本体在网页智能信息采集与检索中的应用研究

江 河
（太原大学 外语师范学院,山西 太原 030012）

【摘 要】互联网技术快速普及使得网络上的信息成几何级数的方式增长,如何找到符合使用者需求的信息,这是目前信息检索的研究热点。鉴于此,本论文研发了本体支持的混合式信息整合采集器,除了能快速整合特定领域的文件数据外,还可以进一步通过混合过滤技术选择出重要的信息进行整合排序。实验验证本系统对整个网页检索效能的信息搜集度与整合度都可达较高的应用水平。

【关键词】知识本体;Protégé;信息检索;本体推理
【中图分类号】TP391 【文献标识码】A 【文章编号】1673-1891(2010)04-0079-03

1 引言

网络信息的爆发增长使得人们在网络搜索想要的信息时,必须使用各个独立的搜索引擎,输入关键词来达成信息搜集的目的,这样的检索方式往往带来了查准率的低下。为解决上述问题,研究中有很多引用本体成为核心技术的信息系统。本体能提供信息系统完整的语意模型,包括:对象、对象属性与对象彼此间的基础知识,具有分享与重复使用的特性;也能自动地处理相关领域间的信息沟通及存取,还能进一步推理出新的知识与结果。在大量数据中找出信息重要的关系或其潜在的规则模式,提供信息系统决策的重要依据。因而本论文的研究主题在于应用知识本体技术设计出相关网页的知识本体;并通过本体建构工具建构与分析程序代码网页的知识本体类别;再配合数据库构建出相关网页关键词知识本体分享平台,进而提高网页的检索查准率。

2 开发技术探讨

2.1 本体的应用

本体(ontology)一词源自哲学理论,它通过词汇来描述一个知识领域内对象们之间的关系,在W3C已提供几个标准来明确的描述资源。

1995年,使用者Gruber提出了构建知识本体时的五条准则,以期能达到重用性及共享性。清楚:知识本体必须使被定义的术语能有效的沟通。一致:知识本体应该是一致,也就是说,它应该认可与其定义一致的推理。可扩展性:知识本体应该为共享的词汇提供被期待的概念基础。最小的编码偏好程度:概念化的描述应该在知识层被具体的描述而不应该依赖于某一种特殊的符号层的表示方法。最小的知识本体约定:知识本体应该只需要充足的最小化知识本体约定以去支持未来的知识分享活动。在系统的构建过程中,笔者也将采用上述五条准则来构建领域本体。

2.2 相关开发技术

本系统的开发工具为MyEclipse,是一个十分优秀的用于开发Java的Eclipse插件集合,功能非常强大,支持也十分广泛。本系统采用MySQL来作为本体知识数据库。MySQL是目前最常被使用的一种关联式数据库管理系统。本体建构工具Protégé是由美国斯坦福大学研究开发的知识本体自由软件。Protégé不但是目前在世界上使用来构建本体最重要的平台之一,更是全球支持本体的平台中最广为使用的一个。

3 本体构建及系统架构

由于系统内检测方式是采用将无障碍规范中的每一条检测码分别以程序代码的方式呈现,所以在检测过程中系统会先将欲检测的网页读入,再依序让拥有检测码检测逻辑的程序代码分别对该网页进行检测。

3.1 构建本体数据库

系统的本体是针对特定领域所构建的知识分享数据库,也就是利用已经构建好的“使用者”本体数据库,支持相关系统运作。使用者本体资料库的构建包括:使用者相关概念统计与分析及本体数据库的建立两阶段。首先,针对相关使用者网页的首页进行统计,进而选取使用者首页会出现的相关概念及其同义字。系统将以此作为比对与过滤信息的基础,由此判断网页是否符合搜集条件的依据。图1为本体构建的第二阶段:使用者本体数据库的构建。此部份主要是将构建于Protégé中的本体转换成MySQL数据库。

3.2 系统架构

3.2.1 网页收集模块

收稿日期:2010-10-20
作者简介:江 河(1976-),男,硕士,讲师,研究方向:多媒体资源库。

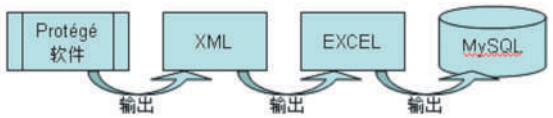

图1 使用者本体数据库转换过程

此部分的功能是输入关键词执行网页搜集的相关预处理，并将关键词转换成URI码嵌入搜索引擎的搜索网址；搜索引擎连结则宣告一个URL组件并把之前好的URI码嵌入Google搜索网址，并使用循环逐行读取回传信息，更将其输出成文字文件，作为分析时的参考；网页内容撷取再逐一检视每一URL连结，再以正确的编码方式读入该网页的html原始文件后，使用正规表达式截取网页标题，并以文件名输出成文本文件，方便信息选取器进行html语法分析，最后，去除html标签后取得纯网页内容文件，支持网页分类模块的后续处理。网页收集模块的各部分功能如图2：

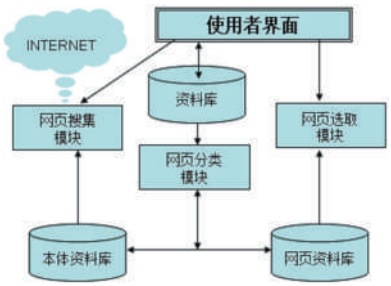

图2 系统架构图

①关键字及下载目录：执行网页搜集的预处理包括：将输出区域清空、将输入位所输入的文字转成URL码嵌入搜索网址、将预设下载预设位置的输入字串转成储存位置的文件名称，并清除该位；最后，系统将提醒使用者输入相关预设作业。

