软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2016,27(7):1626−1644 [doi: 10.13328/j.cnki.jos.005040]

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

数据质量多种性质的关联关系研究*

丁小欧, 王宏志, 张笑影, 李建中, 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)

通信作者: 王宏志, E-mail: wangzh@hit.edu.cn

摘 要: 信息化时代数据海量增长的同时,用户需要利用多种指标从不同性质角度对数据质量进行评价和改善.但在目前数据质量管理过程中,影响数据可用性的多种重要因素并非完全孤立,在评估机制和指导数据清洗规则时,彼此会发生关联.研究了在实际信息系统中适用的综合性数据质量评估方法,将文献所提出以及在实际的信息系统中常用的数据质量性质指标按其定义与性质进行了归纳总结,提出了基于性质的数据质量综合评估框架.之后针对影响数据可用性的 4 个重要性质:精确性、完整性、一致性以及时效性整理出在数据集合上的操作方法,并逐一介绍其违反模式的定义,随后给出其具体关系证明,进而确定数据质量多维关联关系评估策略,并通过实验验证了该策略的有效性.

关键词: 数据质量;数据质量性质;多性质关系;数据清洗;数据管理

中图法分类号: TP311

中文引用格式: 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

英文引用格式: 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). http://www.jos.org.cn/1000-9825/5040.htm

Association Relationships Study of Multi-Dimensional Data Quality

DING Xiao-Ou, WANG Hong-Zhi, ZHANG Xiao-Ying, LI Jian-Zhong, GAO Hong

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

Abstract: Recently, with the rapid growth of data quantity, users are using a variety of indicators to evaluate and improve the quality of data from different dimensions. During the course of data quality management, it is found that many important factors that influence the data availability are not completely isolated. In the evaluation mechanism which can guide data cleaning rules, these dimensions may be associated with each other. In this paper, serveral data quality dimensions researched in the literature as well as being used in the real information system are discussed, and accordingly the definition and properties of the dimensions are summarized. In addition, a multi-dimensional data quality assessment framework is proposed. According to the four important properties of data availability: Accuracy, completeness, consistency and currency, the operation method and the relationships among them on the data set are constructed. Finally, a multi-dimensional data quality accessment strategy is created. The effctiveness of the proposed strategy is verified by experiments.

Key words: data quality; data quality dimension; relationship among dimensions; data cleaning; data management

* 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, 61133002); 黑龙江省留学回国人员基金(LC2016026)

Foundation item: National Program on Key Basic ResearchProject of China (973) (2012CB316200); National Natural Science Foundation of China (U1509216, 61472099, 61133002); Scientific Research Foundation for the Returned Overseas Chinese Scholars of Heilongjiang Provience (LC2016026)

收稿时间: 2015-10-10; 修改时间: 2016-01-12; 采用时间: 2016-02-22; jos 在线出版时间: 2016-03-22

CNKI 网络优先出版: 2016-03-22 13:23:35, http://www.cnki.net/kcms/detail/11.2560.TP.20160322.1323.008.html

数据作为一种重要资源,在诸多学科领域中都有重要的影响,数据的质量对过程管理、决策支持、合作需求分析等活动均有重要的引导作用,高质量的数据能够提供可靠、准确的服务.例如,美国一家小型科技创新公司 Farecast 公司 2012 年对上亿条飞行记录和国内航班票价进行分析处理,其对机票最佳购买时机的预测准确度高达 75%.使用该工具来购买机票的旅客,平均每张机票节省 50 美元.更值得注意的是,这些对于飞行记录的分析亦可有效地预测和评估疾病的流行趋势[1].然而,当人们获取和利用的数据量飞速增加时,由于容错标准不完善、存储数据格式不一致、信息来源可靠性低、数据更新周期过长等原因,造成数据的错误率和混乱程度的增加.例如,I.B.M.、T.J. Watson Labs、International Business Machines 等词在实际情况中都可以用来指代 IBM.在整合不同数据源对 IBM 的记录时,就会因记录格式不一致而导致计算机在处理判断中出现错误.在数据工程中所用到的数据其质量不够优质,那么很可能会给诸多领域带来严重的负面影响.

在数据质量引起关注和重视的同时,研究人员采用直接观测、社会调查、理论推导等方式对数据质量问题展开研究,并分析得出几十种[2,3]主要的数据质量性质(即 data quality dimensions),以及在实际信息系统应用的上百种[4]数据质量的性质.对数据可用性影响较大的性质有:精确性、完整性、一致性、时效性和实体同一性[3].对于数据源,有可靠性、可信度等性质的分析;对数据的具体内容,有切题性、间接性、准确性等性质的分析;对数据的过程管理,有可达性、安全性等性质的分析.我们在研究中发现数据质量性质间存在着关联关系.例如,对“过期”数据的修复结果,可能会引起一致性方面的问题;有些错误数据的产生原因可能正是由于信息失去了时效性或者由于记录字段的不一致;对于存在较多缺失部分的数据进行填充,其填充后的可信度和准确度需要进一步分析.

目前,对于数据质量性质之间的关联关系的研究主要有以下几个技术挑战.

(1) 问题的研究范围广,研究边界难以确定.开展研究以来,研究人员对数据质量的性质进行了多角度的划分.数据的多样性和复杂性也间接导致了数据质量性质问题的多样性和该问题研究的复杂性.

(2) 数据质量性质的指标计算、获取难度大.对于数据质量在不同性质上的满足程度、违反情况和修复处理的理论方法的研究程度不尽相同,难以得到有效、合理的计算方法.

(3) 多种数据质量性质量化统一的难度较大.不同数据质量性质的属性不同,所反映和评判的数据质量的特点也有所不同,这给数据质量性质的统一量化提出难题.

(4) 数据质量性质关联关系的可靠性和有效性分析难度大.数据集合在各数据质量性质上存在的问题繁多,对于其间的关联关系仍然缺乏理论认识.

目前的研究结果仍存在一些不足,对影响数据质量的性质分析得不够全面和深入,未能对各种性质的满足或违反情况进行明确的定义说明.此外,对于几种重要性质也只是对部分问题进行了分析处理[6−9],未能全面说明这几个数据质量性质之间的关联.基于此,本文的研究目的在于找出多种数据质量性质间的明确的关联关系,对数据质量进行更为全面的描述,并避免数据质量管理过程中的重复性工作.本文的主要贡献如下.

(1) 对文献中和实际应用中的数据质量性质进行了深入的统计与分析,提出基于性质的数据质量综合评估框架;

(2) 对数据可用性有重要影响的 4 种数据质量性质为精确性、完整性、一致性、时效性,总结了这 4 种性质上的数据的错误类型和违反模式,并对其进行统一的形式化定义,首次提出对于数据质量多种性质的关联关系研究,理论证明了上述 4 种性质在数据修复(即提高数据质量)的过程中的具体相关关系;

(3) 针对存在多种混合错误类型的数据,提出一种高效、合理的数据清洗和修复策略,并通过实验验证了其有效性、合理性.

本文第 1 节分析数据质量性质的研究现状.第 2 节提出基于性质的数据质量综合评估框架,并对其中 4 个主要部分:时效性、一致性、精确性、完整性的研究范围和错误类型进行分析.第 3 节理论定义并证明上述 4 种性质间的关联关系,并初步提出混合型数据修复策略.第 4 节通过实验验证上述修复策略的有效性和合理性.第 5 节是对本文工作的总结和对未来研究方向和目标的展望.

1 相关研究综述

数据质量(data quality)问题所包括的内容十分广泛,文献[2,5,10]均对数据质量性质进行了详细的统计调查,提出了数据类型、数据质量问题的分类以及数据质量性质及其定义.从 20 世纪 90 年代起,研究表明,用户对数据质量的需求已经远远高于单一准确性的需求,准确性也不再是评估数据质量优劣的唯一指标[4,5,9].文献[5]中针对数据质量性质的潜在可能进行了大量的深入调查工作,通过对 118 种性质的调研,针对数据用户的需求,最终确定了可靠性、切题性、可访问性等 20 种常用数据质量性质.文献[6–9]深入研究了几种性质上(完整性、同一性、一致性等)的判定问题以及在数据清洗中的具体实现策略与算法.文献[10–13]中说明了数据质量对于商业、政府中管理人员的决策操作过程的重要影响.文献[12]提出了在评估数据质量优劣的过程中,解决来自定义、测量、分析和改进等部分的数据质量问题的必要性.文献[14]对几种性质的影响程度、可靠性等方面建立了简单的数据质量分析模型,举例说明其在工业应用中用户需求分析方面应用的有效性.

尽管文献[2]中列举了 40 种数据质量性质,但并未明确它们之间的关联关系.而文献[11]也表明,数据质量的完整性、一致性、时效性等因素并不是完全孤立的,甚至有部分研究内容是有交集的,但并未明确指出这些数据质量性质的关联.因此,本文将把数据质量性质作为研究对象,明确其间的相关关系,以便对数据质量问题进行更加有效、合理的分析.

2 基于性质的数据质量综合评估框架

2.1 框架分析

2.1.1 框架描述

数据质量性质(data quality dimension),是信息归类和数据需求的一种特征或者部分信息片段[2,11],数据质量性质的确定为度量并管理数据(或信息)质量提供了有效的手段.基于文献[2]中数据质量性质的定义和评价方法,以及文献[4,11,14]对不同数据质量性质的归纳和建模分析,我们提出了如图 1 所示的数据质量性质的综合评估框架.根据对数据质量的影响能力强弱程度影响覆盖面广泛程度以及用户的公认重要程度的综合分析,在框架中可研究的数据质量性质分为核心性质与外围性质两部分,在 6 种核心性质中,时效性、一致性、完整性、精确性针对数据进行评估,可靠性与有效性模块将对以上这 4 个性质的评估结果进行分析和评估.

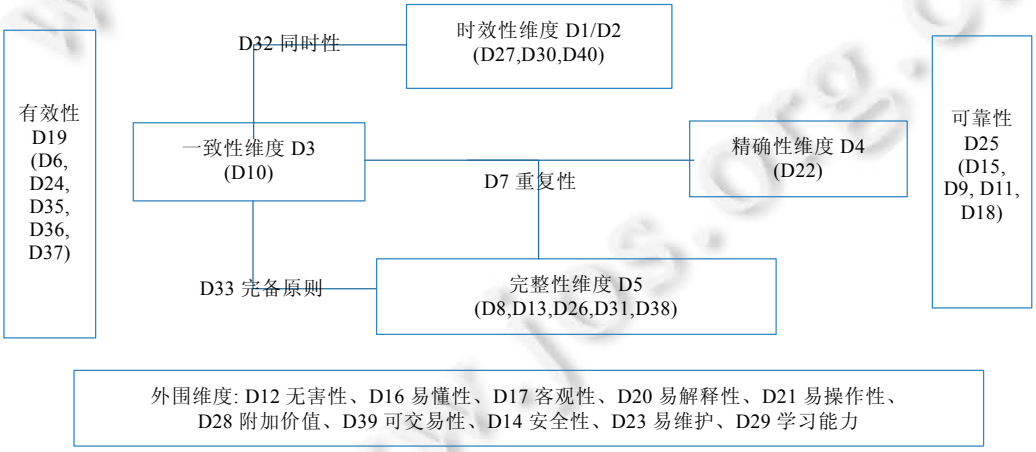


Fig.1 Multi-Dimensional data quality assessment framework

图 1 数据质量综合评估结构

此外,我们对框架的应用条件进行了分析,对于原始研究对象——数据集合,用户应拥有(法律、道德范围或具体准则要求内的)完全操作权限.此外,为保证评估结构的可靠性和客观性,所有数据质量性质指标均为客观评价指标,并且假定数据的内容对用户的现有需求或当前亟待解决的问题具有同等的切题性.同时,为了便于讨

论,本文仅研究结构化的关系型数据.

