DOI:10.16208/j.issn1000-7024.2011.08.037

OWL 本体存储技术研究

常万军[1,2]， 任广伟[1]

(1. 贵州大学 计算机科学与信息学院，贵州 贵阳 550255；
2. 河南机电高等专科学校 计算机科学与技术系，河南 新乡 453002)

摘 要：对本体的存储介质研究的基础上，深入研究本体在关系数据库中的存储模式，指出当前各种本体存储模式的不足之处，提出了新的基于关系数据库的混合存储模式。用教育领域本体测试框架对所提出的存储模式进行实验验证。实验结果表明，设计的基于关系数据库的存储模式具有结构清楚、查询效率高和扩展性能强的优点。用新的混合存储模式要比目前广泛使用的垂直模式在各方面的性能都更优越，而且也适宜于存储大规模本体。

关键词：网络本体语言；资源描述框架；本体；存储；关系数据库

中图法分类号：TP311 **文献标识码**：A **文章编号**：1000-7024 (2011) 08-2893-04

Study on storage technique of OWL ontology

CHANG Wan-jun[1,2], REN Guang-wei[1]

(1. College of Computer Science and Information, Guizhou University, Guiyang 550255, China; 2. Department of Computer Science and Technology, Henan Mechanical and Electrical Engineering College, Xinxiang 453002, China)

Abstract: According to analyzing OWL ontology storage stdium and storage schema of relational database, the shortage of various ontology storage mode is pointed out and a new ontology storage schema based on relational database is given. Finally, the ontology test frame of educate realm is used to test the storage schema. The experiment shows that the new ontology storage schema of relational database in this text has a higher searching efficiency and better expand function. It has more superior function in each aspect than the perpendicular storage schema used extensively now. It is also feat in saving large-scale ontology.

Key words: OWL; RDF; ontology; storage; relation database

0 引 言

语义 Web(semantic web)作为当前万维网的扩展，通过结构化和形式化的方法，以表示 Web 上的资源，使得计算机程序能够对网络资源进行分析和推理[1-2]。语义网作为下一代万维网，目前成为全球领域的研究热点。尤其随着 W3C(world wide web consortium) 的一些语义网相关标准，如资源描述框架(resource description framework，RDF)、RDFS、网络本体语言(web ontology language，OWL)等的制定，越来越多的人利用这些标准和技术，去开发基于 Semantic Web 的应用系统。

在语义网体系的各类技术中，本体是最重要的支撑技术[3]，语义网的发展加速了本体的研究。随着最新本体语言 OWL 的出现，越来越多的人利用 OWL 语言开发基于本体的知识系统，使用本体语言描述某个领域的知识，利用本体工具开发特定领域的知识系统[4]。在这些知识系统中，本体提供了知识库构建的基本结构，是整个系统的骨架和核心。OWL 本体作为一种全新的数据组织形式，选择合适的存储介质和合理的存储模式对其进行有效存储，是当前人们十分关注的问题，也是开发语义本体知识系统的各类技术中最为关键的基础性支撑技术[5-6]。

1 本体存储介质

在本体的存储技术中，存储介质的选择是研究人员首先需要关注的。目前已知的本体存储介质，也是人们经常选择的存储介质有内存、文件、关系数据库等 3 种方式。

1.1 内存存储方法

用内存法存储本体是将本体数据以一定的结构形式直接存储在计算机的主存中，然后在计算机主存中进行数据查询等各种数据操作。这种方法具有较高的运行效率，但囿于物理条件的限制，内存存储方法只能存储很少量的数据，而且记忆能力很差。

1.2 文件存储方法

文件存储的方法简单可行，适宜常久存储，很多本体相关工具都支持用文件格式存储的本体进行存取。但是这种方法

收稿日期：2010-08-23；修订日期：2010-12-21。

作者简介：常万军 (1978－)，男，河南南阳人，硕士研究生，讲师，研究方向为数据库技术； 任广伟 (1962－)，男，贵州贵阳人，副教授，硕士生导师，研究方向为多媒体教育技术、Linux 应用技术和 ORICL 数据库开发技术。E-mail：cswj1979@sina.com

