第 35 卷 第 1 期 计 算 机 工 程 2009 年 1 月
Vol.35 No.1 Computer Engineering January 2009
·软件技术与数据库· 文章编号：1000—3428(2009)01—0076—03 文献标识码：A 中图分类号：TP311

本体复合映射的发现技术

徐德智，陈爱明
(中南大学信息科学与工程学院，长沙 410083)

摘 要：为了提高本体映射的精度，往往需要考虑 1:1, 1:*n*, *n*:*m* 等多种不同情况。以往的算法主要考虑 1:1 映射，忽略了同样重要的复合映射问题，损失了映射的精度。针对该问题，提出一个基于多种关系的复合映射发现算法。实验证明，该算法在对复合映射的发现问题上非常有效，提高了映射结果的精度。

关键词：本体；本体映射；复合映射

Ontology Complex Mapping Discovery Technology

XU De-zhi, CHEN Ai-ming
(School of Information Science and Engineering, Central South University, Changsha 410083)

【Abstract】 In order to improve the accuracy of ontology mapping, multiple different mapping conditions are needed to consider. For example 1:1, 1:*n*, *n*:*m*. As for existed mapping methods, the simple one to one mapping is focused on, while the problem of the complex mapping which also takes important seats has been neglected. Aiming at solving the loss of mapping precision, this article proposes a new complex mapping method based on multiple relationships. Experimental results show good efficiency of the method.

【Key words】 ontology; ontology mapping; complex mapping

随着本体研究的不断发展，本体数目日益增长，本体的重用与共享成为了语义网进一步发展的前提。解决重用与共享的关键是本体的映射技术。目前，国内外学者对本体映射的研究呈白热化趋势，各种算法和系统层出不穷，为此国际上已有专门的学术组织 OAEI 对这些系统进行比较评估，并且已有一些映射系统日趋完善。但是值得关注的是大部分的系统都只考虑了简单的 1:1 映射，并没有涉及 1:*n* 和 *n*:*m* 的复合映射问题。而实际上本体间的复合映射比重不小，因此，进一步提高映射精度的关键在于对复合映射的处理问题。本文提出了一个基于概念间不同关系的复合映射发现方法。

1 相关工作

目前国内外学者对本体复合映射的研究较少。仅有的研究也主要是针对模式的复合映射，主要方法有：使用参考本体[1]的方法，半自动发现数据库模式的imap[2]方法，Rimom[3]，DCM[4]等。但是它们各自都存在其问题，查全率和精度还有待提高。

2 复合映射发现技术

复合映射的概念之间的关系主要为包含关系、平级等价关系或不规则的特殊关系。本文将从这 3 个方面给出不同的发现算法。

2.1 基于包含关系的发现算法

大部分复合映射的概念之间是属于包含关系的，如 *name=firstname+lastname*，实际上后面两者都是属于前者的。

当$C_i \in O_1$，$C_j, C_k, C_m \in O_2$时，存在包含关系的映射$O_1.C_i = f(O_2.C_j, O_2.C_k, O_2.C_m)$满足以下条件：

$$(O_1.C_i \supset O_2.C_j) \cap (O_1.C_i \supset O_2.C_k) \cap (O_1.C_i \supset O_2.C_m)$$

$$(O_2.C_j \cap O_2.C_k) \cup (O_2.C_j \cap O_2.C_m) \cup (O_2.C_k \cap O_2.C_m) = \phi$$

在不同本体中寻找概念之间的关系较为复杂，因为它们之间不存在明显的结构关系。因此，本文使用以下几种方法进行复合映射发现。

2.1.1 利用概念实例的发现技术

实例是概念在现实世界的具体表现，从很大程度上反映出概念的特点。

规则 1 对于概念C_j, C_k, C_m，若它们所有的实例都属于概念C_i的实例集，且它们彼此之间不存在相同实例，则认为存在复合映射关系。即

$$\forall C_j, C_k, C_m (Ins(C_i) \supset Ins(C_j)) \cap (Ins(C_i) \supset Ins(C_k)) \cap (Ins(C_i) \supset Ins(C_m))$$

$$(Ins(C_j) \cap Ins(C_k)) \cup (Ins(C_j) \cap Ins(C_m)) \cup (Ins(C_k) \cap Ins(C_m)) = \phi$$

