

基于路径的网络本体语言存储模型

吕刚^{1,2}, 郑诚², 胡春玲¹

(1. 合肥学院 网络与智能信息处理重点实验室, 合肥 230601; 2. 安徽大学 计算智能与信号处理教育部重点实验室, 合肥 230039)
(lvgang119@126.com)

摘要:为提高信息检索效率,提出基于路径的网络本体语言(OWL)存储模型,首先设计了转换和存储 OWL 数据的方法,实现构建包含有类和属性层次结构关系的数据图,然后通过深度优先搜索(DFS)算法建立从根节点的类和属性信息到每个节点的类和属性信息的路径,再将这些信息存储到设计的关系数据库表中。通过实验与现有方法进行了比较,在查询处理时间和本体更新时间性能方面都有改进,方案具有可行性。

关键词: Web 本体语言; 本体存储; 语义网

中图分类号: TP311.13 **文献标志码:** A

Path-based OWL storage model

LÜ Gang^{1,2}, ZHENG Cheng², HU Chun-ling¹

(1. Key Laboratory of Network and Intelligent Information Processing, Hefei University, Hefei Anhui 230601, China;

2. Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, Anhui University, Hefei Anhui 230039, China)

Abstract: To improve the efficiency of information retrieval, a Path-based OWL Storage (POS) model was proposed. In addition, the structure of the POS system for the translation and storage of OWL data was illustrated. A data schema of inputted OWL and a data graph with hierarchical structural information between classes or properties were analyzed by POS system. Also, paths from the root class or property to all classes or properties were extracted via a Depth-First-Search (DFS) method. The extracted hierarchical structural information was stored in a path attribute in the relational database tables. Compared with the traditional method, the processing time for ontology query and update in the experiment has a feasible improvement.

Key words: Web ontology language; ontology storage; semantic Web

语义网被认为是下一代网络 Web 3.0,信息都被赋予了明确的含义,机器能够自动地处理和集成网上可用的信息。语义网使用 XML 来定义定制的标签格式以及用资源描述框架(Resource Description Framework, RDF)的灵活性来表达数据,采用本体的网络语言来描述网络文档中的术语的明确含义和它们之间的关系。本体随着语义网的提出,受到了广泛关注^[1]。

2004 年网络本体语言(Web Ontology Language, OWL)成为 W3C 推荐标准。如何使 OWL 表示的本体正确、一致、可扩展,以及能够实现高效检索的存储是促进语义网发展的一个关键课题。目前的研究方法主要存在以下不足:存储后容易丢失本体语义信息^[2-3];不能满足本体需要扩展更新的需求^[4-5];或者解决了本体的扩展需求,但是在存储过程中对于保持本体结构一致性方面有所欠缺^[6]。因此,本文通过研究 OWL 结构特点,提出了采用基于路径的存储方法,改善了现有方法的不足。

1 模型定义

1.1 路径创建

OWL 描述的本体结构图中,每个数据节点都由一些元素定义组成,本文定义“路径”表示 OWL 文档中类或属性之间层次结构信息。

定义 1 数据图中节点构造。数据图中每个节点被定义

为五元组表示: $N(name) = (N_n, N_u, V_f, C_f, S_f)$, 其中, N_n 表示每个节点的名称, N_u 表示采用深度优先搜索(Depth-First-Search, DFS)算法得到的节点编号, V_f 表示节点是否被访问和路径是否被创建, C_f 表示节点是否有孩子节点, S_f 表示节点是否有兄弟节点。各属性采用“1”标记表示为“真”,“0”标记表示为“假”。

定义 2 创建路径实例定义。创建路径算法的流程如图 1 所示,从数据图中为每个节点创建路径时,主要有如下几种情况。

Case1: Node.visiting_flag = 0 and Node.child_flag = 1;

Case2: Node.visiting_flag = 0 and Node.child_flag = 0;

Case3: Node.sibling_flag = 1 and Sibling_node.visiting_flag = 0;

Case4: Parent_node.sibling_flag = 0 and Parent_node.sibling_node.visiting_flag = 0;

Case5: Ascendant_node.sibling_flag = 0 and Ascendant_node.sibling_node.visiting_flag = 0;

定义 2 定义了创建路径的不同情况, case 1 表示当前节点没有被访问,并且有孩子节点。这种情况下将保存当前节点名在访问节点数组中,并且继续访问它的孩子节点。 case 2 表示当前节点没有被访问,并且没有孩子节点(即叶子节点),这时可以创建当前节点、中间节点和根节点之间的路径。在这种情况下时因为所有叶子节点的兄弟节点的中间节点路径都一样,所以要检查是否有重复路径。 Case 3 ~ 5 都是回

