大数据质量管理:问题与研究进展

王宏志

哈尔滨工业大学计算机科学与技术学院,哈尔滨 150001

摘要 当前大数据在多个领域广泛存在,大数据的质量对其有效应用起着至关重要的作用,因而需要对大数据进行质量管理。尽管数据质量管理方面已经有一些研究成果,但由于大数据具有规模大、速度快和多样性高的特点,现有的方法难以适用于大数据质量管理。本文针对错误发现、错误修复和劣质数据查询处理,综述了大数据质量管理的问题与挑战,认为大数据质量管理的挑战主要有计算困难、错误混杂和缺少知识3个方面。本文依据这3个方面的解决方法,对大数据质量管理目前的研究进展进行了综述,并展望了大数据质量管理未来的研究方向。

关键词 数据质量;大数据;数据清洗

中图分类号 TP311.13 **文献标志码** A **doi** 10.3981/j.issn.1000-7857.2014.34.011

Big Data Quality Management: Problems and Progress

WANG Hongzhi

Department of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China

Abstract Big data have wide applications. Since the quality of big data plays a crucial role in these data-centric applications, data quality management techniques for big data are in demand. Although some theories and techniques for data quality management have been proposed, due to the volume, variety and velocity of big data, current methods could hardly be applied to data management for big data. This paper discusses the problems and challenges for error detection, error repair and query processing of dirty data in big data management, and identifies intractability, mixed errors and the lack of knowledge as three new challenges to data quality management. The progress of big data quality management in these three aspects is reviewed and open problems for future research are proposed.

Keywords data quality; big data; data cleaning

当前,大数据得到了广泛应用,对科学和产业产生了巨大影响。关于大数据的准确定义,科学界仍缺乏统一认识,从字面上理解,其最本质的特点在于数据量"大",除此之外,还包括了获取、管理及处理时的复杂性。大数据具有明显的时代特征,习惯上将其总结为4个"V":规模性(volume),高速性(velocity),多样性(variety)和价值稀疏性(value)。由于大数据的这些特征,使其有更大可能产生数据质量问题,即出现不一致、不精确、不完整、过时等问题或者描述同一实体的数据出现冲突(简称为实体不同一)等错误[1]。大数据有可能产生数据质量问题,其具体原因如下:

1) 由于规模性,大数据获取、存储、传输和计算过程中可能产生更多错误,如果对其采用人工错误检测与修复,将导致成本极其巨大以致难以有效实施。

2) 由于高速性,数据的大量更新会导致过时数据迅速产生,也更易于产生不一致数据,为人工错误检测与修复带来困难。例如,大型强子对撞机实验设备中包含了15亿个传感器,平均每秒收集超过4亿条实验数据,更新的数据将会导致之前存储数据迅速过时,而在更新速度如此快的情况下,传统方法难以有效用新数据替换对应的旧数据。

3) 大数据的多样性指的是数据来源和形式上的多样,这使得数据有更大的可能产生不一致和冲突。例如,在互联网上不同电子商务网站中获取到的描述同一商品的数据有很大可能存在冲突。

如果没有良好的数据质量,大数据将会对决策产生误

收稿日期:2014-09-25;修回日期:2014-11-06

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

作者简介:王宏志,副教授,研究方向为大数据管理、数据质量、复杂数据管理,电子信箱:wangzh@hit.edu.cn

引用格式:王宏志. 大数据质量管理:问题与研究进展[J]. 科技导报, 2014, 32(34): 78-84.

导,甚至产生有害的结果。根据估算,数据错误每年造成美国工业界经济损失约占GDP的6%[2]。在医疗大数据应用方面,根据美国医疗委员会的统计,由于数据错误引起的医疗事故仅在美国每年就导致高达98000名患者丧生[3]。在电信大数据应用中,数据错误经常导致故障排除的延误、多余设备租用和服务费收取错误,损害了企业信誉并失去很多用户[4]。在商业大数据的应用中,美国零售业每年仅因标价错误就损失25亿美元[5]。在金融大数据的应用中,2008年因数据质量问题导致的信用卡欺诈失察即造成48亿美元的损失[6]。统计数据显示,50%以上的数据仓库项目由于数据质量问题而不得不取消或延迟[7]。

由于大数据存在数据质量问题,并且会带来严重的后果,因此需要对大数据进行质量管理,从而确保基于大数据各种应用的有效实施。由于其重要性,研究人员围绕数据质量管理展开了研究,取得了一系列的研究成果,然而,大数据为数据质量管理带来了诸多挑战问题,进一步增加了数据质量管理的难度,同时也给数据质量管理带来了新的研究机遇。本文对数据质量管理的现有方法进行简要综述,分析大数据为数据质量管理带来的挑战,同时对大数据质量管理研究进行综述,并展望未来的研究方向。

1　大数据质量管理的问题与挑战

质量管理包含错误发现、错误修复和容忍错误的近似查询处理等不同方面,因此依照数据质量管理的不同方面,综述这些方面的问题和现有解决方案,进而分析对大数据进行质量管理时面临的挑战。

