第16卷 第8期
2006年8月

计算机技术与发展
COMPUTER TECHNOLOGY AND DEVELOPMENT

Vol.16 No.8
Aug. 2006

基于本体的XML语义集成研究

王志军,郭学俊
(河海大学 计算机及信息工程学院,江苏 南京 210098)

摘 要:针对具有相同语义的XML数据经常具有不同表达形式的问题,采用了基于本体的语义集成方法来集成异构XML数据源,即通过一系列映射规则将XPath的局部路径与本体关联起来,将用户的RDQL查询重写为基于XML的XQuery查询,从而达到语义集成的效果。本方法的意义在于用户可以通过本体查询异构的XML数据源。

关键词:本体;XML;语义集成
中图分类号:TP301.2;TP312 **文献标识码**:A **文章编号**:1673-629X(2006)08-0057-03

Research on Ontology-Based Semantic Integration of XML Sources

WANG Zhi-jun, GUO Xue-jun
(College of Computer & Information Engineering, Hohai University, Nanjing 210098, China)

Abstract: Some XML data that have the same meaning usually has different presenting style, so can integrate the distributed XML data sources by ontology-based semantic integrated way. It connects XPath and ontology through a series of mapping rules, rewrites users' RDQL query into XQuery. The value of this method is that users can find the data they need through ontology.

Key words: ontology; XML; semantic integration

0 引 言

XML已成为数据交换的标准。相对关系数据,其表达能力更强,几乎可以表达任何数据,从简单的整型到复杂的对象类型。然而XML不能表示语义,只能表示语法,一些具有相同语义的数据经常具有不同的表达形式。因此,在XML的基础上进行语义集成的探索成了一个十分重要的研究课题,并产生了一些研究成果。

直接转换方法是由Klein提出的[1]。此方法把XML数据直接转换为RDF数据,用外部的RDFS说明来注解XML文档。但此方法并没有考虑XML的文件结构,不能把查询从一个数据源传到另一个数据源;Yin/Yang网是由Patel Schneider等人开发的[2],此方法解决了XML和RDF不能互操作的问题,通过把RDF的语义和推理规则集成为XML,为XML和RDF提出了一种集成模型,这样XML的查询可以从RDF的推理中获得,但Yin/Yang网并没有解决XML数据源之间的查询问题。

综合上述方法的优势与不足,文中设计了一种基于本体的XML语义集成方法。即给定两个XML文件,创建其XML Schema,根据XML Schema中的层次信息,将其转化为RDF Schema,同时生成RDF和XML之间的匹配表;通过集成局部本体生成为全局本体,集成过程中建立局部本体和全局本体之间的匹配表;最后,根据匹配表将RDQL的查询转换为Xpath的查询,从而达到语义集成的效果。

收稿日期:2005-11-24

作者简介:王志军(1981-),男,江苏泰州人,硕士研究生,研究方向为本体、语义集成;郭学俊,副教授,研究方向为计算机网络与分布式处理。

1 语义集成方案设计

在集成过程中,局部本体对信息源进行描述,并以全局本体作为参照,建立了局部本体和全局本体之间的映射,因而能够表达复杂的语义。全局本体为用户提供了统一的查询接口,此接口可以依据全局本体中的语义映射关系、局部本体与全局本体之间的匹配表以及XML与局部本体之间的匹配表,把用户提交的RDQL查询重写为对各个XML数据源的查询,经过对查询结果的整合,以统一的格式返回给用户,集成方案如图1所示。

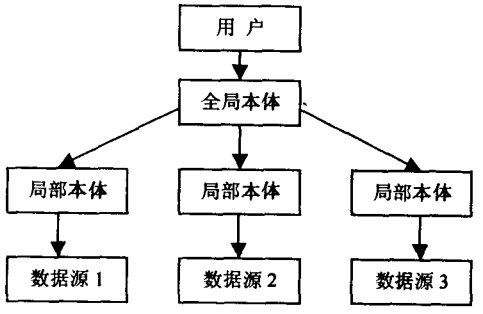

图1 基于本体的语义集成模型

1.1 构建局部本体

构建局部本体就是在保留XML结构的前提下,把XML Schema转换为局部本体。元素和属性是XML

Schema 的两个基本组成部分[3]，元素可以定义为不包含任何内容和属性的简单类型，也可以定义为具有元素内容和属性的复杂类型，属性只能定义为简单类型。

RDF 使用 XML 的语法[4]，以一种人类可理解、机器可读的方式来处理和交换数据。RDF Schema (RDFS)定义了 RDF 的类和属性，这些类和属性可在 RDF 描述中被实例化。

考虑到 XML 的元素、属性及其关系，从 XML 到 RDF 的转换将包含元素层的转换和结构层的转换，利用 Mike Klein 转换算法即可实现此过程。元素层的转换定义了局部本体的基本类和属性，但是元素间的结构关系并没有考虑在内；结构层的转换是从 XML Schema 的层结构向局部本体的转换，把元素－属性关系当作 RDFS 中的类－文本关系进行转换，把元素－子元素关系当作类－类的关系进行转换。

