一种基于 Ontology 的异构数据库语义集成方法

An Ontology – Based Approach to Heterogeneous Database Semantic Integration

吴玲丽 余建桥 孙荣荣 （西南大学 计算机与信息科学学院 重庆 400715）

摘 要： 随着数据的大量增加，数据之间的结构异构和语义异构成为数据集成的重点与难点。本文利用 Ontology 语义集成上的优点，提出了一种基于 Ontology 的异构数据库的语义集成框架，并提出采用基于概念名称语义相似性、属性类型相似度和实例相似度的语义映射方法来重点解决语义集成中的映射问题。

关键词： 本体 语义集成 异构数据库 映射 相似度

随着信息的增长，越来越多的企业部门采用数据库来存储和管理信息，但是这些数据库通常是异构的，要实现异构数据库之间的互操作就必须对异构数据库进行集成。

异构数据库的异构主要有 2 种：结构异构和语义异构。结构异构主要是指存储结构的不同，存储数据的数据库种类不同。通常采用神经网络、XML 等方法可以解决异构数据库集成中的结构异构问题，但却不能解决语义异构。

语义异构主要是：同一概念在不同的数据库中有不同的表示；同一概念可能在不同的数据库有不同的意思；不同的数据库可能用不同的数据结构来表达相同的概念。而本体（Ontology）在概念层上提供了一组用于表达和沟通领域知识的词汇，以及包含这些词汇的一组关系。考虑到本体潜在的描述信息源的语义和解决异构问题的能力，数据集成中利用本体可以很好的解决语义异构的问题。

1 本体在语义集成上的作用

1993 年，Gruber 给出了本体的一个定义，即“Ontology 是概念模型的明确的规范说明”。通俗地讲，本体是用来描述某个领域甚至更广范围内的概念以及概念之间的关系，即提供表示和交流领域知识的词汇，以及在概念层次上提供包含词汇术语的关系集合，使得这些概念和关系在共享的范围内具有大家共同认可的、明确的、唯一的定义，这样，人机之间以及机器之间就可以进行交流。综合本体以上的特点，采用本体来进行语义的集成具有以下的优点：

（1）Ontology 提供丰富的预先定义的词汇，为数据源提供概念视图，而且独立于数据源模式。

（2）Ontology 表示的知识能支持所有相关数据源的转换。

（3）Ontology 支持一致性管理和不一致性数据识别。

（4）Ontology 通过提供共享的术语，能够在语义集成中发挥重要作用。

2 一种基于语义的本体集成方法

2.1 常用的本体映射方法

对异构数据库的集成的最终目标就是为用户建立一个信息集成的模式，在该模式上用户可以更方便、快捷的查找需要的信息。在集成中涉及到两层映射，一层是局部 ontology 到数据源的映射，另外一层是 ontology 之间的映射，通常 ontology 之间映射的方法有以下 4 种：

（1）定义映射的方法。

（2）采用词汇关联的方法。

（3）采用建立顶级本体的方法。

（4）语义对应。

采用定义映射的方法用户可以随意定义映射，在给用户更多的权限的同时导致语义不合理的冲突；采用词汇关联的方法是因为本体间的关系有同义词、交

叠、不相关等，该方法是基于查询策略，即当用户对一个本体中词汇进行查询的时候，系统自动将其扩展到其他本体的词汇库，从而完成语义的映射；采用建立顶级本体的方法是借助形式化语言来建立各个领域的顶级本体，此方法可以消除本体间的冲突和不确定性，但是不能进行直接的对应关系；采用语义对应即建立一个映射词库，在公共词库中存储本体间的语义映射关系，克服了采用建立顶级本体方法带来的对应不明确性。

由此可见单纯的采用任何一种方法都不能很好的解决数据集成中本体间映射的语义问题。采用建立顶级本体可以消除本体间的冲突和不确定性，采用建立映射词库的方法又可以解决采用顶级本体所带来的对应不确定性，因而提出了采用建立顶级本体与映射词库相结合的方法，进行局部 ontology 之间的集成。

2.2 基于语义本体集成的框架

综合以上问题，本文采用建立顶级本体与映射词库相结合的方法，在进行局部 ontology 间的映射的同时，把相互映射的本体概念添加到映射词库中。顶级本体的建立主要有两种方法：自顶向下和自底向上方法。自顶向下方法是在领域专家的参与下建立顶级本体，由本体来统一底层的各信息源语义。自底向上方法是首先抽取底层的各个异构数据源的局部数据模式，再在局部数据模式上抽取局部概念模式（局部本体），然后对局部模式的本体进行集成，最后在局部模式上构建全局的概念模式（即顶级本体）。

