第37卷 第16期 计 算 机 工 程 2011年8月
Vol.37 No.16 Computer Engineering August 2011

·软件技术与数据库· 文章编号：1000—3428(2011)16—0039—03 文献标识码：A 中图分类号：TP311.13

基于 Lucene 的搜索引擎设计与实现

赵 珂，逯 鹏，李永强

(郑州大学电气工程学院，郑州 450001)

摘 要：针对目前教育网庞大的 FTP 资源检索困难的问题，提出一种基于 EdtFTPJ 和 Lucene 的 FTP 搜索引擎的设计和实现方案。该方案整体上采用基于 Struts1.2 框架的模型-视图-控制器设计模式，数据采集模块利用基于正则表达式的有限状态自动机抓取数据，索引模块应用倒排索引方法，系统的分词算法使用基于字典的正向最大匹配中文分词法。实验结果表明，该方案具有较高的资源检索率，同时能够保证检索结果的准确性。

关键词：FTP 搜索引擎；Lucene 框架；模型-视图-控制器；有限状态自动机；倒排索引

Design and Implementation of Search Engine Based on Lucene

ZHAO Ke, LU Peng, LI Yong-qiang

(School of Electrical Engineering, Zhengzhou University, Zhengzhou 450001, China)

【Abstract】The number of File Transfer Protocol(FTP) resources on the China Education and Research Network(CERNET) is quite large. It is difficult to find the resources. Because of this problem, a high-performance FTP search engine is designed based on EdtFTPJ and Lucene. In this engine, Struts1.2 is employed to implement Model View Controller(MVC). Data acquisition module uses finite state machine based on regular expression to grab information. Index module uses inverted index method. Word segmentation algorithm uses maximally match Chinese words segmentation based on dictionary. Query Experimental results indicate that the proposed scheme improves the query efficiency, at the same time to ensure the accuracy of the retrieval results.

【Key words】File Transfer Protocol(FTP) search engine; Lucene framework; Model View Controller(MVC); finite state automata; inverted index

DOI: 10.3969/j.issn.1000-3428.2011.16.013

1 概述

目前教育网 FTP 资源检索方式主要是利用操作系统自身的检索方法进行人工检索，其特点是耗时长且效率低[1]，很难满足日益庞大的 FTP 资源检索的需求。针对该问题，本文提出一种基于 Lucene 的 FTP 搜索引擎的设计和实现方案。该方案采用基于 Struts 框架的模型-视图-控制器(Model View Controller, MVC)设计模式，使系统能快速处理请求和响应[2]，然后通过开源组件 EdtFTPJ，利用基于正则表达式的有限状态自动机，抓取 FTP 服务器信息，并且利用基于字典的正向最大匹配中文分词法来建立倒排索引，使系统能快速处理用户的检索请求。

2 整体设计结构

FTP 搜索引擎的整体设计结构如图 1 所示。该搜索引擎包含数据采集模块、分析模块、Lucene 索引建立模块和 Web 服务器响应检索请求模块。

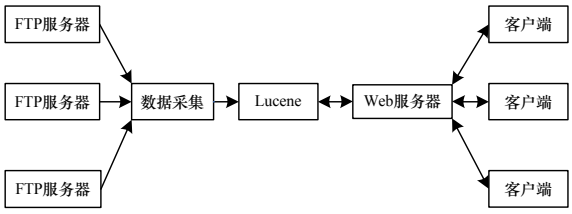

图 1 搜索引擎结构

由于整个系统采用目前较为流行的 B/S(浏览器/服务器)架构，因此将 Struts1.2 框架引入到系统中，这样系统就实现了 MVC 的设计模式，减少了模块之间的耦合性，提高了系统的可重用性和适用性[3]。

3 基于 EdtFTPJ 的信息采集模块的设计

3.1 信息采集

EdtFTPJ 是一个开源的 FTP 客户端软件，它可以方便地嵌入到 Java 开发的系统中。本文利用其提供的 API (Application Programming Interface)，采用递归的方法将 FTP 服务器的相关文件信息抓取到本地 list.txt 文件中。信息片段如下：

```
.\办公软件:
10-29-08  06:34PM  <DIR>       WPS
08-24-08  05:02PM  <DIR>       编辑软件
.\办公软件\WPS:
08-24-08  01:19PM  <DIR>       软件信息
10-17-08  06:56PM  29420592    wps2007.exe
11-12-02  12:00AM  975         使用必读.txt
```

