

信息产品质量测量动态衍变模型研究

刘 婧 赵高正

(西北工业大学管理学院 西安 710072)

摘 要 信息产品质量测量是信息产品质量全面管理 Total Data Quality Management (TDQM) 的重要组成部分, 为日后信息质量提升奠定了基础。基于信息产品质量测量静态传递模型, 在关系数据库基础上, “选择”关系代数作用下, 引入了时间变量, 分析了信息产品及时性对信息产品质量的影响, 构建了信息产品质量测量动态衍变模型。该模型包括信息产品在 t 时刻的及时性矩阵、过时数据集质量衍变映射、基于过时数据集质量衍变映射模型的信息产品质量测量方法。利用应用实例验证了本模型的可行性和有效性。

关键词 信息产品, 质量测量, 选择关系代数, 动态衍变

中图分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.1.054

Research on Dynamic Development Model of Information Product Quality Assessment

LIU Jing ZHAO Song-zheng

(School of Management, Northwestern Polytechnical University, Xi'an 710072, China)

Abstract Information product quality assessment is an important part of Total Data Quality Management (TDQM), lays foundation for potential quality improvement. In this paper, a novel dynamic development model of information product quality assessment that applies to the “Selection” relational algebra for the relational database was established by introducing time variable and analyzing the impact of information product timeliness based on the static propagation model of information product quality assessment. It includes developing the timeliness matrix at time t , proposing dynamic development mapping model of out-of-date dataset and developing a method for information product quality assessment based on the mapping model. The model dataset was used to verify the feasibility and validity of the model.

Keywords Information product, Quality assessment, “Selection” relational algebra, Dynamic development

1 引言

信息系统在企业战略、战术、运营各层次决策中都发挥着重要作用。然而, 通常存储在信息系统或数据仓库中的数据不可能是完美的^[1]。有资料显示, 由低质量的数据产生的成本占到了典型企业收入的 8%~12%, 占到了服务型企业支出的 40%~60%^[2]。低质量的数据会导致高成本、低满意度, 滋生组织内的不信任, 影响决策质量, 阻碍企业重组, 并影响企业长期战略的实施^[3]。相反, 高质量的数据可使企业数据资源可重用性大大增加, 增加企业收益, 有效提升企业效率, 提高决策质量, 为商业智能奠定基础^[4,5]。数据质量的研究不仅在商业领域发挥着重要作用, 在医疗领域也受到越来越多的关注, 如病人地址准确及时的分析、EHR 数据质量的测量^[6,7]。但是识别和改正所有的数据错误是不切实际的, 因为需要耗费大量的企业资源。行之有效的方式是对信息产品质量进行测量评估, 根据测量结果制定合理的质量来提升目标, 评估不同提升方案可能产生的影响^[8], 采取相应的提升措施^[1], 尤其是在异构数据源集成过程中, 如何保证数据质量是亟需研究的问题^[9,10]。因此, 对信息产品质量测量的研究

是十分必要的。

Wang^[11]认为应将信息作为产品, 即信息产品 (Information Product) 进行管理。Ballou^[12]和 Shankaranarayanan^[13,14]分别提出了 IMS (Information Manufacturing System) 和 IP-MAP (Information Product Map) 系统来描述信息产品的生产过程, 认为信息产品的质量是由源数据与信息产品的生产过程共同决定的。本文研究在“选择”关系代数作用下输出的信息产品随时间的推移其数据质量的衍变模型, 因此定义研究对象为“信息产品质量”, 本文对信息质量、信息产品质量、数据质量不予区分, 认为三者可以互换^[15]。

国内外学者对关系数据库下信息产品质量的测量进行了广泛的研究。Kon^[16]构建了查询条件下错误传播的表示模型 EPC (Error propagation calculus), 将错误类型分为归属描述和属性描述, 分别包括误属、不完整、不准确 (包含空值), 定义了错误表示的三元数组 $re = \langle rem, rea, rei \rangle$, 分析了选择和投影作用下, 基于条件概率的错误表示。Reddy^[17]分析了误属等常见的数据质量错误, 将误属包含在了不准确的范围内, 定义了元组准确、关系准确性、元组准确性概率、属性准确性概率, 并提出了在选择、投影、笛卡尔积、并、差等关系代数下关

到稿日期: 2013-12-01 返修日期: 2014-05-06 本文受国家自然科学基金项目 (71172124) 资助。

刘 婧 (1988—), 女, 博士生, 主要研究方向为管理信息系统、信息质量, E-mail: liujing2968@163.com; 赵高正 (1961—), 男, 博士, 教授, 主要研究方向为管理信息系统。

系准确性的计算方法。在 kon 和 Reddy 工作的基础上, Parsian^[18-21] 研究了在关系代数的作用下, 以元组为粒度, 利用源数据及信息产品的生产过程计算输出的信息产品的质量(准确性和完整性), 分别分析了选择、投影、笛卡尔积^[18]、并、差^[19]、聚合函数^[20]、连接^[21]下的质量传递。Parsian 的主要贡献为明确识别了实例主键在输入输出质量传递过程中的影响。国内学者陈卫东^[22] 将 Parsian^[18] 的模型扩展至属性粒度, 认为只有元组中所有属性全部为不准确数据时, 该元组才能认为是不准确的, 否则为模糊元组, 分析了在属性粒度下, 数据质量在选择^[23]、笛卡尔积^[24]、投影^[25]等运算下的传递规律, 并提出了数据质量评价模型^[26]。

