属性粒度数据质量模型及其评价指标研究

陈卫东 张维明
(国防科技大学信息系统与管理学院 长沙 410073)

摘 要 在研究属性粒度关系代数传递影响的基础上,进一步从数据质量正确性和完整性两方面加以总结,提出一个数据质量评价模型,以区分5种元组质量类型,定义正确性和完整性指标同时包含数据项和元组两个方面带来的影响。通过分析空值,即不正确也不完整,建立正确性和完整性指标的相互联系,进而引入属性量化前后错误(空值)率,进一步量化定义评价指标。

关键词 数据质量,质量模型,正确性,完整性,关系数据库
中图法分类号 TP391 **文献标识码** A

Data Quality Model and Metrics Research at Attribute Granularity

CHEN Wei-dong ZHANG Wei-ming
(Information System and Management Dept. of NUDT, Changsha 410073, China)

Abstract Based on the author's previous research at data quality propagation for relational database, the paper sumed up and presented a data quality model at attribute granularity, which both considers the influence of accuracy and completeness. According to the model, the tuples were classified into five categories. The definitions of accuracy and completeness both include the influence of the data item and tuple. After analysing the null data neither correct nor complete, the mutual relationship between them was formed. Then the attribute error rate before and after quantitative was introduced to define metrics further.

Keywords Data quality, Quality model, Accuracy, Completeness, Relational database

1 引言

信息技术的不断进步和广泛应用,使数据库已经成为信息系统的重要组成部分。由于应用的日益复杂和信息量的迅速增加,数据库中出现数据错误、丢失、偏差和延时等现象越来越影响到组织的正确决策并带来经济损失。数据质量问题日益受到重视,促使人们更进一步深入研究数据库数据质量问题。

数据质量研究涉及许多方面,数据库数据质量的传递影响是其中一个重要方面。围绕这个问题,许多学者进行了深入研究。Aebi[1]从"真实-真实视图-模式"出发,形式化描述了正确性、完整性、一致性和最小性质量指标。Kon[2,3]通过比较实际数据库与真实数据库之间的差异,分析元组的质量特征由正确、错误、误属和缺失4种类型组成,并对关系代数运算形成闭集。Motro[4-6]以数据库概念模式的理想实例和存储实例之间的差异提出数据质量模型,定义健全性(Soundness)和完整性指标,定量研究SQL查询(Select)对结果集质量的传递影响。Reddy[7,8]将元组分为正确、错误以及非隶属3种类型,定义正确性、非隶属错误和隶属错误指标,研究选择、投影、笛卡尔运算质量的传递影响,定量地给出质量传递关系。Parssian[9-11]在Kon和Reddy研究基础上,提出数据质量模型,定义正确性、不正确性、误属性和不完整性评价指标,定量研究了选择、投影和笛卡尔集等运算对数据质量的传递影响。Scannapieco[12]从封闭和开放假设思想出发,研究空值对完整性指标的测度影响和评价问题,分别在元组和数据项粒度定义弱(强)元组完整性、弱(强)属性完整性、弱(强)关系完整性指标测度,对并、交和笛卡尔积运算给出空值完整性定量传递性关系。Ballou[13]从信息产品是否可接受角度定义可接受性(Acceptable)评价指标,并运用统计抽样方法研究关系代数运算对评价指标的传递影响,给出定性描述。

关系数据库数据质量传递影响的研究包括元组和属性两个粒度。元组粒度将元组作为一个整体和最小单位,元组任何数据有差错,则判定元组整体有问题。属性粒度以数据项作为最小判定单位,元组质量由其数据项共同反映。显然,元组粒度较属性粒度更为严格。在上述研究者的研究中,多从元组粒度角度展开,属性粒度研究尚不深入。作者从属性粒度出发首先研究了选择[14]、投影[15]和笛卡尔积[16]3种关系运算对正确性指标的传递影响,给出传递计算公式。本文是在上述研究基础上结合正确性和完整性指标的相互作用,对属性粒度评价模型的进一步研究和总结。

2 数据质量模型

2.1 分析

到稿日期:2009-06-18 返修日期:2009-09-10

陈卫东 男,副教授,主要研究方向为信息系统与智能决策技术,E-mail:cwdxlj@263.net;张维明 男,教授,博士生导师,主要研究方向为信息系统与智能决策技术。

