软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2016,27(7):1605−1625 [doi: 10.13328/j.cnki.jos.005038]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

大数据可用性的研究进展*

李建中, 王宏志, 高 宏

(哈尔滨工业大学 计算机科学与技术学院,黑龙江 哈尔滨 150001)

通讯作者: 李建中, E-mail: lijzh@hit.edu.cn

摘 要: 信息技术的迅速发展,催生了大数据时代的到来.大数据已经成为信息社会的重要财富,为人们更深入地感知、认识和控制物理世界提供了前所未有的丰富信息.然而随着数据规模的扩大,劣质数据也随之而来,导致大数据质量低劣,极大地降低了大数据的可用性,严重困扰着信息社会.近年来,数据可用性问题引起了学术界和工业界的共同关注,展开了深入的研究,取得了一系列研究成果.介绍了数据可用性的基本概念,讨论数据可用性的挑战与研究问题,综述了数据可用性方面的研究成果,探索了大数据可用性的未来研究方向.

关键词: 大数据;数据可用性;数据质量;数据清洗;数据管理

中图法分类号: TP311

中文引用格式: 李建中,王宏志,高宏.大数据可用性的研究进展.软件学报,2016,27(7):1605−1625. http://www.jos.org.cn/1000-9825/5038.htm

英文引用格式: Li JZ, Wang HZ, Gao H. State-of-the-Art of research on big data usability. Ruan Jian Xue Bao/Journal of Software, 2016, 27(7):1605−1625 (in Chinese). http://www.jos.org.cn/1000-9825/5038.htm

State-of-the-Art of Research on Big Data Usability

LI Jian-Zhong, WANG Hong-Zhi, GAO Hong

(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

Abstract: The rapid development of information technology gives rise to the big data era. Big data has become an important wealth of information society, and has provided unprecedented rich information for people to further perceive, understand and control the physical world. However, withthe growth in data scale, dirty datacomes along. Dirty data leads to the low qualityand usability of big data, and seriously harms the information society. In recent years, the data usability problems have drawn the attentions of both the academia and industry. In-Depth studies have been conducted, and a series of research results have been obtained. This paper introduces the concept of data usability, discusses the challenges and research issues, reviews the research results and explories future research directions in this area.

Key words: big data; data usability; data quality; data cleaning; data management

信息技术的迅速发展,特别是数据获取技术的突飞猛进,催生了大数据时代的到来,包括我国在内的世界很多国家都在很多领域积累了 PB(10^{15} 字节)级以上规模的大数据.这些大数据对于国民经济发展、社会进步、社会安全和稳定、科学研究模式变革等人类社会的各个方面都具有重大价值,为人类更深入地感知、认识和控制物理世界提供了前所未有的丰富信息.大数据已经开始造福于人类,成为信息社会的重要财富,引起了学术界、工业界和各国政府的高度重视,研究工作风起云涌.

* 基金项目: 国家重点基础研究发展计划(973)(2012CB316200); 国家自然科学基金(U1509216, 61472099)

Foundation item: National Basic Research Program of China (973) (2012CB316200); National Natural Science Foundation of China (U1509216, 61472099)

收稿时间: 2016-05-12; 采用时间: 2016-05-17; jos 在线出版时间: 2016-05-19

CNKI 网络优先出版: 2016-05-18 16:45:38, http://www.cnki.net/kcms/detail/11.2560.TP.20160518.1645.001.html

然而,随着大数据的爆炸性增长,劣质数据也随之而来,导致数据质量低劣,极大地降低了数据的可用性.国外权威机构的统计表明:美国的企业信息系统中,1%~30%的数据存在各种错误和误差[1];美国的医疗信息系统中,13.6%~81%的关键数据不完整或陈旧[2].国际著名的科技咨询机构 Gartner 的调查结果显示:全球财富 1 000 强企业中,超过 25%的企业信息系统中存在数据错误[3].

数据可用性问题及其所导致的知识和决策错误已经在全球范围内造成了恶劣的后果,严重困扰着信息社会.例如:在医疗方面,美国由于数据错误引发的医疗事故每年导致的患者死亡人数高达 98 000 名以上[4];在工业方面,错误和陈旧的数据每年给美国的工业企业造成约 6 110 亿美元的损失[5];在商业方面,美国的零售业中,每年仅错误标价这一种数据可用性问题的诱因,就导致了 25 亿美元的损失[6];在金融方面,仅在 2006 年,在美国的银行业中,由于数据不一致而导致的信用卡欺诈失察就造成 48 亿美元的损失[7];在数据仓库开发过程中,30%~80%的开发时间和开发预算花费在清理数据错误方面[8];数据可用性问题给每个企业增加的平均成本是产值的 10%~20%[9].因而,大数据的广泛应用对数据可用性的保障提出了迫切需求.

数据可用性问题是信息化社会的固有问题,不仅西方发达国家存在,我国也存在.例如,我们通过对我国某个大型医药企业信息中心的大规模数据进行抽样检验,发现 10%以上的数据存在错误.

综上所述,确保数据可用性是关系到信息社会存亡的重大任务,是有效发掘和利用大数据价值的前提.深入开展数据可用性研究,具有重要的战略意义.最近几年,数据可用性的研究引起了全球的关注.中国国家重大基础研究发展计划(973 计划)2012 年启动了为期 5 年的“海量数据可用性基础理论与关键技术研究”项目.目前,大数据可用性的研究蓬勃开展,成为大数据研究的一个重要方面,取得了很多研究成果.本文旨在介绍大数据可用性的基本概念,分析大数据可用性的研究问题,介绍大数据可用性的研究进展,探索未来的研究方向.

1 大数据可用性概念、挑战与研究问题

1.1 数据可用性的基本概念

数据可用性具有很多度量指标.文献[10]列出了 20 个数据可用性指标,文献[11]归纳了 40 个数据可用性指标.我们针对大数据的实际应用,对大数据的可用性度量指标进行了全面而系统的分析,抽象出如下 5 个实际可行的度量指标[12].

- 数据一致性

数据集合中,每个信息都不包含语义错误或相互矛盾的数据.例如,数据(公司=“先导”,国码=“86”,区号=“10”,城市=“上海”)含有一致性错误,因为 10 是北京区号而非上海区号.

- 数据精确性

数据集合中,每个数据都能准确表述现实世界中的实体.例如,某城市人口数量为 4 130 465,而数据库中记载为 400 万.宏观来看,该信息是合理的,但不精确.

- 数据完整性

数据集合中包含足够的数据来回答各种查询,并支持各种计算.例如,某医疗数据库中的数据一致且精确,但遗失某些患者的既往病史,从而存在不完整性,可能导致不正确的诊断甚至严重医疗事故.

- 数据时效性

信息集合中,每个信息都与时俱进,不过时.例如,某数据库中的用户地址在 2010 年是正确的,但在 2011 年未必正确,即数据过时.

- 实体同一性

同一实体的标识在所有数据集合中必须相同而且数据必须一致.例如,企业的市场、销售和服务部门可能维护各自的数据库,如果这些数据库中的同一个实体没有相同的标识或数据不一致,将存在大量具有差异的重复数据,导致实体表达混乱.

设集合 D 的数据一致性、精确性、完整性、时效性和实体同一性分别为 Q_1,Q_2,Q_3,Q_4 和 Q_5,则数据可用性可以定义为

$$usability(D)=\delta_1Q_1+\delta_2Q_2+\delta_3Q_3+\delta_4Q_4+\delta_5Q_5,$$

其中,$\delta_1,\delta_2,\delta_3,\delta_4$ 和 δ_5 是由用户根据实际需要确定的权值,且 $\delta_1+\delta_2+\delta_3+\delta_4+\delta_5=1$.

1.2 大数据可用性的挑战和研究问题

大数据可用性向我们提出了如下 3 个挑战问题.

(1) 量质融合管理:如何实现大数据的数量与质量的融合管理?

现有的大数据管理研究仅关注数据的规模、系统的处理能力和可扩展性,重在“量”的管理,忽视了数据“质”(即质量)的管理.我们面临的第一个挑战是确保大数据的质量,将大数据管理从“量”的管理拓展到“质”的管理,最终实现“量”与“质”的融合管理.为了彻底实现量质融合管理,我们必须研究量质融合管理问题,提出完整的理论体系,解决关键技术问题.

(2) 劣质容忍原理:如何完成劣质数据上的精确或近似计算?

数据错误几乎无处不在已成为不争的事实.“劣质容忍”是指在数据存在错误的情况下,如何完成精确或近似计算.为了实现劣质容忍,我们必须完成如下两个挑战性任务:第一,自动发现并修正大数据的错误,将可校正的劣质数据修复为完全正确的可用数据,支持正确的计算;第二,很多数据错误无法完全修复,经过修复后,这些数据成为部分正确的弱可用数据.我们必须解决如何在弱可用大数据上完成高质量的近似计算.

(3) 深度演化机理:如何认知大数据演化的机理,追索数据错误根源?

数据不是一成不变的,它会随着时间和物理世界的变化而发生演化.现有的大数据研究忽略了按数据的演化机理所进行的研究,使得数据错误的根源难以追索.我们需要探索大数据的深度演化机理,即,以可用性为核心的多源信息集合在时间、空间、形态、粒度等多个维度上正向协同的演化机理.

为了应对上述挑战,我们需要以数据的一致性、精确性、完整性、时效性和实体同一性为核心,以保障信息可用性为目标,研究以下 7 个问题,创建一套完整的确保海量信息可用性的理论、方法和技术.

(1) 大数据可用性的表达机理

首先,探索数据的一致性、精确性、完整性、时效性、实体同一性的语义规则,建立大数据可用性语义表达规则系统,判定规则系统可否公理化,如果可公理化,则建立公理系统;然后,确定从大数据自动发现语义规则问题的计算复杂性,并设计求解算法;最后,确定大数据可用性自动推理问题的计算复杂性,并设计求解算法.

(2) 大数据可用性的判定理论

首先,分别建立大数据的一致性、精确性、完整性、时效性、实体同一性的数学模型;然后,根据大数据研究可用性这 5 项指标之间的相互影响,建立大数据可用性的综合数学模型;最后,确定大数据可用性判定问题的计算复杂性,设计求解算法.

(3) 大数据的演化原理

探索大数据的演化过程,建立大数据演化的世系模型及追踪技术,包括时空、多粒度、多路径和不确定的大数据演化的理论,演化的可逆性判定与近似求解算法,演化描述的复杂性理论和方法,以及网络化、多粒度、随机化的世系追踪和数据错误追踪技术.

(4) 大数据量质融合管理的理论和技术

首先,建立支持大数据量质融合管理的数据模型和相关理论,包括数据的逻辑结构、运算系统、数据的语义约束模型;然后,解决数据质量管理模型和理论与传统数据管理模型和理论的融合问题,建立大数据量质融合管理的模型和理论;最后,研究大数据量质融合管理关键问题的可计算性和计算复杂性,并设计求解算法.

(5) 高质量数据获取与整合的理论和技术

高质量数据的获取,是确保大数据可用性的重要环节.大数据的来源多种多样,类型千差万别,质量参差不齐,整合困难.这些问题在当今突飞猛进的传感网和物联网背景下尤为严重.我们需要解决如下具有挑战性的问题:最优化逼近物理世界的高精度数据获取的理论和技术、最大化数据与问题求解相关性的高相关数据获取的理论和技术、最小化无价值数据的高纯度数据获取的理论和技术.

(6) 数据错误自动检测与修复的理论和技术

数据错误的自动检测和修复是十分困难的问题.我们需要在大数据可用性的表达机理、大数据可用性的判定理论、大数据演化原理的基础上,探索大数据错误自动检测和修复的理论和技术,主要包括:数据错误自动检测和修复问题的可计算性理论、计算复杂性理论、修复结果可信性理论、大数据错误自动检测与修复算法.

