《智能信息处理》课程考试

基于本体的论文检索系统的设计与实现

曹芮

考核	到课[10]	作业[20]	考试[70]	课程成绩[100]
得分				

2021 年 12 月 16 日

基于本体的论文检索系统的设计与实现

曹芮

(大连海事大学 计算机科学与技术，大连 116026)

摘要：为了解决高校等机构学位论文的查询问题，运用本体理论对计算机学科领域中的数据结构进行本体描述，探讨其本体库的构建过程，并在此基础上设计和实现基于本体的试验性论文检索系统，提高系统的检索性能。

关键词：本体；检索系统；语义查询；数据结构

Design and Implementation of Ontology-based Paper Retrieval System

Cao Rui

(Computer science and technology, Dalian maritime university, Dalian 116026, China)

Abstract: In order to solve the query problem of academic dissertations in universities and other institutions, the ontology theory is used to describe the data structure in the field of computer science, and the construction process of the ontology library is discussed. Based on this, the ontology-based experimental paper retrieval system is designed and implemented. Improve the retrieval performance of the system.

Keywords: Ontology;Retrieval system;Semantic search;Data structure

1 引言

基于关键词的检索技术和检索方式严重地割裂了字、词间的语义关联。加之自然语言中一义多词、一词多义现象的广泛存在，导致用户查询获得的检索结果要么包括太多的无关信息，要么返回结果太少。因此，难以满足用户对于文献信息资源查询的需求。针对这些现状，本文运用本体论，通过建立计算机领域中数据结构信息资源本体库，为信息检索在特定领域提供了语义上的支持，在此基础上建立了基于本体的试验性论文检索系统 Psearch，来提高论文文献的检索性能，弥补传统检索方法的不足。

1 本体基本概念

本体（Ontology）的概念源自于哲学领域，在哲学中的定义为“对世界上客观事物的系统描述，即存在论”。哲学中的本体关心的是客观现实的抽象本质。而在计算机领域，本体可以在语义层次上描述知识，可以看成描述某个学科领域知识的一个通用概念模型。德国学者 Studer 在 1998 年给出了本体的相关定义“本体是共享概念模型的形式化规范说明”。这个定义包含了四层含义：即共享（share）、概念化（Conceptualization）、明确性（Explicit）和形式化（Formal）。

（1）共享：指本体中体现的知识是共同认可的，反映在领域中公认的术语集合。

（2）概念化：指本体对于事物的描述表

示成一组概念。

（3）明确化：指本体中全部的术语、属性及公理都有明确的定义。

（4）形式化：指本体能够被计算机所处理，是计算机可读的。

本体通常用来描述领域知识。我们可以这样理解它：本体是从客观世界中抽象出来的一个概念模型，这个模型包含了某个学科领域内的基本术语和术语之间的关系（或者称为概念以及概念之间的关系）。本体不等同于个体，它是团体的共识，是相应领域内公认的概念集合。

1.1 本体的分类

关于本体的研究非常广泛，最为常用的分类方法是根据本体应用主题，将这些为数众多的本体划分为五种类型:领域本体、通用或常识本体、知识本体、语言学本体和任务本体。而依据本体的层次和领域依赖度，Guarino 等人将其分为四类：顶层本体、领域本体、任务本体和应用本体。

（1）顶层本体：研究通用的概念以及概念之间的关系，如空间、时间、事件、行为等，与具体的应用无关，完全独立于限定的领域，因此可以在较大范围内进行共享。

（2）领域本体：研究的是特定领域内概念及概念之间的关系。

（3）任务本体：定义一些通用任务或者相关的推理活动，用来表达具体任务内的概念及概念之间关系。

（4）应用本体：用来描述一些特定的应用，既可以引用领域本体中特定的概念，又可以引用任务本体中出现的概念。

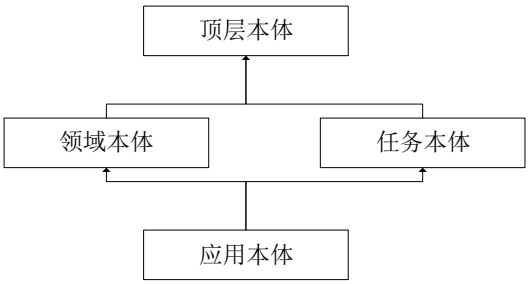

图 1 本体分类

1.2 Psearch 论文检索系统的理论基础

本体是共享概念模型的明确形式化规范说明。本体的目标是捕获相关领域的知识，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。因而它是解决语义层次上信息和知识共享及交换的基础，具有良好的概念层次结构和对逻辑推理的支持，在信息检索，特别是在基于概念的检索中得到了广泛的应用。传统的论文数据库检索系统使用全文和题名关键词检索是基于词的匹配进行检索的,往往存在查不全、查不准、检索质量不高的现象,但是建立了本体后,在知识层面或者说概念层面上就建立了相关领域的知识层次，可以帮助用户获得最佳的检索效果，比如用户可以进一步缩小查询范围或扩大查询，获得上位信息、下位信息以及平级信息等。

