待挖掘的金矿，其价值不言而喻。大数据的核心价值是什么呢？目前人们比较认同的有三个方面的价值。

首先，大数据改变了我们分析和使用数据的思维方式。《大数据时代》一书作者维克托·迈尔—舍恩伯格认为大数据时代处理和分析数据的思维有三大转变：第一个转变是在大数据时代可以分析更多的数据，甚至是相关的所有数据，而不再依赖少量的采样数据。在传统数据分析中，我们所做的是试图通过最少量的样本数据观测来发现规律。由于数据的采集、存储和分析的成本高，因此我们只能采用采样的方法。而在大数据时代，我们收集所有的数据，是与我们所研究的现象相关的所有可获得的数据，因此我们能够基于与某事物相关的所有数据展开数据分析，而不是仅仅依靠分析少量的数据样本。第二个转变是不再追求精确度。大数据时代数据是如此之多，以至于我们不再热衷于追求精确度。适当忽略数据的精确度，可以获得更广泛的数据，将带来更好的洞察力和更大的商业利益。第三个转变是不再热衷于寻找事物之间的因果关系，而是关注事物之间的相关关系。例如，成千上万的电子商务网站可以根据所记录的用户行为习惯，分析出用户喜爱

的产品或服务，然后对用户进行推荐，但是这些网站并不关心用户为什么会对这些产品和服务感兴趣。

其次，大数据提高了决策支持的能力。基于大数据的决策有两个主要特点：第一，不同于传统的基于少量数据样本的数据分析方法。大数据中的海量数据全面覆盖了企业经营以及政治、经济、社会、教育等方面的信息，通过对这些完整的信息进行分析，能够提高决策的质量；第二，决策的技术水平和效率大幅提高。云计算技术是大数据的重要支撑技术，通过云计算强大的计算能力和数据挖掘技术，人类不会被海量数据所淹没，能够高效率驾驭海量数据，获得有价值的决策信息。例如，在企业经营管理中，大数据能够帮助企业分析大量数据而进一步挖掘细分市场的机会，最终能够缩短企业产品研发时间，提升企业在商业模式、产品和服务上的创新力；学校和老师能够在对教学案例进行大数据分析的基础上改进他们的教学方法并合理安排教学内容；交管部门通过整合交通状况、天气以及驾驶员的地点信息等数据，可以更好地管理交通；大数据在政府和公共服务领域的应用，可以有效推动政府工作开展，提高政府部门的决策水平、服务效率和社会管理水平。

最后，通过大数据进行预测。《大数据时代》一书作者维克托·迈尔—舍恩伯格认为预测是大数据的核心，通过对大数据的分析来预测事情发生的可能性和发展的方向。例如，美国加州警方应用大数据进行预测分析，发现了犯罪趋势和犯罪模式，甚至可以对重点区域的犯罪概率进行预测；又如，前面提到的图灵奖获得者、著名数据库专家吉姆·格雷博士提出了第四种科学研究的方法——基于数据探索的方法，这种方法的本质就是基于大数据探索与发现自然和社会的规律。

大数据正日益对生产、流通、分配、消费活动以及经济运行机制、社会生活方式和国家治理能力产生重要影响。大数据的应用已逐步深入我们生活的方方面面，涵盖医疗、交通、金融、教育、体育、零售等各行各业。下面我们列举几个大数据应用的典型案例。

(1) 2014 年最热门的美剧非《纸牌屋》莫属。《纸牌屋》风靡北美乃至全球的一个重要原因，是大数据分析的结果。美国网飞(Netflix)公司是一家在线影片租赁提供商，该公司的网站有近 3 000 万订阅用户，这些用户在网站上收看视频的大量行为数据都被记录下来。据统计，用户每天在网飞上产生 3 000 多万个行为，包括暂停、回放、添加书签以及每天 300 万次搜索、400 万个评分。网飞对这些数据和收视调查等相关数据进行综合分析后发现，喜欢观看 BBC 老版《纸牌屋》的用户，大多喜欢大卫·芬奇导演或凯文·史派西主演的电视剧，于是网飞做出了拍摄《纸牌屋》的决策，投资 1 亿美元拍摄了新版《纸牌屋》，请大卫·芬奇执导、凯文·史派西做主演。结果，大数据技术让网飞公司赚得盆满钵溢。

