文章编号：1005-8451（2010）11-0011-04

基于领域本体的垂直搜索引擎模型的研究

林碧霞，尹治本
（西南交通大学 信息科学与技术学院，成都 610031）

摘 要：用户对于智能化、专业化搜索引擎的需求大力推动了语义搜索的发展。本文在这个需求的环境下提出一种基于领域本体的垂直搜索引擎模型，该模型更加智能化，并且耦合性较低，能满足不同领域的定制和开发。

关键词：领域本体；垂直搜索引擎；主题爬虫；上下文主题描述
中图分类号：TP39 文献标识码：A

Research on vertical search engine based on domain ontology

LIN Bi-xia, YIN Zhi-ben
(School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, China)

Abstract: The needs for the search engine which was more intelligent, professional vigorously promoted the development of semantic search.This paper gave a new module:a vertical search engine based on domain ontology.This model was more intelligent search engine, lower the coupling and could be used in different areas.

Key words: domain ontology; vertical search engine; topic-specific crawler; contextual topic description

随着网络的飞速发展，Web信息呈爆炸性地增长，如何在浩瀚的网络中找到人们需要的信息变得更加的重要。传统的搜索技术虽然满足了人们一定的需要，但是由于其通用的性质，仍然不能满足不同背景、不同时期和不同目的的查询要求。同时如何让计算机理解用户所要查询的信息这也是当今搜索领域面临的一大挑战。传统的通用搜索引擎的不足催生了垂直搜索引擎的发展，同时也使人工智能化的搜索引擎得到了学术界的广泛关注。

垂直搜索就是针对某一行业的专业搜索引擎，是对某类、某行业、某领域的信息的采集和整合，从而为某一类人群或某一领域的用户提供专业和精准的信息[1]。其特点就是“专、精、深”，具有很强的行业和领域特色。

近年来，本体理论的发展、成熟也为搜索引擎的发展带来了新的动力，也为提高检索系统的查全率和查准率提供了进一步的保证。本文在现有

收稿日期：2010-03-31
作者简介：林碧霞，在读硕士研究生；尹治本，教授。

表还可以当作是JPEG图像天然的认证数据，通过检测JPEG图像中量化表的一致性来判断图像的真实性。

参考文献：

[1] 张德峰. MATLAB数字图像处理[M]. 北京：机械工业出版社，2009.
[2] 张汗灵. MATLAB在图像处理中的应用[M]. 北京：清华大学出版社，2008.
[3] Yang En-hui, Wang Long-ji. Joint optimization of run-length coding, Huffman coding, and quantization table with complete baseline JPEG decoder compatibility [J]. IEEE Transactions on Image Processing, March 2009, v 31, n 3: p 552-555.
[4] 何 斌，马天予，王运坚，等. Visual C++数字图像处理[M]. 2版. 北京：人民邮电出版社，2002.
[5] 左 飞，万普森，刘 航. Visual C++数字图像处理开发入门与编程实践[M]. 北京：电子工业出版社，2008.
[6] 赵春江. C#数字图像处理算法典型实例[M]. 北京：人民邮电出版社，2009.
[7] Kornblum, Jesse D.. Using JPEG quantization tables to identify imagery processed by software [J]. Digital Investigation, September 2008, v 5, n SUPPL., p S21-S25.
[8] 于万波. 基于MATLAB的图像处理[M]. 北京：清华大学出版社，2008.

垂直搜索与本体论的研究成果的基础上，提出一种基于本体的上下文主题描述，并在这基础上提出了一种基于本体论的垂直搜索引擎模型。利用该模型，能够快速解决垂直搜索引擎的定制问题，使得这个框架定制于各种不同的领域，同时将提高垂直搜索引擎的性能。

1 本体论及相关研究

1.1 本体论定义

本体是概念模型的明确的规范说明[2]。本体的实质是把本体当作是领域（特定领域，或更广的范围）内部的不同主体（人、机器、软件系统等）之间进行交流（对话、互操作、共享等）的一种语义基础，即由本体提供一种明确定义的共识。

