

基于本体映射的数据交换映射识别技术研究

苏冬娜¹ 李天阳^{2*} 杨 锐² 张旭东²

(东北石油大学计算机与信息技术学院¹, 大庆 163318; 河北汉光重工有限责任公司², 邯郸 056028)

摘要 为了实现局部本体之间数据交换映射关系的自动识别, 重点研究基于全局本体与局部本体之间的数据交换映射技术。技术以本体映射重用角度出发, 结合信息集成中数据交换映射的研究重点分析了本体映射的详细分类与可逆性, 定义了映射传递性与识别缺失率, 研究映射关系组合结果实现映射自动识别, 技术在大庆油田数据中心大型数据交换项目中得到应用。

关键词 本体映射 数据交换 XeOML 映射识别

中图法分类号 TP311. 1; **文献标志码** A

企业在信息集成的过程中, 构建了全局本体及一定数量的局部本体, 并建立了全局本体与各个局部本体之间的映射关系, 实现信息资源的整合与数据共享。在应用中, 各局部本体之间可以通过全局本体或集成应用系统进行简单数据交换^[1], 但对于复杂、大量数据的数据交换, 实现较为困难, 需要明确局部本体之间的数据交换映射, 并基于特定工具实现数据交换。为了实现局部本体间数据交换映射的自动识别, 本文重点研究了基于本体映射识别局部本体之间数据交换映射方法, 提出了基于本体映射的动态识别技术, 该技术通过对本体映射细分类、映射的可逆性与传递性进行分析, 构建映射关系组合识别结果对照表, 实现局部本体间数据交换映射动态识别。

1 技术实现方案

本体映射与数据交换映射存在区别: 本体映射为基于相似度计算等得出的本体间模式层的对应关系^[2], 而数据交换映射为模式层对应关系下数据

层对应关系, 实际表现为数据交换过程中的数据转换行为。局部本体之间数据交换映射识别要求基于模式层上的映射关系, 识别模式层下数据交换映射关系。技术方案复用已经建立的本体映射所描述的各局部本体与全局本体之间的映射关系, 以全局本体为桥梁, 建局部本体间的映射关系组合。通过研究本体映射的逆映射与映射传递性进一步确定映射关系组合识别结果, 实现局部本体间数据交换映射动态识别。方案实现如图 1 所示。

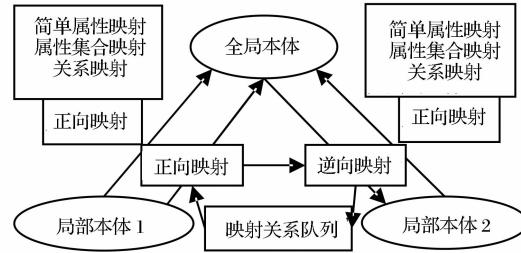


图 1 数据质量评估流程图

2 方案关键技术

2.1 本体映射分类

通过对 XeOML^[3] (An XML-based extensible Ontology Mapping Language) 本体映射分类进行了分析, 结合大庆油田信息集成数据交换映射分类情况及应用需求对本体映射进行详细分类。XeOML 中定义了四类本体元素: 实例、类、属性和关系。因本文主要研究局部本体之间模式层映射的识别技术,

2012 年 3 月 13 日收到 国家自然科学基金项目(61170132)、

黑龙江省自然基金项目(11541008)资助
第一作者简介: 苏冬娜(1980—), 女, 讲师, 研究方向: 数据库、数据挖掘、人工智能等。E-mail: summer135790059@sina.com.cn。

* 通信作者简介: 李天阳(1985—), 男, 硕士, 研究方向: 数据库应用, 数据库等。E-mail: lty66lty@163.com。

对于实例映射^[4]不予以研究。又因关系元素在本体中表现为类之间的引用形式,同时本体描述语言支持对外键属性的描述^[5],对关系元素的映射分类根据主外键引用关系进行了详细的划分。具体分类如图2所示。

