基于Lucene的搜索引擎的研究与应用

张 俊,李鲁群,周 熔
(上海师范大学 信息与机电工程学院,上海 200234)

摘 要:互联网搜索的精确性一直是衡量搜索引擎性能的重要标志。针对普通搜索引擎的固有缺陷,文中提出了一种应用于新闻检索的搜索引擎。该引擎是利用开源的网络爬虫工具将互联网信息抓取到本地,并利用Lucene开放的API,对特定的信息进行索引和搜索。Lucene是基于Java开发的源代码开放的全文检索工具包,具有高性能、可扩展等特性,是实现搜索引擎的核心组件。通过对Lucene的API进行分析,并在此基础上,构建了索引和搜索的模块,并对网上新闻内容进行实时地搜索。通过与普通搜索引擎对比,该新闻搜索引擎提高了搜索的精确性。

关键词:Lucene;网络爬虫;索引;搜索;新闻搜索引擎
中图分类号:TP39 **文献标识码**:A **文章编号**:1673-629X(2013)06-0230-03
doi:10.3969/j.issn.1673-629X.2013.06.059

Research and Application of Search Engine Based on Lucene

ZHANG Jun, LI Lu-qun, ZHOU Rong
(College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai 200234, China)

Abstract: The precision of Internet searching is important signs of weighing the performance of search engine. In order to resolve the inherent vice about the general search engines, present a search engine applied in news search, which uses the web spider to fetch the information to local machine. The search engine also uses the open API of Lucene to index and search the special information. Lucene is a high-performance, extensible full text search kit based on Java, it is the core component for the realization of the search engine. Give an analysis of the API of Lucene. And on this basis, construct the index and search module, then search the news on the web with real time. By comparing with the general search engine, the news search engine improves accuracy in searching.

Key words: Lucene; web spider; indexing; search; news search engine

0 引 言

信息检索技术经过数年的发展,在检索精确程度和检索速度方面有了较大的提高,基于SQL的数据库检索技术因为其结构特征,在大规模搜索领域,尤其是模糊搜索方面是低效的。在大规模搜索领域,有一类开源搜索工具包为Lucene[1]。文中利用Lucene提供的API,结合开源的网络爬虫来实现Lucene对网页新闻进行检索的功能。

1 Lucene简介

Lucene是一个高性能,可扩展的全文检索工具包,其系统结构运用了优秀的面向对象的思想。Lucene的API由7个子包组成,每个包在全文检索中完成特定的功能,如表1所示:

表1 Lucene索引API

Lucene API	功能
org. apache. Lucene. analysis	语言分析器,用于文章中切分词
org. apache. Lucene. document	field的集合,是索引时的文档结构管理
org. apache. Lucene. index	索引管理,包括索引建立,删除
org. apache. Lucene. queryParser	查询分析器,实现查询关键词间运算
org. apache. Lucene. search	检索管理,根据查询语句,返回检索得到的结果
org. apache. Lucene. store	数据存储管理,主要包括一些底层的I/O操作
org. apache. Lucene. util	一些公共类

这7个包提供了完整索引和搜索功能,可以很方便地在目标系统中嵌入检索功能。如果只是想要搜索存储在本地硬盘上的文件,电子邮件,网页等数据,那

收稿日期:2012-09-09 **修回日期**:2012-12-16 **网络出版时间**:2013-03-05
基金项目:国家自然科学基金资助项目(60473092)
作者简介:张 俊(1984-),男,硕士,研究方向为分布式计算;李鲁群,教授,研究方向为多媒体与网格计算。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20130305.0815.007.html

么可以直接用 Lucene 中关于索引和搜索 API 直接调用,所以,Lucene 不是一个完整的搜索引擎,而只是搜索程序的核心索引和搜索模块而已。可是,很多时候,要搜索互联网中的信息,那么就需要一种程序,该程序被称作网络爬虫。要搜索互联网中的信息,还需要网络爬虫对网页中的内容实现抓取。抓取网页内容可采用几类流行的爬虫工具,如 larbin,Nutch,Heritrix,PolyBot,文中采用了 PolyBot 。下面着重阐述 Lucene 的索引和搜索两大功能。

1.1　基于 Lucene 的索引功能

当网络爬虫将网页抓取到本地时,在搜索其中的内容之前,首先得将被搜索的内容进行索引! 构建索引就好比一本书的目录一样。通过目录便可以很快地找到需要的内容,索引的思想也是类似于此。

