软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2015,26(7):1675−1687 [doi: 10.13328/j.cnki.jos.004622]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于本体的智能信息检索系统*

杨月华[1], 杜军平[2,3], 平 源[1]

[1](许昌学院 信息工程学院,河南 461000)

[2](北京邮电大学 计算机学院,北京 100876)

[3](智能通信软件与多媒体北京市重点实验室,北京 100876)

通讯作者: 杨月华, E-mail: yangyuehua0504@126.com

摘 要: 近年来,基于本体的智能信息检索系统已成为智能信息检索系统领域最为活跃的研究方向之一.如何利用本体进一步提高其检索性能和智能性,成为基于本体的智能信息检索系统的主要研究目标.从面向过程的角度,对近几年基于本体的智能信息检索系统的研究进展进行了综述,对其框架、所需本体知识的获取和使用、关键技术、性能评测等进行了前沿概括、比较和分析.最后,对基于本体的智能信息检索系统有待深入研究的难点和热点进行了展望.

关键词: 本体;智能信息检索系统;框架;语义标注;基于本体的查询处理;综述

中图法分类号: TP18

中文引用格式: 杨月华,杜军平,平源.基于本体的智能信息检索系统.软件学报,2015,26(7):1675−1687. http://www.jos.org.cn/1000-9825/4622.htm

英文引用格式: Yang YH, Du JP, Ping Y. Ontology-Based intelligent information retrieval system. Ruan Jian Xue Bao/Journal of Software, 2015,26(7):1675−1687 (in Chinese). http://www.jos.org.cn/1000-9825/4622.htm

Ontology-Based Intelligent Information Retrieval System

YANG Yue-Hua[1], DU Jun-Ping[2,3], PING Yuan[1]

[1](School of Information Engineering, Xuchang University, Xuchang 461000, China)

[2](School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China)

[3](Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia, Beijing 100876, China)

Abstract: Recently, ontology-based intelligent information retrieval systems, aiming to further improve the retrieval performance and intelligence by using ontology, have become one of the hottest topics in the domain of intelligent information retrieval systems. This paper presents an overview of the field of ontology-based intelligent information retrieval systems from a process-oriented perspective, including the system framework, ontology knowledge acquisition and use, key technologies, and evaluation. The prospects for future development and suggestions for possible extensions of the ontology-based intelligent information retrieval systems are also discussed.

Key words: ontology; intelligent information retrieval system; framework; semantic annotation; ontology-based query processing; survey

信息检索系统作为网络信息平台的一个重要组成部分,在网上信息获取方面发挥了不可替代的作用,尤其是在机器学习、自然语言处理、知识表示和推理等人工智能技术被应用到信息采集、信息索引、查询处理、信息检索和排序、结果反馈等基本环节后,使得检索性能得到不断改善,信息检索系统已成为人们获取信息必不可少的工具,相关研究也取得了很大进展.但网上信息资源不断增多,内容和形式多样,信息之间的关联性也比较强,信息组织局部有序而整体无序,这些特点造成用户想要检索到所需要的准确信息仍然十分困难,因为当

* 基金项目: 国家重点基础研究发展计划(973)(2012CB821200, 2012CB821206); 国家自然科学基金(61320106006, 61303232); 河南省教育厅基金(14B520062, 13A413750, 13A413747); 河南省自然科学基金(132300410393); 许昌学院科研基金(2014027)

收稿时间: 2012-09-19; 修改时间: 2013-09-09; 定稿时间: 2014-04-09

前的信息检索系统用户查询表达能力还十分有限,用户查询意图无法得到准确的理解,同时,在用户查询的处理上,要想使信息检索系统能够进行准确的分析,就必须使其具备自然语言理解能力,目前对自然语言的处理虽然已从语法阶段上升到语义阶段,但是仍然限制在一些规范的语句和语法范围内,不可避免地会造成语义上的丢失,而且大都没有考虑到对知识的整合,所以经常会输出大量无用的垃圾信息,智能性也不高.

本体(ontology)作为一种知识建模工具,自被提出以来就引起了国内外众多科研人员的广泛关注.由于本体能够很好地描述概念以及概念与概念之间的关系,具有良好的概念层次结构和对逻辑推理的支持,因而将本体引入信息检索系统后,能够为改进信息检索性能提供组织形式和语义上的保证:首先,在信息检索系统中引入本体后,能够最大限度地保留关键词之间的语义关系,大大增强了用户的检索需求表达能力,使信息检索工具更加人性化,查询变得更加方便、直接、有效.此外,基于本体可以进行语义查询扩展,从而检索出与用户查询语义相关的信息;再者,本体通过公理和属性描述概念之间的逻辑关系和规则,提供了对推理的支持,从而可以在一定程度上实现智能检索[1].因此,研究引入本体的智能信息检索系统意义极大.

近年来,国外许多大学和研究机构已对基于本体的智能信息检索系统理论、方法及应用展开了深入的研究,如美国的马里兰大学[2,3]、路易斯维尔大学[4,5],英国的开放大学[6,7],韩国的高等科技院[8,9]、国防大学[10],西班牙的哈恩大学[11,12]、阿尔卡拉大学[13],土耳其的中东技术大学[14],希腊的雅典大学[15],IBM 研究院[16,17],OntoText 语义研究实验室[18,19]等.他们主要从以下几个方面对基于本体的智能信息检索系统进行研究:(1) 基于本体的用户查询处理方法;(2) 语义标注和索引方法;(3) 基于本体的信息检索方法和模型;(4) 信息检索系统所需本体知识的获取和扩充;(5) 本体推理技术在信息检索中的应用;(6) 基于本体的智能信息检索系统框架;(7) 基于本体的智能信息检索系统评测;(8) 基于本体的智能信息检索系统的应用.取得的代表性研究成果有:文献[10]提出一种混合的查询处理方法,借助所构建的领域本体分别采用查询重写的方法和推理的方法处理频繁变化的知识和不变化的知识,实现了有效的信息检索;文献[15]提出一种面向世界新闻领域信息的语义检索方法,借助所建立的世界新闻本体和领域启发式规则获得了较高的查准率;文献[20]提出一种基于两级不确定模糊本体的智能信息检索方法,在本体构建中引入了模糊逻辑,解决了使用一般的本体不能充分表示领域中不确定性信息的问题;文献[21]提出一种基于语义标注产品族本体的产品信息检索方法,当用户查询涉及产品的多个方面时,采用该方法能够获得较好的检索结果.目前也有了一些初步的实验结果,准确率提高幅度在 5%~20%之间[6,15].

国内也有一些机构和学者针对基于本体的智能信息检索系统进行了相关研究.例如浙江大学[22,23]、大连海事大学[24,25]、复旦大学[26]、中国人民大学[27]、北京邮电大学[28].但是,与国外相比,国内对基于本体的智能信息检索系统的研究起步较晚,目前取得的成果较少,典型的有:文献[24]提出了一种基于 RDF 和模糊本体的语义信息检索方法,实现了大学科学研究信息的模糊语义检索,在该方法中采用了 RDF 来表示科学研究知识;文献[25]基于 RDF 和模糊本体实现了交通信息的语义检索,其中,交通信息用 RDF 表示,建立了模糊本体,基于本体概念间的顺序关系、等价关系、包含关系以及互补关系实现了语义查询扩展,从而能够返回更多满足用户需求的结果文献;文献[29]提出了一种基于本体和标记的语义检索机制,标记的聚类采用语义相似度计算方法实现.

