计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
45(2):300~309, 2008

关系数据库模式和本体间映射的研究综述

瞿裕忠　胡　伟　郑东栋　仲新宇
(东南大学计算机科学与工程学院　南京　210096)
(yzqu@seu.edu.cn)

Mapping Between Relational Database Schemas and Ontologies: The State of the Art

Qu Yuzhong, Hu Wei, Zheng Dongdong, and Zhong Xinyu
(*School of Computer Science and Engineering, Southeast University, Nanjing* 210096)

Abstract Ontologies proliferate with the growth of the semantic Web. To date, however, most of the data on the Web are still stored in relational databases. Therefore, it is important to establish interoperability between relational databases and ontologies for creating a Web of data. An effective way to achieve such interoperability is to discover mappings between relational database schemas and ontologies. In this paper, some definitions in terms of mapping between a relational database schema and an ontology are firstly presented, and two major difficulties are analyzed from the standpoints of model construction and concrete application context. Then, a large number of popular existing solutions are surveyed according to three different facets, i.e., the approaches of model transformation (e.g., transforming ontologies to relational database schemas, transforming relational database schemas to ontologies, or transforming both relational database schemas and ontologies to medium models), the scopes of mapping strategies (e.g., automatization, and number of relational database schemas and ontologies), and the expressions of mapping results (e.g., simple correspondences, or semantic mappings). In addition, six representative mapping tools are introduced and compared with each other based on the three different facets mentioned above, and their unique features are highlighted in details. Finally, some remaining challenges are discussed, and some future research directions are also pointed out.

Key words ontology; relational database schema; ontology mapping; schema matching; semantic Web

摘　要　关系数据库模式和本体间映射是语义网研究中的一个重要问题. 首先,给出关系数据库模式和本体间映射的形式化定义,并从建模思想和应用场景两个方面分析问题的难点. 根据3个不同角度,即模型转换的途径、映射策略的适用范围以及映射结果的表达形式,调研当前存在的多种解决途径. 在此基础上,进一步介绍并比较6个具有代表性的关系数据库模式和本体间映射的工具. 最后,讨论存在的挑战,并指出未来可能的研究方向.

关键词　本体;关系数据库模式;本体映射;模式匹配;语义网

中图法分类号　TP182

在过去的10年里,计算机科学领域中关于语义网(semantic Web)[1]的研究越来越多. 语义网的目标是提供一个通用的语义框架,实现数据在不同应用之间的共享和集成. 本体(ontology)作为语义网

收稿日期:2007-02-05;修回日期:2007-10-24
基金项目:国家自然科学基金项目(60573083);国家"九七三"重点基础研究发展规划基金项目(2003CB317004)

的基础,它是领域知识概念化和模型化的一种途径,可以被用来描述数据的语义信息.目前标准化的本体语言,例如 RDFS[2] 和 OWL[3] 已由万维网联盟 W3C 发布.

尽管语义网在诸如信息检索等许多应用领域取得了阶段性成功,但是它距离真正的实际应用仍有一个很长的过程.其中最为主要的一个原因是目前万维网上绝大多数数据仍然以关系数据库(RDB)的方式存储(约占 77.3%)[4],致使语义网应用程序不能自由地访问和操纵这些数据,从而限制了语义网的发展.万维网的创始人和语义网的倡导者 Berners-Lee 等人在其最新的研究报告中指出:"世界上的数据仍然存储在数据库中,且尚未在万维网上以资源的形式公开发布"[5].因此,如何在语义网环境下使用现有的存储在关系数据库中的数据是语义网相关研究中的一个重要问题.

关系数据库和本体间的数据互操作问题实际上可以归结为关系数据库模式和本体间的映射问题.在传统的关系数据库中,关系表的结构及其完整性约束都是由关系数据库模式(RDB schema)定义的,并且关系数据库模式和本体间存在着许多近似的对应关系,例如关系数据库模式中的表(table)可以对应到本体中的类(class).因此,实现关系数据库和本体之间数据的互操作性可以通过构建关系数据库模式和本体间映射的途径来解决.

通常,不同的关系数据库有着不同的数据模式.而万维网及语义网的分布性使得不同领域甚至相同领域的不同组织也可能定义不同的本体来描述数据.这就导致了在现实世界中不可避免地存在众多异构的(heterogeneous)关系数据库模式和本体.然而,试图通过人工的方法寻找关系数据库模式和本体间的映射是不现实的.

目前,国内外涉及关系数据库模式和本体间映射问题的研究有很多.虽然具体方法有所不同,但通常遵循这样一个过程:它假设分别独立存在一个关系数据库模式和一个本体作为输入,通过应用多种不同的映射策略以及人工辅助参与,构建这对关系数据库模式和本体中对应元素之间的映射.寻找映射的过程又可以进一步细分为 3 个阶段:首先通过模型转换消除关系数据库模式和本体在模型上的异构性;接着根据应用场景选用映射策略,寻找映射结果;最后生成映射结果并以某种形式表达.一个基本的映射框架如图 1 所示:

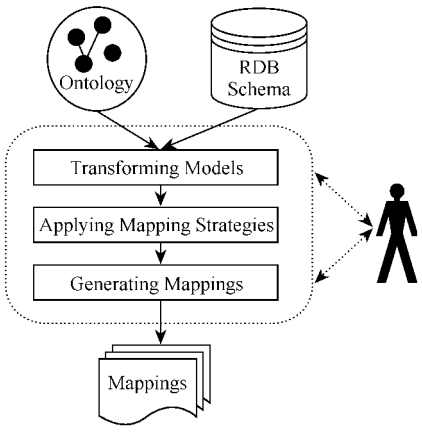

Fig. 1 A basic framework of mapping between RDB schemas and ontologies.