实例发现技术[5]对同领域的异构本体尤为有效，但对于 2 个本体的实例集没有交集时就无能为力。

2.1.2 利用属性的发现技术

复合映射的概念所具有的属性必存在一定的关联。利用属性的发现方法包括以下 2 种：

(1)属性限制

复合映射 *person=man+woman*。*person* 具有属性 sex，对 *person* 而言 sex 的取值可为 male 或 female 中任意一个，而对于后两者它们在 sex 上的取值固定，且取值的合取结果与前者在 sex 上的值域重合。

规则 2 对于概念C_j, C_k, C_m，若它们在属性P上的取值固定不变、互不相交，且取值的合取结果与C_i在P上的值域重合，则认为它们之间存在复合映射关系。即

基金项目：国家自然科学基金资助重点项目(60433020)；湖南省自然科学基金资助项目(06JJ50142)

作者简介：徐德智(1963－)，男，教授、博士，主研方向：Web 计算，语义 Web；陈爱明，硕士研究生

收稿日期：2008-05-18 **E-mail：**hunan.xu@mail.csu.edu.cn

$$R(p(O_1.C_i)) = Valueof(restriction(p(O_2.C_j))) \cup Valueof(restriction(p(O_2.C_k))) \cup Valueof(restriction(p(O_2.C_m)))$$

(2)属性相似

复合映射各概念的属性之间存在相似之处，如复合映射 *address=city+state+street* 可发现各概念都为 located at 属性的值域，如图 1 所示。

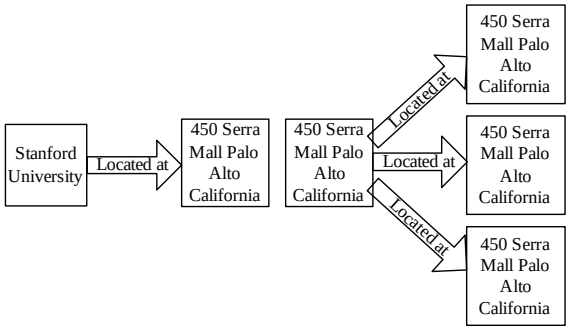

图 1　属性相似

规则 3 概念C_j, C_k, C_m与C_i都为属性P的值域(定义域)且P在本体O_1, O_2中的定义域(值域)类型相同，则认为它们之间存在复合映射关系。

$$R(O_1.P) \supseteq O_1.C_i, R(O_2.P) \supseteq (O_2.C_j \cup O_2.C_k \cup O_2.C_m),$$
$$Typeof(D(O_1.P)) \equiv Typeof(D(O_2.P))$$

2.1.3　利用注释的发现技术

复合映射各概念之间存在相似之处，属于相同领域、描述相似内容，且在注释上也具有相似性。

规则 4 概念C_j, C_k, C_m与C_i的注释相似，在语义上为包含关系，且C_j, C_k, C_m之间也存在互补关系，则认为它们之间存在复合映射关系。即

$$(similarity(comments(C_i), comments(C_j)) \geq \sigma) \cap (similarity(comments(C_i), comments(C_k)) \geq \sigma) \cap (similarity(comments(C_i), comments(C_m)) \geq \sigma)$$

且 $sem(focus(C_i)) \supset (sem(focus(C_j)) \cup sem(focus(C_k)) \cup sem(focus(C_m)))$

2.1.4　使用参考本体的发现技术

对于存在权威的全局参考本体的复合映射，可以将概念分别映射到参考本体中，然后在参考本体中查找它们之间的关系。

规则 5 若概念C_i与C_j, C_k, C_m可映射到同一参考本体，且在该本体中C_i'分别与C_j', C_k', C_m'在结构上为父子关系，语义上为包含关系，则认为它们之间存在复合映射关系。即

$$R(Stru(C_i', C_j')) = R(Stru(C_i', C_k')) = R(Stru(C_i', C_m')) \subseteq R(Stru(father - son))$$
$$R(Sem(C_i', C_j')) = R(Sem(C_i', C_k')) = R(Sem(C_i', C_m')) \subset \{isa, has\ part\}$$

