# 基于本体的语义检索技术研究与实现

王继东,张 瑜,李 娜
(西北师范大学 数学与信息科学学院,甘肃 兰州 730070)

**摘 要**:随着网络技术的发展和Internet上信息量的激增,信息检索系统作为网络信息平台的一个重要组成部分,在用户获取准确的网络信息过程之中发挥着重要的作用。传统的检索技术不能对这些信息提供语义级的组织、理解以及处理,寻找新的方法成为目前研究的热点。在现有语义检索方法的基础上,以本体为依据,提出了基于本体的信息检索系统。通过构建领域本体和推理规则,运用Jena实现了语义推理与检索功能,得出潜在的语义查询结果。提高了检索的查全率与查准率。

**关键词**:本体;语义网;语义推理;语义检索
**中图分类号**:TP391.3 **文献标识码**:A **文章编号**:1673-629X(2009)10-0134-04

# Research and Implementation of Semantic Retrieval Technology Based on Ontology

WANG Ji-dong, ZHANG Yu, LI Na
(College of Mathematics & Information Science, Northwest Normal University, Lanzhou 730070, China)

**Abstract**: With the development of network technology and rapid increasing information on Internet, information retrieval system plays an important role at communication between users and resource on the network. Traditional information retrieval technologies can not meet the needs for better organizations, understanding and processing services in the semantic level; therefore, to find new ways has also become a hotspot of current research. Based on the existing semantic retrieval methods, a novel information retrive system based on ontology is proposed in this paper in order to overcome the limitation. The results of potential semantic were gained by building domain ontology and reasoning rules, and implemented semantic reasoning and retrieval function by means of using the Jena. The experiment resutls have proved that the system has higher recall and precision.

**Key words**: ontology; semantic web; semantic reasoning; semantic retrieval

## 0 引 言

随着全球网络化、信息化的发展,网络上的信息越来越多,对信息检索手段的有效性要求也越来越高。但是,目前的搜索引擎基本都采用基于关键字匹配的全文检索技术,查询经常出现检索不全、答非所问的结果。语义检索[1]正是突破了机械式匹配局限于表面形式的缺陷,从词语所表达的语义层次上来认识和处理用户的检索请求。

## 1 语义网和 Ontology

语义网[2](Semantic Web)可以说是新事物,是由WWW的创始人Berners-Lee在2001年正式提出的。Berners-Lee给出了语义网中的层次关系,它主要基于XML和RDF/RDFS,并在此之上构建本体和逻辑推理规则,以完成基于语义的知识表示和推理,从而能够为计算机所理解和处理。

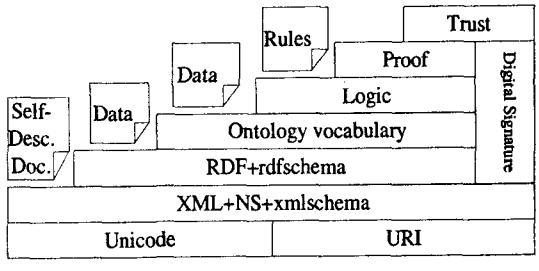


图1 语义网层次结构图

如图1所示,语义网是一套包括网络信息存储、组织、表示、安全认证等各个方面的完整体系,涉及XML、Ontolo-gy、数字签名等技术和方法,它有利于网络信息的基于语义层面的组织和检索,是WWW的

**收稿日期**:2009-02-28;**修回日期**:2009-05-11
**基金项目**:甘肃省科技攻关计划项目(2GS047-A52-002-04)
**作者简介**:王继东(1983-),男,甘肃金昌人,硕士研究生,研究方向为语义Web、Web服务;导师:冯百明,教授,研究方向为计算机体系结构、分布式与并行计算。

发展方向。

Ontology[3]这个术语来自于哲学,它是研究世界上的各种实体以及它们是怎么关联的科学。本体是对应用领域概念化的显示的解释说明,为某领域提供了一个共享通用的理解,从而无论是人还是应用系统之间都能够有效地进行语义上的理解和通讯。Studer 认为,本体是:"共享概念模型的明确的形式化规范说明"。

这包括了 4 层含义:概念模型(conceptualization)、明确(explicit)、形式化(formal)、共享(share)。其中:

(1)概念模型:通过抽象出客观世界中一些现象的相关概念而得到的模型,其表示的含义独立于具体的环境状态。

(2)明确:所使用的概念及使用这些概念的约束都有明确的定义。

(3)形式化:本体应是计算机可读的。

(4)共享:知识本体中体现的是共同认可的知识,反映的是相关领域中公认的概念集,它所针对的是团体而不是个体。

提出本体的目标是捕获相关的领域的知识,提供对该领域知识的共同理解,确定该领域内共同认可的词汇,并从不同层次的形式化模式上给出这些词汇和词汇之间相互关系的明确定义。

## 2　基于 Ontology 的语义检索的基本思想

由于 Ontology 具有良好的概念层次结构和对逻辑推理的支持,因此在信息检索,特别是在基于知识的检索中得到了广泛的应用。

基于 Ontology 的信息检索的基本思想可归纳如下(如图 2 所示):

(1)在领域专家的帮助下,建立相关领域的 Ontology。

(2)利用本体中的概念来标引相关的信息资源并以特定的格式存储,标引的过程与传统的方法类似,可以用手工标引,也可以采取自动和半自动的方式,而结果通常是以 RDF 文档的方式以特定的格式存储。

(3)对 RDF、RDFS、OWL 等相关文件的解析和推理。其目的是为了将以一般文件存储的本体和信息资源信息从文件中读取出来存储在特定的模型中以便于程序处理,并可以根据一定的推理规则基于本体进行语义推理。

(4)对用户检索界面获取的查询请求,查询转换器按照 Ontology 将查询请求转换成规定的格式,在 Ontology 的帮助下从元数据库中匹配出符合条件的数据集合。

(5)检索的结果经定制处理后,返回给用户。

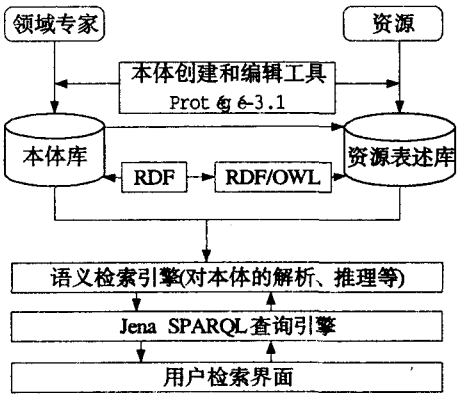


图 2　系统结构图

## 3　检索技术的实现

### 3.1　家族本体的建立

文中用 OWL 语言作为本体的描述语言。OWL 语言建立在 RDF、RDFS 等已有标准的基础上,并通过添加大量的基于描述逻辑的语义原语来描述和构建各种本体。这使得它具有定义良好的语义和表示能力、基于逻辑的推理能力并能保证计算复杂度的可控性和可判定性。为了提高开发效率,在本体的建立过程中,考虑利用现有的开发工具。

斯坦福大学的 Protégé 是最常用的本体开发工具,在文中用到的家庭本体是由 Protégé－3.1 建立的[4],其部分代码如下:

```
……
<rdf:RDF
  <owl:Ontology rdf:about="" />
  <owl:Class rdf:ID="son">
    <rdfs:subClassOf>
      <owl:Class rdf:ID="children" />
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="daughter">
    <rdfs:subClassOf>
      <owl:Class rdf:about="#children" />
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="parents">
    <owl:disjointWith>
      <owl:Class rdf:about="#children" />
    </owl:disjointWith>
  </owl:Class>
  <owl:Class rdf:about="#children">
    <owl:disjointWith rdf:resource="#parents" />
  </owl:Class>
<owl:Class rdf:ID="mother">
```

```
    <rdfs:subClassOf rdf:resource="#parents"/>
  </owl:Class>
  <owl:Class rdf:ID="father">
    <rdfs:subClassOf rdf:resource="#parents"/>
  </owl:Class>
  <owl:ObjectProperty rdf:ID="hasHusband">
    <rdfs:domain rdf:resource="#mother"/>
    <rdfs:range rdf:resource="#father"/>
  </owl:ObjectProperty>
  <owl:ObjectProperty rdf:ID="isFatherOf">
    <rdfs:domain rdf:resource="#father"/>
    <rdfs:range rdf:resource="#children"/>
  </owl:ObjectProperty>
  <owl:ObjectProperty rdf:ID="hasMother">
    <rdfs:domain rdf:resource="#children"/>
    <rdfs:range rdf:resource="#mother"/>
  </owl:ObjectProperty>
  <owl:ObjectProperty rdf:ID="isMotherOf">
    rdfs:domain rdf:resource="#mother"/>
    <rdfs:range rdf:resource="#children"/>
  </owl:ObjectProperty>
.....
  <owl:ObjectProperty rdf:ID="hasDaughter">
    <rdfs:range rdf:resource="#daughter"/>
    <rdfs:domain rdf:resource="#parents"/>
  </owl:ObjectProperty>
</rdf:RDF>
.....
```