图 1 关系数据库模式和本体间映射的基本框架

1 问题描述

本节形式化地给出关系数据库模式和本体间映射问题的相关定义,并分析该问题的难点.

1.1 关系数据库模式、本体以及两者间的映射

数据模型(data model)是用来描述数据的一组概念和定义.对某一类数据的结构、联系和约束的描述称为数据模式(data schema).依据文献[6],简单地给出关系数据库模式的一个定义.

定义 1.关系数据库模式.一个关系数据库模式(S)由一组关系模式组成,其中包含数据库的基表结构(E)和完整性约束(CT)两个部分.基表结构定义关系(表)的结构、属性(列)及其数据类型与长度等;完整性约束定义施加在数据上的语义约束.

本体是对某一概念模型的显式的规范说明[7].关于本体的不同定义有多种[8],参考文献[9]给出本体的一种形式化定义.

定义 2.本体.一个本体(O)可以被表示为一个二元组 $O=(ID, Axiom)$.其中,ID 是本体的词汇集合,且满足 $ID=C\cup R\cup I$.这里,C 表示概念(concept)集合,R 表示关系(relation)集合,I 表示实例(instance)集合;$Axiom$ 是本体的公理(axiom)集合.

关系数据库模式和本体间映射的研究较多.由于研究目的不同,所以也存在多种不同的定义.本文对于关系数据库模式和本体间映射的定义如下.

定义 3.关系数据库模式和本体之间的映射.给定一个关系数据库模式 S 和一个本体 O,S 和 O 之间的映射 map 是由五元组作为元素构成的一个

集合 $\{m\}$. 其中,m 表示一个基本的映射单元,可以写成$\langle u,e,v,rel,f\rangle$的形式. u 为单元标识符,用于惟一标识该五元组;e 和 v 分别为 S 和 O 中的元素,且满足 $map(e)=v$;rel 描述 e 和 v 之间的关系,例如,等价关系"$=$"、包含关系"$\subseteq$"、相交关系"$\cap$"、不相交关系"$\perp$"等;f 标识映射的确信度(或相似度).

这里进一步解释关于关系数据库模式和本体间映射问题的研究范畴.

首先,该研究问题针对的是关系数据库模式到本体,或者本体到关系数据库模式的映射问题,而不是关系数据库模式之间(包括利用本体辅助发现关系数据库模式之间的对应)或者本体之间的映射问题. 关系数据库模式之间的匹配问题(schema matching)请参阅文献[10];而本体间的映射问题(ontology mapping)请参阅文献[9,11].

其次,该研究问题通常假设给定的关系数据库模式和本体是独立存在的. 所以,研究如何把关系数据库模式翻译为本体的表达形式[12-13]并不是本文的关注点. 另外,从存储在数据库的数据中学习出本体的研究[14]也不是本文的关注点.

1.2 关系数据库模式和本体间映射的难点

本质上,关系数据库模式和本体之间的映射问题属于异构数据源的集成问题,但是它又和一般的异构数据源集成问题有所区别. 一般异构数据源的集成问题是基于同种元模型的,例如本体间的映射问题都是以本体为基础. 而关系数据库模式和本体分别属于不同的模型,也就是说两者在建模思想、实现方法、应用场景等方面都有很大的区别. 在已有的一些工作中,已经对关系数据库模式和本体间映射的难点有了部分阐述[15-17]. 我们认为,该研究问题的难点主要包括两个方面.

一方面是由于建模思想不同引起的. 数据库模式是对具体数据的抽象描述;而本体则试图建立领域的共享概念. 因此,关系数据库模式通常表现为局部且规模较小的模型;而本体则表现为相对比较开放且规模较大的模型. 关系数据库模式的语义表达能力较弱,结构也较为简单;而本体的表达能力较强(例如最常见的 OWL 语言的逻辑基础是描述逻辑),结构也较为复杂. 如何协调平衡两种异构模型,是构建关系数据库模式和本体间映射的一个主要难点.

另一方面是由于应用场景不同造成的. 数据库通常仅为有限的几个应用程序服务,因此经常由应用程序的开发者独自创建和管理;而本体是对某一领域中公认的概念知识的建模,所以本体模型和具体应用是分开的. 在实际使用中,数据库模式主要为数据的存储查询服务;而本体则还存在逻辑推理等方面的应用需求. 如何根据应用场景不同,合适地选择构建关系数据库模式和本体间映射的方法,这是该研究问题的另一个难点.

2 分 类

目前的研究工作已从多个方面考察关系数据库模式和本体之间的映射问题,例如描绘关系数据库模式和本体间映射的系统框架[18-21];提出具体的映射算法[16-17,22-23]以及描述映射结果的语法语义[24]. 根据图 1 所示的关系数据库模式和本体间映射的基本框架,以及第 1.2 节所述的关系数据库模式和本体间映射的难点,以下从 3 个角度对已有的解决途径进行分类和归纳(具体请参见图 2).

