基于茶虫害本体的智能检索系统设计

朱利君,张友华,李绍稳,辜丽川,程波波
(安徽农业大学 信息与计算机学院,安徽 合肥 230036)

摘 要:目前信息检索正在向着领域化、智能化方向发展。针对农业领域的分支领域茶虫害为研究对象,通过构建茶虫害领域本体,设计并实现了基于茶虫害本体的智能检索系统。该系统在特定主题的限定下进行信息的搜集和检索,能识别所搜索的网页与主题是否相关,而且能产生智能导航链接在主题最相关的范围内搜索,把信息检索从目前基于关键词层面提升到基于知识的层面,从而能够帮助用户更容易地找到感兴趣的信息,提高信息服务的质量和信息检索的准确率。

关键词:茶虫害本体;概念匹配;本体检索
中图分类号:TP302.1 **文献标识码**:A **文章编号**:1673-629X(2010)06-0130-03

The Design of Intelligent Retrieval System Based on Tea Pest Ontology

ZHU Li-jun, ZHANG You-hua, LI Shao-wen, GU Li-chuan, CHENG Bo-bo
(School of Information Science and Technology, Anhui Agricultural University, Hefei 230036, China)

Abstract: At present information retrieval is developing to the direction of domain and intelligence. With the tea pest of agricultural sub-field as the research object and the construction of the tea pest domain ontology, an intelligent retrieval system is setup based on the ontology, which can collect and retrieve information with a limited subject, and can give the relationship between the collected website and the subject, and the link of intelligent navigation can be produced to retrieve information under the scope of the most relevant subject. It upgrades the information retrieval from the level of keywords-based to the level of knowledge-based. The system can help users to find the interesting information more easily, and enhance the quality of information services and the accuracy of information retrieval.

Key words: tea pest ontology; concept matching; ontology retrieval

0 引 言

传统检索大都是基于关键词匹配的技术,参与匹配的是字符的外在形式,而不是关键词要表达的含义,以致于检索出来的结果不全,不能得到用户满意的答案。关键问题在于大部分 Web 是基于 HTML、XML 等无结构或半结构的数据,计算机不能理解它们的语义。目前信息检索正在向着领域化、智能化方向发展。领域化智能检索为特定领域的信息服务提供了更专业、具体的帮助。基于本体的语义网技术在领域内智能搜索方面的研究是目前信息检索技术研究的热点。

Tim Berners-Lee 在 XML2000 会议上正式提出语义网(Semantic Web)[1],并为未来的 Web 发展提出了基于语义的体系结构,但实现起来却是一项复杂而浩大的工程,到目前还处在发展中。知识是各子领域知识的集合,针对多个相关领域逐个构建其中的单个领域本体,在概念、语义层面上描述各领域的知识,再通过本体合并技术组成较大本体库,最终实现语义网的目标。由此,文中针对农业领域的分支领域茶虫害为研究对象,构建茶虫害领域本体,设计并实现了基于茶虫害本体的智能检索系统,该系统由于在特定主题的限定下进行信息的搜集和检索,能识别所搜索的网页与主题是否相关,而且能产生智能导航链接在主题最相关的范围内搜索。把信息检索从目前基于关键词层面提升到基于知识的层面,能够帮助用户更容易找到他们感兴趣的信息。

收稿日期:2009-10-13;**修回日期**:2010-01-18
基金项目:中国高技术研究发展(863)计划(2006AA10Z249);国家自然科学基金(30800663);安徽省科技攻关项目(8010302170)
作者简介:朱利君(1984-),男,安徽淮南人,硕士研究生,研究方向为本体、推理;张友华,博士,副教授,研究方向为人工智能。