②谷歌链接：提取一个URL组件并把Action方法中转换好的URL加上搜索网址；接着使用一个缓冲读取，并使用循环逐行加入String组件，最后把line输出成文本文件，作为最后分析时的参考，文件内容即为该页面的html原始文件。

③检索链接：使用前述的正规表示法从链接中寻找是否有符合的URL，符合者会下载URL，并且会输出成文本文件提供系统做进一步的处理。

④检索内容：使用匹配方法来判定该网页是否为笔者所设定的网页搜集范围；如果是，就将html原始文件之中的html标签删除，只剩下纯文字内容，方便系统做进一步的后续处理与分析。最后，把搜集过的网页数除以总网页数，该值为目前的搜集进度。

3.2.2 网页分类模块

网页分类模块的主要功能是当使用者网页输入后，还要做剔除无意义的字符等预处理工作，预处理还包括：加载本体数据库、正规化数据库及加载文件列表供后续处理。基于词汇间的组合方式不同就会有截然不同的意义，因此，断词就是个很重要的工作。网页文本交由断词系统处理中文断词，断词后的网页内容包括断词词汇及属性卷标。分类的预处理则包括：格式转换、断词修正、字根还原等，这样可以方便分类处理并补足中文断词系统对于特定领域的断词问题，也可减少文件内容干扰，加强分类效能和查准率。

3.2.3 网页选取模块

系统将网页收集模块与网页分类模块内的资料相结合，获取文件网页html原始文件、URL、文件名等进行重要信息的选取作业。预处理主要包括：URL修补与网页分页撷取。基于网页分页的URL可能写成网站内部链接的形式。网页的特定信息或许存在特定分页中，为此，必须选取特定分页。正规化处理则利用正规表达式针对系统所需重要信息进行分析与选取。本系统通过HTML语法分析卷标选取出特定的信息。最后，将正规化处理后的文件，依照不同类别输出成文字文件。

4 系统验证

为验证本研究开发的系统模型确实能改进网页的检索效率，采取了通过前端浏览接口进行实测评估。评估目标是针对本研究系统运用本体辅助所产生的关键词、多文件摘要、主题结果、网页合并处理、本体分类概念检索等项目，期望能取得比先前研究进步的成果。

笔者以本体词典中的多个关键词进行相关网页搜集为例，并与在谷歌中检索结果进行对比。本系统依然沿用前面研究的变量来评估系统的可用性网页准确率（Precision Rate，RP）及检索率（Recall Rate，RR）。RWT表示所有的检索网页数；RWC为正确的回传网页数；RWR则为相关回传网页数。表1是经领域专家逐一比对回传页面后，得到谷歌的准确率及检索率分别为6.5%与63%，以及本系统输入同样关键词后所得结果。从表中数据比较看，除了显现本系统确实在领域本体相关网页搜集上，比搜索引擎谷歌有较高的准确率及检索率外；也展现及验证本论文提出技术的可行性。

表1 搜索比较结果

	RWC	RWR	RWT	RP	RR
Google	32条	10条	507条	6.5%	63%
本系统	45条	3条	49条	96%	100%

5 结束语

本研究继续先前研究成果（基于本体的网页收

集模块2010),针对网页在检索时的查准率不高,不能较好提供检索结果的问题,提出有效的改善方法。在领域本体的进一步建构下,解决了导致上述问题产生的原因:"检索结果受限于基于关键词语意为主的处理方式"。鉴于名称、内容、URL及时间词在网页本体中扮演重要角色,本研究进一步开发一套网页关键词断词辨识处理系统,能有效将网页文件中的名称、内容、URL及时间词选取出来,作为重要主题关键词;本研究还开发出改良网页本体论架构,开发出事件合并处理机制与重要关键词选取系统,增加事件内容的呈现丰富性,此机制能提升使用者在网页检索事件上的便利性。

注释及参考文献:

[1]武成岗,焦文品,田启家,等.基于本体论和多主体的信息检索服务器[J].计算机研究与发展,2001(6):641-646.

[2]郁书好,郭学俊.基于本体的教学知识库研究与应用[J].计算机研究与发展,2007(8):161-164.

[3]丁晟春,顾德访.Jena在实现基于Ontology的语义检索中的应用研究[J].现代图书情报技术,2005(10):5-9.

[4]黄敏,赖茂生.语义检索研究综述[J].图书情报工作,2008(6):63-66.

Research of the Application of Ontology in Web Intelligent Information Collection and Indexing

JIANG He

(School of Foreign Language Normal College, Taiyuan University, Taiyuan, Shanxi 030012)

Abstract: After the broad popular of Internet technology, the information of Internet does geometric-progressively increase. How to search advantage information to meet users' demands in the internet information torrent of Internet has become the first goal of lots of scholars to make efforts on. For that reason, this paper focused on developing an ontology-supported information integration and recommending system for scholars. Not only can it fast integrate specific domain documents, but also it can extract important information from them through the hybrid filtering technology to take information integration and recommendation ranking. Experimental verification for this system performance on the entire page of information retrieval and integration is possible to collect up to a higher level of application.

Key words: Ontology; Prot é g é ; Information retrieval; Ontology reasoning