基于本体论的论文检索

朱庆生　邹景华
（重庆大学计算机学院　重庆400030）

摘　要　本文首先分析了传统论文检索机制的不足，在此基础上，提出将本体论应用于论文检索中的基本思路，建立了论文本体模型，设计了基于本体论的论文检索系统，最后根据所建立的检索模型，研究了智能化论文检索的相关算法。

关键词　本体，信息检索，元数据

Paper Retrieve Based on Ontology

ZHU Qing-Sheng　ZOU Jing-Hua
(College of Computer Science and Engineering, Chongqing University, Chongqing 400030)

Abstract　In this paper, we first analyze the lack of the traditional paper retrieves mechanism, and then introduce ontology into the paper retrieve system. According to our analysis, we build a paper ontology model and propose a new architecture of paper retrieves system. In the end, we study the intelligent paper retrieve algorithm based on our retrieve model.

Keywords　Ontology, Information retrieve, Metadata

1　引言

随着网络的飞速发展，网上的信息资源日益增多，人们获取信息的方式不再只局限于书本上，更多的时候是在利用网上的电子资源。各类学术论文也逐渐摆脱了传统传媒的限制，以电子文档的形式在网上传播。到目前为止，网上已建成了几大论文索引平台，收录了大量的学术论文，如“CNKI 全文期刊、专利库”，“维普全文科技期刊库”，“万方期刊”等，其中CNKI 已收录8,710,000多篇学术论文。在如此庞大的论文库中，如何快速有效地检索论文资料也就成为当前一项重要而迫切的研究课题。

传统信息检索技术都是基于字词的关键字查找和全文检索技术，主要借助于目录、索引和关键词等方法来实现。此技术的优点是简单、快捷，但其存在四个较突出的问题：第一，“忠实表达”问题。很多情况下，用户很难简单地用关键词或关键词串来忠实地表达他所真正需要检索的内容，表达困难导致检索困难；第二，“表达差异”问题。人类的自然语言中，随着时间、地域或领域的改变，同一概念可以用不同的语言表现形式来表达。因此，对同一概念的检索，不同的用户可能使用不同的关键词来查询，而传统信息检索技术则很难解决同义词查询的问题；第三，“词汇孤岛”问题。在人的大脑中，概念并不是孤立存在的，它总是与其它概念之间存在各种各样的联系。在传统信息检索中，这种概念之间的联系是无法表示的；第四，其过分追求高的查全率导致了检索结果的数量过于庞大，用户根本没有时间和精力处理检索到的所有结果。总之，在信息快速增长的今天，传统信息检索机制由于缺乏必要的智能性，难以满足用户的要求，而论文检索作为信息检索技术的具体应用，也存在同样的问题。

造成这些问题的实质在于传统的信息检索技术所采用的只是基于语法层面上字、词的简单匹配，而缺乏对知识的表示、处理和理解能力。解决这些问题的关键就是要把信息检索从传统的基于关键字层面提升到基于知识（或语义）层面上来。而本体论（Ontology）作为对事物本原研究的方法论，具有良好的概念层次结构和对逻辑推理的支持，已被广泛应用于知识表达、知识共享及重用，这正是本文将本体论应用于论文检索的重要理论依据。

2　本体论（Ontology）

Ontology 最早是一个哲学上的概念。从哲学的范畴来说，Ontology 是对客观存在系统的解释或说明，关心的是客观现实的抽象本质。在人工智能界，最早给出 Ontology 定义的是 Neches 等人，他们将 Ontology 定义为“给出构成相关领域词汇的基本术语和关系，以及利用这些术语和关系构成的规定这些词汇外延的规则的定义[1]”。Studer 等人在对本体论定义进行了深入的研究之后，认为 Ontology 是共享概念模型的明确的形式化规范说明。这包含4层含义[2]：概念模型（conceptualization）、明确（explicit）、形式化（formal）和共享（share）。“概念模型”指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。概念模型所表现的含义独立于具体的环境状态。“明确”指所使用的概念及使用这些概念的约束都有明确的定义。“形式化”指 Ontology 是计算机可读的（即能被计算机处理）。“共享”指 Ontology 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，即 Ontology 针对的是团体而非个体的共识。

