基于本体的主动元数据挖掘系统研究

徐卫军[1],李宝敏[2]
(1. 西安工业大学,陕西 西安 710032;
2. 西安培华学院,陕西 西安 710125)

摘 要:传统网上信息检索是用户被动地依靠浏览超级链接网页而获取的。文中提出基于本体的主动元数据挖掘系统以及在果品领域的应用,在主动搜索、元数据生成、借助本体作用于数据的语义描述等方面,其效果是客观的。使得对信息数据的搜索从被动地获取到主动依靠计算机自动搜索;从依靠关键字作为查询依据到借助本体的作用获取语义描述的信息数据,进而提高了信息查询效率及查询的准确率,这也是当前信息检索研究的热门课题。实验证明,通过主动元数据挖掘实例可以实现语义的扩充,如同义、近义及上下位关系。同时也验证了本体对实施语义智能检索所带来的客观效果。

关键词:本体;领域;主动元数据挖掘;语义
中图分类号:TP391.3 **文献标识码**:A **文章编号**:1673-629X(2012)03-0023-05

Research Based on Ontology Initiative Metadata Mining System

XU Wei-jun[1], LI Bao-min[2]
(1. Xi'an Technology University, Xi'an 710032, China;
2. Xi'an Peihua University, Xi'an 710125, China)

Abstract: Traditional online information retrieval refers to the user browsing the web a way to get information. In this paper, ontology-based data mining systems and active element in the fruit fields of application is proposed, the initiative search, metadata generation and description of semantic on role of body, its effect is objective. The experiments show that, the example of initiative metadata mining can achieve expansion of semantic, like synonymy, nearsynonymy and superordinate and hyponym. Simultaneously it also confirmed the objective effect of the ontology on the intelligent retrieval of semantic.

Key words: ontology; domain; initiative metadata mining; semantics

0 引 言

Internet的信息量是以指数规律在不断膨胀的,其信息的组织是异构的、多元的和分布的。目前的信息检索系统都是以关键字来检索,由于关键字的局限性,不能完全表达用户的检索要求,它存在以下不足:

● 用简单的关键字标引网页,由于其表述的局限性不能准确地反映文档的逻辑语义,常会检索出一些不相关的信息,从而使检索的效率降低;

● 关键字的检索,对于表示同一语义的同义、相似和多义的概念信息无能为力,使得一些有用的网页就会被漏选,其后果查全率不高;

●关键字的检索会引出许多词汇相同而并非用户所需的信息,一般必须通过URL超链接逐个进行浏览,用户经过筛选最终能得到或得不到所需信息,这种"信息过载"是一项既费时又费力的工作,且结果并非能如意;

● URL超链接信息的获取往往是被动的,是通过用户浏览过程实现的。

为解决这一问题可以从两方面入手:一是建立个性化信息服务,根据用户的知识背景、兴趣爱好和历史操作等,执行符合用户个性特色的检索策略;二是将现有的基于关键词层面的信息检索提高到基于知识、语义层面的信息检索,以减少因用户知识水平和认知能力的局限而返回的不准确信息。本课题就是综合以上两方面知识,基于本体语义技术,利用主动元数据挖掘研究语义检索。

1 本体与语义信息模型

本体(Ontology)概念最先来自于哲学范畴(也有将其译成"本体论"的),是用来描述事物的本质[1]。以后在工程领域借用了这个概念。对于本体的定义,

收稿日期:2011-08-02;**修回日期**:2011-11-07
基金项目:国家"星火计划"项目(2004EA850069)
作者简介:徐卫军(1967-),女,浙江金华人,图书馆馆员,研究方向为数字图书馆;李宝敏,教授,研究方向为计算机系统结构、计算机网络与语义网。

人工智能学者 Neches 等人认为:“本体”是“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”。即:摸索出基本的术语和术语间的关系及相应的规则,并给出这些术语和关系的定义。这是一种描述构造本体的过程。而计算机界的 Studer 等人则把本体定义为:是共享概念模型的明确的形式化规范说明[2]。