(7) 弱可用数据上的近似计算的理论和算法

当一个数据集合中的错误不能彻底修复时,我们称其为弱可用数据.于是,弱可用数据上近似计算(如查询、分析、挖掘等)的理论和算法成为重要的研究问题.弱可用数据上的近似计算不同于传统意义下的近似计算,它是在具有一致性错误、完整性错误、精确性错误、时效性错误或实体同一性错误的数据上近似地求解满足给定精度要求的问题的解.现有的近似理论与算法无法支持弱可用数据上的近似计算,因此,我们需要研究弱可用数据近似计算的可行性理论、弱可用数据计算问题的计算复杂性理论和算法、弱可用数据近似计算结果的质量评估理论.

2 大数据可用性的研究进展

近年来,人们开展了大量的研究工作,取得了很多研究成果,文献[12–14]是关于数据可用性早期研究工作的综述.本节综述大数据可用性研究的新进展,包括数据可用性表达机理、数据可用性判定的理论和方法、数据错误检测与修复的理论与方法、弱可用数据近似计算的理论和算法、高质量数据获取的理论和方法、大数据可用性管理系统.

2.1 数据可用性的表达机理的研究进展

数据可用性的表达机理主要解决如何表达一个数据集合的数据一致性、精确性、完整性、时效性和实体同一性及其相关理论问题,为数据可用性的判定和数据错误的自动发现与修复奠定基础.

(1) 数据一致性的表达机理

文献[15]对函数依赖理论进行了扩展,提出了基于条件函数依赖的数据一致性表达机制,证明了存在有穷推理规则集使该机制可有穷公理化,给出了具有 4 条推理规则的公理系统,并证明了公理系统的有效性和完备性.文献[16]研究了如何从数据中有效地发现条件函数依赖的问题,提出了 4 种有效发现条件函数依赖的算法.文献[17–19]研究了条件函数依赖的推理问题、覆盖问题、检测问题、传递问题的计算复杂度及其求解算法.文献[44]研究了条件函数依赖的置信度评估问题,提出了基于抽样的条件函数依赖置信度评估算法,并证明了该算法能够以 $1-\delta$的概率给出相对误差小于ε的估计结果,其中,$\varepsilon>0$,$0<\delta<1$.

文献[20]针对条件函数依赖无法描述“并”语义的问题,提出了表示支持“与”和“并”语义的扩展的条件函数依赖.文献[21]提出了微条件函数依赖.文献[20,21]分别证明了各自的扩展条件函数依赖规则系统是可公理化的,建立了公理系统,证明了公理系统的有效性和完备性,并确定了自动推理问题的计算复杂性,提出了求解算法.文献[20,21]还分别确定了各自的扩展条件函数依赖的自动挖掘问题的计算复杂性,并提出挖掘算法.文献[22]提出了概率数据库中随机条件函数依赖.

文献[23]在有时间戳的数据上提出了序列依赖语义规则,用来描述随时间变化数据的一致性约束,试图解决随时间变化数据的一致性错误的发现和修复问题.

文献[24]针对异构数据源中由数据格式不一致引发的一致性错误,利用属性值的相似性扩展了函数依赖,用来描述异构数据的一致性,发现和修复异构数据的一致性错误.

文献[25]利用统计模型来描述数据的一致性,并通过求解和比较模型参数的方法来发现和修复数据不一致性错误.文献[26]提出了基于统计知识的数据不一致性描述方法,并给出了基于超团的数据一致性提升算法.

文献[71–75]研究了数据一致性规则挖掘问题,分别提出了在数据集合中挖掘各种数据一致性规则的算法.

(2) 数据完整性的表达机理

传统的数据完整性研究工作一般都建立在封闭世界或开放世界假设的基础上.封闭世界假设表示数据库包含了所有表述现实世界实体的元组,这些元组的某些属性值可能遗缺.开放世界假设表示数据库中不仅属性值可能遗失,描述实体的元组也可能完全遗缺.然而,现实世界的数据库经常既不是完全封闭的,也不是完全开

放的.基于这个考虑,文献[27,28]扩展了包含依赖,提出了一种表示数据完整性的包含依赖规则系统,定义了规则的语法和语义,证明了规则系统是可公理化的,建立了公理系统,证明了 22 个相关基础问题的不可计算性、coNP-完全性、Σ_3^p-完全性、Π_2^p-完全性或 EXPTIME-完全性.文献[29]扩展了文献[27,28]的研究结果.

文献[30,31]将传统的完整性理论扩展到 XML 数据上,研究了 XML 数据的问题完整性表达机制.

(3) 数据时效性的表达机理

文献[32]在同一个实体具有多个元组的假设下,提出了一种基于规则的数据时效性表示机制,定义了同一实体对应的不同元组的属性值的时序关系表示方法,提出了基于实体的最新值的时效性查询语义,并给出了应用元组间的时序关系和拷贝关系推导实体最新信息的推理机制.基于这种数据时效性表达机制和时效性查询语义,文献[32]还给出了用户查询的计算复杂性,并研究了在实体最新值缺失的情况下如何扩展元组间拷贝关系以找到实体的最新值.但是,“同一个实体具有多个元组”的假设,使得这种数据时效性表达机制具有很大的局限性.

为了突破文献[32]的数据时效性表达机制的局限性,文献[33]提出了基于不确定规则的数据时效性表达机制,定义了数据时效性规则的语法和语义,证明了规则系统是可有穷公理化的,建立了具有两条推理规则的公理系统,证明了公理系统的有效性和完备性,同时证明了数据时效性规则自动推理问题是 P 问题,确定了问题的时间复杂性下界,并给出求解推理问题的最优化算法.文献[33]还证明了规则挖掘问题是 NP-难的,并给出 $O(n)$和$O(n^2)$时间近似挖掘算法,n 是数据集合的大小.

(4) 实体同一性的表达机理

文献[34]提出了基于规则的实体同一性表达机制,定义了实体同一性规则的语法和语义,证明了对于任意数据集合 D 都存在一个有效、一致、完整和独立的实体规则集合 Σ,同时证明了可满足问题和语义蕴含问题皆为 P 问题,而且它们的时间复杂性下界都是 $\Omega(|\Sigma|^2)$,并给出了求解这两个问题的时间复杂性为 $\Omega(|\Sigma|^2)$的最优化算法.文献[34]还研究了从数据集合 D 中挖掘实体同一性规则的问题,证明了该问题是 P 问题且其时间复杂性下界为 $\Omega(|D|^2)$,给出了时间复杂性为 $O(|D|^2)$的最优化求解算法,并且证明该算法能够从数据集合 D 中挖掘出满足有效性、一致性、完整性和独立性的实体同一性规则集合,即,算法是正确的.

文献[35]提出了实体同一性描述规则,系统地研究了规则的推理问题,提高了描述实体同一性的能力.文献[36]进一步在动态语义下研究了实体同一性规则的相互作用及推理问题.

文献[37]提出用否定规则描述实体的同一性,并研究了否定规则对实体同一性的影响.文献[38]提出了用聚集约束来描述实体同一性的方法.文献[39]结合 EM 算法和无监督学习方法,提出了获取实体同一性描述规则的方法.

(5) 数据精确性的表达机理

文献[40]在“同一个实体具有多个元组”的假设下,提出了一种基于规则的数据精确性表示机制,定义了同一实体对应的不同元组的属性值之间的精确性偏序关系;在此基础上定义了数据精确性规则的语法和语义,确定了规则系统的推理问题的计算复杂性,给出了求解问题的算法,并提出了相应的精确性错误修复框架.

文献[41]把不确定性视为精确度低的现象,提出了一种基于可能世界语义的数据精确性描述方法,并给出了对应的精确性评估算法.

2.2 数据可用性判定的研究进展

数据可用性判定问题定义如下:给定任意数据集合 D,计算 D 的可用性 $usability(D)$.数据可用性判定的关键在于数据的一致性、精确性、完整性、时效性和实体同一性的判定.本节介绍数据一致性、精确性、完整性、时效性和实体同一性的研究进展.

(1) 数据一致性判定的理论和算法

文献[42]系统地研究了数据一致性判定问题.

- 首先,基于数据一致性表达机制建立了数据一致性的数学模型.给定数据集合 D 和 D 上的条件函数依赖集 Σ,D 的一致性定义为 $Consistency(D)=|D'|/|D|$,其中,D'是 D 中满足 Σ 的最大子集,并证明了:如果

D'满足条件函数依赖集合 Σ,则满足 Σ^*,从而降低了求解判定问题的难度;

- 其次,研究了数据一致性判定问题的计算复杂性和近似性,证明:数据一致性判定问题是 NP-完全的;不存在多项式时间$(2-\varepsilon)$-近似算法,除非 unique game 猜想为真;问题是 1.3606 不可多项式时间近似的,除非 P=NP;规则为 3 且属性为 4 时,问题是 1.0625 不可近似的;
- 最后,给出了一种近似比最优化的 $O(n\log n)$时间的 2-近似算法,并给出了一种 $O(\log(n))$时间的$(2+\varepsilon)$-随机近似算法.

文献[43]研究了使用条件函数依赖评价数据一致性的关键问题——最小元组删除集的计算问题,证明了该问题是 NP-完全的,给出了基于冲突图的近似求解算法,算法的近似比为 $2-(1/2)^{|\Sigma|}$,其中,Σ是给定的条件函数依赖集.

(2) 数据时效性判定的理论和算法

数据时效性判定方法可以分为两类,即,基于时间戳的时效性判定和独立于时间戳的时效性判定.

基于时间戳的时效性判定要求数据集合中每个数据值具有时间戳.文献[45−50]把数据从上一次更新到本次使用的时间间隔定义为数据年龄 Age,从不同角度定义了数据的时效性.文献[45,48]假设数据有一个确定的保质期 T.给定一个数据 V,文献[45]把 V 的时效性定义为概率 $\Pr[Age(V)-T(V)>0]$,而文献[48]则在 $T(V)>Age(V)$ 的条件下把数据的 $Age(V)$定义为 V 的时效性.文献[46,47]假设数据时效性随时间流逝的减弱程度可以用时效性衰减函数 f 来刻画,定义数据 V 的时效性为 $e^{-f(V)\times Age(V)}$.文献[49]把数据年龄定义为数据的时效性.文献[50]提出了一种基于模糊逻辑来推断时效性衰减函数的时效性判定方法.

针对实际应用中时间戳常常不存在的情况,文献[33]研究了独立于时间戳的数据时效性判定方法.

- 首先,在数据时效性表达机制的基础上提出了数据时效性的数学模型;
- 然后,证明了时效性判定问题是 P 问题,且其时间复杂性下界为$\Omega(n^2)$;
- 最后,给出了两种基于时效图的数据时效性判定算法:一种是针对一般时效图的 $O(n^2\log n)$时间算法,另一种是针对无环时效图的 $O(n^2)$时间最优化算法.

文献[51]研究了相对于查询的数据时效性判定问题,建立了数据相对时效性的数学模型.针对最新值查询和时效序列查询这两类查询,提出了查询结果的时效性判定方法,并将每类查询作为一个整体,给出了数据集合相对于每类查询的平均时效性判定方法.

(3) 数据完整性判定的理论和算法

文献[52]研究了数据完整性判定问题.

- 首先,给出了一种数据完整性模型,这种数据模型避免了假阳性错误,即,避免能够由函数依赖可导出的值被误判为缺失值;
- 然后,确定了数据完整性判定问题的计算复杂性,证明了该问题是 P 问题,且其时间复杂度下界是$\Omega(n^2)$;
- 最后,给出了时间复杂度为 $O(n^2)$的最优化判定算法,并给出了一种适用于大数据的(ε,δ)-近似算法,其时间和空间复杂性均为 $O(\varepsilon^{-2}\ln(\delta^{-1}))$.

文献[53]进一步扩展了这项研究结果.

文献[54]综述了早期的数据完整判定的研究工作,介绍了不同种类的数据完整度的定义和计算方法.

文献[55]提出了判定地理数据完整性的计算方法.文献[56]给出了时间序列数据完整性判定方法.文献[57,58]给出了其他特定数据集合的数据完整性判定方法.

