数据质量的历史沿革和发展趋势

蔡 莉[1,3] 梁 宇[3] 朱扬勇[1,2] 何 婧[3]
(复旦大学计算机科学技术学院 上海 200433)[1] (上海市数据科学重点实验室 上海 201203)[2]
(云南大学软件学院 昆明 650091)[3]

摘 要 在互联网时代,数据成为了新的生产要素,也成为了基础性资源和战略性资源,同时还是重要的生产力。大数据服务业在全国广泛开展,数据交易所纷纷成立。这时,数据质量就逐渐变成制约数据产业发展的关键问题。首先,按照时间顺序将数据质量的研究内容划分为3个阶段,全面梳理和总结每个阶段的代表性成果,包括理论、方法、技术、工具和框架;然后,分析了在物联网、云计算和大数据环境下,数据质量研究所面临的各种挑战和机遇;最后,从数据质量模型、大数据质量管理、大数据质量相关技术、众包、物联网以及数据开放6个方面对数据质量的研究热点和发展方向进行了展望。

关键词 数据质量,历史沿革,发展趋势,大数据
中图法分类号 TP3-05 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.04.001

History and Development Tendency of Data Quality

CAI Li[1,3] LIANG Yu[3] ZHU Yang-yong[1,2] HE Jing[3]
(School of Computer Science, Fudan University, Shanghai 200433, China)[1]
(Shanghai Key Laboratory of Data Science, Shanghai 201203, China)[2] (School of Software, Yunnan University, Kunming 650091, China)[3]

Abstract In the Internet age, data becomes new factors of production, becomes the basic resources and strategic resources, and are important productive forces. Big data services have been widely carried out in China, and data exchanges have been established. Now, data quality has become a key issue restricting the development of data industry. This paper divided the research issues of data quality into three stages according to the chronological order, and summarized the representative results of each stage, including methodologies, techniques, models, tools and frameworks. Then, it analyzed the challenges and opportunities faced by data quality research in the new environment of big data, the internet of things and cloud computing. Finally, it prospected research focuses and development trend of data quality from six aspects: data quality model, quality management of big data, related quality techniques for big data, crowdsourcing, internet of things and data sharing.

Keywords Data quality, History, Development trend, Big data

1 引言

随着大数据时代的来临,“数据即资产”的概念得到了人们的广泛认同,对数据的重视程度被提到前所未有的高度;然而,不是所有的数据都能成为资产,数据的价值与数据质量密切相关。当下,数据质量在理论与实践中越来越受到关注,不仅成为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。数据质量的研究开始于20世纪70年代前后,在经历了30多年的发展历程后,形成了一系列经典的理论、技术和方法。从时间上来看,数据质量的发展历程大致经历了3个阶段:数据质量的萌芽阶段、形成阶段和繁荣阶段,每一个阶段都涌现出了许多重要的研究成果,如数据质量的定义、MIT的全面数据质量管理、数据质量的评估框架、ISO 8000数据质量标准等内容。2010年以后,物联网、云计算、大数据等一些新技术飞速发展,给数据质量的研究带来了巨大的挑战和机遇。本文综述了传统数据质量研究领域的各种代表性成果,并分析了在新技术和新环境下数据质量研究的热点问题和发展前景。

2 数据质量的历史沿革

数据质量是随着信息系统的发展而出现的,数据质量的好坏会影响信息系统的正常运行。因此,人们意识到数据质量对于信息系统的重要性,并由此揭开了数据质量研究的序幕。30多年来,数据质量逐渐成为一个专业的研究领域并

到稿日期:2017-03-10 返修日期:2017-06-02 本文受国家自然科学基金:基于位置大数据的城市热点区域和居民出行模式的挖掘研究(61663047)资助。

蔡 莉(1975-),女,博士生,副教授,主要研究方向为数据质量、数据挖掘,E-mail:lcai@fudan.edu.cn(通信作者);梁 宇(1964-),男,硕士,教授,主要研究方向为智能交通、云计算;朱扬勇(1963-),男,博士,教授,主要研究方向为数据科学与技术、数据挖掘;何 婧(1978-),女,博士,讲师,主要研究方向为数据索引、数据存储。

涌现出许多重要的研究成果，其发展历程大致经历了以下 3 个阶段。

2.1 数据质量的萌芽阶段

20 世纪 70－90 年代是数据质量的萌芽阶段，这一阶段的特征是：研究人员已经发现劣质的数据质量会影响信息系统的正常运行，但是尚未建立一个关于数据质量的知识体系。

在数据质量萌芽阶段，数据质量问题的研究更多来源于行业应用，如会计领域、管理领域、统计领域和计算机领域[1]。会计必须承担会计信息失真造成的严重后果，因此会计领域对数据质量历来比较重视，从 20 世纪 70 年代起就建立了严格的会计信息质量框架体系[2]。不过，会计领域并没有试图推广自己的数据质量管理经验。统计领域对数据质量也非常重视，20 世纪 60 年代末期建立了国民核算体系中的统计调查制度、抽样调查过程中的质量控制、数据分析中的离群值检验和数据插补等一系列技术手段来保证质量。管理学领域对数据质量的研究始于 20 世纪 80 年代初期，其研究重点在于如何控制数据制造系统，以检测和消除数据质量问题。在计算机领域，Morey，Balloru，Laudon，Mathieu 和 Khalil 等研究人员已经意识到评估和提高信息系统中的数据质量的必要性[3-4]。

上述领域都结合自己的行业需求，提出了保证数据质量的方法和框架，但是相关研究成果和行业经验都仅限于各行业内部，没有推广到其他领域；而且，学科层面并不重视数据质量问题，没有出现一个关于数据质量的统一知识体系[5]。

2.2 数据质量的形成阶段

1990－1999 年是数据质量的形成阶段，这一阶段的特征是：创立了基于相关学科的数据和信息质量理论，并使之成为数据质量管理的统一知识体系。由此，数据质量正式成为一个独立的研究方向。

2.2.1 数据质量的定义及维度的选择

20 世纪 90 年代，美国麻省理工学院（MIT）的数据质量研究小组在 Stuart Madnick 和 Richard Y Wang 教授的带领下提出了全面数据质量管理（Total Data Quality Management，TDQM）的理论，这标志着学科层面首次就数据质量提出了较为完整的知识体系，为今后数据质量的全面发展奠定了良好的理论基础。

这一阶段的研究重点在于定义数据质量、确定质量维度和建立 TDQM 理论。学术界尚未对数据质量形成一个统一的定义，比较认可的是 MIT 提出的定义。他们采纳“fitness for use”（使用的适合性）的概念，将数据质量定义为“数据适合数据消费者的使用”。数据质量的判断依赖于使用数据的个体，不同环境下不同人员的“使用的适合性”不同[6]。数据分析专家 Redman 认为：如果数据在运营、决策和规划中能够满足客户的既定用途，数据便是高质量的。根据这一定义，客户是质量的最终裁决者[7]。

数据质量维度是指一个特征或部分信息，被用于分类信息和确定数据需求。事实上，它提供了一种用于测度和管理数据质量及信息的方式。MIT 采取二阶段调查方法把质量维度划分为固有质量、可访问性质量、语境质量和表达质量 4 种类型，每种类型包含若干个质量维度[6]。文献[8]识别出 6 个 Web 数据特征及 32 个子特征。文献[9]在 MIT 研究的基础上，采用调查方法得到 11 个 Web 数据质量维度。文献[10]以符号学为基础，建立句法层次、语义层次、语用层次、社会层次 4 个符号学层次，共 11 个质量维度。除了学术界定义的维度外，国内外机构和一些行业也制定了相应的维度，如表 1、表 2 所列[11]。

