维普资讯 http://www.cqvip.com

RDF数据存储技术研究*)

姚文琳　刘世栋　张子振
(中国海洋大学　青岛 266071)

摘　要　资源描述框架RDF是一种元数据描述方法,它提供了Web上智能信息服务和语义互操作性的标准,是下一代Web—语义Web应用的基础。大量RDF数据的操作和处理需要数据库的支持,而关系数据库存储RDF数据的方法可以有效处理RDF数据。

关键词　RDF,三元组,关系数据库

1　引言

Tim Berners-Lee提出的语义Web(第二代Web)是面向机器处理Web信息,语义Web的目标是提供Web智能信息服务。而资源描述框架RDF是语义Web的基础,它为Web上的资源描述提供了一种数据模型,是Web上智能信息服务和语义互操作性的标准。充分利用RDF可以有效地处理Web信息,而利用RDF需要解决RDF数据的存储问题。目前,RDF数据的存储形式基本上有3种:XML/RDF文件形式、专门的XML/RDF数据库和传统的关系数据库。对于少量数据,以XML/RDF文件或XML/RDF数据库的形式存储是一种可行的方式,但是对于大量数据,考虑到可扩展性、数据完整性、安全性和查询效率等诸多因素,以关系数据库来存储是较好的选择。

2　RDF数据模型

2.1　RDF简介

元数据即关于数据的数据(Data about data),是相对于对象数据的一种概括性、实质性的描述。而RDF(Resource Description Framework)[1]是W3C推荐的用于描述和处理元数据的一个草案,能为Web上的应用程序之间的交互提供机器可理解的信息,它独立于任何语言,适用于任何领域,是处理元数据的基础,是一个开放的元数据框架,提供了(用于交换Web信息的)语义Web应用程序之间的互操作能力[2,3]。

2.2　RDF语法

RDF定义了一种描述机器可理解的数据语义的数据模型,该数据模型主要包含下面的三个对象类型:

(1)资源(Resource):资源可能是整个网页,网页的一部分或页面的全部集合;或者是不能通过Web直接访问的对象。

(2)特性(Property):特性是描述某个资源特定的方面、特征、属性或关系。

(3)声明(Statement):一个特定的资源和特性名称加上该特性的值一起构成了一个RDF声明。

举个例子,一个声明"张三是这所大学的老师",大学是被描述的资源,它的一个属性是"老师",而这个属性的值是"张三"。下面是用实体—关系图来表示的这个声明(见图1)。

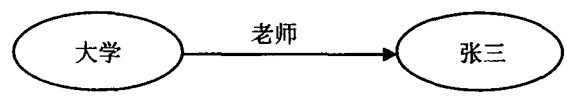

图1　实体—关系图表示声明

3　水平存储和垂直存储

对RDF数据的存储采用了关系型数据库(Oracle 9i),不仅因为其技术的成熟,能够存储大规模的数据,对于数据的存储和查询快速的响应,更因为其在承受能力和稳定性上有很好的表现。

SergeyMelnik[4]等人讨论了如何在关系数据库中存储RDF数据,通过将RDF模型与关系数据库模型比较可以发现,两者非常相似,可以进行模型的一一映射:RDF的节点对应关系数据库的记录;RDF的属性类型对应关系数据库的字段名;RDF的属性值对应关系数据库的字段单元。以上分析表明,RDF数据可以很方便地映射成关系数据库的水平行表示(见表1)。水平行表示的存储方式适合于

*)基金项目:国家自然科学基金项目(60403012),山东省自然科学基金项目(Y2005G06)。姚文琳　副教授,研究方向为人工智能,包括大规模信息的获取、语义Web、信息分类、过滤和存储等;刘世栋　硕士生,主要研究方向为自然语言处理、本体存储、语义Web;张子振　硕士生,主要研究方向为本体论,语义Web。

维普资讯 http://www.cqvip.com

存储稠密型（一个节点有大量的属性填有属性值）的数据，进行查询操作会比较方便，但是存在如下的问题：

(1)字段数目太多，目前的数据库系统对一张表拥有的字段数目都加以限制；

(2)表中会出现很多的空值，这是由于表中的字段是所有数据对象的，因此对某一特定数据对象来讲，它拥有的属性数目很少的话，那就会有很多字段对应的值都是空值；

(3)进化不方便，当需要频繁的更新数据时，水平存储操作起来是很不方便的；

(4)检索性能不高，如果表中的字段太多，而用于检索的字段却很少，会大大地影响检索效率。

在存储 RDF 数据时，数据很可能是稀疏型的（一个节点通常只有少量的属性填有属性值），同时也需要进化（常会对一个节点添加新属性，从而增加了数据库表的属性列），而垂直三元表示[4]（见表 2）正好适合于这种应用，因此在数据库表结构的设计中，经常舍弃水平行表示，而使用垂直三元表示。对于 RDF 数据而言，垂直三元表示恰恰就是 RDF 的三元组表示（Oid--object identifier, Key--attribute name, Val--attribute value）。

