基于文本挖掘的本体自动构建系统架构解析

薛中玉[1],李春梅[2],黄道雄[1]
(1. 北京中机科海科技发展有限公司,北京 100048;
2. 机械工业仪器仪表综合技术经济研究所,北京 100055)

摘 要:本体可以为人与计算机之间的沟通和交流提供语义支撑,在人工智能、知识工程等众多领域有着广泛的应用空间,但现阶段本体主要采用人工构建方法,投入资源大、建设周期长,且质量无法保障,这些成为制约本体应用的主要瓶颈。文中提出了一种基于文本挖掘的本体自动构建系统和方法,详细介绍了用户层、系统工具层和数据资源层中各模块的功能和实现方法,具体分析了系统数据处理的整个流程。该系统和方法对于解决本体构建问题具有借鉴意义。

关键词:文本挖掘;本体构建;系统架构
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-629X(2011)01-0100-04

System Architecture Analysis of Automatic Construction System of Ontology Based on Text Mining

XUE Zhong-yu[1], LI Chun-mei[2], HUANG Dao-xiong[1]
(1. Beijing Zhongjikehai Technology Development Ltd, Beijing 100048, China;
2. Machinery Industry Instrumentation Technology and Economy Institute, Beijing 100055, China)

Abstract: Ontology is able to offer a semantic support for human-computer interaction so that it can be found wide applications in the fields of artificial intelligence, knowledge engineering and so on. However, at present ontology construction mainly uses the manual approach with disadvantage of higher construction cost, long development period, and unsure quality. This becomes a major bottleneck to hinder ontology applications. This paper presents an automatic construction system of ontology and method based on text mining, introduces in detail the functions and implementation method of the user layer, system tools layer and data resource layer in the system, and analyzes the whole system data processing flow. This system and method can be used for reference to solve the similar problems in ontology construction.

Key words: text mining; ontology construction; system architecture

0 引 言

“本体”(Ontology)最初是哲学领域的术语,是关于事物存在及其本质规律的学说[1]。20世纪末,随着信息技术的发展,本体被引入人工智能、知识工程等领域,用于构建大型集成的知识库系统,解决知识概念表示和知识组织体系方面的问题。在新的技术领域,本体被赋予更为具体的定义——共享概念模型的、明确的、形式化的规范说明[2],一般由概念(Concepts)、概念间关系(Relations)和规则(Rules)构成。

本体由其自身的特点,可以将人们广泛认可的各种类型知识转化为规范的、计算机可以理解的形式,为“计算机与人之间正常沟通与交流”提供语义支撑。因此,本体在人工智能、知识工程、图书情报和搜索引擎等众多领域都有广泛的应用空间[3]。但是,目前真正投入使用的本体还很少。其主要原因在于现有本体的构建是以人手工为主[4],利用 Protégé[5] 和 OntoEdit[6] 等常见本体工具,技术的应用实施还很困难,由于该项工作是一项非常复杂、庞大的系统工程,将相关领域的概念和关系进行梳理,并用规范化的模式进行表达,需要领域专家花费大量时间和精力,并且期间涉及了多位专家协作,如果专家间认识和理解不同时,将会出现不一致的现象,需要逐一进行协调和确认,其工作量相当之大。鉴于本体构建工程的复杂性和智力密集性等特点使得本体的构建往往投入资源大、建设周期长,且质量无法保障,这些成为影响本体应用和推广的主要瓶颈和难点[7]。因此,解决本体构建阶段现有技术和方法的瓶颈和难点成为业内人士主要研究方向之一。文中提出了一种基于文本挖掘的本体自动构

收稿日期:2010-04-27;**修回日期**:2010-07-03
基金项目:国家国际科技合作计划项目(2009DFA13110)
作者简介:薛中玉(1981-),男,河南开封人,硕士,工程师,从事文本挖掘、本体和信息检索研究。

建技术,对解决本体构建问题具有很大的借鉴意义。

1　文本挖掘

文本挖掘(Text Mining)是指为了发现知识,从大规模文本库中抽取隐含的、以前未知的、潜在有用的知识(包括概念、模式、规则、规律、约束等形式)[8]。

信息存储与交互最自然的形式是自然语言文本。绝大多数的电子化信息是以无结构自由文本的形式存在的,如 Web 页面、在线新闻、公司档案、研究论文、电子书籍、E-mail 等[9]。这些信息几乎囊括了相关领域所有的概念、知识和专家学者的思想,如果能够利用好这些信息中所包含的知识,完全可以构建非常完整、实用的本体。但是,因为这些信息是由非结构化的自然语言表示的,具有模糊性和歧义性,无法直接获取蕴含的概念和知识,需要运用文本挖掘技术对其进行分析和处理[10]。