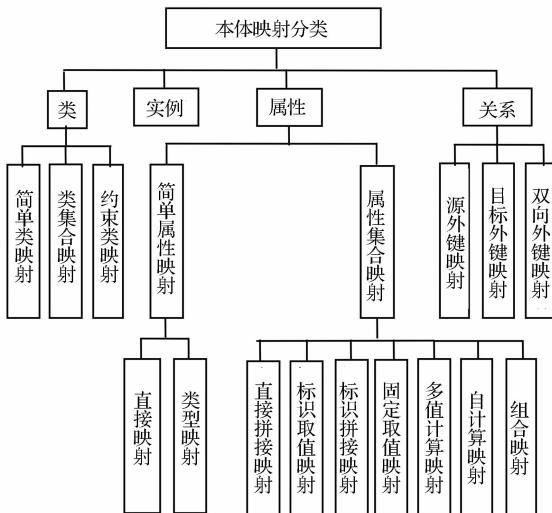


图2 本体映射分类

2.2 映射关系可逆性分析

基于本体映射分类可以进一步研究映射的可逆性,首先给出了逆映射的定义。

定义1 逆映射:源属性(类)A到目标属性(类)B的本体映射关系为R,基于本体映射R确定以B为源到以A为目标的本体映射关系为R',则R与R'互为逆映射。

推论1 可逆映射与不可逆映射:本体映射关系R分为两部分:(1)本体模式层对应关系;(2)基于模式对应的数据转换行为。其中模式对应关系存在逆向对应关系,而数据转换行为过程不完全存在,部分数据转换行为不存在逆向数据转换,导致逆映射R'的数据转换不存在,R'虽为R的逆映射,但仅为模式层的互逆,数据处理互逆中断,R'不具有数据转换,R'不是本体映射。所以,如果映射R与逆映射R'互为已知的本体映射,则映射R为可逆映射,逆映射存在;相反,则映射R不可逆,逆映射不存在。

推论2 映射关系可逆性:基于推论1,本体映射关系的可逆性指本体映射是否可逆,本体映射存

在逆映射则可逆,本体映射不存在逆映射则不可逆。

基于推论2,本文从数据转换处理可逆性角度出发,根据定义1与推论1对本体映射的可逆性进行研究与分析,分析结果如表1。

表1 逆映射对照表

序号	映射名称	可逆性	逆映射
1	简单类映射	可逆	简单类映射
2	类集合映射	可逆	简单类映射
3	约束类映射	可逆	简单类映射
4	直接属性映射	可逆	直接属性映射
5	类型属性映射	可逆	类型属性映射
6	自计算属性映射	可逆	自计算属性映射
7	多值计算属性映射	不可逆	不存在
8	固定取值属性映射	不可逆	不存在
9	标识取值属性映射	不可逆	不存在
10	标识拼接属性映射	可逆	标识取值映射
11	直接拼接属性映射	不可逆	不存在
12	组合映射	不确定	不确定
13	源外键映射	可逆	目标外键映射
14	目标外键映射	可逆	源外键映射
15	双向外键映射	可逆	双向外键映射

2.3 逆映射传递性分析

逆映射的传递性分析主要基于:源局部本体1到全局本体的本体映射,与目标局部本体2到全局本体的逆映射的本体映射关系组合进行研究。逆映射的传递性决定映射关系组合的识别结果,明确逆映射的传递性以便基于全局本体抽取局部本体之间的映射关系组合队列。下面给出映射关系组合与逆映射传递性的定义。

定义2 映射关系组合:源局部类的属性as到全局属性的正向本体映射关系Ra与目标类属性at到全局属性的本体逆映射Rc(目标属性到全局属性映射关系的逆映射)组合Rs为属性的映射关系组合,记为Rs = Ra + Rc。