1.1.1　文档格式转换

在用 Lucene 进行索引操作之前,需先提取文本和创建文档。因为 Lucene 只识别纯文本字符流格式的信息,所以提取文本时需要将原始文档转换成. txt 文件。通常采用一种称为 Apache Tika 的解析类库从不同格式的文档中(例如:HTML,PDF,MS Word Doc)来侦测和提取出元数据,从而实现了对文本格式进行转换[2]。关于该库的更多信息,请参考网页 http://tika.apache.org/ 。

1.1.2　分析文档

当文本格式的数据被转换后,需要对这些数据内容进行分析,将其转换为语汇单元(tokens),之后可以对其执行一些可选的操作,这类操作包括:将这些语汇单位,比如字母转换为小写,提取单词,去除标点符号,去掉字母上的音调符号,去除常用词等。这些处理过程,就是分析(Analyze)。语汇单元与它的域名结合后,就形成了项(Term)[2,3]。这些操作构成了分析器。分析过程会产生大批的语汇单元,随后这些语汇单元将被写入索引文件中。该过程,可以通过 org. apache. Lucene. index 包的 IndexWriter 类实现。

1.1.3　向索引添加文档

Lucene 本身无法对转换后的文档建立索引,而只能识别处理 Document 类型的文件。向索引添加文档,就是通过 org. apache. Lucene. document 包的 Document 类的 addDocument 方法将数据单元传递给 Lucene 进行索引操作。IndexWriter 初始化之后,即可以向索引目录中添加 Document,其本质是对 Document 中 Field 内的数据进行分析和处理,然后将 Field 加入索引中[4]。为了提高索引速度,可以重复使用 Filed 对象,而不是每次都定义新的对象。

1.2　基于 Lucene 的搜索功能

上面概要地分析了如何为搜索建立索引。但是搜索才是全文检索的最终目标,索引只是实现该目标的手段。关于索引的操作在 org. apache. Lucene. search 包中。该包中最重要的是 IndexSearcher 类。在 Lucene 中,所有搜索相关的操作都需要用这个类[5]。当用 Lucene 进行搜索时,可以选择编程来构建查询语句,也可以选择使用 Lucene 的 QueryParser 类将用户输入的文本转换成 query 对象,再用对象引用该类的 parse 方法,parse 方法里的参数作为查询表达式,被 QueryParser 类对象切分为若干项,切分所用的分词器为 SmartChineseAnalyzer,该分词器是 Lucene 自带的分词工具。QueryParser 类作为解析查询表达式类,调用该类的 parse 方法,可以返回一个查询条件类 Query 的对象,该对象作为参数传递到 IndexSearcher 类的 search 方法中,search 方法的调用将返回 Hits 类的对象[6~8],Hits 类是一个存放有序搜索结果指针的简单容器,而程序则通过 Hits 对搜索结果进行访问。

2　将搜索技术添加到新闻搜索系统

互联网信息量浩如烟海,根据用户上网浏览网页的习惯,可能会对某类领域的新闻比较感兴趣,想在互联网上寻找其相关信息,举例来说:用户浏览搜狐网站可能只是为了寻找关于篮球 NBA 赛事的新闻,如果利用通用的搜索引擎,例如百度,搜索火箭队,出现的前几条信息分别是:火箭队的百度百科,百度贴吧等,由此可以发现,利用公用的搜索引擎,很难快速地查找到关于火箭队的实时新闻,从用户体验角度来讲,是一种缺失。一种较好的代替方法,就是搜索某一权威网站上关于火箭队的新闻,而比较权威的新闻综合门户网站不是很多,常用的有:腾讯,新浪,搜狐,网易等。用户可以选择其中一个或多个网站,在里面找到关于火箭队的实时新闻。基于这个需求,文中实现了基于 Lucene 的新闻搜索引擎功能。

在新闻搜索引擎中创建索引时,引入了一个工具类 Index,该类用来实现对内容创建索引,在创建索引时,先创建 IndexWriter 变量,然后构造函数,再通过 AddNews 方法把每条新闻加入索引中,具体代码为创建 Document 对象,向对象中添加字段,再将 Document 对象添加到索引中。

索引工具类创建好后,可以抓取网站上的内容,并对抓取的内容进行索引,用 PolyBot 中的 HTTP 类的对象调用其 getURL 方法,以此来抓取网页的 URL,然后通过 HTMLPage 类的 getLinks 方法获得一个容器的对象,利用该容器的迭代器,去解释 Web 页面,最后用索引类中的 AddNews 方法来对解释的内容建立索引。