目前,信息检索系统领域的国内外综述文献极少涉及基于本体的智能信息检索系统.鉴于基于本体的智能信息检索系统的重要研究意义和实用价值,有必要跟踪学习和总结该领域现阶段的研究成果,并深入分析和预测其发展趋势,以期能够更好地指导后续的研究工作.

本文第 1 节对基于本体的智能信息检索系统进行概述.第 2 节~第 5 节重点介绍基于本体的智能信息检索系统理论与方法研究现状.其中,第 2 节介绍基于本体的智能信息检索系统所需的本体知识的获取与本体扩展方法,第 3 节介绍基于本体的语义标注方法和工具现状,第 4 节对基于本体的查询扩展技术和本体推理机进行归类和对比分析,第 5 节总结基于本体的智能信息检索系统的性能评测数据和指标.第 6 节对有待深入研究的问题和发展趋势进行展望.最后是结束语.

1 基于本体的智能信息检索系统概述

本节首先对智能信息检索系统进行概述,然后介绍基于本体的智能信息检索系统框架,最后对国内外基于

本体的信息检索方法研究情况进行介绍.

1.1 智能信息检索系统

知识的获取与表示、自然语言处理、机器学习、知识推理等人工智能技术是随着时代对社会智能化需求的增加而发展起来的,而人工智能与信息检索的结合则是人们对信息获取智能化的有益尝试.在信息检索系统中融入人工智能技术,使传统的信息检索系统能够更准确地理解用户的查询需求、获得更好的检索性能、智能化程度更高[30,31].总之,人工智能技术的引入,将使传统的信息检索系统向着更加智能化的方向发展.目前,把引入了现代人工智能技术、具有一定程度的智能特征的信息检索系统称为智能信息检索系统.智能信息检索的目标是:在对用户查询内容的处理、信息获取、索引、检索和排序等方面实现检索的智能化,代替人类完成繁杂的信息收集、过滤、分析和处理任务.目前,智能信息检索系统按研究的侧重点不同可以分为以下 3 类.

(1) 语义检索系统

将信息检索从目前基于关键词层面提高到基于知识(或概念)的层面,信息检索建立在概念及概念间关系的基础之上,主要研究如何对用户输入进行语义分析、如何把用户提交的查询通过语义理解和计算转换成语义概念,从语义上真正理解并准确描述出用户的查询需求;为了充分体现信息间的关联,应如何对检索系统所需的知识进行表示;以及通过对知识库的查询和推理,得出用户能够直接加以利用的信息.基于本体的智能信息检索系统就属于语义检索系统[32,33].

(2) 跨媒体信息检索系统

允许用多种媒体信息表达用户查询的需求,同时能够输出多种媒体类型的查询结果,检索功能非常强大,应用范围更广,而且更加符合人类的思维方式,不但能够丰富计算机的服务,更是计算机功能的一种延伸.但是目前,国内外尚未形成较为成熟的跨媒体信息检索算法和技术,跨媒体信息检索效果欠佳[34,35],在跨媒体信息统一表示、跨媒体数据语义标注和内容识别以及跨媒体信息检索结果的排序和相关反馈等方面都有待进行深入研究.

(3) 个性化信息检索系统

能够为具有不同信息需求的用户提供个性化检索结果,即,对不同用户提交的同一种查询词语能够按照不同的用户需求生成不同的检索结果.个性化信息检索系统主要研究如何通过智能代理不断学习、适应信息和跟踪用户兴趣动态变化,如何基于 Web 挖掘技术在网络中提取用户感兴趣的信息或者更高层次的知识和规律,如何通过推送技术使服务器自动通知用户系统中哪些信息是最新更新的,并自动搜集用户可能发生兴趣的信息通过智能代理提交给用户,从而提供个性化信息检索服务.

智能信息检索系统领域的主要国际学术会议和期刊有 SIGIR,IJCAI,UCAI,WSDM,TREC,WWW,AAAI,ECAI,ECIR,ICML,ECML,ISWC,ECWC,AIRS 和 Information Retrieval,IEEE TKDE,IEEE Intelligent Systems,Web Semantics,Knowledge-Based Systems,Knowledge and Information Systems,Data & Knowledge Engineering,Information Processing & Management,Expert Systems with Applications,Knowledge Engineering Review,JSEE,Information Systems,Journal of Artificial Intelligence Research,International Journal of Software Engineering and Knowledge Engineering 等.

1.2 本体和智能信息检索系统

本体在信息检索领域的应用研究始于 20 世纪末~21 世纪初,由于本体提供了一种对信息和知识进行规范化描述和组织的方法,在构建智能信息检索系统的过程中,进行语义标注所使用的词汇、术语以及描述被标注资源之间关系所使用的词汇都可以通过本体给出,当在检索中需要使用推理工具进行推理时,所有资源之间的关系以及对属性的约束等条件也可以由本体给出.因此,通过分析和总结可以发现,本体能够在智能信息检索系统的以下环节发挥作用:

(1) 语义标注:根据本体对检索对象进行语义标注,即,通过分析文档的特征词汇(代表文档内容的词汇、关键字)建立词汇与概念之间的映射关系,从而把文档跟本体关联起来,把文档隐含的语义信息显式地表达出来;

(2) 基于本体的索引:对文档建立基于本体的索引,就是在对文本内容特征提取的基础上生成索引,在索引中反映出文本标引词之间的内在联系,从而在标引过程中过滤文本存在的语言歧义.基于本体的索引由通过语义分析得到的揭示文本内容的特征词汇及其关系构成,通过语义标注完成;

(3) 基于本体的查询扩展:主要是借助本体丰富的语义关系及其推理机制对用户的查询进行语义层次的扩展,从而使检索系统能够更好地理解用户查询意图,帮助用户明确查询目标,能够在一定程度上弥补用户查询表达不够充分的缺陷,因此有助于提高信息检索系统的查全率和查准率.

近年来,研究人员已经实现了一些基于本体的信息检索系统,按照使用本体的方式可将其分成以下两类:

(1) 提供一个低层次的信息空间,本质上类似于传统的分类法和叙词表,利用基于概念层次和概念间关系的查询扩展方式,改进传统搜索的效率,但并没有充分地利用本体语言的优势.例如,利用 WordNet 或基因本体 GO、医学领域本体 MeSH 等进行的检索.AT&T Lab 建立了一个应用本体技术的信息检索系统 FindUR(http://www2.research.att.com/~dlm/findur),通过使用描述逻辑系统规定的语法,表达了 WordNet 中定义的词汇间的同义、上下位关系,以此为基础实现查询扩展.IBM 开发的简易语义搜索(easy semantic search,http://ibm.csdn.net/ISN_J.aspx?action=JMP&pointid=2278)系统能够使用户利用简单的语义搜索功能,通过广为接受的关键字查询模式查询概念,但实际上依然是利用同义词或正则表达式来扩展查询的.

(2) 采用包括上千万条本体实例、概念和任意复杂度关系的知识库进行检索,将信息空间认为是无歧义、无冗余的格式化本体知识.通常的做法是:首先进行前期处理,通过自动、半自动或手动的方法将文档进行标注,映射成本体中的实例,再将用户输入的自然语言或形式化查询语言转化为系统查询语言,通过本体推理与查询得到所需要的实例.最后,将实例对应的文档返回给用户.例如,欧洲科研信息系统 AURIS-MM (http://derpi.tuwien.ac.at/~andrei/documents/AURISMMIntro.htm)以及 OntoText 语义研究实验室开发的 KIM 平台(http://www.ontotext.com/kim)、英国曼彻斯特大学计算机科学系和生物科学学院的 TAMBIS[36]等,都是基于本体知识库的智能信息检索系统的代表.实际上,也就是把传统的文档转化为实例进行搜索.