1.1　错误发现

错误发现指的是发现存在质量问题的数据,根据方法的不同,当前的研究主要有实体识别、基于规则的错误发现和基于主数据的错误发现3类方法。

1）实体识别。实体识别指的是发现描述同一现实世界中实体的不同数据。通过实体识别,可以有效地检测出实体不同一、过时等错误。实体识别是数据质量方面研究最多的问题,研究人员已提出了多个实体识别算法[8]。尽管也有一些工作研究如何提高识别实体的效率[9],但当前实体识别的计算复杂性仍然远超过线性,难以应用于大数据。

2）基于规则的错误发现。基于规则的错误发现指的是利用给定规则捕捉数据中的错误,即找出违反规则的元组作为错误元组。规则有多种形式,包括描述一致性的函数约束[10]、条件函数约束[11]、条件包含约束[12]、描述时效性的时序约束[13]和描述精确性的精确性约束等。文献[11]针对集中存储的关系数据库,设计了基于SQL语言的自动检测算法,用于查找违反条件函数约束和条件包含约束的数据元组。文献[13]在时间戳缺失的情况下,用完整性约束语言来描述同一实体不同信息值间的时序关系,给出了应用时序关系和拷贝关系推导实体最新信息的推理机制。文献[14]提供了一个模型来确定相对准确的数据,提出了精确性确定规则和chase程序的推理系统。

目前基于规则的方法难以直接应用于大数据。一方面,规则通常需要人来给出,而对于大数据而言,人难以了解数据的全貌,故难以给出有效的规则,尽管有一些规则发现算法提出[15~18],但这些算法需要一个数据量较小且质量高的学习集合,在大数据上难以有效找到这样的集合。另一方面,大数据高速的更新会使得规则迅速失效。

3）基于主数据的错误发现。主数据是一个高质量的数据集合,用于给多种应用提供企业核心业务实体的一个同步一致的视图[19],以主数据为基准,可以用来发现数据中的错误。例如,文献[20]提出了信息相对完全理论来表述信息库相对于主数据和用户查询的完整性,相对于主数据和用户查询,判定一个信息库完全与否。然而,当前主数据通常是人工维护的一个小规模的数据集合,对于大数据,其部分数据难以体现出其全貌,而维护大规模的主数据需要更高的成本;而大数据的高速性也需要主数据进行相应的快速更新,从而导致了成本的进一步提高。

1.2　错误修复

数据错误修复是指对存在错误的数据进行修改或者补充,提高其质量。根据数据错误修复思路的不同,数据修复可以分为基于规则的修复、真值发现和基于机器学习的修复。

1）基于规则的修复。基于规则的修复主要指修改数据使其满足给定规则的数据修复方法。文献[21]提出了一种基于函数依赖的修复算法GREEDY_REPAIR,它采用启发式的方法来修复字符串型数据,以修改破坏约束右部属性的取值来纠正不一致。启发式算法BATCHREPAIR[22]由上述方法扩展得到,该方法针对函数依赖修复效果欠佳的问题,采用条件函数依赖进行不一致数据的修复工作。文献[23]从图论方法入手修复不一致数据,提出了通过删除元组解决数据不一致的冲突图模型及相应方法,以删除元组为修复操作,将不一致数据修复问题转化为图上最大独立集问题。文献[24]针对多种约束组合时出现的冲突现象,提出了多规则的修复框架,以及新的语义限制等价生成依赖,基于类间的偏序关系确定修复顺序。对于过时数据的修复问题,文献[25]将冲突消解问题转化为求解集合中最新且一致的值,并提出了基于时间偏序和条件函数依赖的冲突消解方法。文献[20]讨论了数据库不完全时,如何扩展这个信息库以包括足够的信息来回答用户的查询。

2）真值发现。对于实体不同一数据的修复通过发现描述实体属性的真实值实现。文献[26]通过迭代方式计算源的真实度和值的自信度,然后通过值的自信度寻找真值。文献[27]考虑了数据源之间的依赖关系,这种依赖关系需要从值自信度推测出来。文献[28]提出了基于数据源之间的依赖关系的真值发现,使独立的数据源在投票过程中具有更高的权重。文献[29]给出了贝叶斯推理模型,其推理的真值必须满足最大后验概率,但此算法的指数复杂度使其难以实际应用,尽管有基于抽样的近似算法,但该算法的抽样对初始值

比较敏感,测试样本也必须具有很好的质量。

3)机器学习。机器学习主要用于不完整性数据的修复,从数据的完整部分学习相应的模型用于填充缺失的值。基于机器学习技术的缺失值填充方法主要包括决策树[30]、贝叶斯网络[31]及神经网络[32]。文献[33]给出了"伪装缺失值"的检测与清洗方法。

4)错误修复的难题。尽管当前有一些错误修复方法提出,但如下两方面的问题使得这些方法难以应用于大数据。第一,这些方法计算复杂度较高,有的问题甚至是NP(非确定图灵机多项式)难问题,难以应用于大数据。第二,由于大数据中错误存在混杂的情况,这些方法在修复一种错误的同时可能会引入另外一种错误,例如基于机器学习的缺失值填充可能会引入数据的不一致。

1.3 劣质数据查询处理

在一些数据中的错误难以有效修复的情况下,需要容忍数据中的错误,在存在错误的数据上进行查询处理,从存在质量问题的数据上获得高质量的查询结果。这方面的研究主要可以分为近似数据操作、不一致数据查询处理和有空值数据的查询处理。