1.2 构建全局本体

由于局部本体已构建完毕，可通过本体复用的方法来构建全局本体。把多个局部本体当作输入，把一个集成好的本体当作输出[5]，本体集成操作包括：

(1) 类的集成：把多个概念相同的类集成为一个类。

(2) 属性的集成：把类中多个概念相同的属性集成为一个属性。

(3) 类之间的关系集成：把概念上相同的关系从一个类集成到另一个类。

(4) 复制一个类或属性：如果相同的类或属性没有在目标本体中出现，则直接复制这个类或属性。

1.3 匹配表

在局部本体的构建阶段和全局本体的构建阶段，XML 与 RDF 之间的匹配表、全局本体与局部本体之间的匹配表分别记录了 XML 到 RDF、局部本体到全局本体之间的匹配信息。这些匹配表建立了从底层的数据源到局部本体最终到全局本体之间的联系，并构成了查询转换的基础。通过如图 1 所示的语义集成模型，作用于全局本体的 RDQL 查询在全局本体到局部本体的匹配表的指示下可转换为对各局部本体的查询。各局部本体的查询在 XML 与 RDF 间的匹配表的指引下，可转换为对各 XML 数据源的 Xquery 查询。

在全局本体产生过程中建立的匹配表是一个三元组，如果一个类、属性或全局本体中的类 P 之间的关系是由集成不同的局部本体 P_i 和 P_j 而来，则产生的匹配信息形式为 (P, P_i, P_j)；如果全局本体中的类或属性 P 是由局部本体 P_i 复制而来，则产生的匹配信息为 (P, P_i)；在局部本体产生过程中建立的匹配表是一个简单的二元组，记录了 XML 与 RDF 之间的匹配信息。

1.4 查询重写

全局本体连接所有的局部本体，通过向全局本体提交查询，用户可以从数据源中检索出所需数据。此节探讨了从全局本体的 RDQL 查询到 XML 的 XQuery 查询之间的转换。用 M 表示全局本体和局部本体之间的匹配表，Q_g 表示全局本体的查询，Q_r 表示局部本体的查询，Q_x 表示 XML 的查询，则查询重写过程如下：

1) 对在 Q_g 中使用的变量，找到其对应的 RDF 路径表达式，并放入集合 P。根据不同情况，对这些变量进行分类：

(1) 把在 Select 子句中的 RDF 表达式加入集合 P_s。

(2) 把在 Where 子句和受到文本变量或 URI 限制的 RDF 表达式加入集合 P_w。

(3) 把出现在 And 子句中的 RDF 表达式加入集合 P_w。

2) 对参与集成的局部本体 R_i，根据匹配表 M 中的对应信息，利用 R_i 中匹配的 RDF 路径来代替 P 中的 RDF 路径，更新 P_s 和 P_w 也使用此方法。根据下述方法，把 Q_g 重新写入 Q_r：

(1) 对于 Where 子句，遍历 R_i 以发现一个非循环子图，此子图包含了 P 中所有路径表达式，并把所有的边放到集合 E 中去，对每一个 $e_i \in E$，把三层结构 $(?S_i, e_i, ?O_i)$ 加入到 Where 子句中。

(2) 对每一个 Select 和 And 子句，利用它们在 R_i 中相应的路径来代替 RDF 路径，并为每一条路径赋予一个变量。

3) 利用 XML 和局部本体 R_i 间的匹配信息，对 P_s 和 P_w 中的每一个元素 P_i，找到其 Xpath 表达式。用 P_s' 代替 Xpath 集合，用 P_w' 代替 P_w。

4) 为每一个 XML 数据源构建目标查询 Q_x：

(1)let 子句。根据以下格式输出一个 let 子句：let $ < root label > : = doc("< XML source name >")。

(2) For 子句。对于 P_w' 和 P_s' 中的每一个 Xpath P_i，根据以下格式输出一个 for 子句：for $ < node label > in < p_i >。

(3) Where 子句。对于 P_w' 中的每一个 P_i'，根据 Q_r 中的限制构建一个查询，并把这些查询的连接当作一个 where 子句。

(4) Return 子句。把集合 P_s' 中的每一个 P_i' 当作 return 子句中的一个元素。

2 试验与分析

为了验证方案的可行性，文中对语义集成的具体实现做了一点探索，实验的预期目标是能通过本体从数据库中检索出所需数据，过程如下：

(1)在两个 Oracle 数据库中分别存放 20 条关于电子商务的 XML 数据，图 2 定义了这两个数据库中 XML 数据的模式 S_1，虽然这两个模式属于同一领域，但在结构和术语上却截然不同。

(2)根据元素层的转换和结构层的转换，把图 2 中的两个 XML 模式转换为基于 RDF 的局部本体，局部本体的 ER 模型如图 3 所示(以 S_1 为例)，XML 与 RDF 之间的匹

