“十三五”
国家重点出版物出版规划项目

国之重器出版工程
国防现代化建设

航天先进技术研究与应用系列

Big Data Cleaning

大数据清洗技术

王宏志 著

中国工信出版集团
哈爾濱工業大學出版社
HARBIN INSTITUTE OF TECHNOLOGY PRESS

航天先进技术研究与应用系列

大数据清洗技术

Big Data Cleaning

王宏志 著

中国工信出版集团
哈爾濱工業大學出版社
HARBIN INSTITUTE OF TECHNOLOGY PRESS

内 容 简 介

本书主要介绍了大数据清洗方面的研究成果。全书共分7章，重点面向大数据清洗中计算困难、错误混杂、缺少知识等难题，针对实体识别、真值发现、缺失值填充、不一致检测与修复等问题提出了相应的技术和算法，并在第7章提出了多数据质量问题综合清洗与优化技术。

本书可作为高等院校和科研机构大数据、数据质量管理、数据治理等方面的教学和科研参考书。

图书在版编目(CIP)数据

大数据清洗技术/王宏志著. —哈尔滨：哈尔滨工业大学出版社，2020.1

国之重器出版工程. 航天先进技术研究与应用系列

ISBN 978-7-5603-7753-7

Ⅰ.①大… Ⅱ.①王… Ⅲ.①数据处理 Ⅳ.①TP274

中国版本图书馆CIP数据核字(2018)第253125号

大数据清洗技术

DASHUJU QINGXI JISHU

策划编辑 张 荣

责任编辑 刘 瑶 张 荣

出 版 哈尔滨工业大学出版社

社 址 哈尔滨市南岗区复华四道街10号 邮编150006

传 真 0451-86414749

网 址 http://hitpress.hit.edu.cn

印 刷 固安县铭成印刷有限公司

开 本 710mm×1000mm 1/16 印张20 字数390千字

版 次 2020年1月第1版 2020年1月第1次印刷

书 号 ISBN 978-7-5603-7753-7

定 价 88.00元

《国之重器出版工程》
编辑委员会

专家委员会委员（按姓氏笔画排列）：

于　全　中国工程院院士

王少萍　“长江学者奖励计划”特聘教授

王建民　清华大学软件学院院长

王哲荣　中国工程院院士

王　越　中国科学院院士、中国工程院院士

尤肖虎　“长江学者奖励计划”特聘教授

邓宗全　中国工程院院士

甘晓华　中国工程院院士

叶培建　中国科学院院士

朱英富　中国工程院院士

朵英贤　中国工程院院士

邬贺铨　中国工程院院士

刘大响　中国工程院院士

刘怡昕　中国工程院院士

刘韵洁　中国工程院院士

孙逢春　中国工程院院士

苏彦庆　“长江学者奖励计划”特聘教授

苏哲子　中国工程院院士

李伯虎　中国工程院院士

李应红　中国科学院院士

李新亚　国家制造强国建设战略咨询委员会委员、中国机械工业联合会副会长

杨德森　中国工程院院士

张宏科　北京交通大学下一代互联网互联设备国家工程实验室主任

陆建勋　中国工程院院士

陆燕荪　国家制造强国建设战略咨询委员会委员、原机械工业部副部长

陈一坚　中国工程院院士

陈懋章　中国工程院院士

金东寒　中国工程院院士

周立伟　中国工程院院士

郑纬民　中国计算机学会原理事长

郑建华　中国科学院院士

屈贤明　国家制造强国建设战略咨询委员会委员、工业和信息化部智能制造专家咨询委员会副主任

项昌乐　“长江学者奖励计划”特聘教授，中国科协书记处书记，北京理工大学党委副书记、副校长

柳百成　中国工程院院士

闻雪友　中国工程院院士

徐德民　中国工程院院士

唐长红　中国工程院院士

黄卫东　“长江学者奖励计划”特聘教授

黄先祥　中国工程院院士

黄　维　中国科学院院士、西北工业大学常务副校长

董景辰　工业和信息化部智能制造专家咨询委员会委员

焦宗夏　“长江学者奖励计划”特聘教授

前 言

由于数据质量对数据的可用性起着决定性作用，因此一直以来数据质量管理是数据管理和数据治理的重要主题。数据清洗是提升数据质量的一种行之有效的手段，得到了学术界和产业界的广泛关注。20 世纪 90 年代以来，研究人员在数据清洗理论和技术方面开展了大量的研究，取得了很多研究成果。

