基于主题本体的信息采集模型研究

拜战胜[1,2],徐德智[1],彭佳红[2],陈光仪[1,2]
(1.中南大学 信息科学与工程学院,湖南 长沙 410083;
2.湖南农业大学 信息科学技术学院,湖南 长沙 410128)

摘 要:互联网上的海量信息,至今还在快速发展,面向主题的信息检索已成为当前的研究热点之一。在提高信息检索的精度方面,一般认为本体技术是解决方法之一。在对领域本体技术和传统的基于主题的信息采集技术的基础上,设计了一个基于领域本体的信息采集模型,给出了模型的体系结构,提出了一种关键词加权的词性相关性计算方法以及利用领域本体及对应的词典判定主题相关度的算法。通过实验验证了所提出的方法在提高检索的准确率方面具有明显的优势。

关键词:主题本体;领域本体;信息采集;主题相关度
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2009)10-0102-04

Research of a Model of Web Information Acquisition Based on Topic_Ontology

BAI Zhan-sheng[1,2], XU De-zhi[1], PENG Jia-hong[2], CHEN Guang-yi[1,2]
(1.College of Information Science and Engineering, Central South University, Changsha 410083, China;
2.College of Information Science and Technology, Hunan Agriculture University, Changsha 410128, China)

Abstract: There are huge amount of Web pages in Internet, and they are still increasing rapidly. The topic-specific Web information retrieval has been one of the hot spot being studied at present. Ontology technology is considered to be one of the solution in improving retrieval accruacy. In this paper, combine domain_ontology technology with the traditional information retrieval technology. A model of Web information acquisition based domain_ontology is designed and the architecture of the model is given. Proposes an approach for calculating the relevance between the Web pages and the predefined topic utilizing domain_ontology and lexicon. At last, some results of experiment indicate that our model has obviously advantage in improving retrieval accuracy.

Key words: topic_ontology; domain_ontology; information acquisition; topic-specific relevance

0 引 言

随着 Internet 的快速发展,Web 已经成为全球最重要的信息源,如何有效地获取和利用这些资源成了一个重要的研究课题。虽然搜索引擎在 Web 检索信息方面为用户提供了很大的帮助,但是通用性搜索引擎无法很好地为用户提供特定主题的信息,另外基于关键字的技术也使得搜索引擎缺乏提供对语义的支持。要解决这一问题,有研究表明[1~3],基于本体的技术是解决方法之一。

本体不仅为规范化描述资源提供了基础,也为更准确地搜索信息提供了保障。目前已经有很多的研究做了将本体论应用到信息检索领域的尝试[4~6]。文献[7~9]主要研究通用本体在信息检索及 NLP 中的应用,但是目前要建立一个能够涵盖所有领域知识的通用本体几乎是不可能的,比较可行的方法是先建立某个领域的本体去解决该领域的特定的信息检索问题。

文中在对领域本体和主题信息采集技术研究的基础上,设计了一个基于计算机软件的领域本体的信息采集模型,给出了模型的体系结构,介绍了相关模块。

收稿日期:2009-02-24;**修回日期**:2009-05-14
基金项目:湖南省教育基金资助项目(07C008)
作者简介:拜战胜(1979-),男,讲师,硕士,研究方向为语义 Web 方面;徐德智,教授,博士后,研究方向为语义网、Web 计算;彭佳红,教授,研究方向为数据挖掘。