本体的目标是对相关领域知识的获取、描述和表示,是对该领域知识的共同理解,确定其内共同认可的词汇,并明确定义了来自不同层次的形式化模式上的这些词汇(术语)和词汇间相互的关系。本体具有静态性和动态性。静态性是指本体反映的是概念模型,没有涉及动态的行为;动态性则是指本体的内容和服务对象是变化的,对于不同的领域,本体的定义和构造则不同。

美国科学家联合会对信息模型的理解为:信息模型用于描述组织内的信息资源以及这些资源的相互关系。信息是作为一种资源为了某种特定的目的、在一定范围内聚合起来的数据集。数据可以被认为是有含义事件的符号表示,单一的含义可以用于多个不同的事件。信息是由数据构成,而数据是由事实和其代表的含义构成。因此,语义信息模型就是数据所代表的语义及其之间的关系。从知识共享的角度看,本体可以被看作是一种概念化的显式说明,是对客观存在的概念和关系的描述。概念模式的主要目标是提供一个数据含义和相互关系的一致定义,从而保证集成、共享和数据管理的完整性。

基于 Ontology 的信息系统中,Ontology 提供了概念化、可共享的知识体系,成为语义检索实现知识理解和知识推理的基础。

随着网络信息的多样化、网络数据库的异构化,本体论越来越受到了计算机界的重视。在帮助智能体对 Internet 上的种种信息进行领域分类、信息整合以及规范用户信息智能化检索方面,本体论都将发挥越来越重要的作用。

2 主动元数据挖掘

元数据是一种用来描述资源的属性、能够进行精确查询的结构化数据,是一种能够被计算机理解和处理的数据[3]。对于因特网上的数字化文档,在进行必要的预先处理后,一些相对规整的文本文档,经过元数据提取模块的加工处理,产生符合规范定义的文档元数据,并将该文档元数据存到元数据数据库中。因特网上的数字化文档一般质量参差不齐,格式也不统一,但是,经过数字化文档元数据的规范定义后,生成提取元数据的相应模式,再依此模式对数字化文档进行挖掘与匹配。自动抽取数字化元数据,就是说在没有人的干预情况下,依靠文档元数据规范地主动地来获取用户感性趣的元数据信息,并将获取的结果进行合理的组织与存储。这种提取是与 Robert * 的主动挖掘协调进行的。

智能 Agent 又称为智能主体,其作用是在用户没有具体明确的要求情况下,依据用户的需求,替代用户完成各种复杂的操作[4]。例如对信息的查询、筛选以及管理。智能 Agent 还能推测用户的用意、自主制定、调整和自动处理等。在主动元数据挖掘过程中利用智能 Agent 来监控 Robert * 24 小时对网络的搜索,对网页文本进行分词、词法分析、语义分析,根据用户对信息的偏好以及需要进行甄别、归纳和总结,然后独立、自动地代替用户检索其有兴趣的信息;并且能够应用多种通信协议与多个智能 Agent 进行交流,与其它部件一起共同完成数据信息主动挖掘任务。