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，从不同层次的形

朱庆生　教授，博导，主要研究方向为多媒体数据压缩、网络信息系统和软件工程。邹景华　硕士研究生，主要研究方向为数据挖掘、本体论和语义网络。

式化模式给出这些词汇(术语)和词汇间相互关系的明确定义,通过概念之间的关系来描述概念的语义。概念间的语义关系主要包括有同义词关系、上下位关系、类属关系等。

3 基于本体论的论文检索系统框架

在本体论的指导下,本文构建了一个智能化的论文检索原型系统,该系统由如下几个模块组成:本体构建、知识获取、语义词典和用户检索,系统体系结构如图1所示。

3.1 本体构建

为了方便地进行人机交互和概念间的语义转换,本文通过对论文对象的抽象,建立了论文的本体模型。在论文本体模型中声明了论文的元数据结构(metadata schemas)以及论文与其他实体之间的语义网络(semantic Web)。

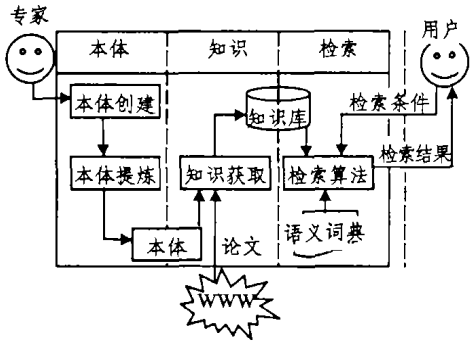

图1 基于本体论的论文检索体系结构

表1 论文的元数据结构

元素	类型说明	说明
标题	字符串	论文的标题
作者	实体对象	论文的作者
摘要	字符串	论文的摘要
关键字	字符串	论文的关键字
中图分类号	字符串	论文的中图分类号
参考文献	实体对象	论文的参考文献
出版时间	日期	论文的出版时间
出版期刊	实体对象	论文的出版期刊
收录情况	实体对象	论文的收录情况

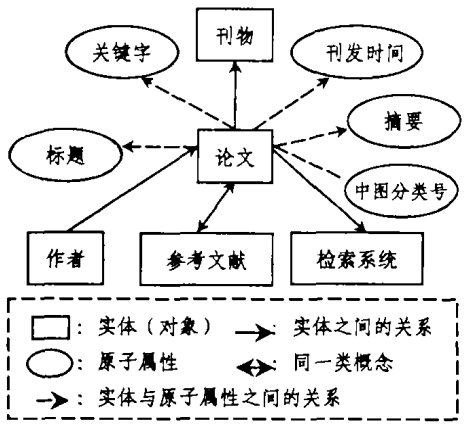

图2 论文语义网

3.2 知识获取

目前网上期刊全文数据库中收录了大量的学术论文,并提供了统一的检索界面,因此可从中收集论文信息,构建论文的知识库。

网页信息抽取的方法较多[3,4]。由于本模型是以网上期刊全文数据库作为论文知识抽取的数据源,其网页上论文信息的表现形式比较固定统一,因此可以采用网页结构分析法[4],获取所需的论文知识。现以CNKI全文数据库为数据源,简单描述网上论文知识的获取过程,在CNKI检索页面中,有关论文元数据信息的HTML代码如图3所示。

```
<table>
    <tr>
        <td>[篇名]</td>
        <td>数据挖掘技术</td>
    </tr>
    <tr>
        <td>[作者]</td>
        <td>李建忠.. </td>
    </tr>
    <tr>
        <td>[刊名]</td>
        <td>中国金融电脑</td>
        <td>1997年08期</td>
    </tr>
    <tr><td></td><td>CJFD收录期刊</td></tr>
    <tr>
        <td><b>[摘要]</b></td>
        <td>概述数据挖掘…</td>
    </tr>
    …
</table>
```

图3 CNKI论文检索结果页面HTML代码片断(已去除显示控制和其它无关代码)