1 基于主题本体的信息采集模型

1.1 体系结构

在对本体技术和主题信息采集研究的基础上,设计了一种基于领域本体的信息采集模型其体系结构,如图 1 所示。

1.2 系统组成

本模型主要包括数据采集、爬虫程序、页面解析和

过滤、分词模块、相关性计算、本体管理器和控制界面几部分。

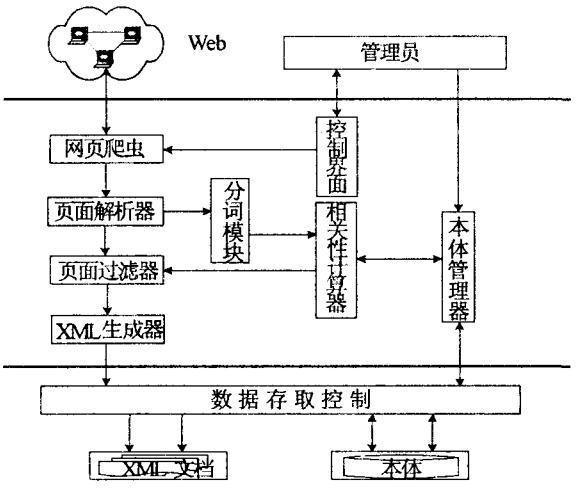

图 1　体系结构

1)控制界面:用户通过控制界面完成相关的设置,包括主题本体的选择、设置初始抓取页面的 URI 和相关参数等。

2)数据采集:系统与外网进行交互的部分,主要根据用户指定的初始 URL,通过 Web 协议来自动采集 Web 站点内有效的信息。

3)页面解析和过滤:主要完成对文档的正文提取、分词、去除停用词以及页面内容分析等操作。其中的内容分析包含了链接的提取和标题的提取等,为相关性计算提供基础。

4)本体和词典:采用相关本体及对应词典表示目标主题。本模块是相关性计算的依据。

5)相关性计算:完成计算采集信息的相关性得分,进而判断处理页面是否同目标内容相关。若相关则保存,否则丢弃。

2　关键技术

文中涉及到的关键技术主要包括 6 个:主题本体、词典、关键词加权、相关性计算、实体权值计算算法、主题相关性判定算法。

2.1　主题本体

主题描述的不准确性常常是造成不良检索结果的重要原因。本体明确地、形式化地描述了某一领域的共享概念,能使用户和计算机更精确地基于语义进行交流。文中的本体采用领域本体的方法构建。所谓领域本体是将领域概念化抽象出来的对象、关系和类等用一个词汇集来表达,该词汇集就是领域本体[5]。当对象、关系和类集中在更小的一个范围之内,比如针对奥运会、计算机网络这样的主题范围时,将这样一些词汇集组成的集合就称为主题本体。主题本体不但能很好地描述主题内容,而且可以揭示概念间的语义关系,提高主题描述的精度,使主题相关度计算以及主题爬行策略计算更准确。

定义 1:主题本体是包含下列元素和操作的集合。

1) C:类。一个类描述了具有某些属性而同属一组的一些个体的集合。

2) Sc:类的层次结构关系。

3) P:属性用来表述个体之间的二元关系。

4) I: 类的实例。具有某些属性而同属一组的个体的集合中的一个个体。

5) M:本体中的概念、属性等到词典的映射。

2.2　词　典

定义主题本体中没有全部表示出来的词汇及同义词,与主题本体一起描述信息。

定义 2:词典是包含下列元素和操作的集合。

1) Lc:词典中的词条。L 中词条详细地描述了它所表示的主题内容。

2) SYN:词条对应的同义词。此操作目的是解决多词同义的问题。

3) M':词典中词条到本体中概念或属性的映射。

2.3　关键词加权技术

利用本体对关键词加权。关键字的权值反映了该关键字表现主题的能力。首先给出该主题范围内的核心词汇集合 Core,然后根据同核心词汇的语义距离来确定其他关键词的权值。有研究表明,词语距离与词语相似度之间有着密切的关系,2 个词汇的距离 Dis 越大,其相似度越低;反之,相似度越大。这里的距离指的是连接两个概念之间的最短的路径的边数。确定了核心实体的权值之后,再根据实体同核心实体的距离来定义非核心实体的权值。

定义 3: $w_{\overline{core}}$ 是非核心实体 $\overline{core}$ 的权值:

$$w_{\overline{core}} = \frac{\alpha}{\alpha + \mathrm{Dis}(core, \overline{core})} \tag{1}$$