表 1 国际机构和政府部门的常用数据质量维度

Table 1 Common data quality dimensions of international institutions and governments

国际机构或者政府部门	数据质量维度
国际货币基金组织	诚信的保证、方法的健全性、准确性和可靠性、适用性以及可获取性
欧盟统计局	相关性、准确性、可比性、连贯性、及时性和准时、可访问性和清晰
联合国粮食及农业组织	相关性、准确性、及时性、准时性、可访问性和明确性、可比性、一致性和完整性、源数据的完备性
美联邦政府（公众传播）	实用性、客观性（准确、可靠、清晰、完整、无歧义）、安全性
美国商务部	可比性、准确性、适用性
美国国防部	准确性、完整性、一致性、适时性、唯一性及有效性
加拿大统计局	准确性、及时性、适用性、可访问性、衔接性、可解释性
澳大利亚国际收支统计局	准确性、及时性、适用性、可访问性、方法科学性

表 2 国内部分领域或行业的数据质量维度

Table 2 Data quality dimensions used in some domestic fields or industries

行业	数据质量维度
烟草	准确性、完整性、一致性、及时性、可解释性、可访问性
气象通信	科学性、标准化、共享性、时效性、稳定性、可维护性
军事	完全性、一致性、准确性、唯一性、时效性、可解释性
医疗	一致性、可靠性、可用性、适用性
交通	完整性、有效性、准确性、实时性
地理信息系统（GIS）	位置精度、现势性、一致性、完整性、可靠性

从表 1 和表 2 可以看出，各个机构和行业对质量维度的要求不尽相同，其中，出现频率较高的维度是：准确性、完整性、一致性、可获得性和及时性。

2.2.2 质量维度度量

在不同的场景下，数据质量维度的度量方式存在差异。下面以准确性、完整性和一致性为例，介绍它们各自的度量模型。模型的相关定义如下[12]。

定义 1 设 $E_1,E_2,\cdots,E_m$ 为需要度量的 m 条记录，它们组成一个数据集合 $D=\{E_1,E_2,\cdots,E_m\}$，E_i 为集合中的任意一条记录，$m\in N^+$。

定义 2 设 $A_1,A_2,\cdots,A_n$ 为 E_i 的 n 个属性，A_{ij} 表示 E_i 在属性 j 上的取值，则 $E_i=\{A_{i1},A_{i2},\cdots,A_{in}\}$，$n\in N^+$。$A_{ij}$ 可能存在缺失、拼写错误、不一致等质量问题。

定义 3 $R=\{R_{11},R_{12},\cdots,R_{mn}\}$ 表示权威性的参考数据源。R_{ij} 表示记录 E_i 在属性 j 上的正确值或者期望值。

(1)准确性(Accuracy)

这里将准确性定义为:准确性=真实值的数量/所有值的数量 。设 $f(\cdot)$ 为评估对象 A_{ij} 的取值结果到(0,1)的映射,若结果正确,则取值为 1,反之为 0,有:

$$f(A_{ij})=\begin{cases}1, & A_{ij}=R_{ij}\\ 0, & \text{其他}\end{cases} \tag{1}$$

那么,D 在属性 j 上的准确性为:

$$Accuracy=\sum_{i=1}^{m} f(A_{ij})/m \tag{2}$$

D 在全部属性上的准确性为:

$$Accuracy_D=\sum_{j=1}^{n}\sum_{i=1}^{m} f(A_{ij})/(m\times n) \tag{3}$$

(2)完整性(Completeness)

这里将完整性定义为:完整性=非空值的数量/所有值的数量。设 $g(\cdot)$ 为评估对象 A_{ij} 的赋值情况到(0,1)的映射,若 A_{ij} 非空,则取值为 1,反之为 0,有:

$$g(A_{ij})=\begin{cases}1, & A_{ij}\text{有值}\\ 0, & A_{ij}\text{为空}\end{cases} \tag{4}$$

那么,D 在全部属性上的完整性为:

$$Completeness_D=\sum_{j=1}^{n}\sum_{i=1}^{m} g(A_{ij})/(m\times n) \tag{5}$$

(3)一致性(Consistency)

这里将一致性定义为:一致性 = 一致性值的数量/所有值的数量。集合 $C_i=\{C_{i1},C_{i2},\cdots,C_{is}\}$ 表示 A_{ij} 可能的取值范围。设 $h(\cdot)$ 为评估对象 A_{ij} 的取值情况到(0,1)的映射,若 A_{ij} 的值为集合 C_i 中的任一值,则取值为 1,反之为 0,有:

$$h(A_{ij})=\begin{cases}1, & A_{ij}=C_{is}\\ 0, & A_{ij}\neq C_{is}\end{cases} \tag{6}$$

那么,D 在属性 j 上的一致性为:

$$Consistency=\sum_{i=1}^{m} h(A_{ij})/m \tag{7}$$

上面介绍了 3 种质量维度的度量方法,在实际应用中,还要根据具体需求细化模型或者对模型加以改进,以适应后续的数据质量评估。

2.2.3　全面数据质量管理

由于缺少一个完整的数据质量知识体系,Stuart Madnick 和 Richard Wang 在 MIT 启动了 TDQM 研究计划,其主要目标是借鉴全面质量管理(Total Quality Management,TQM)的思想创立基于相关学科的数据和信息质量理论,使之成为数据质量管理的统一知识体系,这些学科包括计算机科学、统计学、会计学、管理学和组织行为学等。与质量管理中的 PDCA(P 即 Plan,D 即 Do,C 即 Check,A 即 Action)循环类似,TDQM 也是一个循环过程。为了改善数据质量,需要执行定义、测量、分析和改进 4 个阶段,如图 1 所示[13]。

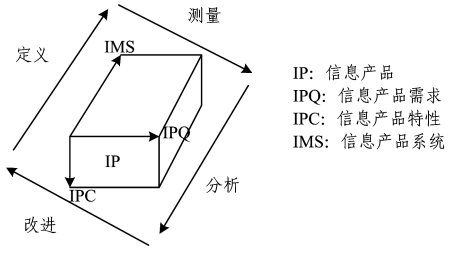

图 1　TDQM 循环

Fig. 1　Total data quality management cycle

定义阶段需要描述数据质量的概念并识别数据质量的维度。TDQM 从数据消费者的角度把数据质量定义为“fitness for use”,而且通过多阶段问卷调查的方法把数据质量维度划分为固有质量、可访问性质量、语境质量和表达质量。测量阶段需要根据数据产品及其质量的定义,确定质量指标体系,跟踪数据的测量并监控数据质量。TDQM 采用简单比率、最大最小运算和加权平均[14] 3 种客观评价数据质量的算法来度量数据质量。分析阶段使用差异分析技术来发现数据维度和数据角色在数据质量方面的差异。改善阶段会根据分析结果采取措施来消除产生数据质量问题的根源,如采用数据清洁、转换等技术改进数据重复、数据缺失、数据不一致等问题,或者制定政策改进数据的生产过程和管理方法。通常,改变产生数据的过程被认为是一种更有效的方法[15]。

在学术界开始建立数据质量知识体系的同时,政府部门也将目光转向这一领域。1995 年,为了规范联邦政府的数据收集管理工作,美国国会要求联邦政府的行政管理预算局(OSM)制定新的政策,给出具体措施,以确保所发布数据的可靠性,即数据要有质量[16]。这也标志着数据质量首次引起了政府层面的关注。

2.3　数据质量的繁荣阶段

2000－2009 年是数据质量研究和应用全面发展的阶段,其主要特征是:数据质量研究继续深入,相关数据质量产品已经大量出现,国际组织开始研究和制定数据质量标准,政府部门颁布数据质量法案。