(4) 数据精确性判定的理论和算法

对于数据集合中的不精确值,其精确值难以预知.于是,数据精确性判定问题是一个非常困难的问题.文献[59,60]提出了一种多模态数据集的精确性判定方法,使用均方误差这一参数衡量数据的精确性.该方法把数据集合中的数据分为 3 类,即可量度型、可比性型、分类型.针对每类数据的特点,建立了不同的精确性数学模型,并组合这些数学模型,最终建立了数据精确性的数学模型.针对精确值可知与不可知两种情况,该方法提供了不

同的数据精确性判定算法.在精确值可知的情况下,直接用均方误差来判定数据的精确性;在精确值缺失的情况下,针对不同的数据类型,分别提出了二次规划算法、迭代算法和 EM 算法,以求解各类数据精确性的判定问题,最后确定整个数据集合的精确性.

(5) 实体同一性判定的理论和算法

文献[61]提出了一种基于实体识别结果的实体同一性判定方法.

- 首先,定义了元组之间的距离,用以描述任意两个元组在所有属性上的值的不一致的程度,并基于元组的距离进一步定义了实体同一性的数学模型;
- 其次,研究了实体同一性判定的计算复杂性,证明了实体同一性判定问题是 NP-难的;
- 最后,分别给出了求解实体同一性判定问题的 4 个子问题的 $O(n\log n)$时间 2-近似算法和 $O(n\log n)$时间 n-近似算法,并最终给出了 $O(n\log n)$时间的 n-近似算法.

2.3 数据错误检测与修复研究进展

数据可用性具有数据一致性、完整性、精确性、时效性和实体同一性这 5 个维度.数据错误检测和修复方面的研究主要围绕一致性、完整性、时效性和实体同一性这 5 个方面开展研究,在考虑了多维度错误的同时检测与修复问题.本节介绍这些研究结果.

(1) 数据一致性错误的检测与修复

基于条件函数依赖,人们提出了很多数据一致性错误的检测与修复方法.文献[62,63]针对集中存储的关系数据库,使用 SQL 语言设计了自动检测算法,用于查找违反条件函数依赖和条件包含依赖的元组.文献[64]研究了在分布式环境下检测数据一致性错误的问题,目标是最小化数据通信量.文献[65]给出了一种增量式的分布式数据库中数据一致性错误的检测方法.

文献[66]给出了基于主数据的数据一致性修复问题:给定数据集合 D、条件函数依赖集 Σ 和主数据 D_m,修复 D 中的数据一致性错误.

- 首先,证明了该问题是 NP-完全的、不存在多项式时间的 $O(\log n)$-近似算法,除非 NP=P;
- 然后,提出了与主数据相关的修复规则,证明修复规则的一致性问题和覆盖问题是 coNP-完全的;
- 最后,基于主数据和修复规则,提出了启发式数据一致性错误修复算法.

文献[67]研究了基于用户反馈的数据一致性错误修复问题,证明了对于多类查询,该问题分别是 NP-完全的、Σ_2^P-完全的或 PSPACE-完全的;而对于选择-投影查询,该问题是 P 问题.针对选择-投影查询,给出了时间复杂性为 $O(n^2)$的精确的问题求解算法.

文献[68]研究了数据一致性修复问题中,数据集合的一致性表达规则可能不正确的问题,提出了相对信任的概念,用以判定是数据可信,还是一致性表达规则可信,从而确定是修改数据还是修改一致性表达规则,并提出了相应算法.

文献[69]设计了一个数据不一致性修复框架,提出了不同的一致性约束条件类型和选择最优值的方法,用以解决数据一致性错误修复问题,提出了新的修复规则语义和一种计算最优修复方案的算法.

文献[70]提出了基于 Hadoop 并行平台的不一致数据检测与修复算法.

(2) 数据时效性错误的检测与修复

文献[76]合并时效函数依赖与近似函数依赖,提出了时效近似函数依赖,并给出其基本定义和一些相关的数据挖掘技术.

文献[77]提出了一种模型,通过部分时间顺序和时间约束来指定数据时效性,并通过不变的条件函数依赖来强化数据时效性.

文献[78]针对网络数据的时效性提出了时效规则发现问题,给出了基于关联规则和离群点识别的机器学习算法,从而实现了数据的时效性检测与修复.

文献[79]提出一类新的时效性修复规则 CRR,将规则和统计的方法结合起来修复过时数据.该规则一方面能够通过规则模式表达领域知识,另一方面还能够使用其特有的分布表来描述数据随时间变化的统计信息.由

于静态数据上的 CCR 模式生成问题是 NP-难的,该文献给出了两种解决该问题的多项式时间近似算法,并提出了修复代价约束条件下的最优修复计划产生算法.文献[80]研究动态数据环境下 CRR 的生成问题以及基于 CRR 的数据修复问题,提出了两条剪枝原则,用于快速产生 CRR 模式.

(3) 数据完整性错误的检测与修复

数据完全性错误检测与修复的研究主要集中在缺失值的填充方法研究上.按照缺失值填充所使用的数据来源,缺失值填充方法可以分为内部数据填充和外部数据填充两类.内部数据填充是利用正在修复的数据集合中的值来填充缺失值;而外部数据填充是指利用正在修复的数据集合之外的数据来填充缺失值,如网上数据.

文献[81]研究缺失值填充的基础理论,分析了 3 种不完整数据修复方法,即,基于确定答案和表现系统的方法、基于逻辑的方法和基于相对值排序的方法,为缺失值填充建立了理论基础.文献[98]证明,在普通条件下缺失值的最小化修复是 NP-难问题,并提出了有效的近似算法.

文献[82−89]提出了利用内部数据填充缺失值的方法.文献[82,83,85−88]研究了连续型缺失值的填充问题.文献[83,87,88]提出了如何填充性别、国家等离散型缺失值的方法.文献[89,90]提出了利用相似性规则填充非数值型缺失值的方法.

文献[91−93]研究了用网络上的数据填充缺失值的问题,提出了多种利用网络数据来填充缺失值的方法.文献[95]提出了以最小化网络查询代价为目标的利用网络数据填充缺失值的方法.文献[96]提出了利用贝叶斯网络与众包相结合,利用网络数据缺失值填充的方法,这种方法在利用贝叶斯网络进行简单概率推理的基础上,结合了众包技术,将部分缺失元组发送至众包平台,利用人工反馈获得可靠的填充值.

文献[94]提出了利用内部数据和外部数据相结合的缺失值填充方法.

文献[97]研究了半结构化数据开闭标签不匹配的问题,提出了一种动态规划算法和一种基于经验主义的组合边界算法,以填充丢失的标签.

(4) 实体同一性错误的检测与修复

实体同一性错误的检测与修复的关键是实体识别,即,从数据集合中发现描述现实世界同一实体的不同数据.实体识别的研究结果很多.

使用众包的方法来提高实体识别精度,是目前的一个研究热点.文献[99]提出一种基于众包的实体识别方法,提高了识别结果的精度.文献[100]针对已有的基于众包的实体识别算法需要开发者参与而导致实体匹配算法扩展性不强的问题,提出了无需开发者参与的众包算法.文献[101]研究了提高众包方法中人类回答出错的问题,提出了 bDENSE 算法,提高了利用众包方法识别实体的准确率.文献[102]提出了一种人机混合的实体识别方法.文献[103]针对实体识别中计算机对一些实体的识别准确率低而需要人工标记相似实体对的问题,提出了一种以最小化人工标记实体对数目的实体识别方法.

非结构数据中的实体识别问题也是目前人们关注的热点,目前的工作主要集中在链接描述同一实体的不同命名上.文献[104]针对微博文本的实体链接问题,通过社交和时间上下文分析,考虑实体流行程度、实体近期流行程度、用户兴趣这 3 个方面,提出了一种实体链接方法.文献[105]针对网络文本和复杂信息网络的关联问题,提出了一种概率模型,以链接各个命名的实体,并使用 EM 算法对模型中的链接权重进行估计.文献[106]针对商业过程中的重复事件检测问题,提出了一个新的事件相似度函数,迭代地计算邻居间的相似度,并给出了相应的算法.

动态变化数据中的实体识别也是一个重要的研究问题.文献[107]针对时序数据建立一个实体变化模型,用来实现动态变化数据中的实体识别.文献[108]提出了一种方法,解决如何利用已有的实体识别结果来节省数据变化所带来的实体识别的冗余工作的问题.文献[109]针对大数据时代数据变化快而导致的描述实体的记录集合快速失效问题,提出了一个新的解决方案,递增、有效地更新实体识别结果,确保实体识别结果的质量.

提高实体识别的效率是实体识别的目标之一.文献[110]针对内嵌数据的重复删除过程中磁盘瓶颈对主存影响的问题,提出了一种能够动态地从磁盘中预先提取指纹并存入缓存的框架,以有效地支持实体识别.文献[111]采用基于抽样的算法来提高实体识别的效率,该算法与其他算法相比,效率提升了 2~3 个数量级.文献[112]

提出了大数据实体识别的近似算法.文献[113]提出了多模态数据的实体识别方法.文献[114]在无需计算元组相似性的前提下,提出了一种新的实体识别规则,有效地描述了实体匹配条件,并提出从数据中发现规则的有效算法,给出了基于规则的实体识别算法.

提高实体识别的准确率,是实体识别的另一个目标.文献[115−118]提出了一系列具有高准确率的实体识别方法.

实体不同一错误的修复,通常称为真值发现,即,发现描述同一实体同一属性的不同值中的真实值.文献[119]针对真值发现中的长尾现象,提出了一种置信度可知的真值发现方法.文献[120]提出了不同数据源的数据冲突的两种解决方法.文献[121]利用概率模型解决了流数据中的真值发现问题.文献[122]提出了一种自动真值发现算法,利用 Sherlock 规则和参考表,发现冲突数据集合中的真值.

(5) 多维度错误的组合修复方法

文献[123]研究了数据可用性各个维度之间的关系,发现:

- 完整性错误修复结果会引起一致性、时效性、精确性的变化;
- 一致性错误修复结果会引起时效性和精确性的变化;
- 时效性错误修复结果会引起一致性和精确性的变化;
- 精确性错误修复结果不会引起其他可用性维度的变化.

这些结果为多维度错误的组合修复奠定了基础.

文献[124]提出了同时修复数据时效性错误和一致性错误的方法,利用最新的和一致的实体数据构造出准确的实体描述数据.

文献[125]提出了同时检测和修复数据的一致性错误、完整性错误和实体同一性错误的优化技术.

(6) 其他错误检测与修复方法

一些学者研究了数据错误根源的发现问题.文献[126]利用贝叶斯方法来建立代价模型,用以诊断大数据中发生的错误,并判断错误发生的源头和原因.文献[127]扩展了文献[126]中提出的基于贝叶斯推理的代价模型,用于发现数据中的错误的共同属性,追寻产生错误的原因.

很多学者研究了如何组合现有方法实现数据错误的检测与修复.文献[128]结合定量数据清洗和逻辑数据清洗方法,将数据清洗问题转化为最小统计失真修复问题,给出了基于 EMD 的修复策略.文献[129]将模式映射和数据修复两个问题综合考虑,给出了模式映射和数据修复问题的综合定义,提出了求解该问题的基于 Chase 的算法.文献[130]提出了一种统一框架来处理数据清洗问题.文献[131]提出了基于正则表达式的结构化数据修复算法.

一些学者研究了特定类型数据的错误清洗问题.文献[132]提出了基于众包的不确定数据的清洗方法.文献[133]提出了事件数据错误的检测与修复方法.

一些学者研究了动态数据上的错误检测与修复问题.文献[134]提出了动态变化数据的数据修复分类器,在考虑数据语法错误和语义错误的基础上加入了用户的修复偏好,根据历史修复记录来预测当前所需的修复方法.文献[135]将增量错误检测技术应用于分布式数据上,试图解决在数据不断变化的情况下数据错误的动态检测问题,证明了该问题是 NP-难的,提出了增量式错误检测算法.文献[136]根据数据分布,利用数据分区和机器学习技术预测可能的数据更新,提出了一种动态数据错误检测与修复方法.