随着大数据时代的到来，数据成为许多系统的核心，数据质量的重要性日益凸显。对大数据而言，由于其规模性、高速性、多样性等特征，其数据质量问题显得尤为突出。数据清洗成为保障大数据可用性、令大数据增值的重要技术。同时，大数据清洗也带来了计算困难、知识缺乏、错误混杂等一系列挑战性问题，这使得研究人员对大数据清洗的研究兴趣日益提升，成为数据管理等领域的热点研究问题之一。

遗憾的是，至今国内外尚无一部面向大数据清洗的学术著作问世。为了满足广大大数据研究、工程和应用工作者的需求，作者集中了国内外和大数据清洗相关的学术论文以及多年来的研究成果，经过近一年的系统研究整理，完成了本著作。

数据清洗的范畴非常广泛，既有语法层面的清洗，又有语义层面的清洗。本书主要介绍针对语义层面的清洗，而针对语法层面的清洗（如格式等）则不在本书的讨论范围之内。本书旨在为已经具备一定数据管理基础知识的读者介绍更系统、更深入的大数据清洗基础理论和使用技术。本书对大数据清洗的最新研究成果进行了梳理，并系统论述了大数据清洗的相关技术，包括面向大数据的实体识别、真值发现、缺失值填充和不一致数据检测与修复，还包括多种类型混合

错误的修复技术。

本书共7章。第1章介绍了大数据和数据质量的概念，分析了大数据为数据清洗带来的挑战，综述了数据清洗的研究进展，并概述了本书的内容。第2章介绍了大数据处理技术，包括大数据计算平台和众包平台。第3章论述了面向大数据的实体识别技术，包括高效串行实体识别算法、并行实体识别算法、增量实体识别算法及基于众包的实体识别技术。第4章论述了面向大数据的真值发现技术，包括并行真值发现、增量真值发现和基于众包的真值发现。第5章论述了面向大数据的缺失值填充技术，包括基于贝叶斯网络的串行缺失值填充算法、并行缺失值填充算法和基于众包的缺失值填充算法。第6章论述了不一致数据检测与修复，包括并行不一致数据检测与修复算法、基于众包的不一致数据检测与修复算法和扫描数据一次的大数据不一致检测算法。第7章论述了多数据质量问题综合清洗与优化技术，包括多数据质量维度的关联、基于任务合并的并行数据清洗优化技术，还介绍了支持多数据质量问题清洗的综合大数据清洗系统。

哈尔滨工业大学的李建中教授、高宏教授以及哈尔滨工业大学海量数据计算研究中心的诸位同事对本书的出版提出了宝贵建议。哈尔滨工业大学的叶晨、张美范、丁小欧、李宁宁、金连、贾立、谢晖、张安珍、门雪莹、甘小楚、李明达、霍然、张笑影、齐志鑫、孙铭、孔欣欣、李东升等同学，为本书的资料收集、整理提供了很大帮助，在此一并表示衷心的感谢。

非常感谢我的爱人黎玲利副教授，在给我家庭温暖的同时，作为一名数据清洗方面的学者，也对我的研究给出许多有益的建议和启发。感谢我的母亲和岳母，感谢她们对我研究工作的持续支持。感谢我的儿子壮壮，为我的生活带来了许多欢乐，也为我的研究带来了许多灵感。