效率很低,而且存储规模有限。当OWL本体数据文件很大时,要把握OWL本体数据全局的结构,必须反复扫描OWL本体文件,进行大量数据存取工作,严重影响了系统的效率。有时为了保证本体系统的并发性,还需建立一定的并发控制机制[7]。

早期的一些本体工具是基于文件系统实现的,这些工具主要用来编辑和建立本体,并不是为大规模本体数据的存储和查询管理服务的,例如Onto,Protégé[5]。

1.3 关系数据库存储方法

关系数据库存储方法是将本体按照一定的规则存储在关系数据库中,利用关系数据库系统对数据的操纵和管理能力来存取本体。相对于内存存储和文件存储先天不足,关系数据库则体现了非常强大的优越性。关系数据库技术比较成熟,应用广泛,存储效率很高,它的自身的事务管理系统非常完善,可以确保操作的正确性,而且可以避免重复开发[8-9]。

2 关系数据库的存储模式

使用关系数据库存储本体,首先需要考虑的是存储模式的设计问题。如何设计关系数据库表的结构,如何用关系数据库与本体知识建立一一对应关系?本体在关系数据库来实际存储方案是什么呢?我们知道,关系数据库的存储模式有4种:水平模式,垂直模式,分解模式和混合模式。本体与关系数据库相结合,大概有以下几种存储模式:

2.1 水平模式

在数据库里把OWL本体的所有类存储在一张通用表中,表的列为类名、类的类型以及所有的属性的名字,数据表中的各条记录对应本体的实例或类。表的概要反映的是OWL本体中所有类的层次。以数据库的水平模式的来存储本体,一个本体对应一个数据表,表的数目很少,建库非常方便。但是水平方式也有很多先天的不足:①数据表允许的列的数目可能不能满足当本体中类的属性非常多的情况;②对于本体中类的多值属性的存储意愿也会因表的二维结构限制而不能满足;③本体中有许多属性为空的类的存储也会造成数据库资源的大大浪费[10]。

2.2 垂直模式

把本体实例的所有信息都用一个RDF三元组(Subject、Object和Predicate)来进行表示,然后将这些三元组存储在数据库的一张三元组数据表中。这种存储模式的实质是将本体中复杂的数据关系分解为多个简单的RDF二元关系来进行表达和存储。这种模式简单易实现,在数据表中添加删除记录对表的结构不会产生影响,比较适合本体中实例的多变性[11]。以垂直模式存储本体数据,对于开发人员来说,可读性非常差,设计准确无误的SQL语句非常困难。更为不利的是,如果以这种模式存储本体数据,当对单个本体进行查询,不仅需要搜索整个数据库,而且还要做很多连接操作,查询效率非常低。

2.3 分解模式

依据数据库的类或属性的不同,可以将水平模式或垂直模式的一个数据表分解成若干张数据表,这就是属性分解和类的分解两种分解模式。类的分解是将每个类分别存储在一个单独的数据表中,表中的字段一一对应类各个属性;性质分解是相同性质的数据归并统一存储同一个数据表中,数据表只有两列数据,即RDF三元组中的主体和谓词。

2.3.1 基于类的分解模式

类的分解模式与水平模式相似,但它的粒度比水平模式的粒度要小许多。这种模式在对一个或一组实例的属性进行查询时效率很高。这种模式的不足之处是,其数据表的结构必须随着本体中类或者属性的动态变化而变化,具有不稳定性。当然,水平模式中大量空域现象问题在基于类分解模式中也是经常发生的。

2.3.2 基于属性的分解模式

按照性质的不同将数据表分解成若干个只包含两列数据的小型数据表,在执行简单查询时的响应时间非常快。然而,当遇到涉及很多属性的复杂查询时,数据库要执行很多join operation操作,查询的效率不会太高。与类的分解模式一样,随着本体的动态变化要不断的创建和删除数据表,这则是数据库管理中效率最低、代价最大的数据操作方法。

2.4 混合模式

关系数据库的最基本的存储模式在解决前面向种本体存储方面都有各自的不足,为此有研究人员尝试将数据库的3种基本模式混合使用,以达到本体存储的实际要求。

目前已有的DLDB系统就是混合模式的一个应用典范。它提出将基于类的分解模式与基于属性的分解模式相结合的存储模式。在本体中定义一个类就要用基于类的分解模式的相关方法为该类创建一个表,在本体中定义一个属性就要用基于属性的分解模式的相关方法为该属性创建一个表[12]。这种模式包含一百个左右类的本体,运行得很好。但是,这种模式在本体的类比较多的情况下,会出现数据库无法容纳的多表或者查询效率很低。

