文章编号:1673 - 0062(2013)04 - 0077 - 05

基于一种文档表示模型的站内搜索引擎设计与实现

蒋 辉,阳小华,刘志明,闫仕宇,马家宇,李晓昀,李 萌,周 座

(南华大学 计算机科学与技术学院,湖南 衡阳 421001)

摘 要:根据全信息理论,认识论信息是语法信息、语义信息和语用信息的三位一体,在信息检索的过程中加入语用信息能有效的提高信息检索的质量. 基于查询与内容的文档表示模型较好的利用了语用信息,对站内搜索引擎的查准率的提高有着很好作用;Lucene 是一个用 java 语言开发的开源的全文搜索引擎架构. 本文利用 Lucene 设计和实现一个基于查询与内容的文档表示模型的站内搜索引擎,实验结果表明该模型能有效的提高信息检索的查准率.

关键词:lucene;站内搜索引擎;搜索引擎;信息检索

中图分类号:TP391 **文献标识码**:B

Website Search Engine Design and Implementation Based on a Document Representation Model

JIANG Hui, YANG Xiao-hua, LIU Zhi-ming, YAN Shi-yu, MA Jia-yu, LI Xiao-yun, LI Meng, ZHOU Zuo

(School of Computer Science and Technology, University of South China, Hengyang, Hunan 421001, China)

Abstract: According to the comprehensive information theory, epistemology information is the trinity of syntactic information, semantic information and pragmatic information. Making better use of pragmatic information in information retrieval can promote the quality of information retrieval. A document representation model based on query and content can make better use of pragmatic information, and it is good to promote the precision of the website search engine. Lucene is a open source full text search engine architecture which is developed using java language. We use lucene to design and implement a website engine based on document representation model using query and content. The experiment results show that this model can effectively improve precision rate in information retrieval.

收稿日期:2013 - 06 - 30

基金项目:湖南省自然科学基金资助项目(11JJ6047);衡阳市科技计划基金资助项目(2011KJ14;2013KG67);湖南省科技计划基金资助项目(2011FJ3087);南华大学计算机科学与技术校级重点学科基金资助项目

作者简介:蒋 辉(1981 -),男,湖南衡阳人,南华大学计算机科学与技术学院讲师,硕士. 主要研究方向:信息检索与知识科学、软件工程.

key words: lucene; website search engine; search engine; information retrieval

0 引 言

随着互联网快速地发展与广泛地普及,互联网上的信息量也在迅猛的增加,这使得大家想方便和快捷地在互联网上查询所需信息成为一种奢望. 搜索引擎技术的出现有效的缓解了这一问题,并越来越成为人们关注的焦点. 随着信息化在我国的逐步推进,许多企事业单位、政府都有了自己的信息化系统,通过一段时间的运行都有了大批量的数据,人们也可以使用站内搜索引擎快捷方便地从中获取所需要的信息. Lucene 是一个用 java 语言开发的开源的全文搜索引擎架构,可以利用其来搭建全文(站内)搜索引擎. 查准率是衡量搜索引擎性能指标之一,查询与内容的文档表示模型[1]较好的利用了语用[2]信息,对提高站内搜索引擎的查准率有很好的作用,本文将利用这个模型和 Lucene 设计和实现一个站内搜索引擎.

1 Lucene 简介

Lucene 是 Apache 软件基金会 Jakarta 项目组的一个子项目,是一个开放源代码的全文检索引擎工具包,即它不是一个完整的全文检索引擎,而是一个用 Java 语言实现的全文检索引擎架构. 它为数据访问和管理提供了简单的函数调用接口,可以方便地嵌入到各种应用程序中实现全文检索功能,提供了强大的全文索引和检索的源动力. 此外,它还提供了完整的查询引擎和索引引擎及部分文本分析引擎,还具有方便的用户接口、面向 WWW 的开发接口、二次应用开发接口等等.

Lucene 以其开放源代码的特性、优异的索引结构、良好的系统架构获得了越来越多的应用. Lucene 作为一个全文检索引擎,其具有优秀的面向对象系统架构;索引文件格式独立于平台;实现分块索引,提升索引速度等突出的优点[3].

Lucene 系统由基础结构封装、索引核心、对外接口三大部分组成. 其中操作索引文件的索引核心是系统的重点. Lucene 将所有源码分为七个模块,其系统的系统结构图如图 1.

Lucene 应用了最基本的一条程序设计准则:引入额外的抽象层以降低耦合性. 在每一个局部细节上,比如某些常用的数据结构与算法上,Lucene 也充分的应用了这一条准则. 在高度的面向对象理论的支撑下,使得 Lucene 的实现容易理解且易于扩展.

