

文章编号:1005-0523(2009)05-0031-04

基于本体的智能化语义信息处理系统研究

宋 岚¹,雷莉霞¹,王 洪²

(1.华东交通大学 信息工程学院,江西 南昌 330013; 2.南昌市体育局,江西 南昌 330038)

摘要:研究了推理机在语义推理、查询分析方面的应用,设计并实现1个基于领域本体的语义信息处理的实验系统,建立了教育领域本体库,并构造教育领域的推理规则,通过测试表明该系统能够较为有效地实现语义检索,提高检索精度。

关键词:语义 Web;领域本体;语义推理

中图分类号:TP391.72

文献标识码:A

随着互联网的飞速发展,它的缺陷也逐渐暴露了出来,如网页功能单调、搜索引擎智能化程度低等,这是因为大部分 Web 上的内容是设计给人阅读的,而不是让计算机程序按其意义进行操作的。语义 Web 为了弥补这方面的不足,为网页扩展了计算机可处理的语义信息^[1]。在语义 Web 中,各种资源被人为地赋予了各种明确的语义信息,计算机可以分辨和识别这些语义信息,并对其自动解释、交换和处理。

海量信息的存储问题是促使语义 Web 研究不断前进的动力之一,因此,在语义 Web 的底层,必须具备完善的应用系统开发工具来实现资源描述框架(RDF)的存储。Jena 由 HP 公司的 Brian McBride 开发,起源于早些时候 SiRPAC API 的工作,主要用于开发语义 Web 的应用系统。很长一段时间以来,Jena 凭借其在数据库支持方面的可靠性和持续稳定性,一直是各类应用系统在开发过程中的首选^[2,3]。

1 Jena 推理机制及推理规则构造

Jena 为 RDF、RDFS 和 OWL 提供了可编程环境,它包含了1个基于规则的推理引擎和方便的外部推理机接口^[4]。设有表示图(graph)、资源(resource)、属性和文本(literal)的对象类表示资源,属性和文本的接口分别称为 Resource, Property, 和 Literal。在 Jena 中,一个图被称为一个模型并被 Model 接口所表示。

Jena 自带了1个简单的基于规则的推理机^[5],OWL 推理机使用规则推理机进行所有规则的推理,因此本体推理比用纯 RDFS 数据推理显得更慢。Jena 基于规则的推理常常被应用于 RDF 和 OWL 推理机上,支持前向推理,后向推理和一种混合执行模式的推理。

基于规则推理机中的规则由1个 Java Rule 对象来定义,是由前提、结论和名字及指定的用法构成,规则解析器允许规则以短小精悍的方式出现在文本文件中。

规则语法的描述如下:

Rule : = bare - rule.

or[bare - rule]

or[ruleName: bare - rule]

bare - rule : = term, ... term - > hterm, ... hterm//前向规则

or term, ... term < - term, ... term//后向规则

hterm : = term

or[bare - rule]

term : = (node, node, node)//三元组

or (node, node, functor)//扩充的三元组模式

收稿日期:2009-06-16

作者简介:宋 岚(1978-),女,江西南昌人,硕士,助教,研究方向为人工智能。

```

or builtin(node, ... node)//调用程序
functor := functorName(node, ... node)//structured literal
node := uri - ref//e.g. http://foo.com/eg
or prefix:localname//e.g. rdf:type
or ? varname//变量
or 'a literal'//1个平凡字符串文字
or 'lex' typeURI//1个类型文字,xsd:* type names
//supported
or number//e.g. 42 或 25.5

```

如需要完备的 DL 推理,可以使用专业化推理机^[6],如 Pellet、Racer 和 Fact 推理机。Jena DIG 接口很容易连接任何支持 DIG 标准的推理机,Racer 是 1 个基于 DIG(Description Logic Interface Group)接口的推理机,把 1 个外部推理机加入到 Jena 应用中是很方便的。

用如下语句可实现加载 Racer 推理机操作:

```

.....
Model cModel = ModelFactory.createDefaultModel();
Resource conf = cModel.createResource(); conf.addProperty(ReasonerVocabulary.EXT-REASONER-URL, cModel.createResource("
http://localhost:8080"));
//假设推理机 Racer 在 8080 端口
DIGReasonerFactory drf = (DIGReasonerFactory) ReasonerRegistry.theRegistry().getFactory(DIGReasonerFactory.URI);
DIGReasoner r = (DIGReasoner) drf.create(conf);
ReasonerRegistry.theRegistry().create(DIGReasonerFactory.URI, null);
OntModel base = ModelFactory.createOntologyModel(OntModelSpec.OWL-DL-MEM, null);
//创建 1 个本体模型
OntModelSpec spec = new OntModelSpec(OntModelSpec.OWL-DL-MEM);
//创建 1 个规范模型
spec.setReasoner(r);
//为这个模型规范设置推理机
OntModel m = ModelFactory.createOntologyModel(spec, base);
.....