2.4 高质量数据获取的研究进展

大数据的来源多种多样,如 Internet 上的丰富数据资源、物联网和科学实验系统中的传感器或传感网.数据源的质量会极大地影响数据的可用性.高质量数据源的选择,是获得高质量数据的基础.高质量的大数据获取方法,是确保数据获取质量的关键.近年来,高质量数据获取的研究主要集中在高质量数据源的选择、Internet 数据的高质量获取方法和基于传感器或传感网的感知数据的高质量获取方法这 3 个方面.本节介绍这 3 个方面的研究进展.

(1) 高质量数据源的选择方法

为了确保高质量地获取数据,高质量的数据源选择是十分重要的.文献[137]研究了网络信息抽取器在数据中引入错误的概率,提出了数据源的正确度和所抽取数据的正确度的联合推理概率模型,用于解决数据源可信度的评估问题.文献[138]针对传统数据集成使用的数据源选择算法仅考虑静态数据源以及仅以数据融合准确度为衡量标准的问题,提出了新的数据源质量评估标准和它们的计算模型,并针对动态数据源给出了一种高质量数据源选择方法.文献[139]针对数据融合问题,提出了基于数据源的质量,通过贝叶斯分析,推导数据源质量的方法.文献[140]针对实际应用中经常出现的同一个实体具有多个冲突数据源的情况,提出了源数据可靠性的估计方法.文献[141]介绍了作者建立的数据源选择系统,很好地解决了数据一致性问题.

(2) Internet 数据的高质量获取方法

文献[142]发现,数据源之间的数据复制关系能够帮助系统更好地选取高质量的数据源、改善集成数据的可用性.针对静态数据,提出了基于贝叶斯分析的方法,判定数据源之间的复制关系,并基于复制关系提出了高质量数据获取与整合的方法,提高了获取与整合后的数据的可用性.文献[143]针对动态数据,提出了利用数据源中的数据更新历史来判定数据源之间复制关系的方法,利用隐马尔可夫模型来判定数据源的复制关系,并利用贝叶斯模型改善数据获取的过程,提高了数据获取质量.文献[144]进一步考虑更复杂的数据复制关系,包括部分数据复制、多个数据源同步复制、多数据源传递复制,给出了判定复制关系、提高数据集成质量的算法.文献[145]给出了一个判定数据复制关系的演示系统原型.

文献[146]对 Internet 数据高质量获取与融合的研究工作进行了系统的综述.

(3) 感知数据的高质量获取方法

文献[147,148]针对无线传感网能量受限的特点,探索了在保障数据精确性的前提下,以最小能量开销获取感知数据的问题,提出了从无线传感网获取数据的(ε,δ)-近似随机算法,确保获取数据的精度大于ε的概率小于δ.

文献[149,150]研究了如何从传感网获取数据,使得物理世界能够被准确近似,从而获取高精度数据.使用 Hermit 插值、三次样条插值、分段线性拟合等方法,提出了多种面向物理过程的高精度数据获取算法,实现了对物理世界的ε-近似[153],其中,$\varepsilon>0$.

文献[151]针对地理位置相近的传感器节点的数据中存在冗余数据的问题,提出了位置敏感的数据获取方法.利用数据源之间的地理关联特征过滤冗余数据,提高了获取的数据在事件监测应用中的可用性,降低了误判的概率.

文献[152]研究了多模态数据获取问题,定义了基于事件模型和代价约束的最优覆盖问题,证明了该问题是NP-完全的,提出了求解该问题的 $O(n^{k-1})$时间准确算法和 $O(n^5)$时间的$(1-e^{-1})$-近似算法,其中,n 是节点个数,k 是节点的类别数.

文献[154,155]分别针对感知大数据和多种应用并行执行这两种情况,提出了感知大数据的支配数据集的获取方法和同时支持多应用的高质量感知数据获取方法.

文献[156,157]针对恶意篡改感知数据的问题,提出了两种确保感知数据不被破坏的方法,从而保证了数据获取的质量.

2.5 弱可用数据近似计算的研究进展

当一个数据集合的数据错误不能被彻底清除时,我们称其为弱可用数据.弱可用数据的计算是一个新研究领域,研究成果还不多,主要集中在弱可用数据的查询、挖掘和查询结果的质量评估上.本节介绍弱可用数据计算的主要研究结果.

(1) 弱可用数据上的查询处理与挖掘

针对具有实体统一性错误的数据,文献[158]研究了弱可用数据的查询处理问题,提出了在具有实体同一性错误的数据上处理选择-投影-连接查询的算法.

文献[159]统一考虑实体识别和数据集成,提出了同时支持实体识别和数据集成的在线查询处理方法.文献[160]提出了在具有实体同一性错误的数据上求解相似性连接的算法.

针对具有完整性数据的数据,文献[161]提出了基于改写关系代数表达式的查询处理方法.文献[162]实现了

一个针对不完整数据的近似查询处理系统,由数据层、推理层、界面层组成,提出了重组原始查询确保返回答案完整的方法.文献[163−168]从不同角度研究了不完整数据上的 Skyline 查询处理问题,提出了一系列 Skyline 查询方法.文献[169]研究了不完整数据上的偏好查询处理问题,提出了一种能够在不破坏偏好支配关系传递性的情况下处理偏好查询的方法.

针对具有一致性错误的数据,文献[170]在主键约束下,提出了一种基于二进制整数规划技术的合取查询处理方法.文献[171]基于匹配依赖,提出了一种数据清洗与查询处理相结合的查询处理方法.

针对弱可用数据上查询结果的质量问题,文献[172]提出利用采样来提高查询的质量,即:清洗小样本集,并利用清洗效果的经验来改善查询结果.文献[173]面向 NoSQL,提出了以满足用户服务质量和数据可用性要求为目标的查询处理方法.

文献[174]研究了弱可用数据挖掘问题,提出了不完全数据上的分类算法.这个研究工作是目前弱可用数据挖掘方面的唯一研究结果.

(2) 弱可用数据查询结果的质量评估

文献[175]研究了评估查询结果一致性的方法,证明了查询结果一致性评估问题是 coNP-完全问题,并针对一致性错误,设计了基于抽样的查询结果一致性评估算法.

文献[176]使用数据的完整性判定和其他查询结果的完整性来判定给定查询的查询结果的完整性,确定了判定问题的复杂性和查询结果完整的充分条件.文献[177,178]分别使用文献[176]的研究结果研制了两个演示系统:一个用来判定一个查询能否得到完整的查询结果,另一个则处理不完整数据上的查询.文献[179,180]扩展了文献[176]中的结果,不但考虑了缺失的元组,而且退出了在元组中包含缺失值的情况下查询结果完整性的判定算法.

文献[181]提出了使用 RDF 描述数据完整性约束的方法,并利用这些完整性约束给出了判定查询结果完整性的方法.文献[182]给出了一个基于文献[181]的演示系统.

文献[183,184]从逻辑编程角度提出了查询结果完整性的判定方法.文献[185]结合逻辑编程和给定的完整性约束,给出了比文献[176]更多的确保查询结果完整的充分条件.

文献[161]提出了完整性模式的概念,通过在完整性模式上进行代数计算,能够极大地简化查询结果完整性判定的难度.

文献[186]给出了在主数据存在的情况下,判定相对于主数据的数据完整性.给定数据集 D 和查询 Q,该文献研究了如下 4 个判定问题.

- D 能够完整地回答 Q 吗?
- D 是能够完整地回答 Q 的最小数据集吗?
- 是否存在一个有限数据集ΔD,使得 $D\cup\Delta D$ 能够完整地回答 Q?
- 存在能够完整地回答 Q 的数据集吗?

文献[187−190]研究了在实际的商务过程中,如何自动地保证和检查数据完整性的方法.

2.6 数据可用性管理系统的研究进展

最近几年,应用数据可用性的基础理论研究成果,人们已经研制了一系列数据错误检测和修复原型系统.本节介绍这些原型系统.

文献[191]研制了一个旨在确保数据一致性、精确性、完整性和实体同一性的数据错误检测与修复系统.

文献[192]研制了一种可扩展的数据错误修复自动工具,通过采用马尔可夫链对结构化修复的选择,实现了在数据抽取的同时进行数据修复.

文献[193]研制了 Wisteria 系统.该系统针对数据错误检测与修复过程的迭代问题,把逻辑操作和物理实现分离,以反馈为驱动,支持数据错误检测与修复流程的迭代实现和优化.

文献[194]研制了一种新的面向查询的数据错误检测和修复系统.该系统使用神喻的方法进行数据清洗.

文献[195]研制了一个针对数据错误检测和修复通用需求,以高效率、可扩展和易用为目标的数据错误检测

和修复系统.该系统可以在多种平台框架上运行.

文献[196]研制了一个基于知识仓库和众包的数据错误检测和修复系统,对于每个错误数据,可以推荐 top-k 种可能的修复值.

文献[197]在 NADEEF 的基础上,构建了一个具有用户可以自定义大量参数的图形界面、丰富的编程接口、支持交互性等特点的实体解析系统 NADEEF/ER.

文献[198,199]研制了 Cleanix 并行数据错误检测和修复系统.该系统充分运用了 Hyracks 的灵活和高可扩展性的特点,实现了多种数据错误检测和修复任务的并行化,包括异常值监测与修复、缺失值填充、实体识别和冲突消解.

文献[200]研制了一种新的系统,它集成了数据源发现、清洗、转换、语义集成和可视化等组件,并使用了机器学习技术来尽量减少人的干预.

文献[201]研制了以电子商务为背景的基于实体识别的商品信息检索系统.

文献[202]研制了一种基于实体识别的劣质数据管理系统.该系统首先对数据进行实体识别,以实体为单位实现了劣质数据的管理,并支持劣质数据上的近似查询.

3 总结与未来研究方向

综上所述,数据可用性是大数据的一个重要研究方面,已经引起了工业界和学术界的极大关注,开展了大量的研究工作,在数据可用性的表达机理、数据可用性判定的理论和算法、数据错误的检测与修复的理论与方法、高质量数据获取的理论与方法、弱可用数据近似计算的理论与方法等方面取得了大量研究结果,并研制了很多数据错误检测和修复系统.

尽管大数据可用性研究已经取得了很大进展,但是还远远不能满足实际应用的需求,特别是大数据应用的需求.我们仍然面临很多颇具挑战性的问题,需要进一步开展如下研究工作:

(1) 现有的数据可用性研究主要以单个数据集合为研究对象,没有考虑相互关联的多数据集族的整体可用性问题.我们需要以相互关联的多数据集族为对象,深入研究数据可用性的理论和关键技术,包括相互关联的多数据集族的数据可用性表达机理、多数据集族的数据可用性判定理论和方法、多数据集族的交叉关联数据错误的检测与修复的理论和算法等;

(2) 现有的数据可用性计算问题(如可用性判定问题、错误基础与修复问题、弱可用数据计算问题等)的计算复杂性和求解算法研究仍然把“确定图灵机在多项式时间内可解”作为问题易解性的标准,致使很多结果不适于大数据.大数据规模巨大,可达 PB 级甚至 EB 级.多项式时间算法(甚至线性时间算法)难以在人们容忍的时间内求解大数据计算问题.我们需要以“亚线性或对数多项式时间”为大数据计算问题的易解性标准,深入研究大数据可用性计算问题的计算复杂性理论,设计求解大数据可用性计算问题的亚线性和对数多项式时间算法,解决大数据可用性的各种计算问题;

(3) 现有的数据可用性研究结果主要以数理逻辑和统计学为基础,导致很多问题成为 NP-完全问题、Σ_3^P-完全问题、PSPACE-完全问题、EXPTIME-完全问题,甚至不可计算问题.我们需要突破数理逻辑和统计学的限制,开拓大数据可用性的新理论基础,寻觅新的大数据可用性研究方法,努力减少难解问题,提出实际可行的数据可用性保障方法;

(4) 继续深入研究第 1 节提出的 7 个问题:大数据可用性的表达机理、大数据可用性的判定理论、大数据的演化原理、大数据量质融合管理的理论和技术、高质量数据获取与整合的理论和技术、数据错误自动检测与修复的理论和技术、弱可用数据上的近似计算的理论和算法.特别值得注意的是:至今,我们还未见到大数据演化机理和大数据量质融合管理这两方面的研究结果;弱可用数据计算研究的结果非常少见,并且仅集中在弱可用数据的查询处理方面.显然,这两方面需要我们付出更大的努力.