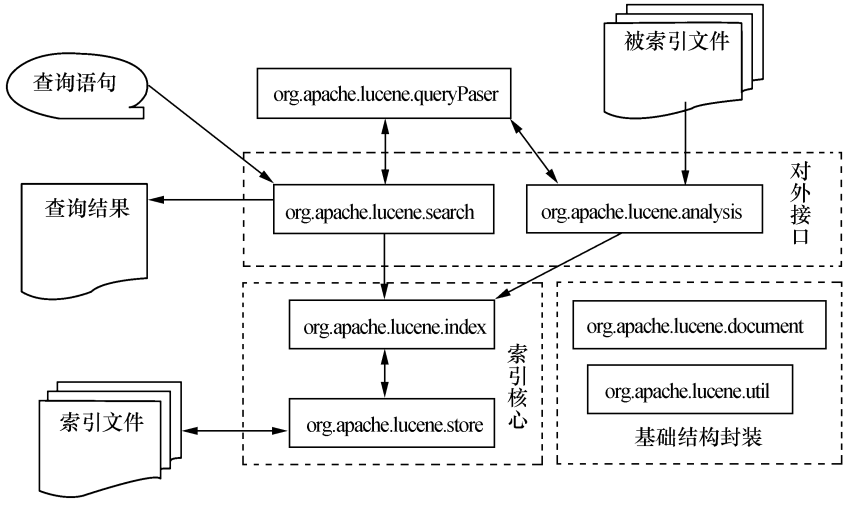

图 1 系统结构图

Fig. 1 System structure diagram

2 基于查询与内容的文档表示模型

基于查询与内容的文档表示模型[1]认为当有较多的用户用相同的关键词搜索,并点击相同的搜索结果时,其查询词和被点击的结果文档间有一定的关系,可以用来表示文档. 这部分语用信息对提

高搜索引擎的查准率是很有帮助的,因此可以收集这部分语用信息,改进文档表示模型,达到提高搜索引擎的查准率的目的. 基本思路是:在信息检索系统初始化时,构建文档的向量空间. 随着搜索引擎的运行,用户查询日志将越积越多,这些日志表明了查询词与被点击文档之间的一种关系. 当其所形成的查询集样本空间足够大时就能够提供有价值的用户隐性反馈信息,与文档关联的高频查询词将在一定程度上反映出用户的意图. 即这些高频查询词较好地描述了对应文档的主题信息. 此时,就可以在文档表示中逐步引入查询集的信息,使得文档的特征空间成为查询空间与文档空间信息的整合,从而提高描述文档主题信息的带权特征词的适应度与可信度,达到改善检索性能的目的.

3　基于查询与内容的文档表示模型和 Lucene 的站内搜索引擎的设计与实现

3.1　系统设计

目前大多数的企事业单位的网站是动态网站,一般的网络爬虫很难抓取其网站内部的信息. 而在开发站内搜索引擎的时候,相关网站的数据库可以获取得到,因此不使用网络爬虫来抓取网页,而选择直接将网站数据库直接转换为固定格式的 xml 文档,这些 xml 文档容易被 lucene 建索引.

Lucene 是通过对查询词与被索引文档之间的评分来排序被索引文档,并输出查询结果的. 因此想改变查询结果的排序只需修改其评分函数,修改评分函数的依据是基于查询与内容的文档表示模型中提出的模型修正公式. Lucene 中虽然带有 CJK 分词工具,但是其分词效果不太理想,所以我们将分词效果较好的极易分词工具替换其自带的 CJK 分词工具. 另外,根据这个模型还需设计一个用户查询行为记录和分析模块,用来记录和分析用户查询行为,根据基于查询与内容的文档表示模型所述,当用户的查询行为所形成的样本空间大于一个阈值时将改变文档的表示,根据 Lucene 的本身特点,将通过修改评分与排序来实现. 系统的整体结构图如图 2 所示.