其中 α 是一个可调节的参数,且 $0 < \alpha \leqslant 1$。另外因为一个实体同核心实体集合中的实体的距离可能不是唯一的。因此对于非核心实体集合中的实体 x,其权值 W:

$$\forall x, x \in \overline{core}, y \in core, w_x = \frac{\alpha}{\alpha + \mathrm{Min}(\mathrm{Dis}(y, x))} \tag{2}$$

2.4　相关性计算

相关性计算是系统的核心部分,Web 上抓取的数据经过预处理之后,由相关度计算模块来计算内容是否与主题定义的内容相关以及相关度是多少。有研究表明网页上面的内容在表现主题的时候的作用是不同

的。例如标题 title 的内容通常说明了当前页面的主要的内容,页面中的粗体斜体很可能表明了作者想着重指出的内容。为了突出这种特点另外不至于把权值太过分散,页面内容被分成两部分——标题和正文,在进行处理的时候对标题中的内容给予更高的权重。

定义 4:页面与主题之间的相似度 $\mathrm{sim}(D,p)$ 为:

$$\mathrm{sim}(D,p) = \lambda_T \mathrm{sim}(D,T) + \lambda_B \mathrm{sim}(D,B) \quad (3)$$

其中 D 定义了主题知识,在这里等价于本体 O,p 表示一个页面,B、T、λ 分别表示页面的正文、页面的标题以及权值。$\lambda_B + \lambda_T = 1.0$,并且 $0 < \lambda_B, \lambda_T < 1.0$。将上式进一步分解得到:

$$\mathrm{sim}(D,p) = \lambda_T \sum_{t \in T} \mathrm{sim}(D,t) + \lambda_B \sum_{b \in B} \mathrm{sim}(D,b) = \lambda_T \sum_{t \in T} W_t + \lambda_B \sum_{b \in B} W_b \quad (4)$$

标题中实体 t 与领域的相关度直接用 t 在本体中的权值表示。同样的,正文实体 b 与领域的相关度也直接用 b 在本体中的权值表示。进行规范化处理之后的相似度如下:

$$\mathrm{sim}(D,p) = \frac{\lambda_T \sum_{t \in T} W_t + \lambda_B \sum_{b \in B} W_b}{\lambda_T N_T + \lambda_B N_B} \quad (5)$$

其中 N_T 和 N_B 分别表示页面中标题实体和正文中实体的数量。利用公式(5)计算,最理想的情况是页面中每一个实体的权值都是 1.0,即:

$$\mathrm{sim}(D,p) = \frac{\lambda_T N_T + \lambda_B N_B}{\lambda_T N_T + \lambda_B N_B} = 1.0$$

2.5 实体权值计算算法

算法 1:本体中实体权值计算算法。

输入:主题本体;

输出:本体中每个实体权值确定的主题本体。

```
ComputeWeight //本体中的实体权值计算
{
指定核心实体集 Core,为每个实体赋值 1.0 即 W_Core = 1.0;
For(∀e ∈ E) //中序遍历本体中节点。E 满足定义 1
{If(e ∈ Core) Continue;
赋值 W_(e) = 0.0;
If(S_C(e,ec) || S_C(e,ec))
{ 修改 W_(e) = α/(α+1); Continue;}
else
根据公式(2)修改权值;
{W_(e) = α/(α + Min(Dis(y,x)))  x ∈ Core, y ∈ Core
}//end ComputeWeight
```

经过该算法的计算,本体 O 被表示成核心实体和非核心实体组成的集合。

2.6 主题相关性判定算法

算法 2:主题相关性判定算法。

输入:主题本体、预处理之后的 Web 页面、调节参数 α(经验值通常取 0.7~0.9);