References:

[1] RedmanT. The impact of poor data quality onthe typical enterprise. Communications of the ACM, 1998,41(2):79−82. [doi: 10.1145/269012.269025]

[2] Miller DW, Yeast JD, Evans RL.Missing prenatal records at a birth center: A communication problemquantified. In: Proc. of the AMIA Annual Symp. Bethesda: American Medical Informatics Association, 2005. 535−539.

[3] Swartz N. Gartner warns firms of dirty data. Information Management Journal, 2007,41(3):6.

[4] To ERR is Human: Building a Safer Health System. Washington: National Academies Press, 2000.

[5] Eckerson W. Data warehousing special report: Data quality and the bottom line. In: Proc. of the Applications Development Trends. 2002.

[6] English LP. Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits. New York: Wiley, 1999.

[7] Woolsey B, Schulz M. Credit card statistics, industry facts, debt statistics. In: Proc. of the Google Search Engine. 2010.

[8] Shilakes C, Tylman J. Enterprise Information Portals. New York: Merrill Lynch, 1998.

[9] Rahm E, Do HH. Data cleaning: Problems and current approaches. IEEE Data Engineering Bulletin, 2000,23(4):3−13.

[10] Wang RY, Strong DM. Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems, 1996,12(4):5−34. [doi: 10.1080/07421222.1996.11518099]

[11] Sidi F, Hassany P, Panahy S, Affendey LS, Jabar MA, Ibrahim H, Mustapha A. Data quality: A survey of data quality dimensions. Faculty of Computer Science and Information Technology, University Putra Malaysia. 2012. [doi: 10.1109/InfRKM.2012.6204995]

[12] Li JZ, Liu XM. An important aspect of big data: Data usability. Journal of Computer Research and Development, 2013,50(6): 1147−1162 (in Chinese with English abstract).

[13] Guo ZM, Zhou AY. Research on data quality and data cleaning: A survey. Ruan Jian Xue Bao/Journal of Software, 2002,13(11): 2076−2082 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/20021103.htm

[14] Batini C, Cappiello C, Francalanci C, Maurino A. Methodologies for data quality assessment and improvement. ACM Computing Surveys, 2009,41(3):75−79. [doi: 10.1145/1541880.1541883]

[15] Bohannon P, Fan WF, Geerts F, Jia X. Conditional functional dependencies for data cleaning. In: Proc. of the ICDE. Piscataway, 2007. 746−755. [doi: 10.1109/ICDE.2007.367920]

[16] Fan WF, Geerts F, Lakshmanan LVS, Xiong M. Discovering conditional functional dependencies. IEEE Trans. on Knowledge and Data Engineering, 2011,23(5):683−698. [doi: 10.1109/TKDE.2010.154]

[17] Bravo L, Fan WF, Ma S. Extending dependencies with conditions. In: Proc. of the VLDB. 2007. 243−254.

[18] Bravo L, Fan WF, Geerts F, Ma S. Increasing the expressivity of conditional functional dependencies without extra complexity. In: Proc. of the ICDE. Piscataway, 2008. 516−525. [doi: 10.1109/ICDE.2008.4497460]

[19] Fan WF, Ma S, Hu Y, Liu J, Wu Y. Propagating functional dependencies with conditions. In: Proc. of the VLDB. 2008. 391−407. [doi: 10.14778/1453856.1453901]

[20] Liu XM, Li JZ. Discovering extended conditional functional dependencies. Journal of Computer Research and Development, 2015,52(1):130−140 (in Chinese with English abstract).

[21] Sun JZ, Li JZ. Micro functionial depenency and reasonning. Chinese Journal of Computers, To Appear (in Chinese with English abstract).

[22] Miao DJ, Liu XM, Li JZ. An algorithm on mining approximate functional dependencies in probabilistic database. Journal of Computer Research and Development, 2015,52(12):2857−2865 (in Chinese with English abstract).

[23] Golab L, Karloff H, Korn F, Saha A, Srivastava D. Sequential dependencies. VLDB, 2009,2(1):574−585. [doi: 10.14778/1687627.1687693]

[24] Koudas N, Saha A, Srivastava D, Venkatasubramanian S. Metric functional dependencies. In: Proc. of the ICDE. Piscataway, 2009. 1275−1278. [doi: 10.1109/ICDE.2009.219]

[25] Korn F, Muthukrishnan S, Zhu Y. Checks and balances: Monitoring data quality problems in network traffic databases. In: Proc. of the VLDB. San Francisco: Morgan Kaufmann Publishers, 2003. 536−547.

[26] Xiong H, Pandey G, Steinbach M, Kumar V. Enhancing data analysis with noise removal. IEEE Trans. on Knowledge and Data Engineering, 2006,18(3):304−319. [doi: 10.1109/TKDE.2006.46]

[27] Fan WF, Geerts F. Relative information completeness. In: Proc. of the ACM SIGMOD-SIGACT-SIGART Symp. on Principles of Database Systems. New York: ACM Press, 2009. 97−106. [doi: 10.1145/1559795.1559811]

[28] Ma S, Fan WF, Bravo L. Extending inclusion dependencies with conditions. Theoretical Computer Science, 2014,515:64−95. [doi: 10.1016/j.tcs.2013.11.002]

[29] Fan WF, Geerts F. Capturing missing tuples and missing values. In: Proc. of the ACM SIGMOD-SIGACT-SIGART Symp. on Principles of Database Systems. New York: ACM Press, 2010. 169−178. [doi: 10.1145/1807085.1807109]

[30] Abiteboul S, Segoufin L, Vianu V. Representing and querying XML with incomplete information. ACM Trans. on Database Systems, 2006,31(1):208−254. [doi: 10.1145/1132863.1132869]

[31] Barceló P, Libkin L, Poggi A, Sirangelo C. XML with incomplete information. Journal of the ACM, 2010,58(1):4. [doi: 10.1145/1870103.1870107]

[32] Fan WF, Geerts F, Wijsen J. Determining the currency of data. ACM Trans. on Database Systems, 2012,37(4):25. [doi: 10.1145/2389241.2389244]

[33] Li MH, Li JZ, Cheng SY. Uncertain rule based method for evaluating data currency. Ruan Jian Xue Bao/Journal of Software, 2014,25(S2):147−156 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/14033.htm

[34] Li LL, Li JZ. Rule-Based method for entity resolution. IEEE Trans. on Knowledge and Data Engineering, 2015,27(1):250−263. [doi: 10.1109/TKDE.2014.2320713]

[35] Fan WF, Jia XB, Li JZ, Ma S. Reasoning about record matching rules. In: Proc. of the VLDB. 2009. 407−418. [doi: 10.14778/1687627.1687674]

[36] Fan WF, Gao H, Jia XB, Li JZ, Ma S. Dynamic constraints for record matching. VLDB, 2011,20(4):495−520. [doi: 10.1007/s00778-010-0206-6]

[37] Whang SE, Benjelloun O, Garcia-Molina H. Generic entity resolution with negative rules. VLDB, 2009,18(6):1261−1277. [doi: 10.1007/s00778-009-0136-3]

[38] Chaudhuri S, Das Sarma A, Ganti V, Kaushik R. Leveraging aggregate constraints for deduplication. In: Proc. of the SIGMOD. New York: ACM Press, 2007. 437−448. [doi: 10.1145/1247480.1247530]

[39] Shen W, Li X, Doan A. Constraint-Based entity matching. In: Proc. of the National Conf. on Artificial Intelligence. Menlo Park: AAAI Press, 2005. 862−867.

[40] Cao Y, Fan WF, Yu WY. Determining the relative accuracy of attributes. In: Proc. of the SIGMOD. 2013. 565−576. [doi: 10.1145/2463676.2465309]

[41] Cheng R, Chen J, Xie X. Cleaning uncertain data with quality guarantees. In: Proc. of the VLDB. 2008. 722−735. [doi: 10.14778/1453856.1453935]

[42] Miao DJ, Li JZ, Liu X. On complexity of sampling query feedback restricted database repair of functional dependency violations. Theoretical Computer Science, 2016,609:594−605. [doi: 10.1016/j.tcs.2015.02.010]

[43] Miao DJ, Li JZ, Liu XM, Gao H. Vertex cover in conflict graphs: Complexity and a near optimal approximation. In: Proc. of the 9th Annual Int'l Conf. on Combinatorial Optimization and Applications. 2015. [doi: 10.1007/978-3-319-26626-8_29]

[44] Decanio SJ. Estimating the confidence of conditional functional dependencies. In: Proc. of the SIGMOD. New York, 2009. [doi: 10.1145/1559845.1559895]

[45] Görz Q. An economics-driven decision model for data quality improvement: Acontribution to data currency. In: Proc. of the AMCIS. Atlanta: AIS, 2011. 1−8.

[46] Heinrich B, Klier M. Assessing data currency: A probabilistic approach. Journalof Information Science, 2011,37(1):86−100. [doi: 10.1177/0165551510392653]

[47] Heinrich B, Klier M, Kaiser M. A procedure to develop metrics for currency andits application in CRM. Journal of Data and Information Quality, 2009,1(1):5. [doi: 10.1145/1515693.1515697]

[48] Cappiello C, Francalanci C, Pernici B. A model of data currency in multi-ChannelFinancial architectures. In: Proc. of the 7th Int'l Conf. on Information Quality. 2002. 106−118.

[49] Cappiello C, Francalanci C, Pernici B. Time related factors of data accuracy, completeness, and currency in multi-channel information systems. In: Proc. of the Forum for Short Contributions at the 15th Conf. on Advanced Information System Engineering. Berlin: Springer-Verlag, 2003. 1−11.

[50] Heinrich B, Hristova D. A fuzzy metric for currency in the context of big data. In: Proc. of the 22nd European Conf. on Information Systems. Atlanta: AIS, 2014. 1−15.

[51] Li MH, Li JZ, Gao H. Evaluation of data currency. Chinese Journal of Computers, 2012,35(11):2348−2360 (in Chinese with English abstract).

[52] Liu YN, Zou ZN, Li JZ. Evaluation of data completeness. Journal of Computer Research and Development, 2013,50(S1):230−238 (in Chinese with English abstract).

[53] Liu YN, Li JZ, Zou ZN. Determining the completeness of data. Journal of Computer Science and Technology, to appear.