表 1 水平行表示

Oid	H1	H2	H3
1	a	b	null
2	null	c	d
3	null	null	a
4	b	null	d

表 2 垂直三元表示

Oid	Key	Val
1	H1	a
1	H2	b
2	H2	c
2	H3	d
3	H3	a
4	H1	b
4	H3	d

4 在关系数据库中存储 RDF 数据

关系数据库是目前数据库应用的主流，用它存储 RDF 数据，可有效利用现有数据库资源。要将 RDF 数据存储于关系数据库中，必须把 RDF 的数据模型转化为关系模型，实现从 RDF 模型到数据库模式的映射。我们在下面将给出存储 RDF 数据的数据库（Oracle 9i）模式，此数据库模式包含五个表，我们还将给出一个视图生成实例。

4.1 数据库模式中的五个表

①资源表（the resource table）：

sql ="CREATE TABLE RDFRESOURCE"+ "("+ "Id INTEGER not null primary key,"+"NS INTEGER not null," + " RoName VARCHAR (255)"+ ")";

该表用来存储 RDF 资源，表中 Id 是资源的内部标识和该表的主键，NS 是指向名空间表中相应资源名空间的指针。

②名空间表（the namespace table）：

sql ="CREATE TABLE RDFNameSpace"+ "("+ "Id INTEGER not null primary key," +" NsName VARCHAR(255)"+ ")";

③RDF 中的文字表（the literal table）：

sql = "CREATE TABLE RDFLiteral"+ "("+ "Id INTEGER not null primary key," + "VAL VARCHAR(4000)"+ ")";

该表用来存储字符串。

④Statement 表（the statement table）：

RDF 模型声明，它是对一个事实的基本描述，也是 RDF 模型的最小有效数据单元，一个声明分为：主体（subject）、谓词（predicate）和对象（object）三个部分。从本质上说，RDF 定义 Object-Attribute-Value 三元组作为基本建模原语并为它们引入了标准的语法，所以存储声明的表是数据库模式中最重要的部分，其结构如表 3 所示。

表 3 用来存储三元组的 RDFStatement 表

Column name	Type	Comment
Id	INTEGER	not null
Subject	INTEGER	not null
Predicate	INTEGER	not null
ObjResource	INTEGER	not null
ObjLiteral	INTEGER	not null
Res	CHAR(1)	flag whether "object" is in literal or resource table

sql="CREATE TABLE RDFStatement"+"("+ "Id INTEGER not null primary key,"+"Subject INTEGER not null," +" Predicate INTEGER not null,"+"ObjResource INTEGER not null,"+"ObjLiteral INTEGER not null,"+"Res CHAR(1)not null"+")";

在三元组中，主体和谓词都是资源，而对象可以是资源也可以是字符串，所以需要附加一个标识位（flag）—Res，标识该对象是资源还是字符串。Subject 对应 RDF 三元组中的 Object，Predicate 对应三元组中的 Attribute，ObjResource 和 ObjLiteral 对应三元组中的 Value。

⑤RDF 模型表（the model table）：

sql="CREATE TABLE RDFModel"+"("+" ModelId INTEGER not null,"+"Statement INTEGER not null."+"Asserted CHAR(1)not null,"+" Reified CHAR(1)not null,"+"primary key(Mode-

维普资讯 http://www.cqvip.com

lld,Statement)"+")";

该表用来存储各个 RDF 模型。

4.2 视图生成实例:产生视图包含数据库中所有的 statement

sql="CREATE OR REPLACE VIEW RootModel" +" AS SELECT UNIQUE Id, Subject, Predicate, ObjResource, ObjLiteral, Res, Asserted, Reified"+"FROM RDFStatement,RDFModel" +"WHERE RDFModel。Statement = RDFStatement. Id";