1)近似数据操作。当前主要的近似操作包括近似搜索和近似连接操作。近似搜索操作在数据库中查找和给定查询相似性大于给定阈值的结果,近似连接操作返回2个数据集合中相似性大于给定阈值的对。这两类操作均可在存在错误的数据集合上得到近似计算结果。针对这两方面问题,有大量研究结果提出,文献[34]对相关研究结果进行了综述。

2)不一致数据查询处理。文献[35]综述数据修复和一致性查询问题。文献[36]首次提出一致性查询问题。文献[37]定义修复语义下的一致性查询,即查询需要满足所有的修复。文献[38]提出EQUIP系统计算合取查询的一致性解,其将一致性查询的补问题归约到0-1规划问题,通过求解规划方程去掉不满足一致性的解。此外,基于析取逻辑程序以及稳定语义模型可以解任意合取查询的一致性回答问题[39],且一致性限制并不局限于主键约束,然而其复杂度为$\prod_2^p$。

文献[40]、[41]研究了*CERTAINTY*(q)问题,即只考虑一致性约束为主键约束,修复类型为子集修复的情况。文献[42]提出一个包含一阶可表达的查询的更大的类,指出不在该类中的涉及到两个不同关系表进行连接操作的查询一定是不可一阶改写的。文献[43]研究了满足函数依赖情况下基于主键约束的一致性查询问题,首先认为数据库是部分一致的(满足函数依赖),基于此研究了*CERTAINTY*(q, Σ)问题,即在满足函数依赖集合Σ的情况下,Σ是否是一阶可表达的,此问题限制Σ不带自连接。文献[44]、[45]研究了*CERTAINTY*(q)的变种:计数的复杂性问题,并证明了对于不带自连接的合取查询q,#*CERTAINTY*(q)是P问题或是#P-完全问题。

3)不完整数据的查询处理。当前不完整数据的查询处理主要集中于skyline查询。针对不完整数据的skyline查询主要针对空缺属性上的支配关系以及基于新支配关系的高效计算开展研究。Khalefa等[46]第一次提出了不完整数据skyline查询的概念,并提出替换算法、桶算法和skyline算法。Alwan等[47]提出对不完整数据进行填充值的skyline查询方法,利用填充后的属性值进一步减少skyline点个数,从而提高查询精度。Bharuka等[48]基于排序搜索算法(SRA)解决不完整数据skyline查询问题,该方法与文献[46]中提出的ISkyline算法相比,可以渐进式输出skyline点,而不需要等全部数据点处理完毕才能一次性输出所有skyline点。Miao等[49,50]提出不完整数据k-Skyband查询问题,并引入失效skyline、阴影skyline和厚度仓库的概念。k-Skyband查询是指查询数据集中被k个其他数据项支配的数据项,一个数据项被支配的次数越少,说明该数据项在各个属性上的总体取值情况越好。

Hadjali等[51]提出了用户偏好存在丢失情况下的skyline查询问题。他们要解决的是根据用户过去的上下文偏好信息,查询当前上下文中存在偏好丢失情况下不被支配的skyline元组。Arefin等[52]考虑数据库中存在数据缺失情况下的skyline集合查询问题,他们提出了基于替换策略的RBSSQ算法,可以有效解决数据库中元组丢失任意数量维度时的问题。Markus等[53]专门提出了针对偏好数据库查询中空缺值进行处理的方法,通过扩展偏好代数,提出了一种标准模型,能够在不破坏偏好支配关系传递性的情况下解决偏好查询问题。

4)劣质数据计算中的难题。对大数据而言,有两方面的难题尚未得到有效解决,一方面是这些计算的时间空间复杂性还较高,难以应用于大数据,另一方面在于当前的方法仅面向一种错误,难以在具有多种混合错误的数据上进行计算。

1.4 大数据为数据质量管理带来的挑战

根据上述讨论,大数据的特点为数据质量管理带来诸多技术挑战,可归纳为:

1)计算困难。大数据规模巨大,达到PB级甚至EB级,而且增长速度快,因此大数据的质量管理需要时间和空间复杂性为线性甚至亚线性的算法,也需要相应并行算法加快计算速度。特别是对于增长速度快的大数据需要在应用允许的时间范围内实施数据质量管理。如何设计时空有效的大数据质量管理算法是第一个挑战性问题。当前数据质量管理方法较少考虑在大规模数据上的可扩展性,其中一些问题甚至被证明是不可计算问题或NP完全问题,当前算法的时间和空间复杂度远超过线性,难以应用于TB级以上的数据,缺少面向大数据的线性或亚线性算法和并行算法。

2)混杂错误。大数据的多样性导致其出现错误的根源复杂,加之大数据在存储和通信过程中造成的错误,可能出现多种类型错误混合并相互影响的情况。而错误的多个方面并非独立,会产生关联,例如精确性会影响一致性、实体同一性和时效性关联。检测与修复相互影响的多种错误是大数据质量管理的第二个挑战性问题。当前的数据质量管理方法通常针对某个特定类型错误提出,缺少对错误之间关联的认知,也缺少多种错误混合发生时的错误检测与修复以及查询处理技术。