[54] Emran NA. Data completeness measures, pattern analysis, intelligent security and the Internet of Things. In: Proc. of the Springer Int'l Publishing. 2015. 117−130. [doi: 10.1007/978-3-319-17398-6]

[55] Razniewski S, Nutt W. Assessing the completeness of geographical data. In: Proc. of the Big Data. Berlin, Heidelberg: Springer-Verlag, 2013. 228−237. [doi: 10.1007/978-3-642-39467-6_21]

[56] Endler G, Baumgärtel P, Wahl AM, Lenz R. ForCE: Is estimation of data completeness through time series forecasts feasible. In: Proc. of the Advances in Databases and Information Systems. Springer Int'l Publishing, 2015. 261−274. [doi: 10.1007/978-3-319-23135-8_18]

[57] Emran NA, Embury S, Missier P, Isa MNM, Muda AK. Measuring data completeness for microbial genomics database. In: Proc. of the Intelligent Information and Database Systems. Berlin, Heidelberg: Springer-Verlag, 2013. 186−195. [doi: 10.1007/978-3-642-36546-1_20]

[58] Emran NA, Embury S, Missier P. Measuring population-based completeness for single nucleotide polymorphism (SNP) databases. In: Proc. of the Advanced Approaches to Intelligent Information and Database Systems. Springer Int'l Publishing, 2014. 173−182. [doi: 10.1007/978-3-319-05503-9_17]

[59] Zhang Y, Wang H, Gao H, Li JZ. Efficient accuracy evaluation for multi-modal sensed data. Journal of Combinatorial Optimization. [doi: 10.1007/s10878-015-9920-8]

[60] Zhang Y, Wang HZ, Yang ZS, Li JZ. Relative accuracy evaluation. PLoS ONE, 2014,9(8):e103853−e103853. [doi: 10.1371/journal.pone.0103853]

[61] Li LL, Li JZ, Gao H. Evaluating entity-description conflict on duplicated data. Journal of Combinatorial Optimization, 2016,31(2): 918−941. [doi: 10.1007/s10878-014-9801-6]

[62] Chen W, Fan W, Ma S. Analyses and validation of conditional dependencies with built-in predicates. In: Proc. of the DEXA. Berlin, Heidelberg: Springer-Verlag, 2009. 576−591. [doi: 10.1007/978-3-642-03573-9_48]

[63] Fan WF, Geerts F, Jia XB, Kementsietsidis A. Conditional functional dependencies for capturing data inconsistencies. ACM Trans. on Database Systems, 2008,33(2):6. [doi: 10.1145/1366102.1366103]

[64] Fan WF, Geerts F, Ma S, Muller H. Detecting inconsistencies in distributed data. In: Proc. of the ICDE. Piscataway, 2010. 64−75. [doi: 10.1109/ICDE.2010.5447855]

[65] Fan WF, Li JZ, Tang N, Yu W. Incremental detection of inconsistencies in distributed data. IEEE Trans. on Knowledge and Data Engineering, 2014,26(6):1367−1383. [doi: 10.1109/TKDE.2012.138]

[66] Fan WF, Li JZ, Ma S, Tang N, Yu WY. Towards certain fixes with editing rules and master data. VLDB, 2012,21(2): 213−238. [doi: 10.1007/s00778-011-0253-7]

[67] Miao DJ, Li JZ, Liu X. On complexity of sampling query feedback restricted database repair of functional dependency violations. Theoretical Computer Science, 2016,609:594−605. [doi: 10.1016/j.tcs.2015.02.010]

[68] Beskales G, Ilyas IF, Golab L, Galiullin A. On the relative trust between inconsistent data and inaccurate constraints. In: Proc. of the ICDE. 2013. 541−552. [doi: 10.1109/ICDE.2013.6544854]

[69] Geerts F, Mecca G, Papotti P, Santoro D. The LLUNATIC data-cleaning framework. In: Proc. of the VLDB. 2013. 625−636.

[70] Zhang AZ, Men XY, Wang HZ, Li JZ, Gao H. Hadoop-Based inconsistence detection and reparation algorithm for big data. Journal of Frontiers of Computer Science & Technology, 2015,9(9):1044−1055 (in Chinese with English abstract).

[71] Papenbrock T, Kruse S, Quiané-Ruiz JA, Naumann F. Divide & conquer-based inclusion dependency discovery. In: Proc. of the VLDB. 2015. 774−785. [doi: 10.14778/2752939.2752946]

[72] Chu X, Ilyas IF, Papotti P, Ye Y. RuleMiner: Data quality rules discovery. In: Proc. of the ICDE. 2014. 1222−1225.

[73] Song SX, Chen L, Cheng H. On concise set of relative candidate keys. In: Proc. of the VLDB. 2014. 1179−1190. [doi: 10.14778/2732977.2732991]

[74] Galárraga LA, Teflioudi C, Hose K, Suchanek F. Amie: Association rule mining under incomplete evidence in ontological knowledge bases. In: Proc. of the WWW. 2013. 413−422.

[75] Abedjan Z, Schulze P, Naumann F. DFD: Efficient functional dependency discovery. In: Proc. of the CIKM. 2014. 949−958. [doi: 10.1145/2661829.2661884]

[76] Combi C, Parise P, Sala P, Pozzi G. Mining approximate temporal functional dependencies based on pure temporal grouping. In: Proc. of the ICDMW. 2013. 258−265. [doi: 10.1109/ICDMW.2013.100]

[77] Fan WF, Geerts F, Tang N, Yu WY. Conflict resolution with data currency and consistency. Journal of Data and Information Quality, 2014,5(1-2):6. [doi: 10.1145/2631923]

[78] Abedjan Z, Akcora CG, Ouzzani M, Papotti P, Stonebraker M. Temporal rules discovery for Web data cleaning. In: Proc. of the VLDB. 2016. 336−347. [doi: 10.14778/2856318.2856328]

[79] Li MH, Li JZ. A minimized-rule based approach for improving data currency. Journal of Combinatorial Optimization, 2015. 1−30. [doi: 10.1007/s10878-015-9904-8]

[80] Li MH, Li JZ. Algorithms for improving data currency. Journal of Computer Research and Development, 2015,52(9):1992−2001 (in Chinese with English abstract).

[81] Libkin L. Incomplete data: What went wrong, and how to fix it. In: Proc. of the PODS. 2014. 1−13. [doi: 10.1145/2594538.2594561]

[82] Liu H, Zhang S. Noisy data elimination using mutual *k*-nearest neighbor for classification mining. Journal of Systems & Software, 2012,85(5):1067−1074. [doi: 10.1016/j.jss.2011.12.019]

[83] Tian J, Yu B, Yu D, Ma S. Missing data analysis: A hybrid multiple imputation algorithm using gray system theory and entropy based on clustering, Applied Intelligence, 2013,40:376−388. [doi: 10.1007/s10489-013-0469-x]

[84] Van Buuren S. Flexible Imputation of Missing Data. Boca Raton: CRC Press, 2012.

[85] Zhang S. Shell-Neighbor method and its application in missing data imputation. Applied Intelligence, 2011,35(1):123−133. [doi: 10.1007/s10489-009-0207-6]

[86] Zhang S. Nearest neighbor selection for iteratively *k*NN imputation. Journal of Systems & Software, 2012,85(11):2541−2552. [doi: 10.1016/j.jss.2012.05.073]

[87] Zhang S, Jin Z, Zhu X. Missing data imputation by utilizing information within incomplete instances. Journal of Systems & Software, 2012,84(3):452−459. [doi: 10.1016/j.jss.2010.11.887]

[88] Zhu X, Zhang S, Jin Z, Zhang Z, Xu Z. Missing value estimation for mixed-attribute data sets. IEEE Trans. on Knowledge & Data Engineering, 2011,23(1):110−121. [doi: 10.1109/TKDE.2010.99]

[89] Song S, Zhang A, Chen L, Wang J. Enriching data imputation with extensive similarity neighbors. In: Proc. of the VLDB. 2015. 1286−1297. [doi: 10.14778/2809974.2809989]

[90] Wu S, Feng X, Han Y, Wang Q. Missing categorical data imputation approach based on similarity. In: Proc. of the IEEE Int'l Conf. on Systems, Man, and Cybernetics (SMC). 2012. 2827−2832. [doi: 10.1109/ICSMC.2012.6378177]

[91] Gummadi R, Khulbe A, Kalavagattu A, Salvi S, Kambhampati S. SMARTINT: Using mined attribute dependencies to integrate fragmented Web databases. Journal of Intelligent Information Systems, 2012,38:575−599. [doi: 10.1007/s10844-011-0169-0]

[92] Koutrika G. Entity reconstruction: Putting the pieces of the puzzle back together. Technical Report, Palo Alto: HP Labs, 2012.

[93] Yakout M, Ganjam K, Chakrabarti K, Chaudhuri S. InfoGather: Entity augmentation and attribute discovery by holistic matching with Web tables. In: Proc. of the SIGMOD. 2012. 97−108. [doi: 10.1145/2213836.2213848]

[94] Li Z, Qin L, Cheng H, Zhang X, Zhou X. TRIP: An interactive retrieving-inferring data imputation approach. IEEE Trans. on Knowledge and Data Engineering, 2015,27(9):2550−2563. [doi: 10.1109/TKDE.2015.2411276]

[95] Li ZX, Shang S, Xie Q, Zhang XL. Cost reduction for Web-based data imputation. In: Proc. of the Database Systems for Advanced Applications. Springer Int'l Publishing, 2014. 438−452. [doi: 10.1007/978-3-319-05813-9_29]

[96] Ye C, Wang HZ, Li JZ, Gao H, Cheng SY. Crowdsourcing-Enhanced missing values imputation based on Bayesian network. In: Proc. of the DASFAA. 2016. 67−81. [doi: 10.1007/978-3-319-32025-0_5]

[97] Korn F, Saha B, Srivastava D, Ying SS. On repairing structural problems in semi-structured data. In: Proc. of the VLDB. 2013. 601−612. [doi: 10.14778/2536360.2536361]

[98] Wang J, Song S, Zhu X, Lin X. Efficient recovery of missing events. In: Proc. of the VLDB. 2013. 841−852. [doi: 10.14778/2536206.2536212]

[99] Wang S, Xiao X, Lee CH. Crowd-Based deduplication: An adaptive approach. In: Proc. of the SIGMOD. 2015. 1263−1277. [doi: 10.1145/2723372.2723739]

[100] Gokhale C, Das S, Doan A, Naughton JF, Rampalli N, Shavlik JW, Zhu X. Corleone: Hands-Off crowdsourcing for entity matching. In: Proc. of the SIGMOD. 2014. 601−612. [doi: 10.1145/2588555.2588576]

[101] Verroios V, Garcia-Molina H. Entity resolution with crowd errors. In: Proc. of the ICDE. 2015. 219−230. [doi: 10.1109/ICDE.2015.7113286]

[102] Vesdapunt N, Bellare K, Dalvi NN. Crowdsourcing algorithms for entity resolution. In: Proc. of the VLDB. 2014. 1071−1082. [doi: 10.14778/2732977.2732982]

[103] Whang SE, Lofgren P, Garcia-Molina H. Question selection for crowd entity resolution. In: Proc. of the VLDB. 2013. 349−360. [doi: 10.14778/2536336.2536337]

[104] Hua W, Zheng K, Zhou XF. Microblog entity linking with social temporal context. In: Proc. of the SIGMOD. 2015. 1761−1775. [doi: 10.1145/2723372.2751522]

[105] Shen W, Han JW, Wang JY. A probabilistic model for linking named entities in Web text with heterogeneous information networks. In: Proc. of the SIGMOD. 2014. 1199−1210. [doi: 10.1145/2588555.2593676]

[106] Zhu X, Song S, Lian X, Wang J, Zou L. Matching heterogeneous event data. In: Proc. of the SIGMOD. 2014. 1211−1222. [doi: 10.1145/2588555.2588570]

[107] Chiang YH, Doan AH, Naughton JF. Modeling entity evolution for temporal record matching. In: Proc. of the SIGMOD. 2014. 1175−1186. [doi: 10.1145/2588555.2588560]

[108] Whang SE, Garcia-Molina H. Incremental entity resolution on rules and data. VLDB, 2014,23(1):77−102. [doi: 10.1007/s00778-013-0315-0]

[109] Gruenheid A, Dong XL, Srivastava D. Incremental record linkage. In: Proc. of the VLDB. 2014. 697−708. [doi: 10.14778/2732939.2732943]

[110] Wildani A, Miller EL, Rodeh O. HANDS: A heuristically arranged non-backup in-line deduplication system. In: Proc. of the ICDE. 2013. 446−457. [doi: 10.1109/ICDE.2013.6544846]

[111] Li X, Dong XL, Lyons KB, Meng W, Srivastava D. Scaling up copy detection. In: Proc. of the ICDE. 2015. [doi: 10.1109/ICDE.2015.7113275]