OWL本体在关系数据库中存储的5个具体模式,如果从结构可读性、结构稳定性和查询速度3个方面来进行比较,可以到得表1的二维表格[13]。

表1 OWL本体存储模式对比

OWL存储模式	结构可读	结构稳定	查询速度快
水平模式	满足	不满足	满足
垂直模式	不满足	满足	不满足
类的分解模式	满足	不满足	满足
属性分解模式	满足	不满足	不确定
混合模式	不确定	不满足	不确定

3 新混合模式的设计与实现

文献[13]指出,可以寻找一种很好的规则和算法,使OWL本体与关系数据库的模式之间建立良好的对应关系。文献[13]两个思想可以为我们借鉴:①用关系数据库来存储本体是可行的;②设计合理混合模式的存储方案,完全可以达到结构稳定、查询迅速和高可读性的最佳存储效果。

3.1 设计思路

混合模式的存储方案,需要建立OWL本体和关系数据库之间良好的对应关系[14]。他们之间元素的对应办法为:用关系数据库中的表和本体中的类相对应,用关系数据库中表的列

与本体中类的属性相对应，对于类的多种数据类型属性可以通过数据库表中相关列的唯一性和主键设置来对应，本体类之间子类-超类的关系可以用数据表中主、外键的设置来对应。他们之间数据类型的对应关系也可以用类似 xsd: integer 与数据库中 INT/INTEGER 对应关系一样一一对应起来。

3.2 设计方案

本体的最基本的概念就是资源。OWL 使用 URI (uniform resource identifier) 来标识资源，这与 RDF 模型的三元组 (Subject、Object 和 Predicate) 完成可以建立相对应的关系[15]。如果文字可以作为 Predicate，而且文字可以作为是一种特殊的资源，那么就允许我们把文字和本体中的URI一起当作资源，存储在一张数据表中，这样 Predicate 为相同文字的各个三元组就能共享相同的 Web 本体资源。

三元组也是本体的一个基本概念，用一个数据表来表示 RDF 基本模型的。从理论上讲，有这两个数据表就可以完全存储本体信息。但为使本体存储的结构更加清晰，并提高数据查询的效率和结构的稳定性，本文专门建立了一些数据表来存放一些常用信息。在数据表的建立过程中有一个原则须共同遵守，就是本体的类作为数据表存储时，一般都要把他们属性名对应的列设置为数据表的主键。如果数据表的主键已经存在，则将须设为主键的列设置为唯一列。采用如下措施把本体实例存储到关系数据库的数据表中：

(1) 建立数据表分别将 subClassOf，subPropertyOf，domain 和 range 等函数属性的定义域存储起来。

(2)本体中，经常出现父子关系的类，因此需要将这种隶属关系的实例存储在一张数据表中。假若若 OWL 本体中类 c 和类d存在subClassOf(c,d)关系，他们对应的数据表分别为tablec 和 tabled，则在将在 tablec 中添加一个与 tabled 的主键同名的列。用此列作为 tablec 的主键和引用数据表 tabled 的外键。

(3)本体存储中，类与类之间出现互逆关系也是需要我们需要考虑的。假设若 OWL 本体中类 d 和类 e 与类 c 之间各一对互逆的对象属性，他们对应的存储数据表分别为 tabled、tablee 和 tablec。我们将 tablec 设置为联系表，在 tablec 中添加两个分别与 tabled 和 tablee 的主键名相同的列，用此列作为数据表 tablec 中的复合主键和引用数据表 tabled 和 tablee 的外键。

(4)将 OWL 中有许多如 sameAs 等用来描述等价的类、性质和实例存储在数据表中。

(5)将 OWL 中使用 objectProperty 等来刻画属性特性的类存储在数据表中。

(6)将 OWL 中使用 allValuesFrom 等用于属性取值的推理的所有类可以存储到属性约束表中。以后如果 OWL 语言又引入了其它的对属性约束的描述，在此数据表中添加相关记录即可，从而减轻程序设计人员修改存储模式的压力。