信息片段以记录的形式存储在本地文件中，而记录又分为 2 种。若记录以“.”开头，则该记录表示父目录，并且此目录是它下面文件或目录的父目录；若以数字开头，则说明该条记录为文件或者目录。该条记录分为 3 段，第 1 段为最后修改时间；第 2 段为标记，若标记为<DIR>，则说明该记录是目录，若标记为数字，则数字表示该文件的大小；第 3 段为文件或目录的名称，若是文件，则包含其后缀。

基金项目：国家自然科学基金资助项目(60841004, 60971110)；郑州大学创新性实验基金资助项目(2009cxsy100)

作者简介：赵 珂(1988－)，男，本科生，主研方向：软件工程，Web 信息处理，数据挖掘；逯 鹏(通讯作者)，副教授、博士；李永强，硕士研究生

收稿日期：2011-02-18 **E-mail：**justke@163.com

根据有限状态自动机的状态设定规则，需要有一个初态和一个终止状态，于是可将该处理过程设为 5 种状态，分别为：初始状态即准备读取状态，处理父目录状态，处理文件或者目录记录状态，处理多余行状态，终止状态即读取结束状态，在编程实现中用变量 state 分别设置 0~4 这 5 个不同值来表示。

3.2 信息处理

匹配 3.1 节信息片段中日期的正则表达式如下：

((\d\d-\d\d-\d\d)\s+(\d\d:\d\d[AP]M))

正则表达式分析：(1)(\d\d-\d\d-\d\d)，匹配日期开头“11-12-02”年-月-日的格式；(2)\s+，匹配“11-12-02”与“12:00AM”之间的一个或多个空白符；(3)(\d\d:\d\d[AP]M)，匹配“12:00AM”时间的格式。

在实际应用中，根据不同系统的需求，利用正则表达式 group 的特性，用 group(0)将整个日期提取出来，也可用 group(1)提取年月日，也可用 group(2)提取时间，使用灵活。

根据 3.1 节中信息格式的分析，可设置 4 个转移函数，这 4 个转移函数分别为处理父目录转移函数、处理记录转移函数、处理结束转移函数和处理多余行转移函数，它们的实现方法如下：

(1)处理父目录转移函数。若该记录以“.”开头，则程序转入处理父目录的处理过程。否则，转入下一个转移函数。

(2)处理记录转移函数。若该记录开头匹配上述日期正则表达式，则程序转入处理记录的的处理过程。否则，转入下一个转移函数。

(3)处理结束转移函数。若该处理过程已到达文件末尾，则转入处理结束的处理过程。否则，转入下一个转移函数。

(4)处理多余行转移函数。若处理过程都不满足前 3 种转移函数，则程序转入处理多余行的处理过程。处理完成后，继续进行下一条记录的处理。

有限状态自动机的实现如图 2 所示。假设有限状态自动机目前处于终止状态即读取结束状态。首先判断是否有新抓取的信息，若有，则将抓取的信息一次性读取到缓存字符串中，这样避免了频繁的 I/O 操作，提高了效率。若成功读取，则将自动机置于信息读取状态，将变量 state 设为 0，进入读取阶段。否则保持 state 不变。

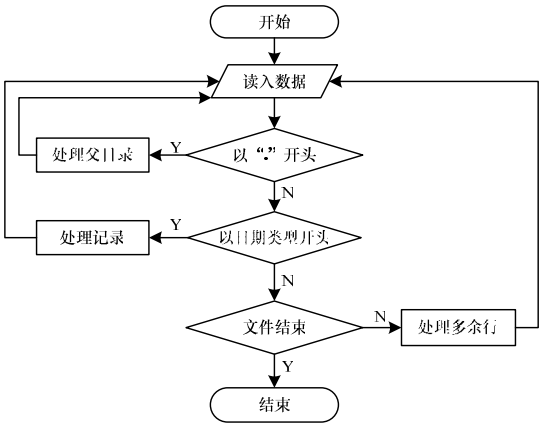

图 2 有限状态自动机流程

进入读取阶段后，每次从缓存字符串中传递给自动机一条记录，根据 4 种状态转移函数来控制过程转移，若以“.”开头，则转入处理父目录过程，将 state 设为 1；若以日期类型开头，则转入处理记录过程，将 state 设为 2；若文件结束，则转入处理结束过程，将 state 设为 4；若不是以上情况，则转入处理多余行过程，将 state 设为 3。

每处理一条记录，则需要创建一个 JAVA 对象的实例来存储该条记录的信息。本文以 File 类型为例，如图 3 所示。File 对象含有 name、size、date 等 Field 来存储信息，每提取出来一条有效的记录，就放入 File 对象的实例中去。保存完 File 类型后修改自动机状态为初态即将 state 设为 0，进行下一条记录的处理。存储完成后将 File 类型转化为 Lucene 可识别 Document 类型，供 Lucene 建立索引。

