基于本体的异构数据共享研究

赵国增[1,2],郭恒川[2]
(1.太原理工大学,山西 太原 030024;
2.洛阳理工学院,河南 洛阳 471023)

摘 要:随着MIS的广泛应用,数据的存储和表示呈现出各种各样的异构性。传统的异构数据共享方法不能对数据的形式化语义进行描述,难以解决异构信息源中的语义异构。"本体"是共享概念模型的明确的形式化规范说明,作为通用语义模型来描述异构数据语义具有较强的概念表达能力和推理的能力。文中利用元数据对分散、异构数据进行明确的规范,给出了一种基于本体的异构数据共享方法。实验表明,该方法能较好地解决异构数据的语义异构问题,为用户提供具有一定语义功能的信息共享。

关键词:本体;语义;元数据;异构;映射
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2010)10-0039-04

Research on Heterogeneous Data Sharing Based on Ontology

ZHAO Guo-zeng[1,2], GUO Heng-chuan[2]
(1. Taiyuan University of Technology, Taiyuan 030024, China;
2. Luoyang Institute of Science and Technology, Luoyang 471023, China)

Abstract: With the extensive application of MIS, data storage and representation showing a variety of heterogeneous nature. The traditional method of heterogeneous data sharing can not describe the formal semantics of data, it is difficult to resolve the semantics of heterogeneous information sources heterogeneous. "Ontology" is a shared conceptual model of a clear formal specification, as a general semantic model to describe the heterogeneous data semantics of the concept has strong expressive power and reasoning capabilities. In this paper, meta-data for distributed, heterogeneous data gives a clear normative ontology-based heterogeneous data sharing methods. Experiments show that the method can better solve the heterogeneous problem of heterogeneous data semantics to provide users with certain semantic features of information sharing.

Key words: ontology; semantic; metadata; heterogeneous; mapping

0 引 言

随着数据库和网络技术的发展和广泛应用,在企业的信息化建设过程中,由于信息管理系统的不同,数据所面向的应用及服务不同,以及其它等因素的影响,积累了大量的异构数据源。主要体现在系统异构、结构异构、语法异构和语义异构。这使得相互联系的部门间不能交换和共享信息,形成了一个个"信息孤岛"。这种现象同时又引起数据冗余以及数据不一致,严重阻碍了其信息化建设的进程。异构数据的共享就是屏蔽信息源的异构性,在保持数据完整的前提下,为用户提供一个统一的访问接口。

近年来,已经有许多技术方案和体系模型应用到异构数据集成中,如联邦数据库方式、数据仓库方法、Mediator-Wrapper和基于语义的数据集成模型[1]等方式。由于XML具有可扩展性、结构化语义以及平台无关性的特点,充分满足了互联网和分布式异构环境的需求,能实现数据的表示和数据内容的分离,能使异构数据源之间通过统一的数据模型交换信息,一定程度上实现了数据共享,但信息源中的语义异构难以得到有效的解决。语义异构的主要原因是术语概念在各信息系统中的表示不同,如不同的术语在多个信息源中可能表示同一个概念,相同的术语在多个信息源中可能表示不同的概念,度量单位不同等等。由于各信息源的分布性、自治性,各信息源中概念之间的隐含联系不能体现出来。本体具有语义表达能力和推理能力,它既可以描述概念术语的含义,又能表达它们之间的内在联系,并且能通过逻辑推理来挖掘概念术语之

收稿日期:2010-02-15;**修回日期**:2010-05-24
基金项目:太原市大学生创新项目(2007010726)
作者简介:赵国增(1977-),男,河南浚县人,太原理工大学硕士研究生,洛阳理工学院讲师,研究方向为计算机网络。

间的隐含关系,因此本体被逐渐应用于异构数据集成中解决语义异构问题[2]。

文中给出了一种基于本体和元数据的异构数据共享方法,来解决异构数据源之间的语义冲突问题,达到了较好的实验效果。

1 相关技术研究

1.1 元数据

为了把分布在不同系统中的异构数据整合起来,实现异构信息资源的共享,需要一种普遍适应的方案对数据进行规范化。元数据用于对信息资源进行描述、解释、定位,是一种结构化的信息,使其易于提取和使用[3]。通常情况下元数据指的是关于资源的结构化数据描述,这些描述是为适应一些广泛的应用而构建的,结构化的目的是让机器不仅能够阅读,而且能够理解这些数据,以便解决信息的映射问题和集成问题。在异构信息资源中,"数据"从本地被传送各个终端,要经过"信息 1—元数据—信息 2"的数据转换。这一过程中,元数据起到了关键的作用。在异构数据资源共享中,元数据能够实现对信息源和信息使用者的明确的描述,通过这些描述为信息源和信息使用者之间转换数据格式,实现异构数据源之间的数据传递。

