“十三五”国家重点图书出版规划项目

大数据技术与应用

丛书主编
朱扬勇 吴俊伟

Big Data

Technology and Application Series

蔡 莉 朱扬勇
编著

大数据质量

上海科学技术出版社

大数据技术与应用

大数据质量

蔡　莉　朱扬勇
编著

上海科学技术出版社

图书在版编目(CIP)数据

大数据质量 / 蔡莉，朱扬勇编著. —上海：上海科学技术出版社，2017.1
(大数据技术与应用)
ISBN 978-7-5478-3374-2

Ⅰ.①大… Ⅱ.①蔡… ②朱… Ⅲ.①数据处理—研究 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字(2016)第 277757 号

大数据质量
蔡　莉　朱扬勇　编著

上海世纪出版股份有限公司
上海科学技术出版社　出版
(上海钦州南路 71 号　邮政编码 200235)
上海世纪出版股份有限公司发行中心发行
200001　上海福建中路 193 号　www.ewen.co
苏州望电印刷有限公司印刷
开本 787×1092　1/16　印张 15.25
字数 320 千字
2017 年 1 月第 1 版　2017 年 1 月第 1 次印刷
ISBN 978-7-5478-3374-2/TP·47
定价：60.00 元

内容提要

数据作为一种基础性与战略性资源得到了广泛认可，数据服务成为很多组织和机构日常营运和活动中必不可少的重要环节。当下，数据质量在理论与实践中越来越受到关注，不仅是制约数据产业发展的关键问题，也是大数据应用研究中绕不开的重大命题。本书汇集了国内外数据质量研究的经典理论、技术和方法，以及最新的前沿发展趋势；首先介绍了传统数据质量研究的各种代表性成果；接着，在此基础上，结合大数据的特性，分析大数据时代下数据质量面临的挑战，并详细介绍基于大数据的数据质量相关技术的实现；最后，通过一个实际案例，提出一套完整的大数据质量解决方案。

本书可作为高等院校相关专业高年级学生和研究生的数据质量课程教材，以及从事数据质量研究和应用的科技工作者的参考书。

本书的相关研究工作得到以下相关项目的资助：
1. 国家自然科学基金项目(编号：61663047)
2. 国家自然科学基金项目(编号：U1636207)
3. 上海市科技发展基金项目(编号：16JC1400801)

大数据技术与应用

学术顾问

大数据技术与应用

编撰委员会

丛书序

我国各级政府非常重视大数据的科研和产业发展，2014 年国务院政府工作报告中明确指出要“以创新支撑和引领经济结构优化升级”，并提出“设立新兴产业创业创新平台，在新一代移动通信、集成电路、大数据、先进制造、新能源、新材料等方面赶超先进，引领未来产业发展”。2015 年 8 月 31 日，国务院印发了《促进大数据发展行动纲要》，明确提出将全面推进我国大数据发展和应用，加快建设数据强国。前不久，党的十八届五中全会公报提出要实施“国家大数据战略”，这是大数据第一次写入党的全会决议，标志着大数据战略正式上升为国家战略。

上海的大数据研究与发展在国内起步较早。上海市科学技术委员会于 2012 年开始布局，并组织力量开展大数据三年行动计划的调研和编制工作，于 2013 年 7 月 12 日率先发布了《上海推进大数据研究与发展三年行动计划（2013—2015 年）》，又称“汇计划”，寓意“汇数据、汇技术、汇人才”和“数据‘汇’聚、百川入‘海’”的文化内涵。

“汇计划”围绕“发展数据产业，服务智慧城市”的指导思想，对上海大数据研究与发展做了顶层设计，包括大数据理论研究、关键技术突破、重要产品开发、公共服务平台建设、行业应用、产业模式和模式创新等大数据研究与发展的各个方面。近两年来，“汇计划”针对城市交通、医疗健康、食品安全、公共安全等大型城市中的重大民生问题，逐步建立了大数据公共服务平台，惠及民生。一批新型大数据算法，特别是实时数据库、内存计算平台在国内独树一帜，有企业因此获得了数百万美元的投资。

为确保行动计划的实施，着力营造大数据创新生态，“上海大数据产业技术创新战略联盟”（以下简称“联盟”）于 2013 年 7 月成立。截至 2015 年 8 月底，联盟共有 108 家成员单位，既有从事各类数据应用与服务的企业，也有行业协会和专业学会、高校和研究院所、大数据技术和产品装备研发企业，更有大数据领域投资机构、产业园区、非 IT

领域的数据资源拥有单位，显现出强大的吸引力，勾勒出上海数据产业的良好生态。同时，依托复旦大学筹建成立了“上海市数据科学重点实验室”，开展数据科学和大数据理论基础研究、建设数据科学学科和开展人才培养、解决大数据发展中的基础科学问题和技术问题、开展大数据发展战略咨询等工作。