1.3 基于本体的智能信息检索系统的框架

基于本体的智能信息检索系统是语义检索系统的一种,如何从面向过程的角度研究和实现语义检索系统,是人们普遍关心的问题之一,不同研究人员持不同的观点.例如:文献[6]提出的语义检索系统框架包括查询处理、语义索引、检索、排序和语义网关几个组成部分,语义网关负责收集、分析和访问语义网上的语义知识;文献[37]则认为,语义检索框架包含 4 部分内容:知识提取、语义学习、知识匹配和检索;文献[1]中实现的语义检索原型系统包括信息智能获取、知识建模、语义标注和信息检索几个阶段.可以看出,知识的获取和使用是语义检索系统的关键组成部分.基于本体的智能信息检索系统作为一种典型的语义检索系统,应当遵循语义检索系统的流程,不仅要考虑本体知识的获取,还要重点研究如何将本体引入信息检索过程中.目前,研究人员也提出了一些框架,例如:文献[14]提出的系统框架包括信息采集、信息抽取、本体和扩展、语义检索几个环节,本体知识是通过信息抽取和推理来获取和不断加以扩充的,并用于索引的建立中;文献[21]提出的系统框架包括本体构建、语义索引、查询处理、检索和排序几个环节,本体知识是通过信息抽取和标注方法进行获取和不断加以扩充的.尽管以上框架能够获得较好的检索结果,但在信息检索过程中还未充分利用本体的优势.我们总结认为,基于本体的智能检索系统应包含信息采集、本体获取和扩展、语义标注、语义索引、查询处理、检索和排序几个部分(如图 1 所示).

(1) 信息采集:使用网络爬虫在互联网上爬取网页并下载到本地磁盘中,然后对网页中的文本内容进行抽取和预处理,为后续进行语义标注等做准备;

(2) 本体获取和扩展:从语义网上获取本体或者根据领域检索需求构建本体,通过本体学习方法自动获取本体中的概念和概念间关系等,或者通过信息抽取和标注的方法构建本体,并对本体库不断进行扩展;

(3) 语义标注:在文本文档中识别出本体中的实体,包括本体中的类、属性、实例等,然后生成相应的标记.与传统的信息检索索引过程类似,只是索引的是本体中的实体,而不是纯关键词.通过语义标注,识别出文本文档中的语义知识;

(4) 语义索引:为文本文档建立基于本体的索引,建立文档和一系列的语义实体和语义关系的连接,给语义实体和关系赋予权重.用领域本体中各种概念的语义关系来描述文档的语义,在语义标注结果的基础上即可完成;

(5) 查询处理:对用户查询进行分词等预处理并与本体的内容进行匹配,基于本体的语义关系和描述逻辑公理进行查询扩展和推理,得到新的更能反映用户查询意图的查询词;

(6) 检索和排序:对新的查询词进行检索,基于语义相关度计算出实例与文档的相关度后,还需要计算查询实例与文档的相似度等,得到各个文档的排序得分;最后,按排序得分高低将排好序的检索结果返回给用户.

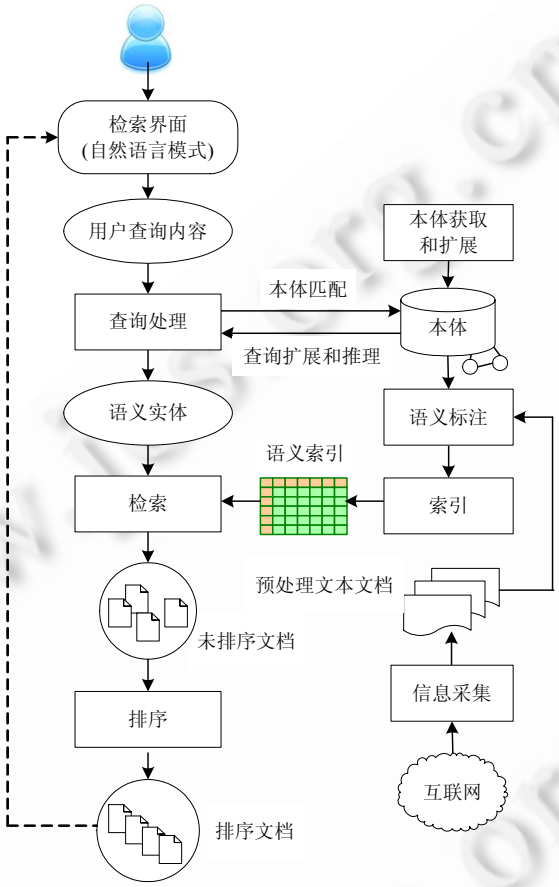

Fig.1 Process-Oriented ontology-based intelligent information retrieval systems

图 1 基于本体的智能信息检索系统面向过程视图

2 本体获取与扩展

基于本体的智能信息检索系统与传统信息检索系统的重要区别在于:前者融入了本体,因此,选择使用什么本体、如何获取检索系统所需的本体知识成为基于本体的智能信息检索系统应考虑和重点研究的问题之一.

国外对本体的研究项目很多,已经建成了许多可以使用的开源本体知识库系统,如较为成熟的通用本体库系统 WordNet,Dbpedia,Cyc,以及生物医学领域本体、企业领域本体.在 Semantic Web 网站上也提供了一些可以使用的本体,网址为 http://semanticweb.org/wiki/Ontology.国内对这方面的研究十分有限,与国外存在很大的差距,比较著名的是通用本体库系统 HowNet(知网,http://www.keenage.com/html/e_index.html),是由中国科学院董振东教授开发的一个以英汉双语所代表的概念以及概念的特征为基础、以揭示概念与概念之间以及概念所具

有的特征之间的关系为基本内容的常识知识库.

在信息检索系统中可以使用已有的一些本体,但是对于一些特定的领域,由于已有的一些本体缺乏专业领域词语及其关联,在将其应用到具体的领域信息检索中时效果并不理想,所以在很多情况下还需要根据领域检索需求自己构建本体.Perez 等学者认为,一个本体由概念(或类)、关系、函数、公理和实例 5 个基本元素构成,也称为本体的建模原语[38].在实际应用中,不一定严格按照这 5 种元素进行建模,而是要结合特定领域的具体情况进行取舍.近年来,在本体建模方面,国内外学者还提出一些新的方法,例如:Iribarne 等人对开放环境(open environment)的知识使用 UML(unified modeling language,统一建模语言)进行了本体建模,给出了将 UML 元素转换为 Web 本体语言的方法[39];Lee 等人基于区间二型模糊集合(T2FS)建立了二型模糊本体模型(type-2 fuzzy ontology,简称 T2FO),用于表示糖尿病人饮食推荐领域的知识[40];Razmerita 提出了一种通用的基于本体的用户行为建模框架(OntobUMf),并给出了其组成部分及用户行为的本体建模过程[41];Liu 等人对中国古代建筑知识进行了本体建模,并实现了一个能在多种建筑中识别出不同元素和风格的建筑知识系统[42].本体知识现在普遍采用 OWL(Web ontology language,Web 本体语言)来进行形式化表示.OWL 已经成为目前公认的开放网络环境下知识表示的规范,是语义网技术中知识表示的标准[43].在信息检索系统中,通过 OWL API (http://mac.softpedia.com/get/Utilities/OWL-API.shtml)来使用本体知识.