输出:页面是否相关的结论。

```
JudgeRelevance//判断页面是否同主题相关
{
初始化页面权值 W、标题权值 W_t 和正文权值 W_b;
while(页面 p 中还存在未处理的实体) //实体已经经过预处理
{if (标题 t 存在)
{ for (标题 t 中的每一个实体 y)
if (y 存在于本体中) 修改标题权
W_t += W_y; //W_y 是实体 y 的权值
}
if (正文存在)
{ for (正文 b 中每一个实体 z)
if (z 存在于本体中) 修改正文权值 W_b += W_z; //W_z 是实体 z 的权值
}
{ 根据(2-5)计算页面主题相关度 W;
If (W >= W_td) 保存页面;//W_td 表示页面是否相关的阈值
else 丢弃页面;
计算前 N 个页面的权值的平均值
W̄ = (∑W_x)/N, 调整 W_th : = W̄
//N 表示已经处理过的页面,阈值反馈是对阈值的一个非人工干预的自动调整
} //end JudgeRelevance
```

3 实 验

3.1 实验设置

实验参考了中图分类法关于计算机软件的分类标准,并查阅了朱三元等编辑的汉英计算机综合词典(修订本)和郑茂松、查良锪编辑的新英汉计算机软件词汇[9,10],使用 Protégé3.1 初步建立了关于计算机软件的主题本体,如图 2 所示。

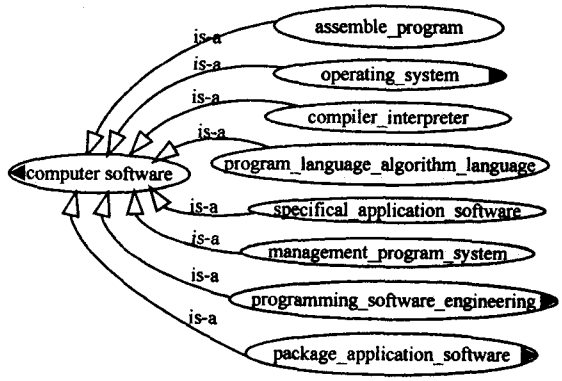

图 2 计算机软件主题本体示意图

经过讨论确定了包含计算机软件、软件、软件包、软件工程和程序设计等一系列的有关计算机软件主题的核心实体。根据定义 3 确定其他实体的权值。实验

中标题的权值 λ_T 和正文的权值 λ_B 分别设置为 0.6 和 0.4。调节参数 α 经过实验最终取值为 0.9。实验选择了人民日报 2004 年 12 月的 4120 篇图文电子版文档作为测试集合,对其进行了主题相关度的计算和判断。表 1 给出两个计算页面相关度的实例。

表 1　页面主题相关度计算实例

相关词汇	文档 1	文档 2	…
软件	1.0 (1+59)	1.0 (1+23)	
微软	1.0 (0+12)	1.0 (0+3)	
操作系统	0.4737	ϕ	
……			
综合权值	0.185	0.140	

表 1 中文档 1 代表《国产软件:在"温室"中长大?》这篇文章,文档 2 则代表了《网友为国产软件发展建言》这篇文章。表格中列出了考察的文档中出现的主题词汇、它们对应的权值以及在标题和正文中出现的次数。ϕ 表示词汇没有出现。

3.2　衡量指标

算法评价标准采用 Web 信息检索常用的评价指标,即采用准确率 P(precision)、召回率 R(recall) 和 F 值:

$$\text{准确率}(P) = \frac{\text{实际收集到的属于该领域的文档数目}(D)}{\text{收集到的所有文档}(A)}$$

$$\text{召回率}(R) = \frac{\text{实际收集到的属于该领域的文档数目}(D)}{\text{测试集中实际属于该领域的文档数目}(T)}$$