3 系统结构模块及各模块功能

基于本体主动元数据挖掘系统结构模块如图 1 所示。

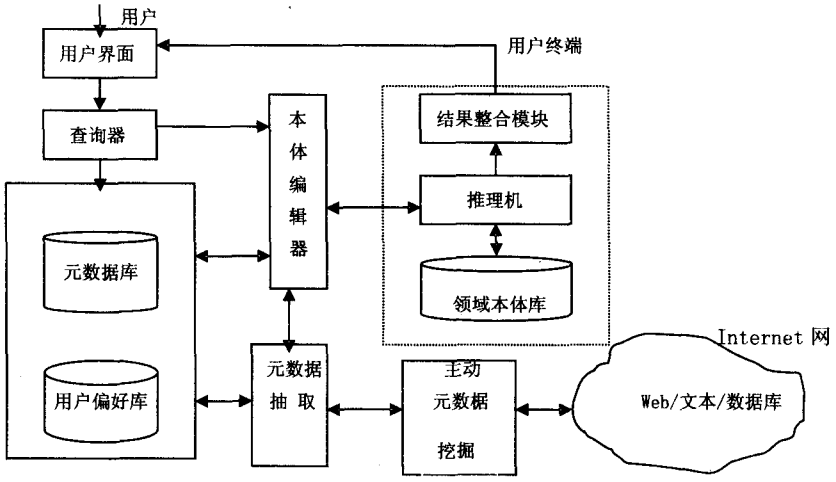

图 1 本体主动元数据挖掘系统结构模型

该结构模块的基本工作原理是:首先由用户提出检索要求,检索器接受请求,然后按照本体的要求经过本体编辑器转换成本体规范的格式提供给本体推理机,进行推理、判断、语义分析等处理,给出用户准确的语义,然后,在本体的协助下检索元数据库和用户偏好数据库中符合条件的并与其相匹配的数据集,经过格式定制及整合后输出给终端用户 。具体讲就是:“兴趣偏好模型+主动挖掘方法、规则+本体协助→元数据”,最终实现:“用户兴趣化的知识+元数据+本体的

智能推理→用户实现准确查询"[5]。

本系统各个模块的主要功能是:

(1) 用户登录界面:接收用户输入的查询信息以及输出系统查询返回的结果。输入的查询信息可以是关键字:对于确定的信息,如轮船、汽车等;不确定的信息,如个人数字计算机、微型电脑(同义不同名)等。也可以是自然语言,如软件工程师、软件讲师、软件教师等等;也可以是嵌套的模式语言,如数理逻辑方面的论文等。

(2) 查询器:第一步将分布在各种存储介质上的信息收集到查询信息库中。语义检索通常是以提供专业解决方案为目标,其查询数据库中的信息也是主要面向具体的专业领域。因此,可以选择专业图书馆或具有较高权威性的专业网站作为信息检索的起点。

(3) 本体编辑器:用语义描述语言描述本体所要求的表达形式。用户查询的信息遵照本体规定经过本体编辑器转换成要求的格式(推理机内部检索语句)并以检索模块提供给推理机。遵照本体规定实现从元数据库中检索与其相匹配的信息,查询不到的信息则通知查询器进行再检索。依据术语含义来修改用户偏好库,同时记录下用户的偏好,作为推理机来推理、判断提供用户个性特征的依据。

(4) 本体推理机:接收本体编辑器提供的检索信息,通过本体推理模块,使用推理规则,经过元数据库的查询后匹配出符合条件的数据集合,与用户偏好库(用户个性特征)共同生成本体内一种表达样式,经过推理、查询、判断,经结果整合模块处理后提供给用户。

例如,用户输入"本体应用方面的论文",推理机从中发现有三方面的知识可展示用户的真实需要。知识1:本体应用;知识2:论文(一般与期刊和文章有关);知识3:所需要的信息为"本体应用"和"论文"两者的结合。所以,经过推理可知:用户需要的是自然科学类的期刊、文章,或科技应用类期刊、文章。

(5) 领域本体库[6]:对于领域的资源、信息依据本体规则和公理、本体内部概念之间关系的逻辑表达进行推理处理,使其在语义上准确全面地表达领域信息的知识和事实。除描述信息本身外,还包含信息间的关系。通常由本体专家和领域专家一起来完成。领域本体构建之后,导入系统作为本体推理机在内部表达的模式。

(6) 元数据挖掘与抽取[7]:主动元数据挖掘有两重含义,一是将现在由网上信息的检索通过 URL 超链接并在用户直接参与下被动地获得用户所需的信息,改变为由系统主动地 24 小时遍历 Web,根据领域的需要、用户查询的需求,在领域本体的协助下遵循一定的策略和方法(挖掘)主动从 Web 上挖掘所需要的领域信息。不需要用户的参与,由系统自动完成。二是对挖掘的信息经过综合、过滤、抽取分别充实到元数据库和领域本体库中,以供用户在领域范围查询与检索。具体过程如下:

●以改进的 Robot(Robert *,机器人)为基础依照领域知识有针对性、有条件地搜索网站或网页,以已建立的本体为参照,把属于该领域的知识及文档带回。

●利用用户偏好库来记录用户的点击历史(选择的信息,数据发生的变化情况),修改、更新信息以及搜索界面的相关度排序情况。

●利用智能代理(Agent),根据领域用户的意愿,在用户偏好库中结合用户最关注的信息,对网络上信息实行实时性检测,发现领域新的信息则通过机器人(Robot *)增补到元数据库中。

