基于 Lucene 全文检索引擎的应用研究

朱雪莲

（新疆艺术学院基础部（思政部），新疆 乌鲁木齐 830049）

摘 要：Lucene 是一个强大的全文索引引擎工具包，它的全文检索技术是信息检索领域广泛使用的基本技术，具有访问索引时间快、多用户访问、跨平台使用的特点。介绍了一个高性能的全文检索引擎——Lucene 开源系统，详细分析了 Lucene 的系统结构、全文索引机制，然后将其引入具体应用，给出了一个基于 Lucene 全文检索技术的具体实例。

关键词：全文检索技术；Lucene；索引

中图分类号：TP319 **文献标识码**：A **文章编号**：1674-7720(2010)22-0003-03

Implementation and research of full-text retrival engine based on Lucene

ZHU Xue Lian

(Foundation Department, Xinjiang Art Institute, Urumqi 830049, China)

Abstract: Lucene is a full-text index/retrieval software package, its full-text retrieval technology is a basic technology used. It has high access speed, supportsmulti-user accesses and can be used in a cross-platformway. Firstly, Lucene, an advance full-text retrieval engine is introduced, system structure, full text indexing are analysed in detail, Then employ it in the application, demonstrate an example based on lucene technology.

Key words: full-text retrieval technology; Lucene; search engine; index; tokenize

Lucene 作为一个开放源代码全文检索工具包，具有优异的索引结构和良好的系统架构，不仅可以通过它来构建具体的全文检索应用，而且能方便地集成到各种系统软件中，本文对 Lucene 进行深入的研究和分析，以此为基础设计实现了一个以商业网站中构建搜索引擎的实例。

1 全文检索引擎 Lucene

1.1 Lucene 概述

Lucene 是用 Java 写的全文检索引擎工具包，并不是一个完整的全文检索引擎，而是一个全文检索引擎的架构，可以提供多个应用程序编程接口函数和数据存储结构，并能方便地嵌入到各种应用中，从而实现针对应用的全文索引/检索功能。

1.2 Lucene 系统结构

Lucene 的系统结构中运用了面向对象的设计思想，定义的索引文件格式与平台无关，并通过抽象将系统的核心组成部分和具体的平台部分设计为抽象类，与具体平台相关的部分例如文件存储也封装为类，经过层层处理，形成了一个低耦合、高效率、容易二次开发的检索引擎系统。系统结构图如图 1 所示。

从图 1 看到 Lucene 系统是由基础结构封装、索引核心、对外接口三大部分组成。其中索引核心部分是系统的重点。Lucene 中共有 7 个子包，每个包的具体功能见表 1，核心类包主要有：org. apache.lucene. analysis；org. apache. lucene. Index；org. apache.lucene. search。

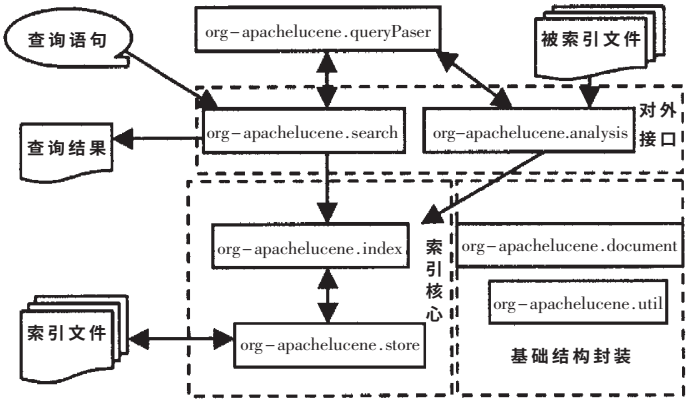

图 1 系统结构图

表 1 Lucene 的组成结构

org.apache.Lucene.search	检索入口,提供了在索引上检索的类
org.apache.Lucene.index	索引入口,提供了用于访问与维护索引的类
org.apache.Lucene.analysis	语言分析器,提供了将文本转化为可索引的词(token)的类
org.apache.Lucene.queryParser	查询分析器,用户可以定制
org.apache.Lucene.document	存储结构,文档(Document)的抽象描述
org.apache.Lucene.store	底层 IO/存储结构
org.apache.Lucene.util	一些公用的数据结构

1.3 Lucene 全文索引机制

Lucene 索引存储结构采用层次结构,主要由索引、字段、文档、字、词组成,在存储正向信息时通常是按层次保存从索引到词的包含关系,即 Lucene 索引文件由若干段(Segment,相当于表)组成,每一段由若干的文档(document,相当于表中记录)组成,每一个文档由若干的域(Field,相当于表中字段)组成,每一个域由若干的项(Term,相当于表中数据)组成;而反向信息则保存了词典到倒排表的映射。因此,索引存储结构设计比较通用,输入输出结构类似于数据库中的表→记录→字段,很多文件、数据库等都能较为方便地映射到 Lucene 的索引存储结构/接口中。

Lucene 访问索引的时间较快,这是因为大部分数据库引擎是用 B 树来维护索引结构的,更新索引时会导致大量的输入和输出操作,而通过 Lucene 构建的索引文件在扩展索引时,是将新创建的小索引文件定期地合并到原先的大索引文件中,从而提高了索引效率。