为了信息抽取的方便,本文采用DOM模型描述论文检索页面的结构。在DOM模型中,非叶节点代表HTML代码中的标签字段,叶子节点表示网页的具体描述信息,从根节点出发到叶子节点的路径则反映了论文数据在网页中的抽取路径,如图3中论文标题元数据的抽取路径为Table.Tr(1).Td(2)。根据已建的论文本体和论文元数据在网页中的抽取路径,可实现网页中论文元数据的自动抽取。

3.3 语义词典

在基于本体的信息检索系统当中,语义词典是用来描述术语之间的语义关系,在整个系统中起着举足轻重的作用。术语之间的语义关系包括有同义关系、上下位关系等。同义关系是指对同一概念的不同表达形式,描述的是术语间的等价关系,如"数据挖掘"和"知识发现"。上下位关系则指的是概念之间的蕴含关系,即某一概念蕴含于另一概念中,如"数据库"和"关系数据库"。

本模型中的语义词典主要描述论文关键词之间、关键词和学术领域之间的语义关系。由于论文涉及的领域众多,对其归类的难度较大,为此本文采用"中图分类法"的思想,对论文关键词所蕴含的语义关系进行描述和刻画。"中图分类法"作为一种广泛认可的图书分类方法,是建立在科学分类的基础上,结合图书资料的特性所编制的分类法,它将学科分成五大类和若干细目,能比较清晰地展示各学科之间的类属关系、上下位关系等。由于在论文元数据中也有中图分类号,因此,"中图分类法"比较适合于论文关键词的分类,能大大减少论文词典构建的复杂度。

语义词典构建时,先根据论文的中图分类号,将论文的关键词填入到中图分类表中对应的位置,形成词典基表,其能反映关键词之间的上下位关系。另外,由经验可知:同义关键词一般只出现在同类论文(具有相同或相近的中图分类号)中。因此,在词典基表的基础上,只需专家对词典基表中的相邻或相近的关键词进行少许划分和调整,即可得出关键词之间的同义关系,并最终形成有关论文关键词的语义词典。

4 基于本体论的论文检索算法

众所周知,传统的信息检索技术依赖于关键字匹配,若匹

(下转第176页)

所以在当前上下文的情况下，“人”应取编码为 Aa01的义项，完成歧义消除。

表1 多义词消歧结果

多义词	义类数	样本数	特征数	正确率
人	5	6499	1223	81.39
等	4	3130	1227	86.12
说	3	3482	1065	89.48
和	9	6519	1739	85.25
是	5	5837	1379	94.67
中	11	7083	1435	86.47
对	12	3851	1159	86.01
要	9	3813	837	86.57
与	6	2456	1027	84.30
在	7	7875	1499	87.27
地	7	4428	1144	89.97
以	6	4293	1225	90.45
将	6	2118	826	92.12
都	5	2218	891	92.76
还	5	2442	1088	86.12
把	12	1538	877	80.37
下	15	2866	873	92.61
于	6	2943	949	93.91
了	6	6524	1419	80.81
上	17	4984	1267	93.24

2. 实验结果 实验测试的20个多义词如表1第1列所示，第2列是每个多义词从《同义词词林》中获取的义类数，第3列是训练阶段的样本数，第4列是提取的特征数，最后1列是消歧的正确率。每个多义词使用了200个测试例。

实验获得了较好的结果，平均将近88%的准确率。从实验结果我们可以看出并不是训练阶段提取的特征越多越好，一个适度的特征数才能取得较好的正确率。

结语 至此，我们详细介绍了本文提出的基于最大熵原理的汉语多义词消歧方法。本文介绍了最大熵的原理，讨论了在最大熵框架下的特征选择，用一种统一的形式解决了各种特征的表示问题，综合利用各种语言信息进行消歧。实验证明本文提出的方法是有效可行的。