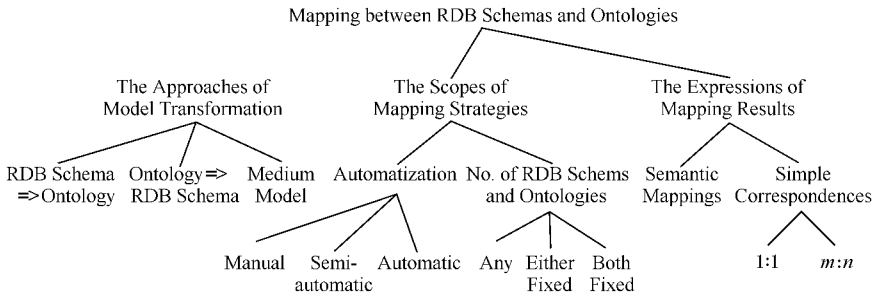

Fig. 2 A classification of the existing solutions for mapping between RDB schemas and ontologies.

图 2 对已有关系数据库模式和本体间映射解决方案的分类

首先是基于模型转换的途径分类. 由于关系数据库模式和本体建模思想不同,所以需要在映射过程中消除这种异构性,即协调两者在表达能力、规模结构等方面的差异. 具体请参见第 2.1 小节.

其次是基于映射策略的适用范围分类. 对于不同的应用场景,需要选择不同的映射方法. 例如在

开放式的万维网环境中,应当选择(半)自动化的、可以处理任意多个关系数据库模式和任意多个本体间映射且时空复杂度较低的方法工具.具体内容请参见第 2.2 小节.

最后是基于映射结果的表达形式分类.传统上,关系数据库模式间映射结果主要以较为简单的对应关系(simple correspondences)的形式表现,而本体间映射结果则试图提供更丰富的语义映射(semantic mappings)[25].从我们第 2.3 小节的分析可以看出,目前的关系数据库模式和本体间映射方法中存在多种不同的映射结果表达形式.

2.1 基于模型转换途径的分类

关系数据库模式和本体间模型上的差异不仅仅表现在它们的语法层次上,更重要地表现在语义层次上.对于语法层次上的差异可以通过一些转换规则较好地消除,例如可以把关系数据库模式中的某些表转换为本体中的类,或者通过观察表之间主外键间某种联系,把某些表之间的关系转换为 RDFS 或 OWL 本体中的"*rdfs*:*subClassOf*"关系[13].

但是在语义层次上,则很难实现两个模型之间的转换.相较于本体而言,关系数据库模式更侧重描述数据库的结构而非语义,即使是使用扩充 E-R 数据模型(extended E-R data model),也只能部分地表达数据库隐含的语义信息.一个具体表现在于关系数据库模式中元素的数目较少,而本体中元素的数目较多.例如,Chen 等人[16,19]构建的中医药本体中,数据属性(data property)和对象属性(object property)的数目大约有 800 个,而每个关系数据库模式中所包含的关系(relationship)和属性(attribute)的数目不到 100 个.另外,在模型转换过程中,对于本体中空白节点(blank node)的语义处理也较为困难.

消除关系数据库模式和本体之间模型上差异的解决途径主要有 3 类:1)把本体转换为类似关系数据库模式的形式表达;2)把关系数据库模式转换为类似本体的形式表达;3)把关系数据库模式和本体分别转换到某种中间模型(medium model).目前已有的研究主要采用第 2 类和第 3 类解决途径.

对于第 1 类模型转换途径到目前为止还尚未有方法采用.主要原因在于本体的表达能力强于关系数据库模式,所以把本体用对应的关系数据库模式的形式表达会丧失本体丰富的语义信息,从而导致在映射过程中很难从语义层次上发掘映射,并且对于映射结果也很难进行语义层次上的分析、验证和调试.

有部分工作采用了把关系数据库模式用本体的形式表达的转换途径[16,19-21].通常这类工作首先通过一些转换规则,例如采用关系数据库的逆向工程(relational database reverse engineering)的思想[26],自动或半自动地把关系数据库模式表达为本体的形式(以 RDFS 或 OWL 最为常见),然后再寻找转换本体和输入本体之间的映射.这类转换思想的优点在于可以在最大限度上重用大量已有的本体映射算法,取得较好的结果.但在实际应用中,由于关系数据库模式和本体之间不存在完美的兼容关系,并且两者在表达能力等方面差距较大,所以这种转换通常是不完备的,转换效果也较差.

现有的研究主要集中在最后一类模型转换途径上,即把关系数据库模式和本体分别转换到某种统一的中间模型,例如文献[17-18,22-23]等.另外,在 XML 模式和本体之间的映射问题中,也有相关工作采用了类似的思想[27].通常这类工作首先定义一个表达能力适中的中间模型,如有根的有向无环图(rooted directed acyclic graph)[17]和 Web-PDDL 中间模型[18]等,然后分别把关系数据库模式和本体转换到中间模型.对于关系数据库模式到中间模型的转换,可以增加某些语义信息,例如通过机器学习和数据挖掘的方法[18]获取更多更复杂的关系;而对于本体到中间模型的转换,则需要裁剪丢弃不兼容的语义信息,例如把本体图模型转换为树型的连接公式(conjunctive formulas)[22].这类转换思想的优点在于中间模型可以平衡关系数据库模式和本体之间的差异,并且灵活度相对较大.但是由于中间模型一般为映射方法本身定义,所以不能直接利用大量已有的数据库模式映射或本体映射的方法,因而重用性较差.