在研究中我们发现,数据质量性质间存在以下几种可能的相关关系,本文将在第 3.2 节中给出明确定义.

(1) 完全正相关:若数据质量性质 D_i 的变化会引起 D_j 的变化($i,j\in N$),且两者变化方向相同,则称 D_j 与 D_i 完全正相关,记作 $D_iR_+D_j$;

(2) 完全负相关:若数据质量性质 D_i 的变化会引起 D_j 的变化($i,j\in N$),且两者变化方向相反,则称 D_j 与 D_i 完全负相关,记作 $D_iR_-D_j$;

(3) 不相关:数据质量性质 D_i 的变化与 D_j 的变化无(明显)关系,记作 $D_i\overline{R}D_j$;

(4) 不完全相关:数据质量性质 D_i 的变化会引起 D_j 以概率 p 进行相同方向的变化,以概率 q 进行相反方向的变化($q+p\leqslant 1$),记作 D_iRD_j;

(5) 蕴含关系:数据质量性质 D_j 是(或等同于)D_i 的第 k 个子性质,记为 $D_j=D_{i_k}$ ($i,j\in N$,且 $k\in[0,i]$).

2.1.2 框架的可靠性分析

本小节基于统计方法分析框架的可靠性.我们采用统计学与社会科学研究中的常用方法,针对文献[5]中提供的千份调查问卷计算得出各性质的重要性分析,如图 2 所示,通过计算克隆巴赫系数[15](Cronbach's alpha)对调查题目的信度进行测试.

$$\alpha=\frac{K}{K-1}\left(1-\frac{\sum_{i=1}^{K}\sigma_{Y_i}^2}{\sigma_X^2}\right),$$

其中,K 为样本数,σ_X^2 为总体样本的方差,$\sigma_{Y_i}^2$ 是已观测样本的方差.在本文中,α为信度系数,K 为调查题目数,σ_X^2 为所有性质总得分的方差,$\sigma_{Y_i}^2$ 是第 i 个性质得分的方差.通常,克隆巴赫系数α值域为(0,1),如果α不超过 0.6,则认为其内部信度不足.达到 0.7~0.8 时,认为量表具有很高信度.

根据图2所示,排名前20种数据质量性质得分(score)的置信区间(score∈[1,10]且重要程度递减)在本项目中所选取的 6 种核心性质,均有很高的用户认可度,且每个性质的可靠性很高(均在 0.77 以上).因此,我们认为本模型的可靠性较强.

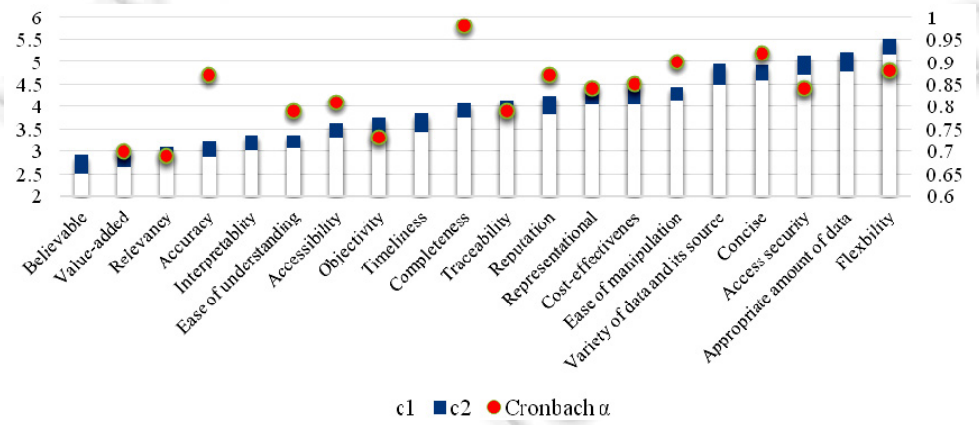


Fig.2 Statistics in the importance of partial data quality dimensions

图 2 部分数据质量性质重要性统计示意图

2.2 数据质量性质的定义

本节详细分析了核心性质中时效性、一致性、完整性、精确性这 4 种性质,确定每种性质与其他相似性质间的关系,并将存在蕴含关系的性质合并到核心性质中,并对每种性质中实际存在的错误类型进行总结.

2.2.1 时效性

数据的时效性是影响数据质量的重要因素之一.时效性性质属于时间相关性质(time-related dimension),用于描述数据的更新程度对数据质量的影响.通过对文献[2,8,9]中时效性相关性质的研究,确定了时效性

(currency)问题的研究包含了新鲜度(freshness)、数据衰败度(data decay)、及时可用性(timeliness and availability)这 3 种性质问题的研究.

本文用 D_{name} 表示具体的数据质量性质,用第 2.1 节的表示形式,时效性相关的几种性质的关系可表示为

$$D_{Currency1}=D_{Freshness},且D_{curr1}=D_{Freshness}R_{+}D_{currency},$$

$$D_{Currency2}=-D_{Data\_Decay},且D_{curr2}=D_{Data\_Decay}R\_D_{currency},$$

$$D_{currency3}=D_{Timeliness\_Avaliability}.$$

2.2.2 一致性

一致性主要评估单个数据集合内,或者多个数据集合之间数据记录、格式、内容等方面的一致情况.通过对文献[4,6,7]中一致性相关性质的研究,一致性表示(consistent representation)、一致性与同时性(consistency & synchronization)这两种性质的研究包含在一致性问题的研究中,其关系可表示为

$$D_{Consistency1}=D_{Consistent\_R},且D_{cons1}=D_{Cons\_R}R_{+}D_{consistency},$$

$$D_{Consistency2}=D_{Cons\&Syn},且D_{cons2}=D_{Cons\&Syn}R_{+}D_{consistency}.$$

根据数据一致性[6]的常见问题,将数据一致性问题具体分为以下 5 种(见表 1),在本文的研究范围内,数据集合发生表 1 中任意一项的违反情况,将视为违反了数据一致性要求.具体的违反模式定义将在第 3.2 节给出.通过初步分析发现,表中格式一致性、内容一致性与数据精确性性质相关,时间一致性与数据时效性性质相关.具体关联关系将在第 3.3 节中给出.

Table 1 An overview of data-quality problems in consistency

表 1 数据一致性的问题分类

问题分类	描述
概念一致性	信息系统与数据集物理结构的匹配、符合程度
格式一致性	多源数据集中的数据结构、属性、关系的匹配程度;单源数据集中字段值格式的统一程度
值域一致性	实际值对值域的符合程度,即值满足值域之间的运算关系
内容一致性	单源数据集中相同属性的具体字段值的准确程度
时间一致性	有序数据(序列)的正确性

2.2.3 精确性

精确性(accuracy)评估了在信息系统中,数据集合每个有意义的记录状态能准确表示物理世界信息的能力[2,4].无错性(free of error)性质包含在精确性的研究中,其关系表示如下:

$$D_{Accurancy1}=D_{Free\_of\_error},且D_{acc1}=D_{Free\_of\_error}R_{+}D_{Accuracy}.$$

此外,通过对文献[4–6]等文献中精确性方面所出现的实际问题的分析,将违反精确性的错误类型总结为表 2 所示的 6 种情况.

Table 2 An overview of data-quality problems in accuracy

表 2 违反数据精确性的错误类型

问题	错误数据	原因
不合法值	出生日期=1984/13/20	日期超出了日期型数据范围
违反数据类型	出生日期=1984/aa/dd	记录中有错误字符
违反属性依赖	生肖:狗,出生日期=1984/3/20	生肖与出生日期不一致
违反属性模式	出生日期=84/3/20	日期格式不准确
违反业务事实	入库日期=2012/07/23,出库日期=2012/06/21	出库日期比入库日期早
精确不够	出生日期=1984	没有具体到日期

2.2.4 完整性

完整性(completeness)是指特征、特征属性和特征关系的多余或缺失情况的处理.完整性模块评估了数据的完整程度,数据规格(data specification)、合适的数据量(appropriate amount of data)、数据量(amount of data)、简洁性(concise)、数据规模(data coverage)的研究均包含在完整性的研究中,其关系表示为

$$D_{Completeness1}=D_{Data\_S},且\ D_{com1}=D_{Data\_S}R_{+}D_{completeness},$$

$$D_{Appropriate\_amount\_of\_data}\in D_{Amount\_of\_data},$$

$$D_{Completeness2}=D_{Amount\_of\_data}, \text{且 } D_{com2}=D_{Amount\_of\_data}R_{+}D_{completeness},$$

$$D_{Completeness3}=D_{Concise},$$

$$D_{Completeness4}=D_{Data\_coverage}, \text{且 } D_{com3}=D_{Data\_coverage}R_{+}D_{completeness}.$$

通过对文献[4,11]中完整性的实际问题的分析,将违反精确性的错误类型总结为表 3 所示的 5 种情况.

Table 3 An overview of data-quality problems in completeness

表 3 违反数据完整性的错误类型

问题分类	描述
规模完整性	保证数据集合的可用性的规模程度
属性完整性	数据表中针对记录最小所需属性量的覆盖程度
内容完整性	数据集合中表、记录、数据项、符号、标记等是否存在缺失及其程度
关系完备性	数据集中的特征、特征属性及特征关系的存在程度
数据完备性	数据集合在各个性质上的完整程度

3 主要数据质量性质关联关系

在用数据质量性质进行描述时,通常采用定量元素方法[16]描述数据集合对于预先设定或理想预期中的质量标准及指标的满足程度,并提供定量的质量信息.本节研究了对数据可用性有重要影响的 4 种主要的数据质量性质.首先根据第 2.2 节中对于信息系统在每个数据质量性质上的违反情况和错误发生情况的具体分析,将每种质量性质细分为多个二、三级质量性质,在此层面上对数据集合违反模式分别给出明确定义,之后在本文的违反模式定义范围内,给出其具体的关联关系.

在下面的讨论中,$X=\{x_1,x_2,x_3,...,x_N\}$ 是一个数据库上的实例,存在一个标准关系模式 $R_{standard}(A_{s1},A_{s2},A_{s3},\ldots,A_{sm})$ 且 $\forall t_i\in R_{standard}, V^s_{t_i[A_j]}$ 准确记录.

3.1 性质违反模式定义

精确性.根据上文对数据精确性错误类型的分析,总结为 3 个二级数据质量性质进行描述,见表 4.这 3 个二级数据质量性质较为全面地描述了精确性的具体评价指标,并且具有较高的计算可行性.在考虑数据精确性错误类型时,我们对精确性违反模式进行如下定义.

定义 3.1(精确性违反模式定义). 对于一个具体的关系模式,$R(A_1:d_1,A_2:d_2,A_3:d_3,...,A_n:d_n)$ 上有 n 个属性 A_i,其对应的域为 d_i,D 是 R 的数据集合实例(此时,准确记录 $R_{standard}$ 重点为"精确而具体"的准确记录).当 R 对 $R_{standard}$ 的相符合程度满足以下条件时:

$$match(R,R_{standard})=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\max match(R,R_{standard})<1,$$

则存在精确性错误.此时,对于 $\exists t_i\in R,t_j\in R_{standard}, V_{t_i[A_j]}\neq V^s_{t_i[A_j]}$,则称 t_i 发生了精确性错误,将 t_i 加到精确性违反模式集合中,即模式 R 中,$t_i\in\sum Vio(D_{acc})$.