自顶向下的方法通常要求领域专家的参与，对专家的依赖程度较高，因而采用自底向上的方法。即首先把局部的数据源进行数据抽取、转换形成局部本体，即将不同数据库，例如 SQL Server、Oracle 等数据库中存储的数据源转换成用 ontology 描述的标准形式；其次，在此基础上进行局部 ontology 间的映射，采用计算概念名称语义相似性、属性类型相似度和实例相似度相结合的方法计算语义相似度；同时把计算所得的相互映射的顶级本体概念及其所包含的局部本体，添加到映射词库表中。其框架结构如图 1。

2.3 语义集成映射

要减少全局语义模式和局部语义模式之间映射的工作量，必须着重解决全局语义模式和局部语义模式之间映射的自动建立，尽量减少人工干预，实现系统的“自动化”。采用自底向上建立顶级本体的方法进行异构数据库的集成，主要完成以下两层映射。

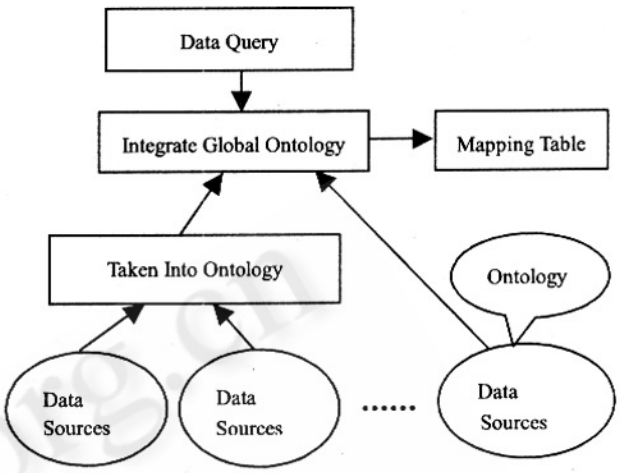

图 1　语义集成框架

2.3.1 局部本体与数据源的映射

从数据源中抽取出数据进行概括形成局部本体，即实现局部本体与数据源的映射。建立局部本体主要包含两个步骤：分析数据源和建立局部本体。首先对数据源进行分析，得到对数据源的概念集合；然后确定建立的本体要表达概念集合中的哪些概念，如何表达；最后选定一种本体表达语言，完成对局部本体的建立。

OWL（Web Ontology Language）能够被用于清晰地表达数据源中的概念以及这些概念之间的关系。而对于数据源概念及它们概念关系的表达即 Ontology。例如我们用 OWL 建立一个花的类：< owl:Class rdf:ID ="Flower" / > ID 为类的名字，并没有指定花类的其他任何信息，例如该类的成员等等。

```
<owl:Class rdf:ID ="Rose" >
<rdfs:subClassOf rdf:resource ="#Flower" />
...
```

</owl:Class > 定义花类的子类玫瑰，此外还可以用 datatype property 来描述类元素和 XML 数据类型之间的关系；object property 来定义两个类元素之间的关系。例如：

```
<owl:ObjectProperty rdf:ID ="HaveGill" >
<rdfs:domain rdf:resource ="#Flower" />
<rdfs:range rdf:resource ="#Rose" />
</owl:ObjectProperty >
```

上面的定义 property HaveGill 给出了类 Flower 的

元素和类 Rose 的元素之间的关系。通过此方法，采用 OWL 建立局部本体，完成数据源到局部本体的映射。

2.3.2 局部本体间的映射

Ontology 之间的映射是异构数据库 ontology 之间互操作的关键问题。异构数据库 ontology 的映射主要解决两种冲突：语义冲突和结构冲突。其中解决语义冲突的算法很少，不能很好的解决语义冲突问题。就此问题，本文提出了以下解决方法。

在 ontology 的映射过程中，采用计算语义相似度的方法可以很好的完成 ontology 之间的映射。考虑 ontology 中的概念属性，确定概率名称的相似性用 WordNet 来计算概率的语义距离。WordNet 就是一部树状的语义字典，树状图上两片树叶之间的距离，就是这两个概念的语义距离。在一开始进行训练操作的时候用以下公式来计算概念的语义距离：

$$Sim(w_1,w_2)=\frac{dis(c_1,c_a)}{2dis(c_1,c_{root})}+\frac{dis(c_2,c_a)}{dis(c_2,c_{root})}$$

其中 c_1,c_2 表示的是概念 w_1,w_2 的具体语义含义，c_a 是二者的父概念节点，c_{root} 是 w_1,w_2 所在分类树的根节点；$dis(c_2,c_2)$ 表示它们在 WordNet 语义树中的路径长度。在计算概念语义距离的同时，根据已有的本体映射词库，再计算本体映射词库中本体概念的语义相似度

$$Sim(w_1,w_2)=\frac{2\log(P(lso(w_1,w_2)))}{\log(P(w_1))+\log(P(w_2))}$$

其中 $lso(w_1,w_2)$：是 (w_1,w_2) 的最近共有祖先，$P(w)$ 是在建立起的本体映射词库中出现的概率。最后综合以上 2 个公式，得出语义距离的综合相似度。