2.2 基于映射策略适用范围的分类

针对不同的应用场景,需要选择不同的映射方法以适应需求.例如,在生物医学领域,由于测量、采集、组织和管理数据的难度很大,所以只存在少量的大型数据库和大型本体.由于这些数据库和本体本身的规模过于巨大(例如解剖学领域中著名本体 FMA[①]中存在大约 10 万个概念),完全依靠手工的方式构建映射是不现实的,所以需要采用半自动或全自动的方法,且这些方法最好能够保证映射的准

① http://sig.biostr.washington.edu/projects/fm

确度(对运行速度可以放宽要求);另外,由于本体是领域通用的,因此也可以选取适用于多个关系数据库模式和一个本体间映射的方法.而对于开放式的万维网环境,由于存在众多可能的数据库和本体,而且存在大量可能的变更(例如有新的数据库加入),所以需要采用可以适用于任意多个关系数据库模式和任意多个本体间映射的方法;另外由于大部分关系数据库模式和本体的规模较小,有时手工的方法也是可行的.

根据上述例子,可以从两个维度分析归纳目前已有的解决方案:1)从方法的自动化程度上分类(手动、半自动、全自动);2)从方法针对的关系数据库模式和本体的数量上分类(两者皆为任意数目、两者之一数目固定、两者数目都固定).

根据第1个维度,目前已有的映射方法中,文献[16,19,20-21]采用的是手动构建映射的方法.这类研究主要针对特定的应用场景而设计,例如文献[16,19]针对于中医药领域,文献[21]针对大学数字图书馆资源.这类手动构建映射的方法有时可以发现复杂的映射,但非常费时费力.而文献[17-18,22]则采用了半自动化的方法,通过和用户的多次交互,提高映射的准确度.例如文献[22]要求用户首先输入或验证一些简单的映射,而文献[17-18]则是一个迭代交互的过程,在整个迭代过程中用户都可以参与映射结果的验证和修改.这类半自动化方法的映射质量在很大程度上受到用户交互质量的影响.目前很少有全自动的映射方法(除文献[23]以外),主要原因在于此类方法实现难度较大,并且通常情况下准确度较低,同时也很难发现复杂的隐式映射关系.

对于半自动或者全自动的方法,我们还可以进一步从方法的主要特性上分析.从发掘映射的算法策略上可以分为单一型算法和集成型算法.例如,An等人[22]提出的基于图的相似度传播的算法是单一型算法的代表;对于集成型算法又可以细分为混合型(hybrid)和组合型(composite)两种.文献[23]首先采用基于语言学的方法找到部分映射,接着把这部分映射输入到基于结构的方法,以寻找更多的映射,该方法属于混合型集成算法;而文献[17]同时使用多种算法,每种算法找到部分映射,再把这些映射组合起来,作为最终的输出,所以该方法属于组合型集成算法.一般认为,采用集成型算法的方法适用面更广、稳定性更强.另外,映射方法的时空复杂度也是一个需要考察的特征,通常自动化的方法要比半自动化的方法速度更快、基于字符串比较的映射方法要比基于相似度传播的映射方法的时间复杂度更低.

根据第2个维度,文献[18,22]主要针对任意数目的关系数据库模式和任意数目的本体间的映射问题,其目标在于提供一种通用的解决方法;文献[17]采用了全局视图(global as view)的方式,要求关系数据库模式和本体的数目都固定;而最为常见的一类方法为面向任意多个关系数据库模式和一个已知本体之间的映射[16,19-21].通常此类方法主要面向某些特殊领域的数据集成问题,在这些领域中存在被普遍认同的通用本体,且这些通用本体覆盖了该领域中绝大多数的概念知识,这时只需要考虑多个关系数据库模式到该通用本体的映射问题,一般采用本地视图(local as view)的方法,例如文献[16,19]把多个关系数据库模式映射到一个通用的中医药本体.

2.3 基于映射结果表达形式的分类

目前已有的解决方案主要包含两个层次的映射结果:关系数据库模式和本体间元素之间的简单对应关系和较复杂的包含语义信息的映射.

对于简单的对应关系,文献[20,23]只考虑简单的1:1的对应关系,而文献[17,22]则允许多对多($m:n$)的对应关系,其中文献[22]采用Horn子句的表达形式,其生成的多对多的映射结果更有助于实现查询重写.无论是1:1的对应关系,还是$m:n$的对应关系,它们共同的特点都是不指明语义关系(例如等价关系、包含关系等),因此它们通常需要用户进一步参与后才能形成最终的映射结果.

一个更高的层次是输出包含语义信息的映射,它不仅仅找到映射,还指明映射关系的语义.这类方法的代表是文献[18].Dou等人[18]采用桥接公理(bridging axioms),指明关系数据库模式和本体间元素之间的语义映射.它的映射结果可以充分利用本体的逻辑推理能力.另外值得注意的是,在本体映射领域,不少映射方法的映射结果是以语义映射的形式表达[25].