2 Psearch 论文检索系统的概述与结构设计

2.1 检索系统概述

Psearch 是基于本体的语义检索试验性系统，包含了计算机学科中数据结构方面的学术论文共 300 篇，这些论文基本涵盖了数据结构中的各个方面。设计该系统的目标一方面是为用户提供专业领域的知识积累，以及查准率和查全率更高的领域论文检索功能。使读者可以获取符合自己要求的各种文献资源;另一方面探索基于本体实现的语义检索的路径和形式,重点在于模拟和分解语义检索过程的内部实现机理和过程,为实现可用于实际应用的论文数据库语义检索系统积累相关理论和技术经验。

Psearch 论文检索系统基于 Web 开发,用户通过 IE 浏览器即可获得信息检索服务,系统界面提供了传统查询和语义查询两种基本查询方式。本系统不仅能检索出在语法形式上和检索条件相一致的结果，而且能通过事先给定的语义关系检索出在语义上和检

索条件相符的结果，这里的语义关系主要包括依赖关系、同义关系、下位关系、反义关系等等，系统借助本体库来完成语义检索功能。

2.2 检索系统的结构设计框架

Psearch 系统的结构框架主要分为三大功能模块：本体库 OWL 数据文档，基于 Jena 设计的 Servlet 后台检索应用程序以及前台基于 JSP 的用户检索界面，检索系统的组成结构框架如图 1 所示。

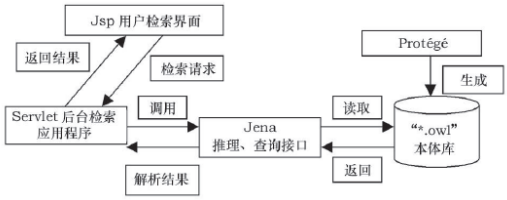

图 1　Psearch论文检索系统结构框架图

3 论文检索系统功能模块的实现

3.1 本体库的生成

Psearch 系统本体库的生成主要包含两个步骤：构建本体模型和本体编码。

（1）构建本体模型

构建领域本体模型是实现 Psearch 系统的第一步，也是整个语义推理和查询的基础。构建领域本体模型是实现 Psearch 系统的第一步，也是整个语义推理和查询的基础。目前，关于本体建模还没有一个统一的方法，比较易于应用的是 Ruben Prirto-Diaz 在他的文章中给出的一种基于方面（Fact）的本体构建方法，基本步骤为：

①在文本中抽取领域词汇。

②根据①中抽取的关键词汇构造领域的方面（Facet），进而把所有词汇分类。

③向①②中构造出的分类词汇系统中加入词汇与词汇之间的关系。

Ruben Prieto - Diaz 方法的优点在于构建条理清晰，是一种递增的构建方法，因此易于实行。但它也存在明显不足：首先，这种构建方法在词汇选择时没有方向性，仅仅是在选择领域词汇后，对于不同的方面把词汇分类而已，因此过程①中领域词汇的获取趋于混乱，很难给出较完整的词汇库；其次，其构建出来的词汇库缺乏层次性，这样在进行检索的时候是对全部词汇逐一匹配，效率很低。笔者借鉴斯坦福大学医学院开发的领域本体建模七步法的一些基本思想，对 RubenPrieto - Diaz 给出的建模方法做了一定改进，弥补上述的不足，形成数据结构信息资源本体库的建模方法：

①在领域分析的基础上，以《中国分类主题词表》和《计算机科学技术汉语叙词表》作为抽取数据结构领域概念的基础，一些核心的数据结构教材为补充，以此来构建数据结构信息资源本体库中数据结构词典所需要的领域词汇。

②使用自上而下的方法，抽取领域中抽象层次最高的词汇，在“数据结构信息资源”领域，首先可以分为“数据结构词典”与“数据结构文章”两大类，然后对抽象层次高的词汇的内涵词汇进行选取，如对“数据结构词典”这一词汇，其内涵词汇包括“数据结构算法”与“数据结构类型”两部分，所以又可以将这些词汇加入领域词汇。按照这一思想，随着词汇抽象层次的降低，本领域中各个方面都被确定，最终就可以得到描述一个领域所需的全部词汇，如无向图、有向图、二叉树、哈夫曼树、广义表、线性表、各种算法等等的关键词。

③在抽取出的领域词汇中选出抽象层次最低的词汇组成底层子类，然后归纳出上层子类，接着再寻找每一个上层子类的父类如此递归，最终构造出一棵类层次树，如图 2 所示。使用这种自底向上的构造方法可以为上一步领域词汇抽取是否完备做一个检验，如果在构造过程中出现问题，可以及早解决以确保在关系构造时的正确性。