```

2 系统的设计与实现

我们以 1 个教育系统领域的本体为例,介绍 1 个智能化语义信息处理系统的设计与实现。

系统采用七步法的本体建设方法,参考《中国图书资料分类法》的一些方法,先确定一些本体中的核心概念和关系,构建完几个顶层的概念之后,通过深入研究数字参考咨询领域知识中类与属性、类与类及其属性与属性来完成本体的建设,再通过后续的进化阶段不断扩展和完善本体,最后进行实例的创建。

为了高效地操作 OWL 本体,系统采用 MySQL 作为本体的持久化存储工具,因为直接每次从 OWL 文件里读取 OWL 的图结构信息是非常费时间的过程,而 Jena API 提供了将 OWL 的图数据结构存储到各种数据库后台中,这样从数据库直接读取本体模型到内存就会节约不少时间。

系统中定义的部份规则文件:

```

[allID:(? C rdf:type owl:Restriction),(? C owl:onProperty ? P),
(? C owl:allValuesFrom ? D) -> (? C owl:equivalentClass all(? P, ? D))]
[all2:(? C rdfs:subClassOf all(? P, ? D)) -> print('Rule for ', ? C)]
[all1b:(? Y rdf:type ? D) < - (? X ? P ? Y),(? X rdf:type ? C)]
[max1:(? A rdf:type max(? P, 1)),(? A ? P ? B),(? A ? P ? C) -> (? B owl:sameAs ? C)]
[transitiveRule:(? A demo:p ? B),(? B demo:p ? C) -> (? A > demo:p ? C)]

```

```
[symmetricRule:(? Y demo:p ? X) -> (? X demo:p ? Y)]
[rule1:(? a pre:teach ? b)(? c pre:elective ? b) -> (? a pre:hasStudent ? c)]
[rule2:(? a ? q ? b) < - (? p rdfs:subPropertyOf ? q), (? a ? p ? b)
[rule3:(? c pre:concatFirst ? p), (? c pre:concatSecond ? q) ->
[rule4:(? x ? c ? y) < - (? x ? p ? z)(? z ? q ? y)]];
.....
```

以选课系统为例,系统里应该有这些对象:课程、学生、教师,它们之间互有联系。对于教师,我们定义了教师的工资号、职称及所教授的课程等属性。学生的学号、所选课程等属性的本体信息在 selectcourse. OWL 文件中。

如果 Wang 教授教计算机科学这门课程,而学生 Yang 选修了计算机科学,那么系统自动增加 1 个教师属性 hasStudent,结果为:-(hhu:Wang hhu:hasStudent hhu:Yang),推理机得出学生 Yang 是教授 Wang 的学生。系统运行结果如图 1 所示。

在信息检索过程中,常会遇到这样的情况,多数检索系统只是呆板地通过关键词的匹配来查找用户所需信息,却不能检索出和关键词同义或等价词的相关信息,导致所查信息资料不全。通过语义查询,不仅可以检索出在语法形式上和检索条件相一致的结果,而且能检索出在语义上和检索条件相符的结果,即将检索条件的等价关系、上位关系和下位关系也一并进行检索^[7],因此在一定程度上提高了查全率。图 2 是系统等价关系推理结果的界面。