在对上述定义所进行的精确性分析过程中,要求关系模式中所有记录值是确定的,仅适用于评估内容精确性.然而,在实际的数据集合中,某些字段值的内容记录是正确的,但记录的精度不够(粒度精确性),或者在处理数据时,由于客观原因无法进行精确计算,而只能采用近似算法给出近似结果(计算精确性).对于这些特殊记录情况,我们不能简单地认为它们是违反精确性的,因此,我们给出精确性违反模式的推广定义.

精确性违反模式推广定义. 在信息系统中,对于一个具体的关系模式 $R(A_1:d_1,A_2:d_2,A_3:d_3,...,A_n:d_n)$ 上有 n 个属性 A_i,其对应的域为 d_i,D 是 R 的数据集合实例.当 R 对 $R_{standard}$ 的匹配程度满足以下条件时:

$$match(R,R_{standard})=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\max match(R,R_{standard})<1,$$

则出现了精确性问题.此时,对于 $\exists t_i\in R,t_j\in R_{standard}, V_{t_i[A_j]}$ 到 $V^s_{t_i[A_j]}$ 编辑距离大于标准编辑距离,即 $d\left(V_{t_i[A_j]},dV^s_{t_i[A_j]}\right)>d^s$ 时,则称 t_i 存在精确性问题.当 R 与 $R_{standard}$ 的近似程度(即精确性评估结果)小于给定的数据质量精

确性要求标准系数,即 $Q_{acc}(R,R_{standard})<q_{acc}$ 时,则称 R 的数据质量精确性低,其对精确性违反程度为 $Vio_{acc}(R,R_{standard})$.将存在违反情况的 t_i 加入精确性性质违反集合中,即模式 R 中的 $t_i\in\sum Vio(D_{acc})$.

Table 4 The category of data quality dimensions: Accuracy

表 4 数据质量分级性质表:精确性

二级数据质量性质	三级数据质量性质描述
内容精确性	有无不合法值 有无数据类型错误 有无属性模式错误
粒度精确性	数据表的记录范围精度是否足够
计算精确性	字段值的精度是否足够

完整性.根据第 2.2.4 节中对于数据完整性的错误类型的分析,用 4 个二级数据质量性质进行描述(见表 5),这 4 个二级性质全面地覆盖了数据在记录、数据表、属性、数据规模上的完整性评估,并且在计算和修复过程中具有较高的可行性[4,10,17].完整性的违反模式定义如下.

定义 3.2(完整性违反模式定义). 对于一个具体的关系模式 $R(A_1:d_1,A_2:d_2,A_3:d_3,...,A_n:d_n)$ 上有 n 个属性 A_i,其对应的域为 d_i,D 是 R 的数据集合实例.R 满足下列任一条件时,则称其发生了完整性错误.

(a) $\exists t_i\in R$, t_i 在全部属性 $A=\{A_1,A_2,A_3,\ldots,A_n\}$ 上的取值:若 $V_{t_i[A_j]}=\text{null}$, 则记为 $v_{t_i[A_j]}=0$,$V_{t_i[A_j]}\neq\text{null}$, 则记为 $v_{t_i[A_j]}=1$; 且 $\wedge v_{t_i[A_j]}=0$;

(b) $A=\{A_1,A_2,A_3,\ldots,A_n\},A^s=\{A_{s1},A_{s2},A_{s3},...,A_{sm}\},A\cap A^s\neq A^s$;

(c) $t_i\in R$, 已知 R 中元组总数 N_{t_i}, $R_{standard}$ 中元组的个数 N^s,$N_{t_i}<N^s$.

此时,满足(a)与(b),称为违反了完整性中的内容完整性,满足(c),称为违反了完整性中的规模完整性.将满足上述条件的 t_i 加入完整性性质违反集合中, 即模式 R 中的 $t_i\in\sum Vio(D_{com})$.

对上述定义所进行的完整性分析是较为理想化的,如果某一个关系模式中存在一项字段值的缺失即被判断为违反了完整性,这样的判断在实际的信息系统中可能过于苛刻.在实际情况中,海量数据里偶有一小部分缺失值但不影响整体数据完整性的情况很常见.此外,一个关系模式中包含的属性集不总是给定的标准属性集的子集,通常是标准属性集与其他相关集的交集,即没有完全覆盖全部标准属性,但记录了一些相关属性.同理,数据量的判断也存在这种问题.因此为了更好地适应客观条件,我们给出了完整性违反模式推广定义.

完整性违反模式推广定义. 在信息系统中,对于一个具体的关系模式 $R(A_1:d_1,A_2:d_2,A_3:d_3,...,A_n:d_n)$ 上有 n 个属性 A_i,其对应的域为 d_i,D 是 R 的数据集合实例(此时,准确记录 $R_{standard}$ 重点为“完备”的准确记录).当 R 满足下列任一条件时,则称其发生了完整性错误.

(a) $\forall t_i\in R$, t_i 在必须考察的 l 个重要属性 $A=\{A_1,A_2,A_3,\ldots,A_l\}$ 上的取值:若 $V_{t_i[A_j]}=\text{null}$, 则记为 $v_{t_i[A_j]}=0$, $V_{t_i[A_j]}\neq\text{null}$, 则记为 $v_{t_i[A_j]}=1$; 且 $\wedge v_{t_i[A_j]}=0$; 或 $\forall t_i\in R$, 已知 t_i 的属性个数 n_{t_i}, t_i 的空记录属性个数 $n_{t_{i\_null}}$, 有 $Q_{com1}=F\left(\sum n_{t_i}-n_{t_{inull}}/n_t\right)<q_{com1}$.

其中, Q_{com1} 表示完整性评估结果 1, q_{com1} 是系统按业务分析而给定的数据质量完整性要求标准系数 1.

(b) $A=\{A_1,A_2,A_3,\ldots,A_n\},A^s=\{A_{s1},A_{s2},A_{s3},...,A_{sm}\},Q_{com2}=F(match(A,A^S))<q_{com2}$, 其中, Q_{com2} 表示完整性评估结果 2, q_{com2} 是系统按业务分析而给定的数据质量完整性要求标准系数 2.

(c) $t_i\in R$, 已知 R 中元组总数 N_{t_i},$R_{standard}$ 中元组的个数 N^s,$Q_{com3}=F\left(\sum N^S-N_{t_i}/N^S\right)<q_{com3}$, 其中, Q_{com3} 表示完整性评估结果 3, q_{com1} 是系统按业务分析而给定的数据质量完整性要求标准系数 3.

此时,满足(a)与(b),称为违反了完整性中的内容完整性,满足(c),称为违反了完整性中的规模完整性. $Q_{comi}<q_{comi}$ 时,称为 R 的数据质量完整性低,其对完整性违反程度为 $Vio_{com}(R,R_{standard})$, 将存在违反情况的 t_i 加入完整性性质违反集合中,即模式 R 中的 $t_i\in\sum Vio(D_{com})$.

一致性.根据第 2.2.3 节中对于数据一致性的错误类型的分析,本节选取了格式一致性这个二级性质,利用现有的理论和技术[6,7],能够对一致性问题进行甄别和修复.讨论的数据一致性主要考虑数据集合对于条件函数依赖(conditional functional dependencies,简称为 CFDs)规则的符合程度.首先我们简单介绍一下 CFDs 所描述的问题:R 是数据表的关系模式,考虑一条应用模式 *Student*(*Sid*,*Name*,*Faculty*,*Class*,*Zipcode*,*Country*,*Region*),其中,每条元组都详细说明了一个学生的学号、姓名、学院、班级、邮编、国籍和行政地区.表 6 给出了此关系模式的部分实例.

Table 5 The category of data quality dimensions: Completeness

表 5 数据质量分级性质表:完整性

二级数据质量性质	三级数据质量性质描述
内容完整性	表中的记录是否完整 数据集合中的表是否完整
规模完整性	数据量是否足够
关系完整性	关系是否足够完整
属性完整性	属性个数是否完整

Table 6 Undergraduate information table

表 6 本科生信息表

SID	Name	Faculty	Class	Zipcode	Country	Region
0320101	Bob	CS	201	19014	USA	PA
0310102	Alice	CS	101	19014	USA	PA
0320107	Ellen	EE	201	10012	USA	NY
0440103	Michael	SE	401	10012	USA	NY
…	…	…	…	…	…	…

在 student 关系模式下的 FDs 有:

FD_1:[SID]→[faculty,class],FD_2:[Zipcode,Country]→[Region].

一个学生的所在院系和班号由其学号唯一确定,而其所在的地区(省份)由其邮编代码和国籍来确定.我们将违反依赖的情况视为违反一致性错误发生.根据上表,我们定义两条 CFD 规则.

φ_1=([SID]→[faculty,class],T_1).T_1 如下表所示:

SID	Faculty	Class
0310102	CS	03101
0410202	SE	04102

φ_2=([Zipcode,Country]→[Region],T_2).T_2 如下表所示:

Zipcode	Country	Region
10011	U.S.A	NY(纽约州)
19014	U.S.A	PA(宾夕法尼亚州)

定义 3.3(一致性违反模式定义). 对于两个具体的关系模式: $R(A_1:d_1,A_2:d_2,A_3:d_3,...,A_n:d_n)$ 上有 n 个属性 A_i,i=1,2,…,n,其对应的域为 d_i,D 是 R 的数据集合实例. $R'(A'_1:d'_1,A'_2:d'_2,A'_3:d'_3,...,A'_n:d'_n)$ 有 n'个属性 $A'_i(i'=1,2,...,n')$, 其对应的域为 d'_i, D'是 R'的数据集合实例.已知一组一致性规则集合ΣCFDs,其中有 l 条规则φ_l,且φ_l是准确无误的.对于一条 CFD α=$(R_{standard}:X\to A,t_p)$, 当满足以下任一条件时,则称其发生了一致性错误.

(a) $\forall t_i\in R$, 假设 $t_p[A]$是一个常量, $t[X]\asymp t_p[X]$, 但是 $t[A]\neq t_p[A]$;

(b) 假设 $t_p[A]$是一个变量,R 中两条元组 t 和 t', $t[X]\asymp t_p[X]$ 且 $t[A]\neq t_p[A]$,$t'[X]\cong t_p[X]$,且 $t[A]\neq t_p[A]$, $t[X]=t'[X]$, 但是 $t[A]\neq t'[A]$;

(c) 假设 $t_p[A]$是一个变量,R 中一条元组 t,与 R'中一条元组 t', $t[X]\asymp t_p[X]$, 且 $t[A]\asymp t_p[A]$, $t'[X]\asymp t_p[X]$ 且 $t[A]\neq t_p[A]$,$t[X]=t'[X]$, 但是,$t[A]\neq t'[A]$.

其中,(a)称为单源单元组一致性违反,(b)称为单源多元组一致性违反,(c)称为多源单元组一致性违反.在满足(a)的情况下,将 t_i 加入一致性性质违反集合中,即模式 R 中 $t_i\in\sum Vio(D_{cons})$. 而在满足(b)的情况下,将 t 和 t'加入一致性性质违反集合中,即模式 R 中 $t_i,t'_i\in\sum Vio(D_{cons})$. 在满足(c)的情况下,将 t 和 t'加入一致性性质违反集合中,即模式 R 与 R'中 $t_i,t'_i\in\sum Vio(D_{cons})$.

对于上述定义,我们也考虑了其在实际信息系统数据质量背景下的具体情况,通过数据集合中对一致性的违反程度来定义其一致性质量.