关系数据库建模是将真实世界(记为 W)映射为关系模式(记为 D)的过程。从模式 D 实例化可得到理想关系(d_0)和存储关系(d)。理想关系准确、完整地包含模式 D 在真实世界中全部实体的实例。存储关系包含通过信息系统获取实体的实例。理想关系可以作为评价基准,通过与存储关系比较来确定质量问题和定义评价指标(如图 1 所示)。

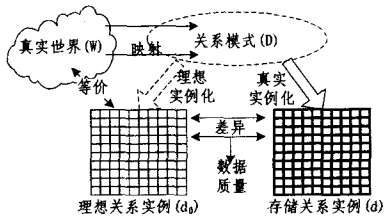

图 1　数据质量产生的原因

在数据质量评价指标中,正确性和完整性是两个基本指标,含义明确,易于鉴别。所谓正确性(Accuracy),是指数据是否真实、无误地反映客观实体属性的真实状态,表达数据值与实体属性真实值的符合程度,强调数据所表达的含义与它代表的客观实体属性的含义是否具有相同、相近(即正确)的语义;所谓完整性(Completeness),是衡量实体状态被完全映射的程度,表现为数据的形式、内容等是否完备、满足要求,更多强调数据外在的形式和表示是否完全。

正确性和完整性关系密切。一般地,越是正确的数据完整性越好,越是完整的数据正确性也越好。但是,正确的数据可以完整,也可以不完整,反之错误的数据可以完整或不完整,正确性较完整性更重要。例如,地名"长沙"可以认为不完整(通常完整为"湖南长沙"),但未必错误。又如通信地址等类型数据在正确和完整之间也普遍存在上述关系。这种正确性和完整性互相交叠现象导致问题的复杂性大大增加,一方面与具体的应用领域、评价目的和评价人素质等多种因素相关,也与主体和上下文的需要相关。为简化问题,本文提出的模型针对评价指标的客观性方面,同时考虑数据库中存在的情况,对正确性和完整性做进一步界定。

正确性与一般理解无差异,数据非对即错。完整性考虑数据库中存在空值(Null)和缺失两种情况。对于空值,一般认为实体属性值必然存在,不会出现无法确定的空值状态。从不严格的角度来看,如果数据项非空,则可认为是完整的(尽管数据的具体表示也可能不完整,但是属性存在值至少比空值的完整性要强),否则就不完整。从正确角度来看,无论如何空值不正确,即空值既不完整,又不正确;对于缺失是指存储关系中缺少了本应存储于关系中的数据。本文讨论空值主要基于 3 方面考虑:首先,正确性和完整性关系密切,使得判断具体数据是否正确、完整存在多种组合,导致问题复杂化,而空值对于正确性和完整性指标易于界定;其次,实际数据库中,空值普遍存在;第三,空值完整性计算可以实现自动化。

表 1 是实例化存储关系示例(斜体下划线表示数据有质量问题)。其中 1#,2# 元组没有错误,3# —8# 元组存在错误和不完整,9#,10# 本不属于关系但包括在存储关系中。表 2 元组原本应属于存储关系但缺失了。

表 1　员工关系实例数据

#	姓名	籍贯	地址	时间
1	舒望清	江苏	扬州市岳口陈场林场	1969.08
2	王运生	上海	宝山区辛庄村	1973.10
3	张德运	北京	怀柔县车埠镇 76 号	*(null)*
4	李可	*浙江*	番禺市成全镇农科村	1980.09
5	王平	四川	隆昌市隆化镇南卫村	*1994.07*
6	周鲁	湖南	*益阳市太平街 3 号*	1992.10
7	赵兴东	*(null)*	九江市邓苍村十一组	*2002.07*
8	张剑	*山西*	*太原市中洲农场 25 号*	*2003.07*
9	*王伟亮*	*湖北*	信阳市花桥镇	*(null)*
10	张萍	湖北	*(null)*	*2003.07*