●周期性地检查元数据库,对已经过时或没有用的信息从元数据库中删除。

(7) 元数据库:存放从各个网站、网页收集来的领域相关信息。元数据不仅包括了通常信息检索系统搜索数据库所得到的信息及文档内容,而且还隐含了信息及文档与相应领域的语义关系,为系统的语义推理、信息搜索等以后操作奠定基础。

(8) 用户偏好库:存放用户每次输入的一类信息及历史记录用户感兴趣的知识、信息、背景和专业等,是用户个性特征记录的数据库[8]。

(9) 结果整合模块:将结果集进行变形整合返回给终端用户。

4　应用实例

现以果品领域构建的本体为例,简述开发应用中的几个实例。

(1)语义标注[9]。

领域本体的构建是一种基于概念的层次方法,另一种是基于描述逻辑的方法[10]。本体论为同一应用领域的成员之间,例如人或者智能代理等,提供了统一的该领域的术语集。在此基础上,就需要一种描述对象进行概念化的表示语言。当前用于这种描述的表示语言和系统一般采用基于一阶谓词逻辑的表示方法,例如本体语言(KIF—based Ontololingua)、Loom、框架逻辑(Frame—Logic)等。虽然,这样的方法可以有不同的计算特性和表达能力,但是,对于因特网上的应用来说,更为重要的是必须有一种具备统一语法的语言,使得系统中的本体能够遵循统一的语法格式进行数据的交换。XML 即可扩展标记语言(eXtensible Markup Language),标记是指计算机所能理解的信息符号,通过此种标记,计算机之间可以处理包含各种信息的文章等。在当前因特网上已经广泛地应用 XML 标记语

言，支持不同系统之间的信息交换，XML具有良好的可扩展性，样式与内容的分离以及有可遵循的严格的语法要求等特点[11]。下面即是采用基于XML语法并且以资源描述框架语言RDF(Resouree Description Framework,)三元组的形式描述信息资源[12]，同时结合元数据库中的元数据而生成实例文件，如图2所示。

```
<网站 rdf:ID="东方林果站">
    <website:regionOf rdf:resource="#山东"/>
    <website:relatedFruit rdf:resource="#苹果"/>
    <website:fullDesc rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >引进葡萄生食品种三十五个，酿酒品种二十二个，苹果十二个品种，桃七个品种，
梨五个品种，樱桃五个品种。</website:fullDesc>
    <website:webAddr rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >http://www.dfgy-qd.com/jianjie.htm</website:webAddr>
    <website:relatedFruit rdf:resource="#樱桃"/>
    <website:relatedFruit rdf:resource="#桃"/>
    <website:websiteName rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >东方林果站</website:websiteName>
  </网站>
```

图2 RDF实例文件(部分)

(2)原代码描述。

这里引用了系统创建推理模型——getInfModel.java的原代码：

```
import com. hp. hpl. jena. rdf. model. *; import java. io. InputStream;
    public class getInfModel{
        public InfModel getInfModel ( ) {      String file = "../../ontology/Fruit. owl";
        Model data = ModelFactory. createDefaultModel( );
        Model model = ModelFactory. createDefaultModel( );
  try{
            InputStream in = FileManager. get( ). open(file);
            data. read(in, "");
        }
        catch(Exception e){
            e. printStackTrace( );
        }
        Resource configuration = model. createResource( );
        com. hp. hpl. jena. reasoner. Reasonerreasoner = GenericRuleReasoner Factory . the Instance( ). create(configuration);
        InfModel infModel = ModelFactory. createInfModel(reasoner, data);
        return infModel;
        }
    }
```