概念类型相似度主要是通过定义概念之间的数据类型的值 $Sim(a_1,a_2)$ 来判断的，其中 a_1,a_2 为本体 O_1，O_2 的概念属性类型。概念类型相同的时候相似度为 $Sim(a_1,a_2)=1$，例如可能的相似值 int 和 int 型的相似度为 $Sim(a_1,a_2)=1$，int 和 float 型的相似度 $Sim(a_1,a_2)=0.6$，bit 型和 small int 型的相似度 $Sim(a_1,a_2)=0.6$ 等。在具体操作的时候具体根据概率定义相似度。

实例相似度可通过以下公式进行计算：

$$RSim(A,B)=\frac{P(A\cap B)}{P(A\cup B)}=\frac{P(A,B)}{P(A,B)+P(A,\bar{B})+P(\bar{A},B)}$$

其中 A，B 为本体 O_1,O_2 的概念 C_1,C_2 的属性，U_i 表示本体 O_i 的全部实例集，$N(U_i)$ 表示 U_i 中实例的大小，$N(U_i^{A,B})$ 表示 U_i 中既属于 A 也属于 B 的实例数目。所以 $P(A,B)=\frac{N(U_1^{A,B})+N(U_2^{A,B})}{N(U_1)+N(U_2)}$。

最后采用 Sigmoid 函数，这是一种对较高的相似度赋予高权重，较低相似度赋予低权重的计算方法。公式

$$Sim=sig_1(Sim(w_1,w_2))+sig_2(Sim(a_1,a_2))+sig_3(RSim(A,B))$$

得到概念的相似度，最后根据相似度完成语义映射，并将计算所得的顶级语义映射关系及其所对应的局部本体的映射关系存储到映射词库中去，更新原有的映射。

2.3.3 映射表的建立

在概念相似度的计算过程中，同时建立起本体映射匹配表，记录顶级本体所包含的局部本体，以及 ontology 之间的映射关系，某一本体概念映射概率等等，以便在搜索查找中能很快的找到数据源所在，以及在对数据源进行扩充的时候，可以参考本地的映射词库，提高映射效率。

2.3.4 优化查询

所有数据的集成都是为了帮助异构数据源的高效准确使用。在建立好顶级本体以及匹配表的基础上，再进行查询。其主要查询步骤如下：

(1) 建立在顶级本体上的查询语句；

(2) 根据建立的顶级本体映射表直接映射到局部本体模式中；

(3) 查找各个局部本体库，完成从局部本体到数据源的转化；

(4) 结果返回，可以选择结果的表现形式，如列表形式、XML 文件形式等。

3 算法分析

由 ontology 在语义集成上的优点可知运用 ontology 进行异构数据库的集成比神经网络，XML 等方法更好地解决语义异构的问题。ontology 语义集成中常用的本体匹配方法有 PROMPT、SAT、S－MATCH 方法。S－MATCH 只考虑了元素级的本体匹配，没有考虑结构级的匹配，同时通常 S－MATCH 方法只是考虑的语义字典中的本体语义匹配，针对性不强，并且学习时间长。

本文提出在集成体系上增加建立映射词库，存储局部 ontology 之间的映射关系，在具体匹配上采用计

（下转第 37 页）

(上接第 33 页)

算词库中本体概念语义距离与映射词库中出现概率语义距离,同时计算本体概念间的数据类型相结合的方法,可以缩短学习的时间。

4 结论

语义集成是进行异构数据库互操作的关键,本文提出了一个比较完整的用 ontology 来实现异构数据库的集成框架,采用自底向上的方法建立其局部本体,再采用计算概念语义相似度以及概念类型相似度和实例相似度的方法来计算本体概念语义相似度从而完成语义映射,同时建立映射表反映领域的本体映射,加快本体匹配速度。用 ontology 来实现异构数据库的互操作是现在一个很有发展潜力的方向。如何更好的提高映射的查全率和准确率还需要进一步的研究。

参考文献

1 Helena Sofia Pinto, Joao P Martins. A Methodology for Ontology Integration . IEEE, 2001:1 - 8.

2 Yannis Kalfoglou, Marco Schorlemmer. Ontology Mapping: the State of the Art. The Knowledge Engineering Review, 2003, 18(1): 1 - 31.

3 An Hai Doan, Jayant Madhavan, Robin Dhamankar, et al. Learning to match ontologies on the Semantic Web . The VLDB Journal, 2003:1 - 17.

4 Qi He, Tok Wang Ling. An ontology based approach to the integration of entity - relationship schemas. Data Knowledge Engineering, 2005:299 - 327.

5 Jie Tang, Juanzi Li, Bangyong Liang, et al. Using Bayesian decision for ontology mapping. Web Semantics, 2006: 243 - 262.

6 Andreia Malucelli, Daniel Palzer, Eugenio Oliveira. Ontology - based Services to help solving the heterogeneity problem in ecommerce negotiations . Electronic Commerce Research and Applications,2006:29 - 43.