3）知识缺少。大数据价值密度低，仅从小部分数据难以得到对数据的完整认识；大数据规模巨大，来源多样，难以认知其全貌，从而难以全面认识大数据的语义。如何有效获取充分的语义信息支持大数据质量管理是第三个挑战性问题。当前大多数数据质量管理方法需要专家用户指定规则和相关参数，而自动错误检测修复和规则学习算法需要主数据或清洁的训练集。就大数据而言，一方面，聘请专家或维护主数据成本很高；另一方面，缺少自动选取有效训练集的算法。因此当前数据质量管理算法难以直接应用于大数据。

2　大数据质量管理研究进展

2.1　针对计算困难的解决方法

针对计算困难的问题，主要有两类解决方案，一是采取并行化技术实施数据质量管理，二是为数据清洗设计线性亚线性的算法。

1）并行数据质量管理。并行数据质量管理的研究当前刚刚起步，研究工作主要集中在并行实体识别和并行相似性连接两个方面。

Dedoop[54]提供一个分块和匹配方法库，支持浏览器输入实体识别策略。为了简化多个相似性策略的实体识别配置，Dadoop支持基于训练的机器学习方法，将特定的实体识别策略自动转化为Hadoop集群上并行执行的MapReduce任务。Dedoop支持无冗余的多次分块以及先进的负载平衡方案[55,56]。文献[57]基于MapReduce平台设计了实体识别算法，该方法首先通过属性值并行计算记录间的相似程度，而后基于图聚类的方法进行实体识别从而输出得到最终结果。

文献[58]基于MapReduce框架设计了分类属性的填充算法，该算法利用基于概率的推理填充缺失值，该推理过程是在一个基于属性相关性而建立起来的贝叶斯网络中进行。

Vernica等[59]提出了MapReduce框架下的前缀过滤和PP连接算法，这种方法也可以应用到基于Jaccard相似性的相似连接。Metwally和Faloutsos[60]提出了V-SMART-Join算法，这种算法在token级别聚集相似性分数的贡献，从而计算相似性函数。Afrati等[61]研究了球散列技术和描点分析法来加速MapReduce上的相似性连接。Okcan和Riedewald[62]设计了Theta-Join框架可以处理任意约束的连接。

文献[63]研究了基于MapReduce相似性字符串连接，支持多种基于集合的相似性函数和基于字母的相似性函数，该方法扩展了现有基于划分的签名来支持基于集合的相似性函数，使用签名来生成key-value对，为了减少通讯开销，这种方法通过合并key-value对来减少key-value对数量。文献[64]提出了ClusterJoin框架，这种方法将数据空间进行基于数据分布的划分，将每条元组分布到其基于距离函数可能产生连接结果的划分中，该方法为不同距离函数设计了一个强候选元素过滤集合，从而每个元组仅需要被分不到少数划分中从而保证正确性，为了解决高维数据中常见的偏斜问题，进而设计了基于采样的动态负载平衡策略，其提供了划分规模的强概率保证，从而确保了可扩展性。

Cleanix[65]是一个基于并行机群的大数据清洗系统，其包含了异常值检测和修复、缺失值填充、实体识别以及冲突消解等并行数据清洗模块。

2）数据清洗的线性亚线性算法。研究主要集中在数据流清洗，即通过扫描数据一次完成数据清洗，其主要应用背景是RFID数据的清洗。

文献[66]、[67]是早期RFID数据清洗工作，提出了基于规则的推理方法。这些方法直接作用于数据流上或者RFID数据已经存储。使用规则的一个例子是将首先识别出的数据[67]或者读取次数最多的值[68]置为真值。文献[69]提出的方法利用参考对象（例如架子等）清洗RFID数据流。文献[70]通过考虑容量约束建立概率模型，提出了基于后验阅读率Metropolis-Hasting采样来从模型中推理隐变量得到对象标签的位置。文献[71]研究了用于对象检测的RFID数据流清洗方法，提出了移动环境下对象检测的概率模型，基于该模型设计了贝叶斯推理用于清洗RFID数据。为了从运动的分布中抽样数据，设计了Gibss采样器快速有效地清洗RFID数据。

文献[72]提出了清洗有噪数据流的问题，其中噪声指的是错误标记的训练样例，目标是精确地表示和去除误导的数据，从而提高基于清洁数据流得到的预测模型的精度。为了达到这个目的，其首先使用偏置方差分解得到用于数据流清洗的最大方差边际（MVM），基于此概念，进一步提出了局部和全局的过滤器框架结合局部（在单一数据块中）和全局（跨越多个连续数据块）过滤器来发现错误数据。

2.2　针对混杂错误的解决方法

数据质量的多个方面相互关联。当前绝大多数研究人员把数据质量的5个方面当作孤立的方向，已经有研究人员开始复合类型错误的检测与修复，文献[73]探讨了信息修复和元组匹配的交互影响，基于条件函数约束和匹配约束提出了一个同时支持信息修复和实体识别的信息清洗框架。文献[74]提出了一种考虑数据时效性的冲突消解模型，该模型利用时序偏序关系和时序约束来描述时效性，利用常数条件函数依赖描述一致性。该论文提出可以利用数据的时效顺序辅助修复不一致数据，反之亦然，还提出同时考虑数据时效性和一致性的统一数据消解算法。NADEEF[75]是一个端到端的数据清洗系统，提供编程界面允许用户输入各种异构的数据质量规则，其中规则包括一致性约束和匹配规则，并提供核心算法检测错误并改正错误。

