基于语义的信息集成系统研究

陈 磊[1], 陈 盈[2]
(1.淮南师范学院 计算机与信息工程系,安徽 淮南 232001;
2.台州学院 数信学院,浙江 台州 317000)

摘 要:语义网技术的发展为传统的信息集成领域带来了新的契机,构建基于语义的信息集成系统对于知识的表示、信息的组织与检索等有着重要的意义。通过对领域本体的构建方法和对基于本体的数据集成方法的分析,提出一种基于语义的信息集成模型,改进了传统的基于视图的查询响应方法,在传统的数据库层上构建了RDF/OWL视图,实现了关系模式向本体模式的转换,将SPARQL向SQL转化的查询处理运用于查询重写中。从技术与应用的角度论证了系统的可行性。

关键词:信息集成;本体;语义;查询重写
中图分类号:TP301.2 **文献标识码**:A **文章编号**:1673-629X(2010)06-0074-04

Research on Semantic-Based Information Integration System

CHEN Lei[1], CHEN Ying[2]
(1.Department of Computer and Information Engineering, Huainan Normal University, Huainan 232001, China;
2.School of Mathematics and Information Engineering, Taizhou University, Taizhou 317000, China)

Abstract: The semantic Web technologies bring new chances to the information integration. It is meaningful to construct a semantic-based information integration system for knowledge representation and information retrieval. On the basis of the construction of domain ontology and the analysis of ontology-based data integration, puts forward a semantic-based information integration model, which improves the traditional method of 'Answering Queries using Views', models an RDF/OWL view on the database framework, and realizes the transformation from relational schema to ontology schema. Also apply the transformation of 'SPARQL to SQL' into rewriting of views. At last, the author argues the feasibility of this model from the angle of technology and application.

Key words: information integration; ontology; semantic; views rewriting

0 引 言

互联网上的信息检索机制给人们带来了极大的方便,但也存在着一些问题:面对用户输入的查询条件,系统或给出海量的检索结果,或遗漏了一些目标文档,其主要原因归结起来有两点:

(1)基于文档的网络。现行的互联网是文档之网(Web of Document)。同一领域和主题的信息依赖超级链接进行关联,普通应用程序只对网页上的内容进行简单处理。由于缺乏语义描述的支持,所以处理的结果是大量的文档,其中绝大部分与查询内容无关。

(2)数据缺乏语义。在各种现行的信息索引系统中,与查询信息相关的数据缺乏必要的语义描述,搜索引擎的搜索机制是基于文本的关键词匹配,缺乏推理的支持,导致了大量工作仍需要用户自己完成。

语义网(Semantic Web)技术的出现可以改变这种情况。语义网强调互联网应是数据之网(Web of Data),信息的关联应该依靠链接数据(Linked Data)。根据语义网的原理,语义建立在领域本体(ontology)之上。本体将词汇赋以语义,在此基础上,信息的检索不再是基于关键词的,而是基于语义的检索。因此,文中主要从以下几个方面进行研究,以构建一个基于语义的信息集成系统模型。

1)引入领域本体。

所谓本体就是"给出构成相关领域词汇的基本术语和关系,以及利用这些术语和关系构成的规定这些词汇外延的规则的定义[1,2]。"因而,领域本体规范了领域中的术语以及它们之间的关系,这些信息的总和就表现为数据的语义。

2)基于语义的检索。

传统的基于关键词的检索方式是导致错误匹配与

收稿日期:2009-10-20;修回日期:2010-01-05
基金项目:安徽省优秀青年人才基金项目(2009sqrz164);安徽省自然科学研究项目(kj2009b146z)
作者简介:陈 磊(1980-),男,讲师,博士研究生,研究方向为语义网技术、数据库技术。

海量检索结果的直接原因。将数据赋与语义后,就可以进行基于语义的检索与推理,这将大大地提高了检索的准确性,不会带来无关的垃圾文档,而且,在数据上进行逻辑推理会提高信息检索的自动处理能力。