文本挖掘的过程一般包括文本数据预处理、文本信息提取和索引、文本知识挖掘及知识后处理等步骤[11]。数据预处理包括数据清洗(如去噪、去重)、数据选择(选择合适的、面向特定领域的文本数据)和文本切分(如中文分词、段落切分)等。数据预处理后,必须提取中文文本的特征信息,包括关键词提取、术语提取、基于模板的信息抽取和基于专业词典的概念转换等操作。经过中文文本特征提取操作后,中文文本数据转换为中文文本信息。在文本信息的基础上进行知识挖掘,包括文本自动摘要、文本聚类、关联规则抽取和语义关系挖掘等。由于知识挖掘得到的结果可能不一致、不新颖、不符合构建本体基本要素的形式要求,因此需要对文本知识进行必要的后处理,包括知识的评价与取舍、知识的规范形式化表达等。

采用文本挖掘技术,通过文本预处理、提取知识中的概念和关系,能够为本体自动构建提供所需的素材。通过开发的文本挖掘结果分析工具和本体自动构建工具,进而能够实现本体的自动构建。

2　系统架构

基于文本挖掘的本体自动构建系统架构如图 1。该系统主要分为用户层、系统工具层和数据资源层等。

2.1　用户层

用户层包括领域资料管理(具体分为本体名称、核心概念、主题词表和语料库等)、规则模版管理、核心概念管理、三元组管理和本体文件管理等接口模块,用于提供丰富的人机交互接口。各模块主要功能如下:

领域资料管理,用于接收用户提交的预构建本体的名称和核心概念、该领域的主题词表和领域相关语料。

规则模版管理,用于接收用户对系统默认设定的领域概念识别、核心语句抽取、本体继承关系提取等规则模版进行的添加、修改和删除等操作。

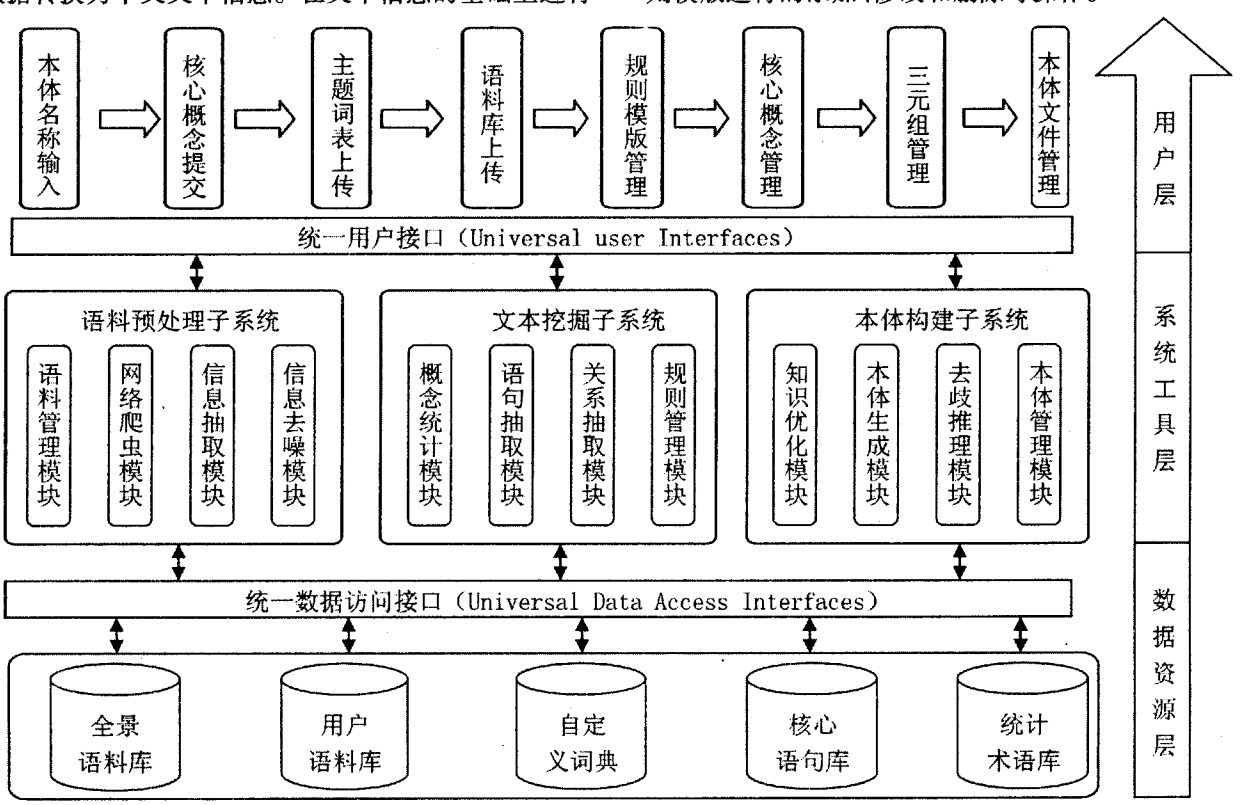

图 1　基于文本挖掘的本体自动构建系统架构

核心概念管理,用于接收用户对系统提取的语料中的备选核心概念进行修改、添加、删除和确认等操作。

三元组管理,用于接收用户对三元组进行的编辑、删除和确认等操作,并返回最终的三元组序列。

本体文件管理,用于展示系统自动生成的本体文件,接收用户对本体的概念和关系进行的添加、修改和删除等操作,并返回给用户最终的本体文件。

2.2 系统工具层

系统工具层包括语料预处理子系统、文本挖掘子系统和本体构建子系统,用于语料分析、知识挖掘和本体构建。

(1)语料预处理子系统。

语料预处理子系统包括语料管理模块、网络爬虫模块、信息抽取模块和信息去噪模块,用于接收和处理用户提供的相关资料。各模块的功能如下:

语料管理模块,用于管理用户上传的各类资源,包括对上传语料的添加、选择、删除和分类等操作。

网络爬虫模块,用于对网页抓取引擎的设置和对网页抓取资源的监控,并实现对相关网页的镜像抓取。

信息抽取模块,用于对选中的多种格式(pdf、word、ppt、txt、xls 和 html 等)文件中的信息进行抽取。

信息去噪模块,用于去除各类文件中的无用信息(包括乱码、标签、页眉、页脚等),并确保有用信息完整保留。

(2)文本挖掘子系统。

文本挖掘子系统包括概念统计模块、语句抽取模块、关系抽取模块和规则管理模块,用于分析和挖掘语料中相关知识。各模块的功能如下:

概念统计模块,用于统计用户上传语料中的单词概念和组合概念的权重和领域相关性,最终识别和确定领域的相关概念,形成领域相关概念集,该模块还通过用户层的核心概念管理接口显示领域概念统计结果,并实现用户对领域概念进行的添加、编辑和删除等操作。

语句抽取模块,用于识别并抽取语料预处理结果中与领域相关的核心语句。

关系抽取模块,用于抽取核心语句中有用的、领域相关的三元组关系,具体包括本体概念间的上下位继承关系、同义关系、属性关系和实例关系。

规则管理模块,用于实现用户对相关规则模版进行的添加、修改和删除等操作,使之更加适合用户所选择的技术领域和所上传的领域资料。

(3)本体构建子系统。

本体构建子系统包括知识优化模块、本体生成模块、去歧推理模块、本体管理模块,用于组织和搭建最终本体。各模块的功能如下:

知识优化模块,用于对包含上下位继承关系、同义关系、属性关系和实例关系的各条三元组进行自动分类整理,并对三元组关系的领域相关性和有用性进行计算推理,识别并排除不相干、歧义和无用的三元组信息,通过用户层的三元组管理接口,返回给用户进行必要的修改和确认。

本体生成模块,用于生成本体知识库。通过调用 Jena、KAON2 等工具添加本体类、属性和实例的接口,或根据 OWL(Ontology Web Language)[12] 的语法格式要求,直接操作本体文件,将三元组关系转化成本体。

去歧推理模块,用于对本体文件进行一致性和完整性检测,找出并反馈本体文件中矛盾、重复、不一致和概念缺失等问题。

本体管理模块,用于对生成的本体文件进行编辑和修改,对本体中的元素进行添加、修改、查询和删除。

2.3 数据资源层

数据资源层包括全景语料库、用户语料库、自定义词典、知识提取库和统计术语库,用于存储提供最初语料、中间产物和分析结果。各模块主要功能如下:

全景语料库,用于存储有代表性的、尽量涵盖国民经济所有领域的各类语料。相关语料来源可以是近年来较为规范的、全国各类期刊杂志的摘要信息。

用户语料库,用于存储用户上传的各类语料信息资源,包括通过用户设定的领域门户网站抓取的网页信息,以及文本预处理的结果信息。

自定义词典,作为系统分词、句法分析的自定义词典,用于记录并通过系统分析挖掘不断更新的领域相关概念集,以提高系统分析的准确率。

知识提取库,用于存储系统抽取的三元组信息。

统计术语库,用于存储全景语料库和用户语料库中各类语料术语统计分析的结果。

3 系统处理流程

基于文本挖掘的本体自动构建系统处理流程如图 2 所示。其中实线空心箭头代表系统处理的正向流向,虚线空心箭头代表当系统处理中间结果不理想时,返回前期步骤进行修改和校正,以期获得更好结果。具体步骤如下:

(1)本体名称输入,接收用户输入的本体名称,根据用户输入的本体名称创建含有一个顶层类概念的初级本体模型。

(2)核心概念提交,接收用户提交的、在预构建本体中占有重要地位的一系列概念。构建的本体必须包含这些核心概念,并需要一定程度的扩展,所输入的核心概念及其下位概念应在本体所有概念中占有一定规

模。另外将用户输入的核心词汇添加到自定义词典。

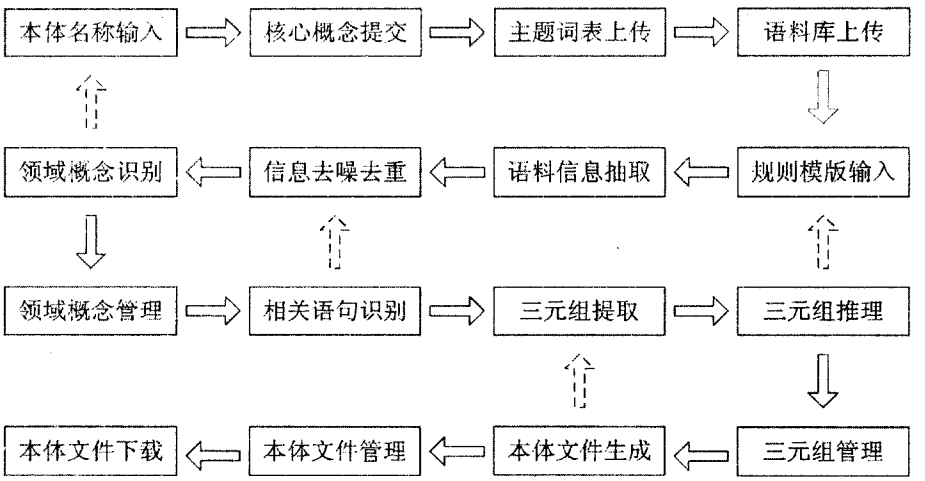

图 2　基于文本挖掘的本体自动构建系统处理流程图

(3)主题词表上传,接收用户上传的、该领域的词汇集,生成主题词表,更新自定义词典,将主题词表中的层次继承关系加入到知识提取库,为系统分词和领域相关度计算等模块提供依据。

(4)语料库上传,接收用户上传的该领域的相关语料。内容包括与领域有关的法律、法规和管理办法等政策文件,以及著作、论文、标准、研究报告和专利等成果文件。上传语料格式包括 pdf、word、ppt、txt、xls 和 html 等,支持对该领域相关门户网站的输入,系统能够通过网络爬虫模块自动抓取网站相关信息,作为语料的一部分存入用户语料库。

(5)规则模版输入,上传完用户语料库后,用户可以通过"规则模版输入",更新当前系统的领域概念识别模版、继承关系表达模版、同义词表达模版、属性表达模版、实例表达模版等。如果处理后发现所提取的三元组关系的有用性和领域相关性均较小时,可以根据需要对规则模版库进行更新。