收稿日期:2010-11-22;修回日期:2011-01-11。 基金项目:安徽省自然科学基金资助项目(11040606M133);安徽省高校省级优秀青年人才基金资助项目(2011SQRL134);安徽省高校省级自然科学基金项目(KJ2011B137)。

作者简介:吕刚(1978-),男,安徽来安人,讲师,硕士,主要研究方向:数据挖掘、知识工程; 郑诚(1956-),男,安徽合肥人,副教授,博士,主要研究方向:数据库、数据挖掘、计算机网络; 胡春玲(1972-),女,安徽合肥人,讲师,博士研究生,主要研究方向:人工智能。

溯实例, Case 3 表示了访问的叶子节点有兄弟节点并且没有被访问,这时回溯访问双亲节点并且从兄弟节点作为下一个序列。Case 4 表示了叶子节点没有兄弟节点也没有被访问过,这时回溯双亲节点,并且访问双亲节点的兄弟节点。Case 5 表示了如果没有被访问的节点和没有被访问的兄弟节点,则继续访问这些节点。这些过程重复循环,直到数据图中每个节点都被访问,并且路径完全建立。

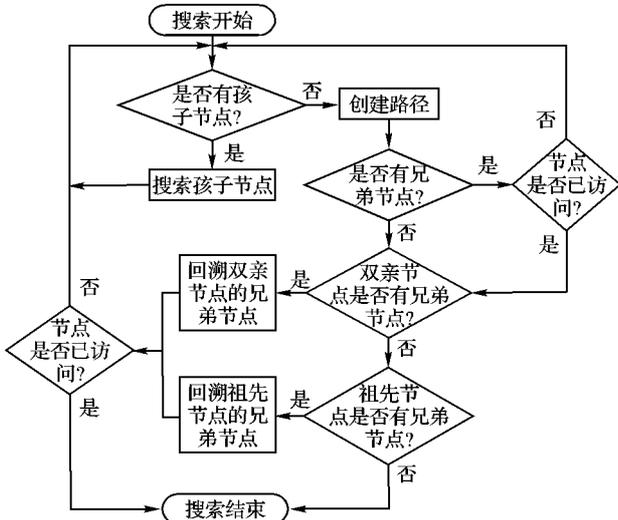


图 1 数据图中创建路径算法流程

1.2 数据库表设计

基于路径的 OWL 存储(Path-based OWL Storage, POS)数据库包括了类或属性之间的层次结构关系信息,能够有效地表示本体的语义信息。图 2 为 POS 数据库表结构,包括 class 表、property 表、triples 表和 instance 表。数据库包含了数据图中表示类和属性的层次结构信息的路径值 class_path 和 prop_path 两个属性,优化了本体层次结构信息表示方法。

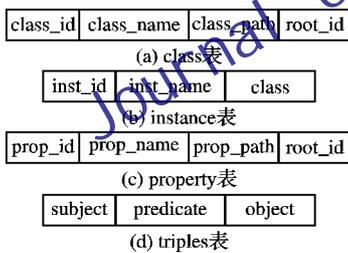


图 2 POS 数据库表结构

为提高访问本体数据的效率,数据库在 class 表和 property 表中增加了 root_id 属性表示 class 和 property 根 ID 号。如果一个 OWL 文档包含多个本体,通过 root_id 属性信息将方便本体的搜索和修改,通过 root_id 属性值,可以减少本体修改和重建的时间。

从 OWL 文档取得类或者属性层次结构信息存储到图 2 所示的 POS 数据库里,具体步骤如下:

- 第 1 步 分析 OWL 文档框架,创建包含类和属性层次关系的数据图^[7]。
- 第 2 步 在数据图上从类/属性根节点开始到叶子节点进行 DFS 遍历,并为每个节点创建路径。
- 第 3 步 当为每一个节点创建了路径以后,从根节点到叶子节点进行搜索,如果到达了叶子节点,则为叶子节点和中间的节点创建一条路径。
- 第 4 步 抽取路径信息存储到 class 和 property 表中。

1.3 基于路径的 OWL 存储系统模型

图 3 描述了将 OWL 文档数据存储到关系数据库的方法。首先,通过 OWL 分析器分析 OWL 文档的句法和语义信息,获得类、属性、结构层次等信息;OWL 文档信息分为层次机构信息和本体信息,一方面层次结构信息用于生成类和属性机构层次的数据图,另一方面本体信息抽取获得类、属性、实例信息以及类、实例之间的关系;最后,整合两个方面的信息存储到本文设计的关系数据库表中。