3)信息集成。

大量信息都分布在网络下层的数据库中,其中异构(heterogeneous)数据源的互操作(interoperability)问题是近年来的领域研究热点。在领域合作化不断扩大的趋势下,分布式数据库的集成有着重要的应用价值。

基于以上分析,笔者拟以领域本体为基础,以数据集成为原理,建立一个基于语义的信息集成系统。用户的查询(SPARQL 查询)是基于语义的;参照基于视图的查询响应(answering queries using views)方法,实施于全局模式(global schema)上的查询转化为数据源上的查询,这就涉及到 SPQRQL 查询向 SQL 查询的转换问题。

1　基于语义的信息集成系统

本节从"本体的构建"、"基于本体的数据集成"、"基于视图的查询设计"、"RDF/OWL 视图及 SPARQL 向 SQL 的转化"四个方面来对模型进行阐述与分析,最后给出系统模型的结构图。

1.1　本体的构建

本体用于描述概念及概念之间的关系,并通过这种描述来定义词汇的语义。作为一种有效表现概念层次结构和语义的模型,本体已经被广泛地应用到计算机科学的众多领域。

根据本体的定义,可以将本体的结构描述成一个五元组[3,4]:

$$O:=\{C,R,H^C,\mathrm{Re}l,A^O\}$$

其中,C 是概念集合,也称为类集合;R 是 C 中元素的关系的集合;H^C 表示 C 中元素的分类关系;$\mathrm{Re}l$ 表示 C 中元素的非分类关系;A^O 表示本体公理。

可以看出,在构造领域本体的时候,最重要的是要确定本体中的概念(类)和它们之间的关系,这种关系称为概念的属性,是概念与概念之间的"桥梁"。领域本体的构建方法很多,属于本体工程的研究范畴,在初步构造的时候,可以选择一种简单可行的办法,文献[5]向人们推荐了一种常见的构建本体的方法,这种方法被称为"七步法":(a)确定本体的范围;(b)考虑复用已存在的本体;(c)列举领域重要术语;(d)定义分类;(e)定义属性;(f)定义侧面;(g)定义实例。

本体的构建是一项复杂的工程,需要一整套的机制来支持和开展。并不存在着一个完全"正确"的本体构建方法,因为至少到目前为止,还没有一个公认的完整的本体评价机制,这也是本体工程中亟待解决的一个问题。

1.2　基于本体的数据集成

数据集成旨在解决异构数据源的互操作问题。一些机构或领域可能拥有很多相互之间有信息关联却无法兼容的异构(heterogeneity)数据源,出于某种需求,用户需要对整个领域进行全局的信息检索,此时,数据的集成就成为解决这一问题的基本手段。

数据的异构主要有结构异构(structure heterogeneity)、语法异构(syntax heterogeneity)和语义异构(semantic heterogeneity)。结构异构和语法异构在传统的数据集成中已经得到较好的解决,但是对于语义上的异构问题,传统的数据集成技术则不能解决。这是因为语义的异构主要来自于不同系统中的数据源使用了不同的概念描述了同一事物,或是使用了相同的概念描述了不同的事物,前者称为异名同义(synonym)问题,后者称为同名异义(homonym)问题。本体的引入有助于解决这一问题。在基于本体的数据集成系统中,本体的作用主要有两个:表示概念和用于构建全局模式(global schema)。本节先介绍表示概念问题,全局模式的构建在 1.3 节中讨论。