近年来,国内外很多研究人员都在基于本体的智能信息检索系统中使用了自己构建的本体,例如:Kara 等人构建了足球本体用于足球领域的信息检索中[14],Lim 等人构建了产品族本体用于多层面产品信息检索中[21],崔金栋在网格信息检索系统中使用了自己构建的网格本体[44].但是,他们在使用自己构建的领域本体时基本上都忽略了一个问题,即,对其有效性的评估.关于本体的评价方法国内外目前都还处于研究阶段,缺乏被广泛认可的评价理论体系和评价方法体系,评价集中于概念、属性、关系等方面,综合评价体系尚未真正建立起来,没有形成权威性的标准.因此,在后续研究中提出有效的本体评价方法是非常必要的.

本体知识获取主要指的是构成本体的概念、概念间语义关系等的自动获取,相关的技术也称为本体学习(ontology learning)技术,即,利用统计和机器学习等方法自动或半自动地从已有的数据资源中获取期望的本体概念和概念间关系.国外的研究人员已经提出了许多有效的方法[45,46],并开发了一些相关的工具,如 Hasti, OntoLearn,Text-To-Onto,OntoBuilder,OntoLiFT[47],这与以英语为代表的西文进行分词处理较为容易有关.国内在本体方面的研究刚刚起步,并且研究重点主要集中在如何利用本体来解决语义问题上.而专门针对本体学习方面的研究成果还比较少,还没有一个能够支持中文的本体学习工具.由于中文语法相对比较复杂,中文本体学习技术确实存在很多困难,单纯依靠统计的手段或现有的与语言无关的算法很难获得令人满意的学习结果,必须结合中文自然语言处理领域的研究成果,使用一些基于规则的方法来改善本体学习的质量,以有效地获取本体知识,实现本体的自动扩展.

3 语义标注

语义标注是信息检索系统中一个非常重要的环节,进行语义标注的目的,就是从文档中提取出计算机能够理解的语义.因此,标注的好坏将直接影响到最后的检索效果.由于本体能够提供对领域知识的一致理解,有效给出词汇和词汇间相互关系的准确定义,基于本体进行语义标注一方面能够极大地提高信息检索的准确性,另一方面还能够消除用户查询的歧义问题.例如,一个使用本体进行语义标注的文档集合能够分辨出用户查询中用同一个词“计算机网络”表示的技术和图书名称,这是因为在本体的支持下两者被标注为不同的概念.因此,通过本体进行语义标注有助于检索系统更加准确地理解用户查询意图.

目前,国外研究人员在基于本体的语义标注方面已经取得了一些研究成果[48,49],开发了一系列工具(http://annotation.semanticweb.org/tools),如自动标注工具 MnM(http:kmi.open.ac.uk/projects/akt/MnM/),Melita (http:nlp.shef.ac.uk/melita/)等,主要借助自然语言处理、信息抽取和机器学习技术来实现语义标注,但是都不支持最新的本体语言 OWL.此外,一个页面上的词汇往往涉及多个本体中的概念,少数原型,如 SMORE,允许用户使用多个本体标注页面,多数原型不支持同时打开、浏览多个本体并使用多个本体来标注页面;而且,语义标注过程中本

体查询、辅助推理支持以及元数据产生的自动化程度还不够.至今的所有原型都是英语运行环境,不支持多语言.

近年来,研究人员又提出一些新的语义标注方法,例如:Rajput 等人提出一种基于贝叶斯网络的语义标注框架[50];Nguyen 等人提出一种基于朴素贝叶斯和关联规则,并结合本体结构图和交互网络的语义标注方法[51];Smine 等人建立了可自动标注含有语义元数据文本的模型[52].国内,关于基于本体的语义标注方法的研究较少,相关文献不多,取得的研究成果较少[53−55].这说明,我国在基于本体的语义标注领域的相关研究还需加强.

4 基于本体的查询处理

在信息检索领域中,用户输入的查询往往与文档中的目标词不相匹配,导致信息检索系统无法返回符合用户查询请求的结果集.如何对用户查询词进行处理以提高信息检索的准确率,是一个开放的问题.查询扩展是其中一种可行的解决方法[56,57],其基本思想是:在原始查询词的基础上加入与用户查询词相关联的词,以组成新的更长、更准确的查询词,可以在一定程度上弥补用户查询信息不足的缺陷,解决查询词不匹配的问题,改善信息检索的效果.但是传统的查询扩展方法存在难以克服的缺陷,即:它们都忽略了语义层面上的扩展,不能从根本上解决用户查询意图与检索资源之间的语义匹配问题.为此,学者们又提出了基于词典的查询扩展方法.基于词典的查询扩展方法的基本思想是:在一个词组或词语集合中找出与原始查询词语义相关的词语,如同义词、反义词作为扩展词实现查询扩展.Nawab 等人在检索可能的抄袭文档的应用中就采用了基于词典的查询扩展方法,并通过实验验证了这种方法较其他常用的查询扩展方法的优越性[58].目前主流的搜索引擎,例如 Google、百度已经实现对原始查询词的同义词或相关词的扩展.另外,还有一些信息检索工具使用了 WordNet,HowNet 等本体作为语义词典来获取原始查询词的同义词、上下位词等,从而实现查询扩展.

基于本体进行查询扩展的思想最早是由 Voorhees 在 1994 年提出来的[59],主要是借助本体中明确形式化的概念定义,利用本体中丰富的语义关系来进行查询扩展.在此之后,基于本体进行查询扩展的研究不断深入.目前,基于本体的语义查询扩展方法主要是把原始查询映射为本体中的概念,根据本体中概念间的各种关系,利用一定的技术提取出查询语义及其语义关联词,从而得到比原查询更长的新查询词,进而改善查询效果.根据用户查询的不同,可以基于本体从以下几方面进行查询扩展[38]:

(1) 同义扩展:这是最基本的扩展类型,即,通过本体定义获取用户查询中概念的同义词;
(2) 属性扩展:从本体中获取用户查询中与主语相关的属性概念;
(3) 层次扩展:用于确定某一个体所属的类或某一类的层次结构,进而可进行属性扩展、扩大或缩小检索范围、提供相关性检索以及利用概念之间的层次关系进行简单的推理;
(4) 公理扩展:利用本体中定义的公理对用户查询进行扩展,从而明确用户查询内容,限制检索输出内容.通过公理扩展可获得关于用户查询的语义信息,如概念的层次结构、属性的定义域和值域、属性的性质以及属性的基数限制和值限制等;
(5) 规则扩展:利用本体中定义的规则扩展用户查询,并基于规则进行推理.

近几年来,基于本体的查询扩展方法已成为一个研究热点并取得了较多的研究成果,具有代表性的有:Díaz-Galiano 等人通过使用医学本体 MeSH 扩展用户查询词,极大地提高了多式信息检索系统的性能[11];Segura 等人提出的基于领域本体的查询扩展方法中使用了多种关系类型,包括通用本体关系、特定领域的本体关系和传统的术语间关系等,并通过实验验证了在准确率相当时,采用基于本体的查询扩展方法能够获得更高的查全率[13];Liu 等人建立了基于领域本体的查询扩展模型,给出了 5 种查询扩展方法并应用于语义检索系统中[60];Chen 等人提出的基于本体的查询扩展方法通过构建概念格和计算用户查询与概念间的相似度来寻找最优的扩展词[61].