以上研究多针对准确性、完整性等信息质量维度, 对于信息产品及时性的研究, 以 Ballou 和 Bernd Heinrich 较为突出。信息产品及时性是指现实世界发生的改变与信息系统中产生变化之间的延误^[27], 是信息不过时的程度^[2]。Ballou^[28] 在属性粒度上, 提出了源数据及时性的测量方法为 $Timeliness = \{ \max[(1 - \text{currency/shelf-life}), 0] \}^s$, 但该方法没有聚合性^[29], 即不适用于元组粒度或关系粒度。Bernd Heinrich^[29-32] 从概率的角度出发, 提出了测量信息产品及时性的方法^[30]。文献[31]对文献[30]进行了扩展, 考虑了补充性数据对信息及时性的影响。文献[32]指出了文献[30, 31] 的局限性, 认为不应事先确定属性生命周期的概率分布, 而应根据真实样本数据进行模拟, 选择最适合的概率分布。文献[29, 32]提出了测量信息产品及时性的过程和对测量方法的要求, 并将其应用到了客户关系管理(CRM)领域。

通过对国内外信息产品质量测量研究的分析, 可以看出, 目前的研究主要集中在静态测量方法的建立及在关系代数下输入输出信息质量的传递过程。对信息产品质量随着时间推移的衍变过程研究较少。因此, 本文主要研究在关系数据库基础上, “选择”关系代数作用下产生的信息产品 R, 其数据质量随时间的动态衍变模型, 即信息产品及时性对数据质量的影响。由于篇幅有限, 本文只研究选择条件作用在非主键属性^[31]且过时数据项在非主键的情况。

2 信息产品质量测量静态传递模型

本文主要参考文献[18]。该文献从元组粒度将记录区分为准确、不准确、误属和不完整 4 种类型。

- (1) 如果元组所有的属性值都是正确的, 则该元组是准确的。
- (2) 如果元组的主键是正确的, 但是在非主键属性上至少存在一个不正确或空的属性值, 则该元组是不准确的。
- (3) 如果一个元组不应该被存储在 S, 但实际存储在 S 中, 则该元组是误属的。
- (4) 如果一个元组应该被存储在 S, 但实际没有, 则该元组是不完整的。

文献[18]中, 选择条件作用在非主键属性上的信息产品质量测量静态传递模型如图 1 所示。该模型虽然描述了选择结果记录集 R 相对 S 的质量传递过程, 但只在一定时间内甚至选择发生时刻有效。随着时间的推移, R 中某些数据会过时, 不能正确反映现实世界的变化, 因此 R 的质量特征会随之发生变化。若 CRM 管理中 R_A 元组 π 所指实体的 Email 发生变化, 但 R 中相应存储属性值并未相应改变, 即该 Email

值过时, 则 π 不应继续存在 R_A 数据集中。因此建立信息产品 R 的质量基于数据及时性的随时间推移的动态衍变模型是必要的。

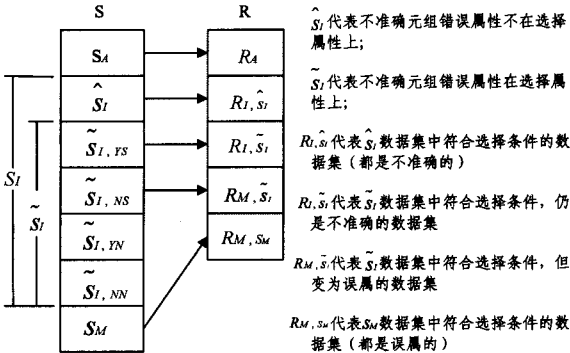


图 1 信息产品质量测量静态传递模型

3 信息产品质量测量动态衍变模型

3.1 信息产品 R 在 t 时刻及时性矩阵的构建

Ballou^[28] 提出了数据项及时性的测量方法。设信息产品递送给客户的时间为 t_D , 该数据进入数据库的时间为 t_I , 数据进入数据库时已有“年龄”为 a , 数据项的生命周期时间为 L 。则位于第 i 个元组第 j 个属性上的数据 d 的及时性为:

$$T(d(i, j)) = \{ \max[(1 - ((t_D - t_I) + a)/L), 0] \}^s \quad (1)$$

容易腐坏的产品, 如食物, 都有保质期的概念, 在这一段时间内该产品会全价出售, 并且认为在保质期内食物不会有衰变或变质。与这些产品一样, 源数据的提供者或数据管理者决定数据的“保质期”, 即在该时间段内, 该数据始终保持有效。数据的“保质期”即称为数据项的生命周期 L 。不同的数据项, 其生命周期 L 不同, 如股票、货币汇率等非常易变的数据, 其生命周期是很短的, 而如某人的出生日期等信息, 其生命周期为无穷大。一般数据的生命周期 L 由信息质量管理者在咨询信息产品消费者和参考必要的统计数据的基础上制定^[28]。