2.3.1　数据库与数据仓库中的质量控制

在学术界,数据质量的研究更加细化,涌现出了许多重要理论和成果。得益于数据库和数据仓库的广泛应用,采用数据库或者数据仓库技术来解决数据质量问题成为当时的研究热点。在数据库领域,数据完整性(Data Integrity),即数据的准确性(Accuracy)、可靠性(Reliability)和一致性(Consistency)[17],是质量度量的基石,可用于防止数据库中存在不规范的数据,避免错误信息造成输入/输出的无效操作。数据库采用多种方法来保证数据完整性,包括主键、外键、约束条件、规则和触发器。除了数据完整性以外,用户进行数据库设计时,通常还需要考虑属性之间的关系,即依赖关系,这是一种语义范畴的概念,只能根据数据的语义来确定。目前,最常用的依赖关系是函数依赖和包含依赖,它能有效避免数据库中出现的各种数据异常情况[18]。2007 年,樊文飞教授等提出条件函数依赖和条件包含依赖[19-20],通过条件表的约束增强函数依赖和包含依赖对实际语义的表达能力。条件函数依赖一经提出,便受到数据库研究者的广泛关注,在数据清洗、提高数据质量上产生了极大的影响。

数据仓库是面向主题的、集成的、非易失的和随时间变化的数据集合,它被用于支持决策经营管理中的决策制定过程[21]。在数据仓库中,最重要的功能可以归纳为“抽取(Extract)－转换(Transfer)－加载(Load)”,即 ETL 操作[22]。数据仓库中数据质量的优劣是数据仓库能否成功的关键因素之一。尽管在 ETL 过程中已经对数据源进行了数据质量的提高操作,但是在将数据加载到数据仓库之前,还需要用各种方法来确认数据的质量。最好的方法是在数据源层就对质量问题加以处理,但是要确认外部数据源的质量存在困难。因此,

需要在 ETL 阶段通过手工方式或者软件自动化的方式完成对数据质量的确认。

2.3.2 数据质量评估框架及其方法

(1)数据质量评估框架

为了实现对数据质量和信息质量的测量和评估,MIT 首先提出了信息管理质量评估(Assessment Information Management Quality,AIMQ)框架;国际货币基金组织(International Monetary Fund,IMF)也提出了通用性数据质量评估框架(Data Quality Assessment Framework,DQAF),该框架可被广泛应用于各成员国的统计数据质量的评价和改善。之后,又出现了十多个评估框架(见表 3)[23],不断丰富着相关研究成果。

表 3 常见的数据质量评估框架

Table 3 Common data quality assessment frameworks

评估框架缩写	全称	主要创建者	创建时间
AIMQ	A methodology for information quality assessment	Lee 等人	2002
CIHI	Canadian Institute for Health Information methodology	Long 和 Seko	2005
DQA	Data Quality Assessment	Pipino 等人	2002
IQM	Information Quality Measurement	Eppler 和 Munzenmaier	2002
ISTAT	ISTAT methodology	Falorsi 等人	2003
DQAF	Data Quality Assessment Framework	IMF 组织	2003
AMEQ	Activity-based Measuring and Evaluating of product information Quality (AMEQ) methodology	Su 和 Jin	2004
COLDQ	Loshin Methodology (Cost-effect of Low Data Quality)	Loshin	2004
DaQuinCIS	Data Quality in Cooperative Information Systems	Scannapieco 等人	2004
QAFD	Methodology for the Quality Assessment of Financial Data	De Amicis 和 Batini	2004
CDQ	Comprehensive methodology for Data Quality management	Batini 和 Scannapieco	2006

表 3 中的一些框架主要针对通用领域的数据质量或者专业领域的数据质量进行评估,其他框架则适用于企业内部的信息系统或者协同信息系统的评估。

(2)数据质量评估方法

数据质量评估方法主要分为定性方法、定量方法和综合方法。定性方法主要依靠评判者的主观判断。定量方法则为人们提供了一个系统、客观的数量分析方法,结果较为直观、具体。综合方法则将定性方法和定量方法结合起来,以发挥两者的优势。

定性评估方法一般基于一定的评估准则与要求,根据评估的目的和用户对象的需求,从定性的角度对数据资源进行描述与评估。具体步骤是:确定相关评估准则或指标体系,建立评估准则及各赋值标准,通过对评估对象进行大致评定,给出各评估结果。评估结果有等级制、百分制等表示方法[24]。通常,定性评估可划分为:用户反馈法、专家评议法和第三方评测法。

定量评估方法是指按照数量分析方法,从客观量化角度对基础数据资源进行的优选与评估。定量方法为人们提供了一个系统、客观的数量分析方法,结果更加直观、具体。目前,传统的纸质印刷品,如报纸、图书、期刊、标准和专利等内容都已经实现数字化,并存放在各种数据库中供用户检索、浏览和下载。为了评价各数据库中文献的数据质量,可以制定用户注册人数、文献下载量、文献在线访问量以及引用率等评估指标来评价各个数据库收录文献质量的优劣。

综合方法将定性和定量两种方法有机地集合起来,从两个角度对数据资源质量进行评估。常用的综合评估方法有:层次分析法(Analytic Hierarchy Process,AHP)[25]、模糊综合评估法(Fuzzy Comprehensive Evaluation,FCE)[26]、云模型评估法(Cloud Model,CM)[27]和缺陷扣分法(Defection Subtraction Score,DSS)[28]。本文从使用的难易程度、使用模型、应用场景和适用范围 4 个方面来对比这 4 种评估方法,如表 4 所列。

表 4 4 种综合评估方法的对比

Table 4 Comparison among four comprehensive assessment methods

评估类别	难易程度	使用模型	应用场景	适用范围
AHP	较简单	层次结构模型	质量指标权重确定	无限制
FCE	复杂	隶属函数	模糊性的质量问题	无限制
CM	复杂	正态云模型	模糊性与随机性共存的质量问题	无限制
DSS	简单	无	产品质量	特定专业领域

AHP 方法需要将复杂问题分解成若干层次,建立阶梯层次结构,然后构成判断矩阵,完成层次单排序一致性检验,最后进行层次总排序和一致性检验,从而得出结论。由于使用较为简单,适用范围广,在数据质量领域,AHP 常用于指标权重的确定。FCE 方法以模糊数学为基础,被用于处理模糊性的质量问题,通过建立隶属函数对事物做出综合评价,可应用在产品质量评定、科技成果鉴定、港口环境评价等领域。在现实世界中,许多事物的概念是不确定的,对于一个模糊性和随机性共存的质量问题,更适合采用 CM 理论。CM 将概率论和模糊集合理论结合起来,通过特定构造的算法,形成定性概念与其定量表示之间的转换模型,并揭示随机性和模糊性的内在关联性。缺陷扣分法指通过计算单位产品(数据或信息)的得分值,由单位产品的得分值来评价产品质量的方法。然而在实际操作中,该方法也存在着许多局限性和不足,仅适用于一部分专业领域,如空间数据等结构化数据的质量评价,而在全面的综合评价方面不完全适用。

2.3.3 数据清洗

如发现经过评估后的数据存在质量问题,就需借助数据清洁技术来改善数据质量。数据清洁(Data Cleaning,Data Scrubbing),也称为数据净化或者数据清洗,是指检测数据集合中存在的不符合规范的数据,并进行数据修复,以提高数据质量的过程[29]。数据清洁一般是自动完成的,只有在少数情况下需要人工参与完成。数据清洁可分为“特定领域(domain-specific)数据清洁”和“领域无关(domain-independent)

数据清洁”两类[30]。“特定领域数据清洁”需要用到相关领域知识,并要求参与清洁过程的人员掌握相关领域知识;“领域无关数据清洁”面向普通数据库用户,适用于不同的业务领域,更方便与传统的 DBMS 相整合。

