

# 基于 Lucene 的个性化搜索引擎研究与实现

丁兆贵, 金 敏

(湖南大学 软件学院, 湖南 长沙 410082)

**摘 要:**越来越多的用户在使用搜索引擎时希望能提供快速有效的个性化的查询结果。根据搜索引擎的工作原理,在研究分析开源的搜索引擎工具 Lucene 的系统架构、模型和索引器的基础上,设计了武警部队网站个性化搜索引擎。通过二阶段数据处理流程实现信息的增量采集,通过采用逆向词典结构实现自动分词以及利用双向分词器进行倒排索引的功能,最后利用 Tomcat 服务器进行了部署实现。文中所设计的个性化搜索引擎提高了原 Lucene 搜索引擎的速度和准确率。

**关键词:**搜索引擎;个性化;中文分词;检索器

**中图分类号:**TP391.3

**文献标识码:**A

**文章编号:**1673-629X(2011)02-0105-04

## Research and Implementation of Personal Search Engine Based on Lucene

DING Zhao-gui, JIN Min

(Software School of Hunan University, Changsha 410082, China)

**Abstract:** More and more users want the search engine to provide personal search result fast and efficiently, but also hope the query result more personally. According to the working principle of search engine, design a personal search engine of armed policemen website, based on research and analysis of the system structure, model and indexer of Lucene which is an open source search engine toolkit. Provide the functions of incremental collecting information by two phases data process, automatically dividing words by using inverse dictionary structure, and creating inverse index by using bidirectional words divider. Finally, deploy this system on Tomcat server. The designed personal search engine improves the speed and veracity of the original Lucene search engine.

**Key words:** search engine; personal; Chinese dividing words; searcher

## 0 引 言

个性化搜索引擎是搜索引擎个性化服务的一种体现<sup>[1]</sup>,越来越多的用户使用搜索引擎查找信息时,不仅要求能快速有效地获得查询结果,而且要求结果能充分体现自己的个性化信息需求。个性化搜索通过对用户的行为进行分析和挖掘,根据用户的特点对信息进行重排、整理,过滤无关或相关度低的信息,达到体现个性化的目的。Lucene<sup>[2-4]</sup>使用 Java 编程语言进行开发,是一个实现全文检索引擎工具功能的开放源代码项目,可以非常方便地嵌入到各种各样的应用中实现针对特定应用的全文索引与检索功能。Lucene 自发布以来,在开放源代码社区引发了巨大反响<sup>[5]</sup>。文中在研究分析 Lucene 的系统架构、引擎模型及其索引器的基础上,设计并实现了一个基于 Lucene 的个性化搜索

引擎—武警部队网站个性化搜索引擎。

## 1 个性化搜索引擎系统架构

武警部队网站个性化搜索引擎的系统架构如图 1 所示。个性化服务组件包括一个对信息进行分类的组件(documentAnalysis),一个对用户模型进行判断的组件(userFeature),一个根据用户模型对信息进行过滤和重排的组件(seqTactic)以及一个对用户行为进行跟踪的组件(userActionAnalysis)。这三个组件中所涉及到的信息分类策略、用户模型判断策略、用户模型种类、信息过滤和重排策略都不是固定不变的,可以由管理员根据实际情况通过配置文件进行配置,以提高适应性和可扩展性。Index 和 store 是索引核心组件,是整个搜索引擎系统的核心部分。analysis 和 search 是对外接口,analysis 是文本分析器,负责对被索引文件进行分析,search 是检索器,对外提供检索服务。

整体上看本系统是一个由基础类库、索引核心层和对外接口构成的三层结构。系统最基本的一个设计准则就是引入额外的抽象层以降低模块之间的耦合

收稿日期:2010-06-11;修回日期:2010-09-29

基金项目:中国移动合作项目(20109143123)

作者简介:丁兆贵(1979-),男,山东平邑人,硕士生,研究方向为电子政务;金 敏,博士,副教授,研究方向为嵌入式系统及应用、软件工程与软件项目管理、分布式控制系统等。