[112] Whang SE, Marmaros D, Garcia-Molina H. Pay-as-You-Go entity resolution. IEEE Trans. on Knowledge and Data Engineering, 2013,25(5):1111−1124. [doi: 10.1109/TKDE.2012.43]

[113] Li LL, Li JZ, Wang HZ, Gao H. Context-Based entity description rule for entity resolution. In: Proc. of the CIKM. 2011. 1725−1730. [doi: 10.1145/2063576.2063825]

[114] Li LL, Li JZ, Gao H. Rule-Based method for entity resolution. IEEE Trans. on Knowledge and Data Engineering, 2015,27(1):250−263. [doi: 10.1109/TKDE.2014.2320713]

[115] Wang FD, Wang HZ, Li JZ, Gao H. Graph-Based reference table construction to facilitate entity matching. Journal of Systems and Software, 2013,86(6):1679−1688. [doi: 10.1016/j.jss.2013.02.026]

[116] Altowim Y, Kalashnikov DV, Mehrotra S. Progressive approach to relational entity resolution. In: Proc. of the VLDB. 2014. 999−1010. [doi: 10.14778/2732967.2732975]

[117] Altwaijry H, Kalashnikov DV, Mehrotra S. Query-Driven approach to entity resolution. In: Proc. of the VLDB. 2013. 1846−1857. [doi: 10.14778/2556549.2556567]

[118] Wang HZ, Li JZ, Gao H. Efficient entity resolution based on subgraph cohesion. Knowledge Information Systems, 2016,46(2):285−314. [doi: 10.1007/s10115-015-0818-7]

[119] Li Q, Li YL, Gao J, Su L, Zhao B, Demirbas M, Fan W, Han JW. A confidence-aware approach for truth discovery on long-tail data. In: Proc. of the VLDB. 2015. 425−436. [doi: 10.14778/2735496.2735505]

[120] Prokoshyna N, Szlichta J, Chiang F, Miller RJ, Srivastava D. Combining quantitative and logical data cleaning. In: Proc. of the VLDB. 2016. 300−311. [doi: 10.14778/2856318.2856325]

[121] Zhao Z, Cheng J, Ng W. Truth discovery in data streams: A single-pass probabilistic approach. In: Proc. of the CIKM. 2014. 1589−1598. [doi: 10.1145/2661829.2661892]

[122] Interlandi M, Tang N. Proof positive and negative in data cleaning. In: Proc. of the ICDE. 2015. 18−29. [doi: 10.1109/ICDE.2015.7113269]

[123] Ding XO, Wang HZ, Zhang XY, Li JZ, Gao H. Association relationships study of multi-dimensional data quality. Ruan Jian Xue Bao/Journal of Software, 2016,27(7):1626−1644 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5040.htm [doi: 10.13328/j.cnki.jos.005040]

[124] Fan W, Geerts F, Tang N, Yu W. Inferring data currency and consistency for conflict resolution. In: Proc. of the ICDE. 2013. 470−481. [doi: 10.1109/ICDE.2013.6544848]

[125] Yang DH, Li NN, Wang HZ, Li JZ, Gao H. The optimization of the big data cleaning based on task merging. Chinese Journal of Computers, 2015,39(1):97−108 (in Chinese with English abstract).

[126] Wang X, Dong XL, Meliou A. Data X-ray: A diagnostic tool for data errors. In: Proc. of the SIGMOD. 2015. 1231−1245. [doi: 10.1145/2723372.2750549]

[127] Wang XL, Feng M, Wang Y, Dong XL, Meliou A. Error diagnosis and data profiling with data X-ray. In: Proc. of the VLDB. 2015. 1984−1995. [doi: 10.14778/2824032.2824117]

[128] Prokoshyna N, Szlichta J, Chiang F, Miller RJ, Srivastava D. Combining quantitative and logical data cleaning. In: Proc. of the VLDB. 2016. 300−311. [doi: 10.14778/2856318.2856325]

[129] Geerts F, Mecca G, Papotti P, Santoro D. Mapping and cleaning. In: Proc. of the ICDE. 2014. 232−243. [doi: 10.1109/ICDE.2014.6816654]

[130] Chu X, Ilyas IF, Papotti P. Holistic data cleaning: Putting violations into context. In: Proc. of the ICDE. 2013. 458−469. [doi: 10.1109/ICDE.2013.6544847]

[131] Li ZY, Wang HZ, Shao W, Li JZ, Gao H. Repairing data through regular expressions. PVLDB, 2016,9(5):432−443. [doi: 10.14778/2876473.2876478]

[132] Zhang CJ, Chen L, Tong Y, Liu Z. Cleaning uncertain data with a noisy crowd. In: Proc. of the ICDE. 2015. 6−17. [doi: 10.1109/ICDE.2015.7113268]

[133] Wang J, Song S, Lin X, Zhu X, Pei J. Cleaning structured event logs: A graph repair approach. In: Proc. of the ICDE. 2015. 30−41. [doi: 10.1109/ICDE.2015.7113270]

[134] Volkovs M, Chiang F, Szlichta J, Miller RJ. Continuous data cleaning. In: Proc. of the ICDE. 2014. 244−255. [doi: 10.1109/ICDE.2014.6816655]

[135] Fan WF, Li JZ, Tang N, Yu WY. Incremental detection of inconsistencies in distributed data. IEEE Trans. on Knowledge and Data Engineering, 2014,26(6):1367−1383. [doi: 10.1109/TKDE.2012.138]

[136] Yakout M, Berti-Equille L, Elmagarmid AK. Don't be SCAREd: Use SCalable automatic REpairing with maximal likelihood and bounded changes. In: Proc. of the SIGMOD. 2013. 553−564. [doi: 10.1145/2463676.2463706]

[137] Dong XL, Gabrilovich E, Murphy K, Dang V, Horn W, Lugaresi C, Sun S, Zhang W. Knowledge-Based trust: Estimating the trustworthiness of Web sources. In: Proc. of the VLDB. 2015. 938−949. [doi: 10.14778/2777598.2777603]

[138] Rekatsinas T, Dong XL, Srivastava D. Characterizing and selecting fresh data sources. In: Proc. of the SIGMOD. 2014. 919−930. [doi: 10.1145/2588555.2610504]

[139] Pochampally R, Sarma AD, Dong XL, Meliou A, Srivastava D. Fusing data with correlations. In: Proc. of the SIGMOD. 2014. 433−444. [doi: 10.1145/2588555.2593674]

[140] Li Q, Li YL, Gao J, Zhao B, Fan W, Han JW. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In: Proc. of the SIGMOD. 2014. 1187−1198. [doi: 10.1145/2588555.2610509]

[141] Chalamalla A, Ilyas IF, Ouzzani M, Papotti P. Descriptive and prescriptive data cleaning. In: Proc. of the SIGMOD. 2014. 445−456. [doi: 10.1145/2588555.2610520]

[142] Dong XL, Berti-Equille L, Srivastava D. Integrating conflicting data: The role of source dependence. In: Proc. of the VLDB. 2009. 145. [doi: 10.14778/1687627.1687690]

[143] Dong XL, Berti-Equille L, Srivastava D. Truth discovery and copying detection in a dynamic world. In: Proc. of the VLDB. 2009. 146. [doi: 10.14778/1687627.1687691]

[144] Dong XL, Berti-Equille L, Hu YF, Srivastava D. Global detection of complex copying relationships between sources. In: Proc. of the VLDB. 2010. 1358−1369. [doi: 10.14778/1920841.1921008]

[145] Dong XL. Solomon: Seeking the truth via copying detection. In: Proc. of the VLDB. 2010. 1358−1369. [doi: 10.1145/1966883.1966887]

[146] Dong XL, Naumann F. Data fusion: Resolving data conflicts for integration. In: Proc. of the VLDB. 2009. 1654−1655. [doi: 10.14778/1687553.1687620]

[147] Cheng SY, Li JZ. Sampling based (ε,δ)-approximate aggregation algorithm in sensor networks. In: Proc. of the IEEE ICDCS 2009. Piscataway, 2009. 273−280. [doi: 10.1109/ICDCS.2009.8]

[148] Li JZ, Cheng SY. (ε,δ)-Approximate aggregation algorithms in dynamic sensor networks. IEEE Trans. on Parallel and Distributed Systems, 2012,23(3):385−396. [doi: 10.1109/TPDS.2011.193]

[149] Cheng SY, Li JZ, Cai ZP. ε-Approximation to physical world by sensor networks. In: Proc. of the INFOCOM. Piscataway, 2013. 3184−3192. [doi: 10.1109/INFCOM.2013.6567121]

[150] Li JZ, Li GH, Gao H. Novel ε-approximation to data streams in sensor networks. IEEE Trans. on Parallel Distrib. System, 2015, 26(6):1654−1667. [doi: 10.1109/TPDS.2014.2323056]

[151] Cheng SY, Li JZ, Liu Y. Location aware peak value queries in sensor networks. In: Proc. of the INFOCOM. Piscataway, 2012. 486−494. [doi: 10.1109/INFCOM.2012.6195789]

[152] Gao J, Li JZ. Composite event coverage in wireless sensor networks with heterogeneous sensors. In: Proc. of the INFOCOM. 2015. 217−225. [doi: 10.1109/INFOCOM.2015.7218385]

[153] Li JZ, Cheng SY, Gao H, Cai ZP. Approximate physical world reconstruction algorithms in sensor networks. IEEE Trans. on Parallel and Distributed Systems, 2014,25(12):3099−3110. [doi: 10.1109/TPDS.2013.2297121]

[154] Cheng SY, Cai ZP, Li JZ, Fang XL. Drawing dominant dataset from big sensory data in wireless sensor networks. In: Proc. of the INFOCOM. 2015. 531−539. [doi: 10.1109/INFOCOM.2015.7218420]

[155] Data collection in multi-application sharing wireless sensor networks. IEEE Trans. on Parallel and Distributed Systems, 2015,26(2): 403−412. [doi: 10.1109/TPDS.2013.289]

[156] Li JZ, Yu L, Gao H, Xiong SG. Grouping-Enhanced resilient probabilistic en-route filtering of injected false data in WSNs. IEEE Trans. on Parallel and Distributed Systems, 2012,23(5):881−889. [doi: 10.1109/TPDS.2011.217]

[157] Yu L, Li JZ, Cheng SY, Xiong SG, Shen HY. Secure continuous aggregation via sampling-based verification in wireless sensor networks. IEEE Trans. on Parallel and Distributed Systems, 2014,25(3):762−744. [doi: 10.1109/TPDS.2013.63]

[158] Altwaijry H, Mehrotra S, Kalashnikov DV. QuERy: A framework for integrating entity resolution with query processing. In: Proc. of the VLDB. 2015. 120−131. [doi: 10.14778/2850583.2850587]

[159] Rezig EK, Dragut EC, Ouzzani M, Elmagarmid AK. Query-Time record linkage and fusion over Web databases. In: Proc. of the ICDE. 2015. 42−53. [doi: 10.1109/ICDE.2015.7113271]

[160] Liu XL, Wang HZ, Li JZ, Gao H. Similarity join algorithm based on entity. Ruan Jian Xue Bao/Journal of Software, 2015,26(6): 1421−1437 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4610.htm [doi: 10.13328/j.cnki.jos.004610]

[161] Razniewski S, Korn F, Nutt W, Srivastava D. Identifying the extent of completeness of query answers over partially complete databases. In: Proc. of the SIGMOD. 2015. 561−576. [doi: 10.1145/2723372.2750544]

[162] Savkovic O, Mirza P, Tomasi A, Nutt W. Complete approximations of incomplete queries. In: Proc. of the VLDB. 2013. 1378−1381. [doi: 10.14778/2536274.2536320]

[163] Bharuka R, Kumar PS. Finding skylines for incomplete data. In: Proc. of the 24th Australasian Database Conf., Vol.137. Australian Computer Society, Inc., 2013. 109−117.