1 茶虫害本体的构建

1.1 设计思路

引入领域专家采用生物分类方法-"界、门、纲、目、科、属、种"来对茶树昆虫进行分类。由于茶树昆虫同属于一个"纲",所以在构建本体的过程中,从"纲"的

下一层——“目”开始来对茶树昆虫进行分类,提出了基于“目”、形体特征和习性特征的茶树害虫分类和领域本体构建方法,使所构建的本体能够正确、明确地区分出茶树的各种害虫[2,3]。利用 Protégé 工具建立了一个的茶虫害本体,完善了顶层概念库,即分类本体,运用本体思想和方法组织茶虫害领域的概念、属性、实例,共构建了 9 个目,44 个科,19 个对象属性,112 个实例,用以表示概念层次、实例声明以及概念实例间的复杂语义关系。

1.2 本体表示

OWL[4](Web Ontology Language)是 W3C 最新推荐的 Ontology 表示语言,是在 WWW 上发布和共享 Ontology 语义标记语言。OWL 有三种不同表达能力的子语言:OWL Lite、OWL DL、OWL Full。

采用的是 OWL DL,使用 owl:Class 声明类,使用 rdfs:subClassOf 表示类之间的层次关系。rdf:Property 用来说明实体间或者从实体到数据值的关系,OWL 的简单属性有 ObjectProperty 和 Datatype - Property。通过 rdfs:subPropertyOf 表示属性的层次关系,通过 rdfs:domain 和 rdfs:range 对二元属性施加限定,通过 TransitiveProperty 声明传递属性,通过 owl:inverseOf 声明逆反属性,通过 SymmetricProperty 声明对称属性,通过 FunctionalProperty 声明属性只有一个单一值。OWL 属性的约束非常丰富,allValuesFrom 指对于每一个有指定属性实例的类实例,该属性的值必须是由 owl:allValuesFrom 从句指定的类的成员;someValuesFrom 指至少有一个是由 owl:someValuesFrom 从句指定的类的成员[5,6]。

```
<owl:Class rdf:about="#半翅目">
<owl:disjointWith>
<owl:Class rdf:about="#等翅目" />
</owl:disjointWith>
<owl:equivalentClass>
<owl:Class>
<owl:intersectionOf rdf:parseType="Collection">
<owl:Restriction>
<owl:onProperty>
<owl:ObjectProperty rdf:about="#hasForeWings" />
</owl:onProperty>
<owl:allValuesFrom rdf:resource="#革质基部和膜质端部的前翅" />
</owl:Restriction>
<owl:Restriction>
<owl:onProperty>
<owl:ObjectProperty rdf:about="#hasForeWings" />
</owl:onProperty>
<owl:allValuesFrom rdf:resource="#半鞘翅" />
</owl:Restriction>
<owl:Class rdf:about="#茶虫害" />
</owl:intersectionOf>
</owl:Class>
</owl:equivalentClass>
</owl:Class>
```

2 基于茶虫害本体的智能检索系统

2.1 功能和系统框架

基于茶虫害本体智能检索系统的主要功能包括:智能检索和分类导航。智能检索主要是利用推理引擎和 Google Ajax Search API 得出与用户输入的关键词更准确的搜索结果。分类导航主要是利用 XML DOM 和 Xpath 技术解析茶虫害本体库,推理出与用户输入的关键词所有相关的概念节点,以分类的形式表示概念间的层次关系,并以链接的形式供用户继续搜索。