尽管基于本体的查询扩展方法取得了一定的进展,但许多方法是将查询映射到本体中的概念,或者说,它们所使用的本体更像是简单树形结构的词表,并没有属性和实例概念,能表达的也主要是上下位关系、同义关系,没有考虑到属性、层次、公理和规则的扩展.因而,这样的本体并不能扩展出很多语义关联词.此外,在某些情况

下,基于本体的查询扩展方法也可能会减弱某些查询的性能.例如,当新查询词与原查询词的相关性不大时,可能导致“查询漂移”,最终导致“主题漂移”.因此,在将来的查询扩展研究中,一方面要考虑更多的语义关系和从多个角度进行扩展,另一方面应提出有效的查询扩展词推荐策略,对扩展出来的查询扩展词进行过滤,以保证最后得到的扩展词是有效的.

此外,在对用户查询进行处理时,还可以基于本体进行推理,获得一些隐含的语义关系,以便使用户查询得到更准确的理解.目前,在信息检索系统中进行的本体推理主要是使用一些推理机来完成的,分别由查询解析器和本体解析器解析用户查询和本体,通过推理引擎完成推理.在信息检索系统中进行的本体推理有以下几类:

(1) 类(概念)/实例关系推理.给定本体 R,C 是 R 中的一个类(概念),I 是 R 中的一个个体.可以通过推理判断一个个体是否是 C 的实例;判断在 R 中 C 的所有实例;判断 R 中个体 I 是哪些类的实例;判断与某个实例有特定关系的实例或判断两个实例之间的关系;

(2) 类(概念)的关系推理.给定两个概念 $C1$ 和 $C2$,判断它们之间的关系,包括子类、成员、部分等;

(3) 在类的体系架构中进行推理.给定类 C,返回在 R 中 C 的所有或相关的超类,或者返回在 R 中 C 的所有或相关子类;

(4) 属性关系的推理.属性与类(实例)推理相似,包括属性/实例关系、属性体系结构、属性包含和属性可满足性等.

近年来,随着本体研究的深入,针对本体推理方面的研究也逐渐开展起来[62−64],研究人员已经开发了一系列可用的本体推理机[65],按采用的推理方法可以分成 4 类.

- 基于传统描述逻辑的推理方法能够提供可判定的推理服务,但推理功能仅限制在分类和包含.典型代表有,Racer(http://www.sts.tu-harburg.de/~r.f.moeller/racer/),Pellet(http://clarkparsia.com/pellet/)和 FaCT++(http://owl.man.ac.uk/factplusplus/).这些本体推理机都是基于传统的 Tableau 算法设计并实现的,同时引入了许多 Tableau 算法的优化技术,因而它们的推理效率很高;
- 基于规则的推理(rule-based reasoning,简称 RBR)是基于领域专家知识和经验的推理,将专业的知识和经验抽象成为推理过程中的推理规则,是一种基于谓词逻辑的产生式系统,比较直观,对推理过程易于理解,同时推理效率比较高.目前已出现了很多由 OWL 到规则的现成转化工具,典型的有 Jess(http://www.jessrules.com/)和 Jena(http://www.hpl.hp.com);
- 逻辑编程的方法基于演绎数据库技术实现推理,典型的项目有基于 XSB 演绎数据库技术实现的 F-OWL,支持部分 OWL Lite 的推理.另外,德国卡尔斯鲁厄大学的 KAON2(http://www.aifb.kit.edu/web/KAON)也是采用该技术方法实现的典型例子;
- 基于一阶谓词证明器的方法实现对 OWL 的推理也非常方便,因为将 OWL 声明语句转化为一阶逻辑很方便,本体推理机 Hoolet(http://owl.man.ac.uk/hoolet)就是利用了 Vampire 一阶谓词证明器来实现本体推理的.但仅就 OWL 本体上的推理而言,如果只使用定理证明器,一阶逻辑工具对推理支持的完备性远不及描述逻辑工具.

5 基于本体的智能信息检索系统的性能评测

性能评测(evaluation)对于检验基于本体的智能信息检索系统的性能和发现其存在的问题来讲十分重要,是不可或缺的步骤,但目前还没有用于基于本体的智能信息检索系统的标准评测框架.已有的一些评测方法都是以用户为中心的,不可扩展,很难复用.因此,目前主要还是采用传统的信息检索系统评测方法进行评测,只是在评测时引入了一些被公认的涵盖了涉及到的查询主题的本体.早期检索系统评测最著名的研究是 Cleverdon 在 20 世纪 50 年代末期开始进行的 Cranfield 实验,它开创了以测试集及评测指标来评测系统的模式[66].目前,国际上比较著名的 TREC 会议(http://trec.nist.gov/)在信息检索评测领域起到了很好的示范作用,提供的评测语料包括:VLC track(VLC,VLC2,WT2g,WT10g),300 个主题,100GB;Web track(.GOV 语料),550 个主题,18GB;Terabyte track(.GOV2 语料),1 800 个主题,400GB.跨语言评测论坛 CLEF(http://clef.isti.cnr.it)、NTCIR 会议(http://

research.nii.ac.jp/ntcir)等也都专注于信息检索评测.国内的学者参考 TREC 多年以来的成功经验,也构建了一些用于中文搜索的评测语料(中文 Web 信息检索论坛:http://www.cwirf.org/),比如:2005 年,北京大学关于信息检索的国家 863 项目中构建的评测语料集是 CWT100G,提供了 50 个查询主题;SEWM 系列会议的信息检索评测更侧重于 Web 信息检索的评测,语料库有 CWT100G,CWT200G 等几个版本;CIRB030 评测的语料是纯文本格式的,用 XML 做了一些标记,提供了 42 个查询主题;搜狗实验室在 2008 年也推出了 3.0 版本的互联网搜索语料,收集了超过 1.3 亿网页数据,总存储规模达到 5TB 以上.除了包括所有的网页原始数据外,还包括了提取出的这部分网页之间的链接关系数据以及 PageRank 数值数据.与该数据同时推出的还包括规模庞大的用于网络信息检索评测的标准评价集合,评价集合规模超过 10 000 个查询.

基于本体的智能信息检索系统的目标也是在消耗较少的情况下尽快返回准确而全面的结果,因此与传统的信息检索系统一样,主要从效率、效果等方面进行评价.在效率方面,主要是对时间开销、空间开销和响应速度进行测试;在效果方面,主要考虑检索返回的文档中有多少相关文档、返回了多少正确的文档以及靠不靠前等,有时还需要对覆盖率、访问量以及数据更新速度进行评价.采用的评价指标可以分为 3 类:对单个查询进行评估的指标,包括召回率、正确率、*F* 值、*E* 值、平均准确率 *AP*、*Precision@N*、Bpref、NDCG;对多个查询进行评估的指标,包括 MAP,GMAP,MRR;面向用户的评价指标,包括覆盖率(coverage)、出新率(novelty ratio),覆盖率表示系统找到的用户已知的相关文档比例,而出新率则表示系统找到的用户已知的相关文档比例.

6 基于本体的智能信息检索系统研究发展的难点与热点

在精准化和智能化信息检索需求的驱动下,随着本体技术、自然语言处理、机器学习、知识推理等人工智能技术的发展,基于本体的智能信息检索系统已取得一定的进展,但仍然是一个充满问题与挑战的新兴研究领域,可以深入并可能取得成果的方向有很多,主要包括:

(1) 信息检索系统中采用的本体的有效性评估

为了提供信息检索所需的知识,不同领域的学者建立信息检索系统时都采用自己所建立的领域本体,领域本体的有效性难于评估,而且也不通用,因此需要建立一些公用的领域本体并提出有效的本体评价方法.关于本体的评价方法国内外目前都还处于研究阶段,还没有提出被广泛认可的评价体系,没有形成权威性的标准.另一方面,对于本体在信息检索系统中所起的作用在已有的文献中也基本上没有进行评估,在这方面可以通过将引入了本体和未引入本体的信息检索结果进行比较来了解.