一致性违反模式推广定义. 在信息系统中,对于两个具体的关系模式: $R(A_1:d_1,A_2:d_2,A_3:d_3,...,A_n:d_n)$ 上有 n 个属性 $A_i(i=1,2,\ldots,n)$,其对应的域为 d_i,D 是 R 的数据集合实例. $R'(A'_1:d'_1,A'_2:d'_2,A'_3:d'_3,...,A'_n:d'_n)$ 上有 n'个属性 $A'_i(i'=1,2,...,n')$, 其对应的域为 d'_i (此时准确记录 $R_{standard}$ 重点为"统一而一致"的准确记录).已知一组一致性规则集合ΣCFDs,其中有 l 条规则 φ_l,且 φ_l 是准确无误的,当满足以下任一条件时,则称其发生了一致性违反.

(a) $\forall t_i\in R$, 假设 $t_p[A]$是一个常量, $t[X]\asymp t_p[X]$, 但 $t[A]\neq t_p[A]$, $Q_{cons1}=F\left(\sum n_{t_i}-n_{t_{iviolate}}/n_{t_i}\right)<q_{cons1}$.

其中, Q_{cons1} 表示一致性评估结果 1, q_{cons1} 是系统按业务分析而给定的数据质量一致性要求标准系数 1. n_{t_i} 代表单元组的属性个数, $n_{t_{iviolate}}$ 代表单元组中违反的属性个数.

(b) 假设 $t_p[A]$是一个变量,R 中两条元组 t_i,t'_i, $t[X]\asymp t_p[X]$, 且 $t[A]\neq t_p[A]$,$t'[X]\cong t_p[X]$,且 $t[A]\asymp t_p[A]$, $t[X]=t'[X]$, 但是 $t[A]\neq t'[A]$,$Q_{cons2}=F\left(\sum N_{t_i}-N_{t_{iviolate}}/N_{t_i}\right)<q_{cons2}$.

其中, Q_{cons2} 表示一致性评估结果 2, q_{cons2} 是系统按业务分析而给定的数据质量一致性要求标准系数 2. N_{t_i} 代表元组总个数, $N_{t_{iviolate}}$ 代表违反的元组个数.

(c) 假设 $t_p[A]$是一个变量,R 中一条元组 t,与 R'中一条元组 t', $t[X]\asymp t_p[X]$, 且 $t'[X]\asymp t_p[X]$ 且 $t[A]\asymp t_p[A]$, $t[X]=t'[X]$, 但是$t[A]\neq t'[A]$, $Q_{cons3}=F\left(\sum N_{t_i}-N_{t_{iviolate}}/N_{t_i}\right)<q_{cons3}$.

其中, Q_{cons3} 表示一致性评估结果 3, q_{cons3} 是系统按业务分析而给定的数据质量一致性要求标准系数 3. N_{t_i} 代表计算的元组 (t_i,t'_i) 总对数, $N_{t_{iviolate}}$ 代表违反的元组 (t_i,t'_i) 总对数.

其中,(a)称为单源单元组一致性违反,(b)称为单源多元组一致性违反,(c)称为多源单元组一致性违反. $Q_{consi}<q_{consi}$ 时称为 R 的数据质量一致性低,其对一致性违反程度为 $Vio_{cons}(R,R_{standard})$. 在满足(a)的情况下,将 t_i 加入一致性性质违反集合中,即模式 R 中 $t_i\in\sum Vio(D_{cons})$. 而在满足(b)的情况下,将 t 和 t'加入一致性性质违反集合中,即模式 R 中 $t_i,t'_i\in\sum Vio(D_{cons})$. 在满足(c)的情况下,将 t 和 t'加入一致性性质违反集合中,即模式 R 与 R'中 $t_i,t'_i\in\sum Vio(D_{cons})$.

时效性.根据第 2.2.1 节对时效性的分析,对于那些被记载在数据集合中的记录所对应的实体,将已知其在实际物理世界中已经发生改变,而数据集合中对应的记录并没有改变的情况定义为违反了数据时效性,我们给出如下定义.

定义 3.4(时效性违反模式定义). 考虑两个数据集合实例,对于两个具体的关系模式: $R(A_1:d_1,A_2:d_2,A_3:d_3,...,A_n:d_n)$ 上有 n 个属性 $A_i(i=1,2,\ldots,n)$,其对应的域为 d_i,D 是 R 的数据集合实例. $R'(A'_1:d'_1,A'_2:d'_2,A'_3:d'_3,...,A'_n:d'_n)$ 有 n'个属性 $A'_i(i'=1,2,...,n')$, 其对应的域为 d'_i ,D'是 R'的数据集合实例.已知一组时效约束(currency constraints)集合ΣCCs,其中有 l 条规则 φ_l,且 φ_l 是准确无误的.当满足以下任一条件时,则称其发生了时效性错误.

(a) $\forall t_i,t_j$ 在 A_k 上,当 $t_i[ID]=t_j[ID]$(或 t_i,t_j 指代同一实体时), $(t_i[ID]=t_j[ID])\wedge\varphi_l\rightarrow t_i\prec_{A_k}t_j$ 时,其中,A_k 是模式 R 上的属性,φ_l 是 $R_{standard}$ 上的时效约束,$\prec_{A_k}$ 是在 R 上的时效判定关系符.

(b) $\forall t_i\in R,t_j\in R',\forall A_k\subset R\cap R'$, 当 $t_i[ID]=t_j[ID]$(或 t_i,t_j 指代同一实体时), $(t_i[ID]=t_j[ID])\wedge\varphi_l\rightarrow t_i\prec_{A_k}t_j$.

此时,满足(a)情况称为单源时效违反,元组 t_i 发生时效性错误,将 t_i 加入时效性性质违反集合中,即实例 D 中 $t_i\in\sum Vio(D_{curr})$. 对于满足(b)的情况称为多源比较时效违反,称为 D 相对于 D'发生时效性错误,则将实例 D 加入精确性性质违反集合中,即实例 $D_i\in\sum Vio(D_{curr})$.

时效性违反模式推广定义. 对于一个具体的关系模式 $R(A_1:d_1,A_2:d_2,A_3:d_3,...,A_n:d_n)$ 上有 n 个属性 $A_i(i=1,2,\ldots,n)$,其对应的域为 d_i,D 是 R 的数据集合实例. $R'(A'_1:d'_1,A'_2:d'_2,A'_3:d'_3,...,A'_n:d'_n)$ 有 n'个属性 $A'_i(i'=1,2,...,n')$, 其对应的域为 d'_i ,D'是 R'的数据集合实例(此时准确记录重点为"最新的"准确记录).已知一组时效约束集合ΣCCs,其中有 l 条规则 φ_l,且 φ_l 是准确无误的.当满足以下任一条件时,则称其发生了时效性错误.

(a) $\forall t_i,t_j$ 在需要判定时效性的属性 A_k 上,当 $t_i[ID]=t_j[ID]$(或可判断 t_i,t_j 指代同一实体时), $(t_i[ID]=$

$t_j[ID]) \wedge \varphi_l \to t_i \prec_{A_k} t_j$ 时, $Q_{curr1} = F\left(\sum n_{t_i} - n_{t_{i\_outofdate}} \Big/ n_{t_i}\right) < q_{curr1}$.

其中,A_k 是模式 R 上的所有需要判定时效性的属性集合,$\prec_{A_k}$ 是在 R 上的时效判定关系符. Q_{curr1} 表示时效性评估结果 1, q_{curr1} 是系统按业务分析而给定的数据质量时效性要求标准系数 1. n_{t_i} 代表单元组的属性个数, $n_{t_{i\_outofdate}}$ 代表单元组记录过时的属性个数.

(b) $\forall t_i \in R, t_j \in R'$, ∀需要判断时效性的属性 $A_k \subset R \cap R'$, 当 $t_i[ID]=t_j[ID]$(或可判断 t_i, t_j 指代同一实体时), $(t_i[ID] = t_j[ID]) \wedge \varphi_l \to t_i \prec_{A_k} t_j$ 时, $Q_{curr2} = F\left(\sum N_{t_i} - N_{t_{i\_outofdate}} \Big/ N_{t_i}\right) < q_{curr2}$.

Q_{curr2} 表示时效性评估结果 2, q_{curr2} 是系统按业务分析而给定的数据质量时效性要求标准系数 2. N_{t_i} 代表实例 D 上元组的属性个数, $N_{t_{i\_outofdate}}$ 代表实例 D 上记录过时的元组个数.

此时, $Q_{curri} < q_{curri}$ 称为 R 的数据质量时效性低,其对时效性违反程度为 $Vio_{curr}(R, R_{standard})$. 满足(a)情况,称为单源时效违反,元组 t_i 发生时效性错误,将 t_i 加入时效性性质违反集合中,即实例 D 中 $t_i \in \sum Vio(D_{curr})$. 对于满足(b)的情况,称为多源比较时效违反,也称为 D 相对于 D'发生时效性错误,则将实例 D 加入精确性性质违反集合中,即实例 $D_i \in \sum Vio(D_{curr})$.

3.2 数据质量性质关联关系分析

根据上一节各个数据质量性质的违反模式,在这一节给出理论上的分析证明.数据质量评估的工作不仅在于单纯地对数据集合的质量进行量化评估,而是在发现数据集合对于上述几种性质的违反情况后,用数据清洗或其他技术手段对其进行有效的修复,以便在后续应用中能发挥更大的价值.本节所讨论的数据质量性质的关联关系,即从修复的角度出发,证明在对数据集合进行修复的过程中,几个重要性质所发生的影响关系.

我们首先给出数据质量性质间影响关系的定义.

定义 3.5. 在数据集合 $X = \{x_{11}, x_{12}, \ldots, x_{mn}\}$ 上,有如下性质 $D_i, D_j \in \Sigma D$ (ΣD 是所有数据质量性质的集合),X 在 D_i 上达到的质量标准记为 $Q_1^{D_i}$, 在 D_j 上达到的质量标准记为 $Q_1^{D_j}$. 利用修复集合 $R^{Q_1} = \{y_{11}, y_{12}, \ldots, y_{m'n'}\}$ 对 D_i 修复后经重新计算 X 在 D_i 上的质量标准变为 $Q_2^{D_i}$, 在 D_j 上的质量标准变为 $Q_2^{D_j}$, 且 $Q_2^{D_i} \neq Q_1^{D_i}$, 记 $\Delta Q^D = Q_2^D - Q_1^D$.

(1) 若 $\Delta Q^{D_2} = 0$, 则 D_i 的修复与 D_j 不相关,即 $D_i \not\prec D_j$;

(2) 若 $\Delta Q^{D_2} \neq 0$, 则 D_i 的修复与 D_j 相关,此时若 $\Delta Q^{D_1} > 0$, 且 $\Delta Q^{D_2} > 0$, 则称 D_i 的修复与 D_j 正相关,即 $D_i \prec_+ D_j$; 若 $\Delta Q^{D_1} > 0$, 且 $\Delta Q^{D_2} < 0$, 则称 D_i 的修复与 D_j 负相关,即 $D_i \prec_- D_j$; 若 $\Delta Q^{D_1} > 0$, 且 $\Delta Q^{D_2} > 0$ 的概率 $P^{D_2} \in (0,1)$, 则称 D_i 的修复与 D_j 不完全相关,即 $D_i \prec D_j$.

在下面的讨论中, $X = \{x_1, x_2, \ldots, x_N\}$ 是一个数据库上的实例,存在一个标准关系模式 $R_{standard}(A_{s1}, A_{s2}, A_{s3}, \ldots, A_{sm})$, 且 $\forall t_i \in R_{standard}, V^s_{t_i[A_j]}$ 为准确记录.X 由 M 个属性组成,即 $A = \{A_1, A_2, A_3, ..., A_m\}$, 本部分进行 4 个数据质量性质的评价,即 $D = \{D_{acc}, D_{com}, D_{cons}, D_{curr}\}$.

3.2.1 精确性与其他数据质量性质关联分析

针对精确性错误的修复会使待修复记录值更精确,不会丢失任何信息,没有违反完整性原则,因此不会引起完整性错误.

