基于数据依赖的数据修复研究进展

胡艳丽 张维明 罗旭辉 肖卫东 汤大权

(国防科学技术大学 C⁴ISR 技术国防科技重点实验室 长沙 410073)

摘 要 介绍了数据依赖理论及如何基于数据依赖修复不一致数据,提高数据质量。首先介绍了数据依赖理论;给出了数据修复的语义假设及对应的修复操作;总结了基于数据依赖修复不一致数据的方法;最后讨论了基于数据依赖修复不一致数据的未来发展方向。

关键词 数据依赖,不一致数据,修复,数据清洗,数据质量

中图法分类号 TP311.131 **文献标识码** A

Dependencies Theory and its Application for Repairing Inconsistent Data

HU Yan-li ZHANG Wei-ming LUO Xu-hui XIAO Wei-dong TANG Da-quan

(C⁴ISR Technology National Defense Science and Technology Key LAB, National University of Defense Technology, Changsha 410073, China)

Abstract The theory of data dependencies was introduced and its application for repairing inconsistent data was discussed. Firstly, dependencies were analysized and the semantics for repairing inconsistent data were introduced. Secondly, the methods for repairing inconsistent data were analyzed. Finally, the future research topics related to repairing inconsistent data were also discussed.

Keywords Data dependency, Inconsistent data, Repair, Data cleaning, Data quality

1 引言

信息技术的普及和发展使得数据成为信息时代最重要的战略资源之一,为科学研究以及辅助政治、经济、商业以及军事等领域的决策提供重要依据。可靠、准确的数据是进行正确决策的基础,而不一致数据是数据集成与数据交换面临的普遍问题,因此有效修复不一致数据是深入开展数据库技术及其应用的迫切要求。

随着数据量的增加和数据应用范围的扩大,数据采集、存储、加工、使用等过程中由于多方面的原因导致数据中出现的错误、偏差和不一致等问题日益严重,造成的决策失误、经济损失不容忽视。据调查显示,2001 年 75%的美国公司都存在数据质量问题,超过三分之一的公司由于数据质量问题不得不推迟或放弃新产品的研发、上市,约 30%的公司由于数据错误、缺失难以正常收缴业务收入[1]。Data Warehousing Institute 在 2002 年的调查中估计,由于数据质量问题导致美国企业每年的经济损失高达 6000 亿美元[2],仅零售业的商品价格错误信息每年就导致 25 亿美元的损失[3]。广泛存在的脏数据使得数据质量低劣,从而引起决策失误或偏差,极大影响数据的应用价值和效益,制约并削弱数据对战略决策的辅助作用。

关系数据库通过数据依赖定义实体和实体间的联系。当属性取值不满足指定的语义约束时就会导致数据不一致问题。如何修复不一致数据成为数据库领域研究的热点和难点。

本文分析总结了数据依赖理论和基于数据依赖修复不一致数据方法的研究情况。首先介绍了数据依赖;其次给出了数据修复的语义基础;然后分析了基于数据依赖修复不一致数据的方法;最后探讨了基于数据依赖修复不一致数据的未来研究方向。

2 数据依赖理论

基于数据依赖修复不一致数据的基础是对数据的语义一致性进行建模。数据依赖(data dependency)定义了关系数据库属性间的语义约束,是现实世界数据间内在联系在数据库中的抽象反映[4,5]。

自从 Codd 提出函数依赖(Functional Dependencies, FDs)以来[6],研究者陆续提出了包含依赖(inclusion dependency, IND)、多值依赖(Multivalued Dependencies)、连接依赖(Join Dependency)等数据依赖类型[7,8]。上述经典关系理论的数据依赖均可采用统一的一阶逻辑语句描述[9]:

$$\forall x_1 \cdots \forall x_n [\varphi(x_1, \cdots, x_n) \rightarrow \exists z_1 \cdots \exists z_k \psi(y_1, \cdots, y_m)] \quad (1)$$