其中，C_i', C_j', C_k', C_m'分别为C_i, C_j, C_k, C_m在参考本体中对应的概念。

2.1.5　基于包含关系的映射发现算法

算法 1 基于包含关系的算法

输入　待映射本体O_1, O_2

输出　复合映射关系

Step1　确定概念C_i与C_j, C_k, C_m之间是否存在包含关系，不存在则终止。

Step2　将概念映射到已有参考本体并符合规则 5 则输出。

Step3　分别使用属性、实例与注释策略将查找出的映射分别放入候选队列 1~队列 3 中。

Step4　对候选队列 1~队列 3 分别采用不同的技术进行筛选，修正映射关系，删除错误映射并输出结果。

2.2　基于平级等价关系的发现算法

虽然大部分复合映射属于包含关系，但是如 *weight-kg*= 2.2×*weight-pounds*，*kg* 与 *pounds* 之间并不存在包含关系，在相同本体中它们属于平级的兄弟关系，都为计重单位的子概念，且相互之间存在着一定的等价关系，本文称之为平级等价关系。此关系常见于重量、长度、金钱等单位代换领域。

规则 6 对于概念C_i和C_j来说，若它们属于单位代换领域且超类相似，而其实例在数值上又存在一一对应的固定比例关系，则认为它们之间存在复合映射关系。即

$$Domain(O_1.C_i, O_2.C_j) \subset Unit$$
$$superclassof(O_1.C_i) = superclassof(O_2.C_j)$$
$$\frac{Valueof(Ins(O_1.C_i))}{Valueof(Ins(O_2.C_j))} = \omega$$

2.3　基于不规则特殊关系的发现算法

诸如 *list-price=price*×(1+*fee-rate*), *sports-car=car+constrains*(*speed*>250 km/s)等不规则的复合映射关系，选择重用过去已有的复合映射匹配或者在指定的形式模板中搜索。

前者是在该领域已有的数据库中搜索匹配，后者可设计模板为C_i=C_j+*constraints*(x)进行挖掘。

3　实验结果及分析

3.1　实验

由于OAEI的测试集中，还没有专门针对复合映射的测试数据，为了评估本文提出的方法的有效性，寻找合适的实验本体进行测试是关键。鉴于复合映射的领域性与特殊性，选择了university数据集和company数据集作为实验本体。前者包含的 2 个实验本体分别是SWRC与LUBM，描述的是大学本体，分别简写为$U1$ 和$U2$。后者包含的实验本体描述的是某公司信息，分别简写为C_1与C_2。实验数据集的基本信息如表 1 所示。

表 1　测试集的统计数据

数据集	本体	概念	属性	实例	注释
university	U_1	56	74	471	/
	U_2	43	32	1 623	/
company	C_1	240	169	121	39
	C_2	153	74	218	56

本文采用查全率和查准率作为评价标准对实验结果进行评估。笔者手工建立了测试数据集的复合映射关系，用它们作为测评标准。

3.2　实验设计与分析

根据复合映射概念之间的不同关系，分 3 种情况分别进行了实验测试，并对实验结果进行了分析。

(1)对本文提出的基于包含关系的方法的实验验证，实验结果如表 2 所示。

表 2　company 数据集上的实验结果　(%)

策略	查全率	查准率
单属性	30.4	34.5
单实例	32.5	37.8
单注释	57.4	61.5
综合	80.5	72.9
算法 1	94.5	91.2

1)使用单属性策略进行挖掘。

属性是本体本身所具有的特征，是本体必不可少的部分，但是实验表明将它单一作为映射挖掘方法，结果并不理想，所得的查全率和查准率都相对较低。但是在其映射结果中加入实例及注释信息对其进行筛选即可将错误映射降低40%~50%。

2)使用单实例策略进行挖掘。

实例是本体概念在现实世界的具体表现，实例的多少及2 本体实例差距的大小对基于实例的发现技术的影响相当大。实例技术有效的前提是 2 本体的实例来源于同一个实例库。实验证明，实例技术作为辅助发现技术相当有效，而作为单一的决定因素则可能会导致错误映射的产生。

3)使用单注释策略进行挖掘。

注释是本体的辅助信息，在复合映射发现中起着举足轻重的作用。其他方法无法发现的映射关系一旦加入适当的注释信息便变得显而易见了。在 university 数据集中人工加入注释结果发现所有未发现的映射关系都可以被找到。

4)多种技术方法同时考虑。

同时使用不同的策略对不同类型的复合映射进行挖掘，实验证明结果还比较理想。但是由于对一个复合映射只是使用一种策略进行挖掘，容易出现错误的映射关系，因此使得实验的查准率不高。