(6)语料信息抽取,实现对用户语料库中的 pdf、doc、ppt、html、excel、txt 等常见文件中的信息进行抽取。

(7)信息去噪去重,将抽取的信息进行去噪处理,去除标签、乱码、页眉和页脚等无用信息,同时确保有用信息被完整保留。

(8)领域概念识别,通过对预处理后的、语料中的词汇进行统计分析,最终找出领域概念,并更新自定义词典。通过领域概念识别后,如果发现所识别的领域概念和预构建领域本体的相关性均不大,则可以选择对本体名称、核心概念的重新输入以及领域主题词表的重新上传和修改。

(9)领域概念管理,实现对识别的领域单词概念、组合概念、主题词表上传概念等进行分类、修改、添加和删除等管理,最终保存与该领域最相关的领域概念。

(10)相关语句识别,实现对含有领域概念的相关语句的识别和抽取,并将抽取结果返回用户。如果发现所抽取的相关语句含有乱码、杂乱符号等问题,则返回信息去噪去重进行重新处理。

(11)三元组提取,三元组关系主要包括本体层次继承关系、本体同义词关系、本体属性关系和本体实例关系等,基于各种关系对应的模版进行提取。

(12)三元组推理,对提取的三元组关系进行推理,通过设定规则,进行一致性、冗余性检测,自动发现并删除多余、矛盾或错误的三元组关系,并将结果返回用户。经用户判断,如果整体符合要求,则进行下一步处理,如果所提取的三元组整体相关性较差,则返回规则模版输入进行重新调整和校正。

(13)三元组管理,对经自动推理判断后的三元组进行人工添加、修改和删除等操作,以提高生成本体文件的质量。

(14)本体文件生成,利用系统前期步骤生成的三元组关系,搭建初步的本体文件,并返回用户进行确认,如果与本体需求相差较大,则返回三元组提取步骤重新处理,如果基本符合本体需求,则进行下一步处理。

(15)本体文件管理,实现对初步生成的本体文件的在线编辑,实现对本体中概念、关系和实例的添加、修改和删除等操作,最终保存修改后的本体文件。

(16)本体文件下载,将最终生成的本体文件下载到本地。

4　结束语

文中提出的基于文本挖掘的本体自动构建系统和方法能够在很少人工干预的情况下,完成本体的自动构建,缩减本体构建周期的同时,能够充分利用互联网上信息和用户拥有的电子资源,进行很好地融合、推理和消歧,避免个别专家观点对本体知识的影响。该系统和方法对解决人工构建本体的问题有很大的借鉴意义,对推动本体在人工智能、知识工程、图书情报和搜索引擎等众多领域的广泛应用有很大帮助。

参考文献:

[1]　邓志鸿,唐世渭,张　铭,等. Ontology 研究综述[J]. 北京大学

(下转第 128 页)