数据之所以会出现质量问题,是因为“脏数据”的存在。本文将脏数据分为单数据源模式层问题、单数据源实例层问题、多数据源模式层问题和多数据源实例层问题 4 种类型[31-32]。表 5 列出了“脏数据”类型与其出现的原因。

表 5　“脏数据”类型、实例及出现原因

Table 5　Categories, examples and reasons of “dirty data”

类别	脏数据层次	类型	出现原因
单源脏数据	模式层	缺少完整性约束	不在约束范围内
		唯一性冲突	两个不同记录的主键重复
		参照完整性冲突	超出设定的值范围,没有相应的对象数据
	实例层	拼写错误	数据输入错误、数据传输过程中发生的错误
		重复/冗余记录	现实中的同一个实体在数据集合中用多条不完全相同的记录来表示,由于它们在格式、拼写上的差异,导致数据库管理系统不能正确识别;或者模式未做规范化处理
		空值	字段空值设计不合理,或者用户不愿意填写
		数据失效	原有数据经过一段时间后变成无效数据
		噪声数据	由于采集设备异常,造成接收的数据取值不合理
多源脏数据	模式层	命名冲突	同一实体在不同的来源中存在不同的名称
		结构冲突	属性类型不一致、一个代码有不一致的含义、相同的意义不同的代码,格式不同
	实例层	时间不一致	不同时间层次上的数据在同一层次进行比较与计算
		粒度不一致	不同层次上的数据在同一层次进行比较与计算
		数据重复	相同的数据在合并后的数据库中出现两次及以上

如表 5 所列,“脏数据”的类型有很多种。从实例层来说,单数据源的“脏数据”就是不完整数据、不正确数据、不可理解数据、过时数据、数据重复等;单数据源的数据清洁需要在属性上对数据进行检测与处理。多数据源的“脏数据”更为复杂,主要指大量的重复数据、冲突数据;多数据源的数据清洁重点是对重复数据的检测与处理、解决数据冗余和数据冲突问题。

为了提高数据质量,按照表 5 所列举的“脏数据”类型,可将数据清洁方法划分为基于模式层和基于实例层的方法,具体如表 6 所列[32-35]。

表 6　常见的数据清洗方法

Table 6　Common data cleaning methods

脏数据层次	类型	清洁方法
模式层	属性约束	人工干预法和函数依赖法
	避免冲突	数据重构,元数据方法
实例层	拼写错误	拼写检查器来检错和纠错
	重复/冗余记录	基于字段和基于记录的重复检测后删除重复值
	空值	忽略元组,人工填写空缺值,使用一个全局变量填充空缺值,使用属性的中心度量(均值、中位数等),复杂的概率统计函数值填充空缺值
	数据不一致	指定简单的转换规则,使用领域特有的知识(如,邮政地址)对数据作清洁
	噪声数据	分箱(Binning)法、回归(regression)法、计算机和人工检查相结合处理、使用简单规则库检测和修正错误、使用不同属性间的约束检测和修正错误、使用外部数据源检测和修正错误

2.3.4　数据溯源

溯源一词源自法语“provenir”,意思是出处、发源。在计算机领域,溯源也称为世系(lineage)或者谱系(pedigree),用于描述数据的起源或者出处。许多科学应用和数据管理应用通常需要收集和处理大量不同来源的数据,由于来源复杂、质量参差不齐,数据和应用结果的可信度会受到质疑[36]。因此,能有效识别和查询数据来源的溯源技术成为数据质量研究领域的一项重要技术,它可以帮助用户理解数据并处理结果的可信度。

数据溯源可以划分成两种类型,即粗粒度的工作流溯源(Workflow Provenance)和细粒度的数据溯源(Data Provenance)[37]。当科学家使用工作流系统设计和运行科学实验时,工作流执行的结果数据集可能需要与报告或论文一起发布,以供其他科学实验的输入重复使用。此时,要求科学家在发布数据的同时发布其溯源元数据,包括数据的演变历史、起源和所有权,这一过程称为工作流溯源[38]。细粒度的数据溯源是指某个转换步骤结果中的片段数据是如何衍生的,它更加关注结果数据集的推导。早期的数据溯源通常细分为 Where 和 Why 型溯源[39];之后,在此基础上引入了 How-provenance[40];2010 年出现了 W7 模型[41],该模型是指数据溯源信息应该包括 Who, When, Where, How, Which, What, Why 7 个部分。

目前,数据溯源追踪的主要方法有标注法和反向查询法。此外,还有通用的数据追踪方法、双向指针追踪法、利用图论思想和专用查询语言的追踪法,以及以位向量存储定位等方法[42]。标注法通过记录处理相关的信息来追溯数据的历史状态,即用标注的方式来记录原始数据中的一些重要信息,并让标注和数据一起传播,通过查看目标数据的标注来获得数据的溯源。采用标注法来进行数据溯源虽然简单,但存储标注信息需要额外的存储空间。反向查询法主要用于数据库追溯,即在一定的限制条件下,可以通过分析数据库操作语句得出任意粒度的逆查询语句,追溯数据起源。与标注法相比,反向查询法更加复杂,但需要的存储空间更小。

2.3.5 ISO 8000 数据质量标准

为了满足数据产品的质量需求，2005 年，ISO 下设的委员会开始组织撰写 ISO 8000 标准，以突破数据质量没有国际标准的困境。ISO 8000 数据质量标准主要由 4 个部分组成[43]：1)通用数据质量，Part1－Part99。其中 Part1 介绍 ISO 8000 标准创建的目标、适用范围和进展，Part 2 介绍相关的术语。2)主数据质量，Part100－Part199。主数据下面又分为几个部分：Part100，Part110，Part120，Part130，Part140 和 Part150。所有部分都已经在 2012 年之前制定完成。3)事务数据质量，Part200－Part299。4)产品数据质量，Part300－Part399。

自从 2008 年底发布了第一部标准(ISO 8000-110、2008 主数据、语法、语义编码和数据规范的一致性)以来，至今已陆续发布了 6 部 ISO。已发布的 ISO 8000 标准都集中在 ISO 8000 100 系列部分，即主数据质量部分。目前，ISO 8000 的整体框架如图 2 所示[44]。

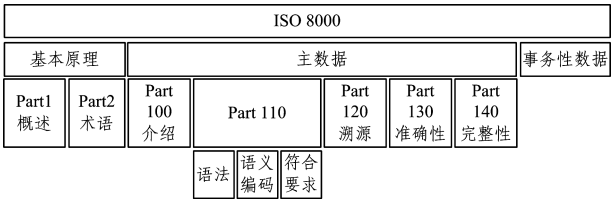

图 2 ISO 8000 体系架构图

Fig. 2 Architecture of ISO 8000 standard

2.3.6 数据质量的相关立法

2001 年美国“信息质量法”(Information Quality Act，IQA)获得国会批准并正式生效[45]。“信息质量法”并不是一个独立成章的法律文件，它只是美国财政部和政府机构拨款法案中的一个条款规定；而且其重点关注的也仅是在美国联邦政府机构中的信息传播活动。欧盟认为公共领域的信息是数字内容产品和服务的原材料，对公共领域信息的再利用和挖掘，有助于提高欧盟区域内的就业率并促进该地区经济的发展。为了协调欧盟成员国公共信息再利用的市场秩序，促进公民使用新的方法获取和利用知识，欧盟于 2003 年制定了《公共信息再利用指令》(简称指令)。再利用公共部门信息的企业应该按照合法的途径和方法加工制作信息产品，保证数据的准确、完整、权威、及时。相关公共部门应该履行一定的监督责任，防止企业滥用公共部门信息，建立一系列对信息质量的评审和评价体系。