2.3　针对知识缺少的解决方法

当前针对知识缺少的主要解决方法是引入用户的工作。特别是通过众包技术进行数据质量管理。

目前在数据质量管理领域众包技术使用的最为广泛的问题是实体识别问题。Demartini等[76]开发了一个人机交互系统，并增加了一个实体识别结果筛选概率框架。Wang等[77]提出了一个以预算为基础的方法，假设没有足够的金钱标记所有记录，讨论如何利用有限的资金标识最有用的比较对。

Wang等[77]开发了一种人机混合系统先用机器方法剔除一些明显不匹配的记录,将剩下的匹配对利用众包完成。随后又提出了利用传递关系来减少可众包的记录对,并提出了一种可优化的标记排列顺序算法[78]。

文献[79]提出了利用众包填充缺失值的策略,首先选择适用于众包填充的缺失值,继而根据属性类型选择不同缺失值填充方法。文献[80]将主动学习和众包相结合进行真值发现,该方法采用迭代方法进行真值发现,在每一次迭代中通过主动学习发现真值不确定性最高的属性进行众包,并基于返回结果进行投票,根据投票进一步判定真值不确定属性。

CrowdCleaner[81]是一个适用于Web上多版本数据的清洗系统,该系统使用基于众包技术来检测和修复传统数据清洗方法难于解决的问题,并结合主动和被动众包方法纠正多版本数据中的错误。

Lofi等[82,83]提出了采用众包平台数据库技术提高不完整数据skyline查询结果质量的方法。提出了精细的错误处理模型,在关注正确元组的同时,更加重点关注那些最有可能出错的元组。通过利用众包平台结合启发式技术,尽可能消除错误值,集中处理最可能产生用户期待结果的元组。

另外一种方法是通过提取互联网信息获取相应的语义信息。WebPut[84]是一个基于Web信息的数据填充信息,该系统扩展了信息提取方法用于形式化向Web搜索查询以高效检索出缺失值。WebPub使用了基于置信的方法自动为每个缺失值选择最有效的填充查询,并设计了贪心的迭代算法确定数据填充顺序,并按顺序依次发布相应的查询。该论文还提出一些优化策略用于在元组级别和数据库级别降低估计填充查询置信度的代价。

3 未来的工作

随着大数据的广泛应用,数据质量管理将越来越重要,而面向大数据质量管理的研究刚刚起步,还存在诸多亟待解决的问题。

1)数据质量多维度相互影响的认知。当前尽管有一些工作涉及到两种不同种类数据质量问题的协同处理,然而,尚无综合考虑数据质量多个维度的方法提出,缺乏对这种相互影响的深入认识,而对数据质量问题的全面解决需要对这种相互影响进行定量分析,而且需要在统一逻辑框架下对不同数据质量问题的统一表达,这方面现在处于空白状态,有待深入研究。

2)高效数据错误检测与修复算法设计。当前的错误检测与修复算法普遍计算复杂性超过线性,而且缺少有效的并行算法,难以适用于大数据,而仅有的线性复杂性算法和并行算法只集中在相似性连接、实体识别、RFID错误检测等几个问题,对于大多数错误检测与修复的问题尚无适用于大数据的高效算法提出,给研究人员很大的进一步研究的空间。

3)劣质大数据近似计算理论与算法。对于大数据,在很多情况下,错误难以完全修复,而且修复过程中经常存在无法了解属性真实值的情况,因此在很多情况下需要容许错误的存在,在存在错误的劣质数据上进行近似计算。现在已经存在一些劣质数据查询处理算法,然而有两方面的工作做得还比较初步,有待进一步探索,一方面是当前劣质数据的计算对算法可扩展性考虑较少,难以应用到大数据,另一方面是当前劣质数据近似计算的研究成果主要集中在查询处理,计算的其他重要方面(如数据挖掘等)[85]研究成果较少,存在大量需要研究的问题。

4)支持数据质量管理的数据语义信息获取。目前数据质量管理的重要问题之一在于缺乏对数据语义信息的充分了解。因此支持数据质量管理的数据语义信息获取成为一个亟待解决的问题。当前尽管有基于众包和互联网信息方法用于获取知识以支持数据清洗,但这两方面的研究还刚刚起步,仅覆盖了数据质量中实体统一性、完整性等少数几个维度和部分问题,有很多问题需要研究人员进一步研究和探索。

4 结论

由于具有规模大、多样性高和更新速度快的特点,大数据存在数据质量问题的可能性更大。数据质量对大数据应用起着至关重要的作用,因此数据质量管理是大数据管理的核心步骤之一。与传统数据质量管理相比,大数据质量管理存在计算困难、错误混杂和缺少知识3方面的技术挑战亟待进一步研究。

参考文献(References)

[1] Li J Z, Liu X M. An important aspect of big data: Data usability[J]. Journal of Computer Research and Development, 2013, 50(6): 1147-1162.

[2] Eckerson W W. Data quality and the bottom line: Achieving business success through a commitment to high quality data[R]. Renton, WA: The Data Warehousing Institute, 2000: 12-20.

[3] Institute of Medicine. To err is human: Building a safer health system [M]. Washington: The National Academies Press, 1999.