系统框架图[7]如图 1 所示。

系统运行效果如图 2 所示。

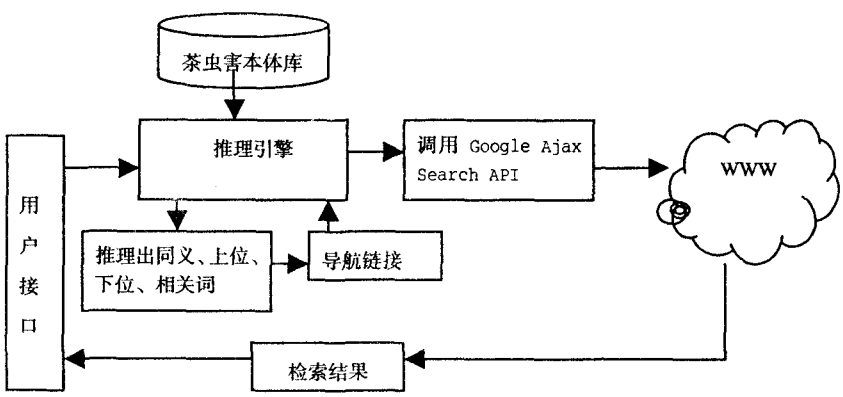

图 1　系统框架图

2.2 领域本体解析

系统采用了 XML DOM 和 Xpath 技术解析茶虫害本体库,System.Xml 命名空间使用 XmlDocument 或 XPathDocument 类提供内存中 XML 文档、片断、节点或节点集的编程表示形式。XPathDocument 类使用 XPath 数据模型提供 XML 文档在内存中的快速只读表示形式。XmlDocument 类提供实现 W3C 文档对象模型(DOM)级别 1 核心和核心 DOM 级别 2 的 XML 文档在内存中的可编辑表示形式。这两个类均实现 IXPathNavigable 接口,并返回 XPathNavigator 对象,用于选择、计算、浏览和(在某些情况下)编辑基础 XML 数据[8]。