3 数据质量的发展趋势

2010 年至今，物联网、云计算和大数据等技术已成为 IT 领域的发展热点，这些技术的出现给数据质量领域带来新的研究课题。随着数据的不断积累和政府及各行业数据的相继开放，组织和机构获取数据的来源和类型越来越丰富。大数据时代，有 80％的新增数据属于半结构化和非结构化数据，这给以结构化数据为研究对象的传统数据质量理论带来了不小的挑战。同时，如何在合理时间内完成对海量数据的质量评估和清洗也成为一个亟待解决的问题。此外，物联网、众包模式和公共数据开放中的数据质量问题也成为人们关注的焦点。

在新环境和新技术下，数据质量的研究方向在于：1)从大数据的特征出发，研究针对数据模型和数据质量管理的新方法和新技术；2)研究基于云计算平台(Hadoop，Storm 和 Spark)的数据质量相关技术的实现，如数据清洁和数据溯源；3)研究众包模式下数据的评估技术和质量保证机制；4)研究物联网数据验证方法、数据清洗算法和数据质量保证技术；5)研究数据质量与数据开放的关联性。下面简要介绍各研究热点所涉及的相关理论和技术。

3.1 大数据质量

由于大数据本身呈现出一些新的特性，如 4V 特性，因此如何从海量的、快速变化的、来源丰富的大数据中提取出高质量且真实的数据成为企业处理大数据过程中亟待解决的问题。

3.1.1 数据质量模型

传统数据质量的理论、技术和方法都是以结构化数据为研究对象，半结构化和非结构化数据质量的研究一直是一个难点，尤其是在当前的大数据时代，非结构化数据占据了新增数据总量的 80％以上。通常，有两种方式对非结构化数据进行处理：一种采用“非结构化数据－半结构化数据－结构化数据”这种逐步转换的方式，实现了非结构化数据向结构化数据转换的功能，最终将数据存入关系数据库中进行管理；另一种则采用“非结构化数据－结构化数据”的转换方式，借助元数据完成相应的处理。

为了评估半结构化数据(如 XML 数据格式)的质量，Scannapieco 和 Virgillito 等人[46]提出了数据和数据质量模型(Data and Data Quality，D^2Q)。D^2Q 模型可用于检验数据的准确性、一致性、完整性和实时性。D^2Q 模型是半结构化的，允许每个组织采用一定的灵活性导出数据的质量。此外，质量维度可以被关联到数据模型的不同元素上，范围从单个数据值到整个数据源。

对于非结构化数据，学术界并未提出有效的数据质量模型，现有研究更多集中在采用什么模型来表示这些数据。例如：Siadat 和 Chu 等人基于 E-R 图提出一种用于非结构化数据的模型[47-48]。针对多媒体检索系统，Marcus 和 Amato 等人[49-50]提出了一个多层数据模型，用来表示语义描述、底层特征和原始数据。四面体数据模型(Tetrahedral Data Model)是由 Li 和 Lang 于 2010 年提出的一个面向非结构化数据的模型[51]，它可以为不同类型的非结构化数据提供统一、集成和相关描述，并支持检索和数据挖掘等智能数据服务。

3.1.2 大数据质量管理

在大数据环境下，质量管理会面临如下挑战和问题。

问题 1：分散的信息。对于管理工作而言，信息孤岛一直是一个棘手的问题，且已存在很长时间。以前，企业常用的数据仅仅涵盖自己业务系统所生成的数据，如销售系统、CRM 系统等。但是，现在企业所能采集和分析的数据已经远远超越这一范畴。由于缺乏可以直接应用到这些数据上的统一的规则和方法，企业要保证从多个数据源获取结构复杂的大数据并有效地对其进行整合，异常艰巨[52]。

问题 2：复杂的异构数据源。不同来源的数据在结构上差别很大。一些数据是非结构化数据，如文本、视频、图像等，一些数据是半结构化数据，如电子邮件、电子表格等，还有一些是结构化数据，如数据仓库/商业智能数据、传感器/机器数

据记录、关系型数据库管理系统的数据等。

问题 3:海量数据。1970 年以后,信息量大约每 3 年就增加一倍;如今,全球信息总量每两年就可以增加一倍。2011 年,全球被创建和被复制的数据总量为 1.8ZB,要对如此大体量的数据进行采集、清洁、整合,最后得到符合要求的高质量数据,在一定时间内是很难实现的[53]。

问题 4:信息的可信度。由于企业采集到许多外部数据源提供的数据,用户无法理清这些数据之间的关系,更不了解这些信息是如何被交付的,因此更无法得知这些数据的质量处于什么水平。

为了有效应对上述挑战,大数据质量管理需要一个能提供多源异构数据整合、支持分布式和非结构化数据处理以及完成数据质量相关操作的大数据质量管理解决方案。对于分散的信息,可以使用 Linkedin 的 Kafka、Facebook 的 Scribe、Cloudera 的 Flume 和 Hadoop 的 Chukwa 等技术实现分布式的数据采集;对于复杂的异构数据源,可采用基于本体的方法完成异构数据的集成与融合;对于海量数据,可以采用 Hadoop 和 Storm 等分布式系统架构进行质量评估和数据清洗等操作;对于信息可信度的问题,则通过建立统一的数据质量标准和数据溯源来加以解决。

中国电信股份有限公司上海研究院从数据质量维度评估指标角度切入,论述运营商该如何构建大数据的数据质量管理体系,并提出了一个理想的实例,如图 3 所示[54]。

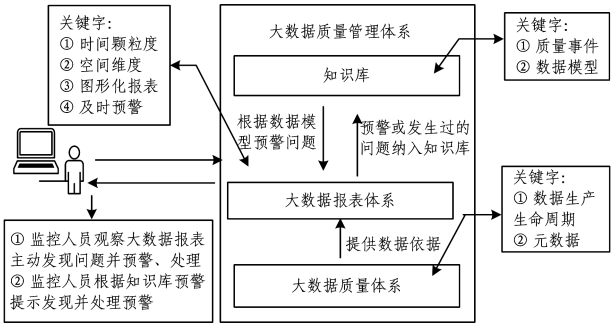

图 3　运营商的大数据质量管理体系

Fig. 3　Quality management architecture of big data for operators

该质量管理体系的核心在于大数据质量体系、大数据质量报表体系和大数据质量知识库体系的构建。大数据质量体系用来“发现质量问题”,大数据质量报表分析体系需要从已经发现的问题中寻找解决方法,而大数据质量知识库则针对数据质量领域的问题求解提供帮助。

3.2　基于云平台的数据质量相关技术的实现

随着云计算技术的广泛应用,早期基于单机版的数据质量技术已经无法满足海量和实时数据处理的需求,因此,如何根据各种云计算平台的特点开发相应的系统就成为了数据质量研究的热点内容之一。下面将介绍数据清洁和数据溯源技术在云平台上的解决方案。

基于语义的大数据清洁系统(简称为 SPF)是运行在 Map/Reduce 批处理工作流和 Storm 流数据框架上的数据清洁系统,能同时支持实时和非实时的大数据清洁,主要用来清洁日志文件、Internet 数据和视频流。其系统架构如图 4 所示[55]。