(2) 新的本体知识的获取和使用

在基于本体的信息检索系统中,本体在领域专家的帮助下通过手工或者自动化的方式建立,这在很大程度上依赖于现有的词汇知识.如果本体中不存在用户查询对应的概念或者实例,就不可能查到含有它们的文档.因此,获得新词、生成新实例并将它们及时加入本体中,是保证信息检索系统准确性的一项重要工作.但是,目前在基于本体的信息检索系统领域还没有明确提出解决以上问题的有效方法.一方面,由于基于本体的信息检索理论还不够成熟,本体论与传统 IR 技术的结合有待进一步研究;另一方面,本体中包含很多复杂的语义关联,将新知识自动扩充到本体中也是一项不容易的工作.

(3) 基于本体的语义标注方法

基于本体的语义标注就是借助本体将检索文档中的语义提取出来的过程,对提高信息检索系统的查全率和查准率具有重要作用,是一个值得关注的研究方向.目前的多数标注方法在语义标注过程中能够较好地完成类型标注,识别出文本中的实例,但却很难抽取出实例间的关系;通过对自然语言语句的句法结构进行分析,可以从句法关系出发,寻找实例间的属性关系,但是当前方法仅局限于对主谓宾句法关系的分析,而对其他句法关系却未能有效利用;在映射谓语动词到本体属性时,需要利用外部领域知识,但是通用语言本体却不能完全涵盖领域内词汇.此外,多数研究工作是面向英文的,国内的语义标注研究还不够充分,也未开发出有效的本体标注工具,因此,这一方向还有待深入研究.

(4) 基于本体的查询扩展技术

作为改善信息检索效果的一种方法,基于本体的查询扩展技术最关心的是能否扩展出有效的查询词.尽管目前已经取得了一定的进展,但是由于忽略了属性和实例方面的扩展,扩展出的查询词还是有限的.为了更加准确和全面地反映用户查询意图,需要研究如何基于本体进行和用户查询相关的实例和属性的扩展.但是在扩展出更多查询词的同时,还有可能导致最终的检索结果主题偏移,可以考虑在将来的研究中结合自然语言处理技术并引入查询词相关性计算等,对扩展出的查询词进行过滤和推荐.

(5) 系统的评测方法

目前缺乏基于本体的信息检索系统的评估标准和框架,没有提出有效的评估方法.当前文献给出的基于本体的信息检索方法几乎都没有经过严格验证,尽管已建立了大规模的评估数据集、查询集和相关结果集,但是测试集并不适合基于本体的信息检索系统.测试文档来自专业领域也来自通用领域,并且许多文档带有语义标记,这些都是已有的测试集无法提供的.此外,缺乏合理的评估标准对语义标注、基于推理的检索结果等进行有效的评测.

(6) 信息检索系统中的本体匹配和映射机制

随着本体应用的快速发展,本体的数量和规模会变得越来越庞大,越来越复杂,一些领域中的本体可能包括成千上万个概念.由于本体的创建者不同,所使用的本体建模方法不同,即使针对同一领域知识建模,不同的专家开发出的本体也存在差别,不同的人和组织倾向于使用不同的本体.因此,为了在信息检索系统中使用更多的本体,需要对不同的本体进行集成.而目前的匹配和映射算法还无法处理大规模本体集成问题,提出有效的用于信息检索系统中的本体匹配和映射的机制是非常必要的.

7 结束语

随着信息技术与互联网技术的发展,网上的信息资源越来越丰富,造成用户想要找到所需要的信息往往十分困难.智能信息检索系统被认为可以有效缓解这一难题,得到学术界和工业界的广泛关注和应用,并取得许多研究成果.近年来,基于本体的智能信息检索系统将本体引入到信息检索中来,由于本体具有良好的概念层次和语义表达能力,能够进一步改进检索性能,也促进了学科交叉融合的研究,成为智能信息检索系统研究领域最为活跃的分支之一,并被逐渐应用于多个领域.目前,国内外关于基于本体的智能信息检索系统的综述性文献还极少见,因此,本文对该领域的研究进展和发展趋势等进行了归纳总结和预测,包括基于本体的智能信息检索系统的框架、关键技术、性能评价以及研究热点和难点,希望本文所做的工作能够为从事该领域研究工作的学者提供一些有益的信息,从而进一步推进我国在该领域的研究进展.

References:

[1] Dong JX. Semantics-Based Service-Oriented Knowledge Management and Processing. Hangzhou: Zhengjiang University Press, 2009 (in Chinese).

[2] Han LS, Finin T, Joshi A. Schema-Free structured querying of DBpedia data. In: Proc. of the 21st ACM Int'l Conf. on Information and Knowledge Management (CIKM 2012). Maui, 2012. 2090−2093. [doi: 10.1145/2396761.2398579]

[3] Rubin DL, Flanders A, Kim W, Siddiqui KM, Jr CEK. Ontology-Assisted analysis of Web queries to determine the knowledge radiologists seek. Journal of Digital Imaging, 2011,24(1):160−164. [doi: 10.1007/s10278-010-9289-2]

[4] Zhuhadar L, Nasraoui O, Wyatt R. Visual ontology-based information retrieval system. In: Proc. of the 13th Int'l Conf. on Information Visualisation. Barcelona, 2009. 419−426. [doi: 10.1109/IV.2009.47]

[5] Zhuhadar L, Nasraoui O, Wyatt R, Romero E. Multi-Language ontology-based search engine. In: Proc. of the 3rd Int'l Conf. on Advances in Computer-Human Interactions (ACHI 2010). Barcelona, 2010. 13−18. [doi: 10.1109/ACHI.2010.43]

[6] FernÁNdez M, Cantador I, LÓPez V, Vallet D, Castells P, Motta E. Semantically enhanced information retrieval: An ontology-based approach. Web Semantics: Science, Services and Agents on The World Wide Web, 2011,9(4):434−452. [doi: 10.1016/j.websem.2010.11.003]

[7] Allocca C, D'Aquin M, Motta E. Impact of using relationships between ontologies to enhance the ontology search results. In: Proc. of the 9th Extended Semantic Web Conf. (ESWC 2012). LNCS 7295, Heraklion, 2012. 453−468. [doi: 10.1007/978-3-642-30284-8_37]

[8] Lee J, Park JH, Park MJ, Chung CW, Min JK. An intelligent query processing for distributed ontologies. Journal of Systems and Software, 2010,83(1):85−95. [doi: 10.1016/j.jss.2009.06.008]

[9] Jung M, Jun HB, Kim KW, Suh HW. Ontology mapping-based search with multidimensional similarity and bayesian network. Int'l Journal of Advanced Manufacturing Technology, 2010,48(1-4):367−382. [doi: 10.1007/s00170-009-2268-4]

[10] Yoo D. Hybrid query processing for personalized information retrieval on the semantic Web. Knowledge-Based Systems, 2012, 27:211−218. [doi: 10.1016/j.knosys.2011.10.004]

[11] DÍAz-Galiano MC, MartÍN-Valdivia MT, Ureña-LÓPez LA. Query expansion with a medical ontology to improve a multimodal information retrieval system. Computers in Biology and Medicine, 2009,39(4):396−403. [doi: 10.1016/j.compbiomed.2009.01.012]