(7) 将 OWL 中如 allDifferent 等用来描述资源间的二元关系但使用频率又比较低的类全部存储到一张二元关系的数据表。同(6)一样，这种存储办法也可以减轻程序设计人员修改存储模式的压力。

3.3 实 验

3.3.1 实验环境和实验数据

我们选择的实验环境硬件为 CPU P4 1.7GHz，1024M 内存和 80G 硬盘，软件采用 Windows XP 系统和 Mysql 数据库，编程语言为 Java (JDK1.6)。本体测试框架采用比较著名的关于高等学校的领域本体 LUBM。这里我们选择 LUBM(l,0)作为实验数据。

3.3.2 实验方法

在现有的存储模式中，除垂直模式外，其它存储模式的表结构都不稳定，在实际应用中具有很大的局限性。所以，实验只比较本文的存储模式和垂直模式性能方面的不同。在LUBM(l,0)中，分别对 3 个本体文件、9 个本体文件和 16 个本体文件使用 7 个具体的查询进行性能测试：

(1)类 person 的所有子类

(2)属性 subPropertyOf 的所有子属性

(3)实例 Undergraduatestudent63 的属性 takesCourse 的值是什么

(4)图

(Undergraduatestudent63 takesCourse ? x)(? x type Course)

(5)属性 teacherof 的 range 是什么

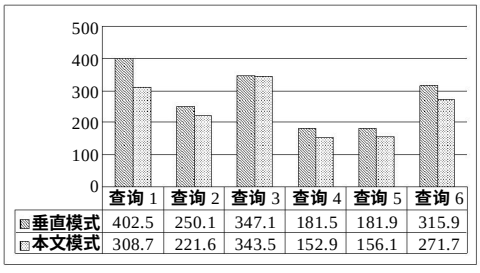
(6)属性 advisor 的 domain 是什么

(7)Faculty 的所有实例

3.3.3 实验结果

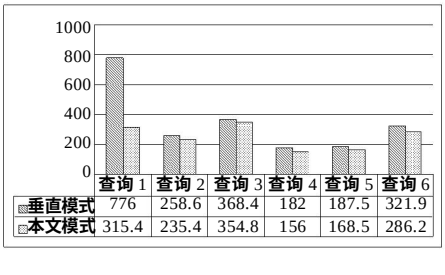
取 5 次连续 100 次查询的所需平进行比较，以 ms 作为时间单位。得到用垂直模式和本文模式进行前 6 个查询的时间对比如图 1、图 2 和图 3 所示，将 3 类本体文件进行查询 7 得到的时间对比如图 4 所示。

由时间对比图可以看出，对于相同数量的本体文件和相关的查询条件，本文设计的混和存储模式的查询时间要比垂直模式的查询时间短，说明本文实现的混合存储模式查询性能更好；从图 4 还可以看出，当本体数量增加时，对于相同的查询，本文的存储模式时间增长速度慢于垂直模式的查询时



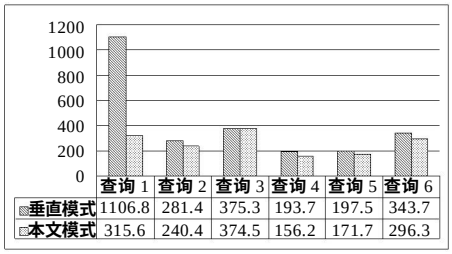
	查询 1	查询 2	查询 3	查询 4	查询 5	查询 6
垂直模式	402.5	250.1	347.1	181.5	181.9	315.9
本文模式	308.7	221.6	343.5	152.9	156.1	271.7

图 1 3 个文件的查询时间对比



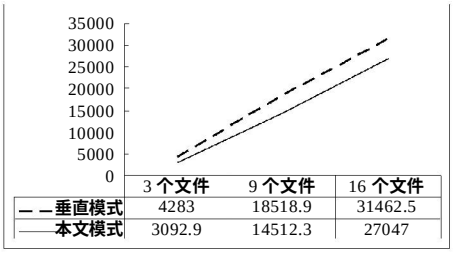
	查询 1	查询 2	查询 3	查询 4	查询 5	查询 6
垂直模式	776	258.6	368.4	182	187.5	321.9
本文模式	315.4	235.4	354.8	156	168.5	286.2