本文对式(1)进行扩展, 提出 t 时刻数据项及时性为:

$$T(d(i, j)) = \{ \max[(1 - ((t - t_I) + a)/L), 0] \}^s \quad (2)$$

由式(2)可知, 数据项的及时性值域为 $[0, 1]$ 。及时性为 0 表示该数据项已过时, 此时的数据项成为过时数据项。及时性为 1 说明该数据是最新的。对每一个数据项进行计算, 可以得到信息产品 R 的及时性矩阵。值得说明的是: 1) 如果在数据进入数据库时, 时间戳 t_I 和 a 等数据完整, 那么及时性矩阵可以通过计算机程序自动计算。2) 如果信息产品 R 中主键过时, 则整个元组各数据项的及时性为 0。3) 如出生年月等不会过时的固有属性, 其生命周期时间 L 为 $+\infty$ 。信息产品 R 及时性矩阵的示例见表 1。

表 1 信息产品 R 及时性矩阵示例

顾客编码	姓名	出生年月	地址	电话	Email	收入
1	0.7	1	0	0.5	0.9	0.7
1	0.8	1	0.7	0.6	0	0.7
1	0.7	1	0.8	0.6	1	0.7
0	0	0	0	0	0	0

3.2 过时数据集质量衍变映射模型

定义 1 A 为元组中的属性, w 为元组中的属性值。如果元组中 $\exists T(A, w) = 0$, 则该元组是过时的; 否则, 称该元组

是及时的。在信息产品 R 中,由及时元组组成及时数据集 U,由过时元组组成过时数据集 O。

定义 2 π 为任一元组,如果属性 A 中 $\exists T(\pi, w)=0$,则称

属性 A 是过时的。属性 A 的及时性 $U_{curr} \cdot (A) = \frac{\sum_{i=1}^{|R|} T(i, A)}{|R|}$ 。

属性 A 的过时性为 $O_{curr} \cdot (A) = \frac{\sum_{i=1}^{|R|} (1-T(i, A))}{|R|}$ 。

表 2 过时数据集质量衍变映射模型

	映射条件	映射结果	注释	基
O _A	1 作用在查询条件上且最新数据满足查询条件	$W_1 O_A \partial$	不准确	0.07
	2 作用在查询条件上且最新数据不满足查询条件	$W_M O_A \bar{\partial}$	误属	0.02
	3 没有作用在查询条件上	$W_1 O_A \hat{\partial}$	不准确	0.56
O _{I, S_I}	4 作用在查询条件上且最新数据满足查询条件	$W_1 O_{I, S_I} \partial$	不准确	0
	5 作用在查询条件上且最新数据不满足查询条件	$W_M O_{I, S_I} \bar{\partial}$	误属	0
	6 作用在错误属性上且最新数据与存储错误数据相同	$W_A O_{I, S_I} I$	准确	0
	7 作用在错误属性上且最新数据与存储错误数据不同	$W_1 O_{I, S_I} \bar{I}$	不准确	0
	8 没有作用在错误属性上	$W_1 O_{I, S_I} \hat{I}$	不准确	0
O _{I, S_I}	9 作用在查询条件和错误属性上且最新数据与存储错误数据相同	$W_A O_{I, S_I} I$	准确	0
	10 作用在查询条件和错误属性上且最新数据与存储错误数据不同且满足查询条件	$W_1 O_{I, S_I} \partial$	不准确	0.02
	11 作用在查询条件和错误属性上且最新数据不满足查询条件	$W_M O_{I, S_I} \bar{\partial}$	误属	0.004
	12 没有作用在查询条件上	$W_1 O_{I, S_I} \hat{\partial}$	不准确	1
O _{M, S_M}	13 作用在查询条件和错误属性上且最新数据满足查询条件	$W_1 O_{M, S_I} \partial$	不准确	0.01
	14 作用在查询条件和错误属性上且最新数据不满足查询条件	$W_M O_{M, S_I} \bar{\partial}$	误属	0.002
	15 没有作用在查询条件上	$W_1 O_{M, S_I} \hat{\partial}$	不准确	0.5
O _{M, S_M}		$W_M O_{M, S_M}$	误属	1

设在信息产品质量测量静态传递模型中 $R_A, R_{I, S_I},$

$$\beta_R = \frac{|U_{I, S_I}| + |U_{I, \sim I}| + |W_1 O_A \partial| + |W_1 O_A \hat{\partial}| + |W_1 O_{I, S_I} \partial| + |W_1 O_{I, S_I} \bar{\partial}| + |W_1 O_{I, S_I} I| + |W_1 O_{I, S_I} \bar{I}| + |W_1 O_{I, S_I} \hat{I}| + |W_1 O_{I, S_I} \partial| + |W_1 O_{I, S_I} \bar{\partial}| + |W_1 O_{M, S_I} \partial| + |W_1 O_{M, S_I} \bar{\partial}|}{|R|}$$

$$\mu_R = \frac{|U_{M, S_I}| + |U_{M, S_M}| + |W_M O_A \bar{\partial}| + |W_M O_{I, S_I} \bar{\partial}| + |W_M O_{I, S_I} \partial| + |W_M O_{M, S_I} \bar{\partial}| + |W_M O_{M, S_I} \partial|}{|R|}$$