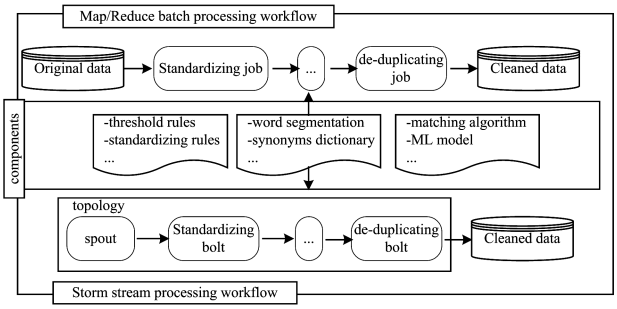

图 4　基于语义的大数据清洁框架

Fig. 4　Semantic-based intelligent data cleaning framework for big data

在图 4 中,各种来源的非实时大数据存储在 HDFS 系统中,这样,Map/Reduce 数据清洁工作流可以访问 HDFS 以获得原始的海量数据。对于实时应用,SPF 系统构建一个 Topology连接到这些应用上,并使用 Spout 获取持续的原始流数据。发现和清洁异常数据的规则采用 Drools 引擎进行管理,相关规则有阈值规则、标准化规则等。该系统能够检测到的数据异常情况包括错误数据、不一致性数据和重复数据。SPF 的并行清洁是由一系列的 Map/Reduce 或者 Bolt 功能来执行的。在整个数据清洁过程中,数据变化的日志为所有的行动和引起这些行动的原因提供了一个审计追踪。最后,清洁后的海量数据存储在 HDFS 或者 Storm 框架中作为输出结果。

大数据溯源是一种服务于大数据的科学计算和工作流的溯源类型。在大数据时代,一些有效的溯源系统被开发出来,如 Kepler 分布式溯源系统[56]、RAMP 溯源系统[57]、Hadoop-prov 溯源系统[58]和 Pig Lipstick 系统[59]。Kepler 分布式溯源框架不仅能在 MapReduce 任务中捕获溯源信息,而且还能在非 MapReduce 任务中溯源信息。在 MySQL 集群中,这一框架以一种分布式的方法——Kepler DDP 架构来记录和查询溯源。同时,它也提供了一个 API 来查询所收集的溯源信息。收集和查询溯源的可扩展性可以使用 WordCount 程序和生物信息学中的 BLAST 应用来评估。其分布式溯源框架包括 3 部分:Kepler GUI、Hadoop Mater 节点和 Hadoop Slaves 节点,如图 5 所示[56]。

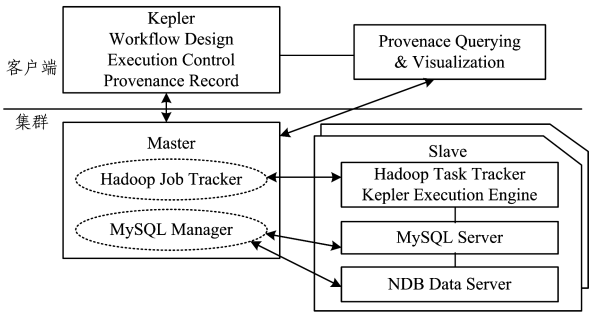

图 5　Kepler 分布式溯源框架

Fig. 5　Kepler MapReduce provenance architecture

3.3　众包模式下的数据质量

众包模式是一种新的生产组织形式,它使得数据的产生和获取不再局限于组织或者平台,个人用户借助一些硬件设备和软件也能产生和收集数据。通过众包模式来采集数据降低了数据获取成本,但是数据质量也参差不齐。维基百科是

众包应用的一个成功案例，由于维基百科建立了一个比较权威的数据质量审核团队来负责输入数据的监控和数据质量的控制，因此其发布的信息内容的质量都非常高。

开放街道地图（OpenStreetMap，OSM）是众包模式应用于 GIS 领域的典型项目。从 2004 年诞生至今，已经拥有上百万用户；而且，其中约 30%的用户至少都在 OSM 地图中提供了一个准确的地点；美国苹果公司甚至将 OSM 地图嵌入 iOS 的 iPhoto 中，使其成为它的地图数据源之一。不同于收费较高的专业测绘地图，OSM 上的地图数据都可以免费下载和使用，因此，这些数据成为许多科研人员的研究首选。OSM 数据质量的研究热点集中在 3 个方面：1）OSM 地图数据质量的评估维度、评估模型和评估方法。文献[60-62]等使用参考地图（如测绘地图、Google 地图和 Bing 地图）与 OSM 地图进行对比，分析 OSM 地图数据的准确性、完整性、时效性和一致性，其中一致性的评估是难点。2）改善 OSM 地图数据质量的方法和系统。文献[63]通过开发一个名为 OSMantic 的标签插件来提高 OSM 数据的语义质量并降低语义的异构性。3）OSM 质量保证工具的开发与应用。OSM 网站提供了一些质量保证工具来帮助使用者和数据贡献者提高数据质量，这些工具包括：bug 报告工具、错误检测工具、可视化工具、监控工具、帮助工具和标签统计[12]。但是，这些工具还存在诸如语言版本限制、检测错误类型少和不提供错误修复等问题，需要继续对其进行完善或者开发出更实用的工具。

3.4 物联网中的数据质量

物联网是由数十亿或数万亿无线识别的“物体”彼此联结和整合而成的动态网络，这些数量庞大的智能设备进行实时数据采集和彼此之间的信息交互，产生了巨大的数据量。物联网中常用的传感设备，如射频识别器和传感器网络，由于自身的局限和电磁环境的影响，采集到的数据中往往存在差错，如重读、漏读、数据冗余和数据错误。此外，物联网中的数据具有多源和多模态的特征，需要进行数据融合和数据集成，即在多源异构数据之间进行交互和转换。由于实现方法不同，这也会影响到集成数据的质量。目前，物联网中的数据质量的研究热点集中在 3 个方面：1）传感器数据验证的方法；2）数据流清洗技术；3）数据质量保证框架。

传感器数据验证的方法主要包括[64]：采用统计分析检测异常值，使用指数加权移动平均法检查漂移，空间一致性方法，解析冗余（检查相关物理模型中的数量），粗差检测，多元统计方法和数据挖掘技术。针对传感器数据的错误和不确定性，常用的数据清洗技术有[65]：基于静态窗口的时间平滑策略、基于自适应窗口的时间平滑策略、基于对象关联度的数据清洗模型、基于 SQL 查询模型的方法、基于时空冗余信息的清洗方法、基于行为模式的数据清洗模型和基于代价评估的最优清洗框架等。为了提高 RFID 数据的质量，研究者提出了一些数据质量保证模型，ESP（Extensible Receptor Steam Processing）是一个基于 SQL 语句的在线清洗五层管道框架[66]，主要由点机制、平滑机制、归并机制、仲裁机制、虚拟化机制 5 个连续的清洗阶段组成。在 ESP 框架的基础上，文献[67]设计了一个并行的质量评估框架。质量评估框架中的每个模块都与一个特定的清洗模块相关联。随着数据流清洗模块的处理，质量保证框架相应地评估出 RFID 应用中的数据流质量水平。

3.5 数据质量与数据开放

在大数据浪潮的推动下，数据开放，尤其是政府公共数据开放成为席卷全球的运动。2009 年，美国联邦政府的数据开放门户网站——Data. gov 正式上线，全面开放政府拥有的 1000 多项公共数据。2010 年，英国政府的数据开放网站 Data. gov. uk 正式出炉，除去地理信息，该网站公布了 3000 多项民生数据。随后，全球五大洲的 30 多个国家都正式建立数据开放门户网站。公共数据开放需要具备八大基本原则[16]：1）数据必须是完整的；2）数据必须是原始的；3）数据必须是及时的；4）数据必须是可读取的；5）数据必须是机器可处理的；6）数据的获取必须是无歧视的；7）数据格式工序是通用非专有的；8）数据必须是不需要许可证的。在这 8 项原则中，第 1），3）－5），8）项都是与数据质量有关的内容。公共数据开放中最棘手的质量问题在于数据的一致性，即开放的数据来自各个政府部分，难免存在交叉，如果两个部门的数据不一致，应该以谁的数据为准？又该如何向公众解释？即使对同一性质、同一类别的数据，新数据还在源源不断地产生和收集，如何保证新旧数据之间的一致性？现有研究未针对一致性问题给出很好的解决方式，还需要继续探索。