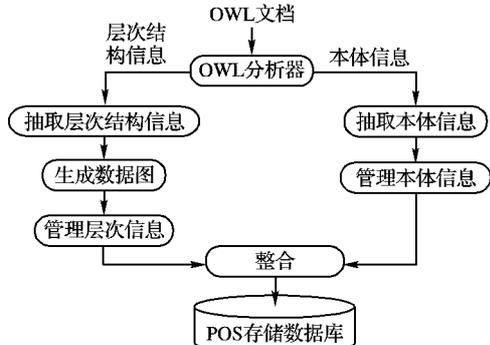


图 3 POS 系统模型

2 实验与性能分析

为检验设计方案的科学性与合理性,进行了实验验证。实验环境采用 Pentium Dual CPU 2.66 GHz, 1 GB 内存,采用 Oracle 9i 作为数据库管理系统,算法使用 Java 语言实现。实验分别使用了本文提出的 POS 模型、文献[8]提出的存储方法 Sesame 系统和文献[9]提出的基于 Jena 的存储方法进行了比较,实验本体数据由 UBA 系统创建^[10],实验本体是关于 University、Department 和 University Activity 的内容,包含了 43 个类、32 个属性。采用 UBA 系统创建了 LUMB(1,0)、LUMB(5,0)和 LUMB(10,0),实验在查询处理时间和本体更新两个性能进行了比较。

2.1 查询处理时间

本文根据查询关于类和属性层次结构,进行存储系统的性能分析。如图 4~5 表示了分别使用 Sesame、Jena 和 POS 系统查询类与属性层次结构实例的时间性能,通过实验结果可以看出,随着 OWL 文件的增加,查询时间也增加。由于 Sesame 和 Jena 在查询时,这两种采用的方法是将本体信息存储在各自不同的类和属性表中,所以查询时要多次重复查询数据库,耗费时间。而本文提出的 POS 模型只需要访问一个表,所以查询速度较快,节约了时间。

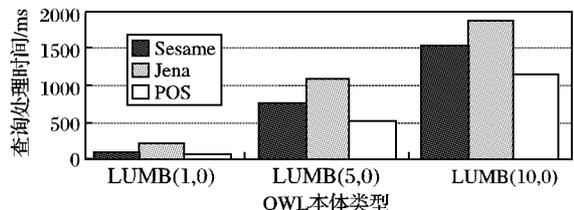


图 4 查询指定类的子类及相关实例

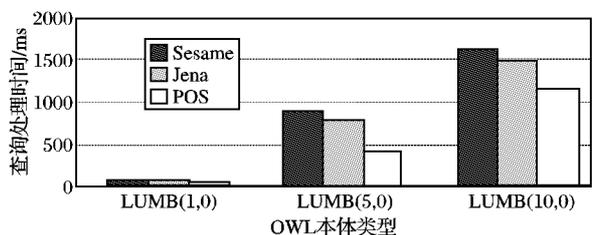


图 5 查询指定属性的子属性及关联的类和对象

2.2 本体更新

本体数据在网络环境中需要经常改动和移动,本体里的层次结构信息或者实例属性可以改动,但是,在存储工程中需要重新构建层次结构信息,耗费大量时间,如图 6 表示了本体更新实例。图 7 显示了分别采用 3 种方法消耗时间,在本体结构进行改动时,Sesame 和 Jena 两种方法对应的数据库需要修改多个数据库表,才能反映出本体所表达的类、属性,以及实例之间的关系本文提出的存储模型,数据库模型存储了每个本体的类、属性根节点。在数据库模型中 class 表和 property 表都包含了 root_id 属性,系统通过管理根节点可以快速修改更新本体内容,从图 7 可看出 POS 方法在本体更新性能优越于其他两种方法。

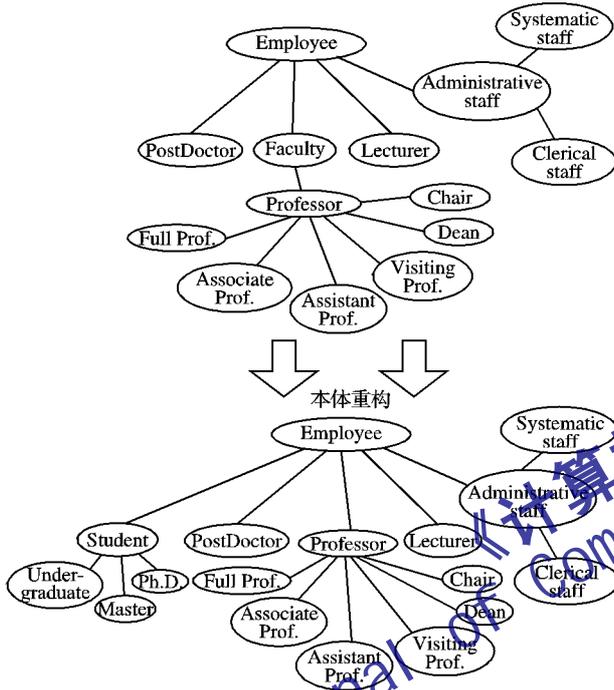


图 6 本体更新方案

3 结语

随着网络信息数据量的快速增长,如何提高检索的准确度已经成为一个重要研究课题。基于 OWL 描述的本体在语义网发展中发挥着重要作用,而在本体中表示的类和属性之间的层次结构关系是关键因素。