定理 1. 精确性违反模式的修复对完整性不相关,即 $D_{acc}, D_{com} \in D$, 有 $D_{acc} \not\prec D_{com}$.

证明: $\exists t_i \in X, t_k \in R_{standard}$, t_i 在属性 $A = \{A_1, A_2, A_3, ..., A_m\}$ 上的取值 $V_{t_i[A_j]} \neq V^S_{t_i[A_j]}$, 且 $V_{t_i[A_j]}, V^S_{t_i[A_j]}$ 皆不为空.

(1) 当对 $V_{t_i[A_j]}$ 以 $V'_{t_i[A_j]}$ 进行更新时, $V'_{t_i[A_j]} \neq \text{null}$, 且 $A' = A = \{A_1, A_2, A_3, ..., A_m\}$, 未对完整性造成影响, $V'_{t_i[A_j]} = V^S_{t_i[A_j]}$ 的概率为 $p^{D_{acc}}_{t_i[A_j]}, p^{com}_{t_i[A_j]}$ 与 $p^{D_{acc}}_{t_i[A_j]}$ 无关;

(2) 当对足够多个 $V_{t_i[A_j]}$ 以与之对应的 $V'_{t_i[A_j]}$ 进行更新时,由定义 3.2 可知,空记录 N 的个数不变,且 $Q'_{com1} =$

$Q_{com1}, Q'_{com2}=Q_{com2}, Q'_{com3}=Q_{com3}$ 未发生完整性违反,此时发生完全正确更新的概率为 $\prod^{n}_{i=1}\prod^{m}_{j=1} p^{D_{acc}}_{t_i[A_j]}$, P^{com} 不变,且 $\Delta Q^{D_{com}}=0$.

综上,由定义 3.5 可知, $D_{acc} \not\asymp D_{com}$ 成立,证毕. □

对于粒度精确性错误的修复会改变某些属性的值的范围,一致性约束规则中的数据的精确性保持与数据库中数据同步,因此精确性修复对一致性没有影响.

定理 2. 精确性违反模式的修复对一致性不相关,即 $D_{acc}, D_{cons} \in D$, 有 $D_{acc} \not\asymp D_{cons}$.

证明: $\exists t_i \in X, t_k \in R_{standard}$, t_i 在属性 $A=\{A_1, A_2, A_3, ..., A_m\}$ 上的取值 $V_{t_i[A_j]} \neq V^{S}_{t_i[A_j]}$, 且 $V_{t_i[A_j]}, V^{S}_{t_i[A_j]}$ 皆不为空.

(1) 当对 $V_{t_i[A_j]}$ 以 $V'_{t_i[A_j]}$ 进行更新时, $d\left(V'_{t_i[A_j]}, V^{S}_{t_i[A_j]}\right) < d\left(V_{t_i[A_j]}, V^{S}_{t_i[A_j]}\right)$, 根据定义 3.3 可知,如果对于 $t_{i[A_j]}$ 满足在一致性规则集合ΣCFDs 中有对于一条 CFD $\alpha=(R_{standard}: X \to A, t_p)$ 与其对应,则 t_i 的修复未对一致性造成影响, $V'_{t_i[A_j]}=V^{S}_{t_i[A_j]}$ 的概率为 $p^{D_{acc}}_{t_i[A_j]}, p^{cons}_{t_i[A_j]}$ 与 $p^{D_{acc}}_{t_i[A_j]}$ 无关;

(2) 当对足够多个 $V_{t_i[A_j]}$ 以与之对应的 $V'_{t_i[A_j]}$ 进行更新时,根据定义 3.3 可知,违反记录 N 的个数不变,即 $Q'_{cons}=Q_{cons}$, 未发生一致性违反,此时发生完全正确更新的概率为 $\prod^{n}_{i=1}\prod^{m}_{j=1} p^{D_{acc}}_{t_i[A_j]}, P^{com}$ 不变,且 $\Delta Q^{D_{cons}}=0$.

综上,由定义 3.5 可知, $D_{acc} \not\asymp D_{cons}$ 成立,证毕. □

针对粒度精确性错误的修复使待修复记录值更精确,使信息系统中数据集合里的属性状态值与现实物理世界中的真实情况的状态更为接近,因此不会引起时效性错误.

定理 3. 对于精确性违反模式的修复对时效性不相关,即 $D_{acc}, D_{curr} \in D$, 有 $D_{acc} \not\asymp D_{curr}$.

证明: $\exists t_i \in X, t_k \in R_{standard}$, t_i 在属性 $A=\{A_1, A_2, A_3, ..., A_m\}$ 上的取值 $V_{t_i[A_j]} \neq V^{S}_{t_i[A_j]}$, 且 $V_{t_i[A_j]}, V^{S}_{t_i[A_j]}$ 皆不为空.

(1) 当对 $V_{t_i[A_j]}$ 以 $V'_{t_i[A_j]}$ (不为空)进行更新时, $d\left(V'_{t_i[A_j]}, V^{S}_{t_i[A_j]}\right) < d\left(V_{t_i[A_j]}, V^{S}_{t_i[A_j]}\right)$, 根据定义 3.4 可知,如果对于 $t_{i[A_j]}$ 满足在时效约束集合ΣCCs 中有对于一条时效约束 φ_l 与其对应,则 t_i 的修复对时效性未产生影响.此时 $V'_{t_i[A_j]}=V^{S}_{t_i[A_j]}$ 的概率为 $p^{D_{acc}}_{t_i[A_j]}, p^{curr}_{t_i[A_j]}$ 与 $p^{D_{acc}}_{t_i[A_j]}$ 无关;

(2) 当对足够多个 $V_{t_i[A_j]}$ 以与之对应的 $V'_{t_i[A_j]}$ 进行更新时,根据定义 3.4 可知,违反记录 N 的个数不变,即 $Q'_{curr}=Q_{curr}$, 未发生时效性违反,此时发生完全正确更新的概率为 $\prod^{n}_{i=1}\prod^{m}_{j=1} p^{D_{acc}}_{t_i[A_j]}, P^{curr}$ 不变,且 $\Delta Q^{D_{curr}}=0$.

综上,由定义 3.5 可知, $D_{acc} \not\asymp D_{curr}$ 成立,证毕. □

综上所述,对精确性错误的修复与其他 3 个质量性质不相关.

3.2.2 完整性与其他数据质量性质关联分析

完整性错误的修复过程会对记录中缺失的字段值、数据表中缺失的属性乃至不足的数据量进行填充,新增加的数据部分可能由于来自不同数据源而存在不一致错误,进而引发修复后的数据违反一致性约束,因此可能会引起一致性错误.

定理 4. 对于完整性违反模式的修复对一致性不完全相关,即 $D_{com}, D_{cons} \in D$, 有 $D_{com} \prec D_{cons}$.

证明: $\exists t_i \in X, t_k \in R_{standard}$, t_i 在属性 $A=\{A_1, A_2, A_3, ..., A_m\}$ 上的取值 $V_{t_i[A_j]}=\text{null}$.

(1) 当对 $V_{t_i[A_j]}$ 以 $V'_{t_i[A_j]}$ 进行填充修复时,若 $V_{t_i[A_j]}=V^{S}_{t_i[A_j]}$, 如果对于 X 上的属性 A_j,满足在一致性规则集合ΣCFDs 中有对于一条 CFD $\alpha=(R_{standard}: X \to A, t_p)$ 与其对应,则 t_i 的修复符合一致性规则,即其发生概率为 $p^{cons}_{t_i[A_j]}=p^{D_{com}}_{t_i[A_j]}$;

若 $V_{t_i[A_j]} \neq V^{S}_{t_i[A_j]}$, 由定义 3.3 可知:根据一致性规则集合ΣCFDs 中有对于一条 CFD $\alpha=(R_{standard}: X \to A, t_p)$, $t[X] \asymp t_p[X]$, 但 $t[A] \neq t_p[A]$, 则 t_i 发生一致性违反,即 $t_i \in \sum Vio(D_{cons})$, 其发生概率为 $\overline{p^{cons}_{t_i[A_j]}}=\overline{p^{D_{com}}_{t_i[A_j]}}$;

(2) 当对足够多的$V_{t_i[A_j]}$以与之对应的$V'_{t_i[A_j]}$进行填充修复时,将满足$V_{t_i[A_j]}=V^S_{t_i[A_j]}$的$t_{i[A_j]}$个数记为$n_{correct}$,将满足$V_{t_i[A_j]}\neq V^S_{t_i[A_j]}$的$t_{i[A_j]}$个数记为$n_{error}$,计算$Q_{cons}=F\left(\sum n_{t_i}-(n_{t_{violate}}-n_{correct}+n_{error})/n_{t_i}\right)$,若$n_{correct}-n_{error}\geqslant 0$,则$Q_{cons}\geqslant q_{cons}$,即$X$的一致性得到提高;若$n_{correct}-n_{error}<0$,则$Q_{cons}<q_{cons}$,则根据定义3.3的推广定义可知,$X$发生了一致性违反,即$\Delta Q^{D_{cons}}>0$的概率$P^{D_{cons}}\in(0,1)$.

综上,由定义3.5可知,$D_{com}\prec D_{cons}$成立,证毕. □

在对完整性错误进行修复时,新填充的数据部分可能是根据之前已发现的规则推理得到,但此过程可能由于补充的数据源不同或者所采用的缺失值填充方法的原理不同,填充部分数据的时效性不能得到保障,因此这一过程可能会引入时效性错误.

定理5. 对于完整性违反模式的修复对时效性不完全相关,即$D_{com},D_{curr}\in D$,有$D_{com}\prec D_{curr}$.

证明:$\exists t_i\in X,t_k\in R_{standard}$,$t_i$在属性$A=\{A_1,A_2,A_3,...,A_m\}$上的取值$V_{t_i[A_j]}=\text{null}$.

(1) 当对$V_{t_i[A_j]}$以$V'_{t_i[A_j]}$进行填充修复时,若$V_{t_i[A_j]}=V^S_{t_i[A_j]}$,如果对于$X$上的属性$A_j$,满足在时效约束集合ΣCCs中有对于一条时效约束$\varphi_l$与其对应,则$t_i$的修复符合时效性规则,即其发生概率为$p^{curr}_{t_i[A_j]}=p^{D_{com}}_{t_i[A_j]}$;

若$V_{t_i[A_j]}\neq V^S_{t_i[A_j]}$,由定义3.4可知:根据时效约束集合ΣCCs中有一条时效约束φ_l,可以得出$(t_i[ID]=t_k[ID])\wedge\varphi_l\rightarrow t_i\prec_{A_k}t_k$,则$t_i$发生时效性违反,即$t_i\in\sum Vio(D_{curr})$,其发生概率为$\overline{p^{curr}_{t_i[A_j]}}=\overline{p^{D_{com}}_{t_i[A_j]}}$;

(2) 当对足够多的$V_{t_i[A_j]}$以与之对应的$V'_{t_i[A_j]}$进行填充修复时,将满足$V_{t_i[A_j]}=V^S_{t_i[A_j]}$的$t_{i[A_j]}$个数记为$n_{correct}$,将满足$V_{t_i[A_j]}\neq V^S_{t_i[A_j]}$的$t_{i[A_j]}$个数记为$n_{error}$,计算$Q_{cons}=F\left(\sum n_{t_i}-(n_{t_{ioutofdate}}-n_{correct}+n_{error})/n_{t_i}\right)$,若$n_{correct}-n_{error}\geqslant 0$,则$Q_{curr}\geqslant q_{curr}$,即$X$的时效性得到提高;若$n_{correct}-n_{error}<0$,则$Q_{curr}<q_{curr}$,则根据定义3.3的推广定义可知,$X$发生了时效性违反,即$\Delta Q^{D_{curr}}>0$的概率$P^{D_{curr}}\in(0,1)$.