源类与目标类的所有属性映射关系组合的有序队列为源类到目标类的映射关系组合队列。

定义3 逆映射传递性:映射关系组合Rs组合结果为C,若C为已存在的本体映射关系逆映射具有传递性,相反则不具有传递性。

基于定义3得出推论3如下:

推论3 当R1 * 具有传递性,则映射关系能够识别;当R1 * 不具有传递性,则映射关系不能识别。

通过推论 3 可知,当基于逆映射具有传递性,则能够确定局部数据库之间的映射关系队列及其组合结果,识别映射关系。本体映射中类映射的逆映射是具有传递性的,类映射层次较高,如果本体类的逆映射不具有传递性,则属性映射关系组合队列不存在,本体类逆映射传递性为属性映射关系组合的进一步识别提供了基础。属性本体映射的逆映射中部分不具有传递性,不能够有效识别属性映射关系队列,导致属性映射关系的识别出现缺失,为了研究映射关系识别缺失出现比率,以及对映射关系队列识别的影响,给出映射关系识别缺失率概念。

概念 1 映射关系识别缺失率:映射关系组合队列中组合识别失败的比率,用 M 表示映射关系组合总数, N 表示识别失败的组合数量,则映射关系识别缺失率 = $(M/N) \times 100\%$,下文中简称为缺失率。

通过对缺失率研究表明,当两个局部本体之间的数据模型差异较大,缺失率较高,约 20% 左右;当数据模型基于同一标准构建差异不大情况下,缺失率只有 7% ~ 12% 左右。在两个局部本体之间映射关系队列识别之前可以基于数据模型之间的对应关系进行缺失率估算。缺失率是检验映射关系动态识别的可行性以及异构数据模式之间是否存在对应关系的判断标准,当缺失率 > 20%,可能异构数据模型之间不存在对应关系,即使存在对应关系,考虑映射关系正确性问题,需要手工建立映射关系。当缺失率 < 20% 时,缺失部分映射关系需要手工构建。

2.4 映射关系组合结果

映射关系动态识别需要对基于全局本体的局部本体之映射关系组合结果进行分析研究,确定组合的结果为何种映射关系。本体映射关系组合分为类映射组合与属性映射组合,根据映射关系的可逆性与传递性对映射关系组合的映射关系结果进行分析,给出映射关系组合结果对照表。

表 2 类映射与类逆映射之间组合结果表

2.4.1 类映射组合结果

映射关系

设简单类映射用 s 表示,类集合映射用 rp 表

	正向映射	逆映射
s	s	
rp	rp	
c	c	

示,约束映射用 c 表示。正向类映射有三种,而逆向类映射只有一种,组合结果如表 2:

如表 2 所示,在正向映射与逆映射交叉的单元格为类映射组合结果映射关系, $rp + rp$ 、 $rp + rj$ 、 $rj + rp$ 与 $rj + rj$ 在发现处理中需要参考全局结构,可能会出现不存在对应属性的问题,如果属性对应关系不存在则按照映射缺失处理。

2.4.2 属性映射与关系映射组合结果映射关系

设直接映射用 d 表示,类型映射用 t 表示,标识拼接映射用 mj 表示,标识取值映射用 mp 表示,自计算映射用 sc 表示,多值计算映射用 ml 表示,固定取值映射 lp 表示,直接拼接映射用 dc 表示,组合映射 cc 表示,不存在的逆映射用 o 表示,缺失映射用 n 表示,属性映射与属性逆映射之间组合结果如表 3:

表 3 属性映射与属性逆映射之间组合结果表

正向映射	逆映射						
	d	t	mp	mj	sc	cc	o
d	d	t	mp	mj	sc	cc	n
t	t	d/t	cc	cc	cc	cc	n
mj	mj	cc	d/n	$mj + mj$	n	cc	n
mp	mp	cc	$mp + mp$	d/n	n	cc	n
sc	sc	cc	n	n	d/sc	cc	n
lp	lp	cc	cc/n	cc/n	n	cc	n
dc	dc	cc	cc/n	cc/n	n	cc	n
ml	ml	cc	n	n	cc	cc	n
cc	cc	cc	cc	cc	cc	d/cc	n