在利用本体进行数据集成的方法中,本体可以对各数据源的数据进行语义描述,主要有三种方法[6],如图 1 所示。

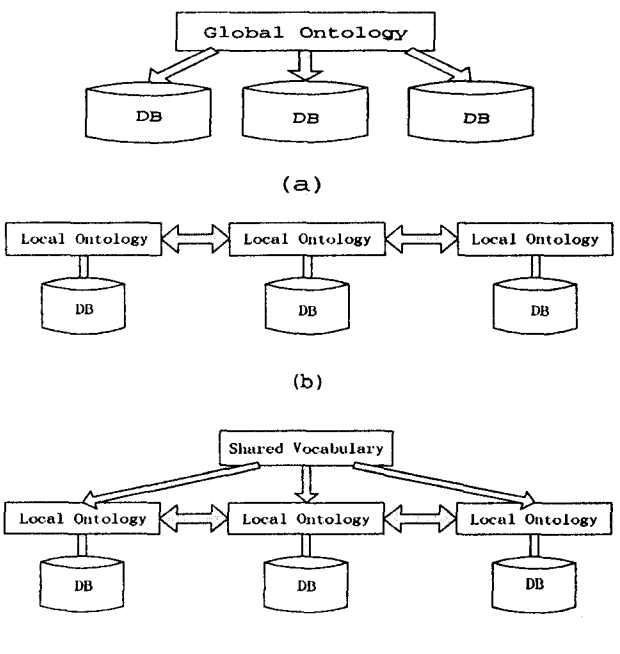

图 1　数据集成中的本体描述数据的三种方法

(a)所表示的是单一本体方法(single ontology approach),所谓单一本体法是指不同的数据源利用同一个全局本体提供共享词汇表,所有的数据源都必须按

照这个全局本体词汇来描述自己的源数据。

(b)所表示的是多本体法(multiple ontologies approach),在多本体法中,每个数据源的数据都用它的本地本体(local ontology)进行描述,各本地本体之间的映射(ontology mapping)是这种方法必须解决的关键问题。

(c)所表示的是混合法(hybrid approach),混合本体法的主要思想是各数据源可以有自己的本地本体,但仍然要建立一个领域共享词汇(shared vocabulary)。本体映射是以共享词汇为基础的。

三种本体描述方法的比较见表 1。

表 1 三种常用本体方法的比较

	单一本体法	多本体法	混合法
实现的难易程度	直接实现	高成本	适中
语义异构	支持域的相近统一视图	支持异构的视图	支持异构的视图
添加/删除数据源	需要改变全局本体	添加一个新的本地本体,提供与其他本体的映射	添加一个新的本地本体
与多本体的比较	——	困难,缺少一个通用的词汇	简单,所有本体使用一个通用词汇

1.3 基于视图的查询响应及 RDF/OWL 视图

在数据集成的应用领域,利用视图进行查询响应(answering queries using views)的主要目的是建立一个全局视图[7]。

数据集成的方法主要有全局视图法(global - as - views,GAV)和局部视图法(local - as - views,LAV)。全局视图法中的全局模式是在数据源视图基础上建立的,它由一系列元素组成,每个元素对应一个(组)数据源,表示相应数据源的数据结构和操作;局部视图法中先构建全局模式,数据源的数据视图则是参照全局模式而定义,由全局模式按一定的规则推理得到[8]。相比较而言,全局视图法容易实现,因为实施于全局模式上的查询只需简单地按规则展开便可转换成各个数据源上的子查询,类似于普通数据库上的查询操作。但是,全局视图法不能很好地支持数据源的更新,因为任一个数据源的更新都可能影响到全局视图。

与全局视图法相比,局部视图法的优点是它较好地支持了数据源的更新,有着良好的可扩展性。基于局部视图法的数据集成系统可简单地定义如下:

设数据集成系统为:

$$D = < G, S, M >$$

这里,G 代表全局模式(global schema),在本模型中可用一种本体描述语言(如 RDF/OWL)表示;S 是源模型(source schema),例如关系模式等;M 是 G 与 S 之间的映射,实际上是一个断言(assertions)集合,形式如下:

$$s \sim > q_G$$

其中,s 代表着数据源里的一个元素,q_G 代表全局模式上的一个视图。从这个描述可以看出,一个 LAV 映射就是一组断言,每个断言描述如何将数据源里的元素描述成全局模式上的一个视图。因此,在全局模式看来,原来的数据源就是一组视图(views),用户实施在全局模式上的查询最终要依靠这组视图进行演算和响应,这正是"answering queries using views"的主要思想。