$$F = 2 * \frac{(R * P)}{R + P}$$

3.3　结果及分析

文中基于主题本体的采集方法同页面提供的基于关键字匹配的方法进行了对比,结果如图 3 所示。

图中方法 1 表示:页面提供的基于关键词匹配的方法,使用计算机软件作为搜索词汇;方法 2 表示:页面提供的基于关键词匹配的方法,使用软件作为搜索词汇;方法 3 表示:文中提出的方法。图 3(a)~(d)分别表示了三种方法的实际收集到的属于该领域的文档数目、召回率、准确率以及 F 值的对比结果。横坐标表示方法,纵坐标表示相应的测试指标。每一种方法上面的数字表示了该方法在对应的指标中的表现。从图 3 的对比可以看出:方法 1 在各个方面的表现都比较差,主要原因是计算机软件这个关键词虽然也能表达需求信息,但是使用率不高。同样是基于关键词的方法,使用软件这个词汇表达需求信息的时候效果就比方法 1 好的多。因此,基于关键字的方法的优劣很大程度上取决于表达需求的关键词。

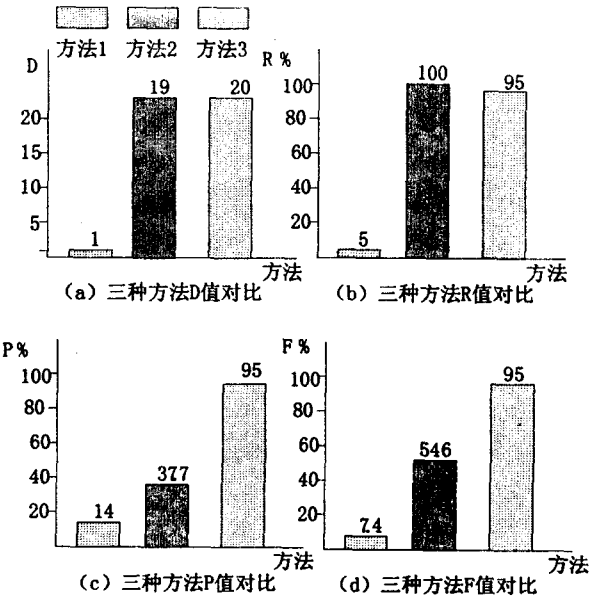

图 3　三种方法实验结果对比

同时,从方法 2 和方法 3 的对比可以看出,在 D 和召回率 R 上两种方法差别不太多,说明文中的方法在提高召回率方面没有显著的优于传统方法。但是对比准确率和 F 值文中的方法的效果就明显优于另外的方法。总的来说所提出的方法是有效的。

4　结束语

用文中所提出的技术来进行主题分析是所进行的一个尝试,建立在本体论基础上的语义分析可以初步实现人机语义交互,有效提高主题发现过程中的准确率。目前本体中的概念一般都是通过人工提取的并且面向一个特定领域,远没有成为一种工程性的活动,这使得基于本体的应用不能大规模开展,因此,在以上工作的基础上,下一步的工作重点是进一步完善本体以及如何设置参数以获得更好的主题判断的准确率。

参考文献:

[1] Tijerino Y A, Sanati R. Onto TEMAS: an ontology based teaching materials search engine[J]. Journal of Computing Sciences in Colleges, 2005, 20(4): 177 - 182.

[2] 宋峻峰,张维明,肖卫东,等.基于本体的信息检索模型研究[J].南京大学学报:自然科学版,2005,41(2):189 - 197.

[3] 陈　康,武港山.基于 Ontology 的信息检索技术研究[J].中文信息学报,2005,19(2):51 - 57.

[4] 王　进.基于本体的语义信息检索研究[D].合肥:中国科学技术大学,2006.

[5] 贾自艳.Web 信息智能获取若干关键问题研究[D].北京:中国科学院,2004.