表 2　员工关系缺失数据

#	姓名	籍贯	地址	工作时间
11	陈清	湖南	长沙市北区四方坪	1971.04
12	王二	湖南	邵阳市城南路 112 号	1969.03

2.2　评价模型

在属性粒度理想关系按照正确性和完整性实例化映射到存储关系时,可划分得到元组的质量类型。其中,正确性映射可划分为正确(S_A)、模糊正确(S_F)、错误(S_I)、误属(S_M)和缺失(S_C)5 种类型,完整性映射可划分为完整(S'_A)、模糊完整(S'_F)、不完整(S'_I)、误属(S_M)和缺失(S_C)5 种类型。于是得到图 2 所示的数据质量模型。

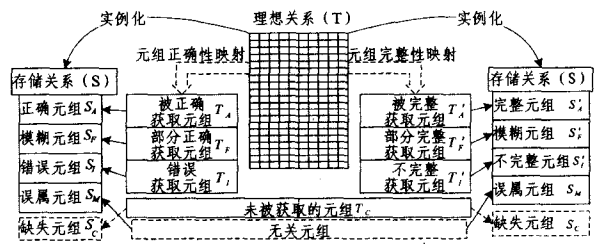

图 2　数据质量模型

记 $v_{i,j}^T$,$v_{i,j}^S$ 分别表示关系 T,S 的第 i 个元组、第 j 个属性(i 行 j 列)数据项的值。T,S 同构。不失一般性,假设关系 S 中前 $1..k$ 个属性为键属性,记为 K_S;$k+1..m$ 为非键属性,记为 Q_S。根据关系数据库理论,键属性是实体的唯一标识,于是元组质量类型可以通过键属性来描述。

定义 1(正确元组,S_A)　任意数据项都没有错误的元组构成的集合。元组 t_i 属于 S_A,当且仅当

$$S_A=T_A=\{t_i \mid \forall j,1\leqslant j\leqslant m,v_{i,j}^T=v_{i,j}^S\} \tag{1}$$

定义 2(模糊正确元组,S_F)　即元组键属性数据无误、非键属性存在数据项错误的元组构成的集合。元组并非完全正确或完全错误,而是介于正确和错误之间,具有模糊性。元组 t_i 属于 S_F,当且仅当

$$S_F=\{t_i \mid \forall r,v_{i,r}^T=v_{i,r}^S,1\leqslant r\leqslant k \text{ and } \exists j>k,v_{i,j}^T\neq v_{i,j}^S\} \tag{2}$$
$$|S_F|=|T_F|$$

定义 3(错误元组,S_I)　即任意非键属性值完全错误的元组构成的集合。键属性数据无误,非键属性数据全部错误。元组 t_i 属于 S_I,当且仅当

$$S_I=\{t_i \mid \forall r,v_{i,r}^T=v_{i,r}^S,1\leqslant r\leqslant k \text{ and } \forall j>k,v_{i,j}^T\neq v_{i,j}^S\} \tag{3}$$
$$|S_I|=|T_I|$$

定义 4(误属元组,S_M)　由本不应属于关系 S 的元组构成。由于键属性是实体的惟一标识,因此无论非键属性是否正确,只要键属性中存在错误,则该元组就是误属元组。元组 t_i 属于 S_M,当且仅当

$$S_M=\{t_i \mid 1\leqslant \exists j\leqslant k,v_{i,j}^T\neq v_{i,j}^S\},|S_M|=|T_M| \tag{4}$$

定义 5(缺失元组,S_C)　由本应在关系 S 中存储的元组

组成。元组认为是完整和正确的(未被获取而已)。元组 t_i 属于 S_C,当且仅当

$$S_C=T_C=\{t_i\mid T-S\} \tag{5}$$

完整性质量类型参考式(1)－式(5)定义,不再赘述。需要说明 S_M 和 S_C 对元组正确性和完整性划分相同。其余部分有 $S_A+S_F+S_I=S'_A+S_F'+S_I'$,$S_A\neq S_A'$,$S_F\neq S_F'$,$S_I\neq S_I'$,$S_A$、$S_F$、$S_I$ 与 S_A'、S_F'、S_I'存在交叉和重叠。