(3)检索结果。

这里引用了实现语义扩充查询的实例，即：不仅能够得到包含关键字的信息，而且能够得到与关键字同义、近义和上下位关系的信息，从一定程度上避免了

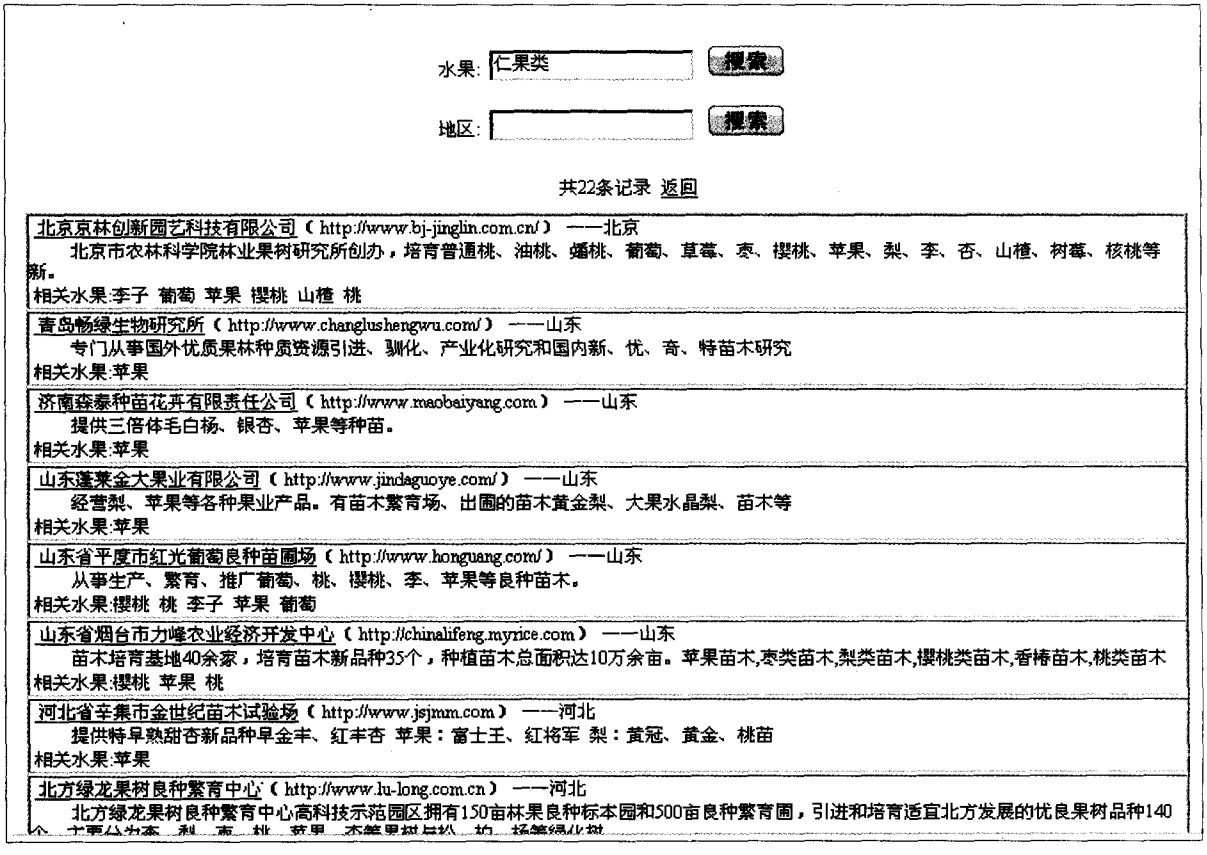

图3 语义扩充查询的实例

"信息过载"。如图3,在条件检索中输入"仁果类",与导航检索中点击"仁果类"超链接,都会出现与"仁果类"下位关系的概念,如李子、苹果等,测试结果界面(部分)如图3所示。

5 结束语

网上信息检索从用户被动地依靠浏览超级链接网页获取,到主动依靠计算机自动搜索;从依靠关键字作为查询的依据,到借助本体的作用获取语义描述的信息数据是当前信息检索研究的热门课题。文中通过基于本体主动元数据挖掘系统构建以及在果品领域的应用作了初步的尝试,在主动搜索、元数据生成、借助本体作用于数据的语义描述等方面,其效果还是客观的。但是在果品领域知识的获取、领域规则规范等方面还有待今后通过研究和实践进一步完善,进而使该系统得到充实和提高,实现最终的推广应用。

参考文献:

[1] 李宝敏,韩岳松. 本体环境下用户偏好库的查询算法扩展[J]. 西安工业大学学报,2007,27(5):480-484.