性。在图 1 中,包 org.apache.lucene.store 包含了对索引文件的操作封装,在包 org.apache.lucene.index 中完成对索引核心的抽象,在包 org.apache.lucene.search 与包 org.apache.lucene.analysis 中以索引核心为基础实现对外查询和索引接口。

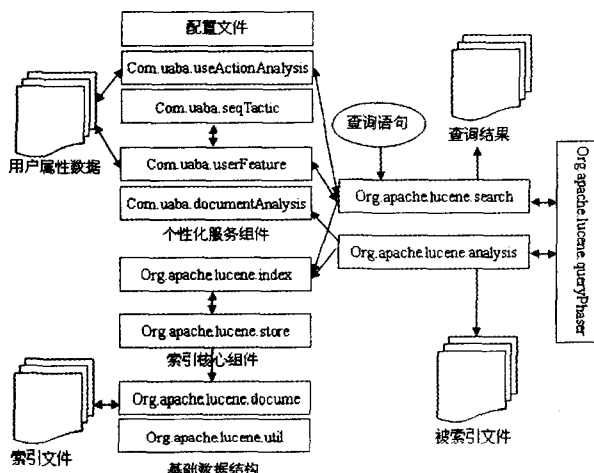


图 1 搜索引擎系统架构

在系统架构上文中引入了传统的 C/S 架构以外的应用结构:可将 Lucene 仅作为一个运行库包含到应用本身,而不需将其当作一个单独的索引服务器。这样一个架构特色体现了 Lucene 编写的本意,即提供一个全文搜索引擎的架构指南,而不仅仅是实现。

## 2 基于 Lucene 的个性化搜索引擎模型

Lucene 为全文检索引擎提供了参考架构<sup>[5]</sup>。本武警部队网站系统以此架构做为开发基础,新增了站内信息的增量采集与文件类型的转换功能,同时对索引模块的中文分词功能进行了扩展。

### 2.1 系统模型

本系统基于 Lucene 开源项目,采用 B/S (Browser/Server) 架构。整个系统分成三个子系统:站内数据采集器、Lucene 全文索引器和检索器。系统的工作模型如图 2 所示。

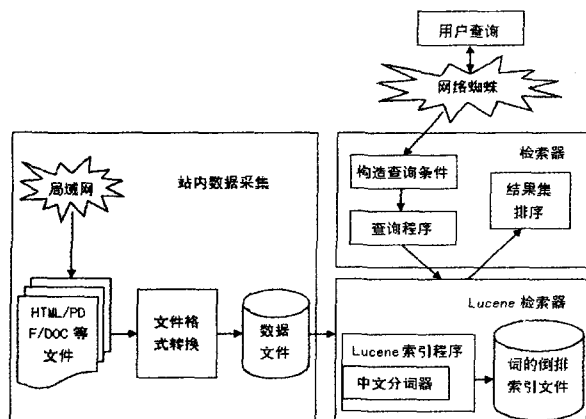


图 2 个性化搜索系统工作模型

其中站内数据采集器将数据进行转换和初步整理,并提交给 Lucene 索引器进行分词和索引,然后形成结构化的数据提供给检索器<sup>[6]</sup>。

### 2.2 系统工作原理

本武警部队网站系统的个性化搜索引擎将工作流程分为数据准备和数据查询两个阶段。

第一阶段:数据准备阶段。站内数据采集器<sup>[6,7]</sup>首先从局域网中武警部队网的主页(或局域网网站内任一指定页面)开始,对网站的全部信息进行采集,收集到的信息经过初步地处理后,存储到 XML 格式的数据文件中;然后对 Lucene 中文分词器提供的分词结果加以扩展,进行倒排索引<sup>[8]</sup>,索引的结果作为数据输入源提供给检索器进行处理。

个性化搜索系统在首次运行时对所有数据均进行处理,处理的数据量大,所需时间长,但在以后的运行中可对新增的数据进行增量式处理,从而使数据处理时间大为减少<sup>[9]</sup>。

第二阶段:数据查询。当数据准备就绪后,用户可使用通用的浏览器进行检索,检索器接收到用户的请求后,首先将对请求的查询内容进行分词处理,构造出规范化的查询表达式,再将该表达式提交给查询程序进行正式地检索处理,最后将检索到的结果集排序后返回给用户。