对于局部视图法而言,当其中的数据源更新时,只需参照全局模式的要求更新这组视图的定义即可,而无需更改系统的其他部分。但是,局部视图法的映射算法实现起来比较复杂,所以,全局模式上的查询转化成视图上的运算后,所得到的结果可能只是原始查询理论解的一个子集。

以 RDF/OWL 视图的形式定义好各数据源以后,系统就具备将语义查询(SPARQL 查询)转化为 SQL 查询的条件了。

1.4 SPARQL 查询向 SQL 查询的转化

在局部视图法中,两个关键问题要解决:一是将各数据源用一种本体描述语言视图化;二是在此基础上将实施于全局模式上的 SPARQL 查询转化为各数据源上的 SQL 查询。

SPARQL 查询转化为 SQL 查询具体说来有下面四步:

(1)构造语义 SPARQL 查询。用户在全局模式上以 SPARQL 查询的形式向系统发出请求。

(2)语义查询重写。参照视图定义,SPARQL 查询被转化成相应 SQL 子查询。

(3)执行 SQL 查询。在各数据源上执行相应的 SQL 子查询,得出各子查询的相应结果。

(4)返回查询结果。将各 SQL 查询所得到的结果返回,并转化为 RDF 元组。

其中,语义查询重写是最核心的任务。

1.5 基于语义的信息集成系统

至此,可以构建一个基于语义的信息集成系统模型。系统的特点在于以本体为基础,并以本体作为系统的全局模式。用户的操作实施在全局视图上,即对用户来说,真正的物理数据源是透明的,它只需要在全局视图上提出 SPARQL 语义查询,通过处理,SPARQL 查询被转化成子 SQL 查询集合,系统收集各 SQL 查询执行的结果,以统一的格式将完整的查询结果返回给用户。图 2 是系统的结构图。

在图 2 中,组件 Mediator 是中间件,Wrapper 是包装器。它们是数据集成系统中的关键组件,中间件负

责处理用户的查询请求,按照相应的视图定义,将用户的SPARQL查询进行转化分解,最终得到SQL子查询。然后,中间件将各SQL子查询交付到相应的数据源所对应的包装器上,包装器在数据源的基础上执行子查询并得到相应的查询结果。接着,包装器将其执行的结果传送至中间件,中间件将各子查询得到的结果进行整合,并最终转化为用户需要的查询结果。

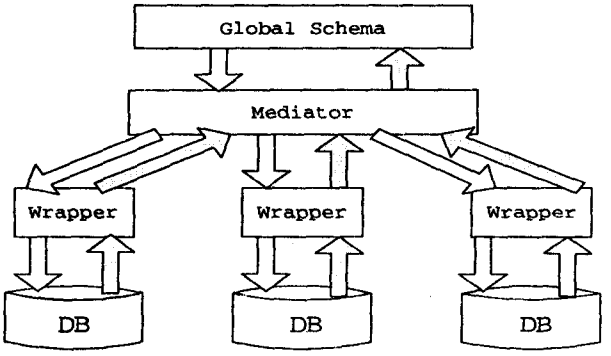

图2　基于语义的信息集成系统

2　分析与总结

"基于语义的信息集成系统"属于数据集成应用领域,它与传统的数据集成相比有着以下一些特点:

(1)全局模式构建在领域本体的基础上。

随着语义网的发展,各应用领域都在加强建立通用本体,从目前的发展来看,本体是保证领域资源最大程度共享的重要前提;另外,正是因为有了领域本体,才允许各数据源提供一种较为松散的集成机制,而这种松散的集成机制一般不会打破各个源系统之间的固有结构。也就是说,领域本体不仅促使了领域资源的最大化集成,也有效地保证了各数据源的相对独立性,这个特性将有助于解决"企业数据独立"与"企业信息共享"之间的矛盾。

另外,由于本体使数据具有了"语义(semantics)",因而使得知识推理变得可行,例如W3C推出的OWL-DL就是基于描述逻辑的,在这种本体表示语言上,系统可以做到很好的逻辑推理。