一般，最典型的本体应具有一个分类系统和一系列推理规则，分类系统定义对象的类别和类目之间的关系；借助推理规则，可以提供更强的推理能力。本体的类型有如下几种：

（1）顶级本体：描述的是最普通的概念及概念之间的关系，如空间、时间、事件、行为等，与具体的应用无关。

（2）领域本体：描述的是特定领域中的概念及概念之间的关系。

（3）任务本体：描述的是特定任务或行为中的概念及概念之间的关系。

（4）应用本体：描述的是依赖于特定领域和任务的概念及概念之间的关系。

1.2 本体在信息检索领域的应用现状

本体是一种技术，它可以在许多涉及知识表示与共享的环境下应用。由于本体具有良好的概念层次结构，并且支持逻辑推理，这使得本体在信息检索，特别是知识检索中得到了广泛的应用。

在国外，目前本体应用在信息检索领域中的著名项目包括（Onto）Agent、Ontobroker和SKC（Scalable Knowledge Composition）。这3个项目的方向不同：（Onto）Agent目的是为了帮助用户检索到所需要的WWW上已有的本体，主要采用了参照本体。参照本体是以WWW上已在的本体为对象建立起来的本体，它保存有各类本体的元数据。Ontobroker面向的是WWW的网页资源，目的是为用户检索到所需要的网页，这些网页含有用户所关心的内容。SKC目标是解决信息系统语义异构的问题，实现异构的自治系统之间的互操作。该项目希望通过在本体上建立异构代数系统，用这个代数系统来实现各本体之间的互操作，从而实现异构系统之间的互操作。

国内也有一些学者正在研究如何将本体应用于信息检索领域，但是，基于本体的信息检索还处于实验原型阶段，还没有真正进入商业化实施阶段。国内主要的相关研究包括：

（1）基于本体的KMSphere知识管理平台[3]。充分利用智能主体自主性、社会性和反应性。KMSphere知识管理平台能通过概念空间检索出用户真正想要的文本信息，实现了基于概念的智能互动语义查询。

（2）Falcons系统。Falcons是一个面向领域的语义搜索系统。在事先构建的领域本体和知识库的基础上，在索引过程中提取并存储语义信息，在传统的基于向量空间的索引模型上增加了显式的语义信息。在此语义索引的基础上提供了本体驱动的搜索和浏览机制，提供了一种新颖的基于图的查询机制。

2 基于领域本体论的垂直搜索模型

为了提高搜索引擎的准意性、语义性和针对性，加强搜索信息过程中的智能化处理是本框架要解决的关键问题。

基于领域本体的信息检索的基本思想是：在领域专家的帮助下，建立相关领域的本体；收集信息源中的数据，并参照已建立的本体把收集来的数据按规定格式存储在元数据库中；对从用户检索界面获取的查询请求，查询转换器按照本体把查询请求转换成规定的格式，在本体的帮助下从元数据库中匹配出符合条件的数据集合，检索的结果经过处理后返回给用户[4]。

本文设计一种基于本体的垂直搜索模型，见图1，其主要思路是用户提出垂直搜索的请求，智能检索器接受请求，经过查询分析，按照要求转换成规定的格式提供给推理机，进行推理、判断和语义分析等，得出用户在这个领域里的准确语义，在领域本体的帮助下查询索引数据库中与其相匹配的、符合条件的数据集，经整合、排序后输出给用

户终端。

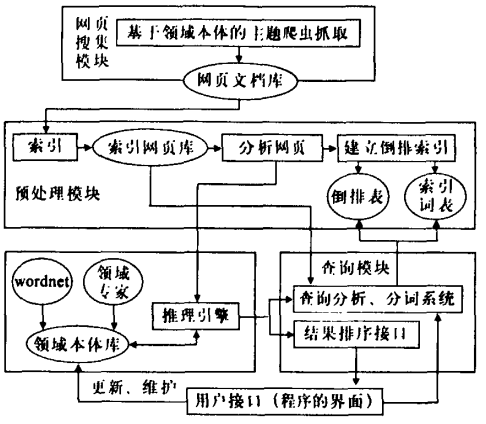

图1　基于领域本体论的垂直搜索模型

2.1　领域本体库的构建

本体是实现领域知识共享、集成和重用的基础。本体的目标就是获取、描述和表示相关领域的知识，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（术语）和词汇间相互关系的明确定义。本框架利用WordNet与领域专家来构建本体。