其中, $|U_A|, |U_{I, S_I}|, |U_{I, \sim I}|, |U_{M, S_I}|, |U_{M, S_M}|$ 可由 3.1 节得到。因此本部分只对过时数据集映射结果的数据集的基进行探讨,从概率的角度出发,分别分析相关映射条件发生的概率,进而分析每一项映射条件发生的概率。

3.3.1 相关映射子条件发生的概率

1) 过时元组中过时数据项作用在属性 A 的概率:

$$P(O, A) = \frac{O_{curr} \cdot (A)}{\sum_{j=1}^{|A|} O_{curr} \cdot (j)} = \frac{\sum_{i=1}^{|R|} (1-T(i, A))}{|R|} / \frac{\sum_{j=1}^{|A|} O_{curr} \cdot (j)}{|R|}$$

2) A 属性是查询条件的概率: $P(A, Q) = \frac{p}{|A|}$ 。

$R_{I, S_I}, R_{M, S_I}, R_{M, S_M}$ 分别对应的及时数据集为 $U_A, U_{I, S_I}, U_{I, S_I}, U_{M, S_I}, U_{M, S_M}$; 过时数据集为 $O_A, O_{I, S_I}, O_{I, S_I}, O_{M, S_I}, O_{M, S_M}$ 。以过时数据项是否作用在查询条件上(对应的最新数据是否满足查询条件)、是否作用在错误属性上为区分条件,构建过时数据集质量衍变映射模型,如表 2 所列。假设“选择”关系代数的查询条件所在的属性集合 $A = \{A_1, A_2, A_3, \dots, A_k\}$, 其中, $A_1, A_2, \dots, A_k$ 代表具体属性;某存储在数据库中的元组 π 的过时数据所属属性为 $B_1, B_2, \dots, B_i$, 如 $B_i \in A$, 则称该过时数据作用在查询条件上;元组 π' 为元组 π 对应的实际中的实例,二者之间存在唯一的映射关系,如 $\pi' \in \sigma_{A_1 \wedge A_2 \wedge A_3 \wedge A_k}(s)$, 则称最新数据满足查询条件;反之,则称最新数据不满足查询条件。假设数据集 S 中元组 π 为不准确记录,即 $\pi \in S_I$, 且错误数据项所在的属性集合为 $A' = \{A_1', A_2', A_3', \dots, A_k'\} (k \geq 1)$ 。如果 $B_i \in A'$, 则称过时数据项作用在错误属性上。

因篇幅限制,本文只研究选择条件在非主键属性且过时数据项作用在非主键属性的情况。

关于 $O_A, O_{I, S_I}, O_{I, S_I}, O_{M, S_I}, O_{M, S_M}$ 对应的 $R_A, R_{I, S_I}, R_{I, S_I}, R_{M, S_I}, R_{M, S_M}$ 的含义参考图 1。根据这些定义可以推断:若 O_{I, S_I} 中最新数据与错误数据相同,则肯定满足查询条件; O_{I, S_I} 过时数据项没有作用在查询条件上,则必定不会作用在错误属性上; O_{M, S_I} 中最新数据与错误数据不可能相同。

3.3 基于过时数据集质量衍变映射模型的信息产品质量测量相关符号定义如表 3 所列。

表 3 相关符号表示

符号	α_{Rt}	β_{Rt}	μ_{Rt}	QR	P
含义	R 数据集在 t 时刻的准确性	R 数据集在 t 时刻的不准确性	R 数据集在 t 时刻的错误属性	R 数据集非主键个数	查询条件列的个数
符号	$ U $	$ O $	$ R $	$ D $	
含义	R 中及时数据集的基	R 中过时数据集的基	R 数据集的基	A 属性值域的个数	

结合 Parssian^[14]对相关质量的测量方法,信息产品 R 在 t 时刻的相关质量为:

$$\alpha_R = \frac{|U_A| + |W_A O_{I, S_I} I| + |W_A O_{I, S_I} \bar{I}|}{|R|} \quad (3)$$

$$\beta_R = \frac{|W_1 O_{I, S_I} \partial| + |W_1 O_{I, S_I} \bar{\partial}| + |W_1 O_{I, S_I} I| + |W_1 O_{I, S_I} \bar{I}| + |W_1 O_{I, S_I} \hat{I}| + |W_1 O_{I, S_I} \partial| + |W_1 O_{I, S_I} \bar{\partial}| + |W_1 O_{M, S_I} \partial| + |W_1 O_{M, S_I} \bar{\partial}|}{|R|} \quad (4)$$

$$\mu_R = \frac{|U_{M, S_I}| + |U_{M, S_M}| + |W_M O_A \bar{\partial}| + |W_M O_{I, S_I} \bar{\partial}| + |W_M O_{I, S_I} \partial| + |W_M O_{M, S_I} \bar{\partial}| + |W_M O_{M, S_I} \partial|}{|R|} \quad (5)$$

Parssian^[18]提出了两个假设:

假设 1 每一个正确的属性值是值域中的随机数。

假设 2 R 中不准确数据不是由系统性的原因造成的。

3) 基于假设 1、假设 2 可推断过时数据项对应的最新数据满足查询条件的概率: $P(a \in \delta) = \frac{|R|}{|S|}$ 。

4) 依据 Parssian^[18]命题 1 可知非主键属性 A 的不准确性即 A 属性是错误属性的概率为:

$$P(I, A) = 1 - (\alpha_R / (\alpha_R + \beta_R))^{1/q_R}$$

具体推导过程详见文献[18]。

5) 过时数据项对应的最新数据与存储的错误数据相同的概率: $P(a = prestore) = 1 / |D|$ 。

3.3.2 每一项映射条件发生的概率

假设 3 映射子条件:属性 A 过时,属性 A 为查询条件,属性 A 中包含错误属性值,过时数据项对应的最新数据与存储的错误数据相同,过时数据项对应的最新数据满足查询条件,它们是相互独立的。

依据这一假设,在表 2 中提到的各项映射条件发生的概率如表 4 所列。

表 4 映射条件发生概率

	发生概率
1	$P(1)=P(O,A) * P(A,Q) * P(a \in \delta)$
2	$P(2)=P(O,A) * P(A,Q) * [1-P(a \in \delta)]$
3	$P(3)=P(O,A) * [1-P(A,Q)]$
4	$P(4)=P(O,A) * P(A,Q) * [1-P(I,A)] * P(a \in \delta)$
5	$P(5)=P(O,A) * P(A,Q) * [1-P(I,A)] * [1-P(a \in \delta)]$
6	$P(6)=P(O,A) * P(I,A) * [1-P(A,Q)] * P(a=prestore)$
7	$P(7)=P(O,A) * P(I,A) * [1-P(A,Q)] * [1-P(a=prestore)]$
8	$P(8)=P(O,A) * [1-P(I,A)] * [1-P(A,Q)]$
9	$P(9)=P(O,A) * P(A,Q) * P(I,A) * P(a=prestore)$
10	$P(10)=P(O,A) * P(A,Q) * P(I,A) * [1-P(a=prestore)] * P(a \in \delta)$
11	$P(11)=P(O,A) * P(A,Q) * P(I,A) * [1-P(a=prestore)] * [1-P(a \in \delta)]$
12	$P(12)=P(O,A) * [1-P(A,Q)] * [1-P(I,A)]$
13	$P(13)=P(O,A) * P(A,Q) * P(I,A) * P(a \in \delta)$
14	$P(14)=P(O,A) * P(A,Q) * P(I,A) * [1-P(a \in \delta)]$
15	$P(15)=P(O,A) * [1-P(A,Q)] * [1-P(I,A)]$

3.3.3 计算映射结果相关数据集的基

由于篇幅限制,本文只以 $W_1O_A \partial$ 为例介绍映射结果相关数据集基的计算方法 $|W_1O_A \partial| = |O_A| * P(1)$,其他数据集类似。

将映射结果相关数据集的基代入式(3)一式(5)即可计算出信息产品 R 在 t 时刻的准确性、不准确性及误属性。

4 应用实例

本文以文献[18]中的数据集为应用对象,人为增加了不准确数据集中不准确的数据项,以便验证本模型的正确性。在该数据集中,Rec# 为主键,该数据集共有 7 个非主键属性(“Status”不是该数据集的属性列)。表 5 数据集 S 表示存储在数据库中的数据集,表 6 表示数据集 S 中错误的数据项。

如查询条件为“Street 中门牌号 ≥ 1500 ”,则对应于图 1 所示的 R 中各组成部分 $R_A, R_{I, s_1}, R_{I, \sim s_1}, R_{M, \sim s_1}, R_{M, s_M}$ 包含的记录集如表 7 “选择”条件下 R 各组成部分所列。随着时间的推移,依据 3.1 节信息产品 R 在 t 时刻及时性矩阵的构建方法,构建的及时性矩阵如表 8 所列。其中, $A_1, A_2, A_3, A_4, A_5, A_6, A_7$ 分别代表属性“First name”,“Last name”,“Street”,“City”,“State”,“Zip”,“DOB”。

表 5 数据集 S

Rec#	First name	Last name	Street	City	State	Zip	DOB	Status
1	Pat	Anderson	1234 Coit	Los Angeles	CA	90456	25-01-77	A
2	Roberta	Williams	5540 Melrose Pl	New York	NY	12587	12-04-78	A
3	Sarah	Johnson	7628 Walnut Hill	San Diego	CA	90006	22-05-80	I
4	Mary	Smith	8277 Royal Lane	Los Angeles	CA	90322	08-08-78	A
5	Georgia	Campbell	9999 Park Blvd.	Los Angeles	CA	90554	14-02-79	I
6	Sandra	Mathews	1414 Valley View	New York	NY	12587	23-10-78	A
7	Steve	McMullin	1111 Belt Way	Los Angeles	CA	96581	05-12-80	M
8	Susan	McLane	1550 Coit Rd	Los Angeles	CA	90554	22-02-81	I
9	Barbara	Boyce	1745 Beverly Hills	San Diego	CA	90016	01-05-77	I

表 6 数据集 S 中错误的数据项

唯一标识符	元组状态	错误描述
3	不准确	City 应该是“Los Angeles” DOB 应该是“03-10-77” Street 应该是“628 Walnut Hill”
5	不准确	DOB 应该是“09-09-80” Last name 应该是“Smith”
7	误属	不应属于关系 S
8	不准确	City 应该是“San Diego” Street 应该是“1600 Coit Rd”
9	不准确	DOB 应该是“13-11-79” Street 应该是“1745 Walnut Hill”