### 3.2 对本体的解析和推理的实现

对 RDF、RDFS、OWL 等相关文件的解析和推理是实现语义检索的最关键和最直接的一步，主要通过 Jena 开发包来实现。

#### 3.2.1 Jena 在实现语义检索中的功能

Jena[5,6]是 HP 公司开发的一个基于 Java 的开放源代码语义网工具包，为解析 RDF、RDFS 和 OWL 本体提供了一个编程环境及一个基于规则的推理引擎。文中主要讨论 Jena 的推理功能和查询功能。

语义检索[7~9]是基于概念及其概念之间的关系进行的语义层面的检索，其关键在于概念之间的推理。Jena 提供基于规则的推理机，它包含了一般的推理功能，此外用户还可以根据需要自定义推理规则。如图 3 所示，推理机的工作原理是：推理机注册机制根据基本 RDF 三元组描述和 Ontology 模型创建出推理机，由此推理机可以生成包含推理机制的模型对象（Inference Graph，InfGraph），在 Jena 中，图（Graph）也被称为模型（Model），而表现形式为模型界面（Model Interface），然后可以使用 Model API 和 Ontology API 对此模型进行操作和处理，从而实现语义层面的信息检索。

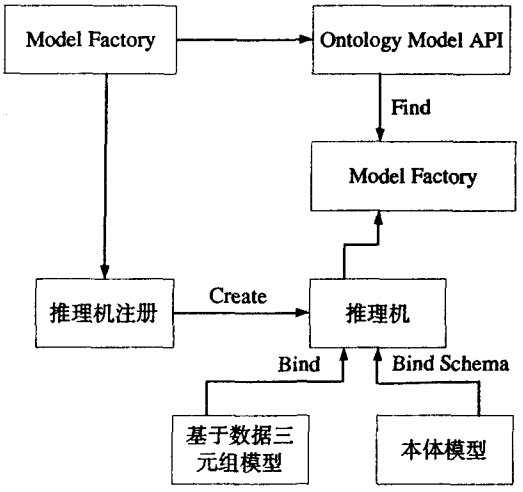


图 3 推理机工作机制

#### 3.2.2 Jena 推理及查询

Jena 自身提供的基于推理的推理机，不能支持所有的 OWL－DL 推理功能。因此，笔者在使用 Jena 所提供的推理规则的基础上，结合查询需要构造了家庭成员关系的推理规则库，把建立好的规则库放入 Jena 推理机中，基于前面建立的本体库进行推理，可以很方便地推理和查询出更多更精确的信息。

使用 Jena 推理机制来实现查询功能的前提是：

（1）家庭成员关系库用 owl 表示。通过 Protégé，已经建立了 family.owl 本体库。

（2）实例需要转换为 RDF 的表示形式。使用 D2R MAP Processor 将数据从关系数据库表示转换到 RDF 表示，建立 members.rdf。

（3）建立家庭成员之间的推理规则。按 Jena 提供的推理规则构造语法来构造家庭成员关系推理规则，实现检索功能。

规则如下：

Rule1(? X hasHusband ? Y) (? X isMotherOf ? Z) －> (? Y isFatherOf ? Z)

即：X 有丈夫 Y，X 是 Z 的母亲，则 Y 是 Z 的父亲。

Rule2(? X hasHusband ? Y)(? X hasDaughter ? Z) －> (? Y isFatherOf ? Z)

即：X 有丈夫 Y，X 有女儿 Z，则 Y 是 Z 的父亲。

Rule3(? X hasHusband ? Y) (? X hasDaughter ? Z) －> (? Y hasDaughter ? Z)

即：X 有丈夫 Y，X 有女儿 Z，则 Y 有女儿 Z。

Rule4 (? X hasHusband ? Y) (? X hasDaughter ? Z) －> (? Z hasFather ? Y)

即：X 有丈夫 Y，X 有女儿 Z，则 Z 有父亲 Y。

Rule5 (? Y hasFather ? Z) (? X hasFather ? Y) －> (? X hasGrandPa ? Z)