5)使用算法 1 进行挖掘。

算法 1 在综合考虑使用多种策略的前提下，对候选映射综合使用其他技术进行筛选和修正，删除错误映射并修正有误映射。实验证明这样做大大减少了错误映射的概率，提高了映射的查全、查准率。

对测试数据集使用算法 1 所得的实验结果如图 2 所示。

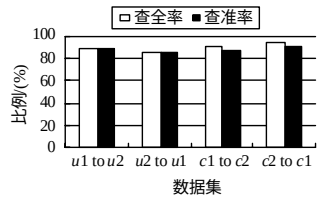

图 2　查全率、查准率结果

(2)对于平级等价关系而言，主要出现在单位代换领域，在普通本体中比重不大。可将经常出现的单位存储在固定的数据库中，实验证明这样可以极大提高搜索发现的效率。

(3)不规则特殊关系比较罕见，它的挖掘也较困难，往往受限于已有的匹配模版，在此不多作分析。

3.3　实验结果分析

实验结果表明本文提出的方法在对复合映射的发现方面是十分有效的，查全、查准率也比较理想。实验中出现的错误匹配主要是由于缺乏有效信息而导致遗漏了构成复合概念的个别概念。

4　结束语

本体映射是促进本体共享与重用的关键。针对迄今为止大部分研究停留在 1:1 映射上的问题，本文提出了一种全新的复合映射发现方法，并通过实验证明了其有效性与精确性。

未来主要从以下方面进行改进：(1)提高映射精度的同时提高映射的效率；(2)综合考虑 1:*n*, *n*:*m* 映射；(3)提高映射发现的自动化程度。

参考文献

[1] Dragut E, Lawrence R. Composing Mappings Between Schemas Using a Reference Ontology[C]//Proc. of International Conference on Cooperative Information Systems. Agia Napa, Cyprus: Springer, 2004: 783-800.

[2] Dhamankar R, Lee Y, Doan A H, et al. iMAP: Discovering Complex Semantic Matches Between Database Schemas[C]//Proc. of the 23th International Conference on Management of Data. Paris, France: [s. n.], 2004: 383-394.

[3] Tang Jie, Liang Bangyong, Li Juanzi, et al. Risk Minimization Based Ontology Mapping[C]//Proc. of the Advanced Workshop on Content Computing. Zhenjiang, China: [s. n.], 2004: 469-480.

[4] He Bin, Chang K C C. Automatic Complex Schema Matching Across Web Query Interfaces: A Correlation Mining Approach[J]. ACM Transactions on Database Systems, 2006, 31(1): 1-45.

[5] 王家琴, 李仁发, 李仲生, 等. 一种基于本体的概念语义相似度方法的研究[J]. 计算机工程, 2007, 33(11): 201-203.

（上接第 75 页）

情况 $n \geq 1$，因此，期望中这部分项的值大于等于原有系统。n 取其他值时的情况与此相同，总体期望备份系统大于原有系统。

5　结束语

仿真和分析的结果表明，本文提出的通过备份模式映射绕过离开节点的方法，在解决由对等系统节点自治性引发的映射路径断裂问题上，查询可达节点数增加，能够取得比原有系统更好的效果。下一步的研究可以放在优化模式映射网络的拓扑结构，增加结构对于节点离开的健壮性，最小化节点离开的影响。在离开发生前就优化结构，效果预期要好于离开发生时才启动绕过通信机制。

参考文献

[1] Serafini L, Giunchiglia F, Mylopoulos J, et al. The Local Relational Model: Model and Proof Theory[R]. Information Telecomunication, University of Trento, Tech Rep: DIT-02-009, 2001.

[2] Giunchiglia F, Zaihrayeu I. Making Peer Database Interact ——A Vision for an Architecture Supporting Data Coordination[C]//Proc. of the 6th International Workshop on Cooperative Information Agents VI. Loudon, UK: Spring-Verlay, 2002: 23-25.

[3] Tatarinov I, Halevy A. Efficient Query Reformulation in Peer Data Management Systems[C]//Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data. New York, USA: ACM Press, 2004: 539-550.

[4] Zhao Z. Dynamic Coordination Rules in Peer-to-peer Database[C]// Proceedings of SPIE'06. Bellingham, Washington, USA: [s. n.], 2006.

[5] Halevy A Y, Ives Z G, Mork P, et al. Piazza: Data Management Infrastructure for Semantic Web Applications[C]//Proceedings of the 12th International Conference on World Wide Web. New York, USA: ACM Press, 2003: 556-567.