

语义检索及其关键技术研究

张娜, 李宝敏

(西安工业大学 计算机学院, 陕西 西安 710032)

摘要:传统的基于关键字的搜索引擎由于忽视了关键词本身所含的语义信息而得到较低的查全率和查准率。文中结合万维网的具体特点,提出了基于语义万维网的智能信息检索系统的功能结构,详细描述了智能信息检索系统的设计思想和检索流程。并且对智能检索模型中所涉及到的若干关键技术进行了分析研究。为智能信息检索系统的顺利实施奠定了良好的基础。

关键词:本体;语义检索;语义网;智能信息检索

中图分类号:TP391.3

文献标识码:A

文章编号:1673-629X(2006)11-0022-04

Research of Semantic Retrieve and the Key Technologies

ZHANG Na, LI Bao-min

(Institute of Computer Science and Engineering, Xi'an Technology University, Xi'an 710032, China)

Abstract: Because of ignoring the semantic information inside the keywords, the traditional searching engine based on the key words has lower recall and precision. Proposes a new intelligent information retrieval function structure based on the semantic Web as a result of some detail characters. Also describes design idea and process of the intelligent information retrieval system Web in detail. Moreover, aims at key techniques of intelligent system which establishes the theory basis for implementation of an intelligent retrieval system.

Key words: ontology; semantic retrieve; semantic Web; intelligent information retrieve

0 引言

目前,Internet上的搜索引擎主要是基于关键字的查找,这种方法带来了以下缺点^[1]:

1) 用户需要输入关于查找的精确语义的关键字,否则搜索引擎会返回很多大量无用的信息。

2) 由于Internet上的信息是多种多样并且是不断变化的,搜索引擎的索引数据库必须能针对这些不同的Web站点做出不同的反应。

3) 由于Internet上不同种类的信息所建立的索引数据库也不同,每个搜索引擎能独立完成搜索,所以这种彼此独立的数据库必然会存储重复的信息,搜索效率较低。

4) 搜索引擎根据用户所输入的关键字返回的是相关的URL地址,用户通过超链接来获得具体的信息。

针对目前搜索引擎存在的诸多缺点,文中提出了一个基于语义的智能检索系统,旨在提高查找的效率。

1 语义检索系统模型

1.1 研究的背景

目前国内外对搜索引擎的研究已经成为一项热门课

题,尤其在智能搜索方面。实际应用上也研制出一些大家熟知的搜索工具,比如网络信息采集大师、hohojob等。但这些搜索工具存在的关键问题在于所能搜索到的信息量不是很大,查全率和查准率需要进一步的提高。

1.2 研究的主要内容

在服务器端,其应用程序实现在线搜索,并将搜索结果存入数据库中。客户端输入检索信息,可以是关键字,也可以是自然语言或嵌套的模式语言。将结果提交给检索器,检索器按照本体要求,经格式转换后提交给推理机,经过推理判断、语义分析后,得到用户准确的语义。然后在本体的帮助下查询数据库中与其相匹配的符合条件的数据集,经整合、格式定制后返回给用户。

1.3 拟解决的关键问题

(1) 通过搜索引擎搜集到的信息量是很大的,要将这些信息分类整理后存入数据库中,需要按照一定的格式,通常是按照数据表的格式列出。

(2) 对检索到的信息通常是在语义的基础上进行分类,用户可以提取出精确语义的信息,提高检索的效率。

(3) 应及时地保持数据库的更新,对于长期没有用到的信息应使用淘汰策略进行淘汰。

(4) 领域本体库的构建,需要在领域专家和领域专家的协助下,捕获相关领域的知识,提供对该领域知识的共同理解,确定该领域内共同认可的词汇,并从不同层次的形式化模式上给出这些词汇(术语)和词汇之间相互关

收稿日期:2006-07-12

基金项目:国家“星火计划”资助项目(2004EA850069)

作者简介:张娜(1980-),女,河南人,硕士研究生,研究方向为语义网络;李宝敏,教授,研究方向为计算机系统结构、网络和语义网。

系的明确定义^[2]。

1.4 语义检索功能结构

本课题在以往信息检索的基础上增加了智能检索的功能,本体(ontology)能将领域中的各种概念及其属性、概念之间的关系表达出来,在语义方面发挥着重要作用^[3]。其语义检索功能结构如图 1 所示。