本书的出版得到了2016年度黑龙江省精品图书出版工程资助项目的资助。作者关于大数据清洗方面的研究和本书的写作还得到了国家自然科学基金项目(编号：U1509216，U1866602，61472099)、国家重点研发计划项目(编号：2016YFB1000703)、黑龙江省留学回国人员基金(编号：LC2016026)和微软－教育部语言语音重点实验室的资助，在此表示感谢。

由于作者水平有限，本书疏漏及不妥之处在所难免，敬请同行和读者提出宝贵的建议，以期改进本书。作者邮箱：wangzh@hit.edu.cn。

作　者

2019年2月

目 录

第1章 绪 论

大数据指的是所涉及的数据量规模巨大到无法通过人工在合理时间内达到截取、管理、处理，并整理成为人类所能解读的形式的信息。数据质量有多种定义和多个不同的维度。数据清洗是解决数据质量问题行之有效的方法之一。大数据规模性、高速性、多样性、价值稀疏性等特点为大数据清洗带来了新的挑战。围绕这些挑战，本书提出了基于规范化流程的混杂错误修正技术、基于任务整合的分布式流程优化技术和基于群智计算的候选修复决策技术。

1.1 大数据的定义及其应用

回望21世纪以来的科学技术发展史，数据驱动的方法发挥了越来越重要的作用。2007年，图灵奖得主Jim Gray在“科学的第四个范式”中“预测数据探索式科研”将成为继实验、理论、计算之后的第四个人类科学研究范式。微软研究院出版的《第四范式：数据密集型科学发现》，从科学研究模式角度来分析大数据及其深远影响。正如预测所言，大数据计算在自然科学与社会科学的各个领域得到了广泛的应用。据英国著名报刊《每日电讯》报道，在2012年美国总统大选中，统计学家Silver依靠海量民意调查结果以及历史数据精准地预测出50个州的票数与得票率，精确率超过多组资深政治学专家和观察者。据《自然》杂志刊出的文章显示，谷歌公司通过分析每天数十亿条搜索指令，测试了4.5亿个不同的数学模型，准确及时地预测出流感传播情况。众多证据表明，大数据正以前所未有的速度和方式改变着科学研究与社会生活。

关于“大数据”的准确定义，科学界现在仍然缺乏统一的认识。从字面上理解，它最本质的特点在于数据量“大”，除此之外，还包括获取、管理及处理时的复杂性。“大数据”的概念起源于2008年9月《自然》杂志刊登的名为“Big Data”的专题。2011年《科学》杂志也推出专刊 *Dealing with Data* 对大数据的计算问题进行讨论。谷歌、雅虎等著名企业在此基础上，总结它们利用积累的海量数据为用户提供更加人性化服务的方法，进一步完善了“大数据”的概念。

根据维基百科上的定义，大数据指所涉及的数据量规模巨大到无法通过人工在合理时间内达到截取、管理、处理，并整理成为人类所能解读的形式的信息。

大数据具有明显的时代特征，习惯上将其总结为四个“V”：规模性(Volume)、高速性(Velocity)、多样性(Variety)和价值稀疏性(Value)。

1. 规模性(耗费大量存储、计算资源)

大数据之“大”，体现在数据的存储和计算均需耗费海量规模的资源：美国宇航局收集和处理的气候观察、模拟数据达到32 PB；谷歌公司索引的网页总数超过1万亿个；美国个人消费信用评估公司(FICO)的信用卡欺诈检测系统保护全世界超过18亿个活跃信用卡账户。

2. 高速性(增长迅速，亟须实时处理)

大数据另一个特点在于速度快：大型强子对撞机实验设备中包含了15亿个传感器，平均每秒收集超过4亿条实验数据；每秒钟超过3万次用户查询提交到谷歌；等等。

3. 多样性(来源广泛,形式多样)

在大数据背景下,数据在来源和形式上的多样性愈加凸显:除大量以非结构化形式存在的文本数据,也存在位置、图片、音频、视频等信息。除信息形式的多元化,信息的来源也表现出多样性:从网络日志、物联网、移动设备、传感器到基因图谱、医疗影像、天体运行轨迹、交通物流数据等。