配表见表 1(以 S_1 与 S_1' 之间的匹配表为例)。

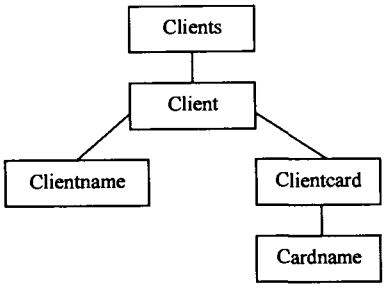

图 2　XML 的数据模式

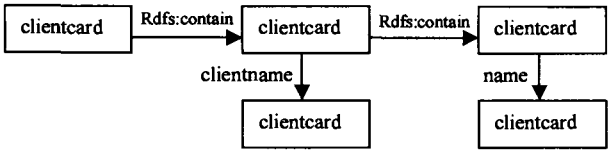

图 3　基于 RDF 的局部本体

表 1　XML 与 RDF 之间的匹配表

S_1 中的 Xpath 表示	S_1'中的 RDF 表示
/clients	clients
/clients/client	client
/clients/client/clientname	client. clientname
/clients/client/clientcard	clientcard
/clients/client/clientcard/cardname	clientcard. cardname

(3)Protege－2000 是一种半自动化的本体集成工具,在用户指定参与集成的本体后,系统会自动找出本体间的冲突,并提供可供选择的解决方案。在用户选择相应方案后,系统会继续寻找本体间的其余冲突,反复循环,直至集成过程的结束。利用此工具把图 3 中的两个局部本体集成为全局本体。

(4)用户根据界面提示输入"查询所有客户所持卡的卡号"的请求,根据上文中的查询转换算法对用户的查询请求进行重写,并向数据库提交查询请求。实验结果表明,可从数据库中检索出 28 条相应数据。虽然集成效率仍有待提高,但也说明此方案是可行的。

3　结束语

文中针对具有相同语义的 XML 数据经常具有不同表达形式的问题,通过在语义集成中融入本体的思想,构建了基于本体的语义集成方案,使用户不需了解 XML 数据的结构和模式,就可以实现查询。方案着重对局部本体、全局本体、匹配表、查询转换等几部分进行了设计和探索,并通过一个实验证明此方法是可行性的。方案采用 RDF 作为本体描述语言,然而 RDF 的表达能力毕竟有限,在 OWL 趋于成熟时应过渡到 OWL。

参考文献:

[1] HP Labs. RDQL－RDF Query Language[J]. The art of Semantic Web,2001(5):64－69.

[2] Schneider P. The Yin/Yang Web:XML Syntax and RDF Semantics[A]. AAAI/IAAI－2000[C]. 北京:清华大学出版社,2000. 146－153.

[3] 李丽萍,马文阁,梁　勇. XML 深入剖析[J]. 辽宁工程技术大学学报,2002(4):41－45.

[4] Brickley D. RDF Schema Specification[EB/OL]. http://www.w3.org/TR/PR2rdf2schema, 2004.

[5] Stumme G. Ontology Merging for Federated Ontologies on the Semantic Web[A]. FMII－2001[C]. 北京:机械工业出版社,2001. 413－418.

(上接第 56 页)

适合需要高度保密的无线局域网应用环境,可以采用基于 OCB(Offset CodeBook,分支编码本)模式的 AES(Advanced Encryption Standard,高级加密标准)的保密机制。AES 加密算法是美国的标准加密算法,其抗攻击型已经得到验证和检验。

一个加密系统的核心是密钥管理,而 WEP 协议的一个主要问题就是不存在密钥管理机制,使得系统的安全性得不到保证。所以,采用高效、合理的密钥管理机制,是解决问题的根本方法。

传统 WEP 协议规定的是单向的身份认证,其身份认证机制存在种种问题,解决问题的关键是完善认证机制,建立双向的、性能良好的身份认证机制。

4　结束语

无线局域网不需要有线连接就能收、发数据,人们能够自由地把计算机设备放置在最合适的地方。并且无线局域网很大程度上的灵活性,使人们能够及时对有变动的要求做出应对。虽然无线局域网拥有诸多优势,但同样面临着一些阻碍其发展的问题,而安全性就是最主要的问题之一。作为一个不断改善和升级的过程,只有采取一套严密的安全方案以确保无线局域网的安全,才能让无线局域网得到更大范围的发展。

参考文献:

[1] 钱　进. 无线局域网技术与应用[M]. 北京:电子工业出版社,2004.

[2] Haslestad T, Telenor R D. Wireless Local Area Network IEEE 802.11[EB/OL]. http://www.tele.ntnu.no,2004.

[3] Jani K. Wireless Local Area Network Security－Obscurity Trough Security[EB/OL]. http://www.ee.oulu.fi,2004.

[4] Madge Limited. Wireless LAN Security White Paper[EB/OL]. http://www.madge.com,2003.

[5] 钟晓珊,刘　旭. 无线局域网接入的安全性问题[J]. 信息技术,2004,12(28):10－13.