综上,由定义3.5可知,$D_{com}\prec D_{curr}$成立,证毕. □

完整性错误的修复会对记录中缺失的字段、数据表中缺失的属性乃至不足的数据量进行填充,在基于规则或基于概率的缺失值填充方法中,对缺失值的判定结果不是唯一确定的,在此过程中,新的填充数据可能会影响精确性的违反模式.

定理6. 对于完整性违反模式的修复对精确性不完全相关,即$D_{com},D_{acc}\in D$,有$D_{com}\prec D_{acc}$.

证明:$\exists t_i\in X,t_k\in R_{standard}$,$t_i$在属性$A=\{A_1,A_2,A_3,...,A_m\}$上的取值$V_{t_i[A_j]}=\text{null}$.

(1) 当对$V_{t_i[A_j]}$以$V'_{t_i[A_j]}$进行填充修复时,若$V_{t_i[A_j]}=V^S_{t_i[A_j]}$,则$t_i$的精确性得到提高,即其发生概率为$p^{acc}_{t_i[A_j]}=p^{D_{com}}_{t_i[A_j]}$.若$V_{t_i[A_j]}\neq V^S_{t_i[A_j]}$,根据定义3.1可知:$t_i$发生了精确性违反,即$t_i\in\sum Vio(D_{acc})$,其发生概率为$\overline{p^{acc}_{t_i[A_j]}}=\overline{p^{D_{com}}_{t_i[A_j]}}$;

(2) 当对足够多的$V_{t_i[A_j]}$以与之对应的$V'_{t_i[A_j]}$进行填充修复时,若$0<d\left(V_{t_i[A_j]},V^S_{t_i[A_j]}\right)<d^S$,则$X$的精确性得到提高,即$Q_{acc}(X,R_{standard})\geqslant q_{acc}$,其发生概率为$P^{acc}=\prod^n_{i=1}\prod^m_{j=1}p^{D_{com}}_{t_i[A_j]}$;

若$d\left(V_{t_i[A_j]},V^S_{t_i[A_j]}\right)<d^S$,则根据定义3.1的推广定义可知,发生了精确性违反,此时$Q_{acc}(X,R_{standard})<q_{acc}$,其概率为$\overline{P^{acc}}=1-\prod^n_{i=1}\prod^m_{j=1}p^{D_{com}}_{t_i[A_j]}$,即$\Delta Q^{D_{acc}}>0$的概率$P^{D_{acc}}\in(0,1)$.

综上,由定义3.5可知,$D_{com}\prec D_{acc}$成立,证毕. □

综上所述,完整性与其他3个质量性质的关系如图3所示.

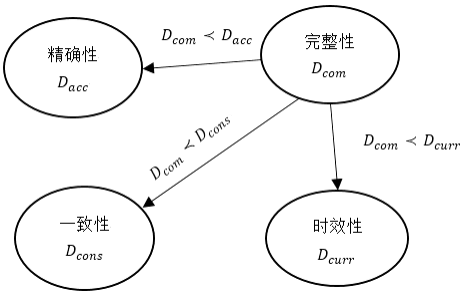


Fig.3 The relationship between completeness and other dimensions

图 3 完整性与其他性质的关系

3.2.3 一致性与其他数据质量性质关联分析

一致性错误的修复操作会将违反语义约束的数据修改为满足一致性约束,不会对数据部分进行删减,因此不会带来数据的丢失,不会引入完整性错误.

定理 7. 对于一致性违反模式的修复对完整性不相关,即 $D_{cons}, D_{com} \in D$, 有 $D_{cons} \not\prec D_{com}$.

证明: $\exists t_i \in X, t_k \in R_{standard}$, t_i 在属性 $A=\{A_1, A_2, A_3, ..., A_m\}$ 上的取值 $V_{t_i[A_j]} \neq V^S_{t_i[A_j]}$, 且 $V_{t_i[A_j]}$ 不为空.已知一组一致性规则集合ΣCFDs,其中有 l 条规则 φ_l,且 φ_l 是准确无误的.t_i 在属性 $A=\{A_1, A_2, A_3, ..., A_m\}$ 上的取值 $V_{t_i[A_j]}$ 违反了ΣCFDs 中的 φ_n.

(1) 当对 $V_{t_i[A_j]}$, 利用一条 CFD $\alpha=(R_{standard}: X \to A, t_p)$ 以 $V'_{t_i[A_j]}$ (不为空)进行填充或者更新时,因为 $V'_{t_i[A_j]} \neq \text{null}$, 且 $A'=A=\{A_1, A_2, A_3, ..., A_m\}$, 未对完整性造成影响, $V'_{t_i[A_j]}=V^S_{t_i[A_j]}$ 的概率为 $p^{D_{cons}}_{t_i[A_j]}, p^{com}_{t_i[A_j]}$ 与 $p^{D_{cons}}_{t_i[A_j]}$ 无关;

(2) 当对足够多的 $V_{t_i[A_j]}$, 利用规则集合ΣCFDs,以与之对应的 $V'_{t_i[A_j]}$ 进行填充或更新时,根据定义 3.2 可知,空记录 N 的个数不变,且 $Q'_{com1}=Q_{com1}, Q'_{com2}=Q_{com2}, Q'_{com3}=Q_{com3}$ 未发生完整性违反,此时发生完全正确更新的概率为 $\prod^n_{i=1}\prod^m_{j=1} p^{D_{cons}}_{t_i[A_j]}, P^{com}$ 不变,即 $\Delta Q^{D_{com}}=0$.

综上,由定义 3.5 可知, $D_{cons} \not\prec D_{com}$ 成立,证毕. □

一致性错误的修复会修改数据,数据的改变可能导致时效性的改变,因此可能会引起时效性错误;当修复一致性函数依赖规则中的属性值因时效性改变而产生变化时,那么利用此项修复得到的结果会引起时效性错误.

定理 8. 对于一致性违反模式的修复对时效性不完全相关,即 $D_{cons}, D_{curr} \in D$, 有 $D_{cons} \prec D_{curr}$.

证明: $\exists t_i \in X, t_k \in R_{standard}$, 已知一组一致性规则集合ΣCFDs,其中有 l 条规则 φ_l,且 φ_l 是准确无误的.t_i 在属性 $A=\{A_1, A_2, A_3, ..., A_m\}$ 上的取值 $V_{t_i[A_j]}$ 违反了ΣCFDs 中的 φ_n.

(1) 当对 $V_{t_i[A_j]}$ 利用一条 CFD $\alpha=(R_{standard}: X \to A, t_p)$ 以 $V'_{t_i[A_j]}$ 进行填充或者更新时,若 $V_{t_i[A_j]}=V^S_{t_i[A_j]}$, 如果对于 X 上的属性 A_j 满足在时效约束集合ΣCCs 中有对于一条时效约束 φ_l 与其对应,则 t_i 的修复符合时效性规则,即其发生概率为 $p^{cons}_{t_i[A_j]}=p^{D_{curr}}_{t_i[A_j]}$;

若 $V_{t_i[A_j]} \neq V^S_{t_i[A_j]}$, 由定义 3.4 可知:根据时效约束集合ΣCCs 中有一条时效约束 φ_l,可以得出 $(t_i[ID]=t_k[ID]) \wedge \varphi_l \to t_i \prec_{A_k} t_k$, 则 t_i 发生时效性违反,即 $t_i \in \sum Vio(D_{curr})$, 其发生概率为 $\overline{p^{cons}_{t_i[A_j]}}=\overline{p^{D_{curr}}_{t_i[A_j]}}$;

(2) 当对足够多的 $V_{t_i[A_j]}$, 利用规则集合ΣCFDs,以与之对应的 $V'_{t_i[A_j]}$ 进行填充或更新时,将满足 $V_{t_i[A_j]}=V^S_{t_i[A_j]}$ 的 $t_{i[A_j]}$ 个数记为 $n_{correct}$, 将满足 $V_{t_i[A_j]} \neq V^S_{t_i[A_j]}$ 的 $t_{i[A_j]}$ 个数记为 n_{error}, 计算 $Q_{cons}=F\left(\sum n_{t_i}-(n_{t_{ioutofdate}}-n_{correct}+n_{error})/n_{t_i}\right)$, 若 $n_{correct}-n_{error} \geqslant 0$, 则 $Q_{curr} \geqslant q_{curr}$, 即 X 的时效性得到提高;若 $n_{correct}-n_{error}<0$, 则 $Q_{curr}<q_{curr}$,

根据定义 3.4 的推广定义可知,X 发生了时效性违反,即 $\Delta Q^{D_{curr}}>0$ 的概率 $P^{D_{curr}}\in(0,1)$.

综上,由定义 3.5 可知, $D_{cons}\prec D_{curr}$ 成立,证毕. □

一致性错误的修复会修改数据,数据的改变会导致精确性的变化,在规则(如 FD[7]、CFD[6,7]、CIND[18])的约束下,在对违反一致性问题的修复过程中,其修复结果可能不是确定的,而是根据最小修复代价判定得到的一个有概率的修复值,因此,这一过程可能会引起精确性错误.

定理 9. 对于一致性违反模式的修复对精确性不完全相关,即 $D_{cons},D_{acc}\in D,D_{cons}\prec D_{acc}$.

证明: $\exists t_i\in X,t_k\in R_{standard}$, 已知一组一致性规则集合ΣCFDs,其中有 l 条规则 φ_l,且 φ_l 是准确无误的.t_i 在属性 $A=\{A_1,A_2,A_3,...,A_m\}$ 上的取值 $V_{t_i[A_j]}$ 违反了ΣCFDs 中的 φ_n.

(1) 当对 $V_{t_i[A_j]}$, 利用一条 CFD $\alpha=(R_{standard}:X\to A,t_p)$ 以 $V'_{t_i[A_j]}$ 进行填充或者更新时,若 $V_{t_i[A_j]}=V^S_{t_i[A_j]}$, 则 t_i 的精确性得到提高,即其发生概率为 $p^{acc}_{t_i[A_j]}=p^{D_{cons}}_{t_i[A_j]}$; 若 $V_{t_i[A_j]}\neq V^S_{t_i[A_j]}$, 根据定义 3.1 可知:则 t_i 发生精确性违反,即 $t_i\in\sum Vio(D_{acc})$, 其发生概率为 $\overline{p^{acc}_{t_i[A_j]}}=\overline{p^{D_{cons}}_{t_i[A_j]}}$;

(2) 当对足够多的 $V_{t_i[A_j]}$, 利用规则集合ΣCFDs,以与之对应的 $V'_{t_i[A_j]}$ 进行填充或更新时,若 $0<d\left(V_{t_i[A_j]},V^S_{t_i[A_j]}\right)\leqslant d^S$, 则 X 的精确性得到提高,即 $Q_{acc}(X,R_{standard})\geqslant q_{acc}$, 其发生概率为 $P^{acc}=\prod^n_{i=1}\prod^m_{j=1}p^{D_{cons}}_{t_i[A_j]}$, 若 $d\left(V_{t_i[A_j]},V^S_{t_i[A_j]}\right)>d^S$, 则根据定义 3.1 的推广定义可知,$X$ 发生了精确性违反,此时 $Q_{acc}(X,R_{standard})<q_{acc}$, 其概率为 $\overline{P^{acc}}=1-\prod^n_{i=1}\prod^m{}_{j=1}p^{D_{cons}}_{t_i[A_j]}$, 即 $\Delta Q^{D_{acc}}>0$ 的概率 $P^{D_{acc}}\in(0,1)$.