## 3 基于 Lucene 个性化搜索引擎设计与实现

### 3.1 基于 Lucene 的系统索引器设计与实现

索引的目的是为了提高检索速度。Lucene 本身提供了对文本串以 Token 为单位建立倒排索引文件的功能<sup>[10]</sup>。

采用 Lucene 开源框架构建索引文件需要完成实例化 Document 文档对象和创建索引文件库两个主要的工作。下述的部分关键代码以 Lucene2.0 为基础进行改进。

#### (1) 实例化 Document 文档对象。

首先通过对数据文件进行实例化建立解析对象,以不分词的方式建立 URL、Modified 和 Title 等对象域,为索引提供了前期的准备基础,以下为相关的部分关键代码:

```
//实例化数据文件,建立解析对象
DataFileParsedfparser=new DataFileParse(sourcefilename);
//将 Document 文档对象进行实例化
Document document=new Document();
//在文档解析对象中建立 URL 域,下述各新增域的过程均只在索引文件中存储,而不分词
document.add(new Field("URL",dfparser.getURL(),Field.Store.YES,Field.Index.UN_TOKENIZED));
//在文档解析对象中建立 Modified 域
```

```
document.add(new Field("Modified", dfpaser.getLastModified(),
Field.Store.YES, Field.Index.UN_TOKENIZED));
//在文档解析对象中建立 Title 域
document.add(new Field("Title", dfpaser.getTitle(), Field.Store.
YES, Field.Index.UN_TOKENIZED));
//在文档解析对象中建立 text 域
document.add(new Field("Text", dfpaser.getText(), Field.Store.
YES, Field.Index.UN_TOKENIZED));
//在文档解析对象中建立 FullText 域
document.add(new Field("FullText", dfpaser.getTitle()+dfpaser.
getText(), Field.Store.NO, Field.Index.TOKENIZED));
```

### (2) 创建索引文件。

在上述数据文件对象的基础上建立相关索引文件,并写入索引库中,其关键代码如下:

```
//实例化索引文件库的路径
static final File INDEX_DIR=new File("index");
//实例化双向最小分词器
IndexWriter indexWriter=new IndexWriter(INDEX_DIR,new Re-
verseMinAnalyzer(),true);
//将文档加入到索引库
IndexWriter.addDocument(document);
```

## 3.2 中文分词器

### (1) 中文分词分析。

中文分词方法作为中文信息处理的基础研究课题<sup>[11]</sup>,当前主要分为经验主义和理性主义两大类。经验主义的方法建立在对已用词语统计分析的基础上,再依此推导出当前最可能的分词结果。主要包括基于概率的分词和机械分词;理性主义则通过对中文的语法和语义进行归纳分析,即按照语言的规则进行分词。其经典的分词方法为基于语义的分词法<sup>[12]</sup>。

本系统对中文分词进行改进,从而提高检索的速度和准确率。系统在设计分词算法时主要考虑了以下一些指标:分词速度、分词的粒度和准确率以及编程接口。基于速度考虑,本系统采用了速度最快的机械分词;同时采用最小分词策略来降低分词粒度,从而提高最终查询的准确率,此外还选用了逆向分词策略来进一步增强查询的准确率;在对以上分析进行综合后,本系统使用了逆向最小分词算法进行分词。

### (2) ReverseMinAnalyzer 分词算法。

由于系统是在 Lucene 的基础上进行开发,因而所设计的中文分词器接口必须与 Lucene 在接口上保持一致。文中采用了继承的方式对 Lucene 的 Analyzer 类进行复用,从而保证了接口的一致性,并且在自定义的子类中对 tokenStream 方法进行了重载;同时设计了两个辅助类:一个是继承自 Lucene 中 Tokenizer 类的 ReverseMinTokenizer,其通过设计逆向词典结构来完成逆向最小分词算法;另一个是继承自 Lucene 中 TokenFilter 类的 ReverseMinFilter,它通过使用 XML 格式的