其中,$\{x_1, \cdots, x_n\}$,$\{y_1, \cdots, y_m\}$和$\{z_1, \cdots, z_k\}$是属性变量集合,且$\{z_1, \cdots, z_k\} = \{y_1, \cdots, y_m\} - \{x_1, \cdots, x_n\}$;$\varphi$是一阶逻辑语句的前提,为空集或关系谓词、等式原子公式的合取,ψ是

到稿日期:2009-06-04 返修日期:2009-09-01

胡艳丽(1979-),女,博士研究生,主要研究领域为信息管理、数据库技术,E-mail:smilelife1979@163.com;张维明(1962-),男,教授,博士生导师,主要研究领域为信息管理、智能决策;罗旭辉(1980-),男,博士研究生,主要研究领域为信息管理、数据库技术;肖卫东(1968-),男,教授,博士生导师,主要研究领域为信息管理、指挥决策技术;汤大权(1971-),男,副教授,硕士生导师,主要研究领域为数据库、指挥决策技术。

一阶逻辑语句的结论，为关系谓词或等式原子公式的合取。关系谓词的形式为 $R(w_1,\cdots,w_l)$，等式原子公式的形式为 $w=w'$，其中 R 是关系模式，$w,w',w_1,\cdots,w_l$ 是属性变量。

对式(1)进行变形得到几种常用的依赖，如完全依赖(full dependency)[10]、元组产生依赖(tuple-generating dependency, tgd)和等值产生依赖(equality-generating dependency, egd)[11]等。

经典关系理论中的数据依赖只涉及关系谓词及表示属性取值是否相等的等式原子公式，无法进一步表达对属性取值的约束。而实际应用中普遍存在这样的约束，例如销量超过50万份、100万份和1000万份的唱片分别为金唱片、白金唱片和钻石唱片；我国男性已婚人员的年龄应大于23岁，女性已婚人员的年龄应大于21岁等。上述语义约束与属性取值密切相关，如根据唱片销量取值范围决定唱片级别，婚姻状态为已婚的公民需要大于一定的年龄等。由于经典数据库理论中的数据依赖无法表达这类语义约束，导致基于上述数据依赖无法修复这类破坏取值约束的不一致数据。因此需要具有更强表达能力的数据依赖表达这类语义约束。

否定约束(denial constraints)[12]通过不等式增强描述语义冲突的不一致数据，形式为：

$$\forall \bar{x}_1,\cdots,\bar{x}_m\rightarrow[P_1(\bar{x}_1)\wedge\cdots\wedge P_m(\bar{x}_m)\wedge\varphi(\bar{x}_1,\cdots,\bar{x}_m)] \tag{2}$$

其中，$\bar{x}_1,\cdots,\bar{x}_m$ 表示属性变量序列，$P_1,\cdots,P_m$ 表示关系谓词，φ 是内置谓词合取公式。

约束产生依赖(constraint-generating dependencies, CGDs)[13,14]是对等值产生依赖的扩展，通过内置谓词原子公式的合取式表达属性取值约束。形式为：

$$\forall \bar{x}(R_1(\bar{x})\wedge\cdots\wedge R_k(\bar{x})\wedge\xi(\bar{x})\rightarrow\xi'(\bar{x})) \tag{3}$$

其中，R_i 是关系谓词，ξ,ξ' 是具有内置谓词的属性变量约束。

全称约束(universal constraints)[15]是对完全依赖的扩展，允许采用不等式的析取式指定语义约束，形式为：

$$\forall x_1\cdots\forall x_n[\varphi(x_1,\cdots,x_n)\rightarrow\phi(x_1,\cdots,x_n)] \tag{4}$$

其中，φ 是关系谓词的合取，ϕ 是关系谓词和内置谓词(built-in predicates)原子公式的析取，其中内置谓词原子公式形式为 $x\theta y$，x 和 y 是属性变量或全域内的常量，内置谓词 $\theta\in\{\leqslant,<,=,\neq,>,\geqslant\}$。