良好的元数据定义是构建资源描述并实现快速查询和检索的关键,目前,国际上最有影响的元数据标准是 Dublin Core,它已经为许多行业制定了领域相关的元数据标准,目前 Dublin Core 已经被 W3C 接受为标准,由 UKOLN(英国图书馆与信息网络协会)维护管理,除此之外,其它标准如关于文档格式的 TEI 与 MARC、关于几何空间的 FGDC 标准、档案编码标准(EAD)和数据文档编码(DDI)等[4]。

元数据是解决资源形态多样的主要手段,但即使有了标准化的元数据,国际上仍然会存在多种不同的标准。所以,在现代资源共享中,一般都会根据需要,遵循一些大标准,根据实际需要和资源特点等建立自己的资源元数据标准。

1.2 本 体

本体(Ontology)有着与元数据类似的哲学定义:"描述存在的存在",在知识的共享过程中,"本体是共享概念模型的明确形式化规范说明"[5]。自语义网被提出后,本体成为解决网络资源内容互联的一个重要途径,在共享资源的语义表述及推理方面发挥重要作用。当前,利用本体建立各种公共共享知识库,利用本体推理技术进行知识推理的研究正在展开。

异构数据共享中,本体是作为一个中介存在的,它屏蔽了异构数据源的底层结构。用户在执行查询时,可以不知道各种数据源的结构,也不用知道如何去查询数据,只需要知道需要什么数据,给系统提供一个针对于本体的查询就可以了。系统可以根据语义的定义及其映射关系自动将对本体的查询分解为针对各数据源的查询。通过本体的逻辑推理能力,在分布式环境下的异构数据共享中,可以大大提高信息的查准率和查全率[1]。

利用元数据标准规范实现资源数据表现层的统一和互理解,利用本体技术实现资源语义互联,在文献[6]中都有一些相关的报道。文中对异构资源进行分析,结合两种技术,为异构数据资源的共享提供了一种解决方法。

2 基于本体的异构数据共享方法

2.1 本体构建

本体中的知识是使用类(Classes),实例(Instances),关系(Relations),函数(Functions)和公理(Axioms)五种元语来规范化的。所以,一个本体可以用一个五元组来表示[7]:O=(C,I,R,F,A)。C、I、R、F、A 分别表示概念、实例、关系、函数和公理。在本体中的类常常分类组织,以概念的形式出现。以此五个最基本的元语对概念进行描述[8]。

(1) 类(class)或概念(concept):表示的是对象的集合,是对现实事物的抽象。一般采用框架结构进行定义,其中包括有概念的名称,以及概念与概念之间的关系集合,并使用自然语言对其进行描述。

(2) 关系(relation):表示在领域中概念与概念之间的相互作用,形式上可以用 n 维笛卡尔积的子集来定义:$R:C_1\times C_2\times\cdots\times C_n$。

(3) 函数(function):是指特殊的关系。可以通过此关系的前面所有的元素来唯一确定下一个元素。形式化定义为 $F:C_1\times C_2\times\cdots\times C_{n-1}\rightarrow C_n$。

(4) 公理(axiom):就是在一个理论系统中被公认为真的命题。

(5) 实例(instance):实例表示具体的元素,就是对象。

一旦本体的关键元语表示完成之后,本体就可以选用符合实际需求的语言来表示。而本体的概念间还具有不同的关系,基本关系的具体形式如表 1 所示。

表 1 概念间关系

关系名称	关系描述	举例
part-of	部分与整体关系	轮胎与汽车
kind-of	继承关系	交通工具与汽车
instance-of	实例与概念关系	桑塔纳 2000 与汽车
attribute-of	属性关系	车颜色与车

本系统也采用基于混合本体的方法[9],集成方法如图 1 所示。全局本体和局部本体,以及它们之间的映射关系利用本体描述语言(OWL)描述。使用这种方法需要构建一个全局的词汇集,使其中尽可能包含领域中的所有术语(原语)。用户面对的是系统能够集成的数据集合,它由全局本体组成。用户将根据这个全局的概念集合产生查询请求。局部本体描述的是各具体数据源中的概念及其相互关系。局部本体概念与全局本体中的概念相对应(映射关系)。这样系统中就存在着全局本体和局部本体、局部本体和数据源两种映射关系,系统根据建立的映射关系将全局的查询分解成一个个针对具体数据源的子查询,经过合并汇总各子查询的结果生成总的查询结果呈现给用户。