即：Y 有父亲 Z，X 有父亲 Y，则 X 有爷爷 Z。

将以上 5 条规则加入 Jena 推理规则中，进行基于

本体库 family.owl 和数据实 members.rdf 推理。

(4)用 SPARQL 查询语言[10]对包含推理关系的数据模型进行查询。SPARQL 从 RDF Model (包括 InfModel 和 OntModel 等)中检索数据,检索的结果存储在一定的数据结构中可供调用,如图 1 所示,将其与用户的检索界面实现联接和集成,则可解决语义检索应用系统与用户交互的问题。

下面是对家族本体进行语义检索的关键代码:

```
String file="../../ontology/family.owl";
  Model data=ModelFactory.createDefaultModel();
    InputStream in;
//读取当前路径下的文件,加载模型
  try{
      in =FileManager.get().open(file);
    data.read(in,"");
    }
    catch (Exception e ){
    }
    Resource configuration=data.createResource();
//将上述推理规则放在 reasoning.rules 文件中,用 reasoning.rules
创建一个符合 RDF 规范的向前链引擎出发的推理机
configuration. addProperty ( ReasonerVocabulary. PROPruleMode,
"forward");
configuration. addProperty ( ReasonerVocabulary. PROPruleSet,
"../../rules/reasoning.rules");
Reasoner reasoner = GenericRuleReasonerFactory. theInstance ( ).
create(configuration);
Model model=ModelLoader.loadModel("../../ontology/member.
rdf");
//根据自定义的推理机创建包含推理关系的数据模型
   InfModel infmodel = ModelFactory. createInfModel ( reasoner,
model);
String searchString ="
PREFIX info: <http://family/memberInfo#>
SELECT ? fanther ? mother WHERE
  { OPTIONAL {wangle info:
hasFather ? father}
OPTIONAL {wangle info:
hasMother ? mother}.}";
//创建一个查询
Query query = QueryFactory.create(searchString);
//执行查询,获得结果
QueryExecution qe=QueryExecutionFactory.create(query, infmod-
el);
ResultSet results=qe.execSelect();
//向控制台输出结果
ResultSetFormatter.out(System.out, results, query);
//释放资源
qe.close();}
```

## 4　结束语

传统的信息检索技术不能满足用户对大量的网络信息资源的检索需求的情况下,研究语义检索技术对信息检索技术的改进有重要的实用价值。文中在研究了惠普实验室开发的语义 Web 应用系统开发工具 Jena-2.5.6 及本体检索语言 SPARQL 的基础上,在手工建立的家族本体上实现了语义的推理与检索功能。相对于传统的检索而言,由于语义检索首先将用户的输入转换为系统所能认知的知识,因而使得查准率和查全率也大大提高。

如何自动建立本体,处理复杂查询将是下一步研究的重点。

**参考文献:**


[1] 黄　敏,赖茂生.语义检索研究综述[J].图书情报工作,2008,52(6):63-66.

[2] 李　洁,丁　颖.语义网关键技术概述[J].计算机工程与设计,2007,28(8):1831-1836.

[3] 张丽坤,蒋　波.基于本体的语义 Web 研究[J].计算机技术与发展,2007,17(6):116-119.

[4] Matthew H. A Practical Guide to Building OWL Ontology Using the Protégé - OWL Plugin and Code Tools[EB/OL]. [2007-04-10]. http:// www. co - ode. org/resources/Protégé-OWLTutorial. pdf.

[5] BRESTOL. Jena2: A Semantic WebFramework[EB/OL]. 2004-10-12[2008-10-12]. http://Jena. Sourceforge. net.

[6] Jena2OntologyAPI[EB/OL]. 2003-08-20[2008-10-12]. http://Jena. Soureforge. net.

[7] 曹志松,曹文君.基于语义 Web 实现有效 Web 信息检索的研究[J].复旦学报:自然科学版,2004(6):422-427.

[8] Lel Y,Uren V,Motta E. Semsearch: A search engine for the semantic web[C]//Proceedings of the 15th International Conference on Knowledge Engineering and Knowledge Management (EKAW), LNAI. Heidelberg: Springer, 2006: 238-245.

[9] 胡必云,黄因生.基于语义的 Web 信息检索[J].计算机技术与发展,2006,16(10):71-73.

[10] Eric P,Andy S. Hewlett P L. Sparql query language for RDF [EB/OL]. 2008-01-15[2008-10-12]. http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/.