作为最重要的数据依赖，如何增强函数依赖的表达能力受到研究人员的重视。受限函数依赖(constrained functional dependencies, CFDs)[16,17]通过一组约束限定函数依赖成立的前提，指定函数依赖只在满足约束的数据子集上成立，形式为：

$$\xi\rightarrow(Z\rightarrow W) \tag{5}$$

其中，ξ 是具有内置谓词的属性变量约束的析取，函数依赖($Z\rightarrow W$)只适用于满足条件 ξ 的数据子集。

受限函数依赖只能够表达函数依赖成立的条件，文献[18]进一步提出受限元组产生依赖(constrained tuple-generating dependencies, CTGDs)，通过不等式原子公式表达依赖所含任意属性的约束，形式为：

$$\forall \bar{x}(R_1(\bar{x})\wedge\cdots\wedge R_k(\bar{x})\wedge\xi(\bar{x})\rightarrow\exists\bar{y}(R'_1(\bar{x},\bar{y})\wedge\cdots\wedge R_k'(\bar{x},\bar{y})\wedge\xi'(\bar{x},\bar{y}))) \tag{6}$$

其中，$R_i,R_j'(i,j\in[1,k])$ 是关系谓词，ξ,ξ' 是具有内置谓词的属性变量约束。

英国爱丁堡大学樊文飞教授提出条件依赖[19-22]，通过等值约束增强函数依赖和包含依赖的表达能力，并对其性质进行了系统研究。

综上所述，经典关系理论中的数据依赖难以表达具有属性约束的语义关联，因此无法修复这类不一致数据。为此，研究人员提出增强传统数据依赖表达能力的多种依赖类型，为有效修复不一致数据提供了基础。

3 基于依赖修复不一致数据的语义假设

数据的一致性是信息系统对数据的基本要求[36-39]。在从现实世界向数据世界转换的过程中，完整性约束(integrity constraints)是基于现实世界客观事物及其关联定义的数据语义规则，用以维护数据库中数据与现实世界的一致性，破坏完整性约束的数据被称为是不一致的[40,42]。在实际的数据库应用中，由于各种原因，数据常常违反完整性约束，导致存在大量不一致数据。从理论上分析，造成这一问题的原因在于：

(1)数据建模过程中对客观事物及其关联缺乏深入理解、数据库设计阶段数据模式设计不规范、数据模型表达能力不足以及缺乏有效的完整性约束，使得数据库缺乏对数据的限制，无法检测到不符合语义约束的数据记录，从而导致单数据源中存在不一致数据[45,46]。

(2)当多数据源集成时，由于独立设计的各数据源具有异构的数据语义，导致即使每个数据源中的数据是一致的，但集成时却存在大量不一致数据的现象[43]。

由于不一致数据破坏了客观事物及其关联，无法正确反映现实世界客观事物的真实状态，导致根据不一致数据会得出错误的结果和结论，因而极大降低了数据的应用价值，制约并削弱数据对决策的辅助作用[41]。不一致数据成为影响数据应用的严重瓶颈。有效检测和纠正不一致数据成为数据库应用和信息技术发展的迫切要求。然而，目前对数据质量问题的研究主要集中在填充缺失值[48]、平滑噪声[47]以及识别消除相似重复记录[44,49-51]等方面，对于纠正破坏完整性约束的不一致数据缺乏深入研究。实际应用中主要通过人工纠正不一致数据，根据人工或底层程序的辅助检测不一致数据，然后由数据生产者或提供者纠正不一致[52]。随着实际应用日益复杂、数据库规模不断增大以及对数据处理时效性要求的提高，通过人工检测和纠正不一致数据的方法暴露出自身的缺陷：

一是人工发现和纠正不一致数据需要耗费大量的人力和时间。例如，中等规模的统计数据需要12名统计工作者花费3个月的时间进行编辑和修补，以保证不存在不一致数据[31]。完全诉诸于人工无法满足及时高效纠正大量不一致数据的需求。

二是人工方法受经验和对领域熟悉程度的限制。完整性约束通常涉及多个属性间的联系，纠正破坏完整性约束的不一致数据使其满足语义约束的方式并不唯一。由于人的知识、经验和分析能力的差异和限制，人工发现和纠正不一致数据受人员知识和经验的限制较大。实际应用复杂程度的不断增加也给人工对不一致数据的认知带来极大挑战。因此人工纠正不一致数据已不能适应复杂应用和海量数据的需求。