3 评价指标描述

3.1 分析

为了量化衡量表 1 的正确性和完整性,需要量化处理。数据项正确量化为 1,属性非空(完整)量化为 1,否则均为 0(如表 3 所列)。然而,由于键属性对元组类型的决定性作用,还需更进一步细致分析数据项的量化结果。从表 1 可知,9＃元组地址数据是正确的,量化为 1,如表 3 所列。但由于 9＃元组本不属于关系,整个元组对关系 S 来说无疑是错误的,量化结果应为 0,而不是 1。也就是说,表 1 中由于误属元组本不属于关系,无论 9＃,10＃的数据项是否正确,都应量化为 0,得到表 4 的量化结果。

表 3 员工关系正确性量化表(直接)

＃	姓名	籍贯	地址	时间
1	1	1	1	1
2	1	1	1	1
3	1	1	1	0
4	1	0	1	1
5	1	1	1	0
6	1	1	0	1
7	1	0	1	0
8	1	0	0	0
9	0	0	1	0
10	0	1	0	0

表 4 员工关系正确性量化表(修正)

＃	姓名	籍贯	地址	时间
1	1	1	1	1
2	1	1	1	1
3	1	1	1	0
4	1	0	1	1
5	1	1	1	0
6	1	1	0	1
7	1	0	1	0
8	1	0	0	0
9	0	0	0	0
10	0	0	0	0

可以看出,直接量化结果反映关系数据项的正确性,修正量化结果反映关系的正确性。关系正确性评价应以表 4 为准(类似对完整性分析,略)。

为区分表 3 和表 4 存在的差异,引入属性量化前错误率(记为 E_j)和属性量化后错误率(记为 E_j')。由元组的质量类型分析可得非键属性 A_j 量化前后的数据构成,如图 3 所示。

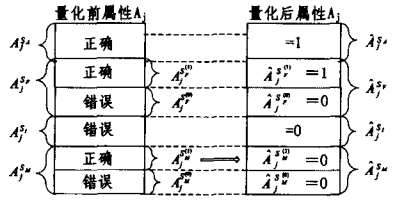

图 3 数据质量模型

定义 6(属性量化前错误率,E_j) 即关系 S 的属性 A_j 在量化之前非空错误数据项个数与总个数之比,表示为

$$E_j=\frac{|A_F^{S(0)}|+|A_I^{S}|+|A_M^{S(0)}|}{|A_j|}-N_j \tag{6}$$

式中,N_j 为属性量化前空值率(定义为量化前属性空值数据项个数与属性总数据项个数之比,公式略)。

属性量化前错误率定义中没有包含空值数据项。这是由于属性空值率不仅易于计算,而且不含空值的错误率更能反映属性正确性分布(空值仅仅为数据库管理系统中的一种特殊数据类型,实际中并不存在)。

定义 7(属性量化后错误率,E_j') 即关系 S 的属性 A_j 在量化之后错误数据项(含空值错误)个数与总个数之比,表示为

$$E_j'=\frac{|\hat{A}_F^{S(0)}|+|\hat{A}_I^{S}|+|\hat{A}_M^{S}|}{|\hat{A}_j|}=\frac{|A_F^{S(0)}|+|A_I^{S}|+|A_M^{S}|}{|A_j|}$$

对于完整性指标,类似地可以定义量化前空值率(N_j)和量化后空值率(N_j')为量化前后属性空值数据项个数与属性总数据项个数之比(公式略)。

如表 1 中属性列量化前错误率分别为(0.2,0.4,0.3,0.6),量化后错误率分别为(0.2,0.5,0.4,0.6)。

属性量化前后错误率的含义分别是量化前错误率反映数据的质量分布,量化后错误率反映数据的质量评价,这说明属性粒度下数据的质量分布和数据的质量评价并不等价。质量分布不能直接作为评价的结果,而一般直觉上认为质量分布就直接可以评价数据,事实上并非如此。为了得到二者的数量关系,这里假设属性出现错误和空值具有随机性,进一步可以证明