4. 价值稀疏性(价值总量大,知识密度低)

大数据以其高价值吸引了广泛关注。据全球著名咨询公司麦肯锡报告:"如果能够有效地利用大数据来提高效率和质量,预计美国医疗行业每年通过数据获得的潜在价值可超过 3 000 亿美元,能够使得美国医疗卫生支出降低 8%。"虽然大数据的价值高,但是知识密度非常低。谷歌公司首席经济学家 Hal Varian 指出:"数据是广泛可用的,所缺乏的是从中提取出知识的能力。"IBM 副总裁兼 CTO Dietrich 表示:"可以利用 Twitter 数据获得用户对某个产品的评价,但是往往上百万条记录中只有很少的一部分真正讨论这款产品。"

大数据管理与分析是关系国民经济发展与国家安全的重大决策,是把握信息产业的制高点。2011 年 5 月,全球著名咨询公司麦肯锡撰文《大数据:下一个创新、竞争和生产力的前沿》提出:"通过利用大数据提高政府行政管理方面的运作效率,估计欧洲发达经济体可以节省开支超过 1 000 亿欧元;充分利用大数据的零售商有可能将其经营利润提高 60%以上。"2012 年 3 月,美国奥巴马政府科技办公室发布《大数据的研究与发展计划》,计划共投资两亿多美元,推动包括美国国家科学基金、美国国家卫生研究院、美国能源部、美国国防部、美国国防部高级计划局、美国地质勘探局在内的六个部门支持大数据研究。大数据被视为美国在"信息高速公路"之后又一个国家科技战略层面的重大决策。

概括地讲,开展大数据研究的意义在于:

(1)辅助社会管理决策。

从大数据中可以更加及时地获得利用其他方式获取时有延迟、误差的统计特征,进而建立相应的决策模型,辅助政策制定者有效地制定决策、观察反馈、调整优化。近年来,大数据在辅助社会管理决策方面的应用也逐渐为人们所接受。例如,2009 年甲型流感在美国爆发,受限于个人习惯与消息传播,美国国家疾控中心收集到的数据与流感病毒实际传播的情况有 1～2 周的延迟。针对这一情况,谷歌公司的工程师提出了一类通过分析搜索日志掌握甚至预测流感的爆发与传播情况的方法,他们利用对谷歌每天数十亿条搜索指令的分析,测试了 4.5 亿个不同的数学模型,得到了较为准确并比疾控中心数据及时的流感传播预测结果。

(2)推动科学发展。

大数据处理技术能够推动自然科学领域展开更好的研究。例如,2012 年 3 月,日本爆发大地震,9 分钟后,美国国家海洋和大气管理局即发布详细的海啸预警,并利用计算机模拟建立海啸模型。这项技术就是依托于从全球范围内遍布海底与海面的传感器中收集的数据,辅以合理的数学模型以及大数据处理分析策略。在医疗卫生领域,英国癌症研究院研发出一种新设备,用于分析大量癌症病例数据,以此了解细胞在癌症治疗过程中的功能及性能变化。在能源行业,Opower 公司与多家电力公司合作,通过数据分析向用户提供分析报告来提高消费用电的能效,预计全年节省电耗 2 亿美元。

(3)提高企业效益。

全球第二大零售商 Tesco 公司利用大数据分析技术,了解用户的购买习惯并进行分类,再根据用户的类别组织开展相应的业务推销活动。通过这样的方式,Tesco 公司在保持盈利的同时,每年节省了 3.5 亿英镑的宣传推广费用。它保存了大量的用户数据,利用这些数据为用户制订个性化推荐,从仓储、物流及广告等各方面获取了丰厚的回报。SAS 公司提供海量数据挖掘服务,帮助客户更快、更准地进行商业决策,完善产品及服务,改进企业绩效。