面对提高数据质量的迫切需求和现有方法处理不一致数据的局限性，Arenas 等人于1999年在 ACM SIGMOD Con-

ference on Principles of DataBase Systems(PODs)提出非一致性数据库(inconsistent database)和数据修复(data repairing)的概念。所谓修复,是指给定数据库上定义的一组完整性约束Σ,当数据库实例I不满足Σ时,通过修复操作获得满足Σ的一致修复实例I',即$I' \vdash \Sigma$,且I'与I差距最小[23]。

基于数据依赖的数据修复方法根据数据库上成立的一组数据依赖,检测破坏数据依赖的不一致数据,采用相应的修复操作处理不一致数据,获得满足数据依赖且与初始不一致实例最接近的修复实例。基于数据依赖修复不一致数据建立在对非一致性数据库(inconsistent database)的语义假设基础上,修复不一致数据所采取的操作取决于数据依赖类型和语义假设。

关系数据库通常采用基于封闭世界假设(closed world assumption,CWA)的精确释义(exact interpretation),即数据库实例包含的数据是可靠且完全的[24]。然而基于封闭世界假设的精确释义不适用于包含不一致数据的非一致性数据库,根据数据状态可将非一致性数据库的语义假设分为可靠语义(sound semantics)、完全语义(complete semantics)或松散语义(loose semantics)等。

可靠语义假设非一致性数据库包含的数据是可靠的,当存在不一致数据时只能通过插入新的元组进行修复,由此得到的修复实例是原实例的超集(superset)[25,26]。完全语义假设非一致性数据库包含的数据是完全的,当存在不一致数据时通过删除导致不一致的部分数据进行修复,因此得到的修复实例是原实例的子集(subset)[27]。通常来讲,可靠语义和完全语义过于严格,因此研究人员进一步提出松散语义,即认为非一致性数据库中的数据既非可靠也非完全的,因此可以同时通过插入和删除元组修复不一致数据[28]。

从理论上讲,插入或删除元组的操作可用于任意类型的数据依赖。但元组粒度的修复操作具有粒度较大的缺陷,而且会导致额外的问题,例如删除包含不一致数据的元组的同时丢失正确数据。因此文献[29]提出数据项粒度的修复操作,通过修改不一致数据项的取值修复不一致数据,使其满足数据依赖。

4 基于数据依赖修复不一致数据

数据依赖的表达能力和性质直接决定能否有效检测破坏语义约束的不一致数据,同时影响如何修复不一致数据。面向修复的数据依赖应该具有下列特性:

(1)具有较强的表达能力,不仅能够表示模式层的语义约束,同时能够表示应用相关的、实例层的语义约束;

(2)为修复不一致数据提供辅助信息,从而有效计算满足数据依赖的一致修复实例。

对破坏数据依赖的不一致数据进行修复,得到一致修复实例的过程由一系列修复操作序列构成,每一个修复操作包括如何选择待修复的不一致数据,以及如何修改选中的待修复数据项。破坏数据依赖的不一致数据通常存在多种修复方式,并且修复可能导致新的不一致。上述问题使得基于数据依赖修复不一致数据是一个非常困难的问题。对于目前用于修复不一致数据的数据依赖类型,研究表明求解与初始实例最接近的一致修复实例均是NP难的。因此数据修复方法主要研究如何设计有效修复不一致数据并且可在多项式时间内完成启发式算法或近似算法来求解修复实例。

文献[30]针对字符型或串型数据,提出基于函数依赖的启发式修复算法GREEDY-REPAIR,通过修改破坏函数依赖的元组在依赖右部属性上的取值为相等进行修复。主要思想是建立等价类,根据函数依赖将取值应该相等的数据项归并为同一等价类,为等价类统一赋值作为其所含数据项的修改值,从而通过修改得到满足函数依赖的一致数据项。每次修复首先穷举所有可以归并的等价类,计算归并前后等价类所含数据项取值变化产生的加权距离之和,然后选择距离最小的归并等价类并确定其赋值,重复上述过程直至得到一致的修复实例。函数依赖表达能力的局限性使得修复不一致数据的效果欠佳。