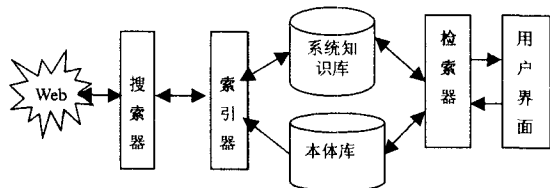


图 1 智能信息检索结构

系统的主要功能模块如下:

(1) 用户界面。输入用户的查询信息,显示查询结果,提供用户反馈机制。

(2) 搜索器。其功能如下:

① 利用主页中的超文本连接遍历 Web,发现和搜集信息,一般是根据领域的需要,有条件地搜索(即:基于主题的搜索)。

② 24 小时全天搜索,周期性地将信息返回给索引器,在周期搜索的同时,对已搜索爬行过的不予理睬,对已经更新的网页重新搜索。对已删除的网页通知索引器,删除信息库中该 URL 的相关信息。

(3) 索引器。其功能如下:

① 将从 Internet 搜索到的信息结合领域知识的相关主题分类,如工程师、高级工程师、教授等,并将分类化的信息以及按照领域知识要求的内容甚至表示的形式写入数据库。

② 写信息是对网页中的关键词中的异名同义词,如电脑、计算机、微电脑等,和语义相近的词写入数据库,以供在检索器工作时使用。

(4) 检索器。其功能为:

① 将用户输入的所需要的查询信息结合本体库规则进行充分的理解和推理,在语义上完全正确地表达用户的需求。

② 将推理的精确语义的信息送搜索器。

③ 搜索器从系统知识库中查找,查找时依据索引器分类化的信息进行检索,将检索到的内容返回用户界面,没有找到则通知搜索器进行再次搜索,最终将结果返回给用户。

(5) 本体库,即领域知识库。按照领域要求提供领域知识的词典、规则,为语义表述匹配提供依据。

(6) 系统数据库。用来保存按照领域知识要求搜索得到的网页及信息。

2 系统所使用的关键技术

2.1 关键技术

采用基于 Web 的三层 B/S 结构,数据服务器主要负

责原始信息的收集以及对原始信息的加工处理等,最终得到有关原始信息的系统知识库(也称元数据库);Web 终端主要是获取用户的查询请求、查询条件处理;Web 服务器通过搜索引擎查询系统库、对检索的结果进行排序以及将检索的最终结果返回给用户。其所涉及到的关键技术如下:

(1) 信息资源的收集。信息库是信息检索系统的基础设施之一,在信息检索时,首先要确保信息库中存在有足够多的可供检索的信息,然后才能考虑如何有效地检索。信息源有数据库、网页和文档等多种形式。由于实际工作导致了信息的多样性和信息存储分布性。因此,为了确保信息检索的性能,必须通过一个信息检索器事先将分布在各种存储媒质中信息搜集到系统库中。传统的信息收集由 robot 程序自动地获取每个站点上有用的信息和文档,按照“活”站点获取的不同分为盲目检索和定向检索,由于智能检索系统针对的是专业领域,它在进行原始信息的收集时,采用的是定向检索,一般选择的是专业网站作为信息搜索的起点,根据宽度、深度优先和启发式信息获取算法在万维网上循环地收集信息。

(2) 本体库的创建。在实践中探索了不同的方法路径,比如:Uschold 与 King 方法,Grüniger 与 Fox 方法、METHONTOLOGY 方法等,这些方法都体现了本体的 4 层含义,即:概念模型,抽象出客观世界中的相关概念而得到的模型;明确,所使用的概念及使用这些概念的约束都有明确的定义;形式化,ontology 是计算机可读的;共享,ontology 中体现的是共同认可的知识,反映的是相关领域中公认的概念集。目前被语义 Web 研究者们所广泛接受的创建 ontology 一般有以下几个步骤^[4]:确定 ontology 覆盖的领域和范围;考虑使用现存的 ontology 资源;列出 ontology 中的重要词汇;确定类和类的层次关系;确定类的属性;确定对属性值的描述;创建实例;检查一致性。