[12] DÍAz-Galiano MC, MartÍN-Valdivia MT, Ureña-LÓPez LA, Perea-Ortega JM. Using wordnet in multimedia information retrieval. In: Proc. of the 10th Workshop of the Cross-Language Evaluation Forum (CLEF 2009). LNCS 6242, 2010. 185−188. [doi: 10.1007/978-3-642-15751-6_21]

[13] Alejandra SN, Salvador-Sanchez, GarcÍA-Barriocanal E, Prieto M. An empirical analysis of ontology-based query expansion for learning resource searches using MERLOT and the gene ontology. Knowledge-Based Systems, 2011,24(1):119−133. [doi: 10.1016/j.knosys.2010.07.012]

[14] Kara S, Alan Ö, Sabuncu O, Akpinar S, Cicekli NK, Alpaslan FN. An ontology-based retrieval system using semantic indexing. Information Systems, 2012,37(4):294−305. [doi: 10.1016/j.is.2011.09.004]

[15] Kallipolitis L, Karpis V, Karali I. Semantic search in the world news domain using automatically extracted metadata files. Knowledge-Based Systems, 2012,27(3):38−50. [doi: 10.1016/j.knosys.2011.12.007]

[16] Luo G. Design and evaluation of the iMed intelligent medical search engine. In: Proc. of the Int'l Conf. on Data Engineering (ICDE 2009). Shanghai, 2009. 1379−1390. [doi: 10.1109/ICDE.2009.10]

[17] Dolby J, Fokoue A, Kalyanpur A, Kershenbaum A, Schonberg E, Srinivas K, Ma L. Scalable semantic retrieval through summarization and refinement. In: Proc. of the 22nd AAAI Conf. on Artificial Intelligence (AAAI 2007). Vancouver: British Columbia, 2007. 299−304.

[18] Popov B, Kiryakov A, Ognyanoff D, Ognyanoff D, Goranov M. KIM—A semantic platform for information extraction and retrieval. Natural Language Engineering, 2004,10(3-4):375−392. [doi: 10.1017/S135132490400347X]

[19] Popov B, Kiryakov A, Ilian K, Angelov K, Kozhuharov D. Co-Occurrence and ranking of entities based on semantic annotation. Int'l Journal of Metadata, Semantics and Ontologies, 2008,3(1):21−36. [doi: 10.1504/IJMSO.2008.021203]

[20] Hourali M, Montazer GA. An intelligent information retrieval approach based on two degrees of uncertainty fuzzy ontology. In: Proc. of the Advances in Fuzzy Systems, Vol.2011. Article ID 683976, 2011. 11. [doi: 10.1155/2011/683976]

[21] Lim SCJ, Liu Y, Lee WB. Multi-Facet product information search and retrieval using semantically annotated product family ontology. Information Processing & Management, 2010,46(4):479−493. [doi: 10.1016/j.ipm.2009.09.001]

[22] Cai M, Zhang WY, Zhang K. Manuhub: A semantic Web system for ontology-based service management in distributed manufacturing environments. IEEE Trans. on Systems, Man, and Cybernetics—Part A: Systems and Humans, 2011,41(3):574−582. [doi: 10.1109/TSMCA.2010.2076395]

[23] Wang C, Zhuang L, Wu JQ, Zhou F. An ontology-based fuzzy matching approach to semantic retrieval of historical place names. Lecture Notes in Computer Sciences, 2012,7634:19−28. [doi: 10.1007/978-3-642-34752-8_3]

[24] Zhai J, Li JF, Lin Y. Semantic retrieval based on SPARQL and fuzzy ontology for electronic commerce. Journal of Computers, 2011,6(10):2127−2134.

[25] Zhai J, Chen Y, Yu Y, Liang YD, Jiang JT. Fuzzy semantic retrieval for traffic information based on fuzzy ontology and RDF on the semantic Web. Journal of Software, 2009,4(7):758−765.

[26] Dai WH, You Y, Wang WJ, Sun YM, Li T. Search engine system based on ontology of technological resources. Journal of Software, 2011,6(9):1729−1736.

[27] Gu YQ, Yan J, Liu HY, He J, Ji L, Liu N, Chen Z. Extract knowledge from semi-structured websites for search task simplification. In: Proc. of the 20th ACM Conf. on Information and Knowledge Management (CIKM 2011). Glasgow, 2011. 1883−1888. [doi: 10.1145/2063576.2063847]

[28] Yang YH, Du JP, He BW. A novel ontology-based semantic retrieval model for food safety domain. Chinese Journal of Electronics, 2013,22(2):247−252.

[29] Zhou H, Liu BW, Liu J. Research on mechanism of the information retrieval based on ontology label. Procedia Engineering, 2012, 29:4259−4266. [doi: 10.1016/j.proeng.2012.01.654]

[30] Fan WG, Pathak P, Zhou M. Genetic-Based approaches in ranking function discovery and optimization in information retrieval—A framework. Decision Support Systems, 2009,47(4):398−407. [doi: 10.1016/j.dss.2009.04.005]

[31] Maria G, Marcin P, Jakub S. Combining information from multiple search engines-preliminary comparison. Information Sciences, 2010,180(10):1908−1923. [doi: 10.1016/j.ins.2010.01.010]

[32] Maedche A, Staab S, Stojanovic N, Studer R, Sure Y. SEmantic portAL: The SEAL approach. In: Proc. of the Spinning the Semantic Web. MIT Press, 2003. 317−359. https://mitpress.mit.edu/index.php?q=books/spinning-semantic-web

[33] Guha R, McCool R, Miller E. Semantic search. In: Proc. of the 12th Int'l World Wide Web Conf. (WWW 2003). Budapest, 2003. 700−709. http://www2003.org/cdrom/index.html

[34] Zhang H, Wu F, Zhuang YT, Chen JX. Cross-Media retrieval method based on content correlations. Chinese Journal of Computers, 2008,31(5):821−825 (in Chinese with English abstract).

[35] Zhuang YT, Yang Y, Wu F. Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval. IEEE Trans. on Multimedia, 2008,10(2):221−229. [doi: 10.1109/TMM.2007.911822]

[36] Stevens R, Baker P, Bechhofer S, Ng G, Jacoby A, Paton N, Goble CA, Brass A. TAMBIS: Transparent access to multiple bioinformatics information sources. Bioinformatics, 2000,16(2):184−185. [doi: 10.1093/bioinformatics/16.2.184]

[37] Pan X, Zhang SY, Ye XZ. A survey of content-based 3D model retrieval with semantic features. Chinese Journal of Computers, 2009,32(6):1069−1079 (in Chinese with English abstract). [doi: 10.3724/SP.J.1016.2009.01069]

[38] Dai WM. Semantic Web Information Organization Technologies and Methods. Shanghai: Academia Press, 2008 (in Chinese).

[39] Iribarne L, Padilla N, Asensio JA, Criado J, Ayala R, Almendros J, Menenti M. Open-Environmental ontology modeling. IEEE Trans. on Systems, Man, and Cybernetics—Part A: Systems And Humans, 2011,41(4):730−745. [doi: 10.1109/TSMCA.2011.2132706]

[40] Lee CS, Wang MH, Hagras H. A type-2 fuzzy ontology and its application to personal diabetic-diet recommendation. IEEE Trans. on Fuzzy Systems, 2010,18(2):374−395. [doi: 10.1109/TFUZZ.2010.2042454]

[41] Razmerita L. An ontology-based framework for modeling user behavior—A case study in knowledge management. IEEE Trans. on Systems, Man, and Cybernetics—Part A: Systems And Humans, 2011,41(4):772−783. [doi: 10.1109/TSMCA.2011.2132712]

[42] Liu Y, Xu CF, Zhang Q, Pan YH. The smart architect: Scalable ontology-based modeling of ancient Chinese architectures. IEEE Intelligent Systems, 2008,23(1):49−56. [doi: 10.1109/MIS.2008.16]

[43] Gan JH, Jiang Y, Xia YM. Ontology and Its Application. Beijing: Science Press, 2011 (in Chinese).