针对修复效果不理想的问题,文献[31]对文献[30]的方法进行扩展,提出基于条件函数依赖修复不一致数据的启发式算法BATCHREPAIR。与文献[30]的不同之处在于,因为条件函数依赖指定了属性取值的等值约束,修复时当不一致数据项无法同时满足不同条件函数依赖对同一属性的等值约束时,需要修改元组在依赖左部属性上的取值进行修复。实验表明,基于条件函数依赖修复不一致数据较函数依赖效果好。

以利用图论方法为目标,研究人员用图来表示非一致性数据库包含不一致数据的情况,进而将修复不一致数据转化为求解相应的图论问题。

文献[12]首先提出冲突图(conflict graph)的概念,研究通过删除元组修复破坏函数依赖的不一致数据的方法。冲突图以实例中的元组作为顶点,破坏函数依赖的任意两条元组之间的连线作为一条边,表示实例破坏函数依赖的情况。在此基础上,将计算修复实例转化为求解冲突图的最大独立集(maximal independent set)问题,即冲突图中彼此间不存在边的最大顶点子集,对应于实例中不破坏任意函数依赖的最大元组子集。文献[27]进一步提出冲突超图(conflict hypergraph)的概念,研究通过删除元组修复破坏否定约束的不一致数据的方法。冲突超图以元组作为顶点,破坏一条否定约束的一组元组作为一条超边,得到实例破坏否定约束的超图表示。与文献[12]类似,将修复不一致数据转化为求解超图的最大独立集问题,由此得到不包含于任何超边的顶点集,对应于的元组子集就是通过删除不包含于最大独立集的不一致元组得到的满足否定约束的一致修复实例。由于删除元组会导致丢失正确数据,因此上述方法有待改进。

文献[32,33]基于标记的冲突超图(labeled conflict hypergraph)研究通过修改操作修复破坏一类特殊否定约束的不一致数据的方法。与文献[27]类似,标记冲突超图采用元组作为顶点,用被破坏的否定约束标记破坏它的一组元组构成的超边。然后通过修改每条超边中元组在不一致数据项上的取值,求解使得超边所含元组满足对应否定约束的局部修复(local repair)。所有局部修复构成原实例的候选修复集,在此基础上,文献[32,33]将修复不一致数据转化为求解超图的最小加权集合覆盖优化问题(minimum weighted set cover optimization problem,MWSCP),得到覆盖所有超边的元组集,对应的局部修复作为原实例的修复。文献[32]采用的否定约束具有特殊性,对破坏这类否定约束的元组进行修复不会导致新的不一致,因此可直接根据局部修复得到一致的修

复实例，而其它数据依赖不具有这一特性。

文献[29,34]研究通过将数据项取值修改为变量修复不一致数据的方法。以数据项作为顶点，破坏矛盾生成依赖(contradiction-generating dependency)的一组元组作为一条超边，文献[29,34]将修复不一致数据转化为求解超图的最小击中集(minimal hitting sets)，包含于最小击中集的数据项对应于修复不一致数据所要修改的最小数据项集。在此基础上，将最小击中集中的数据项取值替换为互异的变量，由此所得的造型表(tableau)对应于不一致实例的修复模板，根据依赖仔细选择变量的赋值可得一致的修复实例。引入变量的方法将确定待修改数据项和如何修改数据项的问题分离开，只完成了数据修复的第一步。