[164] Lofi C, El Maarry K, Balke WT. Skyline queries over incomplete data-error models for focused crowd-sourcing. In: Proc. of the Conceptual Modeling. Berlin, Heidelberg: Springer-Verlag, 2013. 298−312. [doi: 10.1007/978-3-642-41924-9_25]

[165] Lofi C, El Maarry K, Balke WT. Skyline queries in crowd-enabled databases. In: Proc. of the 16th Int'l Conf. on Extending Database Technology. ACM Press, 2013. 465−476. [doi: 10.1145/2452376.2452431]

[166] Miao X, Gao Y, Chen L, Chen G, Li Q, Jiang T. On efficient *k*-skyband query processing over incomplete data. In: Proc. of the Database Systems for Advanced Applications. Berlin, Heidelberg: Springer-Verlag, 2013. 424−439. [doi: 10.1007/978-3-642-37487-6_32]

[167] Gao Y, Miao X, Cui H, Chen G, Li Q. Processingk-Skyband, constrained skyline, and group-by skyline queries on incomplete data. Expert Systems with Applications, 2014,41(10):4959−4974. [doi: 10.1016/j.eswa.2014.02.033]

[168] Arefin MS, Morimoto Y. Skyline sets queries from databases with missing values. In: Proc. of the 22nd Int'l Conf. on Computer Theory and Applications. IEEE, 2012. 24−29. [doi: 10.1109/ICCTA.2012.6523542]

[169] Markus E, Patrick R, Florian W, Alfons H, Werner K. Handling of null values in preference database queries. In: Proc. of the 6th Multidisciplinary Workshop on Advances in Preference Handling.

[170] Kolaitis PG, Pema E, Tan WC. Efficient querying of inconsistent databases with binary integer programming. In: Proc. of the VLDB. 2013. 397−408. [doi: 10.14778/2536336.2536341]

[171] Bertossi LE, Kolahi S, Lakshmanan LVS. Data cleaning and query answering with matching dependencies and matching functions. In: Proc. of the ICDT. 2011. 268−279. [doi: 10.1145/1938551.1938585]

[172] Wang J, Krishnan S, Franklin MJ, Goldberg K, Kraska T, Milo T. A sample-and-clean framework for fast and accurate query processing on dirty data. In: Proc. of the SIGMOD. 2014. 469−480. [doi: 10.1145/2588555.2610505]

[173] Xu C, Xia F, Sharaf MA, Zhou MQ, Zhou AY. AQUAS: A quality-aware scheduler for NoSQL data stores. In: Proc. of the ICDE. 2014. 1210−1213. [doi: 10.1109/ICDE.2014.6816743]

[174] Chen YC, Li JZ, Luo JZ. ITCI: An information theory based classification algorithm for incomplete data. In: Proc. of the WAIM. 2014. 167−179. [doi: 10.1007/978-3-319-08010-9_19]

[175] Liu XL, Li JZ. Consistent estimation of query result in inconsistent data. Chinese Journal of Computers, 2015,38(9):1727−1738 (in Chinese with English abstract).

[176] Razniewski S, Nutt W. Completeness of queries over incomplete databases. In: Proc. of the VLDB. 2011. 749−760.

[177] Savković O, Paramita M, Paramonov S, Paramonov S, Nutt W. MAGIK: Managing completeness of data. In: Proc. of the 21st ACM Int'l Conf. on Information and Knowledge Management. ACM Press, 2012. 2725−2727. [doi: 10.1145/2396761.2398741]

[178] Savkovic O, Mirza P, Tomasi A, Nutt W. Complete approximations of incomplete queries. In: Proc. of the VLDB. 2013. 1378−1381. [doi: 10.14778/2536274.2536320]

[179] Nutt W, Razniewski S. Completeness of queries over SQL databases. In: Proc. of the 21st ACM Int'l Conf. on Information and Knowledge Management. ACM Press, 2012. 902−911. [doi: 10.1145/2396761.2396875]

[180] Nutt W, Razniewski S, Vegliach G. Incomplete databases: Missing records and missing values. In: Proc. of the Database Systems for Advanced Applications. Berlin, Heidelberg: Springer-Verlag, 2012. 298−310. [doi: 10.1007/978-3-642-29023-7_30]

[181] Darari F, Nutt W, Pirrò G, Razniewski S. Completeness statements about RDF data sources and their use for query answering. In: Proc. of the Semantic Web (ISWC 2013). Berlin, Heidelberg: Springer-Verlag, 2013. 66−83. [doi: 10.1007/978-3-642-41335-3_5]

[182] Darari F, Prasojo RE, Nutt W. CORNER: A completeness reasoner for SPARQL queries over RDF data sources. In: Proc. of the Semantic Web: ESWC 2014 Satellite Events. Springer Int'l Publishing, 2014. 310−314. [doi: 10.1007/978-3-319-11955-7_40]

[183] Paramonov S. Query completeness—A logic programming approach. Technical Report, KRDB13-2, KRDB Research Center, Free University Bozen-Bolzano, 2013. http://www.inf.unibz.it/krdb/pub/tech-rep.php

[184] Nutt W, Paramonov S, Savkovic O. An ASP approach to query completeness reasoning. Theory and Practice of Logic Programming, 2013,13(4-5):1−10.

[185] Nutt W, Paramonov S, Savkovic O. Implementing query completeness reasoning. In: Proc. of the 24th ACM Int'l Conf. on Information and Knowledge Management. ACM Press, 2015. 733−742. [doi: 10.1145/2806416.2806439]

[186] Cao Y, Deng T, Fan W, Geerts F. On the data complexity of relative information completeness. Information Systems, 2014,45: 18−34. [doi: 10.1016/j.is.2014.04.001]

[187] Razniewski S, Montali M, Nutt W. Verification of query completeness over processes. In: Proc. of the Business Process Management. Berlin, Heidelberg: Springer-Verlag, 2013. 155−170. [doi: 10.1007/978-3-642-40176-3_13]

[188] Marengo E, Nutt W, Savkovic O. Towards a theory of query stability in business processes. In: Proc. of the 8th Alberto Mendelzon Workshop on Foundations of Data Management. Cartagena de Indias, 2014.

[189] Savkovic O, Marengo E, Nutt W. Query stability in data-aware business processes [Extended Version]. In: Proc. of the CoRR. 2015.

[190] Savkovic O, Marengo E, Nutt W. Query stability in monotonic data-aware business processes. In: Proc. of the ICDT. 2016.

[191] Wang HZ, Li JZ, Huo R, Jia L, Jin L, Meng XY, Xie H. HITCleaner: A light-weight online data cleaning system. DASFAA, 2013,2: 481−484. [doi: 10.1007/978-3-642-37450-0_41]

[192] Ortona S, Orsi G, Buoncristiano M, Furche T. WADaR: Joint wrapper and data repair. In: Proc. of the VLDB. 2015. 1996−2007. [doi: 10.14778/2824032.2824120]

[193] Haas D, Krishnan S, Wang JN, Franklin MJ, Wu E. Wisteria: Nurturing scalable data cleaning infrastructure. In: Proc. of the VLDB. 2015. 2004−2015. [doi: 10.14778/2824032.2824122]

[194] Bergman M, Milo T, Novgorodov S, Tan WC. Query-Oriented data cleaning with oracles. In: Proc. of the SIGMOD. 2015. 1199−1214. [doi: 10.1145/2723372.2737786]

[195] Khayyat Z, Ilyas IF, Jindal A, Madden S, Ouzzani M, Papotti P, Quiané-Ruiz JA, Tang N, Yin S. Big dansing: A system for big data cleansing. In: Proc. of the SIGMOD. 2015. 1215−1230.

[196] Chu X, Morcos J, Ilyas IF, Ouzzani M, Papotti P, Tang N, Ye Y. KATARA, a data cleaning system powered by knowledge bases and crowdsourcing. In: Proc. of the SIGMOD. 2015. 1247−1261. [doi: 10.1145/2723372.2749431]

[197] Elmagarmid AK, Ilyas IF, Ouzzani M, Quiané-Ruiz JA, Tang N, Yin S. NADEEF/ER: Generic and interactive entity resolution. In: Proc. of the SIGMOD. 2014. 1071−1074. [doi: 10.1145/2588555.2594511]

[198] Wang HZ, Li MD, Bu YY, Li JZ, Gao H, Zhang JC. Cleanix: A big data cleaning parfait. In: Proc. of the CIKM. 2014. 2024−2026. [doi: 10.1145/2661829.2661837]

[199] Wang HZ, Li MD, Bu YY, Li JZ, Gao H, Zhang JC. Cleanix: A parallel big data cleaning system. SIGMOD Record, 2015,44(4): 35−40. [doi: 10.1145/2935694.2935702]

[200] Stonebraker M, Bruckner D, Ilyas IF, Beskales G, Cherniack M, Zdonik SB, Pagan A, Xu S. Data curation at scale: The data tamer system. In: Proc. of the CIDR. 2013.

[201] Wang HZ, Zhang XD, Li JZ, Gao H. ProductSeeker: Entity-Based product retrieval for e-commerce. In: Proc. of the SIGIR. 2013. 1085−1086. [doi: 10.1145/2484028.2484205]

[202] Wang HZ, Liu XL, Li JZ, Tong X, Yang L, Li YK. EntityManager: An entity-based dirty data management system. DASFAA, 2013,2:468−471. [doi: 10.1007/978-3-642-37450-0_38]

附中文参考文献:

[12] 李建中,刘显敏.大数据的一个重要方面:数据可用性.计算机研究与发展,2013,50(6):1147−1162.

[13] 郭志懋,周傲英.数据质量和数据清洗研究综述.软件学报,2002,13(11):2076−2082. http://www.jos.org.cn/1000-9825/20021103.htm

[20] 刘显敏,李建中.一种扩展条件函数依赖的发现算法.计算机研究与发展,2015,52(1):130−140.

[21] 孙继洲,李建中,微函数依赖及其推理.计算机学报,录用待发表.

[22] 苗东菁,刘显敏,李建中.概率数据库中近似函数依赖挖掘算法.计算机研究与发展,2015,52(12):2857−2865.

[33] 李默涵,李建中,程思瑶.一种基于不确定规则的数据时效性判定方法.软件学报,2014,25(S2):147−156 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/14033.htm

[51] 李默涵,李建中,高宏.数据时效性判定问题的求解算法.计算机学报,2012,35(11):2348−2360.

[52] 刘永楠,邹兆年,李建中.数据完整性的评估方法.计算机研究与发展,2013,50(S1):230−238.

[70] 张安珍,门雪莹,王宏志,李建中,高宏.大数据上基于 Hadoop 的不一致数据检测与修复算法.计算机科学与探索,2015,9(9): 1044−1055.

[80] 李默涵,李建中.数据时效性修复问题的求解算法.计算机研究与发展,2015,52(9):1992−2001.

[123] 丁小欧,王宏志,张笑影,李建中,高宏.数据质量多种性质的关联关系研究.软件学报,2016,27(7):1626−1644. http://www.jos.org.cn/1000-9825/5040.htm [doi: 10.13328/j.cnki.jos.005040]

[125] 杨东华,李宁宁,王宏志,李建中,高宏.基于任务合并的并行大数据清洗过程优化.计算机学报,2015,39(1):97−108.

[160] 刘雪莉,王宏志,李建中,高宏.基于实体的相似性连接算法.软件学报,2015,26(6):1421−1437. http://www.jos.org.cn/1000-9825/4610.htm [doi: 10.13328/j.cnki.jos.004610]

[175] 刘雪莉,李建中.不一致数据上查询结果的一致性估计.计算机学报,2015,38(9):1727−1738.

李建中(1950－),男,黑龙江哈尔滨人,博士,教授,博士生导师,主要研究领域为海量数据管理与计算,无线传感器网络,数据质量.

高宏(1966－),女,博士,教授,博士生导师,CCF 高级会员,主要研究领域为海量数据计算,无线传感器网络.

王宏志(1978－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为数据库,大数据,数据质量.