$$E_j'=\gamma_S+(1-\gamma_S)(E_j+N_j) \tag{8}$$

$$N_j'=\chi_S+(1-\chi_S)N_j \tag{9}$$

由式(8)、式(9)可以看出,由于误属和缺失元组影响,属性的错误率(空值率)会增大,这将导致关系正确性(完整性)进一步变差,质量评价结果有恶化的趋势。

3.2 指标定义

在元组质量类型中,由于键属性的作用产生了误属和缺失元组。误属元组影响正确性,定义误属性(记为 γ_S)指标反映:$\gamma_S=|S_M|/|S|$。缺失元组影响完整性,定义缺失性(记为 χ_S)指标反映:$\chi_S=|S_C|/|T|=|S_C|/(|S|-|S_M|+|S_C|)$。显然,误属性和缺失性是元组粒度的定义。如表 1、表 2 中,$\gamma_S=0.2$,$\chi_S=0.2$。

另一方面,关系中数据项和元组的数量很大,单纯通过数据项或元组来定义正确性并不合适,因此用数量较少的属性来定义,关系正确性可以用属性正确性反映。属性正确性(α_j)定义为属性量化后正确数据项个数与属性总数据项之比,即 $\alpha_j=1-E_j'$。考虑到关系属性又分为键属性和非键属性两部分,且不相交,于是定义键属性和非键属性正确性(完整性定义类似,略)。

定义 8(键属性部分正确性,α_{K_S}) 即关系 S 键属性部分正确性占总属性部分的比率,表示为

$$\alpha_{K_S}=\frac{1}{m}\sum_{j=1}^{k}(1-E_j') \tag{10}$$

可以证明任意键属性 A_j 正确性相等,即 $\alpha_j=1-\gamma_S(1\leqslant j\leqslant k)$,进而可得 $\alpha_{K_S}=\frac{k}{m}(1-\gamma_S)$。

定义 9(非键属性部分正确性,α_{Q_S}) 即关系 S 非键属性部分占总属性部分的比率,表示为

$$\alpha_{Q_S}=\frac{1}{m}\sum_{j=k+1}^{m}(1-E_j') \tag{11}$$

关系 S 正确性显然由键属性和非键属性两部分正确性组成,即 $\alpha_S=\alpha_{K_S}+\alpha_{Q_S}$。

对于完整性(定义略),亦有键属性 A_j 的完整性 $\beta_j=1-\chi_S$,键属性部分有 $\beta_{K_S}=\frac{k}{m}(1-\chi_S)$,关系完整性有 $\beta_S=\beta_{K_S}+\beta_{Q_S}$。

由式(8)、式(10)和式(11)可得

$$\alpha_S=\alpha_{K_S}+\alpha_{Q_S}=\frac{1}{m}(1-\gamma_S)(k+\sum_{j=k+1}^{m}(1-E_j-N_j)) \tag{12}$$

$$\beta_S=\beta_{K_S}+\beta_{Q_S}=\frac{1}{m}(1-\chi_S)(k+\sum_{j=k+1}^{m}(1-N_j)) \tag{13}$$

由式(12)、式(13)可以看出,关系正确性(完整性)评价计算与误属性(缺失性)和错误(空值)的分布相关,这样就可以利用统计抽样方法推断获得(空值率可以自动计算),从而有利于评价的实施。

结束语 本文是在文献[14-16]研究基础上的进一步深入和总结。属性粒度数据质量评价模型从正确性映射和完整性映射两个方面分别分析和描述了元组的质量类型。正确性指标定义包含数据项错误和元组误属两个方面的影响,完整性指标定义包含数据项空值和元组缺失带来的影响。通过分析空值即不正确也不完整,在模型中建立了正确性和完整性指标相互联系,进而引入属性量化前后错误(空值)率,进一步量化定义评价指标。

笔者还将以数据质量评价模型为基础继续就关系代数运算对正确性和完整性评价指标的传递影响的理论方面进行更为深入的研究。

参 考 文 献

[1] Aebi D, Perrochon L. Towards Improving Data Quality[C]// Proceedings of the International Conference on Information Systems and Management of Data. 1993:273-281

[2] Kon H B. A process view of data quality (TDQM working paper)[M]. Total Data Quality Management Research Program. Sloan School of Management, Massachusetts Institute of Technology, 1993:17

[3] Kon H B, Madnick S E, Siegel M D. Good Answers from Bad Data: A Data Management Strategy[R]. Massachusetts Institute of Technology (MIT), Sloan School of Management, 1995:1-16