综上,由定义 3.5 可知, $D_{cons}\prec D_{acc}$ 成立,证毕. □

综上所述,一致性与其他 3 个质量性质的关系如图 4 所示.

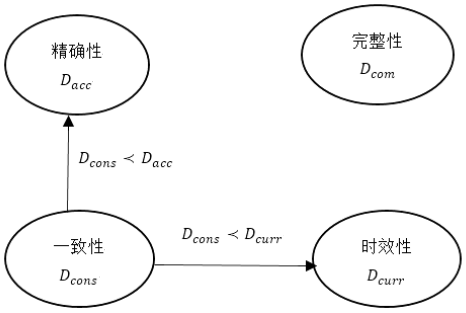


Fig.4 The relationship between consistency and other dimensions

图 4 一致性与其他性质关系

3.2.4 时效性与其他数据质量性质关联分析

在实际的信息系统中,在关系模式下,有一些人为制定或者机器训练产生得到的时效判定或者常识原理可以作为时效约束[8,9]进行时效性修复.时效性错误的修复不会对数据的字段值、属性和数据量进行删改,因此不会引起任何信息的丢失,从而也不会引起完整性错误.

定理 10. 对于时效性违反模式的修复对完整性不相关,即 $D_{curr},D_{com}\in D,D_{curr}\equiv D_{com}$.

证明: $\exists t_i\in X,t_k\in R_{standard}$, t_i在属性 $A=\{A_1,A_2,A_3,...,A_m\}$ 上的取值 $V_{t_i[A_j]}\neq V^S_{t_i[A_j]}$, 且 $V_{t_i[A_j]}$ 不为空.已知一组时效性规则集合ΣCCs,其中有 l 条规则 φ_l,且 φ_l 是准确无误的.t_i在属性 $A=\{A_1,A_2,A_3,...,A_m\}$ 上的取值 $V_{t_i[A_j]}$ 违反了ΣCFDs 中的 φ_n.

(1) 当对 $V_{t_i[A_j]}$, 利用一条 CC $\alpha=(R_{standard}:X\to A,t_p)$ 以 $V'_{t_i[A_j]}$ (不为空)进行填充或者更新时,因为 $V'_{t_i[A_j]}\neq\text{null}$, 且 $A'=A=\{A_1,A_2,A_3,...,A_m\}$, 未对完整性造成影响, $V'_{t_i[A_j]}\neq V^S_{t_i[A_j]}$ 的概率为 $p^{D_{curr}}_{t_i[A_j]},p^{curr}_{t_i[A_j]}$ 与 $p^{D_{com}}_{t_i[A_j]}$

无关;

(2) 当对足够多的$V_{t_i[A_j]}$,利用规则集合ΣCCs,以与之对应的$V'_{t_i[A_j]}$进行填充或更新时,根据定义 3.2 可知,空记录 N 的个数不变,且$Q'_{com1}=Q_{com1},Q'_{com2}=Q_{com2},Q'_{com3}=Q_{com3}$未发生完整性违反,此时发生完全正确更新的概率为$\prod^n_{i=1}\prod^m_{j=1}p^{D_{curr}}_{t_i[A_j]},P^{com}$不变,即$\Delta Q^{D_{com}}=0$.

综上,由定义 3.5 可知,$D_{curr}\not\prec D_{com}$成立,证毕. □

时效性错误的修复会改变某些属性的值,使其结果更匹配物理世界里的实体的状态,但通过时效约束得到的过时数据修复结果不是确定的,而是有概率的近似结果,因此在这一过程中会导致精确性的错误.

定理 11. 对于时效性违反模式的修复对精确性不完全相关,即$D_{curr},D_{acc}\in D$,有$D_{curr}\prec D_{acc}$.

证明:$\exists t_i\in X,t_k\in R_{standard}$,已知一组时效性规则集合ΣCCs,其中有 l 条规则φ_l,且φ_l是准确无误的.t_i在属性$A=\{A_1,A_2,A_3,...,A_m\}$上的取值$V_{t_i[A_j]}$违反了ΣCFDs 中的$\varphi_n$.

(1) 当对$V_{t_i[A_j]}$,利用一条 CC $\alpha=(R_{standard}:X\to A,t_p)$以$V'_{t_i[A_j]}$进行更新时,若$V_{t_i[A_j]}=V^S_{t_i[A_j]}$,则$t_i$的精确性得到提高,其发生概率为$p^{acc}_{t_i[A_j]}=p^{D_{curr}}_{t_i[A_j]}$;若$V_{t_i[A_j]}\neq V^S_{t_i[A_j]}$,根据定义 3.1 可知:$t_i$发生了精确性违反,即$t_i\in\sum Vio(D_{acc})$,其发生概率为$\overline{p^{acc}_{t_i[A_j]}}=\overline{p^{D_{curr}}_{t_i[A_j]}}$;

(2) 当对足够多的$V_{t_i[A_j]}$,利用规则集合ΣCCs,以与之对应的$V'_{t_i[A_j]}$进行更新时,若$0<d\left(V_{t_i[A_j]},V^S_{t_i[A_j]}\right)\leqslant d^S$,则 X 的精确性得到提高,即$Q_{acc}(X,R_{standard})\geqslant q_{acc}$,其发生概率为$P^{acc}=\prod^n_{i=1}\prod^m_{j=1}p^{D_{curr}}_{t_i[A_j]}$;

若$d\left(V_{t_i[A_j]},V^S_{t_i[A_j]}\right)>d^S$,则根据定义 3.1 的推广定义可知,$X$ 发生了精确性违反,此时$Q_{acc}(X,R_{standard})<q_{acc}$,其概率为$\overline{P^{acc}}=1-\prod^n_{i=1}\prod^m_{j=1}p^{D_{curr}}_{t_i[A_j]}$,即$\Delta Q^{D_{acc}}>0$的概率$P^{D_{acc}}\in(0,1)$.

综上,由定义 3.2 可知,$D_{curr}\prec D_{acc}$成立,证毕. □

有时修复部分属性上的部分时序关系,对于相同元组修复后的属性部分,其与其他属性之间的一致性可能会发生改变,影响到一致性约束的判定与评估.此外,如果对于可修复部分与不可修复部分的一致性统一操作方面未考虑全面,则在时效性修复的过程中会在时效约束上引入一致性错误.

定理 12. 时效性违反模式的修复对一致性不完全相关,即$D_{curr},D_{cons}\in D,D_{curr}\prec D_{cons}$.

证明:$\exists t_i\in X,t_k\in R_{standard}$,已知一组时效性规则集合ΣCCs,其中有 l 条规则φ_l,且φ_l是准确无误的.t_i在属性$A=\{A_1,A_2,A_3,...,A_m\}$上的取值$V_{t_i[A_j]}$违反了ΣCFDs 中的$\varphi_n$.

(1) 当对$V_{t_i[A_j]}$,利用一条 CC $\alpha=(R_{standard}:X\to A,t_p)$以$V'_{t_i[A_j]}$进行更新时,若$V_{t_i[A_j]}=V^S_{t_i[A_j]}$,如果对于 X 上的属性A_j满足在一致性规则集合ΣCFDs 中有对于一条 CFD $\alpha=(R_{standard}:X\to A,t_p)$与其对应,则$t_i$的修复符合一致性规则,即其发生概率为$p^{cons}_{t_i[A_j]}=p^{D_{curr}}_{t_i[A_j]}$,

若$V_{t_i[A_j]}\neq V^S_{t_i[A_j]}$,根据定义 3.3 可知:根据一致性规则集合ΣCFDs 中一条 CFD $\alpha=(R_{standard}:X\to A,t_p)$,$t[X]\asymp t_p[X]$,但$t[A]\neq t_p[A]$,则$t_i$发生一致性违反,即$t_i\in\sum Vio(D_{cons})$,其发生概率为$\overline{p^{cons}_{t_i[A_j]}}=\overline{p^{D_{curr}}_{t_i[A_j]}}$;

(2) 当对足够多的$V_{t_i[A_j]}$,利用规则集合ΣCCs,以与之对应的$V'_{t_i[A_j]}$进行更新时,将满足$V_{t_i[A_j]}=V^S_{t_i[A_j]}$的$t_{i[A_j]}$个数记为$n_{correct}$,将满足$V_{t_i[A_j]}\neq V^S_{t_i[A_j]}$的$t_{i[A_j]}$个数记为$n_{error}$,计算$Q_{cons}=F\left(\sum n_{t_i}-(n_{iviolate}-n_{correct}+n_{error})/n_{t_i}\right)$,若$n_{correct}-n_{error}\geqslant 0$,则$Q_{cons}\geqslant q_{cons}$,即 X 的一致性得到提高;若$n_{correct}-n_{error}<0$,则$Q_{cons}<q_{cons}$,则根据定义 3.3 的推广定义可知,X 发生了一致性违反,即$\Delta Q^{D_{cons}}>0$的概率$P^{D_{cons}}\in(0,1)$.

综上,由定义 3.5 可知,$D_{curr}\prec D_{cons}$成立,证毕. □

综上所述,时效性与其他 3 个质量性质的关系如图 5 所示.

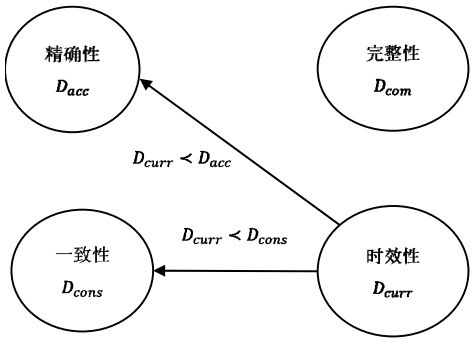


Fig.5 The relationship between currency and other dimensions

图 5 时效性与其他性质关系

3.3 基于数据质量性质关联的数据修复顺序选择

根据上文对 4 种数据质量性质的逐一分析,我们将其表示为图 6 所示的有向图形式.通过分析,当对完整性错误进行修复后,其修复结果会引起一致性、时效性、精确性性质方面的变化;对一致性错误进行修复后,其修复结果会引起时效性和精确性性质方面的变化;对时效性错误进行修复后,其修复结果会引起一致性和精确性性质方面的变化;而对精确性错误进行修复后,则不会引起其他 3 个性质的变化.

我们定义一个数据清洗的规则集合ΣRepr,当数据质量性质 D_i 上的修复处理操作要先于 D_j 上的操作时,我们记为 $t_{D_i} \prec t_{D_j}$. 因此我们根据第 3.2 节的关系得到以下结论,当ΣRepr 中考虑的性质:

$$D=\{D_{completeness}, D_{Consistency}, D_{Currency}, D_{Accuracy}\} \text{时}, t_{D_{com}} \prec t_{D_{cons}} \asymp t_{D_{Curr}} \prec t_{D_{Acc}}.$$

分析得知数据集合存在 $D=\{D_{completeness}, D_{Consistency}, D_{Currency}, D_{Accuracy}\}$ 上 4 个性质的混合型错误,或者在业务需要对数据集合进行以上 4 个性质的全面修复时,我们将分为以下 3 个步骤进行修复.

Step 1. 完整性违反项修复;

Step 2. 一致性和时效性违反项修复;

Step 3. 精确性违反项修复.

当不需要进行完整性修复时,系统只需按序执行 Step 2,Step 3;当不需要进行精确性修复时,系统只需按序执行 Step 1,Step 2;当不需要进行一致性或者时效性违反项修复时,系统需按序执行 Step 1,Step 2,Step 3,在 Step 2 的步骤中去除时效性或一致性分析与修复过程.