如表 3 所示,正向映射与逆向映射交叉单元格中为属性映射组合结果映射关系,其结果映射为两种情况的(如: cc 或 n),具体为情况需要根据实际映射关系组合确定。

设源外键映射用 sp 表示,目标外键映射用 tp 表示,双向外键映射用 dp 表示,则属性映射与关系逆映射组合结果如表 4。

表 4 属性映射与关系逆映射组合结果表

正向映射	逆映射		
	sp	tp	dp
d	sp	tp	tp
t	cc	cc	cc
mj	cc	cc	cc
mp	n	n	n
sc	cc	cc	cc
lp	cc	cc	cc
dc	cc	cc	cc
ml	cc	cc	cc
cc	cc	cc	cc

关系映射与关系逆映射之间组合结果如表 5:

表 5 关系映射与关系逆映射之间组合结果表

正向映射	逆映射		
	sp	tp	dp
sp	sp + sp	d 或 n	n
tp	d 或 n	tp + tp	n
dp	n	n	d 或 n

关系映射与属性逆映射之间组合结果如表 6:

表 6 关系映射与属性逆映射之间组合结果表

正向 映射	逆映射						
	d	t	mp	mj	sc	cc	o
sp	sp	cc	cc	cc	cc	cc	n
tp	tp	cc	cc	cc	cc	cc	n
dp	mj	n	n	n	n	n	n

3 技术应用

该技术在大庆油田数据交换项目中得到了实际的应用,基于大庆油田建立的本体映射对勘探开发库与 EPDM 库之间的映射关系进行了发现,共发现 95 张表映射,3556 个字段的数据交换映射,其中包含由于直接拼接映射不可逆产生的 10 个映射缺

失,取值映射不可逆 170 个缺失,多值计算不可逆 120 个,组合不可逆 45 缺失,缺失率为 9.7%。分析所缺失的映射的本体属性,缺失映射属性之间不存在映射关系,实际应用证明了技术可行性与不可行性。

4 结论

本文通过对本体映射分类及可逆性、逆映射传递性等分析给出了基于本体映射的数据交换映射识别技术,并给出实际应用的结果,结果证明了技术的可行性。

参 考 文 献

- 1 邓志鸿,唐世渭,张 铭,等. Ontology 研究综述. 北京大学学报(自然科学版),2002;38(5):730—737
- 2 杨 彩,贾松浩,张海玉,等. 基于本体的信息集成的研究与应用. 计算机应用与软件,2008;25(11):58—60
- 3 于春生. 基于 XML 的领域数据集成的研究与实现. 大庆:大庆石油学院,2009
- 4 张红宇. 数据集成中本体映射的研究. 长沙:中南大学,2005
- 5 赵荣娟,王 丹. 一种从关系数据库提取本体的方法. 微电子学与计算机,2006;23(增刊):116—118

Research on the Identification Technique of Data Exchange Mapping Based on Ontology Mapping

SU Dong-na¹, LI Tian-ying^{2*}, YANG Rui², ZHANG Xu-dong²

(College of Computer and Information Technology, Northeast Petroleum University¹, Daqing 163318, P. R. China;
Hebei Hanguang Heavy Industry Co., Ltd², Handan 056028, P. R. China)

[Abstract] In order to achieve the automatic identification of data exchange mapping between the local ontology, the research focuses on the automatic identification technique of data exchange mapping based on the global ontology and local ontologies mapping. The technique reuse the ontology mapping, combined with the study of information integration data exchange mapping analyzes the detailed breakdown of ontology mapping and reversible, defined ontology mapping transitivity and the identify miss rate, study the mapping combined results achieve automatic identification. The technical are applied well in large-scale data exchange project of Daqing Oil field data centers.

[Key words] ontology mapping data exchange XeOML mapping identification