停用词来实现对分词结果中停用词的过滤。

下面是所采用的逆向最小分词算法部分关键代码:

```
//词典采用了逆向词典结构,共收录了6万词汇
ReverseDictionary reverseDic=new ReverseDictionary();
//对分词数据进行格式化
List sentence = StructText.section toSentence(inputString);
//对所有语句进行向最小分词
for(Iterator iter = sentence.iterator(); iter.hasNext();){
//采用逆向最小分词算法进行划分;
List words = reverseDic.rMinm((String)iter.next());
//使用停用词表来过滤掉停用词
List resultWords = TokenFilter.filter(words);
}
```

## 3.3 检索器

Lucene 在建索引时对文档格式没有要求,只要能从中提取出文本信息即可。以静态 HTML 页面文件为例,其建立索引的过程为:(1)通过一个 HTML 解析器对 HTML 文档进行解析,过滤掉其中的 HTML 标签,提取出重要信息,如标题等;(2)将这些信息传送给 Lucene 分析器,将其转换成单独的索引项并存储到 XML 格式的索引文件中。其过程如图3所示。

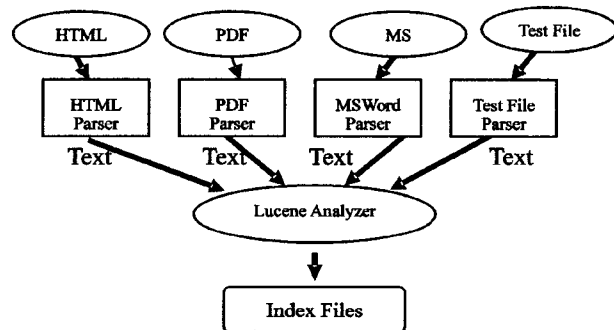


图3 个性化搜索引擎的索引过程

检索器实现对检索关键字的分词,通过创建 Query\_Parser 查询解析器实例对象,利用 Index\_Searcher 类的子类对象进行查找,最后利用自定义的排序函数(或缺省的排序函数)对查找结果进行排序。

## 3.4 个性化搜索引擎的部署

文中基于 Lucene 设计并实现了一个支持全文搜索的个性化搜索引擎,可以从指定的网页中通过超链接进行解析和搜索,并把搜索到的每条信息在建立索引后存储到 SQL Server 数据库中。用户可通过浏览器对 Web 服务器提出请求并从该索引数据库中搜索到所匹配的信息。

### (1) 服务器端。

Tomcat 服务器上以 War 包的方式进行部署,搜索引擎的服务端采用 J2EE 的 Servlet 技术实现,用户使

用 GET 方法向服务端提交查询条件,服务器端通过 Servlet 容器接受并分析所提交的条件参数,再调用基于 Lucene 的接口进行搜索操作,最后以 HTTP 消息包的形式将搜索结果发送至客户端。

#### (2) 客户端。

客户端的安装部署十分简单,只要提供一个浏览器即可进行搜索,本系统在设计的时候采用了类似于 Google 的界面风格,充分考虑了实用性和简洁性<sup>[13]</sup>,并可将查询结果(多个 Hits)分页显示在页面。

## 4 结束语

文中给出了个性化搜索引擎的体系结构,综合运用 Java、Lucene、XML 和 SQL Server 等技术实现了该系统原型,提供了信息的增量采集、自动分词以及建立倒排索引的功能。同时系统也存在着一些缺陷:对搜索信息的主题相关度判断较简单,将来还可进一步研究的内容包括:更精细的主题相关度判断算法的改进、网页去噪及消重算法的引入及应用等。

### 参考文献

- [1] Jing Yanan, Zhang Chunwang, Wang Xueping. An Empirical Study on Performance Comparison of Lucene and Relational Database[C]//Proceedings of the 2009 International Conference on Communication Software and Networks. [s. l.]: [s. n.], 2009:336-340.
- [2] 郑榕增,林世平. 基于 Lucene 的中文倒排索引技术的研究

(上接第 104 页)