下一步的工作需要放在对汉语语言知识的研究上，怎样获取新的汉语语言知识，把它形式化后加入到模型中去，进一步提高消歧的正确率。

参 考 文 献

1 屈刚，陆汝占. 基于特征的汉语词性标注模型. 计算机研究与发展，2003，40(4)：556～561
2 李涓子，黄昌宁. 语言模型中一种改进的最大熵方法及其应用. 软件学报，1999(3)
3 李涓子，黄昌宁，杨尔弘. 一种自组织的汉语词义排歧方法. 中文信息学报，13(3)
4 A Maximum Entropy Approach to Natural Language Processing
5 A Maximum Entropy Approach to Adaptive Statistical Language Modeling
6 A Maximum Entropy-based Word Sense Disambiguation system

（上接第173页）

配失败则返回空记录，若进一步查找，则需用户手动更改查询条件，系统不具备自动推理功能。本文在基于本体和知识的论文检索体系的基础上，设计了一种智能化的论文检索算法：在

输入：查询关键词
输出：Φ或者满足条件的论文

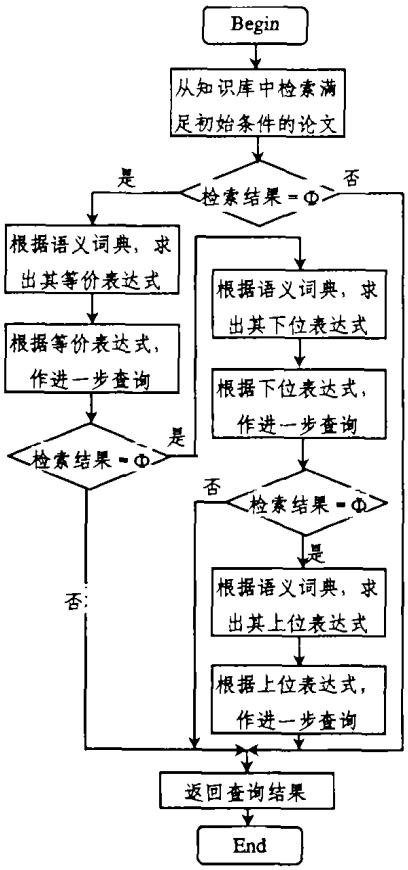

图4 智能化论文检索算法流程图

一般匹配失败的情况下，借助于本体和语义词典对论文关键词语义的刻画，系统能够自动寻找另一条合理的路径，进一步查找。

假设要检索有关“本体论在信息集成方面应用”的相关论文，先根据用户的检索要求，提取检索原始关键词，即“本体论”和“信息集成”。由于论文数据库中没有“精确”包含这两个关键词的相关论文，因此，基于传统信息检索技术的查询结果则为空。但根据语义词典，可知“本体论”和“信息集成”与其它关键词之间存在有丰富的语义联系，如“本体论”和（“本体”、“本体理论”、“信息知识本体论”、“元数据”、“语义网络”）之间存在同义关系等，因此，系统则根据关键词的语义关系（等价关系、上下位关系），自动更改查询条件，做进一步的查询。其具体的检索算法框图描述如图4。

结束语 有关本体论的研究在计算机界受到越来越多学者的重视，然而关于本体论的应用却还处在雏形阶段，没有统一的定义和固定的应用领域。本文在分析现有论文检索机制缺陷的基础上，探讨了将本体论应用于论文检索的新思路，建立了描述论文的本体，提出了基于本体论的论文检索系统模型，并在此基础上，研究了智能化论文检索算法，利用这种检索机制能有效提高系统的检索性能。

参 考 文 献

1 Neches R, Fikes R E, Gruber T R, et al. Enabling Technology for Knowledge Sharing. AI Magazine, 1991, 12(3):36～56
2 Studer R, Benjamins V R, Fensel D. Knowledge Engineering, Principles and Methods. Data and Knowledge Engineering, 1998, 25(122):161～197
3 Zhang N, Chen H, Wang Y, et al. Odaies: ontology-driven adaptive Web information extraction system. IAT 2003. 454～460
4 Buttler David, Liu Ling, Pu Calton. A Fully Automated Object Extraction System for the World Wide Web. In: Proc. of the 21st Intl. Conf. on Distributed Computing Systems, 2001. 361～371