[2] 秦玄铮. 基于本体的个性化信息检索系统的设计与实现[D]. 北京:北京邮电大学,2006.

[3] 顾德访. 语义web环境下基于ontology的语义检索应用研究[D]. 南京:南京理工大学,2005.

[4] 曹志松,曹文君. 基于语义web实现有效web信息检索的研究[J]. 复旦大学学报(自然科学版),2004,43(3):422-427.

[5] 陈　杰. 主题搜索引擎中网络蜘蛛搜索策略研究[D]. 杭州:浙江大学,2006.

[6] 贾学峰,王建新,齐建东,等. 基于领域本体的智能检索模型[J]. 计算机工程,2010,36(23):174-176.

[7] 方卫东,袁　华,刘卫红. 基于Web挖掘的领域本体自动学习[J]. 清华大学学报(自然科学版),2005,5(1):1729-1733.

[8] Chen L,Sycara K. A Personal Agent for Browsing and Searching[C]//Proceedings of 2nd International Conference on Autonomous Agents. New York,USA:[s. n.],1998.

[9] 李宝敏,张　娜. 语义智能检索在果品领域的应用[J]. 西安工业大学学报,2008,28(3):301-306.

[10] Baader F,McGuinness D L,Nardi D,et al. The Description Logic Handbook: Theory, Implementation and Applications [M]. Cambridge:Cambridge University Press,2003.

[11] 李宝敏,张　娜. 基于领域本体的语义智能检索研究[J]. 情报杂志,2007(12):124-126.

[12] Verzulli J. Using the Jena API to Process RDF[EB/OL]. 2001-05-23[2007-03-10]. http://www.xml.com/pub/a/2001/05/23/jena.html.

(上接第22页)

宣告国外前缀的异常数占总数的比例约为9%。对于国外非法宣告国内前缀的异常中,国内视图未看到的异常数占总数的比例约为26%。因此,从上面的分析结果可以看出国内视图监测出异常具有不完整性,国外视图是国内视图监测的有力补充。

4 结束语

文中提出了一个基于国际视图的域间路由异常监测方法,即综合国内和国外视图监测路由异常,以弥补国内单一视图的不足。文章中设计了国家级非法宣告前缀异常监测算法和国家级多源冲突异常监测算法。通过实验结果中监测出的非法宣告前缀异常可以看出,国内视图看到的异常是不全面的,国外视图是对国内视图的有力补充。

参考文献:

[1] Rekhter Y,Li T. A border gateway protocol 4(BGP-4)[S]. IETF Internet RFC,RFC4271,2006.

[2] Oliveira R,Zhang Beichuan,Zhang Lixia. Observing the Evolution of Internet AS Topology[C]//SIGCOMM. Kyoto, Japan:[s. n.],2007.

[3] 胡湘江,朱培栋,龚正虎. 域间路由协议BGP安全性研究[J]. 计算机工程与科学,2007,29(9):5-8.

[4] FeldMann A,Maennel O,Mao Z M,et al. Locating internet routing instabilities[C]//SIGCOMM. [s. l.]:[s. n.],2004.

[5] Mayer D. University of oregon route views project[EB/OL]. 2003. http://www.routeviews.org/.

[6] RIPE RIS Project[EB/OL]. 2002. http://data.ris.ripe.net/.

[7] Telstra CIDR Report[R/OL]. 1997. http://bgp.potaroo.net/as1221/.

[8] Cymru T. The team cymru bogon route server project[EB/OL]. 2004. http://www.cymru.com/Documents/bogon-list.html.

[9] Ripe's MyASN[EB/OL]. [2008-03-04]. http://www.ris.ripe.net/myasn.html.

[10] Lad M,Massey D,Pei D,et al. PHAS:A prefix hijack alert system[C]//Proc USENIX Security Symp. [s. l.]:[s. n.],2006:153-166.

[11] 朱培栋,邓文平,刘　欣,等. ISPView一种域间路由可视化监测系统[J]. 计算机工程与科学,2008,30(2):34-36.

[12] Deng Wenping,Zhu Peidong,Lu Xicheng. ROUSSEAU:A Monitoring System for Inter-domain Routing Security[C]//Communication Networks and Services Research Conference. [s. l.]:[s. n.],2008.