(3) 语义推理。语义推理按照语义 Web 所处的层次不同,分为公理推理和定理推理^[5]。公理推理是建立在人们对客观世界的共同的认识之上,常常是有关常识性知识的推理,定理推理是从具体的逻辑规则出发推理出相应的结论。在语义检索中,公理推理是通过子类、子属性、属性定义域、属性值域、基数限制和互不相交等规范化的术语来实现的,它通过专门的通用处理程序来实现。如 Jena 所提供的本体推理方法。定理推理也要通过专门的处理程序来实现,在语义检索领域运用不是很广泛。

2.2 检索思路

语义 Web 环境下的语义检索。这种思路主要是在语义 Web 环境下基于面向特定领域的 ontology 来实现语义检索,对信息资源标引的结果一般是以 RDF(Resource Description Frame)三元组的结构存储在基于 XML 语法的 RDF 文档中,在具体处理的过程中可以对 RDF 文件进行解析从而创建 RDF 模型,这种模型可以存储在计算机内存中也可以依照 RDF 三元组的结构存储在关系数据库

中。这个过程的信息检索使用的是面向 RDF 模型的特定的检索语言 RDQL(RDF Data Query Language)。

具体来讲,对于网络信息资源中的语义检索过程如下:

(1)首先是提取文档的元数据,描述文档数据的数据为元数据,智能信息检索系统中的文档元数据是按照领域本体的结构进行组织安排的,不仅反映了该文档的内部信息,而且还反映了该文档和其他文档之间的关系。如描述领域的上下位关系、相似关系等。因此,可以说智能信息检索系统中的元数据不仅囊括了传统信息检索系统的索引数据库所能描述的文档内容信息,而且还体现了文档与具体领域的语义关系,为语义推理、信息检索等后续操作奠定了基础。

(2)其次是对文档的元数据进行语义编码,使用 W3C 所发布的资源描述框架(RDF)作为元数据编码的参考模型。对从 XML 文档中提取出来的文档特征短语编码成 RDF/XML 格式,以便于计算机高效地处理这些元数据。RDF 旨在描述事物与事物之间的联系,RDF 的核心是三元组,即,任何复杂的事物描述均可以描述成一系列的三元组。这与哲学上“联系是永恒存在的”的思想相吻合^[5]。

(3)再次是元数据的语义处理,它是根据领域本体和推理规则来完成对有关元数据的推理处理,得出隐含的信息。语义推理的过程就是让计算机识别和理解领域本体的结构和元数据信息,并根据相关的逻辑规则对现有信息的闭包^[5]。以经过语义编码的元数据为推理的起点,根据规则进行扩充以求得其所蕴涵的更为丰富的信息。

(4)最后,根据所建立的领域本体对用户提出的查询条件进行规范化处理,即查询条件预处理,此外还要对查询条件进行编码,编码的过程和语义编码的过程类似,即在领域本体组织框架的指导下,按照资源描述框架模型将经过预处理的查询条件序列转化为 RDF/XML 的查询表达式。

这样信息语义检索的过程在以上过程的配合下,仅仅只需要将经过处理后的查询条件和元数据库中的信息进行匹配,将满足条件的元数据选出。并将检索的结果经整合、格式定制后返回给用户。