表 8 及时性矩阵

	A_1	A_2	A_3	A_4	A_5	A_6	A_7
1	1	1	0.8	1	1	1	1
2	1	1	0	1	1	1	1
3	1	1	0	1	1	1	1
4	1	0.6	0.8	1	1	1	1
5	0.7	0.8	0.7	0.8	1	1	1
6	1	1	1	1	1	1	1
7	0.8	1	0	0.7	1	1	1
8	1	0.5	0	0.3	1	1	1
9	0.9	0.8	0	1	1	1	1

表 7 “选择”条件下 R 各组成部分

R 的组成部分	唯一标识符集	元组个数	过时元组个数	及时元组个数
R_A	2,4	2	1	1
R_{I, s_1}	5	1	0	1
$R_{I, \sim s_1}$	8,9	2	2	0
$R_{M, \sim s_1}$	3	1	1	0
R_{M, s_M}	7	1	1	0
R 的数据质量	α_R	β_R		
值	0.2857	0.4286		

根据及时性矩阵及 3.3.1 节中所述过时元组中过时数据项作用在属性 A 的概率公式,可知,

(1) $P(O, A_3) = (0.2 + 1 + 1 + 0.2 + 0.3 + 0 + 1 + 1 + 1) / ((0.3 + 0.2 + 0.1) + (0.4 + 0.2 + 0.5 + 0.2) + (0.2 + 1 + 1 + 0.2 + 0.3 + 0 + 1 + 1 + 1) + (0.2 + 0.3 + 0.7)) = 0.65$;

(2) A_i 属性是查询条件的概率: $P(A_i, Q) = 1/7 = 0.14$;

(3) 过时数据项对应的最新数据满足查询条件的概率: $P(a \in \delta) = 7/9 = 0.78$;

(4) 主键属性 A 的不准确性即 A 属性是错误属性的概率 $P(I, A) = 1 - (0.2857 / (0.2857 + 0.4286))^{1/7} = 0.12$;

(5) 过时数据项对应的最新数据与存储的错误数据相同的概率: $P(a = prestore) = 1/\infty = 0$ 。

将各概率代入表 4 映射条件发生概率的公式,得到各映射结果的基,填入表 2。由 3.3 节信息质量测量公式,结合表 2 中各组成部分的基与表 7 中的及时元组的基,将其代入式 (3)一式(5)可得到数据集 R 在 t 时刻的信息质量为:

$$\alpha_R = (1+0+0)/7=0.1429$$

$$\beta_R = (1+0+0.07+0.56+0+0+0+0.02+1+0.01+0.5)/7=0.4514$$

$$\mu_R = (0+0+0.02+0+0.004+0.002+1)/7=0.1466$$

由实例注意到, $\alpha_R + \beta_R + \mu_R \neq 1$, 误差来源主要为:

(1)计算相关映射子条件发生的概率时产生的误差,如四舍五入,假设 1、假设 2、假设 3 产生的误差。

(2)计算表 2 中各组成部分的基时产生的误差,如四舍五入。

(3)本实例数据集 S 只有 9 条元组,经选择条件过滤后,元组个数为 7,如此小的数据集一定程度上扩大了本模型的误差。但由于本模型解决的是实际中存储的较大数据集的数据质量的测量,因此该误差会相应减少。

结束语 本文扩展了 Parssian^[18]信息产品质量静态传递模型,引入了时间变量,分析了随着时间的变化,产生的过时数据项对信息产品质量的影响,构建了信息产品质量测量的动态衍变模型,其可以根据原始信息产品的质量及过时数据项的分布概率计算在特定时刻该信息产品的质量特征,并结合应用实例验证了本模型的可行性和有效性。本文只研究了“选择”关系代数及过时数据项作用在非主键属性的情况,对于“选择”关系代数作用在主键属性上或过时数据项作用在主键属性上及在其他关系代数(投影、笛卡尔积、并、交等)作用下质量的动态衍变还有待进一步研究。

参考文献

- [1] Parssian A. Managerial decision support with knowledge of accuracy and completeness of the relational aggregate functions [J]. Decision Support Systems, 2006, 42(3): 1494-1502
- [2] Redman T C. Data Quality for the Information Age[M]. Artech House, Norwood, MA, 1996
- [3] Redman T C. The impact of poor data quality on the typical enterprise[J]. Communications of the ACM, 1998, 41(2): 79-82
- [4] DeLone W H, McLean E R. Information systems success: the quest for the dependent variable[J]. Information systems research, 1992, 3(1): 60-95
- [5] Davenport T H. Competing on analytics[J]. harvard business review, 2006, 84(1): 98
- [6] Barr P J, Shuttlesworth I. Reporting address changes by migrants: The accuracy and timeliness of reports via health card registers[J]. Health & place, 2012, 18(3): 595-604
- [7] Byrd J B, Vigen R, Plomondon M E, et al. Data quality of an electronic health record tool to support VA cardiac catheterization laboratory quality improvement: the VA Clinical Assessment, Reporting, and Tracking System for Cath Labs (CART) program[J]. American heart journal, 2013, 165(3): 434-440
- [8] Even A, Shankaranarayanan G. Dual assessment of data quality in customer databases [J]. Journal of Data and Information Quality (JDIQ), 2009, 1(3): 15
- [9] Boufares F, Ben Salem A. Heterogeneous data-integration and data quality: Overview of conflicts[C]// 2012 6th International

Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT). IEEE, 2012: 867-874

- [10] Joglekar N R, Anderson E G, Shankaranarayanan G. Accuracy of aggregate data in distributed project settings: Model, analysis and implications[J]. Journal of Data and Information Quality (JDIQ), 2013, 4(3): 13
- [11] Wang R Y, Lee Y W, Pipino L L, et al. Manage your information as a product[J]. Sloan Management Review, 1998, 39(4): 95-105
- [12] Ballou D P, Pazer H L. Modeling data and process quality in multi-input, multi-output information systems[J]. Management science, 1985, 31(2): 150-162
- [13] Shankaranarayanan G, Wang R Y. IPMAP: Current state and perspectives[C]// Proceedings of the 12th International Conference on Information Quality. 2007: 500-517
- [14] Shankaranarayanan G, Wang R Y, Ziad M. IP-MAP: Representing the Manufacture of an Information Product[C]// IQ. 2000. 2000: 1-16
- [15] Madnick S E, Wang R Y, Lee Y W, et al. Overview and framework for data and information quality research[J]. Journal of Data and Information Quality (JDIQ), 2009, 1(1): 2
- [16] Kon H B, Madnick S E, Siegel M D. Good answers from bad data: a data management strategy[M]. Sloan School of Management, Massachusetts Institute of Technology, 1995
- [17] Reddy M P, Wang R Y. Estimating data accuracy in a federated database environment [M] // Information Systems and Data Management. Springer Berlin Heidelberg, 1995: 115-134
- [18] 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
- [19] Parssian A, Sarkar S, Jacob V S. Impact of the union and difference operations on the quality of information products[J]. Information Systems Research, 2009, 20(1): 99-120
- [20] Parssian A. Managerial decision support with knowledge of accuracy and completeness of the relational aggregate functions [J]. Decision Support Systems, 2006, 42(3): 1494-1502
- [21] Parssian A, Sarkar S, Jacob V S. Assessing Information Quality for the Composite Relational Operation Join [C] // IQ. 2002. 2002: 225-237
- [22] 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
- [23] 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
- [24] 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
- [25] 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
- [26] 陈卫东, 张维明. 属性粒度数据质量模型及其评价指标研究 [J]. 计算机科学, 2010, 37(5): 139-142
- [27] Wand Y, Wang R Y. Anchoring data quality dimensions in ontological foundations[J]. Communications of the ACM, 1996, 39(11): 86-95
- [28] Ballou D, Wang R, Pazer H, et al. Modeling information manufacturing systems to determine information product quality[J]. Management Science, 1998, 44(4): 462-484

(下转第 252 页)

$(f1)C^{\Delta}(a)$ 。

例3 本体 $O = \{M3\}$, 符号序列 $S = \{\text{Fibrosis, Pancreas}\}$, 应用规则2判断本体 O 是否是 S 的句法本地公理, 如下所示:

$$\underbrace{M3 \text{ Fibrosisi} \sqcap \exists \text{ located_In. Pancreas}}_{\substack{\in \text{Cm}^{\emptyset}(S) \text{ 规则 } (\exists R^{\emptyset}, C) \\ \in \text{Cm}^{\emptyset}(S) \text{ 规则 } (C \sqcap C^{\emptyset})}} \sqsubseteq \underbrace{\text{Genetic_Fibrosis}}_{\in \text{Cm}^{\emptyset}(S) \text{ 规则 } (A^{\emptyset})}$$

应用规则2(1)得到 $M3 = \{\perp \sqsubseteq \perp\}$, 它属于句法本地公理。

命题3 句法本地公理集包含于语义本地公理集。

证明:(1)设满足 S 的句法本地性公理集为 $Ax'_{A \rightarrow \emptyset} \leftarrow^{\emptyset}(S)$, 若 $O \sqsubseteq Ax'_{A \rightarrow \emptyset} \leftarrow^{\emptyset}(S)$, 则称本体 O 满足句法本地性。

(2)由规则2的(1)和(2)对 $Con^{\emptyset}(S)$ 和 $Con^{\Delta}(S)$ 的定义可推导出, 对于 $I_{A \rightarrow \emptyset}^{\leftarrow^{\emptyset}}(S)$ 中的每一个解释 $I = (\Delta^I, \cdot^I)$, 都有 $(C^{\emptyset})^I = \emptyset$, 且对于每一个 $C^{\emptyset} \in Con^{\emptyset}(S)$ 和 $C^{\Delta} \in Con^{\Delta}(S)$, 都有 $(C^{\emptyset})^I = \emptyset^I$ 和 $(C^{\Delta})^I = \Delta^I$ 。因此, 每一条句法本地性公理 α 都满足 $I \in I_{A \rightarrow \emptyset}^{\leftarrow^{\emptyset}}(S)$ 。