[4] Bohannon P, Fan W F, Flaster M, et al. A cost-based model and effective heuristic for repairing constraints by value modification[C]. ACM SIGMOD International Conference on Management of Data, Baltimore, Maryland, June 14-16, 2005.

[5] English L. Plain English on data quality: Information quality management: The next frontier[J]. DM Review Magazine, 2000.

[6] Ben W, Schulz S. Credit card statistics, industry facts, debt statistics [EB/OL]. 2010-03-19, [2014-09-25]. http://www.creditcards.com.

[7] Gartner. Gartner says more than 50 percent of data warehouse projects will have limited acceptance or will be failures through 2007[EB/OL]. 2005-02-24, [2014-09-25]. http://www.gartner.com/newsroom/id/492112.

[8] Elmagarmid A K, Ipeirotis P G, Verykios V S. Duplicate record detection: A survey[J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(1): 1-16.

[9] Christen P. A survey of indexing techniques for scalable record linkage and deduplication[J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(9): 1537-1555.

[10] Rahm E, Do H H. Data cleaning: Problems and current approaches[J].

Bulletin of the Institute of Electrical and Electronics Engineers Data Engineering Bulletin, 2000, 23(4): 3-13.

[11] Fan W F, Geerts F, Jia X B, et al. Conditional functional dependencies for capturing data inconsistencies[J]. ACM Transactions on Database Systems, 2008, 33(2): 1-48.

[12] Bravo L, Fan W F, Ma S. Extending dependencies with conditions[C]. The 33rd International Conference on Very Large Data Bases, University of Vienna, Austria, September 23-27, 2007.

[13] Fan W F, Geerts F, Wijsen J. Determining the currency of data[C]. The 30th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS), Athens, Greece, June 12-16, 2011.

[14] Cao Y, Fan W F, Yu W Y. Determining the relative accuracy of attributes[C]. 2013 International Conference on Management of Data, New York, USA, June 23-28, 2013.

[15] Chiang F, Miller R J. Discovering data quality rules[J]. The Proceedings of the VLDB Endowment, 2008, 1(1): 1166-1177.

[16] Fan W F, Geerts F, Li J Z, et al. Discovering conditional functional dependencies[J]. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(5): 683-698.

[17] Chu X, Ilyas I F, Papotti P. Discovering denial constraints[J]. The Proceedings of the VLDB Endowment, 2013, 6(13): 1498-1509.

[18] Bauckmann J, Abedjan Z, Leser U, et al. Discovering conditional inclusion dependencies[C]. The 21st ACM International Conference on Information and Knowledge Management, Maui, Hawaii, October 29-November 2, 2012.

[19] Loshin D. Master data management[M]. San Francisco: Morgan Kaufmann, 2008.

[20] Fan W F, Geerts F. Relative information completeness[J]. ACM Transactions on Database Systems, 2010, 35(4): 27-35.

[21] Bohannon P, Fan W, Flaster M, et al. A cost-based model and effective heuristic for repairing constraints by value modification[C]. ACM SIGMOD International Conference on Management of Data, Baltimore, Maryland, USA, June 14-16, 2005.

[22] Cong G, Fan W, Geerts F, et al. Improving data quality: Consistency and accuracy[C]. The 33rd International Conference on Very Large Data Bases, University of Vienna, Austria, September 23-27, 2007.

[23] Arenas M, Bertossi L E, Chomicki J, et al. Scalar aggregation in inconsistent databases[J]. Theoretical Computer Science, 2003, 296(3): 405-434.

[24] Geerts F, Mecca G, Papotti P, et al. The LLUNATIC data-cleaning framework[J]. The Proceedings of the VLDB Endowment, 2013, 6(9): 625-636.

[25] Fan W F, Geerts F, Tang N, et al. Inferring data currency and consistency for conflict resolution[C]. 29th IEEE International Conference on Data Engineering, Brisbane, April 8-12, 2013.

[26] Galland A, Abiteboul S, Marian A, et al. Corroborating information from disagreeing views[C]. The third ACM International Conference on Web Search and Data Mining, New York, USA, February 3-6, 2010.

[27] Dong X L, Berti-Equille L, Srivastava D. Integrating conflicting data: The role of source dependence[J]. The Proceedings of the VLDB Endowment-PVLDB, 2009, 2(1): 550-561.

[28] Dong X L, Berti-Equille L, Srivastava D. Truth discovery and copying detection in a dynamic world[J]. The Proceedings of the VLDB Endowment-PVLDB, 2009, 2(1): 562-573.

[29] Zhao B, Rubinstein B I P, Gemmell J, et al. A bayesian approach to discovering truth from conflicting sources for data integration[J]. The Proceedings of the VLDB Endowment, 2012, 5(6): 550-561.

[30] Lakshminarayan K, Harp S A, Goldman R, et al. Imputation of missing data using machine learning techniques[C]. The Second International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, August 2-4, 1996.

[31] Mayfield C, Neville J, Prabhakar S. ERACER: A database approach for statistical inference and data cleaning[C]. ACM SIGMOD International Conference on Management of Data, Indianapolis, Indiana, USA, June 6-10, 2010.

[32] Setiawan N A, Venkatachalam P, Hani A F M. Missing attribute value prediction based on artificial neural network and rough set theory[J]. Biomedical Engineering and Informatics, 2008, 1: 306-310.