文献[35]提出通过修改操作修复破坏函数依赖不一致数据的近似修复算法 FINDVREPAIR。借鉴文献[29,32]的思想，文献[35]以数据项作为顶点，破坏函数依赖(contradiction-generating dependency)的任意两条元组为单位构造超边，修改破坏函数依赖的不一致数据至少要修改每条超边中的一个数据项。在此基础上，文献[35]将选择待修改数据项转化为求解超图的最小加权覆盖，至少要修改最小加权覆盖所含的不一致数据项才可能获得一致的修复实例。文献[35]引入变量修改不一致数据项的取值：若破坏函数依赖 $f_1: X \rightarrow A$ 的元组$\{t_1, t_2\}$中 $t_1[A]$和 $t_2[A]$不同时包含于最小加权覆盖，不妨设 $t_1[A]$属于最小加权覆盖，$t_2[A]$不属于最小加权覆盖，则根据 $t_2[A]$修改 $t_1[A]$的取值；否则，FINDVREPAIR 算法将最小加权覆盖中不对应上述情况的数据项全部修改为互异变量。最小加权覆盖修改完毕后，如果仍存在不一致或上述修改导致了新的不一致，FINDVREPAIR 算法进一步选择包含于最小加权覆盖且破坏依赖数最多的不一致数据项及其基于函数依赖的左部覆盖属性集全部修改为互异变量，直至得到一致的实例。这里的一致是指实例中非变量数据项不破坏任何函数依赖，因为文献[35]假设变量赋值不等于实例中任何常数。FINDVREPAIR 算法的近似比与数据库上成立的函数依赖集合 Σ 相关。当 Σ 固定时，FINDVREPAIR 算法具有确定的近似比 $(n+2)*(2m-1)$，其中 n 是 Σ 中具有相同右部的函数依赖左部属性集最小覆盖的规模，至多为 $O(|\Sigma|)$，m 是 Σ 中函数依赖包含的最大属性数，至多为 $O(|sort(R)|)$，其中 $sort(R)$是模式 R 包含的属性规模。因此 FINDVREPAIR 算法的近似比与 Σ 中函数依赖的规模和相互关系密切相关。对于规模较大且具有复杂函数依赖的数据集，FINDVREPAIR 算法得到的修复实例与原实例的差距非常大。同时，文献[33]选择包含于超边的不一致数据项时未考虑数据可信度的区别；引入变量作为修改值虽然简化了修复，但大量的变量和对变量赋值的假设使得 FINDVREPAIR 算法难以满足实际应用对修复的需求。

结束语 随着人们对数据质量问题认识的逐步深入，如何提高数据质量逐渐成为数据库领域的研究热点。目前，关于数据质量的研究刚刚开始，对于提高数据质量的理论、技术和方法缺乏深入、有效的研究。数据清洗技术作为提高数据质量的重要方法，长期以来一直只是数据仓库、数据库中知识发现等领域数据预处理过程的一个步骤，没有得到足够的重视。已有的清洗方法也主要是针对空缺值、相似重复记录等数据问题，对于如何解决数据不一致缺乏有效的处理方法。

数据依赖通过对关系数据语义完整性的形式化建模，定义一致的数据应该满足的模式，因此可以用于数据清洗，对不一致数据进行修复，获得一致的修复实例，为实现数据不一致性的自动检测、减少清洗过程的人工干预、提高大规模数据清洗的效率提供了基础。

目前基于数据依赖修复不一致数据的研究刚刚开始，有很多问题亟待进一步深入探讨。首先，数据依赖的表达能力和计算复杂性相互制约，表达能力强的依赖虽然能够表示更丰富的语义，但其计算复杂性也相应提高，难于有效修复不一致数据。因此设计实现具有合适表达能力和计算复杂性的数据依赖是亟待解决的问题。其次，基于数据依赖修复不一致数据的核心是如何准确定位导致不一致的数据并对其进行正确修复。因此，有效的数据修复方法是基于数据依赖修复不一致数据的研究重点。此外，如何基于数据依赖实现自动检测和修复不一致数据的工具是修复不一致性数据的关键，将在数据清洗及提高数据质量的研究中发挥重要作用。

参 考 文 献

[1] Benge J, Jordan G M W, Smith P, et al. Global Data Management Survey: The new economy is the data economy[R]. Coopers, Price Waterhouse, 2001