(3)因此, 若根据3.2节 SEMLOC 规则1得到的公理集合为 $Ax_{A \rightarrow \emptyset}^{\leftarrow^{\emptyset}}(S)$, 那么对于 S 就有 $Ax'_{A \rightarrow \emptyset} \leftarrow^{\emptyset}(S) \sqsubseteq Ax_{A \rightarrow \emptyset}^{\leftarrow^{\emptyset}}(S)$ 。

由命题3可知, 句法本地公理一定属于语义本地公理。但是, 语义本地公理并不一定是句法本地公理。

例4 有符号序列 $S = (C_1, C_2)$, 公理 $\beta = (C_1 \sqsubseteq C_1 \perp C_2)$ 是重言式, 所以它是语义本地公理。但它只涉及 S 的相关符号, 因此它不是句法本地公理。

从例4可看出, SYNLOC 规则不能识别每一条重言式, 但是由于在实际应用中, 不能被 SYNLOC 识别的重言式公理并不经常出现, 因此“句法本地公理集并不比语义本地公理集规模大很多”这样的假设是合理的。

根据规则2匹配每一条公理, 使用二进制编码约束函数数量, 则复杂度是多项式阶。若 $|O'|$ 和 $|S|$ 分别表示本体 O' 和符号序列 S 中出现的符号数目, 那么抽取本体 O' 的 SYNLOC 模块的复杂度是 $|O'| + |S|$ 的多项式。

结束语 本体重用对于快速构建高质量的本体具有十分重要的意义, 是目前本体领域研究的热点。与传统的使用结构化方法抽取本体模块相比, 使用描述逻辑的方法保证了抽取模块的完整性与正确性。本文首先证明本地性规则是基于逻辑的模块定义的充分条件, 即“若本体 $O' \setminus M$ 关于符号序列 $S \cup \text{Sig}(M)$ 是本地的, 那么 M 就是 O' 关于 S 的一个模块”; 然后在现有的 SHQ 本体的保守扩展和本地性研究基础上, 提出了适用于 SHQ 本体的本地性规则——SEMLOC 规则和

SYNLOC 规则, 证明了 SEMLOC 规则是重言式的充分必要条件, 及较高效率的 SYNLOC 规则具备 SEMLOC 规则的所有特征, 且其复杂度为多项式阶。这些都为在基于 SHQ 的本体中抽取本体模块提供了理论支持。

下一步工作将在本文提出的本地性规则基础上, 提出对应的本体模块抽取方法。

参考文献

- [1] Berners-Lee T. Semantic web road map[EB/OL]. <http://www.w3.org/DesignIssues/Semantic.html>
- [2] Shvaiko P, Euzenat J. Ontology matching: state of the art and future challenges[J]. Knowledge and Data Engineering, 2012, 25(1), 158-176
- [3] Stuckenschmidt H, Klein M. Structure-based partitioning of large concept hierarchies[M]// The Semantic Web-ISWC 2004. Springer Berlin Heidelberg, 2004: 289-303
- [4] Doran P, Tamma V, Iannone L. Ontology module extraction for ontology reuse: an ontology engineering perspective[C]// Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management. ACM, 2007: 61-70
- [5] Seidenberg J, Rector A. Web ontology segmentation: analysis, classification and use[C]// Proceedings of the 15th international conference on World Wide Web. ACM, 2006: 13-22
- [6] Noy N F, Musen M A. Specifying ontology views by traversal[M]// The Semantic Web-ISWC 2004. Springer Berlin Heidelberg, 2004: 713-725
- [7] Krötzsch M, Simancik F, Horrocks I. A description logic primer[J/OL]. <http://arxiv.org/pdf/1202.4089.pdf>
- [8] Grau B C, Parsia B, Sirin E, et al. Modularity and Web Ontologies[C]// KR. 2006: 198-209
- [9] Kontchakov R, Pulina L, Sattler U, et al. Minimal Module Extraction from DL-Lite Ontologies Using QBF Solvers[C]// IJ-CAI. 2009, 9: 836-841
- [10] Horrocks I, Sattler U. A Tableau Decision Procedure for [J]. Journal of Automated Reasoning, 2007, 39(3): 249-276
- [11] Horrocks I, Sattler U. A description logic with transitive and inverse roles and role hierarchies[J]. Journal of logic and computation, 1999, 9(3): 385-410
- [12] Grau B C, Horrocks I, Kazakov Y, et al. Modular reuse of ontologies: Theory and practice[J]. Journal of Artificial Intelligence Research, 2008, 31(1): 273-318
- [13] Armas Romero A, Cuenca Grau B, Horrocks I. Modular combination of reasoners for ontology classification[C]// 11th International Semantic Web conference. 2012: 1-16
- [31] Buhl H. A novel data quality metric for timeliness considering supplemental data[C]// Proceedings of the 17th European Conference on Information Systems. 2009: 2701-2713
- [32] Heinrich B, Klier M, Kaiser M. A procedure to develop metrics for currency and its application in CRM [J]. Journal of Data and Information Quality (JDIQ), 2009, 1(1): 5

(上接第248页)

- [29] Heinrich B, Kaiser M, Klier M. DQ metrics: a novel approach to quantify timeliness and its application in CRM[C]// ICIQ. 2007: 431-445
- [30] Heinrich B, Klier M. Assessing data currency — a probabilistic approach[J]. Journal of Information Science, 2011, 37(1): 86-100