基于本体的信息集成框架研究

刘 萍,李绪蓉
(南京航空航天大学 信息科学与技术学院,江苏 南京 210016)

摘 要:为了解决信息集成中的语义异构问题,引入本体技术,借鉴"Mediator/Wrapper"体系结构,结合混合本体以及Web Services技术,提出了分布式网络环境下的基于本体的信息集成框架,阐述了如何利用本体技术解决语义异构问题,给出了框架的层次结构以及关键技术,包括本体构建、查询处理和服务注册中心,然后利用原型系统验证了框架是可行的。该框架解决了信息集成中的语义异构问题。

关键词:本体;信息集成;OWL;Web Services
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2007)12-0034-03

Research on Ontology-Based Information Integration Framework

LIU Ping, LI Xu-rong
(Department of Information Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

Abstract: Ontology technique resolves the problems of semantic heterogeneity of information integration, proposes the ontology-based information integration framework based on the architecture of "Mediator/Wrapper", mixed ontology and Web Services, then expounds how to use ontology to solve the problems of semantic heterogeneity, the structure and key technique of the framework are given. The key technique includes building ontology, inquiry processing and the registered center of services. the framework is verified by the prototype system. The framework resolves the problems of semantic heterogeneity of information integration.

Key words: ontology; information integration; OWL; Web services

0 引 言

信息集成可以在不更改原有系统的前提下建立全局的信息视图,集中式地查看分布在不同系统、不同平台和分布式的网络环境下的信息资源。解决信息集成的方法有很多,其中有一种是基于XML的信息集成[1],此方法能较好地解决信息源语法上的异构,但是对于语义上的异构解决能力较弱。文中引入本体技术来解决语义异构问题,提出了分布式网络环境下基于本体的信息集成框架,给出了框架的层次结构、关键技术和框架的验证。

1 基于本体的信息集成框架

目前基于本体的信息集成方法主要有三种[2]:单本体、多本体和混合本体方法。混合本体,一方面不同的用户建立局部本体与各信息源相联,避免了局部结构改变对全局的影响;另一方面,在各个局部本体之上使用共享的词汇集合,即全局本体,它包含了领域中的基本术语,它是构建局部本体的基础,通过它,局部本体之间能够实现互操作。本体映射保证了全局本体与局部本体间语义的一致性。因此文中使用混合本体来解决语义异构问题;因Web Services技术提供网络服务,可跨平台、跨语言、跨操作系统并能穿越防火墙,这就可解决信息源在硬件设备、运行平台、实现语言、通信协议等方面的异构问题;国外著名信息集成项目[3]如TSIMMIS, DISCO, SIMS等大多采用了Mediator/Wrapper体系结构。图1给出了基于本体的信息集成框架,该图也是借鉴"Mediator/Wrapper"体系结构,结合混合本体以及Web Services技术设计而成的。该框架的最大特点是能实现信息源的动态增长,真正实现了信息源的"即插即用"。

1.1 框架语义异构的解决方案

语义异构包括很多种[4],文中主要研究命名、属性、外延异构等。这些类型异构中,有的可以在定义本体时解决,有些需要在本体与本体、本体与信息源的映射中解决。文中采用OWL来描述信息源以及本体映射,语义异构问题的主要解决方案如下:

收稿日期:2007-02-22
作者简介:刘 萍(1982-),女,江苏江阴人,硕士研究生,研究方向为系统集成;李绪蓉,副教授,博士,研究方向为系统集成与重构。

(1)命名异构。

①不同的信息源使用多种术语表示同一概念,该类型异构可通过本体映射中 owl:equivalentClass 和 owl:equivalentProperty 进行解决;

②同一术语在不同的信息源中表达不同的含义,该种类型的异构通过不同信息源对应不同局部本体得到解决;

③对于同一信息源里的不同实体使用相同的名字,在本体定义中对概念采用附加上层概念加以解决;

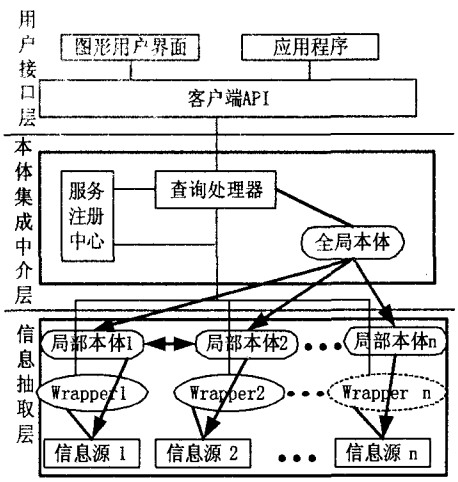

图 1　基于本体的信息集成框架

(2)属性异构,可通过使用本体定义中的 rdfs:subPropertyOf 语法加以解决。在 OWL 中可以通过(owl:hasValue)限制某一属性的值。

(3)外延异构,包括四种:

①等价元素,可通过本体映射中 owl:equivalentClass 和 owl;equivalentProperty 进行解决;

②交叉元素,可通过本体定义语法中 owl:intersectionOf 语法进行解决;

③包含元素,可通过本体定义语法中 rdfs:subClassOf 以及 rdfs:subPropertyOf 进行解决;

④不相交元素,可通过全局本体中使用 owl:unionOf 语法进行解决。

1.2　框架的层次结构

整个框架从下到上分为三个层次:信息抽取层、本体集成中介层、用户接口层。

(1)信息抽取层(Information Collection Layer)。该层采用 Wrapper。Wrapper 功能有:当服务发布时,它负责将底层异构信息源提供数据的行为封装成服务,服务的描述、注册、服务输入参数等按照局部本体进行描述,然后将服务注册到服务注册中心;当服务被调用时,从服务调用请求中取出服务输入参数,并根据信息源和局部本体的映射关系将输入参数转换为系统内部使用的信息形式,然后执行服务中的具体操作,并将查询结果返回给本体集成中介层中的查询处理器。

(2)本体集成中介层(Ontology Integration Mediation Layer)。主要功能有对上接受客户端 API 向系统提交的查询请求,查询处理器首先把查询请求按照全局本体进行语义扩展,形成全局查询,再根据全局本体与局部本体的映射关系,将该全局查询转换为针对各个局部本体的子查询,并把子查询发送给服务注册中心,结合注册中心提供的信息将子查询与服务匹配,然后分别执行,服务被调用的过程中驱动信息源进行查询。最后,将查询结果进行整合,以统一的 XML 格式传回用户接口层。

(3)用户接口层(User Interface Layer)。负责将用户的查询命令提交给本体集成中介层,并将 XML 结果显示给用户。

1.3　框架的关键技术

1.3.1　本体构建

1) 建立局部本体。从各信息源抽取出各种数据模式,在此基础上形成各局部本体,两个主要步骤如下:

(1)分析数据源:对每个信息源做全面分析,这种分析是独立进行的,不用考虑其他的信息源。

(2)定义局部本体:在分析信息源和查找相关术语(原语)的基础上,建立各信息源对应的局部本体,同时也形成局部本体到信息源的映射。

本框架中有一信息源为关系数据库,基于关系数据库模式的局部本体创建策略如下:每个数据库对应一个局部本体,局部本体的 OWL 文件的文件名为预先定义的信息源名;表对应局部本体里的类,表名即为类名;表中的字段对应局部本体中该类的属性,属性名即为字段名,其中表的外键字段定义为类的 ObjectProperty,并定义其 domain 和 range 值以指定具有外键关系的两个表。其余字段都定义为类的 DatatypeProperty。另一信息源为 XML 数据,局部本体可根据 XML Schema 直接构建,还有一个信息源为 OO 数据库,可根据数据库模式构建,类似于关系数据库。

2)形成全局本体。该阶段的任务是在对各信息源所涉及的领域知识的调研和对各局部本体的分析的基础上建立共享的全局本体。包含如下 3 个主要步骤:

(1)分析信息源:对所有信息源进行一个完整分析(不能只考虑单个信息源)。

(2)查找术语:研究分析局部本体中的共享概念及概念的关系,选择应该写入共享词汇库的概念或术语。

(3)定义全局本体:在上一步基础上利用共享概念或术语创建全局本体。

3)定义映射。在这个阶段的任务是定义全局本体和局部本体之间的概念的映射和关系,这是解决语义异构问题的途径。

1.3.2 查询处理

(1)全局查询为 SQL 格式的查询,查询中的概念和属性来源于全局本体。全局查询定义[5]如下:设 Q 为全局查询,O_g 为全局本体,定义 Q 为元组 $< S, F, W >$。其中:$S = \{gc_i \cdot gp_{ij} \mid \forall i = 1, \cdots, n, \forall j = 1, \cdots, m\}$ 为 SELECT 子句,gc_i 为 O_g 中的概念,gp_{ij} 为 gc_i 的属性,$F = \{gc_i \mid \forall i = 1, \cdots, n\}$ 为 FROM 子句,gc_i 为 O_g 中的概念;$W = W_{con} \cup W_{var}$ 为 WHERE 子句,WHERE 子句分为 W_{con}, W_{var} 两类。$W_{con} = \{W_{con\,i} \mid i = 1, \cdots, s\}$,$w_{con\,i} = gc_k \cdot gp_{kl} \Theta \text{Constant}$ 表示将概念属性同常量相比较;$W_{var} = \{w_{var\,i} \mid i = 1, \cdots, t\}$,$w_{var\,i} = gc_k \cdot gp_{kl} \Theta gc_p \cdot gp_{pq}, k \neq p$ 表示将概念属性相比较。$\Theta \in \{=, <, >, \neq, \leq, \geq\}$。

(2)全局查询分解。通过全局-局部本体映射,将全局查询分解为多个子查询,每个子查询对应一个信息源。在文中提出的框架中,全局本体中概念与局部本体中概念之间的映射关系,参考文中 1.1,在全局本体 OWL 中加入全局-局部本体映射。采用 Jena OWL 推理机对全局本体进行推理。推理可得到全局本体和局部本体的映射,即可得到针对各个局部本体的概念等,按照全局查询 $< S, F, W >$,由查询处理器对全局查询中的 S, F, W 按照推理结果组装出针对各个信息源的子查询。

(3)服务注册中心。服务注册中心存储的仅是服务的描述信息和调用信息,而服务的实现是在信息源内部完成的。因此,当服务的具体实现发生变化时,只要服务的描述信息和调用信息不变,那么就不需要对服务注册中心存储的信息进行修改,将系统的维护工作尽量减少。

1.4 框架的验证

为验证该框架的可行性,采用 J2EE 平台开发了一个信息集成原型系统。该原型系统中,信息源提供的服务是以服务输入参数为查询条件的查询操作,并且设定该输入参数唯一。信息源取自某航空研究所,取其中三种信息源:关系数据库、XML 数据、OO 数据库。

原型系统的开发环境采用 Eclipse3.1 + JBoss4.0。采用 JSP, Struts, EJB, Hibernate, Web Services 技术和 MySql 数据库实现原型系统,本体采用 Protégé 工具开发,采用 Jena OWL 推理机对全局本体进行推理。原型系统实现的用户界面如图 2 所示。

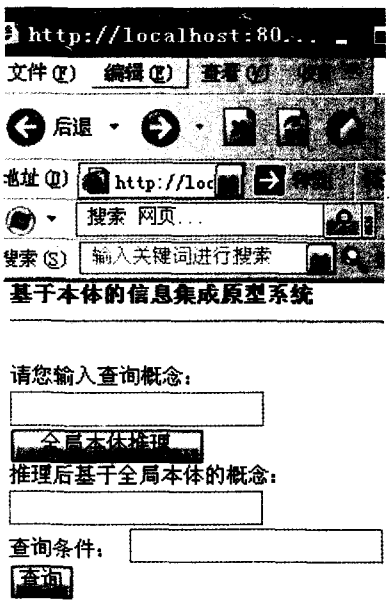

图 2 原型系统实现的用户界面

2 总 结

文中提出的基于本体的信息集成框架,可以屏蔽后台多种异构信息源,全局本体为用户提供了统一的查询接口。利用本体技术,真正实现了语义级的信息资源无缝集成。该信息集成框架处于探索阶段,将在以下几方面继续进行研究:本体映射中概念之间语义相似度的计算;Wrapper 构建的完备性;查询优化等问题。

总之,异构信息集成是当今的一个热点问题,文中所做的工作会促进国内这一领域的研究发展。

参考文献:

[1] 李绪蓉.面向业务构件的可重构信息系统的模型研究[D].南京:南京航空航天大学,2003:99-102.

[2] Wache H, Ogele V, Visser T, et al. Ontology-based integration of information - a survey of existing approaches[C]// USA: IJCAI-01Workshop: Ontologies and Information Sharing. [s.l.]: [s.n.], 2001: 108-117.

[3] Paton N W, Goble C A, Bechhofer S. Knowledge-based Information Integration Systems[J]. Information and Software Technology, 2000(42): 299-312.

[4] Goh Cheng Hian. Representing and reasoning about semantic conflicts in heterogeneous information systems[D]. USA: Massachusetts Institute of Technology(MIT), 1997: 15-56.

[5] 刘文斌.基于本体的信息集成[D].南京:南京航空航天大学,2006.