结束语 在过去的 30 多年，数据质量的研究取得了一系列成果，但是在新技术和新环境下，仍然存在一些问题：1）数据质量标准化尚未统一，虽然 ISO 制定了数据质量标准 ISO 8000，但是仅有 20 个发达国家参与相关工作，这一标准还存在争议，有待进一步成熟和完善；2）有效的半结构化和非结构化数据模型较少，非结构化数据的质量评估没有得到很好的解决；3）许多物联网应用和大数据应用都需要融合多源异构数据，不过，对这些数据建立一个统一的质量测量标准和模型的研究才刚刚起步；4）一些数据密集型企业（如通信运营商）提出了基于云平台的大数据质量管理系统，但是这些系统的运行效率和实施效果还有待验证；5）随着数据开放脚步的加快，数据作为商品进行交易的趋势会越来越明显，参与交易的数据有些是原始数据，有些是经过清洗、分析和建模的数据，因此如何评估交易数据的质量、数据质量与数据定价的关系以及不同来源数据间的一致性问题都值得我们深入探讨。

参 考 文 献

[1] SCANNAPIECO M，CATARCI T. Data Quality under the Computer Science Perspective [J]. Archivi & Computer，2002，2：1-13.

[2] Financial Accounting Standards Board. Qualitative Characteristics of Accounting Information，Statement of Financial Accounting Concepts No. 2 [R]. Financial Accounting Standards Board，2008：6.

[3] 曹建军，刁兴春，徐永平，等. 信息质量 [M]. 北京：国防工业出版社，2013.

[4] NAUMANN F，ROLKWE C. Assessment Methods for Information Quality Criteria[C]// Proceedings of 5th International Conference on Information Quality. 2000：148-162.

[5] HUANG X Y，ZHANG H. Statistical Data Quality Management：From a Multidisciplinary Perspective [J]. Journal of Business Economics，2011，239(9)：90-96. (in Chinese)
黄向阳，张皓. 多学科视角下的统计数据质量管理 [J]. 商业经济与管理，2011，239(9)：90-96.

[6] WANG R Y，STRONG D M. Beyond accuracy：What data quali-

ty means to data consumers [J]. Journal of management information systems,1996,12(4):5-33.

[7] REDMAN T C. Data quality:the field guide [M]. Boston:Digital Press,2001.

[8] ZEIST R H J,HENDRIKS P R H. Specifying software quality with the extended ISO model [J]. Software Quality Journal,1996,5(4):273-284.

[9] KATERATTANAKUL P,SIAU K. Measuring information quality of web sites:Development of an instrument[C]// Proceedings of the 20th International Conference on Information Systems. North Carolina:ACM,1999:279-285.

[10] DEDEKE A. A Conceptual Framework for Developing Quality Measures for Information Systems[C]// Conference on Information Quality. DBLP,2000:126-128.

[11] FAN B W. Study on the quality of crowdsourcing geographic data－a case of Kunming [D]. Kunming: Yunnan University,2015. (in Chinese)
范博文. 众源地理数据质量研究——以昆明市为例 [D]. 昆明:云南大学,2015.

[12] CAI L,ZHU Y Y. Big Data Quality [M]. Shanghai:Scientific & Technical Publishers,2017. (in Chinese)
蔡莉,朱扬勇. 大数据质量 [M]. 上海:科学技术出版社,2017.

[13] ZOOK M,GRAHAM M,SHELTON T,et al. Volunteered geographic information and crowdsourcing disaster relief: a case study of the Haitian earthquake [J]. World Medical & Health Policy,2010,2(2):7-33.

[14] PIPINO L L,LEE Y W,WANG R Y. Data quality assessment [J]. Communications of the ACM,2002,45(4):211-218.

[15] BALLOU D,WANG R,PAZER H,et al. Modeling information manufacturing systems to determine information product quality [J]. Management Science,1998,44(4):462-484.

[16] 徐子沛. 大数据[M]. 桂林:广西师范大学出版社,2013.

[17] SILBERSCHATZ A. Database System Concepts:Fifth Edition [M]. Beijing:China Machine Press,2010.

[18] CHENG L Q. Data Constraints on the Impact of Data Quality [J]. Journal of Yangtze University (Natural Science Edition),2011,8(5):100-102. (in Chinese)
程录庆. 数据约束对数据质量的影响研究 [J]. 长江大学学报(自然科学版),2011,8(5):100-102.

[19] BOHANNON P,FAN W,GEERTS F,et al. Conditional functional dependencies for data cleaning[C]// IEEE 23rd International Conference on Data Engineering, 2007 (ICDE 2007). IEEE,2007:746-755.

[20] CONG G,FAN W,GEERTS F,et al. Improving data quality: Consistency and accuracy[C]// Proceedings of the 33rd International Conference on Very Large Data Bases. VLDB Endowment,2007:315-326.

[21] INMON W H. Building the data warehouse (2nd ed)[M]. John Wiley & Sons,1996.

[22] 李志刚,马刚. 数据仓库与数据挖掘的原理及应用[M]. 北京:高等教育出版社,2007.

[23] BATINI C,CAPPIELLO C,FRANCALANCI C,et al. Methodologies for data quality assessment and improvement [J]. ACM Computing Surveys (CSUR),2009,41(3):16-68.

[24] Chinese Academy of Sciences Computer Network Information Center. Data Quality Evaluation Method and Index System[EB/OL]. [2015-10-17]. http://www. nsdata. cn/pronsdchtml/1. compservice. standards/pages/3123. html. (in Chinese)
中国科学院计算机网络信息中心. 数据质量评测方法与指标体系[EB/OL]. [2015-10-17]. http://www. nsdata. cn/pronsdchtml/1. compservice. standards/pages/3123. html.

[25] SAATY T L. Decision making with the analytic hierarchy process [J]. International Journal of Services Sciences,2008,1(1):83-98.

[26] 陈水利,李敬功,王向公. 模糊集理论及其应用 [M]. 北京:科学出版社,2005:156-207.

[27] LI D Y,LIU C Y. Study on the Universality of the Normal Cloud Model [J]. Engineering Sciences,2004,6(8):28-34. (in Chinese)
李德毅,刘常昱. 论正态云模型的普适性 [J]. 中国工程科学,2004,6(8):28-34.

[28] LIU C. Sampling Theory and Method of Accuracy Measurement and Quality Assurance for GIS Attribute Data [D]. Shanghai: Tongji University,2000. (in Chinese)
刘春. GIS 属性数据的精度度量及质量控制的抽样原理与方法 [D]. 上海:同济大学,2000.

[29] FAN W,GEERTS F. Foundations of data quality management [J]. Synthesis Lectures on Data Management,2012,4(5):1-217.

[30] MONGE A E,ELKAN C. The Field Matching Problem:Algorithms and Applications[C]// KDD. 1996:267-270.

[31] WANG Y F,ZHANG C Z,ZHANG B B,et al. A Survey of Data Cleaning [J]. New Technology of Library & Information Service,2007,2(12):50-56. (in Chinese)
王曰芬,章成志,张蓓蓓,等. 数据清洗研究综述 [J]. 现代图书情报技术,2007,2(12):50-56.

[32] CAO J J,DIAO X C,CHEN S,et al. Data Cleaning and its General System Framework [J]. Computer Science, 2012, 39(S3):207-211. (in Chinese)
曹建军,刁兴春,陈爽,等. 数据清洗及其一般性系统框架 [J]. 计算机科学,2012,39(S3):207-211.