可以看到,关系数据库模式和本体之间映射的目标是构建复杂的查询重写(例如,从本体查询语言SPARQL[28]到数据库查询语言SQL[6]的查询重写),使得现存储在关系数据库中的数据可以在语义网环境中被查询和集成.而语义映射可以更好地利用本体在逻辑推理等方面的优势,所以包含语义信息的映射相对于简单的对应关系更符合语义网的特点.但是也应该看到,由于目前还没有一种统一

的映射结果表达形式,各种表达方式之间又不存在显式的兼容关系,所以很难使用一个统一的框架集成这些包含语义信息的映射.

3 关系数据库模式和本体间的映射工具

第2节已经从3个不同角度对部分已有工作进行了分类和归纳.在本节中,首先将对6个较具代表性的系统工具做简要介绍,接下来再对它们做进一步的比较和分析.

3.1 工具简介

OntoGrate[18,29]是由美国Oregon大学于2006年开发的一个关系数据库模式和本体间映射的系统.系统主要包括6个功能模块:语法转换器、映射生成模块、推理模块、学习模块、挖掘模块、用户界面模块.系统的执行过程为首先利用语法转换器分别将关系数据库模式和本体转换到用Web-PDDL语言描述的中间模型,然后辅之于人工参与,通过映射生成模块构建两个中间模型之间的映射,最后输出桥接公理.另外可以借助于推理、学习以及挖掘模块等进行更深入的处理.该系统的优点在于提供了一个较全面的映射框架并充分利用了多种类型的外部知识辅助构建语义映射.

MAPONTO[22]是由加拿大Toronto大学于2005年实现的一种基于树的相似度传播思想的映射工具.它采用树状结构作为数据库模式和本体的中间转换模型.在执行过程中,首先寻找简单的关系数据库模式的属性和本体的数据属性之间的简单对应关系,然后利用这些对应,在两个中间模型(即两棵树)上迭代地传播这种对应,最终找到关系数据库模式中多个元素(表、关系)和本体中多个元素(类、对象属性)之间的多对多的对应,以Horn子句的形式输出最终映射结果.该工具较全面地考虑了扩充E-R数据模型到树状结构的转换规则.但是,由于该工具基于迭代算法实现,因此工具的时间开销较大.

DL04①[17]是由美国Iowa大学于2004年开发的一个映射工具.它包含两个主要功能:半自动化地实现关系数据库模式和本体之间的映射;以及自动化地实现关系数据库模式之间的映射.这里主要考查前者.它首先把关系数据库模式和本体都转换为COMA[30]图格式(有根的有向无环图),然后再利用工具COMA实现映射.COMA的一个基本执行过程为首先并行地执行多个单元映射器,每个单元映射器分别返回一些对应关系,再通过组合策略(combination strategy)集成这些对应关系,整个过程可以循环多次,且在每次迭代中均允许用户干预.最终的输出为关系数据库模式中的基表结构和本体中词汇之间的简单对应(既允许1:1的对应关系,又允许多对多的对应关系).该工具最大的特点是重用已有的数据库模式映射工具COMA实现关系数据库到本体间的映射.

FDR2#Kit[20]是由荷兰Vrije大学于2004年开发的一个基于万维网访问方式的映射工具集.它构建映射的过程包括3个步骤:1)根据关系数据库建立对应的关系数据库模式;2)自动地把该关系数据库模式转换为本体的形式表达;3)用户手工构建关系数据库模式和本体间的简单对应,最终以RDMap的形式表达.值得注意的是,该工具是关系数据库模式和本体间映射早期研究的代表,许多后续工作继承了它的基本思想.

DartGrid[16,19]是由国内浙江大学于2006年提出的一套针对中医药领域的集成系统.它主要包括3个组件:1)DartMapping是一个可视化的映射工具,用来辅助用户定义关系数据库模式和本体间的映射;2)DartQuery提供一个基于本体的查询界面,帮助用户创建语义查询和实现本体SPARQL到数据库SQL的查询转换;3)DartSearch作为一个基于本体的搜索引擎,允许用户在所有数据库上执行全文搜索以及查询结果之间的语义导航.以下主要考察DartMapping组件.DartMapping采用本地视图的方法,人工地把任意多个关系数据库映射到一个特定的通用中医药本体上,最终以类似datalog语法的形式表达对应关系(允许多对多的对应).DartGrid的特色在于提供了一整套关系数据库和本体之间数据集成的解决方案,并且已经成功地应用在中医药领域;其有待改进之处在于目前只能手工地构建简单对应,尚未考虑关系数据库模式和本体间的模型转换问题.

MASON[23]是由国内东南大学于2006年开发的一种通用的关系数据库模式和本体间映射的工具.它首先利用预先定义的转换规则,把关系数据库模式和本体分别转换到中间模型(有向无环图DAG).然后采用基于字符串比较的映射算法[31]和

① 由于该工具的作者尚未给此工具起名,本文组合该工具两位作者的姓氏首字母及工具发布年代暂作为该工具的名称.

基于结构的 SF 算法[32],寻找元素之间的简单对应并输出. 该工具的优势在于综合了两种具有不同特性的映射算法,可以较好地保证映射结果的准确性;但是关系数据库模式和本体到中间模型的转换规则还有待进一步完善.