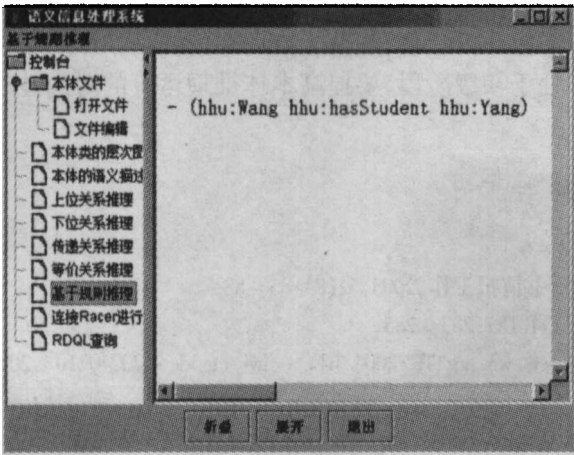


图 1 推理运行结果界面

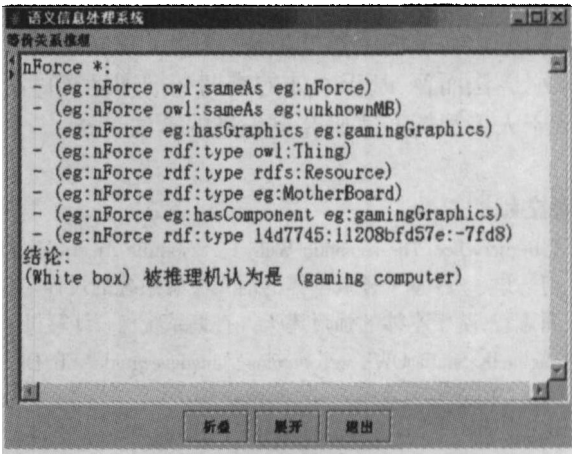


图 2 等价关系推理的界面

3 测试结果

语义推理增强了信息检索的功能,但另一方面,推理过程中的语义扩展、外延和蕴含等操作可能会导致资源的原始特征信息被过分弱化,反而会使检索质量下降,因此,为了兼顾查全率和查准率两方面的因素,在系统中我们规定语义推理所导致的语义扩散距离不得大于 1,即语义扩展、外延和蕴含操作仅局限于具有直接关联关系的词之间。

测试集采用描述软件领域本体的实例文件,数据来源于 SourceForge。信息统计如表 1 所示。

表 1 测试集统计信息

数据集	大小/MB	三元组/个	文本字节/MB	索引大小/MB
SourceForge	10.7	57 699	0.561 505	8.39

实验根据不同的检索方式在不同查全率时的平均查准率绘制了查准率—查全率曲线,如图 3 所示。

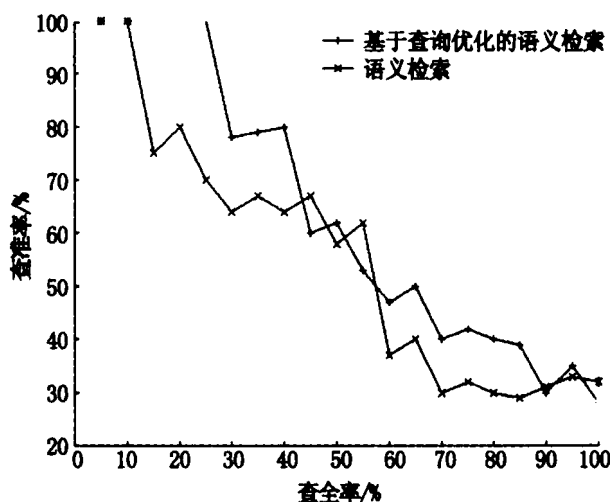


图3 查准率—查全率曲线图

4 结束语

基于本体的检索,查准率和查全率明显高于一般的检索,最重要的原因是我们对 Web 上资源进行了语义标注,资源的文本内容主题被显示地表示出来,因此,可以说基于本体的检索效果的提高,其根本在于人工对检索对象的干预性处理,花费了建立领域本体和语义标注的代价,收获了查询率的提高。相对于全文索引算法而言,基于本体的查询和推理运算算法研究还处于初级阶段,要提高本体推理运算的时间性还有待于人工智能专家研究更先进的算法。

参考文献:

- [1] T Berners-Lee. The semantic Web[J]. Scientific American, 2001, (6): 1 - 6.
- [2] 何琳. 一种基于领域本体的语义检索系统的设计与实现[J]. 图书情报工作, 2008, 52(8): 85 - 88.
- [3] 屈志毅. 基于本体的强对流天气查询系统[J]. 计算机工程, 2008, 34(18): 283 - 285.
- [4] Michael K Smith. OWL web ontology language guide[EB/OL]. <http://www.w3.org/TR/2004/REC-owl-guide-20040210/>, 2004 - 02 - 10.
- [5] Philip McCarthy. Introduction to Jena[EB/OL]. <http://www.ibm.com/developerworks/xml/library/j-jena/>, 2004 - 06 - 23.
- [6] F Baader. Handbook of Description Logics[M]. England: Cambridge University Press, 2003. 47 - 100.
- [7] 宋岚. 基于本体的语义推理研究[J]. 华东交通大学学报, 2007, 24(2): 78 - 81.

A Study of Intelligent Semantic Information Processing System Based on Ontology

SONG Lan¹, LEI Li-xia¹, WANG Hong²

(1. School of Information Engineering, East China Jiaotong University, Nanchang 330013, China; 2. Administration of Sport, Nanchang 330038, China)

Abstract: This paper studies the application of inference engine in semantic reasoning and querying analysis. It also designs and implement a semantic information processing system based on domain ontocogy, establishes an education domain ontology system, constructs reasoning rules in this domoin. The test proves that the system is effective to improve retrieval accuracy.

Key words: semantic web; domain ontology; semantic reasoning

(责任编辑:刘棉玲)