(2) 无论是在国内还是国外，体育行业都蕴含巨大的商机。例如，美国职业篮球联赛(NBA)的纽约尼克斯队在 2013 年就产生了 2.87 亿美元的收入。各支球队为了最大化自己的收入，必须在球场上不断赢球，因此教练组和相关人员必须一直做出正确的决策。而在这些决策中，体育的大数据分析扮演了一个日益重要的角色。

2015—2016 美国 NBA 赛季，骑士从 1∶3 落后，到 4∶3 夺冠，创造了 NBA 总决赛的历史。但球员的爆发，大劣势下的逆转，这一切的发生都不是偶然的。大数据文摘发现，在 2015—2016NBA 总决赛最后一场，骑士队的后卫 JR 史密斯在场上很好地充当了球队

第三得分点，13 投 5 中得到 12 分 4 篮板 2 助攻，其中三分球 8 中 2。这样的例子其实在 NBA 的赛场上比比皆是，球员并不是机器，他们的语言、行为其实都无时无刻不在透露大量可被分析和深度挖掘的信息。如何有效地将这些信息转化为知识，又如何利用这些知识来帮助人们做正确的决策？

运用大数据的体育数据分析包括运用统计工具来分析球员的历史表现。球队老板凭借分析结果来组建球队，教练组结合分析结果和他们的专业知识来调整上场阵容，提高球员的赛场表现。比如，利用非结构化社交媒体数据来提升现有体育分析模型效率，通过自然语言处理和文本挖掘技术来分析 NBA 球员的推文以了解他们的赛前情绪，从而提高对球员赛场表现的预测的准确性。

比如，2016 年 5 月 9 日西部半决赛第四场，雷霆主场战胜马刺，成功扳平大比分。而当地时间是母亲节，杜兰特全场出场 43 分钟，拿下 41 分，5 篮板，4 助攻，成为球队取胜的关键。众所周知，杜兰特与母亲感情非常好，其第一次荣获常规赛 MVP 发表演讲时，更是着重描述了童年时母亲的不易以及与母亲感情的深厚。而在比赛前，两队的明星球员中，只有杜兰特特意发表推文“So proud of my mama”，以此来表达对母亲的感谢，这也就不难解释杜兰特在本场比赛的爆发了。①

(3) 2015 年 5 月，美国费城外一列美国铁路公司火车在一处急转弯路段发生脱轨事故，造成 5 人死亡和超过 200 人受伤。在费城到纽约的这一常用路段上，此次事故的发生显得非比寻常。次日早晨，半岛电视台美国频道发布了脱轨前火车的准确行驶速度：每小时 106 英里(约合每小时 170 千米)，这超过了该路段限速(每小时 80 千米)的 2 倍之多。

之所以能如此迅速地做到这一点，是因为在此事发生的一年之前，他们就已经开始仔细调查美铁列车，设计了追踪其行驶的地图，每隔 5 分钟收集和存储一次数据。数据可以提供国内每列火车的实时定位和行驶速度。因此，通过找到事故发生之前的定位，他们在一张交互式的注释图中准确定位了该趟列车的行驶轨道。在后续追踪和分析从同一弯道通过的几百趟火车的行驶数据后，他们发现大部分火车的行驶速度都低于 50 英里/小时，而出事的火车却是一个特例。

该报道获得了 2016 年全球数据新闻奖(DJA)年度最佳突发新闻数据使用奖。

(4) 淘宝目前占据中国网络购物 75%的市场份额，每天产生的数据量达到了 7 T (7 000G)。这些数据当中大部分是由消费者、商家产生的交易数据，包括交易时间、商品价格、购买数量等，更重要的是，这些信息可以与客户和商家的年龄、性别、地址，甚至兴趣爱好等个人特征信息相匹配。阿里巴巴集团董事局主席马云表示，阿里巴巴公司本质上是一家数据公司，做淘宝不是为了卖货，而是为了获得所有零售的数据和制造业的数据；做物流不是为了送包裹，而是为了将这些数据融合在一起。淘宝数据魔方是淘宝网的大数据分析平台，通过这一平台，商家可以了解淘宝网上的行业宏观情况和自己品牌的市场状况，也可以分析竞争对手，探究消费买卖行为等，并据此进行生产、库存决策，而与此同时，更多的消费者也能以更优惠的价格买到更心仪的宝贝。另外，阿里信用贷款则是阿里巴巴通过所掌握的企业交易数据，借助大数据技术自动分析判定是否给予企业贷款，全程

① 该示例引自《大数据文摘》2016 年 6 月 21 日，“如何利用 NBA 球员推文预测其球场表现？”