(4)改善人民生活。

大数据的价值还与日常生活息息相关,通过大数据,可以帮助人们提高生活质量。IBM 公司曾经帮助加州太平洋医疗中心研发了一套解决方案,通过对大数据的处理分析来跟踪接受心脏移植的患者后续的治疗效果,为科学家与医生提供了丰富的病例来源,帮助医生发现新的治疗方式,为病人提供更好、更全面的治疗方案,并为病人缩短住院时间和治疗成本。美国加利福尼亚州的警务部门正在使用大数据预测可能的犯罪场所与时间,以此减少特定地区的整体作案率。2013 年 1 月,中国住建部公布首批国家智慧城市试点名单,希望通过大数据来改善交通、医疗及城市建设等。

1.2 数据质量问题

数据质量的定义有很多。维基百科上有两种定义:①如果数据适用于在操作、决策制定和计划中的角色,则其看作是高质量的;②如果数据正确描述其指示现实世界中的对象,则称其为高质量的。数据质量可以从多个角度进行描述。

研究人员利用问卷调查,获得了数据消费者方面的数据质量属性列表。

调查问卷的内容是,选择两组调查对象(IT 内业的数据消费者和美国的 MBA 学生)各 112 人,平均年龄为 30 岁,进行头脑风暴,要求他们列出当他们想

到数据质量首先浮现在脑海的维度，即对数据质量维度的第一反应，列出表格。表 1.1 给出了 179 个基本维度。

表 1.1 179 个基本维度

可加入能力	下载能力	识别错误能力	上传能力
可接受性	竞争访问	可访问性	精确性
适应性	充足的详细介绍	充足的空间	审美主义
年龄	聚集能力	可变性	数据量
可审核	权威性	可用性	可信性
数据的广度	简洁	已认证数据	清晰度
来源清晰度	清晰的数据职责	紧凑性	兼容性
竞争优势	完整性	综合性	可压缩性
简洁度	简练性	保密性	整合性
一致性	内容含量	上下文	连续性
便利	正确性	损坏	成本
精确性成本	收集成本	创造性	批判性
当前	可定制性	数据的层次结构	数据提高指标有效性
数据重载	可定义性	可靠性	数据深度
细节	详细的源	分散性	可区分的更新文件
动态	轻松访问	易于比较	易关联性
易于数据交换	易维护性	易检索性	易理解性
易于更新	易用性	易改变性	易于怀疑
效率	耐力	启发性	符合人体工学
无误性	可扩展性	代价支出	易扩展性
扩展性	规模范围	最终确定	天衣无缝
灵活性	演示的形式	格式	完整性
友好性	一般性	习惯性	历史兼容性
重要性	不一致性	一体化	完整性
交互式	感兴趣	抽象程度	标准化程度

续表 1.1

局部化	逻辑连接	可管理性	可操纵性
可测量性	中等	符合要求	最小性
模块化设计	狭窄定义	没有信息丢失	正态性
新颖性	客观性	最优性	有序性
起源	简约性	可分割性	以往经验
血统	个性化	相关性	可移植性
严谨性	精确率	专有性	目的性
数量	合理性	冗余度	格式化的规律性
相关度	可靠性	重复性	再生性
声誉	图形分辨率	职责	可检索性
揭示	可审查性	硬度	稳健性
信息适用范围	保密性	安全性	自我修正
语义解读	语义	规模	来源
特殊性	高速	稳定性	存储
同步性	时间独立性	时效性	可溯源
可翻译	可移植性	明确性	公正无偏
可理解性	唯一性	无组织	最新
可用性	有用性	用户界面友好	有效性
价值	可变性	种类多变	可验证
波动性	证据充分	完善提交	

进一步，根据问卷调查的结果，收集这些维度中对于数据消费者来说重要的维度，然后分析维度的重要性，建立中间数据质量维度模型。

调查内容是，选定了 1 500 个住在美国的 MBA 校友，这些校友分布在不同行业、不同部门和不同管理层，这些人经常使用数据做决策，从而满足了“具有不同观点的数据消费者”这一要求。根据因素分析，找出了 20 个对数据消费者来说重要的维度。表 1.2 给出了这 20 个维度。