[33] GALHARDAS H,FLORESCU D,SHASHA D,et al. AJAX:an extensible data cleaning tool[J]. ACM Sigmod Record,2000,29(2):590.

[34] RAMAN V,HELLERSTEIN J M. Potter's wheel: An interactive data cleaning system[C]// VLDB. 2001:381-390.

[35] VASSILIADIS P,VAGENA Z,SKIADOPOULOS S,et al. ARKTOS:towards the modeling,design,control and execution of ETL processes [J]. Information Systems,2001,26(8):537-561.

[36] CUI Y,WIDOM J,WIENER J L. Tracing the lineage of view data in a warehousing environment [J]. ACM Transactions on Database Systems (TODS),2000,25(2):179-227.

[37] BUNEMAN P,TAN W C. Provenance in databases[C]// Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data. ACM,2007:1171-1173.

[38] DENG Z H,WEI Y Z. Study on the Method of Provenance in Science Workflow for Data Publishing [J]. Library & Information,2014,158(3):61-66. (in Chinese)
邓仲华,魏银珍. 面向数据发布的科学工作流数据溯源方法研究 [J]. 图书与情报,2014,158(3):61-66.

[39] BUNEMAN P,KHANNA S,WANG C T. Why and where: A characterization of data provenance[C]// International Con-

ference on Database Theory. Springer Berlin Heidelberg, 2001: 316-330.

[40] GREEN T J, KARVOUNARAKIS G, TANNEN V. Provenance semirings[C]// Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. ACM, 2007: 31-40.

[41] RAM S, LIU J, GEORGE R T. PROMS: A System for Harvesting and Managing Data Provenance [EB/OL]. [2010-11-01]. http://kartik. eller. arizona. edu/WITS_DEMO_final. pdf.

[42] MING H, ZHANG Y, FU X H. Survey of Data Provenance [J]. Journal of Chinese Computer Systems, 2012, 33(9): 1917-1923. (in Chinese)
明华, 张勇, 符小辉. 数据溯源技术综述 [J]. 小型微型计算机系统, 2012, 33(9): 1917-1923.

[43] WANG J L, LI H, WANG Q. Research on ISO 8000 Series Standards for Data Quality [J]. Standard Science, 2010, 439(12): 44-46. (in Chinese)
王军玲, 李华, 王强. ISO8000 数据质量系列标准探析 [J]. 标准科学, 2010, 439(12): 44-46.

[44] RADACK G. Improving Data Portability and Long Term Data Retention through ISO Standards 8000 and 22745 [C]// The Fifth MIT Information Quality Industry Symposium. 2011: 13-15.

[45] SONG L R, PENG J. Introduction and Inspirations of the "Information Quality Act" in the American Federal Government [J]. Journal of Intelligence, 2012, 31(2): 12-18. (in Chinese)
宋立荣, 彭洁. 美国政府"信息质量法"的介绍及其启示 [J]. 情报杂志, 2012, 31(2): 12-18.

[46] SCANNAPIECO M, VIRGILLITO A, MARCHETTI C, et al. The DaQuinCIS architecture: a platform for exchanging and improving data quality in cooperative information systems [J]. Information systems, 2004, 29(7): 551-582.

[47] SIADAT M R, SOLTANIAN-ZADEH H, FOTOUHI F, et al. Data modeling for content-based support environment (C-BASE): Application on epilepsy data mining[C]// Seventh IEEE International Conference on Data Mining Workshops, 2007 (ICDM Workshops 2007). IEEE, 2007: 181-188.

[48] CHU E, BAID A, CHEN T, et al. A relational approach to incrementally extracting and querying structure in unstructured data[C]// Proceedings of the 33rd International Conference on Very Large Data Bases. VLDB Endowment, 2007: 1045-1056.

[49] MARCUS S, SUBRAHMANIAN V S. Foundations of multimedia database systems [J]. Journal of the ACM (JACM), 1996, 43(3): 474-523.

[50] AMATO G, MAINETTO G, SAVINO P. An approach to a content-based retrieval of multimedia data[C]// Multimedia Information Systems. Springer US, 1998: 9-36.

[51] LI W, LANG B. A tetrahedral data model for unstructured data management [J]. Science China Information Sciences, 2010, 53(8): 1497-1510.

[52] MCGILVRAY D. Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information (TM) [M]. California: Morgan Kaufmann, 2007.

[53] CAI L, ZHU Y. The challenges of data quality and data quality assessment in the big data era [J]. Data Science Journal, 2015, 14(2): 2-10.

[54] YANG D, MA Y A, WANG Z, et al. Exploration and reflection of data quality management system of operators under the big data background [J]. China Internet, 2016(1): 73-79. (in Chinese)
杨迪, 马怡安, 王铮, 等. 运营商在大数据背景下对数据质量管理体系的探索及思考 [J]. 互联网天地, 2016(1): 73-79.

[55] WANG J, SONG Z, LI Q, et al. Semantic-based Intelligent Data Clean Framework for Big Data[C]// 2014 International Conference on Security, Pattern Analysis, and Cybernetics. IEEE, 2014: 448-453.

[56] CRAWL D, WANG J, ALTINTAS I. Provenance for mapreduce-based data-intensive workflows[C]// Proceedings of the 6th Workshop on Workflows in Support of Large-scale Science. ACM, 2011: 21-30.

[57] PARK H, IKEDA R, WIDOM J. RAMP: A System for Capturing and Tracing Provenance in MapReduce Workflows [C]// Proceedings of the VLDB Endowment, 2011, 4(12): 1-4.

[58] AKOUSH S, SOHAN R, HOPPER A. HadoopProv: Towards Provenance as a First Class Citizen in MapReduce[C]// TaPP. 2013.

[59] AMSTERDAMER Y, DAVIDSON S B, DEUTCH D, et al. Putting lipstick on pig: Enabling database-style workflow provenance [J]. Proceedings of the VLDB Endowment, 2011, 5(4): 346-357.

[60] HAKLAY M. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets [J]. Environment and Planning B: Planning and Design, 2010, 37(4): 682-703.

[61] CIEPŁUCH B, JACOB R, MOONEY P, et al. Comparison of the accuracy of OpenStreetMap for Ireland with Google Maps and Bing Maps[C]// Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resuorces and Enviromental Sciences. University of Leicester, 2010: 337.

[62] GIRRES J F, TOUYA G. Quality assessment of the French OpenStreetMap dataset [J]. Transactions in GIS, 2010, 14(4): 435-459.

[63] ARSANJANI J J, ZIPF A, MOONEY P, et al. An introduction to OpenStreetMap in Geographic Information Science: Experiences, research, and applications [M] // OpenStreetMap in GIScience. Springer International Publishing, 2015: 1-15.

[64] SUN S, KRAJWEWSKI J L B, LYNGGAARD-JENSEN A, et al. Literature review for data validation methods [EB/OL]. [2011-6-8]. http://www. prepared-fp7. eu/viewer/file. aspx?fileinfoID=215.

[65] FAN H. Study on Unreliable RFID Data Cleaning and Storage techniques for Internet of Things[D]. Changsha: National University of Defense Technology, 2013. (in Chinese)
樊华. 面向物联网的 RFID 不确定数据清洗与存储技术研究 [D]. 长沙: 国防科学技术大学, 2013.

[66] JEFFERY S R, ALONSO G, FRANKLIN M J, et al. A pipelined framework for online cleaning of sensor data streams[C]// Proceedings of the 22nd International Conference on Data Engineering. IEEE, 2006: 140.

[67] WANG C. Study on Quality Assurance Method for Internet of Things of Location-based Service[D]. Nanjing: Nanjing University of Science and Technology, 2015. (in Chinese)
王川. 面向位置服务的物联网数据质量保证方法研究 [D]. 南京: 南京理工大学, 2015.