[2] Eckerson W W. Data Quality and the bottom line: achieving business success through a commitment to highquality data. Data Warehousing Institute, 2002

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

[4] Mullins C S. Database Administration: The Complete Guide to Practices and Procedures[M]. Addison Wesley

[5] 李建中，王珊. 数据库系统原理[M]. 北京：电子工业出版社，1998

[6] Codd E F. Relational Completeness of Data Base Sublanguages [C]// Rustin R J, ed. Data Base Systems, Courant Computer Science Symposia. Vol. 6, Englewood Cliffs, N. J: PrenticeHall, 1972

[7] Korth, A. S. a. H. F. Database System Concepts[M]. McGraw-Hill, 1986

[8] Ullman J D. Principles of Database Systems[M]. Computer Science Press, 1982

[9] Abiteboul S, Vianu R H V. Foundations of Databases[M]. Addison Wesley, 1995

[10] Beeri C M Y V. The implication problem for data dependencies [C]// Proc. Intl. Conf. on Algorithms, Languages and Programming. Berlin: Springer-Verlag, 1981

[11] Beeri C M Y V. A proof procedure for data dependencies[J]. Journal of ACM, 1984, 31(4): 718-741

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

[13] Baudinet M J C, Wolper P. Constraint-Generating Dependencies [J]. JCSS, 1999, 59(1): 94-115

[14] Baudinet M J C P W. Constraint-Generating Dependencies [C] // Proc. 5th International Conference on Database Theory. 1995: 322-337

[15] Staworko S. Declarative inconsistency handling in relational and semi-structured databases [D]. the State University of New York at Buffalo. 2007

[16] Maher M J. Constrained dependencies[J]. Theoretical Computer Science,1997,173(1):113-149

[17] Maher M J. Constrained Dependencies [C] // Proc. Conj. on Principles and Practice oj Constraint Programming. 1995:170-185

[18] Maher M J, Srivastava D. Chasing Constrained Tuple-Generating Dependencies [C]//PODS. 1996

[19] Fan Wenfei, Jia Xibei, Kementsietsidis A, et al. Conditional Functional Dependencies for Capturing Data Inconsistencies[J]. ACM Transactions on Database Systems(TODS),2008,33(2)

[20] Bohannon P, Fan W, Geerts F, et al. Conditional Functional Dependencies for Data Cleaning [C]//The 23rd International Conference on Database Engineering (ICDE) (the best paper award). 2007

[21] Bravo L W F, Ma Shuai. Extending Dependencies with Conditions [C]// The 33rd International Conference on Very Large Data Bases(VLDB). 2007

[22] Bravo L W F, Geerts F, Ma Shuai. Increasing the Expressivity of Conditional Functional Dependencies without Extra Complexity [C]// The 24th International Conference on Database Engineering(ICDE). 2008

[23] Arenans M, Bertossi L E, Chomicki J. Consistent query answers in inconsistent databases[C] // Proceedings of the 18th ACM Symposium on Principles of Database Systems. ACM Press, 1999

[24] Reiter R. Logic and Databases: On closed world databases[M]. New York: Plenum Press, 1978

[25] Cali A, Lembo D, Rosati R. Query rewriting and answering under constraints in data integration systems [C]// Proceedings of the International Joint Conference on Artificial Intelligence. 2003:16-21

[26] Cali A, Lembo D, Rosati R. On the decidability and complexity of query answering over inconsistent and incomplete databases [C]// Proceedings of the Symposium on Principles of Database Systems(PODS). 2003:260-271

[27] Chomicki J A M J. Minimal-change integrity maintenance using tuple deletions[J]. Inform. Comput. ,2004,197(1/2):90-121

[28] Lin J A O M. Merging databases under constraints[J]. Int. J. of Cooperative Information Systems,1998,7(1):55-76

[29] Wijsen J. Database repairing using updates[J]. ACM Trans. Database. System,2005,30(3):722-768

[30] Bohannon P, et al. A Cost-Based Model and Effective Heuristic for Repairing Constraints by Value Modification [C] // SIGMOD. Baltimore, Maryland, USA, 2005