[4] Motro A, Rakov I. Estimating the Quality of Data in Relational Databases[C]// 1996 Conference on Information Quality. Cambridge, Massachusetts, 1996:94-106

[5] Motro A, Rakov I. Estimating The Quality of Databases[C]// The 3rd International Conference on Flexible Query Answering Systems (FQAS). Cambridge, MA, 1998:298-307

[6] Motro A, Rakov I. Not All Answers Are Equally Good: Estimating the Quality of Database Answers[M]. Kluwer Academic Publishers, 1997:1-21

[7] Reddy M P, Wang R Y. A Data Quality Algebra for Estimating Query Result Quality[C]//CISMOD, Conference. Bombay: 1996

[8] Wang R Y, Ziad M, Lee Y W. Data Quality [M]. New York: Kluwer Academic Publishers, 2002

[9] Parssian A, Sarkar S, Jacob V S. Assessing Data Quality for Information Products: Impact of Selection, Projection, and Cartesian Product[J]. Management Science, 2004, 50(7):967-982

[10] Parssian A, Sarkar S, Jacob V S. Assessing Information Quality for the Composite Relational Operation Join[C]// The 7th International Conference on Information Quality. MIT Cambridge, MA USA, 2002:225-237

[11] Parssian A, Sarkar S, Jacob V S. Assessing data quality for information products. [C]// ICIS-99. 1999:428-433

[12] Scannapieco M, Batini C. Completeness in the Relational Model: a Comprehensive Framework[C]// 9th International Conference on Information Quality. Cambridge, MA, USA, 2004:333-345

[13] Ballou D P, Chengalur-Smith I N, Wang R Y. Sample-based Quality Estimation of Query Results in Relational Database Environments[J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(5):639-650

[14] 陈卫东,张维明. 数据质量模型及选择运算中的质量传播研究[J]. 计算机工程与应用,2007,43(27):1-3

[15] 陈卫东,张维明. 投影运算对数据库数据质量的传递影响研究[J]. 计算机应用研究,2008,25(9):2751-2753

[16] 陈卫东,张维明. 笛卡尔积运算对数据库数据质量的传递影响[J]. 计算机科学,2008,35(6):210-212

(上接第 114 页)

[2] Gao D, Reiter M K, Song D. Behavioral distance measurement using hidden markov models[C]// Zamboniand D, Kruegel C, eds. Research Advances in Intrusion Detection, LNCS 4219. Berlin Heidelberg: Springer-Verlag, 2006:19-40

[3] Ghosh A, Schwartzbard A. A study in using neural networks for anomaly and misuse detection[C]// Proceedings of the 8th USENIX Security Symposium. 1999

[4] Giffin J T, Jha S, Miller B P. Efficient context-sensitive intrusion detection[C]// Proceedings of Symposium on Network and Distributed System Security. 2004

[5] Wagner D, Dean D. Intrusion detection via static analysis[C]// Proceedings of the 2001 IEEE Symposium on Security and Privacy. May 2001:156-168

[6] Gao D, Reiter M K, Song D. Gray-box extraction of execution graphs for anomaly detection[C]// Proceedings of the 11th ACM Conference on Computer & Communication Security (CCS 2003). 2003

[7] Gao D, Reiter M K, Song D. On gray-box program tracking for anomaly detection[C]// Proceedings of the 13th USENIX Security Symposium. 2004

[8] Sharif M S, Singh K, Giffin J, et al. Understanding precision in host based intrusion detection[C]// Proceedings of the International Symposium on Recent Advances in Intrusion Detection (RAID). 2007:21-41

[9] Buck B, Hollingsworth J K. An API for Runtime Code Patching [J]. The International Journal of High Performance Computing Applications, 2000(14):317-329

[10] Parampalli C, Sekar R, Johnson R. A practical mimicry attack against powerful system-call monitors[C]// Proceedings of the 2008 ACM Symposium on information, Computer and Communications Security. March 2008:156-167

[11] Lu Wei, Zeng Qingkai. A control flow based program behavior extended model[J]. Journal of software, 2007, 18(11):2841-2850