(2)基于视图的查询。

在商业化的今天,各个信息库都是企业赖以生存的资本,要构建一个访问各个物理数据源的领域集成系统是不现实的。在基于局部视图法的数据集成中,视图有利于数据源与全局模式之间保持较好的物理独立性[9]。这种视图可以是物化视图(materialized views),也可以是虚拟的,全局模式上转化过来的各个SQL子查询就是基于这样一组视图而演算的。另外,这些视图可以由各个数据源的管理员提供,他们可以根据领域本体的规范和要求,在视图中展现出可以共享的数据,一些敏感信息可以通过视图来屏蔽,这极大地保证了数据的安全和独立。

(3)SPARQL查询向SQL查询的转换。

模型中,SPARQL查询向SQL查询转换一般基于特定的视图定义规则。虽然SPARQL向SQL转换技术研究的时间并不长,但从文献[10]可以看出,这个问题已经得到突破性的进展,文献[11~14]等都不同程度地实现了SQARQL向SQL查询的转换。

总之,在传统的数据集成的基础上,"基于本体的数据集成"方案是近年来数据集成机制的新方向。在语义网的发展中,要充分利用已经广泛存储于关系数据库中的各种数据就必须要解决语义查询向传统关系模式查询的转化和重写,"基于语义的信息集成系统"正是建立在这种思想之上。通过分析可以看到,系统在理论与技术实现上都是可行的。以后的工作将主要放在RDF视图的构建及SPARQL向SQL查询转换的语义保持上。

参考文献:

[1] Gruber T R. A Translation Approach to Portable Ontology Specifications[J]. Knowledge Acquisition, 1993,5(2):199-220.

[2] 邓志鸿,唐世谓,张　铭,等. Ontology研究综述[J].北京大学学报:自然科学版,2002,38(5):730-738.

[3] 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.

[4] Maedche A. Ontology Learning for the Semantic Web[J]. IEEE Intelligent Systems, 2001,16(2):72-79.

[5] Noy N F, McGuinness D L. Ontology Development 101: A Guide to Creating Your First Ontology[R]. US: Standford Knowledge Systems Laboratory and Standford Medical Informatics, 2001.

[6] Wache H, Vgele t, Visser U, et al. Ontology-Based Integration of Information-A Survey of Existing Approaches[C]//Proceedings of the Workshop Ontologies and Information Sharing, IJCAI 2001. Seattle, Washington, U. S. A:[s. n.],2001:108-117.

[7] Levy A Y, Mendelzon A O, Sagiv Y, et al. Answering queries using views[C]//In Proc. of the 14th ACM SIGART Symp. on Principles of Database Systems (PODS'95). [s. l.]:SpringerLink, 1995:113-124.

[8] 陈跃国,王京春. 数据集成综述[J].计算机科学,2004,31(5):48-51.

[9] Calvanese D, De Giacomo G, Lenzerini M, et al. Description logic framework for information integration[C]//In Proc. of the 6th Int. Conf. on Principles of Knowledge Representation and Resoning(KR'98). Berlin:Springer,1998:2-13.

(下转第83页)

的开发插件协作生成,减少工作量。配置文件 hibernate.cfg.xml 代码如下所示:

```
<? xml version='1.0' encoding='UTF-8'? >
<! DOCTYPE hibernate-configuration PUBLIC
"-//Hibernate/Hibernate Configuration DTD 3.0//EN" "http://
hibernate.sourceforge.net/hibernate-configuration-3.0.dtd">
<hibernate-configuration>
<session-factory>
<property name="connection.username">system</property>
<property name="connection.url">
jdbc:oracle:thin:@huijx:1521:reso
</property>
<property name="dialect">
org.hibernate.dialect.Oracle9Dialect
</property>
<property name="myeclipse.connection.profile">Oracle</property>
<property name="connection.password">huijx</property>
<property name="connection.driver_class">
oracle.jdbc.driver.OracleDriver
</property>
<property name="show_sql">true</property>
<mapping resource="cn/edu/nuist/model/IdAssignment.hbm.xml"/>
<mapping
resource="cn/edu/nuist/model/TempResoUser.hbm.xml"/>
.....
</session-factory>
</hibernate-configuration>
```