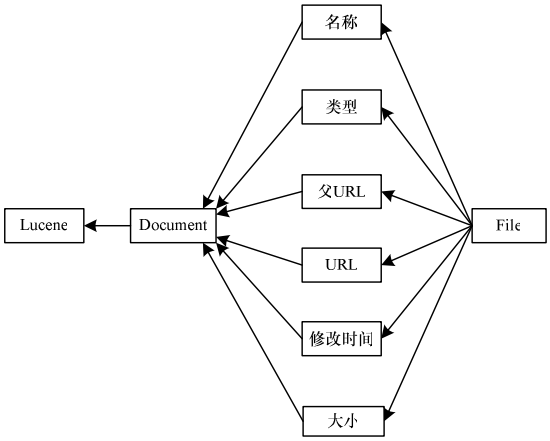

图 3 File 类型与 Document 类型的转化

4 基于 Lucene 的索引建立及检索模块的设计

4.1 倒排索引

倒排索引是一种面向单词的索引机制，通常由词(关键字)和出现情况两部分组成，对于索引中的每个词(关键字)，都跟随一个列表(位置表)和出现频率。

假设 2 篇文档：

A: Its color is red.

B: Red is a kind of color.

由表 1 可知，关键词的数量随文本内容的增长而线性增长[4]。而在表 2 的倒排索引中，若出现与上文相同的关键词时，则仅修改相应的记录信息，由此可见，关键词数量并非随着文本内容的增长也线性增长，这样有利于降低索引文件的占用空间，提高检索效率。

表 1 一般索引

出现单词	出现文档	出现次数
color	A	1
red	A	1
red	B	1
color	B	1
kind	B	1

表 2 倒排索引

出现单词	出现文档	出现次数
red	A,B	1,1
color	A,B	1,1
kind	B	1

4.2 检索模块

FTP 搜索引擎接收到来自客户端的检索请求时，对客户端输入的查询关键字进行处理，如去除多余空格、识别特殊符号等，然后将处理过的关键字输入到 Lucene 的查询分析器 QueryParser 中，该查询分析器将调用由 JE 分词组成的分析器实现分词，分析后将生成 Lucene 内部的查询条件类 Query 的对象，然后将 Query 对象传递给 IndexSearcher 类进行搜索，最后将搜索结果以 Hits 类型返回。Hits 类型为一链式存储结构的容器，可以进行类似于链表的各种操作，方便对检索结果进行人性化的管理[5-6]，该系统检索模块与传统数据库的模糊匹配对比见表 3。

表3 检索模块对比

关键技术	本文设计的搜索引擎	传统数据库模糊查询
索引	建立倒排索引，检索速度快、效率高	无索引，逐个遍历进行匹配，比有索引的检索效率有多个数量级的下降
可定制性	通过分析接口可方便的定制出符合需要的索引规则	无接口或者接口复杂，无法定制
匹配效果	通过词元进行匹配，由语言分析接口实现中文支持	匹配方法单一，可能检索出无关信息
匹配度	通过匹配度控制算法输出按匹配度排序的结果集	无匹配度控制算法，检索后输出所有的无序结果集
适用情况	高负载、大索引量、效率准确率高的模糊检索应用	使用率低，匹配规则简单或者查询量较小

4.3 分词算法

系统分词算法采用开源组件JE分词来实现，它是基于最大匹配算法的开源组件，它全面兼容Lucene，提供简单实用的API，可以方便地被Lucene调用。

例如，对于句子“据路透社报道，印度尼西亚社会事务部一官员星期二(29 日)表示，日惹市附近当地时间 27 日晨5时53分发生的里氏6.2级地震已经造成至少5 427人死亡，20 000余人受伤，近20万人无家可归”，分词效果如下：

据 | 路透社 | 报道 | 印度尼西亚 | 社会 | 事务 | 部 | 官员 | 星期二 | 29日 | 表示 | 日惹 | 市 |附近 | 当地时间 | 27日 | 晨 | 5时 | 53分 | 发生 | 里氏 | 6.2级 | 地震 | 已经 | 造成 | 至少 |5427人 | 死亡 | 20000 | 余人 | 受伤 | 近 | 20万人 | 无家可归 |