[44] Cui JD. The study of grid information retrieval model based on ontology [Ph.D. Thesis]. Changchun: Jilin University, 2011 (in Chinese with English abstract)

[45] Du XY, Li M, Wang S. A survey on ontology learning research. Ruan Jian Xue Bao/Journal of Software, 2006,17(9):1837−1847 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/17/1837.htm

[46] Zouaq A, Gasevic D, Hatala M. Towards open ontology learning and filtering. Information Systems, 2011,36(7):1064−1081. [doi: 10.1016/j.is.2011.03.005]

[47] Segev A, Sheng QZ. Bootstrapping ontologies for Web services. IEEE Trans. on Services Computing, 2012,5(1):33−44. [doi: 10.1109/TSC.2010.51]

[48] Ju YH, Liu C. Comparative study on foreign representative semantic annotation platforms. Journal of Modern Information, 2009, 29(1):215−217 (in Chinese with English abstract).

[49] Liao SH. Comments on ontology-based semantic annotation prototypes. Computer Engineering Science, 2006,28(9):123−125 (in Chinese with English abstract).

[50] Rajput Q, Haider S. BNOSA: A Bayesian network and ontology based semantic annotation framework. Web Semantics: Science, Services and Agents on the World Wide Web, 2011,9(2):99−112. [doi: 10.1016/j.websem.2011.04.002]

[51] Nguyen CD, Gardiner KJ, Cios KJ. Protein annotation from protein interaction networks and gene ontology. Journal of Biomedical Informatics, 2011,44(5):824−829. [doi: 10.1016/j.jbi.2011.04.010]

[52] Smine B, Faiz R, Desclés JP. A semantic annotation model for indexing and retrieving learning objects. Journal of Digital Information Management, 2011,9(4):159−166.

[53] Yuan L, Li ZH, Chen SL. Ontology-Based annotation for deep Web data. Ruan Jian Xue Bao/Journal of Software, 2008,19(2): 237−245 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/19/237.htm

[54] Chen KR, Zuo WL, Zhang F, He FL, Chen YH. Robust and efficient annotation based on ontology evolution for deep Web data. Journal of Computers, 2011,6(10):2029−2036. [doi: 10.4304/jcp.6.10.2029-2036]

[55] Jing T, Zuo WL, Sun JG, Che HY. Semantic annotation of Chinese Web pages: From sentences to RDF representations. Journal of Computer Research and Development, 2008,45(7):1221−1231 (in Chinese with English abstract).

[56] Carpineto C, Romano G. A survey of automatic query expansion in information retrieval. ACM Computing Surveys, 2012,44(1): 1−50. [doi: 10.1145/2071389.2071390]

[57] Selvaretnam B, Belkhatir M. Natural language technology and query expansion: Issues, state-of-the-art and perspectives. Journal of Intelligent Information Systems, 2012,38(3):709−740. [doi: 10.1007/s10844-011-0174-3]

[58] Adeel NRM, Mark S, Paul C. Retrieving candidate plagiarised documents using query expansion. In: Proc. of the 34th European Conf. on IR Research (ECIR 2012). LNCS 7224, Barcelona, 2012. 207−218. [doi: 10.1007/978-3-642-28997-2_18]

[59] Voorhees EM. Query expansion using lexical-semantic relations. In: Proc. of the SIGIR'94 (Proc. of the 17th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval). Dublin, 1994. 61−69. http://link.springer.com/book/10.1007/978-1-4471-2099-5

[60] Liu ZY, Chen JX, Li XH, Qu Y, Li FC. Design and application for the model of semantic query expansion based on domain ontology. Int'l Journal of Modelling, Identification and Control, 2012,16(3):277−284. [doi: 10.1504/IJMIC.2012.047739]

[61] Chen SM, Du YJ, Peng QQ. Ontology-Based query expansion in formal concept analysis. Journal of Computational Information Systems, 2009,5(3):1603−1611.

[62] Pan JZ, Thomas E, Ren Y, Taylor S. Exploiting tractable fuzzy and crisp reasoning in ontology applications. IEEE Computational Intelligence Magazine, 2012,7(2):45−53. [doi: 10.1109/MCI.2012.2188588]

[63] Liu C, Qi GL, Wang HF, Yu Y. Reasoning with large scale ontologies in fuzzy pD* using MapReduce. IEEE Computational Intelligence Magazine, 2012,7(2):54−66. [doi: 10.1109/MCI.2012.2188589]

[64] Juan GR, Patricio MA, García J, Molina JM. Ontology-Based context representation and reasoning for object tracking and scene interpretation in video. Expert Systems with Applications, 2011,38(6):7494−7510. [doi: 10.1016/j.eswa.2010.12.118]

[65] Pan C, Gu H. Ontology reasoner and its application. Computer System & Applications, 2010,19(9):163−167 (in Chinese with English abstract).

[66] Li JJ, Yan HF. Chinese Web retrieval test collections: Construction, analysis and application. Journal of Chinese Information Processing, 2008,22(1):30−36 (in Chinese with English abstract).

附中文参考文献:

[1] 董金祥.基于语义面向服务的知识管理与处理.杭州:浙江大学出版社,2009.

[34] 张鸿,吴飞,庄越挺,陈建勋.一种基于内容相关性的跨媒体检索方法.计算机学报,2008,31(5):821−825.

[37] 潘翔,张三元,叶修梓.三维模型语义检索研究进展.计算机学报,2009,32(6):1069−1079. [doi: 10.3724/SP.J.1016.2009.01069]

[38] 戴维民.语义网信息组织技术与方法.上海:学林出版社,2008.

[43] 甘健侯,姜跃,夏幼明.本体方法及其应用.北京:科学出版社,2011.

[44] 崔金栋.基于本体的网格信息检索模型研究[博士学位论文].长春:吉林大学,2011.

[45] 杜小勇,李曼,王珊.本体学习研究综述.软件学报,2006,17(9):1837−1847. http://www.jos.org.cn/1000-9825/17/1837.htm

[48] 鞠彦辉,刘闯.国外典型语义标注平台的比较研究.现代情报,2009,29(1):215−217.

[49] 廖述梅.基于本体的语义标注原型评述.计算机工程与科学,2006,28(9):123−125.

[53] 袁柳,李战怀,陈世亮.基于本体的 Deep Web 数据标注.软件学报,2008,19(2):237−245. http://www.jos.org.cn/1000-9825/19/237.htm

[55] 荆涛,左万利,孙吉贵,车海燕.中文网页语义标注:由句子到 RDF 表示.计算机研究与发展,2008,45(7):1221−1231.

[65] 潘超,古辉.本体推理机及应用.计算机系统应用,2010,19(9):163−167.

[66] 李静静,闫宏飞.中文网页信息检索测试集的构建、分析及应用.中文信息学报,2008,22(1):30−36.

杨月华(1983－),女,吉林松原人,博士,讲师,主要研究领域为智能信息检索,知识工程.

平源(1981－),男,博士,讲师,CCF 学生会员,主要研究领域为机器学习,信息安全.

杜军平(1963－),女,博士,教授,博士生导师,CCF 高级会员,主要研究领域为智能信息处理,人工智能.