[33] Hua M, Pei J. Cleaning disguised missing data: A heuristic approach [C]. The 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California, USA, August 12-15, 2007.

[34] Lin X M, Wang W. Set and string similarity queries: A survey[J]. Chinese Journal of Computers, 2011, 34(10): 1853-1862.

[35] Leopoldo B. Database repairing and consistent query answering[M]. California: Morgan & Claypool, 2011.

[36] Bry F. Query answering in information systems with integrity constraints[M]//Integrity and Internal Control in Information Systems. New York: Springer, 1997: 113-130.

[37] Arenas M, Bertossi L, Chomicki J. Consistent query answers in inconsistent databases[C]. Eighteenth ACM SIGACT- SIGMOD- SIGART Symposium on Principles of Database Systems, Philadelphia, Pennsylvania, May 31-June 2, 1999.

[38] Kolaitis P G, Pema E, Tan W C. Efficient querying of inconsistent databases with binary integer programming[J]. The Proceedings of the VLDB Endowment, 2013, 6(6): 397-408.

[39] Barceló P, Bertossi L. Logic programs for querying inconsistent databases [M]//Practical Aspects of Declarative Languages. New York: Springer, 2003: 208-222.

[40] Fuxman A, Fazli E, Miller R J. Conquer: Efficient management of inconsistent databases[C]. ACM SIGMOD International Conference on Management of Data, Baltimore, Maryland, USA, June 14-16, 2005.

[41] Fuxman A, Miller R J. First-order query rewriting for inconsistent databases[J]. Journal of Computer and System Sciences, 2007, 73(4): 610-635.

[42] Wijsen J. Consistent query answering under primary keys: A characterization of tractable queries[C]. The 12th International Conference on Database Theory, St Petersburg, Russia, March 23-25, 2009.

[43] Greco S, Pijcke F, Wijsen J, et al. Certain query answering in partially consistent databases[J]. Proceedings of the VLDB Endowment, 2014, 7 (5): 32-65.

[44] Maslowski D, Wijsen J. Counting database repairs that satisfy conjunctive queries with self-joins[C]. The 17th International Conference on Database Theory, Athens, Greece, March 24-28, 2014.

[45] Maslowski D, Wijsen J. On counting database repairs[C]. The 4th International Workshop on Logic in Databases, San Miniato, March 25, 2011.

[46] Khalefa M E, Mokbel M F, Levandoski J J. Skyline query processing for incomplete data[C]. 2008 IEEE 24th International Conference on Data Engineering (ICDE '08), Cancun, April 7-12, 2008.

[47] Alwan A A, Ibrahim H, Udzir N I, et al. Skyline queries over incomplete multidmensional database[C]. The 3rd International Conference on Computing and Informatics, Bandung, June 8-9, 2011.

[48] Bharuka R, Kumar P S. Finding skylines for incomplete data[C]//Proceedings of the Twenty- Fourth Australasian Database Conference. Gold Coast, Queensland: Australian Computer Society, 2013, 137: 109-117.

[49] Miao X, Gao Y, Chen L, et al. On efficient k-skyband query process-

ing over incomplete data[M]//Database Systems for Advanced Applications. Berlin Heidelberg: Springer, 2013: 424-439.

[50] Gao Y, Miao X, Cui H, et al. Processing k-skyband, constrained skyline, and group-by skyline queries on incomplete data[J]. Expert Systems with Applications, 2014, 41(10): 4959-4974.

[51] Hadjali A, Pivert O, Prade H. Possibilistic contextual skylines with incomplete preferences[C]//Proceeding of 2010 International Conference of Soft Computing and Pattern Recognition. New York, USA: Institute of Electrical and Electronics Engineers, 2010: 57-62.

[52] Arefin M S, Morimoto Y. Skyline sets queries from databases with missing values[C]//Proceeding of 22nd International Conference on Computer Theory and Applications. Chengdu: Institute of Electrical and Electronics Engineers, 2012: 24-29.

[53] Markus E, Patrick R, Florian W, et al. Handling of NULL values in preference database queries[C]. 20th European Conference on Artificial Intelligence, Montpellier, France, August 27-31, 2012.

[54] Kolb L, Thor A, Rahm E, et al. Efficient deduplication with hadoop[J]. The Proceedings of the VLDB Endowment, 2012, 5(12): 1878-1881.

[55] Kolb L, Thor A, Rahm E. Load balancing for MapReduce-based entity resolution[C]. International Council for Open and Distance Education, Washington D C, April 1-5, 2012.

[56] Kolb L, Thor A, Rahm E. Block-based load balancing for entity resolution with MapReduce[C]. The 20th ACM International Conference on Information and Knowledge Management, Glasgow, United Kingdom, October 24-28, 2011.

[57] Huo R, Wang H Z, Zhu R, et al. Entity identification in big data based on MapReduce[J]. EIBM, 2013, 50(S2): 20-35.

[58] Jin L, Wang H Z, Huang S B, et al. Missing value imputation in big data based on Map-Reduce[J]. Journal of Computer Research and Development, 2013, 50(Sl): 312-321.

[59] Vernica R, Carey M J, Li C. Efficient parallel set-similarity joins using mapreduce[C]. ACM SIGMOD International Conference on Management of Data, Indianapolis, Indiana, USA, June 6-10, 2010.