变换图像去噪算法相比,实验结果中可以消除 Contourlet 变换缺乏平移不变性而引起的失真,达到了更好的去噪效果和更高的 PSNR 值。

### 参考文献:

- [1] 成礼智,王红霞,罗永. 小波的理论与应用[M]. 北京:科学出版社,2004:76-92.
- [2] Do M N, Vetterli M. The contourlet transform: An efficient directional multiresolution image representation[J]. IEEE Trans. Image Processing, 2005, 14(12): 2091-2106.
- [3] 梁栋,沈敏,高清维,等. 一种基于 Contourlet 递归 Cycle Spinning 的图像去噪方法[J]. 电子学报,2005,33(11):2044-2046.
- [4] Bradley A P. Shift-invariance in the Discrete Wavelet Transform[C]//Proc. VIIth Digital Image Computing: Techniques and Applications. [s. l.]: [s. n.], 2003:29-38.
- [5] Cunha A L, Zhou Jianping, Do M N. The nonsubsampling contourlet transform: theory, design, and applications[J]. IEEE Trans. Image Processing, 2006, 15(10): 3089-3101.

- [J]. 计算机技术与发展,2010,20(3):80-83.
- [3] 李永春,丁华福. Lucene 的全文检索的研究与应用[J]. 计算机技术与发展,2010,20(2):12-15.
- [4] 林碧英,赵锐,陈良臣. 基于 Lucene 的全文检索引擎研究与应用[J]. 计算机技术与发展,2007,17(5):184-186.
- [5] 车东. Lucene: 基于 Java 的全文检索引擎简介[EB/OL]. 2002-08-06. <http://www.chedong.com>.
- [6] 马志强,刘利民,苏依拉,等. 基于 Lucene 的站内搜索引擎研究[J]. 内蒙古工业大学学报,2009,28(1):52-59.
- [7] 陈立. 全文检索引擎的设计研究[J]. 现代情报,2007(10):223-225.
- [8] Wan Jian, Pan Shengyi. Performance Evaluation of Compressed Inverted Index in Lucene[C]//Proceedings of the 2009 International Conference on Research Challenges in Computer Science. [s. l.]: [s. n.], 2009:178-181.
- [9] 印鉴,陈忆群. 搜索引擎研究与发展[J]. 计算机工程,2005,31(14):54-56.
- [10] Lucene[EB/OL]. 2006-09-01. [http://lucene.apache.org/java/does/lucene\\_2\\_0\\_0/index.html](http://lucene.apache.org/java/does/lucene_2_0_0/index.html).
- [11] 孙茂松,左正平,黄昌宁. 汉语自动分词词典机制的实验研究[J]. 中文信息学报,2000,14(1):1-6.
- [12] 马志强,周长胜,杨娜,等. 基于中文搜索引擎的分词词典的设计与实现[J]. 铁路计算机应用,2006,15(12):45-47.
- [13] Zhang Hongbin, Liu Juefu. Search Engine Design Based on Web Service and Lucene[C]//Proceedings of the 2009 WASE International Conference on Information Engineering. [s. l.]: [s. n.], 2009:458-461.

- [6] 焦李成,谭山. 图像的多尺度几何分析:回顾和展望[J]. 电子学报,2003,31(12):1975-1981.
- [7] Shensa M J. The discrete wavelet transform: Wedding the à trous and Mallat algorithms[J]. IEEE Trans. Signal Processing, 1992,40(10):2464-2482.
- [8] Wang X H, Robert S H. Microarray image enhancement by de-noising using stationary wavelet transform[J]. IEEE Trans. Nanobioscience, 2003,2(4):184-189.
- [9] Bruce A G, Gao H Y. Understanding waveshrink: variance and bias estimation[J]. Biometrika, 1996, 83(4): 727-745.
- [10] Donoho D L, Johnstone M. Adapting to unknown smoothness via wavelet shrinkage[J]. Journal of the American Statistical Assoc, 1995, 90(432): 1200-1224.
- [11] 刘英霞,王欣. 最佳软门限去噪[J]. 电子学报,2006(1): 167-169.
- [12] Donoho D L. Wavelet Thresholding and W. V. D: A 10-minute Tour[C]//Int. Conf. on Wavelets and Applications. Toulouse, France: [s. n.], 1993.