(下转第 109 页)

```
……
<! -定义 MailSender Bean,用于发送邮件 ->
<bean id = "mailSender" class = "org. springframework. mail.
javamail. JavaMailSenderImpl">
……
<! -定义 MailMessage Bean ->
<bean id = "mailMessage" class = "org. springframework.
mail. SimpleMailMessage">
……
<! -配置业务逻辑 Bean ->
<bean id = "businessManager" class = . . .
. . . . . .
</beans>
```

2.5 实现 Web 层(Struts 框架)

使用 Struts 作为前端 MVC 框架,要求让 Struts 的 ActionServlet 拦截用户请求。本系统进行的非常彻底,所有的客户端请求都采用 * . do 的模式。所有的请求都被 ActionServlet 拦截,系统的 JSP 页面放在 WEB - INF/jsp 路径下,直接输入 JSP 页面地址无法访问该页面,从而提供了更好的安全性。

配置 Struts 加载 Spring 的 ApplicationContext,应在 struts - config. xml 文件中增加如下配置:

```
<plug - in className = "org. springframework. web. struts. Con-
textLoaderPlugIn">
    <set - property property = "contextConfigLocation"
        Value = "/WEB - INF/applicationContext. xml,
                 /WEB - INF/daoContext. xml,
                 /WEB - INF/action - servlet. xml"/>
</plug - in>
```

本系统使用 DelegatingActionProxy 的整合策略,在这种策略下,系统中所有的 Action 不再配置真正的 Action 实现类,而是使用 Spring 提供的 DelegatingActionProxy 类,该类将请求转发给 Spring 容器内同名的控制器 Bean。因此,struts - config. xml 内的 Action 配置都是如下格式:

```
<action path = "/processAddKind" type = "org. springframework.
web. struts. DelegatingActionProxy"
    name = "kindForm"
    scope = "request"
    validate = "true"
    ……
    <forword name = "login" path = "/WEB - INF/jsp/login.
jsp"/>
</action>
```

3 结束语

文中分析了传统重量级 J2EE 架构的缺点,提出了整合 Struts + Spring + Hibernate 的 J2EE 轻量级框架,并应用于电子拍卖系统中。实际的应用表明:该系统具有良好的交互性、可扩展性和可维护性,表现出良好的性能。必将成为开发 Web 应用的一个有力的解决方案。

参考文献:

[1] 陶以政,吴志杰,唐定勇,等 . 基于 J2EE 的应用框架技术研究[J]. 计算机工程与设计,2007,28(4):826 - 828.

[2] Johnson R. Expert One - on - One J2EE Design and Development[M]. [s. l.]:Wrox,2002.

[3] Struts Reference Documentation. Introduction to the Struts Framework [EB/OL]. 2008. http://struts. apache. org/primer. html.

[4] 闰智敏,王　力,杜军朝,等. 邮政 11185 业务系统持久层的 Hiebmate 解决方案[J]. 计算机技术与发展,2008,18(4):178 - 181.

[5] Johnson R. Introduction to the Spring Framework[EB/OL]. 2005 - 05. http://www. theserverside. com/tt/articles/article. tss? l = SpringFramework.

[6] 胡启敏,薛锦云,钟林辉. 基于 Spring 框架的轻量级 J2EE 架构与应用[J]. 计算机工程与应用,2008,44(5):115 - 118.

[7] 李腊元,徐　鹏. 基于 MVC 模式的 JSF,Spring 和 Hibernate 整合[J]. 计算机技术与发展,2008,18(3):46 - 49.

[8] 杨蕴石,王　颖,吕　科,等. 基于 J2EE 的集成开源框架研究与应用[J]. 微计算机信息,2008,24(5 - 3):220 - 223.

[9] L1 Kang - rong, MIAO Fang. Study on E - commerce system architecture based on MVC model and J2EE platform[J]. Journal of Communication and Computer, 2008, 5(2): 46 - 50.

(上接第 105 页)

[6] 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,2001,38(6):641 - 647.

[7] 潘宇斌,陈跃新. 基于 Ontology 的自然语言理解[J]. 计算机技术与自动化,2003,22(4):71 - 74.

[8] 廖乐键,曹元大,李新颖. 基于 Ontology 的信息抽取[J]. 计算机工程与应用,2002,38(4):110 - 113.

[9] 朱三元. 汉英计算机综合词典(修订本)[M]. 上海:上海科学技术文献出版社,1998.

[10] 郑茂松,查良钿. 新英汉计算机软件词汇[M]. 北京:电子工业出版社,1999.