[31] Cong G W F, Geerts F, Jia Xibei, et al. Improving Data Quality: Consistency and Accuracy[C] // VLDB 2007. Vienna, Austria, 2007

[32] Lopatenko A, Bravo L. Efficient approximation algorithms for repairing inconsistent databases[C] // Proceedings of the International Conference on Data Engineering(ICDE'03). 2003

[33] Bertossi L E, et al. Complexity and approximation of fixing numerical attributes in databases under integrity constraints[C]// DBPL. 2005

[34] Wijsen J. Making More Out of an Inconsistent Database[C]// East-European Conference on Advances in Databases and Information Systems(ADBIS). Springer, 2004

[35] Kolahi S, Lakshmanan L V S. On Approximating Optimum Repairs for Functional Dependency Violations[C] // ICDT 2009. Saint Petersburg, Russia, 2009

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

[37] Ponniah P. Data Warehousing Fundamentals[M]. Wiley

[38] Batini C, Scannapieco M. Data Quality: Concepts, Methodologies and Techniques[M]. New York, USA: Springer, 2006

[39] Lionel A, Galway C H H. Data Quality Problems in Army Logistics[C]// RAND. 1996

[40] Date C J. 数据库系统导论[M]. Addison Wesley, 2007

[41] 陈卫东. 数据质量模型及关系代数运算下质量传递理论与方法研究 [D]. 长沙: 国防科学技术大学, 2007

[42] Ullman J D, Widom J. 数据库系统基础教程[M]. 北京: 机械工业出版社, 2003

[43] Lenzerini M. Data Integration: A Theoretical Perspective [C]// pods'02. 2002

[44] A D, P D, A H. Reconciling Schemas of Disparate Data Sources: A Machine-Learning Approach[J]. ACM SIGMOD, 2001:509-520

[45] Silberschatz A, Korth H F. Database System Concepts [M]. McGraw-Hill, 1986

[46] Ullman J D. Principles of Database Systems[M]. Computer Science Press, 1982

[47] Han J, Kamber M. 数据挖掘: 概念与技术[M]. 北京: 机械工业出版社, 2007

[48] M K W. Rough set approach to incomplete information systems [J]. Information Sciences, 1998, 11(4):39-49

[49] Haidarian H, Shahri A A Z B. Data Mining for Removing Fuzzy Duplicates Using Fuzzy Inference [C] // Annual Conference of the North American Fuzzy Information Processing Society. 2004:419-424

[50] Guyon I M N, Vapnik V. Discovering Information Patterns and Data Cleaning [C]// H. M. R Agrawal, R Srikant, H Toivonen, eds. Advances in Knowledge Discovery and Data Mining. USA, 1996:181-203

[51] Hernandez M A, Stolfo S. Real-world data is dirty: Data cleansing and the merge/purge problem[J]. Data Min. Knowl. Discov. ,1998,2(1):9-37

[52] Heiko Muller J C F. Problems, Methods and Challenges in Comprehensive Data Cleaning[OL]. http://www.Dbis.informatik.hu-Berlin.de/fileadmin/research/papers/techreports/2003_hub_ib_164-mueller.pdf, 2003-09-10/2006-11-21

(上接第 4 页)

[21] Hassin R, Levin A. A better - than - greedy approximation algorithm for the minimum set cover problem[J]. SIAM J. Comput. ,2005,35(1):189-200

[22] Downey R G, Fellows M R, Stege U. Parameterized complexity: A framework for systematically confronting computational intractability[J]. DIMACS Series in Discrete Mathematics and Theoretical Computer Science, 1999, 49:49-99

[23] Fernau H. A top-down approach to search-trees: Improved algorithmics for 3-Hitting Set[R]. ECCC. 2004

[25] Abu - Khzam F N. Kernelization Algorithms for d - Hitting Set Problems[J]. LNCS page, 2007, 4619:434-445

[26] Fernau H. Parameterized Algorithms for d - Hitting Set : the Weighted Case[R]. Informatik/Mathematik No. 08-6. July 2008