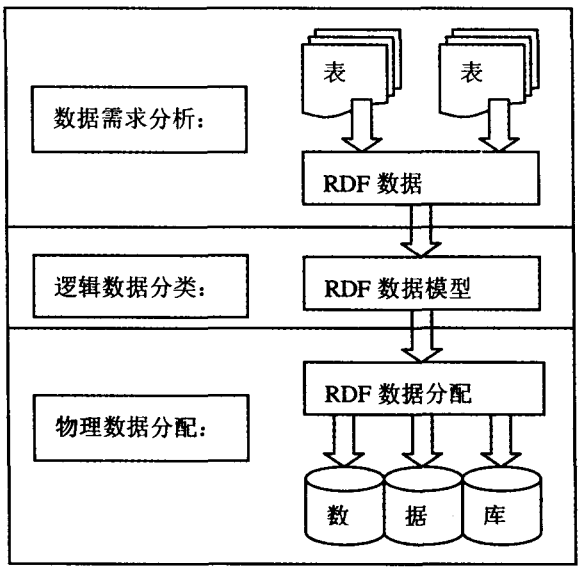

图 2 数据库设计过程

5 关系数据库的设计过程

在设计数据库时,可以从数据库表中提取和日常操作相关的字段,根据字段特点对表中的字段进行分类,在分类的过程中将具有相似特点的字段自下而上分组构造抽象类。通过字段的分类和抽象存储 RDF 数据,抽象类就是数据库中的表,基类就是表的属性。数据库处理设计分为三个阶段:数据需求分析、逻辑数据分类和物理数据分配。在数据需求分析阶段通过应用需求分析存储了 RDF 数据,逻辑数据分类阶段进一步将结构和 RDF 数据集成构造了 RDF 数据模型,物理数据分配阶段获得了满足应用需求的最小数据处理成本,并将数据分配合适。设计过程见图 2。

结束语 随着 RDF 在 Web 上的应用,对 RDF 的操纵需要对 RDF 数据进行存储,本文在介绍 RDF 的基本概念和语法后比较了 RDF 数据存储时水平行表示与垂直三元表示的优缺点,给出了存储 RDF 数据的数据库的表和视图,研究了数据库的设计过程。分析表明关系数据库是一种有效存储 RDF 数据的方法,我们将在以后对 RDF 存储技术作进一步探讨以推进 RDF 在 Web 上应用的发展。

参 考 文 献

1 Lassila, Swick R R. Resource description framework (RDF) model and syntax specification. W3C. http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/,1999-02-22

2 Lassila, Swick R R. Resource Description Framework (RDF) schema specification1.0. [EB/OL]. http://www.w3.org/TR/1999/REC-rdf-syntax

3 Brickley D,Guha R V. Resource Description Framework(RDF) schema specification1.0. [EB/OL]. http://www.w3.org/TR/2000/CR-rdf-schema-2000-03-27

4 Jensen C J, Jeffery K G, Pokorny J. Storage and querying of E-commerce data[A]. Advances in Database Technology-Edbt, 2002.409

(上接第 85 页)

Mname},且 R_3 的码为 Sdept。从而得到 $R_4(U_4, F_4)$,其中 U_4={Sno, Sname, Sdept},F_4={Sno→Sname,Sno→Sdept}。

(3)取 $R_2(U_2,F_2)$,由于 $Cno_{F_2}^{+}$={Cno,Cname}≠U_2,得 $R_5(U_5,F_5)$,其中 U_5={Cno,Cname},F_5={Cno→Cname},且 R_5 的码为 Cno。从而得到 $R_6(U_6, F_6)$,其中 U_6={Sno, Cno, Grade},F_6={(Sno,Cno)→Grade}

(4)由于 R_3,R_4,R_5,R_6 中都不存在函数依赖不包含码,因此它们都属于 BCNF,也即为 R 满足 BCNF 的一个分解。

结束语 模式分解是减小关系数据库数据冗余,排除操作异常的有效工具,同时也是关系数据库数据模型设计的难点。针对该问题,本文基于分层递阶的思想,提出了一种基于 BCNF 的数据模型的层次分解算法。通过实例分析,该算法简化了模式分解的步骤,并解决了文[2]提出的算法不能处理属性之间存在多级函数依赖的情况,为模式分解提供了新的方法。

参 考 文 献

1 萨师煊,王珊. 数据库系统概论(第三版)[M]. 北京:高等教育出版社,2003

2 马雪英,冯睿. 基于函数依赖的模式分解方法[J]. 计算机应用与软件,2004,21(4):31~33

3 Ullman J D, Widom J, Widom J D. A First Course in Database Systems(second edition)[M]. 2001

4 覃遵跃,徐洪智,冯峻松,蔡国民. 利用函数依赖图寻找关系模式的候选码[J]. 安庆师范学院学报(自然科学版),2004,10(1):3~5

5 张永,顾国庆. 关系模式中候选码的求解[J]. 上海电力学院学报,2002,18(1):33~35