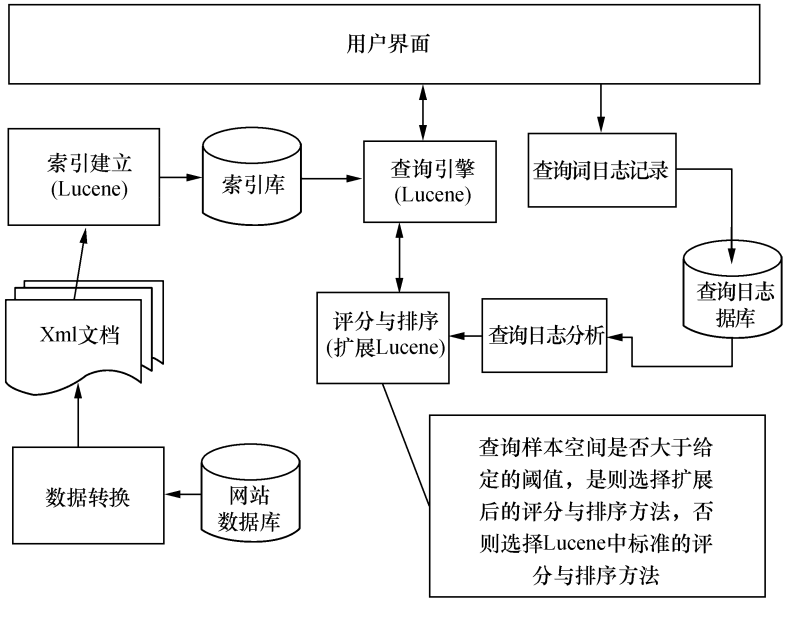

图 2　系统整体结构图

Fig. 2　Overview diagram of system

综上所述,需要开发查询行为日志记录与分析模块、数据转换模块和用户界面模块,需要修改是 lucene 的评分函数以及替换其自身所带的 CJK 分词工具.

3.1.1　数据转换与索引

1)将数据库中的数据转换为 xml 文档

通过 JDBC 链接指定的数据库,利用 W3C 规定的标准 DOM[4](文档对象模型)进行 XML 的读

写工作,因为本系统使用 XML 文档的主要目的是提供数据存储,所以 XML 结构应避免过于复杂. DOM 的基本对象有 5 个: Document, Node, NodeList, Element 和 Attr, 其中 Document 对象代表了整个 XML 的文档,所有其它的 Node 都以一定的顺序包含在 Document 对象之内且排列成一个树形结构,通过遍历这颗树来就可以得到 XML 文档的所有内容.

为了方便 lucene 建索引,我们将 xml 文件的格式定义为如下格式(xml 中 1 个节点):

< Node >

< URL > …… …… </URL > //Web 文档的 url

< Title > …… </Title > //标题

< SubTitle > …… </SubTitle > //子标题

< Author > …… </Author > //作者

< Keywords > …… </ Keywords > //关键字

< Content > …… </ Content > //正文内容

</Node >

2)给导出的 XML 文件建立索引

由于中文在词与词之间没有分界,以及 Lucene 自带的分词工具的局限,在建立索引的过程中我们需要扩展 Lucene 的语言分析包(org. apache. lucene. analysis)实现对中文的处理. 通过对多种中文分词工具(如 StandardAnalyzer、ChineseAnalyzer、CJKAnalyzer、IK _ CAnalyzer、MIK _ CAnalyzer、MMAnalyzer)综合性能的比较,我们选择 JE 分词 1.5 版本作为本搜索引擎中文分词工具.

对 XML 文件建立索引分为两个步骤:利用 SAX 将 XML(Simple API for XML)文档中的每条记录转化为 org. apache. lucene. document. Document 对象和利用 org. apache. lucene. index. IndexWriter 对象循环地对第一步转化得到的每个 Document 对象建立索引.

3.1.2 用户查询行为的获取

该部分主要分为两个步骤:获取用户行为数据和分析并保存所获取的数据.

1)获取用户行为数据

当用户浏览搜索引擎返回结果的标题和摘要时,遇到自己关心的部分一定会点击其链接进入相应的页面. 我们可以设定一个触发事件捕获其点击行为,获取相关信息,比如查询串和点击链接对应的页面 ID 号.

2)分析并保存所获取的数据

接收上一步骤获取的数据(查询串和点击链接对应的文档 ID 号),利用 InsertArgs 类将其插入到数据库 SearchInf(有表 SearchWords 和 Analysis)中. InsertArgs 类继承于 DB 数据操作类,他可以自由调用父类的相关数据操作方法,其类图如图 3.

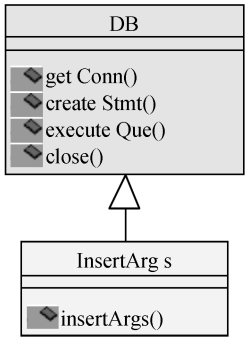

图 3 InsertArgs 类图

Fig. 3 Class diagram of InsertArgs

获取用户行为数据及分析并保存所获取的数据的过程可以用下面的流程图如图 4.