3.2 基于 Jena 设计的 Servlet 应用程序的实现

Psearch 试验性系统的后台检索程序的开发主要是使用惠普实验室发布的 Jena2.4 开发包，它为 RDF、RDFS、OWL 文档提供了一个程序开发环境以及比较完整的解析、持续储

存、推理和查询的函数调用和处理接口。根据实际检索的需要，后台应用程序的开发使用了部分函数接口，具体实现过程如下：

（1）本体推理模型的建立

将本体文件和 Jena 自带的推理机捆绑成一个 InfoModel 对象，它是本体模型推理的结果，是下一步查询的基础。考虑到已经建立的本体文件中含有中文字符，为了 Jena 以后能顺利地解析该对象，本系统采用具有可输入编码格式的 Input-StreamReader 对象读取本体文件，InfoModel 对象建立的方法如下：

```
FileInputStream file = new File InputStream (owlFile);
InputStreamReader in = new InputStreamReader( file, "UTF - 8" ) ;
Model tempModel = ModelFactory. createDefaultModel( ) ;
tempModel.read ( in, null) ;
Reasoner owlReasoner= ReasonerRegistry. getOWLReasoner( ) ;
Reasoner reasoner = owlReasoner. bindSchema ( tempModel) ;
Model infmodel = ModelFactory.create InfModel( reasoner, tempMod2el) ;
```

4 论文检索系统的性能评价

笔者结合传统文档检索系统的评价标准，对该检索系统的查准率、查全率[6]进行性能分析。实验中，考虑到本系统检索的语义扩展性，通过下位语义扩展得到的文档与原始查询间有一定的语义距离，所以在测试查准率和查全率这两个指标时，笔者不把此部分文档算入正确检索文档集合，但需要考虑由等价语义扩展得到的论文，因为它们与原始查询间语义距离为零。

根据查准率和查全率的定义，由测试的结果可以得到题名关键字检索的查准率 100% 略高于语义检索的查准率 90% ，其原因有两点：一是实验检索的文档集合是来自于计算机领域中的专业文档库，其所包含的全部是与数据结构知识有关的文档，这就大大降低了领域核心概念在检索结果中出现二义性的几率，查准率在语义检索中的优势体现不足；二是本系统的检索结果中包含由语义扩展得到的部分文档，这在计算查准率时，由查准率的定义和公式，增加了其分母“检索出的文档数”的值，使得查准率有所降低。但是对于查全率来说，语义查询的查全率 100% 大大优于题名关键词查询的查全率 66. 7% ，通过语义查询可以获得传统查询中丢失的等价语义和下位语义扩展得到的论文，这些文档对于用户来说也非常有价值。

5 结 语

本文给出了基于本体的论文检索系统的设计与实现过程及其查询性能分析。从试验结果上来看，在系统中引入本体论，能够实现计算机领域中对于数据结构方面论文的语义检索，拓展了查询条件，让用户得到了基于关键词匹配无法检索到的一些有价值的论文文献，改善了传统论文查询方法性能的不足。通过该检索系统的实现，探索了基于本体实现的语义检索的路径及其过程的实现机理，为实现可用于实际应用的论文数据库语义检索系统提供了有价值的参考。该实验性系统可以从两方面进一步扩展和改进：一方面，丰富数据结构本体，给出更多的可以作为扩展依据的关系，本体的丰富可以为查询条件的扩展提供更多路径。另一方面，扩大本体的范围，由于受本体论技术发展的限制，目前尚难以构造通用的本体论知识模型。因此，构建更为通用的领域本体库，使检索系统不仅仅限制在某一领域还需要作进一步研究。

参考文献：

[1]D.Fensel,S.Decker,M.Erdmann,R.Studer.Ontobroker:How to make the WWW intelligent,KAW,Banff,1988, 11(4):23-26

[2] 彭鹏，基于本体的信息检索策略优化研究，吉林大学硕士学位论文，2007;4

[3] 刘仁宁，李禹生. 领域本体构建方法[J].

武汉工业学院学报,2008(01):50-53+57.
[4] 宋峻峰,张维明,肖卫东,唐九阳.基于本体的信息检索模型研究[J]. 南京大学学报(自然科学版),2005(02):79-87.
[5] 邓志鸿,唐世渭.Ontology 研究综述.北京大学学报(自然科学版），2002（5）：730 - 738
[6] Prieto-Diaz R. A Faceted Approach to Building Ontologies Information Reuse and Integration. In: IR I 2003.IEEE Infernational Conference. 2003. 458 - 465
[7] Natalya F.Noy,Deborah L.McGuinness Ontology Development 101:A Guide to Creating Yourfirst Ontology.August,2001.