基于字典的正向最大匹配中文分词法的原理如下：对于一个字符串S，按从前到后的顺序扫描，对扫描的每一个字，从词库中寻找最长的匹配。例如：S=“印度尼西亚社会事务部”，词库中有“印度尼西亚”、“印度”、“社会”、“事务”等词。当扫描到“印”字，那么从“印”字开始，向后分别取1个、2个、5个字(“印”，“印度”，“印度尼西亚”)，词库中最长的匹配为“印度尼西亚”，所以分词取最长的匹配，就从“亚”字后面分开，扫描器下一次扫描“社”字。此外，JE分词具有学习功能，可以将新词加入到词典当中，这个功能对于FTP搜索引擎分词模块的扩展具有非常重要的意义。

5 实验结果与分析

基于本文的理论分析，在JDK1.6、Tomcat6.0、Struts1.2的软件平台和Intel Core Due CPU T5750(2.0 GHz)、2 GB内存的硬件平台下对郑州大学5个FTP服务器大约500 000个文件和文件夹的数据规模下建立了索引，并提供了相应的检索服务。该搜索引擎与传统搜索引擎的模块对比见表4。由表4可知，该搜索引擎在数据源、索引、分词算法和查询分析等模块都比传统的搜索引擎有很大的改进和提高。

表4 搜索引擎模块对比

关键技术	本文设计的搜索引擎	传统的搜索引擎
数据源	通过组件EdtFTPJ提供的接口实现对数据源的定制	采用引擎自身抓取模块，功能单一
索引内容	对数据源的信息进行过滤，仅对需要的信息进行索引	缺乏通用性，对所有的数据源进行索引
索引	采用倒排索引，有专用接口用于增量式索引	采用一般索引，不支持增量式索引
分词算法	采用基于字典的正向最大匹配中文分词算法，分词准确，具有学习功能	采用二分法等简单分词算法，效率和准确率较低
查询分析	通过查询分析接口实现定制的查询语法，内容丰富，查询结果匹配度高	缺乏通用查询分析接口，采用字符串匹配等简单查询语法，无法定制

5个服务器的信息采集时间和建立索引的时间见表5。由表5可知，FTP站点信息的采集能在短时间内完成，但是从抓取的时间来看，效率并不是很高，采集算法还有待改进。但是由于FTP服务器站点信息的采集效率并不直接影响用户检索效率，因此该方法能够满足系统要求。

表5 信息抓取时间和文件数量

FTP服务器	信息抓取时间/s	文件和文件夹数量
mtv.zzu.edu.cn	137.438	7 339
ebook.zzu.edu.cn	235.140	30 702
tv.zzu.edu.cn	510.375	52 753
soft.zzu.edu.cn	1 633.245	170 057
media.zzu.edu.cn	3 415.167	386 510

信息检索的时间见表6，随即选择频率出现相对较高的关键词进行检索，由表6的检索时间可知，检索结果都能在1 s之内完成，能够快速响应用户的检索请求。

表6 关键词检索的时间和数量

检索关键词	检索时间/s	检索出的数量
测试	0.047	767
电影	0.078	1 307
书籍	0.281	222
音乐	0.078	1 732
新闻	0.041	648

检索关键词“电影”的结果如图4所示，结果上方显示检索的结果数量和检索时间，检索结果中包含了该文件的名称、格式、大小、类型，最后修改时间以及URL，检索结果比较全面。

图4 检索结果

6 结束语

本文基于EdtFTPJ和Lucene等开源组件、运用MVC、有限状态自动机和基于字典的正向最大匹配中文分词法等相关理论，设计并实现了FTP搜索引擎。该搜索引擎已经对郑州大学5个FTP服务器约500 000个文件和文件夹建立了索引并提供检索服务。从实验结果可以看出，该方案具有较高的检索效率和准确的检索结果。

参考文献

[1] Almpanidis G, Kotropoulos C, Pitas I. Combining Text and Link Analysis for Focused Crawling——An Application for Vertical Search Engines[J]. Information Systems, 2006, 9(4): 1-23.

[2] 张 宇, 王映辉, 张翔南. 基于Spring的MVC框架设计与实现[J]. 计算机工程, 2010, 36(4): 59-62.

[3] Cavaness C. Programming Jakarta Struts[M]. [S. l.]: Oreilly & Associates Inc., 2004.

[4] 郭立力, 赵春江. 高效FTP搜索引擎的设计与实现[J]. 华南理工大学学报, 2009, 37(1): 135-139.

[5] Shepherd S J. Concepts and Architectures for Next-generation Information Search Engines[J]. International Journal of Information Management, 2007, 27(1): 3-8.

[6] Leroy G. An End User Evaluation of Query Formulation and Results Review Tools in Three Medical Meta-search Engines[J]. International Journal of Medical Informatics, 2007, 76(11/12): 780-789.

编辑 任吉慧