3 系统的实现

3.1 系统构建

系统构建如图 2 所示。

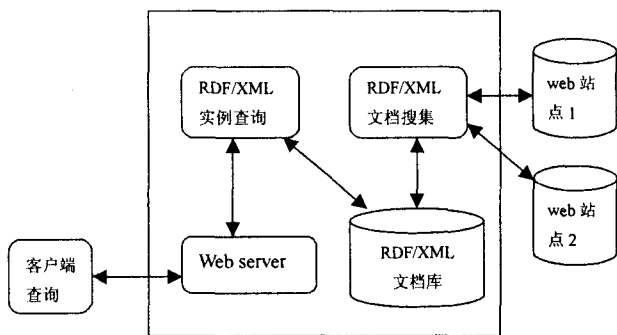


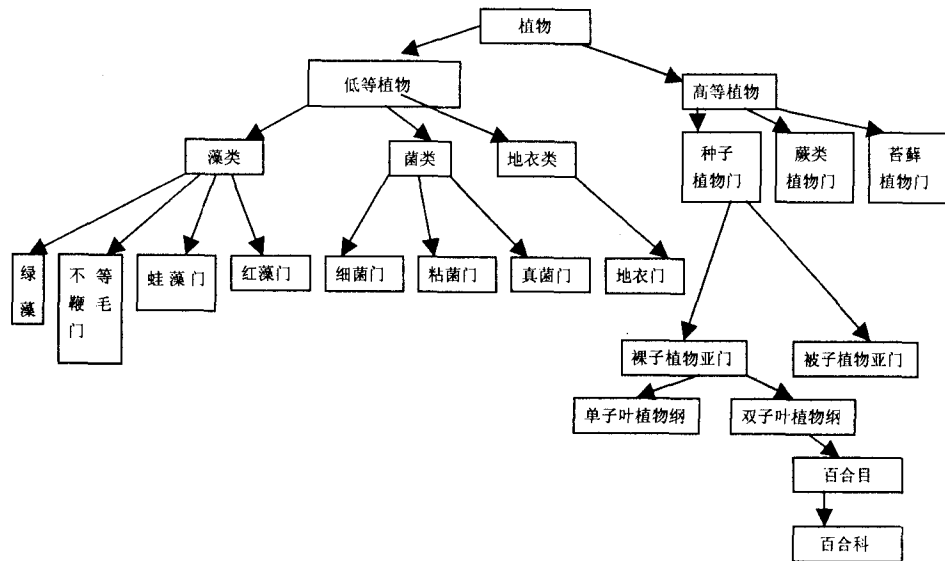
图 2 语义检索网站结构

RDF/XML 实例查询用 JavaBean + JSP + Servlet, 读入并解析 RDF/XML 文件, 然后读入文档库中的多个 RDF/XML 文件, 并向 JSP 服务器提供查询功能; RDF/XML 文档搜集, 主要是负责从相关网站上搜集符合该网站中 RDF/XML 的实例文档, 将这些文档存入文档库; Web server 采用了 Tomcat 作为 JSP 服务器, 提供给用户查询界面及将结果返回给用户; RDF/XML 文档实例库是所搜集到的 RDF/XML 实例文档, 它是对各网站所有的资源的规范描述, 能为计算机所理解和查询。

3.2 本体模型的建立

本体是由一些标准的术语的词汇表以及描述术语语义的规范组成^[3], 作为一个实例, 这里描述了农业植物领域的本体模型的构建方法。

植物本体如图 3 所示。



为了表示这样一个模式体系, 文中选用了 RDF/XML, 下面是用 RDF/XML 描写的植物模式的部分内容:

```
<rdf:RDF
  xmlns:rdf = http://www.w3.org/1999/02/22 - rdf - syntax
  - ns#
  xmlns:rdfs = http://www.w3.org/2000/01/rdf - schema#
  xml:base = http://localhost/ns>
  <rdfs:Class rdf:ID = '植物' />
  <rdfs:Class rdf:ID = '低等植物'>
```

```

< rdfs: subclass rdf: resource = ' #植物' />
< /rdf: Class>
< rdfs: Class rdf: ID = '高等植物'>
< rdfs: subclass rdf: resource = ' #植物' />
< /rdf: Class>
< rdfs: Class rdf: ID = '地衣类'>
< rdfs: subclass rdf: resource = ' #低等植物' />
< /rdf: Class>
...//以上是描述植物类的层次关系
< rdfs: data rdf: about = "http://www. w3. org/2001/
XMLSchema# String"/>
< rdfs: Property rdf: ID = "name">
< rdf: domain rdf: resource = " #植物"/>
< rdf: range rdf: resource = "&xsd: string"/>
< /rdfs: Property>
...//描述类的属性体系
< /rdf: RDF>