[60] Metwally A, Faloutsos C. V-smart-join: A scalable mapreduce framework for all-pair similarity joins of multisets and vectors[J]. The Proceedings of the VLDB Endowment, 2012: 213-300.

[61] Afrati F N, Sarma A D, Menestrina D, et al. Fuzzy joins using mapreduce[C]. International Council for Open and Distance Education, Washington D C, April 1-5, 2012.

[62] Okcan A, Riedewald M. Processing theta-joins using mapreduce[C]. ACM SIGMOD International Conference on Management of Data, Athens, Greece, June 12-16, 2011.

[63] Deng D, Li G L, Hao S, et al. MassJoin: A mapreduce-based method for scalable string similarity joins[C]. 2014 IEEE 30th International Conference on Data Engineering, Moscow, Russia, March 31-April 4, 2014.

[64] Sarma A D, He Y Y, Chaudhuri S. ClusterJoin: A similarity joins framework using MapReduce[J]. The Proceedings of the VLDB Endowment, 2014, 7(12): 1059-1070.

[65] Wang H Z, Li M D, Bu Y Y, et al. A big data cleaning parfait[C]. The 23rd ACM International Conference on Information and Knowledge Management, Shanghai, Nov 3-7, 2014: 10-23.

[66] Bornhövd C, Lin T, Haller S, et al. Integrating automatic data acquisition with business processes experiences with sap´s auto-id infrastructure[J]. The Proceedings of the VLDB Endowment, 2004, 30: 1182-1188.

[67] Rao J, Doraiswamy S, Thakkar H, et al. A deferred cleansing method for rfid data analytics[C]. The 32nd International Conference on Very Large Data Bases, Seoul, Korea, September 12-15, 2006.

[68] Jeffery S, Garofalakis M, Franklin M. Adaptive cleaning for rfid data streams[C]. The 32nd International Conference on Very Large Data Bases, Seoul, Korea, September 12-15, 2006.

[69] Tran T, Sutton C, Cocci R, et al. Probabilistic inference over rfid streams in mobile environments[C]. The 25th International Conference on Data Engineering, March 29-April 2, 2009.

[70] Chen H, Ku W, Wang H, et al. Leveraging spatio-temporal redundancy for rfid data cleansing[C]. ACM SIGMOD International Conference on Management of Data, Indianapolis, Indiana, USA, June 6-10, 2010.

[71] Zhao Z, Ng W. A model-based approach for RFID data stream cleansing[C]. The 21st ACM International Conference on Information and Knowledge Management, Maui, Hawaii, October 29- November 2, 2012.

[72] Zhu X Q, Zhang P, Wu X D, et al. Cleansing noisy data streams[C]. The IEEE International Conference on Data Mining, Cancún, México, December 15-19, 2008.

[73] Fan W F, Li J Z, Ma S, et al. Interaction between record matching and data repairing[C]. ACM SIGMOD International Conference on Management of Data, Athens, Greece, June 12-16, 2011.

[74] Fan W F, Geerts F, Tang N, et al. Inferring data currency and consistency for conflict resolution[C]. The 29th IEEE International Conference on Data Engineering, Brisbane, April 8-12, 2013.

[75] Ebaid A, Elmagarmid A K, Llyas I, et al. NADEEF: A generalized data cleaning system[J]. The Proceedings of the VLDB Endowment, 2013, 6(12): 1218-1221.

[76] Demartini G, Difallah D E, Cudr´e-Mauroux P. ZenCrowd: Leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking[C]. The 21st World Wide Web Conference, Lyon, France, April 16-20, 2012.

[77] Wang J, Kraska T, Franklin M J, et al. CrowdER: Crowdsourcing entity resolution[J]. The Proceedings of the VLDB Endowment, 2012, 5(11): 1483-1494.

[78] Wang J N, Li G L, Kraska T, et al. Leveraging transitive relations for crowdsourced joins[C]. International Conference on Management of Data, New York, USA, June 22-27, 2013.

[79] Ye C, Wang H Z. Capture missing values based on crowdsourcing[J]. Lecture Notes in Computer Science, 2014, 8491: 783-792.

[80] Ye C, Wang H Z, Gao H, et al. Truth discovery based on crowdsourcing[J]. Lecture Notes in Computer Science, 2014, 8485: 453-458.

[81] Tong Y X, Cao C C, Zhang C J, et al. CrowdCleaner: Data cleaning for multi-version data on the web via crowdsourcing[C]. 2014 IEEE 30th International Conference on Data Engineering, Moscow, Russia, March 31-April 4, 2014.

[82] Lofi C, El Maarry K, Balke W T. Skyline queries over incomplete data-error models for focused crowd-sourcing[M]//Conceptual Modeling. Berlin: Springer, 2013: 298-312.

[83] Lofi C, El Maarry K, Balke W T. Skyline queries in crowd-enabled databases[C]. The 16th International Conference on Extending Database Technology, Genoa, Italy, March 18-22, 2013.

[84] Li Z X, Sharaf M A, Sitbon L, et al. A web-based approach to data imputation[J]. World Wide Web, 2014, 17(5): 873-897 .

[85] Chen Y C, Li J Z, Luo J Z. ITCI: An information theory based classification algorithm for incomplete data[J]. Lecture Notes in Computer Science, 2014, 8485: 167-179.

（责任编辑　王媛媛）