3.2 工具的比较分析

除了以上提到的这 6 个具有代表性的工具外,还有许多各具特色的系统工具,譬如 OBSERVER[33],MOMIS[34],OOML[21] 等. 由于目前还没有统一的评价标准,很难对它们进行定量的评价. 一般来说,映射工具之间的主要区别在于:1)对于输入的关系数据库模式和本体的模型转换途径;2)构建映射的具体方法;3)映射结果的表达形式.

根据以上 3 个方面,现将上述的 6 种关系数据库模式和本体之间的映射工具进行总结(见表 1 所示). 表中给出的工具名称表示该工具采用相应的模型转换途径和生成相应的映射结果表达形式,工具名称后的括号内给出的是该工具针对关系数据库模式和本体间映射问题所采用的具体方法.

表 1 表明,首先,目前采用把关系数据库模式和本体分别转换为中间模型的映射工具比较多(4 个工具),且这些工具都是半自动化或全自动化的. 不进行模型转换或者把关系数据库模式转换为本体的工具较少,且都用手工的方式构建映射简单对应. 这种情况表明,目前越来越多的方法和工具都注意到协调关系数据库模式和本体这两种异构模型的重要性.

其次,支持语义映射的映射工具还很少(只有 OntoGrate). 虽然关于构建语义映射的理论重要性已被研究人员普遍认同[25],但是在实际应用中发掘语义映射依然是十分困难的. 虽然 OntoGrate 生成的映射仅仅包含等价、包含等几种简单的语义映射,但是它为今后的相关研究提供了一条思路.

再者,大部分工具使用的映射算法都比较单一,还没有一个整合的映射系统. 说明目前的这些工具都只能在某些情况下取得较好的效果,因为任何一种映射算法都无法适用于所有情况,所以必须考虑多种映射算法的整合问题,从而保证在大部分情况下都能获得较理想的结果. 资料表明,OntoGrate 和 MASON 都在朝着这个方向努力.

Table 1 Summary of the Mapping Tools

表 1 代表性映射工具概述

The Expressions of the Mappings		The Approaches of Model Transformation		
		No Transformation	RDB schema→Ontology	Medium Models
Simple Correspondences	1:1		FDR2 # Kit {manual}	MASON {string comparison, similarity flooding}
	$m:n$	DartGrid {manual}		MAPONTO {similarity propagation in trees}
				DL04 {reuse the algorithms of COMA}
Semantic Mappings				OntoGrate {machine learning, data mining, knowledge inference}

4 总结与展望

本文形式化地定义了关系数据库模式和本体间映射问题,并分析了主要研究难点;从模型转换的途径、映射策略的适用范围以及映射结果的表达形式这 3 个角度,阐述了关系数据库模式和本体间映射问题的基本特征、常用解决方案和研究进展;介绍并比较了现有的 6 个关系数据库模式和本体间映射的工具. 从中可以看出,虽然关系数据库模式和本体间映射是一个新兴的研究课题,但是许多相关领域的研究成果可以供其借鉴,例如重用数据库模式匹配或本体映射领域的一些方法. 然而,由于关系数据库模式和本体间映射问题具有自身的特殊性,该领域仍然存在许多有待解决的问题. 总结起来有以下 3 个方面.

1) 对映射方法的改进:目前虽然已经提出了不少关系数据库模式和本体间映射的方法,但是大部分方法都不理想. 首先针对模型转换途径来说,现有的大多数方法倾向于使用中间模型协调关系数据库模式和本体间的差异,但是这些方法目前只考虑了部分不完全的转换规则,而且没有进一步从理论上分析这些转换规则的正确性和完备性;针对具体方法的适用范围来说,现有的方法往往重用已有的数据库模式映射或本体映射领域的思想和方法,还尚未真正提出创新的针对关系数据库模式和本体间映射的方法;从映射结果的表达形式上看,目前能够输出包含语义信息的映射的方法还几乎没有. 所

以,在今后的研究中有必要提出一些新的映射方法,而对于现有的方法也需要进一步地改进.

2) 对映射工具的完善:目前的关系数据库模式和本体间映射的研发工作中,半自动化或自动化的工具提供的功能都非常有限.这些工具仅仅集成了一个或很少的几个映射算法,因此在实际使用中存在局限性.未来需要的是一个完善的、整合的、能够完成多种映射任务的关系数据库模式和本体间的映射工具,因此如何将各种不同的算法进行集成从而构建更好的映射工具,是未来的一个工作方向.另外,虽然由于缺乏客观的评价标准,无法准确地对这些现有的映射工具进行定量的评价,但是通过使用可以感觉到它们无论在功能,还是在稳定性、易用性等方面都离实际使用还有一段距离.

3) 对映射结果的表示和评价:目前还没有公认的评价关系数据库模式和本体间映射的"标准"的结果表示方式和测试集.在传统的信息抽取领域或者本体映射领域,都已经开发出一些标准化的表示方式和测试集,可以用于定量的考察映射方法或工具的性能.例如,本体映射领域中,映射工具比赛OAEI(ontology alignment evaluation initiative)[35-36]提供了多组测试用例,较全面地评测各个本体映射工具的性能.对于关系数据库模式和本体之间的映射,也迫切需要建立一个"标准"的结果表示方式和测试集,使其能够统一地评价映射结果,这样有利于关系数据库模式和本体间映射的方法和工具的进一步发展.所以,如何对映射结果进行标准化的表示和定量的评价也是一个重要的研究方向.