当客户端发出请求后,服务器端首先通过 get( ) 方法获取请求中的查询条件,然后调用搜索的开发包

进行搜索操作,最后把搜索的结果以 HTTP 消息包的形式发送至客户端,从而完成一次搜索操作[9]。以下是搜索操作的部分关键代码:

```
…
…
public void NewsSearch(String qc, ……) throws Exception {
        IndexSearcher ins = new IndexSearcher(…);
        Analyzer a = new SmartChineseAnalyzer();
        String l = qc;
        Query qy = QueryParser.parse(l, "tit", a);
        …
        …
        Hits h = ins.search(qy);
        …
        …
        for (int s = 0; s < h.length(); s = s + 10) {
                for (int j = s; i < h.length(); i++) {
                Document d = h.doc(j);
                String u = d.get("u");
                …}
            …
            …
```

其 Web 前端界面用 html 语言来实现,生成的界面如图 1 所示:

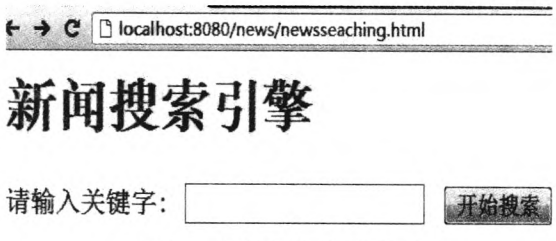

图 1 新闻搜索引擎界面

当用户想要查询关于火箭队的新闻,只要输入“火箭”,便会弹出相应的收索结果,如图 2 所示,新闻搜索引擎会只搜索到用户输入的关键字对应的新闻,而不会出现一些和关键字对应的新闻无关的链接内容。

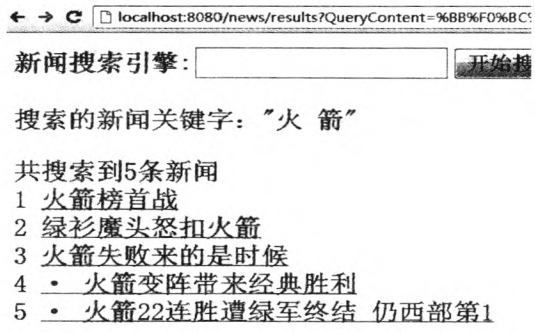

图 2 使用新闻搜索引擎搜索结果

3 结束语

Lucene 提供的 API,为搜索引擎的开发带来便捷。文中提出了一种基于 Lucene 的新闻搜索引擎的应用思路,为新闻网站编辑等工作人员搜索具体的感兴趣的新闻带来了便捷。

参考文献:

[1] 丁兆贵,金 敏. 基于 Lucene 的个性化搜索引擎研究与实现[J]. 计算机技术与发展,2011,21(2):105-108.

[2] 赵 柯,逯 鹏,李永强. 基于 Lucene 搜索引擎设计与实现[J]. 计算机工程,2011,37(16):39-41.

[3] 罗 刚. 解密搜索引擎技术实战 Lucene&java 精华版[M]. 北京:电子工业出版社,2011.

[4] 邱 哲,符滔滔. 开发自己的搜索引擎-Lucene 2.0+Heritex[M]. 北京:人民邮电出版社,2007.

[5] Gospodnetic O,Hatcher E. Lucene 实战[M]. 第 2 版. 北京:人民邮电出版社,2011.

[6] 栾 静,李军锋. 基于 Lucene 全文检索引擎的应用研究[J]. 计算机与数字工程,2010,38(12):184-186.

[7] 宋 佳,诸云强,刘润达. 一种基于 Lucene 改进的全文检索工具包[J]. 计算机工程与应用,2008,44(4):172-175.

[8] 龚 磊,武友新. Lucene 全文检索系统的研究与实现[J]. 计算机与数字工程,2010,38(5):64-67.

[9] 管建和,甘剑峰. 基于 Lucene 全文检索引擎的应用研究与实现[J]. 计算机工程与设计,2007,28(2):489-491.

(上接第 229 页)

[6] 周肖彬,曹存根. 基于本体的医学知识获取[J]. 计算机科学,2003(10):35-39.

[7] van Harmelen F,Hendler J,Horrocks I,et al al. OWL Web Ontology Language Reference[EB/OL]. 2004-02-10. http://www.w3.or g/tr/owl2ref.

[8] 李 薇. 基于本体的知识组织问题研究[D]. 长春:东北师范大学,2007.

[9] Noy F N,McGuinness D I. Ontology Development 101:A Guide to Creating Your First Ontology[R]. Stanford:Stanford University,2001.

[10] 张 良,周长胜. 基于概念本体的视频内容分析框架[J]. 计算机技术与发展,2011,21(12):117-121.

基于Lucene的搜索引擎的研究与应用

作者： 张俊， 李鲁群， 周熔， ZHANG Jun， LI Lu-qun， ZHOU Rong
作者单位： 上海师范大学信息与机电工程学院,上海,200234
刊名： 计算机技术与发展 ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2013,23(6)