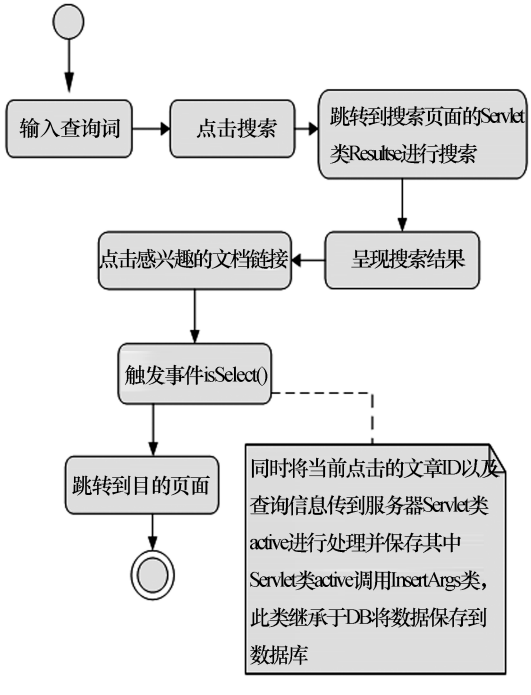

图 4 获取并保存用户行为的数据

Fig. 4 The flow chart of get and save user's behavior

3.1.3 评分函数的修改

用户提交查询串后,通过 lucene 的内部评分机制得到一个文档队列,该队列保存了获分较高的文档及其评分值. 将文档队列中的文档逐个取

出,以文档 ID 号为依据,用查询串分词后的各个词项在表 Analysis 中进行匹配,计算查询空间中每个文档对应的查询词项的频率. 在表 SearchWords 中扫描包含查询词项的查询串个数及查询串的总个数. 计算出文档队列中各个文档在查询空间中的特征词权重,进而计算出每个文档的评分. 将上一步得到的文档评分与 lucene 的内部评分依据预定义规则相加,得到文档的最后评分,然后排序返回给用户.

在 Lucene 架构中,这个保存文档评分的队列在 Lucene 的 TopDocCollector 中,我们的主要目的就是修改这个类,我们的方法是在该类里边增加一个修改评分的模块. 修改评分依据为文献[1]中的公式 4,其中参数 α 在系统中可以配置. 再次调用检索器时,IndexSearcher 就会调用这个类(同时包括修改评分的模块)进行评分.

3.2　实现

采用 Lucene2.4.1 工具包[5],并在其基础上进行扩充与修改,利用 MyEclipse 为开发平台,使用 JSP 开发用户界面,tomcat 为应用服务器. 利用 W3C 规定的标准 DOM(文档对象模型)将数据库中的数据导出成 XML 文档,然后利用 SAX 将 XML 文档中的每条新闻转化为 Lucene 中的文档对象,进而利用 Lucene2.4.1 工具包中的 IndexWriter 对象对各个文档对象建立索引,然后应用 IndexSearcher 对象进行检索. 在建立索引前,先要进行中文分词,本实验采用 JE1.5(极易)中文分词工具对文档对象进行分词.

4　总　结

我们在 MyEclipse 开发平台下用 Java 语言和 Lucene2.4.1 工具包开发的一个基于查询和内容的文档表示模型的站内搜索引擎,然后将其链入校园网供全校师生使用,并收集用户查询日志;重点分析其中的用户查询日志,并从中挖掘有用信息重构校园网文档的表示模型,从而达到提高检索效率的目的. 我们利用收集到的日志进行实验,并将其与传统的向量空间模型的查准率进行对比,得到当参数 $\alpha \in [0.5, 0.9]$ 时系统将取得较好的查准率. 实验结果如表 1.

表 1　两种文档表示模型的查准率比较

Table 1　The precision comparison table of two kinds of document representation model

$\alpha \in [0.5, 0.9]$	General VSM	Query-Content VSM
top-10	0.762	0.892
top-20	0.646	0.794

其中 top-10 表示返回结果的前 10 个,top-20 表示返回结果的前 20 个. 上述实验结果表明,当参数 $\alpha \in [0.5, 0.9]$ 时,基于查询和内容的文档表示模型可以较好地提高站内搜索引擎的查准率.

参考文献:

[1] 阳小华,周座. 基于查询与内容的文档表示模型[J]. 南华大学学报(自然科学版),2010,24(1):39-42.

[2] 钟义信. 自然语言理解的全信息方法论[J]. 北京邮电大学学报,2004,27(4):1-12.

[3] Cutting D. The Lucene Search Engine, Powerful Flexible and Free[J]. Java World,2000(9):15-19.

[4] W3C. DOM Standards[EB/OL]. Internet:csail of MIT, ERCIM, Keio university,2012[2013]. http://www.w3.org/TR/#tr_DOM.

[5] Otis Gospodnetic, Erik Hatcher. Lucene in action[M]. 谭鸿,译. 北京:电子工业出版社,2007.