总之,关系数据库模式和本体间映射是语义网研究中的一个重要问题.国际上在关系数据库模式和本体间映射方面的研究很活跃,并已有多个原型工具.目前国内也有一些相关研究,并且取得了部分进展.可以预见,随着国内外在语义网和本体方面研究的增多,将会有更多针对关系数据库模式和本体间映射的相关方法和工具涌现.

参 考 文 献

[1] T Berners-Lee, J Hendler, O Lassila. The semantic Web [J]. Scientific American, 2001, 284(5): 34-43

[2] D Brickley, R V Guha. RDF vocabulary description language 1.0: RDF schema [OL]. http://www.w3.org/TR/rdf-schema/, 2004

[3] P F Patel-Schneider, P Hayes, I Horrocks. OWL Web ontology language semantics and abstract syntax [OL]. http://www.w3.org/TR/owl-semantics/, 2004

[4] K C Chang, B He, C Li, Z Zhang. Structured databases on the Web: Observations and implications [J]. SIGMOD Record, 2004, 33(3): 61-70

[5] T Berners-Lee, W Hall, J Hendler, *et al*. Creating a science of the Web [J]. Science, 2006, 313: 769-771

[6] Wang Nengbin. Database System Tutorial [M]. Beijing: Publishing House of Electronics Industry, 2004 (in Chinese)
(王能斌. 数据库系统教程[M]. 北京:电子工业出版社, 2004)

[7] T Gruber. A translation approach to portable ontologies [J]. Knowledge Acquisition, 1993, 5(2): 199-220

[8] Li Shanping, Yin Qiwei, Hu Yujie, *et al*. Overview of researches on ontology [J]. Journal of Computer Research and Development, 2004, 41(7): 1041-1052 (in Chinese)
(李善平, 尹奇□, 胡玉杰, 等. 本体论研究综述[J]. 计算机研究与发展, 2004, 41(7): 1041-1052)

[9] Y Kalfoglou, M Schorlemmer. Ontology mapping: The state of the art [J]. Knowledge Engineering Review, 2003, 18(1): 1-31

[10] E Rahm, P A Bernstein. A survey of approaches to automatic schema matching [J]. VLDB Journal, 2001, 10(4): 334-350

[11] P Shvaiko, J Euzenat. A survey of schema-based matching approaches [G]. In: LNCS 3730. Berlin: Springer, 2005. 146-171

[12] L Stojanovic, N Stojanovic, R Volz. Migrating data-intensive Web sites into the semantic Web [C]. In: Proc of the 17th ACM Symp on Applied Computing. New York: ACM Press, 2002. 1100-1107

[13] Xu Zhuoming, Dong Yisheng, Lu Yang. Semantics-preserving translation from ER schema to OWL DL ontology [J]. Chinese Journal of Computers, 2006, 29(10): 1786-1796 (in Chinese)
(许卓明, 董逸生, 陆阳. 从ER模式到OWL DL本体的语义保持的翻译[J]. 计算机学报, 2006, 29(10): 1786-1796)

[14] Du Xiaoyong, Li Man, Wang Shan. A survey on ontology learning research [J]. Journal of Software, 2006, 17(9): 1837-1847 (in Chinese)
(杜小勇, 李曼, 王珊. 本体学习研究综述[J]. 软件学报, 2006, 17(9): 1837-1847)

[15] A Sheth. {Ontology: resource} * {Matching: mapping} * {Schema: instance} = components of the same challenge [C]. Int'l Workshop on Ontology Matching, Athens, 2006. http://www.dit.unitn.it/~p2p/OM-2006/OntologyMapping-ISWC06-Sheth-Keynote.ppt

[16] H Chen, Z Wu, H Wang, *et al*. RDF/RDFS-based relational database integration [C]. In: Proc of the 22nd Int'l Conf on Data Engineering. Los Alamitos, CA: IEEE Computer Society Press, 2006

[17] E Dragut, R Lawrence. Composing mappings between schemas using a reference ontology [C]. In: Proc of Int'l Conf on Ontologies, Databases and Applications of Semantics. Berlin: Springer-Verlag, 2004. 783-800

[18] D Dou, P LePendu, S Kim, *et al*. Integrating databases into the semantic Web through an ontology-based framework [C]. In: Proc of the 3rd Int'l Workshop on Semantic Web and Databases. Los Alamitos: IEEE Computer Society Press, 2006

[19] H Chen, Y Wang, H Wang, *et al*. Towards a semantic Web of relational databases: A practical semantic toolkit and an in-use case from traditional Chinese medicine [C]. In: Proc of the 5th Int'l Semantic Web Conference. Berlin: Springer-Verlag, 2006. 750-763

[20] M Korotkiy, J Top. From relational data to RDFS models [C]. In: Proc of Int'l Conf on Web Engineering. Berlin: Springer-Verlag, 2004. 430-434

[21] X Luo, X Chen, Q Zhao. OOML-based ontologies and its services for information retrieval in UDMGrid [C]. In: Proc of the 6th Int'l Workshop on Advanced Parallel Processing Technologies. Berlin: Springer-Verlag, 2005. 342-352

[22] Y An, A Borgida, J Mylopoulos. Inferring complex semantic mappings between relational tables and ontologies from simple correspndences [C]. In: Proc of Int'l Conf on Ontologies, Databases and Applications of Semantics. Berlin: Springer-Verlag, 2005. 1152-1169