2 基于 Lucene 构建搜索引擎的具体应用

在实现利用 Lucene 构建搜索引擎的具体应用时,以在商业网站中构建一个搜索引擎为例,通过爬虫将各大 IT 门户网站提供的商品信息抓取下来,然后对网页内容进行数据信息抽取并转换为统一格式的文本文件,并构建专业数据库和主题词典,同时将词典内容扩充到中文分词模块中;中文分词模块对文件处理器处理生成的文本文件进行分词处理,并提供词元序列供索引器索引,并将索引结果保存到索引数据库中;使用 Tomcat Web 服务器发布系统的检索页面,当用户通过 Web 界面输入要查询的关键词并提交后,搜索器到索引数据库中进行检索,检索到的结果经过处理之后,作为响应发送给用户。系统结构如图 2 所示。本文将介绍与 Lucene 相关的设计与实现,并利用 Lucene 工具包提供的类对其扩展来实现具体的应用。

首先对抓取下来的网页内容进行结构化的抽取,并对抽取的内容按固定格式保存、完成主题词库的构建、产品数据库的构建、数据库处理类的构建等几项任务,为后续的索引入库、检索打下基础。

2.1 索引库的建立

在索引阶段需要定义 Lucene 的 Document 格式和构

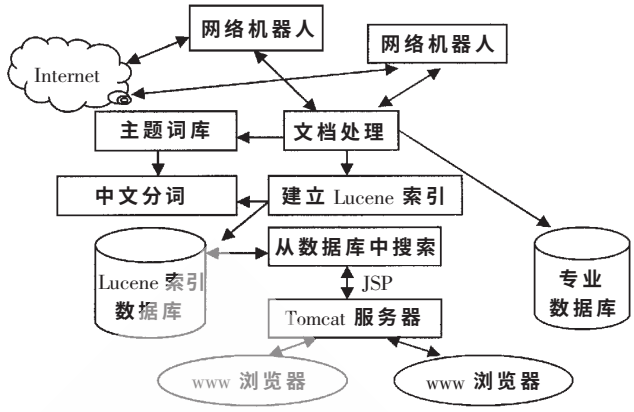

图 2 系统结构图

建索引的处理类。Lucene 的索引库是通过接口添加一条条索引记录实现的,首先需要构造一个 Document 文档对象,确定 Document 的各个域,IndexWriter 负责接收新加入的文档,并写入索引库中。

本实例中 ProductDocument 类静态方法为一个 Product 对象构建 Lucene 的 Document,当中包含了 7 个 Field,分别为 identifier(产品 ID)、indextime(索引时间)、producturl(产品 URL)、category(产品分类)、name(产品名称)、type(产品型号)、all。前 6 个 Field 与数据库中的内容有直接的对应关系,而 all 则是将 category(产品分类)、name(产品名称)、type(产品型号)拼接起来,为用户搜索时提供一个默认 Field。具体定义如表 2 所示。

表 2 ProductDocument 类中 Field 属性配置情况

Field	Store	Index	Tokenized
identifier	√	√	
indextime	√	√	
producturl	√	√	
category	√	√	√
name	√	√	√
type	√	√	√
all	√	√	√

关键代码如下:

```
import org. apache. lucene. document;
public class ProductDocument
{
    //Field 名称,当前产品在数据库中的 ID
    private static final String PRODUCT_ ID="p roductid";
    ……
    public static Document buildProductDocument(Porduct pro-
duct, int id)
    {
        Document doc = new Document( ) ;
        // 此处构建 6 个 Field
        Field identifier;
        Field indextime;
```

```
        Field p roducturl;
        Field category;
        Field name;
        Field type;
        //最后一个 Field 将 category、name、type 的信息综合起来,默认在此进行检索
        Field all;
        // add all
        doc. add ( identifier) ;
        ……
        return doc;
    }
}
```

下面的代码是索引类的代码,它用于向 Lucene 索引中添加 Document。

```
public classProductIndexer{
    ……
private void initialize() throws Exception{
    analyzer = new MMAnalyzer();
    FileReader reader = new
    FileReader(dictionary_file);
    ((MMAnalyzer)analyzer). addDictionary(reader);
    writer = new IndexWriter( indexPath, analyzer,
    true);
}
public void close(){…… }
public void addProduct (Product product, int id) throws Exception{
    writer. addDocument ( ProductDocument.
    buildProductDocument(product, id));
        }
        ……
}
```

在 initialize 方法中,初始化了一个 JE 分词的 MMAnalyzer 实例,然后将生成的主题词库添加到该实例中。addProduct 方法将两个参数 Product 和 id 传递到 ProductDocument.bulidProductDocument 方法里,然后调用 IndexWriter 的 addDocument 方法,把生成的产品加入到索引中。至此,数据库与索引的建立结束。

2.2 检索

Lucene 的检索接口主要由 QueryParser、IndexSearcher、Hits 三个类构成,QueryParser 是查询分析器,IndexSearcher 是索引搜索器,检索时,用户提交检索关键字,先调用 Lucene 的查询分析器分析用户提交的查询,然后调用 IndexSearcher 类进行搜索;返回结果为 Hits 类,通过它再访问 Document=>Field 中的内容。

本文在深入剖析 Lucene 的系统结构和索引机制的基础上,实现了一个商业领域搜索引擎的实例,检索的结果由结构化的数据组成,描述针对性很强,响应速度快、查准率高。今后,将进一步增加对动态页面的索引和语义分析来提高搜索的精度。

参考文献

[1] 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,2007,1.28(2):489-491.

[2] 车东.在应用中加入全文检索功能—基于 Java 的全文索引引擎 Lucene 简介[EB/OL].http://www.chedong.com/tech/lucene.html.2009.03.20.

[3] 邱哲,符滔滔.开发自己的搜索引擎[M].北京:人民邮电出版社,2007.

[4] 李广丽,刘觉夫.垂直搜索引擎的研究与实现[J]. 情报杂志,2009,10.28(10):144-147,169.

(收稿日期:2010-07-22)

作者简介:

朱雪莲,女,1976 年生,讲师,硕士,主要研究方向:信息处理与检索。