本文提出的解决方案克服了传统存储系统的局限性,通

过实验验证了存储模型在完好保存本体语义信息的基础上,提高了查询处理效率和本体更新性能。在今后的工作中,我们将研究继续如何减少抽取的信息存储到数据库中消耗的时间,以提高整个系统性能。

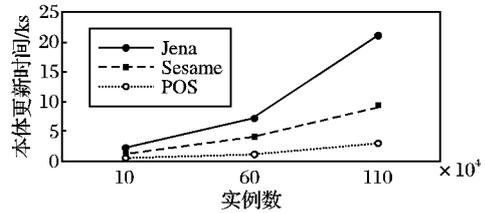


图 7 本体更新时间

参考文献:

- [1] HENDLER J. Ontologies on the semantic Web [J]. IEEE Intelligent Systems, 2002, 17 (2): 73 - 74.
- [2] 朱姬凤, 马宗民, 吕艳辉. OWL 本体到关系数据库模式的映射 [J]. 计算机科学, 2008, 135(18): 165 - 169.
- [3] ASTROVA I, KORDA N, KALJA A. Storing OWL ontologies in SQL relational databases [J]. International Journal of Electrical Computer and Systems Engineering, 2007, 1(4): 1307 - 1379.
- [4] 文敦伟, 樊小虎. 一个基于语义模块的交互式本体匹配框架 [J]. 计算机工程与应用, 2007, 43(29): 1 - 2.
- [5] 陈尧清, 薛建武, 崔璇. 一种异步本体系统的实现框架 [J]. 计算机应用研究, 2009, 26(2): 641 - 644.
- [6] PAN ZHENGXIANG, ZHANG XINGJIAN, HEFLIN J. DLDB2: A scalable multi-perspective semantic Web repository [C]// IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. Washington, DC: IEEE, 2008: 489 - 475.
- [7] ZHOU J T, WANG M W, ZHANG S S, et al. Semi-structure data management by bi-directional integration between XML and RDB [C]// Proceedings of CSCWD'06. Washington, DC: IEEE, 2006: 1 - 5.
- [8] BROEKSTRA J, KAMPAN A, van HARMELEN F. Sesame: An architecture for storing and querying RDF data and schema information [EB/OL]. [2010 - 09 - 12]. <http://www.cs.vu.nl/~frankh/postscript/MIT01.pdf>.
- [9] McKENZIE C, PREECE A, GRAY P. Implementing a semantic Web blackboard system using Jena [EB/OL]. [2010 - 05 - 10]. http://eprints.aktors.org/569/01/ImplSWBbUsingJena_juc2006.pdf.
- [10] GUO Y B, PAN Z X, HEFLIN J. LUBM: A benchmark for OWL knowledge base systems [J]. Web Semantic, 2005, 3(2): 158 - 182.

全国抗恶劣环境计算机第二十一届学术年会征文通知

全国抗恶劣环境计算机第二十一届学术年会将于 2011 年 8 月 23—26 日在山东青岛召开。本次会议由中国计算机学会主办,中国计算机学会抗恶劣环境计算机专委会、中国船舶重工集团公司第 716 研究所承办。会议将通过学术报告、专题讨论等多种形式,充分交流我国抗恶劣环境计算机科研成果,介绍抗恶劣环境计算机技术与产品,展望国内外抗恶劣环境计算机发展趋势和前景。会议将邀请著名专家学者到会做专题报告。

一、征文范围(包括但不限于)

- 抗恶劣环境计算机发展现状与趋势
- 抗恶劣环境计算机需求分析
- 用于自主可控信息系统的计算机技术
- 复杂电磁环境下的计算机技术
- 安全防护与可信计算技术
- 物联网的计算机技术

二、联系信息

投稿邮箱:songly706@sina.com(请注明“2011 抗恶劣征文”)

联系电话:宋凌云(010-68387785);张淑萍(010-68388715)

征文投稿截止日期:2011 年 5 月 30 日 录用通知发出日期:2011 年 7 月 10 日 学术年会召开日期:2011 年 8 月 23—26 日

详情请见:<http://www.ccf.org.cn/sites/ccf/nry.jsp?contentId=2600536795851>