本系统为实验系统,没有独立的搜索引擎,通过采

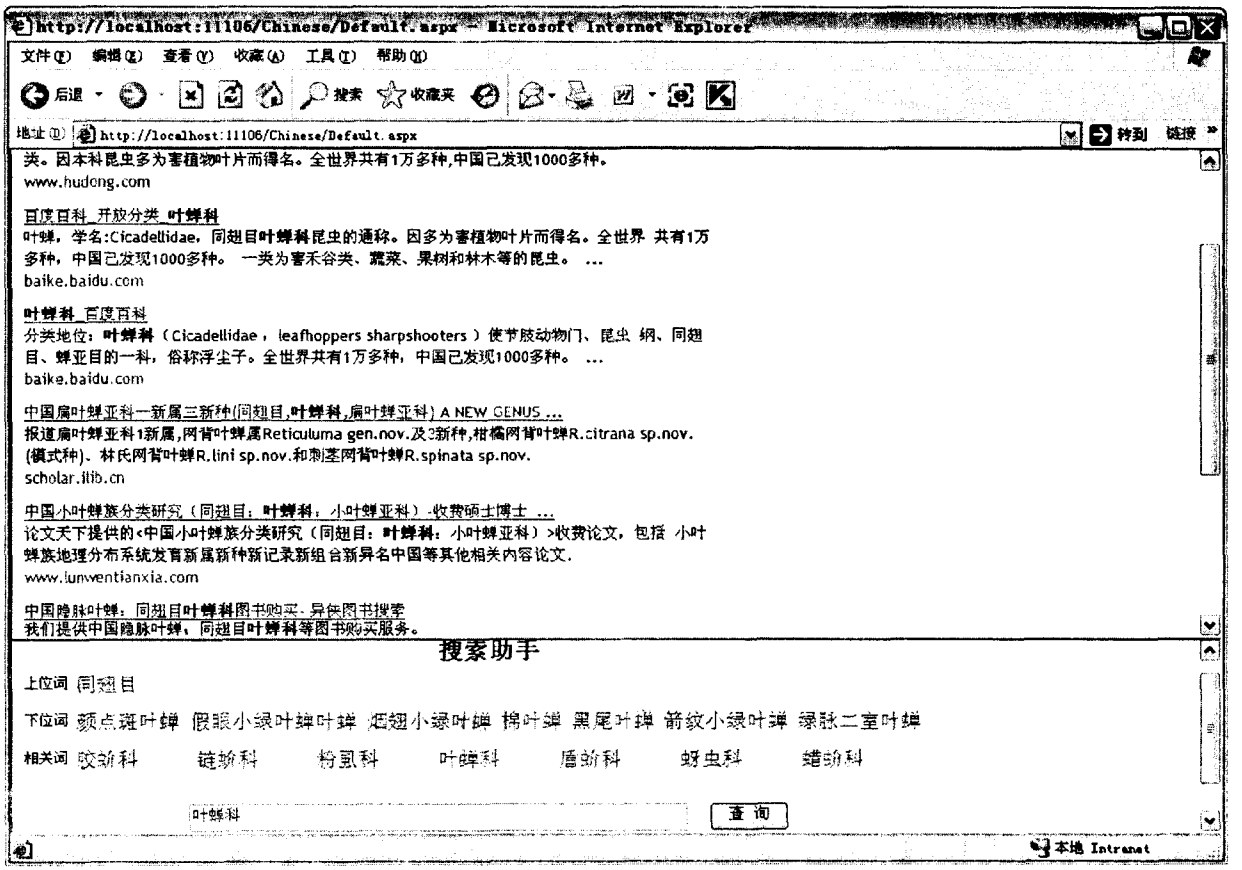

图 2 系统运行效果

用 Google Ajax Search API 技术来检索。Google Ajax Search API[9]是一个 Javascript 库,可以使 Google 搜索嵌入您的网页和其他网络应用程序中。

下面以在茶虫害本体中查找检索词的上、下位词为例,对上、下包含层次关系进行推理。代码片段如下所示:

```
DataSet mySetUp = new DataSet(); //上位词表
DataSet mySetDown = new DataSet(); //下位词表
XmlDocument doc = new XmlDocument();
doc.Load("*.owl"); //装在 XML 数据库文件
//加载命名空间
XmlNamespaceManager myNamespace = new XmlNamespaceManager(doc.NameTable);
myNamespace.AddNamespace("rdf","http://www.w3.org/1999/02/22-rdf-syntax-ns#");
//查询以用户输入的关键字命名的节点
XmlNodeList myNodeList = doc.SelectNodes("用户输入的关键字",myNamespace);
if(myNodeList.Count>0) {//找到该节点
myNodeList = doc.SelectNodes("上位词",myNamespace); //上位词
foreach(XmlNode temp in myNodeList) {
upWord=temp.Attributes.GetNamedItem("上位词").Value;
mySetUp.Tables["上位词"].Rows.add(upWord); } //加载到上位词表中
myNodeList = doc.SelectNodes("下位词",myNamespace); //下位词
foreach(XmlNode temp in myNodeList) {
downWordd=temp.Attributes.GetNamedItem("下位词").Value;
mySetDown.Tables["下位词"].Rows.add(downWord); } //加载到下位词表中
}
```

3 结束语

文中通过设计和实现了基于茶虫害本体的智能检索系统,在茶虫害领域知识内能够做到一定程度上提高了检索结果的准确率,并能以分类导航的形式显示概念层次关系,便于用户进行进一步检索。但系统有领域知识的局限性,下一步工作需要继续扩大领域本体知识内容,及完善茶虫害本体库中概念间的语义关系,使得具有更强大的推理功能。

参考文献:

[1] Berners-Lee T, Hendler J, Lassila O. The Semantic Web [J]. Scientific American,2001,284(5):34-43.

[2] 朱利君,张友华,李绍稳,等. 基于描述逻辑的领域本体知

(下转第 178 页)

数可写为：

produceFromURL("http://localhost:8081/myprj/ShowDetail.jsp?id=166",request.getRealPath("/JT/WX /Detail166"))

项目中实际应用中，在后台管理增加修改或删除记录时，随之进行记录的页面生成操作，特别地，为了顺利生成某条记录的静态页，在服务端后台增加或修改记录时，当时获取记录的 id，便写入 sourcePath 和 destPath 参数，此时便能生成具体的某条记录的静态页面。

例如，在增加或修改记录时只需这样调用：

produceFromURL("http://localhost:8081/myprj/ShowDetail.jsp? id=" + id, request.getRealPath("/JT/WX/Detail") + id + ".htm")，便生成了该记录的静态页面。

在浏览器客户端，若用户浏览首页，直接链接到首页静态页的地址就可以；若用户点击相应的信息标题，需要通过链接地址找到静态页面的路径，在首页中信息标题的链接设置为相应子页所在路径，例：<A href=""/JT/WX/Detail<%=id%>.html">，便能链接到相应的静态子页。

3 结束语

现在，静态页生成技术已经很成熟，应用也日益广泛，许多大型网站都相应采用这种技术。在设计网站时，在 JSP 页面时可以进行处理，先检测是否存在相应的静态页面，如果有则转向，没有则立即生成，然后转向，这个检测的频度可以用临时文件处理。这样的话，即使后台有些页面没有实现或者不是很好实现静态页面与实时数据的影响问题，也可以生成静态页面，从而起到提高网页访问速度的目的。

参考文献：

[1] Morrison M, Morrison J, Anthony. Keys Integrating Web Sites And Database[J]. Communications of the ACM, 2002, 45(9): 81-86.