```

有了植物资源的模式,就可以依据它建立对具体植物的描述,即实例文件,比如对资源 <http://localhost//ns # plant. jsp> 的描述如下:

```

< ? xml version = "1. 0"? >
< rdf: RDF
  xmlns: rdf = http://www. w3. org/1999/02/22 - rdf -
syntax - ns #
  xmlns: terms = http://localhost/ns>
< rdf: Description rdf: about = "http://localhost//ns # plant.
jsp">
  < terms: name>植物< /terms: name>
< /rdf: Description>
< /rdf: RDF>

```

(上接第 21 页)

Servlet 接受 execute 方法返回的 ActionForward 对象,转发到 Action Forward 指定的源。这个源可以是 JSP 页面、其它 Action 或另一个 Servlet。

3 结 论

使用 DAO 模式和业务代理联合模式整合两个框架,在最大程度上实现了模型层与数据持久层的有机分离,较好地实现了层之间的解耦,使得应用系统层次关系清晰。应用新框架开发 Web 系统,其步骤明确易学,缩短了整个 Web 系统的开发周期,起到了快速有效地构建 Web 应用系统的作用。

框架整合是一个反复的过程,需要在应用中不断发现其缺点,在改进中逐步成熟。文中所构建的新框架仍有待于进一步改进。例如,可增加可视化机制用于解决 Struts 与 Hibernate 配置文件编写复杂的问题;也可把新框架与其它软件工具(分析、测试等)的集成建立统一的开发环境,进一步减轻开发者负担。

4 结 论

在对网络资源的检索中,由于语义方法要求用规范化的方法描述网上的资源,并且对所有语义概念的联系给出了机器可以理解的形式化描述,所以只要相关的专业网站采用这种标准模式对资源进行描述,就可以让用户有效地检索资源,而采用传统的文本搜索方式时,由于没有统一的术语,搜索效率很大程度上取决于不同的网页编辑者对资源描述的详细程度和搜索者本人的经验,所以检索的效果就不太理想。

语义 Web 是近年来提出的万维网的新模式,要建立全球的语义 Web,首先是在各个专业领域内建立其本体模型,为专业网站加注语义,然后用本体描述语义的解析器 Jena 和 JSP 技术建立网站,提高网上信息检索的效率,通过实验证实,表明这种方法是可行的。

参考文献:

- [1] Chen Junjie, Liu Lizhen, Song Hantao, et al. An Intelligent Information Retrieval System Model[C]//Proceedings of the 4th World Congress on Intelligent Control and Automation. Shanghai: [s. n.], 2002: 2500 - 2503.
- [2] Li Wenjie, Feng Zhiyong, Li Yong, et al. Ontology - Based Intelligent Information Retrieval System[C]//CCECE 2004 - CCGEI 2004. Niagara Falls: [s. n.], 2004: 373 - 376.
- [3] 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(6): 422 - 427.
- [4] 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.
- [5] 邹景华. 语义万维网在智能信息检索中的应用研究[D]. 重庆: 重庆大学, 2005.

参考文献:

- [1] Tate B. 轻量级开发的成功秘诀[EB/OL]2005 - 10. <http://www - 128. ibm /developworks/cn>.
- [2] 李 扬. 扩展与整合 Web 应用框架的研究与实践[D]. 西安: 西安建筑科技大学, 2005: 29 - 33.
- [3] 冯国土. 基于 Struts 与 Hibernate 集成架构的项目管理系统[J]. 计算机应用, 2005, 25(8): 1884 - 1889.
- [4] 黄烟波. 基于 Struts 和 Hibernate 的 J2EE 架构[J]. 计算机时代, 2004(10): 29 - 30.
- [5] 孙卫琴. 精通 Struts: 基于 MVC 的 Java Web 设计与开发[M]. 北京: 电子工业出版社, 2004: 135 - 138, 155 - 156.
- [6] Crawford W, Kaplan J. J2EE™ Design Patterns[M]. [s. l.]: O'Reilly Media, Inc, 2003.
- [7] Example # 1: Struts with Hibernate[EB/OL]2000. <http://homepage. mac. com/edhand/projects/java/example1. html>.
- [8] 薛 冰. 设计模式和数据持久层框架在 Web 系统中的应用[J]. 天津理工学院学报, 2004(3): 72 - 74.