[23] Zheng Dongdong, Hu Wei, Qu Yuzhong. An approach to matching between relational database schemas and ontologies [C]. In: Proc of the 2nd Jiangsu Computer Conference. Nanjing: Southeast University Press, 2006. 209-213 (in Chinese)
(郑东栋,胡伟,瞿裕忠. 一种关系数据库模式和本体间的匹配方法[C]. 见:第二届江苏计算机大会论文集. 南京:东南大学出版社, 2006. 209-213)

[24] J Barrasa, O Corcho, A Gomez-Perez. R2O, an extensible and semantically based database-to-ontology mapping language [C]. Int'l Workshop on Semantic Web and Databases, Toronto, 2004

[25] A Zimmermann, J Euzenat. Three semantics for distributed systems and their relations with alignment composition [C]. In: Proc of the 5th Int'l Semantic Web Conference. Berlin: Springer-Verlag, 2006. 16-29

[26] I Astrova. Reverse engineering of relational databases to ontologies [C]. In: Proc of the 1st European Semantic Web Symposium. Berlin: Springer-Verlag, 2004. 327-341

[27] Y An, A Borgida, J Mylopoulos. Constructing complex semantic mappings between XML data and ontologies [C]. In: Proc of the 4th Int'l Semantic Web Conference. Berlin: Springer-Verlag, 2005. 6-20

[28] E Prud'hommeaux, A Seaborne. SPARQL query language for RDF [OL]. http://www.w3.org/TR/rdf-sparql-query/, 2006

[29] D McDermott, D Dou. Representing disjunction and quantifiers in RDF [C]. In: Proc of the 1st Int'l Semantic Web Conference. Berlin: Springer-Verlag, 2002. 250-263

[30] H H Do, E Rahm. COMA—A system for flexible combination of schema matching approaches [C]. In: Proc of the 28th Int'l Conf on Very Large Data Bases. Los Alamitos: IEEE Computer Society Press, 2002. 610-621

[31] G Stoilos, G Stamou, S Kollias. A string metric for ontology alignment [C]. In: Proc of the 4th Int'l Semantic Web Conference. Berlin: Springer-Verlag, 2005. 623-637

[32] S Melnik, H Garcia-Molina, E Rahm. Similarity flooding: A versatile graph matching algorithm and its application to schema matching [C]. In: Proc of the 18th Int'l Conf on Data Engineering. Los Alamitos: IEEE Computer Society Press, 2002. 117-128

[33] E Mena, A Illarramendi, V Kashyap, *et al*. OBSERVER: An approach for query processing in global information systems based on interoperation across pre-existing ontologies [J]. Distributed and Parallel Databases, 2000, 8(2): 223-271

[34] D Beneventano, S Bergamaschi, F Guerra, *et al*. Synthesizing an integrated ontology [J]. IEEE Internet Computing, 2003, 7(5): 42-51

[35] J Euzenat, H Stuckenschmidt, M Yatskevich. Introduction to the ontology alignment evaluation 2005 [C]. Int'l Workshop on Integrating Ontologies, Banff, 2005

[36] J Euzenat, M Mochol, P Shvaiko, *et al*. Results of the ontology alignment evaluation initiative 2006 [C]. Int'l Workshop on Ontology Matching, Athens, 2006

Qu Yuzhong, born in 1965. Professor and Ph. D. supervisor of the School of Computer Science and Engineering, Southeast University. His main research interests include software engineering, semantic Web and Internet computing.

瞿裕忠,1965 年生,教授,博士生导师,主要研究方向为软件工程、语义网及网络计算.

Hu Wei, born in 1982. Ph. D. candidate in the School of Computer Science and Engineering at the Southeast University. His current research interests are Semantic Web, ontology mapping, and data integration.

胡 伟,1982 年生,博士研究生,主要研究方向为语义网、本体映射及数据集成(whu@seu.edu.cn).

Zheng Dongdong, born in 1981. Received his M. S. degree in the Computer Software and Theory from Southeast University. His main research interests are semantic Web and Web service.

郑东栋,1981 年生,硕士,主要研究方向为语义网及 Web 服务.

Zhong Xinyu, born in 1968. Lecturer of the School of Computer Science and Engineering, Southeast University. Her main research interests are semantic Web and knowledge representation.

仲新宇,1968 年生,讲师,主要研究方向为语义网及知识表示.

Research Background

The vision of a next generation Web is, despite all the efforts to build up the semantic Web, still more a dream than a reality. The main reason for this situation is the lack of data, since the vast majority of data are still stored in relational databases and thus unavailable for most semantic Web applications. To date, quite a lot of researches try to solve this problem by discovering mappings between relational database schemas and ontologies, which have been proven to be an effective way to establish interoperability between relational databases and ontologies. This paper firstly presents some formal definitions of mapping between relational database schemas and ontologies, and analyzes major difficulties in this research issue. Then, this paper surveys popular existing solutions according to three different facets, i.e., the approaches of model transformation, the scopes of mapping strategies, and the expressions of mapping results. Furthermore, this paper compares six existing mapping tools, and highlights their unique characteristics in details. Finally, this paper summaries and discusses several remaining challenges and research directions in the future. This work is supported in part by the NSFC under grant 60573083, and in part by the 973 Program of China under grant 2003CB317004.