[2] 白金牛，李慧萍，王培吉. ASP.NET 下利用动态网页技术生成静态 HTML 页面的方法[J]. 计算机应用与软件，2008(1):79-81.

[3] Eldridge L. Dynamic vs Static web Pages[EB/OL]. 2001-07-05. http://www.loriswebs.com/dynamicstatic.html.

[4] Knight J. Knightnet Site Design - Static vs Dynamic Web Pages[EB/OL]. 2008-07. http://www.knightnet.org.uk/site-design/static-vs-dynamic-pages.htm.

[5] Scanlon V. Dynamic Or Static Web Pages, Which Way Should You Go [EB/OL]. 2006-11-30. http://ezinearticles.com/? Dynamic-Or-Static-Web-Pages,-Which-Way-Should-You-Go? &id=373266.

[6] 董 斌. 静态页面生成的网站系统研究[J]. 福建电脑，2009(8):160-161.

[7] ExSite Webware Content Management. Static vs. Dynamic Content [EB/OL]. 2008-12-24. http://support.exsitewebware.com/cgi/page.cgi/articles.html/Content-Management/Static_vs_Dynamic_Content.

[8] 马 强，宋 玲. 基于 Web 的社保新闻发布系统的设计与实现[J]. 计算机技术与发展，2007，17(12):31-33.

[9] 高 翔，何立军，李国兴. JSP 动态网站开发技术与实践[M]. 北京：电子工业出版社，2007:245-293.

[10] 汪孝宜，刘中兵，徐佳晶，等. JSP 数据库开发实例精粹[M]. 北京：电子工业出版社，2005:145-209.

[11] 曾春华，江南雨. 动态生成静态网页技术探索[J]. 计算机与网络，2008(24):511-512.

[12] 许冀伟，李广霞，傅王月. 一种基于 ASP.NET 技术生成新闻静态页的方法[J]. 计算机与网络，2007(33):214-214.

[13] Decoder. JSP 技术揭秘[M]. 北京：清华大学出版社，2001:140-156.

[14] 荣钦科技. JSP 动态网站开发与实例[M]. 第 3 版. 北京：清华大学出版社，2006:279-287.

[15] Liang Y D. Java 编程原理与实践[M]. 第 4 版. 北京：清华大学出版社，2005:665-716.

(上接第 132 页)

识逻辑检测[J]. 农业网络信息，2008(9):138-141.

[3] 吉 喆，李绍稳，张友华，等. 基于本体的茶虫害诊断系统构建的研究[J]. 农业网络信息，2008(9):112-116.

[4] OWL Web Ontology Language Reference[EB/OL]. 2004-10. http://www.w3.org/TR/owl-ref/.

[5] 王晓东，张 合，王红涛. 基于 Ontology 的语义信息检索模型研究[J]. 计算机工程与设计，2008，29(11):2939-2941.

[6] 钱 平，郑业鲁. 农业本体论研究与应用[M]. 北京：中国农业科学技术出版社，2006:89-98.

[7] 鲜国建，孟宪学，常 春. 基于农业本体的智能检索原型系统设计与实现[J]. 中国农学通报，2008，24(6):470-474.

[8] XMLDOM, XPath[EB/OL]. 2008-10. http://msdn.microsoft.com/zh-cn/87274khy(VS.80).aspx.

[9] Googel Ajax Search API[EB/OL]. 2008-04. http://code.google.com/intl/zh-CN/apis/ajaxsearch/.