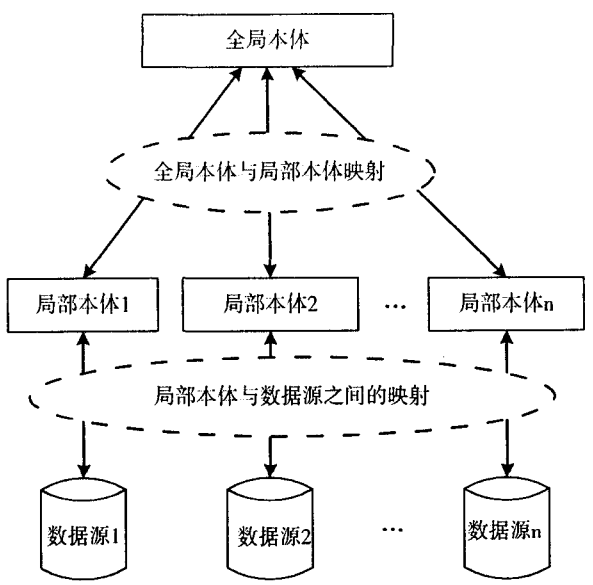

图 1　混合本体与数据源的映射

在关系型数据源中,可能有着不同的数据模型,相同的数据模型中也可能存在语义差异,针对这种情况,通过对关系数据库模式 E/R 模型的分析,确定元数据,根据元数据信息创建本体。把关系数据库中的关系名、字段及实体间的关系创建本体中的类、属性和角色。同时在创建本体过程中把对应关系记录下来,作为局部本体与各数据源之间的映射信息。

假设数据源 DataSource1 和 DataSource2 均表示课程与授课者的信息,其结构如下:

DataSource1:Course(ID,name,time,lecture)

Teacher(TeacherID,name,sex,PresentPosition)

DataSource2:kc(number,kcmc,xs,skz)

Js(bh,xm,xb,zc)

通过对数据源 DataSource1 和数据源 DataSource2 进行分析,分别建立对应的局部本体。DataSource1 的局部本体如图 2 所示。根据元数据对局部本体集成构建全局本体。

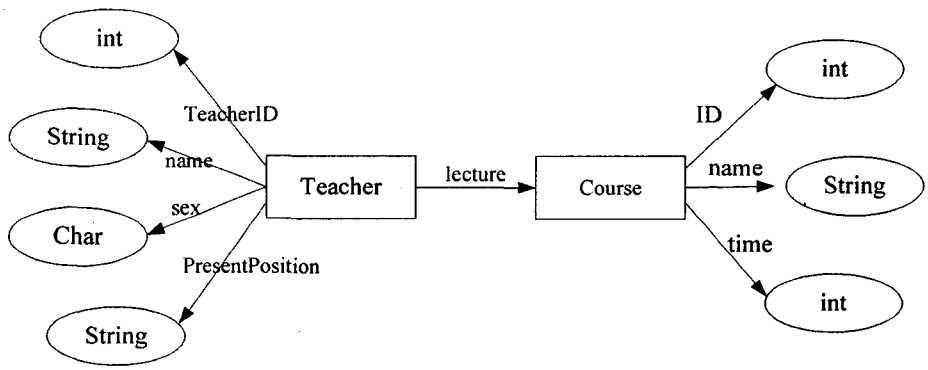

图 2　DataSource1 的本体

通过对异构数据源的分析,发现这里的 Course 和 kc、Teacher 和 js 其实表达的是同一概念。同样,Course 中的 ID、name、time、teacher 和 kc 中的 number、kcmc、xs、skz 具有同样的含义。综合数据源 DataSource1 和 DataSource2 得到全局本体,全局本体的 owl 语言描述[10]部分内容如下。

```
<owl:Ontology rdf:about=""/>
<owl:Class rdf:ID="Course">
<rdfs:subClass of>
<owl:Restriction>
<owl:onProperty rdf:resource="#lecture"/>
<owl: minCardinality rdf: datatype = "&xsd; non-NegativeInteger">
</owl:minCardinality>
</rdfs:subClassof>
</owl:Class>
</owl:Class rdf:ID="Teacher"/>
</rdfs:domain rdf:resource="#name"/>
</rdfs: range rdf: resource = "http://www. w3.org/2001/XMLSchema#int"/>
</owl:Datatypeproperty>
<owl:Datatypeproperty rdf:ID="Teacher ID">
<rdfs:domain rdf:resource="#teachenamer"/>
<rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#int"/>
</owl:Datatypeproperty>
```

2.2　本体映射

本体映射是将两个或多个不同的本体作为输入,

然后按照语义关联关系为这些本体中的元素(概念、属性、关系)建立相应的语义关系。各局部本体和全局本体建立后,还需要将各局部本体与全局本体关联起来。在全局本体的构建过程中,对局部本体综合包括本体中类和角色的映射,形成全局本体的过程就是映射进一步明确的过程。