WordNet是由普林斯顿大学的George Miller等人开发的电子词典系统[3-4]。WordNet中的每个单词都具有一个或若干个含义，而每个含义都有与其它不同的同义词集，由不同的连接关系组成不同的单词集合。

本文根据WordNet系统的结构特点，利用WordNet中语义关系与OWL本体语言中的语言成份的对应关系，可以直接将处理自由文本中单词信息的WordNet资源映射成OWL本体的定义。通过这种方法将WordNet系统中的Synset等概念映射到OWL概念和关系上，将WordNet电子词典系统转换为OWL格式的本体库系统。由于整个WordNet系统的规模十分庞大，而且WordNet词典是与领域无关的词汇资源，所以只要根据需求转换与具体应用密切相关WordNet中的某些子部分。然后加入领域专家根据实际应用对生成的本体库作进一步的细化和扩充。

2.2　基于领域本体的主题爬虫模块设计

2.2.1　主题爬虫的概念与关键问题

网络爬虫模块的目的是按照一定的搜索策略在网络中发现与领域相关的新的网页信息，这个模块的关键在于主题爬虫的设计。本模块不但利用关键字，爬行算法中还依靠概念和关系等高层次的背景知识来对比搜索网页的文本。主题爬虫的基本思路是按照事先给出的主题，分析超链接和已经下载的网页内容，预测下一个要爬行的URL，保证尽可能地多下载与主题相关的网页、尽可能少下载无关网页。因此，主题爬虫需主要解决以下3个关键问题：

（1）判断一个已经下载的网页是否与主题相关。对于已经下载的网页，因为可以知道它的文字内容，所以可用传统的文本挖掘技术来实现。

（2）决定URL的访问次序。许多主题爬虫是根据已下载的网页的相关度，按照一定的原则，将相关度进行衰减，分配给该网页中的超链接，而后插入到优先级队列中。此时的爬行次序就不是简单的以深度优先或者广度优先为序，而是按照相关度大小排序，优先访问相关度大的URL。不同主题爬虫之间的主要区别在于如何决定URL的爬行次序。

（3）提高主题爬虫的覆盖度。这个问题要解决的就是如何穿过质量不够好（与主题不相关的）网页，得到我们所感兴趣的网页，从而提高主题资源的覆盖度。

2.2.2　基于本体的上下文主题描述

主题爬虫在爬行中要不断地去下载的网页进行判断，判断是否与主题相关，同时还要判断该网页里的其他链接是否与主题有关，并选择相关主题网页进行爬行。所以如何进行主题的描述非常关键。

当前有3种主题描述方面：基于关键词的主题描述、基于自然语言格式文本的主题描述和基于层次分类法的主题描述。前面两种方法都是假设主题之间是相互独立的，但是不能体现出多主题之间的关系。第3种方法是以一种树型结构的分类法ODP[6]（Open Directory Project，开放式分类目录搜索系统）为表示主题，文献[5]证明了ODP能进一步提高主题爬虫发现主题相关的链接，并进行爬行，但是分类法没有很好地体现各主题之间的语义关系，因此可能会出现语义方面相关的主题没有办法被爬行到的情况发生。本体在语义描述方面有它特有的优势，通过本体进行上下文

的主题描述不仅能够实现用分类法进行描述的好处，还能进一步将语义相关的主题加入爬行中去。将基于本体的上下文主题描述应用于主题爬虫判断主题相关中，并且防止了其他爬行中出现“隧道”现象，而且进一步提高了主题爬虫的查准率和查全率。