在“汇计划”引领下，由联盟、上海市数据科学重点实验室、上海产业技术研究院和上海科学技术出版社于2014年初共同策划了“大数据技术与应用”丛书。本丛书第一批已于2015年初上市，包括了《汇计划在行动》《大数据测评》《数据密集型计算和模型》《城市发展的数据逻辑》《智慧城市大数据》《金融大数据》《城市交通大数据》《医疗大数据》共八册，在业界取得了广泛的好评。今年进一步联合北京中关村大数据产业联盟共同策划本丛书第二批，包括《大数据挖掘》《制造业大数据》《航运大数据》《海洋大数据》《能源大数据》《大数据治理与服务》《大数据质量》等。从大数据的共性技术概念、主要前沿技术研究和当前的成功应用领域等方面向读者做了阐述，作者希望把上海在大数据领域技术研究的成果和应用成功案例分享给大家，希望读者能从中获得有益启示并共同探讨。第三批的书目也已在策划、编写中，作者将与大家分享更多的技术与应用。

大数据对科学研究、经济建设、社会发展和文化生活等各个领域正在产生革命性的影响。上海希望通过“汇计划”的实施，同时也是本丛书希望带给大家一个理念：大数据所带来的变革，让公众能享受到更个性化的医疗服务、更便利的出行、更放心的食品，以及在互联网、金融等领域创造新型商业模式，让老百姓享受到科技带来的美好生活，促进经济结构调整和产业转型。

上海市科学技术委员会副主任

2015年11月

前 言

质量是关于符合性的一种度量，即符合国际/国家标准或者符合使用者需求的程度。ISO 9000 系列质量体系是一个公认的国际标准，被全球 110 多个国家采用，既包括发达国家，也包括发展中国家。这一标准的执行使得市场竞争更加激烈，产品和服务质量得到日益提高。

国际标准化组织制订的国际标准——《质量管理体系基础和术语》(ISO 9000：2008)中指出：产品质量是指产品的一组固有特性满足要求的程度。与通常的有形产品不同，数据常常被认为是无形的，数据质量的评价要困难很多。1980 年以来，学术界、工业界和国际组织针对数据质量的测量、评估和管理提出了许多理论、技术和方法，却缺乏一个广泛认可的标准。ISO 正在开发的数据质量国际标准(ISO 8000)，目前也只有 20 多个国家接受它。

除了数据是无形的之外，建立数据质量标准的又一难点在于数据具备资源性、产品性和服务性。数据的资源性是指数据类似于矿藏和原矿，强调的是可开采性和可利用性；数据的产品性是指数据经过加工后可以形成数据产品，进入市场流通；数据的服务性是指数据能够以提供服务的方式进入市场，使用者不需要购买和拥有数据，只是使用了数据服务。因此，从这三个大类的性质来看，数据质量的评价体系就存在很大差异，而且每个类别都会面临不同的需求符合性。

数据作为一种基础性资源和一种战略性资源，已经获得广泛认可，数据服务业已广泛开展，各地数据交易所纷纷成立；这时，数据质量就逐渐成为制约数据产业发展的关键问题。此外，由于大数据自身特性，直接采用传统的、面向结构化数据的质量理论和方法来处理质量问题并不合适，数据质量的研究在新环境下面临着更大的挑战。

数据作为一种特殊资源，其质量应当符合真实性、合法性和可用性的基本要求。本书主要从数据的资源性来阐述数据质量，在传统数据质量研究的基础上，结合大数据的

特性，阐述基于大数据的数据质量相关技术的实现，并通过一个实际案例，提出一套完整的大数据质量解决方案。

本书共7章。第1章叙述数据质量的概况，列举出数据质量的影响和产生因素、数据质量的定义及面临的挑战，以及数据质量与信息质量的关系。第2章介绍了与数据质量有关的各种国际标准和行业标准。第3章讨论了数据分类和数据模型，并针对半结构化和非结构化数据，给出了一些数据模型和质量模型。第4章详细阐述数据质量的相关技术，包括：数据集成、数据剖析、数据清洁和数据溯源，并给出它们在大数据环境下的实现技术和方案。第5章详细论述了数据质量评估维度的选取，质量维度的测量和评估方法，同时每一种常用的评估方法都给出具体的评估案例。第6章描述数据质量的管理方法和质量管理成熟度模型。第7章以位置大数据为例，详细分析了位置大数据的来源、质量问题，评估模型和质量控制，给出确实可行的数据质量解决方法。

本书可作为高等院校相关专业高年级学生和研究生的数据质量课程教材，以及从事数据质量研究和应用的科技工作者的技术参考。

特别感谢国内外数据质量专著、教材和许多高水平论文报告的作者们，他们是黄伟、刁兴春、曹建军、黎建辉、樊文飞、Richard Y. Wang、Yang W. Lee、Elizabeth M. Pierce、Danette McGilvray、John Talburt、Carlo Batini、Monica Scannapieca等教授。在本书中引用了他们的部分成果，使本书较全面地反映数据质量各个研究领域的最新进展。感谢李英姿、李永轩和周怡帆三位硕士研究生提供的支持。

本书由朱扬勇教授和蔡莉副教授共同策划并拟定框架内容，并由蔡莉副教授执笔，朱扬勇教授审阅修订。限于作者学术水平，错误之处难免，恳请读者不吝指教。任何意见和建议，请发至电子邮件：caili@ynu.edu.cn。对此，我们将深为感激。

蔡　莉　朱扬勇

2016年12月6日

目 录

第1章

理解数据质量