图 2 9 个文件的查询时间对比



	查询 1	查询 2	查询 3	查询 4	查询 5	查询 6
垂直模式	1106.8	281.4	375.3	193.7	197.5	343.7
本文模式	315.6	240.4	374.5	156.2	171.7	296.3

图 3 查询 16 个文件的时间对比



	3 个文件	9 个文件	16 个文件
垂直模式	4283	18518.9	31462.5
本文模式	3092.9	14512.3	27047

图 4 3 类本体进行查询 7 的时间对比

间增长速度，说明用本文设计的存储大规模本体具有更好的扩展性。

4 结束语

对本体的存储介质进行了分析研究，特别是对关系数据库 4 种存储模式进行深入研究分析后，提出了一种结构清楚，查询效率比较高，扩展性比较好的混合存储模式来实现 OWL 本体的存储，用 OWL 测试框架 LUBM[24]进行了实验测试。实验表明本文设计的基于关系数据库的混合存储模式在本体知识的查询效率和扩展性等方面都具有很强的优越性。特别是当数据量增大时，本文设计的存储模式的性能优势体现的更为明显，说明对于大规模本体的存储来说，该模式也是可行。

参考文献:

[1] 贾琦,郭绍忠,丁志芳.基于本体的元数据管理系统的研究[J].计算机工程与设计,2009,30(1):199-202.
[2] 徐立广,金芝,易利军.一个本体语言及本体构造工具的设计[J].计算机工程与应用,2006,42(25):74-79.
[3] 白伟华,朱嘉贤.语义网中基于 Web 资源本体的数据中介服务[J].计算机工程与设计,2010,31(11):2654-2658.
[4] 朱勤斯,虞慧群.一种基于语义网技术和本体的数据集成方法[J].华东理工大学学报(自然科学版),2009,34(1):199-215.
[5] 史一民,李冠宇,刘宁.语义网服务中本体服务综述[J].计算机工程与设计,2008,29(23):5976-5980.
[6] 何召卫,陈俊亮.基于本体关系匹配的信息抽取[J].计算机工程,2007,33(21):207-209.
[7] 李勇,李跃龙.基于关系数据库存储 OWL 本体的方法研究[J].计算机工程与科学,2008,30(7):105-107.
[8] 陈光仪,陈德智.RDFS 本体在关系数据库中的存储研究[J].计算机与数字工程,2008,36(12):188-190.
[9] 汤庸.高级数据库技术与应用[M].北京:高等教育出版社,2008:16-59.
[10] 陈布伟,李冠宇,张俊,等.基于语义网规则语言的推理框架设计[J].计机工程与设计,2010,31(4):847-853.
[11] 郁君甫.基于本体的关系数据库关键词语义查询扩展方法[J].燕山大学学报,2010,34(3):231-236.
[12] 李曼,王琰.基于关系数据库的大规模本体存储的存储模式研究[J].华中科技大学学报(自然科学版),2005,33(增刊):217-220.
[13] 许卓明,黄永菁.从 OWL 本体到关系数据库模式的转换[J].河海大学学报(自然科学版),2006,34(1):95-99.
[14] 唐富年.一种关系数据库模式到本体映射的失效检测方法[J].计算机科学,2010,37(3):170-174.
[15] 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.

(上接第 2892 页)

[9] 岳青,朱利明.基于健康监测系统的东海大桥桥梁结构养护管理体系的构建[J].桥梁建设,2006(2):171-173.
[10] 张莹.桥梁健康监测预警的实现[D].北京:北京工商大学,2007:45-50.
[11] 顾洪博,张继怀.改进的 k-均值算法在聚类分析中的应用[J].西安科技大学学报,2010(4):156-160.
[12] 顾洪博,张继怀.基于孤立点和初始质心选择的 k-均值改进算法 [J]. 长江大学学报 (自然科学版) 理工卷, 2009,6 (1):60-62.
[13] 黄孝.数据流聚类算法分析[J].池州学院学报,2007(5):56-63.
[14] 董一鸿.动态数据库增量式挖掘算法及其应用的研究[D].杭州:浙江大学,2007(11):163-175.
[15] 唐银敏.基于相对密度的聚类算法研究[J].重庆科技学院学报(自然科学版),2010,12(2):166-169.