映射过程[7]分为三步:

(1)建立元数据标准;

(2)读取异构数据源信息;

(3)建立二者之间的映射。

如,根据建立元数据标准,可以将课程 ID 和 number 映射为 kcID, TeacherID 和 bh 映射为 TeacherID 等。

本体之间的映射主要包括 1:1、1:n、n:1、1:null、null:1 和 n:m 共 6 种映射模型。在具体映射过程中,可以通过元数据的相似度方法进行映射。当本体中的元素 $e_{1,i}$(概念、属性、实例)与本体中的元素 $e_{2,j}$ 的相似度 $\mathrm{Sim}(e_{1,i}, e_{2,j}) > s$ 时,两个元素符合映射关系,有:$\mathrm{map}(e_{1,i}) = e_{2,j}$。其中 s 设定相似度的阈值[11]。

2.3 查询处理

查询处理是异构数据共享的主要研究内容,其目标是从多个分布异构数据源中查询数据,并把这些查询结果进行融合,然后将最终的结果提交给用户。根据文献[12],用户查询是针对全局本体中的概念和属性的查询,可以将查询 Q 定义为三元组<S,F,W>,其中 S 为 select 子句, F 为 from 子句,W 代表 where 子句。查询处理模块接收来自用户的查询请求后,使用全局本体的术语表示用户查询语句,再通过对全局本体进行推理得到相近或相似的概念。再通过全局与局部本体的映射关系将全局查询中的概念、属性转化为局部本体中对应的概念和关系。查询子句通过转换分解为对异构信息源的子查询。子查询返回查询结果后,由包装器进行格式转换,然后查询处理模块对查询结果进行语义分析、汇总、排序和除去冗余等操作,完成查询结果的组装集成,最终提交给用户。

例如,查询课程名称为"Java 程序设计"的授课教师姓名、课程编号、授课学时,则提交的全局本体查询为:

Select c,d,e

from Teacher = a, a. lecture = b, a. name = c, b. kcID = d, b. time = e

where b. name = 'Java 程序设计'

根据映射内容将其分解为子查询:

DataSource1: select name, id, time from Course, eacher

Where name = 'Java 程序设计' and course. lecture = Teacher. TeacherID

DataSource2: select xm, number, xs from kc, Js

Where kcmc = 'Java 程序设计' and kc. skz = Js. bh

3 结束语

利用本体技术和元数据对分散、异构的数据的规范解决了异构信息源之间的共享问题。在本体映射过程中,仍采用人工参与的方法。在今后的研究中,要更深入地探究异构数据语义共享中的本体映射问题,提高映射准确度和自动化程度。随着语义网技术的不断成熟,本体将会在解决异构数据集成的语义异构问题中得到越来越多的应用。

参考文献:

[1] 于 琦,周 勇. 一种基于本体的异构数据源模式集成[J]. 计算机技术与发展,2008,18(2):35-36.

[2] 李星毅,高文浩,施化吉. 基于本体的异构数据集成方法[J]. 计算机工程与设计,2009,30(8):1931-1934.

[3] 张 宇,蒋东兴,刘启新. 基于元数据的异构数据集整合方案[J]. 清华大学学报:自然科学版,2009,49(7):1021-1022.

[4] 俞 时. 异构资源中基于本体的信息互操作性研究[D]. 上海:东华大学,2003:13-15.

[5] 宋 炜,张 铭. 语义网简明教程[M]. 北京:高等教育出版社,2004:56-60.

[6] 王 莉,高仲利. 基于 Web Services 的异构数据源集成研究[C]//第三届中国智能计算大会论文集. 济南:[出版者不详],2009:79-80.

[7] Gruber T R. A Translation Approach to Portable Ontology Specifications[J]. Knowledge Acquisition, 1993(5): 199-220.

[8] Sheth. Changing focus on interoperability in information system: Form system, syntax, structure to semantics[M]//Interoperating Geographic Information Systems. Boston: Kluwer Academic Publisher, 1999:13-14.

[9] Wxche H, Vogele W T, Visser U, et al. Ontology-based integration of information-a survey of existing approaches[C]// Proceedings of IJCAI-01Workshop: Ontologies and Information Sharing. Seattle, WA:[s. n.], 2001:108-117.

[10] 语义网基础教程[M]. 陈小平译. 北京:机械工业出版社,2008:89-107.

[11] 郭 鑫. 基于本体的异构数据集成技术研究与实现[D]. 中国航空第二研究院,2008:28-29.

[12] 严小泉,刘 渊. 基于 XQuery 的异构数据源查询处理[J]. 计算机工程,2009,35(14):87-89.