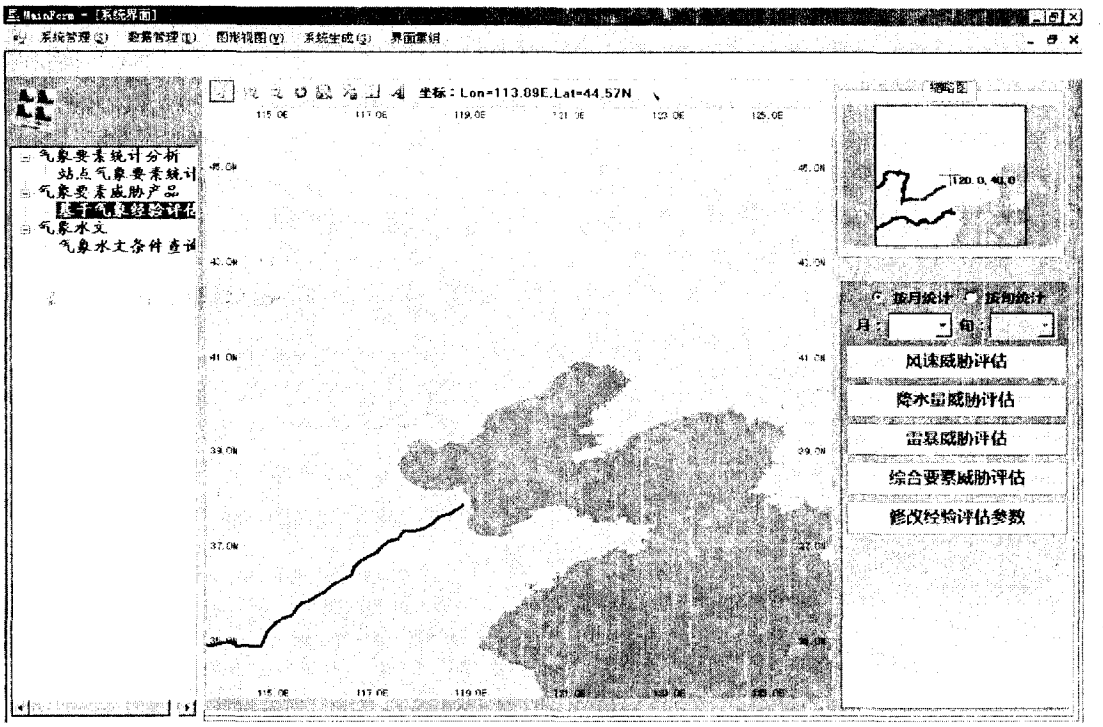

图 5 生成新系统界面

析[J]. 软件工程与标准化,2006(3):38-41.

[2] 严洪森,刘 飞. 知识化制造系统——新一代先进制造系统[J]. 计算机集成制造系统,2001,7(8):7-11.

[3] Yan Hongsen. A new complicated-knowledge representation approach based on knowledge meshes[J]. IEEE Transactions on Knowledge and Data Engineering,2006,18(1):47-62.

[4] 孙中桥,陈菊红. 知识化制造单元协同工作模型的研究[J]. 计算机集成制造系统,2007,13(8):1534-1538.

[5] 范玉顺. 中国运用现代集成制造技术改造传统产业的经验和前景[J]. 制造业自动化,2002,24(4):1-8.

[6] 黄 琛,范玉顺. 基于知识的企业 CIMS 框架及关键技术研究[J]. 计算机集成制造系统,2003,9(10):830-833.

[7] 任开银,黄 东. MIS 中知识的数据库表示及应用[J]. 工业控制计算机,2003,16(1):10-11.

[8] 薛朝改,严洪森. 基于 Agent 网的知识网的自重构研究[J]. 计算机集成制造系统,2003,9(11):995-1000.

[9] 薛朝改,严洪森. 知识化制造系统自重构的研究[D]. 南京:东南大学,2005.

[10] 薛朝改,严洪森. 基于组件技术的知识化制造系统自重构的实现[J]. 计算机集成制造系统,2004,12(10):39-45.

[11] 罗 景,张 路,孙家骕. 构建提取技术综述[J]. 计算机科学,2005,32(12):1-7.

[12] 张 平,严洪森,余晓光. 基于混合算法的知识网运算表达式优化[J]. 计算机技术与发展,2009,19(3):32-35

(上接第 103 页)

学报(自然科学版),2002,38(5): 730-738.

[2] Studer R, Benjamins V R, Fensel D. Knowledge Engineering, Principles and Methods[J]. Data and Knowledge Engineering, 1998,25(1-2):161-197.

[3] 赵 丽. 本体的理论及其应用研究[D]. 兰州:兰州理工大学,2006.

[4] 王晓盈,王晓璇,刘 鹏. 中文本体构建及可视化研究[J]. 计算机技术与发展,2010,20(2):121-124.

[5] Rubin D L, Noy N F, Musen M A. Protégé: A Tool for Managing and Using Terminology in Radiology Applications[J]. Journal of Digital Imaging,2007,20(1):34-46.

[6] Sure Y, Angele J, Staab S. OntoEdit: Guiding Ontology Development by Methodology and Inferencing[C]//In: Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2008:1205-1222.

[7] 薛中玉. 工业自动化仪表领域本体及其在数据共享中的应用研究[D]. 北京:机械科学研究总院,2009.

[8] 谌志群,张国煊. 文本挖掘研究进展[J]. 模式识别与人工智能,2005,18(1):65-74.

[9] 薛为民,陆玉昌. 文本挖掘技术研究[J]. 北京联合大学学报(自然科学版), 2005,19(4):59-63.

[10] 梁颖红,曹 军. 文本语块识别典型方法的比较与分析[J]. 计算机技术与发展 2008,18(11):76-79.

[11] 谌志群,张国煊. 文本挖掘与中文文本挖掘模型研究[J]. 情报科学, 2007,25(7):1046-1051.

[12] Horrocks I. OWL: A Description Logic Based Ontology Language [C]//In: Logic Programming. Berlin, Heidelberg: Springer ,2005:1-4.