Hibernate 为 Java 提供了 ORM 持久化机制和查询服务,Hibernate 将我们在 Java 类里使用的 HQL 语句转换成 SQL 语句,利用 JDBC 驱动进而操作数据源,完成数据的增加、删除、修改、查询等最底层的数据库操作。

4　结束语

文中提出和开发的 MLRLMS 系统在传统的语言资源库管理系统的基础上采用了 SSH 框架,探索了基于 SSH 框架给 Web 应用开发带来的巨大好处。研究建立汉、英、阿、俄对照词汇和短语资源库,为多语种软件开发提供规范的软件界面和用户文档用语,减少重复翻译工作,提高翻译质量,具有重大意义;而且能使多语种软件产品符合阿拉伯语和俄语的使用习惯。系统建设填补了多语种语言民族地区相关软件业领域建设的空白,对软件开发提供重要技术支持。

参考文献:

[1] 图格木勒.蒙古语语言资源库建设相关技术研究[D].呼和浩特:内蒙古大学,2007.

[2] 郭向勇,吴光斌.构建基于跨平台检索技术的校园网多媒体资源库[J].计算机工程,2002(5):252-254.

[3] 钟　珞,王　辉,李锐弢,等.基于语义 Web 的网络学习资源库本体实现[J].计算机工程,2007(8):282-284.

[4] 孙启勤,周　卫.计算机在线翻译快速入门[M].北京:中国水利水电出版社,2008.

[5] 李　刚.Struts2 权威指南——基于 WebWork 核心的 MVC 开发[M].北京:电子工业出版社,2007.

[6] 曹　国,姚建初,吴义忠,等.支持协同设计的资源库的开发研究[J].计算机工程,2003(16):182-184.

[7] 蔡雪燕.Hibernate 开发及整合应用大全[M].北京:清华大学出版社,2006.

[8] 张德育.蒙古语远程教育平台中蒙古文教育论坛和资源库系统的研究与实现[D].呼和浩特:内蒙古大学,2008.

[9] 曾　皓.多语种软件构件库的分类与检索[D].北京:中国科学院研究生院,2008:36-43.

[10] 徐　明,黄云森,陈可期.教学资源库建设策研究[J].中山大学学报:自然科学版,2002,41:114-117.

[11] 李　刚.整合 Struts + Hibernate + Spring 应用开发详解[M].北京:清华大学出版社,2007.

(上接第 77 页)

[10] Sahoo S S, Sis Center K E, Halb W, et al. A Survey of Current Approaches for Mapping of Relational Databases to RDF [DB/OL]. W3C RDB2RDF Incubator Group Reporter, http://esw.w3.org/topic/Rdb2RdfXG/StateOfTheArt, 2009,01.

[11] Chen Huajun, Wu Zhaohui, Mao Yuxin, et al. DartGrid: a Semantic Infrastructure for Bilding Database Grid Applications [J]. Concurrency and computation: Practive and Experience, 2006,18:1811-1828.

[12] Cullot N, Ghawi R, Yetongnon K. DB2OWL: A Tool for Automatic Dabase-to-Ontology Mapping[C]// Ceci M, Malerba D, Tanca L, et al. SEBD. DBLP. [s.l.]:[s.n.], 2007:491-494.

[13] Bizer C, Cyganiak R. D2RQ-Lessons Learned[DB/OL]. Position paper for the W3C Workshop on RDF Access to Relational Databases, 2007, 09. http://www.w3.org/2007/03/RdfRDB/papers/d2rq-positionpaper.

[14] Fisher M, Dean M. Automapper: Relational Database Semantic Translation using OWL and SWRL[C/OL]. Proc. of the IASK Int Conf. - E - Activity and Leading Technologies 2007(E-ALT2007), 2007, http://www.progeny.net/People/MattFisher/files/papers/AutomapperIASKEALT07.pdf.