2.2.3 基于本体的主题爬虫框架

根据上下文主题描述，实现爬虫的语义爬行。在这里提出了一个基于本体的主题爬虫框架，见图2。这个框架主要包括如下功能：

（1）用户初始化统一资源定位符（URLs，Uniform/Universal Resource Locator），进行领域网页信息的搜索。

（2）根据基于本体的上下文主题描述，计算网页、网页里的链接以及主题的相关度。

（3）根据计算出的相关度扩充主题爬行目标。

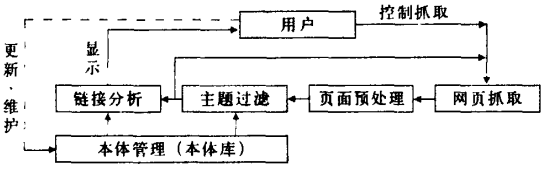

图2 基于本体的主题爬虫框架

（4）这里还将根据爬行出的网页进行分析，进一步扩展本体库。

一个基于本体的主题爬行过程包含两个内在的循环：本体循环和爬行循环。爬行循环从网页抓取开始，经过页面预处理、主题过滤、链接分析等，不断地从网络中取得与主题相关的信息，根据链接分析出来的URLs再进行页面的抓取工作。主题过滤与链接分析都是由本体管理提供语义分析，同时根据在实际爬行过程中出现的高频率新概念又对本体库进行更新与维护，这就是本体循环。最初的本体库是WordNet与领域专家相结合构建的。在这个模块中，用户可以控制抓取页面。

2.3 预处理模块

该模块主要完成对爬虫下载的网页建立索引库的功能，并分析网页，建立对应的倒排表与索引表。建立索引库是这个模块的关键。索引库的建立是通过理解爬虫下载的页面，从中抽取索引项生成文档的索引表与关键字索引表来完成的。在生成索引时，本模块通过建好的领域本体，将相关的关键字根据上、下位的关系在索引网页库中进行标识，这样有助于提高搜索速度。

2.4 查询模块

对用户提交的查询请求，根据领域本体进行查询分析并分词，将关键字提交到倒排表与索引表进行搜索。同时，对于查询出来的结果，根据与主题的相关度进行排序处理。

2.5 用户接口

用户接口用于输入用户查询、显示查询结果。基于本体的垂直搜索的用户接口可以接受用户输入关键字（确定信息与不确定信息），也可以是自然语言或嵌套的模式语言（如搜索引擎方面的论文）。系统使用分词系统对用户的需求进行切分，并辨别是不是属于本领域的概念，然后运用领域本体库与推理机对用户查询的信息语义化，实现语义搜索。

3 结束语

随着网络信息的快速增长，人们对搜索引擎的智能化要求越来越迫切。语义化的搜索引擎、专业化的搜索引擎已成为商业界和学术界共同关注的课题。语义化的搜索引擎将解决通用搜索难以搜到的动态网页和实时信息，而且高度的自动化和智能化。本文基于领域本体，提出垂直搜索引擎模型，利用该模型，可实现专业化信息的语义化，从而满足专业领域用户对搜索的智能化需求。

参考文献：

[1] 余 森，杨 丹，赵俊芹. 垂直搜索引擎的关键技术研究[J]. 软件导报，2007（12）.

[2] Gruber Cf T R. A translation approach to portable ontologies [J].Knowledge Acquisition，1993，5（2）：199-220.

[3] 李 景.本体理论在文献检索系统中的应用研究[M]. 北京：北京图书馆出版社，2005.

[4] 张 红. 语义网中的本体推理及其应用研究 [D]. 吉林：吉林大学，2004.

[5] 陈竹敏. 面向垂直搜索引擎的主题爬行技术研究[D]. 济南：山东大学，2008.

[6] ODP Categories[EB/OL]. http://rdf.dmoz.org/rdf/categories.txt.

[7] 罗 娜. 基于本体的主题爬行技术研究[D].吉林：吉林大学，2009.