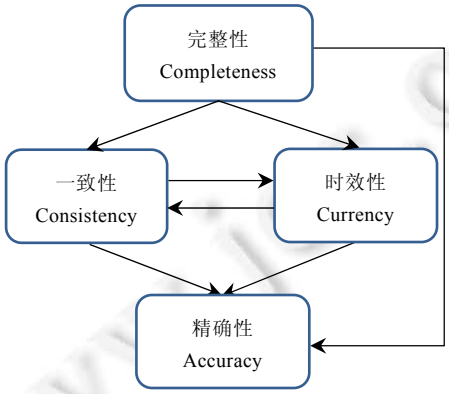


Fig.6 The relationship of multi-dimensions in data quality

图 6 多维数据质量关系图

4 实 验

根据上面的分析,为了实现保证数据清洗质量的同时,简化清洗流程并减小清洗代价,我们提出了如图 7 所示的混合型数据质量错误的数据清洗修复模型.

为了验证上述数据清洗修复流程的有效性,在以下 4 个数据集合上进行了实验测试,其规模见表 7.

(1) 英国大学地理信息真实数据集合(下文简写为 UD);

(2) 历史气候天气真实数据集合 Worldwide Historical Weather Data(下文简写为 WHWD);

(3) 意大利社会安全评测真实数据集合 Italian Social Security Contributors List(下文简写为 ISSCL);

(4) 学生毕业信息虚拟数据集合(下文简写为 SCD).

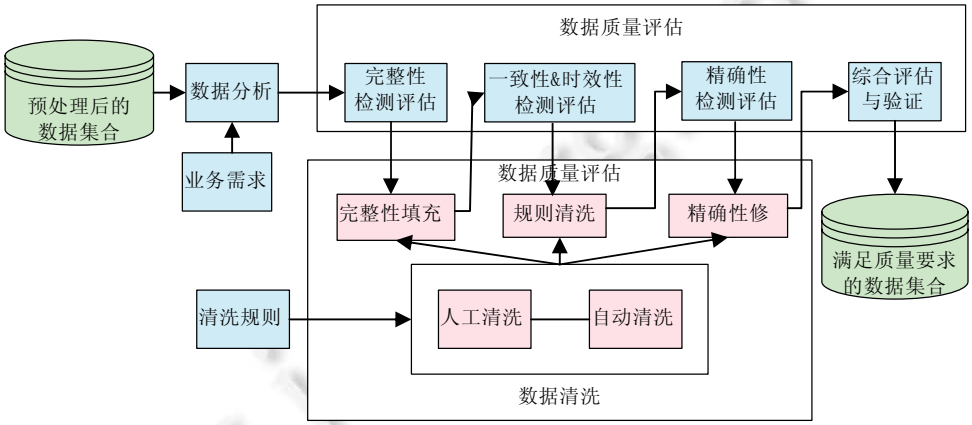


Fig.7 Comprehensive data cleaning process

图 7 综合型数据清洗流程图

Table 7 The datasets scale in experiment

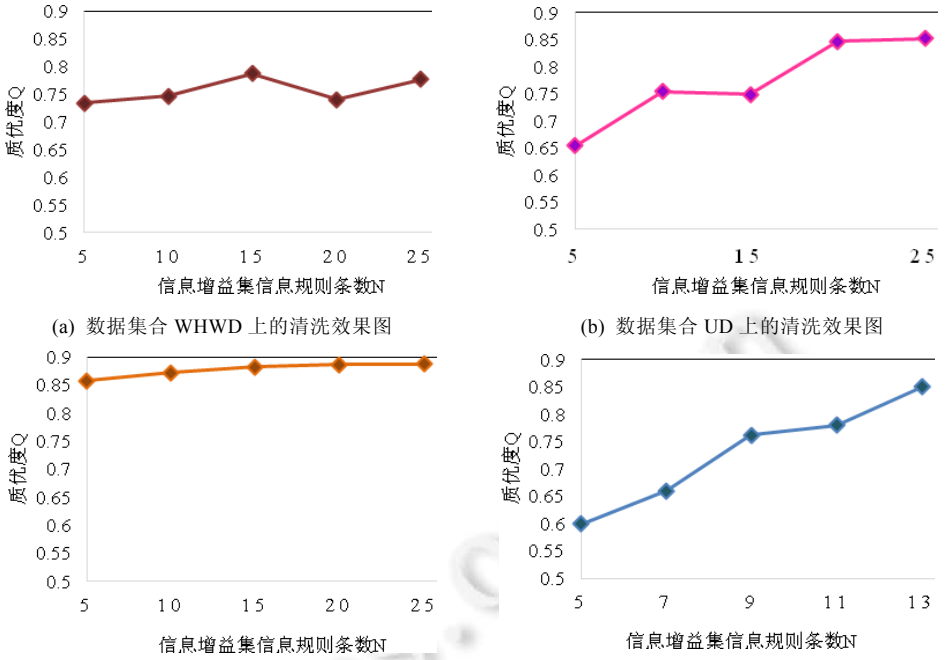
表 7 实验所用数据集合的规模示意表

	UD	WHWD	ISSCL	SCD
行数	580	20 000	1 483 712	220
列数	12	16	6	10

实验程序采用 C++编程,实验环境为内存 6GB,Inter(R) Core™ i5 CPU 处理器,64 位 Win7 操作系统.实验中,在进行实验结果分析时,我们定义了信息增益规则集合,存储针对不同性质的修复规则;采用质优度 $Q=w_1Q_{com}+w_2Q_{acc}+w_3Q_{curr}+w_4Q_{cons}$ 作为评价指标,Q 为第 3 节定义的数据质量标准系数,w 是为每种性质分配的权值.我们逐渐增加信息增益集中的规则个数进行测试,图 8 展示了在 4 个数据集合上的清洗修复效果.

在精确性方面,主要考察了错误值和孤立点的错误问题;在完整性方面,我们用空值发现来判断内容完整性.人工判断选择一组属性组作为标准属性集合,加入我们的信息增益集;在一致性和时效性方面,首先发现数据集合中的一致性规则和时效约束规则,在进行数据清洗的处理时使用这些规则,并记录修复的比例来判断一致性和时效性违反程度.

观察 4 个实验结果,发现随着清洗系统中的信息增益规则条数的增多,对数据集合的 4 个维度的修复粒度的增大,其质优度呈不同程度的增加.其中,数据集合 WHWD 上属性关联性并不强,时效性变化也并不是很大,并且数据量和属性均较为完备,因此修复效果增长较缓.在 UD 集合中,原始数据存在大量的重复、时效性低的数据,并且完备性较低,通过增加规则的个数修复,其质优度增幅较大.实验证实了本文的方法能够有效、合理地修复数据集合上的错误,并且对于原始数据集合质量较低,另外,数据集合中存在较多属性关系的数据集合处理效果更好.



(a) 数据集合 WHWD 上的清洗效果图 (b) 数据集合 UD 上的清洗效果图

(c) 数据集合 ISSCL 上的清洗效果图 (d) 数据集合 SCD 上的清洗效果图

Fig.8 Cleaning effect of different datasets

图 8 在数据集合上的清洗效果图

5 总结和展望

本文通过对数据质量的多种性质的研究,建立了数据质量多种性质模型,详细阐述并总结了数据质量在 4 种重要性质上存在的实际问题,并对这 4 个性质:完整性、精确性、一致性、时效性的违反模式给出定义,理论证明了在数据修复背景下它们之间的关联关系,基于此制定了混合型错误情况下的数据清洗修复策略,并通过实验验证了其有效性和合理性.

四维数据质量关系模型有助于数据质量的综合评估,这对于数据集合上开展数据清洗有着重要的影响力.综合评估对于数据清洗策略和具体步骤的制定在效果、效率等方面均有指导意义.我们今后的研究重点将放在四维数据质量关系模型在混合型错误的数据集合上的清洗实现,整合并优化数据清洗算法.此外,我们也将更为全面地研究关联关系理论在不同数据类型(如非结构化数据和半结构化数据)中的应用.

References:

[1] Mayer-Schonberger V, Cukier K. Big Data: A Revolution That Will Transform How We Live, Work, and Think. London: Houghton Mifflin Harcourt, 2013. 19−31.

[2] Sidi F, Shariat PPH, Affendey LS, Jabar MA, Ibrahim H, Mustapha A. Data quality: A survey of data quality dimensions. In: Proc. of the 2012 Int'l Conf. on Information Retrieval & Knowledge Management. IEEE, 2012. 300−304. [doi: 10.1109/InfRKM.2012.6204995]

[3] 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/ch/reader/view_abstract.aspx?flag=1&file_no=20021103&journal_id=jos

[4] Batini C, Cappiello C, Francalanci C, Maurino A. Methodologies for data quality assessment and improvement. ACM Computing Surveys, 2009,41(3):No.16. [doi: 10.1145/1541880.1541883]

[5] Wang RY, Strong DM. Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems, 1996,12(4):5−33. [doi: 10.1080/07421222.1996.11518099]

[6] Cong G, Fan W, Geerts F, Jia XB, Ma S. Improving data quality: Consistency and accuracy. In: Proc. of the 33rd Int'l Conf. on Very Large Data Bases. VLDB Endowment, 2007. 315−326. http://dl.acm.org/citation.cfm?id=1325890&preflayout=flat

[7] Bohannon P, Fan W, Geerts F, Jia XB, Kementsietsidis A. Conditional functional dependencies for data cleaning. In: Proc. of the 23rd IEEE Int'l Conf. on Data Engineering. Istanbul: IEEE, 2007. 746−755. [doi: 10.1109/ICDE.2007.367920]

[8] Fan W, Geerts F, Wijsen J. Determining the currency of data. ACM Trans. on Database Systems, 2012,37(4):25−41. [doi: 10.1145/2389241.2389244]

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

[10] McGilvray D. Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information. Burlington: Elsevier, 2008. 16−59.

[11] Fan W, Ma S, Tang N, Yu WY. Interaction between record matching and data repairing. Journal of Data and Information Quality, 2014,4(4):16. [doi: 10.1145/2567657]

[12] Tee SW, Bowen PL, Doyle PH. Rohde F. Factors influencing organizations to improve data quality in their information systems. Accounting & Finance, 2007,47(2):335−355. [doi: 10.1111/j.1467-629x.2006.00205.x]

[13] Eckerson W. Data quality and the bottom line, Vol.1. TDWI Report, Data Warehouse Institute, 2002. 1−31.

[14] Pipino LL, Lee YW, Wang RY. Data quality assessment. Communications of the ACM, 2002,45(4):211−218. [doi: 10.1145/505248.506010]

[15] https://en.wikipedia.org/wiki/Cronbach%27s_alpha

[16] Yue K. Data Engineering: Processing, Analysis and Service. Beijing: Tsinghua University Press, 2013. 169−180 (in Chinese).

[17] Fan W, Geerts F. Relative information completeness. ACM Trans. on Database Systems, 2010,35(4):97−106. [doi: 10.1145/1862919.1862924]

[18] Bravo L, Fan W, Ma S. Extending dependencies with conditions. In: Proc. of the 33rd Int'l Conf. on Very Large Data Bases. VLDB Endowment, 2007. 243−254. http://dl.acm.org/citation.cfm?id=1325882&CFID=627672245&CFTOKEN=70772333

附中文参考文献:

[3] 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/ch/reader/view_abstract.aspx?flag=1&file_no=20021103&journal_id=jos

[9] 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.

[16] 岳昆.数据工程——处理、分析与服务.北京:清华大学出版社,2013.169−180.

丁小欧(1993－),女,黑龙江哈尔滨人,硕士,CCF 学生会员,主要研究领域为数据质量管理,数据清洗.

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.

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.

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.

张笑影(1994－),女,学士,主要研究领域为数据质量.