基于 Lucene 的全文检索系统模型的研究及应用*

梁 弼,王光琼,邓小清

(四川文理学院 计算机科学系,四川 达州 635000)

摘 要:分析了 Lucene 的系统结构及检索原理,设计了一个基于 Lucene 的全文检索系统模型,并将该系统模型应用到自动答疑系统中进行实验。实验结果表明,以 Lucene 作为核心的检索系统不仅建立索引的效率高,而且检索速度也较快。

关键词:Lucene;全文检索;索引;搜索速度

中图分类号:TP393 **文献标识码**:A **文章编号**:1674-7720(2011)01-0044-03

Research and application of full-text retrieval model based on Lucene

Liang Bi, Wang Guangqiong, Deng Xiaoqing

(Department of Computer Science, Sichuan University of Arts and Science, Dazhou 635000, China)

Abstract: This paper analyses the structure of Lucene system and the search theory of Lucene at first. Then a Lucene-based full-text retrieval model is designed and applied to the automatic answering system. The experimental results show that Lucene-based search system not only has the high efficiency of creating index, but also has faster search speed.

Key words: Lucene; full-text retrieval; index; search speed

随着信息技术的快速发展,互联网上的信息呈爆炸式增长,这种趋势使得用户在得到更多信息的同时,也不可避免地加剧用户筛选信息的难度,为了使用户在海量数据中能快速地找到有效数据,高性能的信息检索系统显得越来越重要。

对海量数据检索而言,全文检索是唯一高效的解决方案。以前借助商业数据库(如 SQL Server)提供的全文检索,由于其运行环境、效率或商业成本方面的原因无法满足需求。现在一般通过 Lucene 将数据库中的数据全部索引,然后提供查询,进而实现对数据的全文检索[1]。本文通过剖析开放源码的全文检索技术——Apache Lucene,构建了一个基于 Lucene 的全文检索系统模型,并结合 Struts、Ajax 等相关技术,设计并实现了一个以 Lucene 作为核心的自动答疑系统。

1 Lucene 全文检索

Lucene 是一个用 Java 语言实现、成熟、开源的软件项目,是一个高性能、可扩展的信息检索工具集,可以方便快捷地融入到应用程序中,以增加索引和搜索功能。

* 基金项目:四川文理学院 2009 年科研项目(2009B02Z)

1.1 Lucene 的系统结构

Lucene 的系统结构采用分层的方式构建,各模块间基于协议进行交互,形成了具有松耦合特征的体系结构,这大大增强了系统的弹性。Lucene 系统主要由基础结构封装、索引核心和对外接口三大部分组成,其结构如图 1 所示[2]。

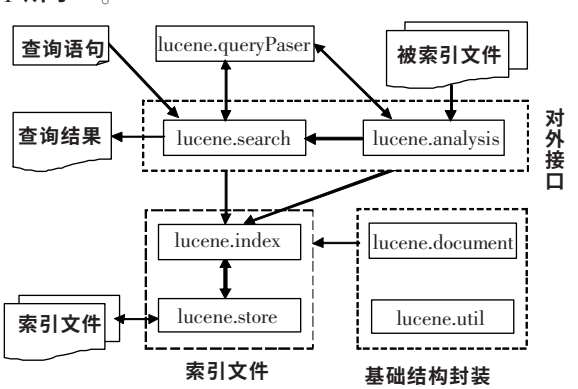

图 1 Lucene 系统结构图

从图 1 可以看出,Lucene 的源码主要分为 7 个模块,其核心类有 3 个,分别是:analysis、index 和 search。analysis 主要用于切分词,切分词的具体工作由 analyzer

的扩展类来实现；index 主要提供库的读写接口，通过它可以创建库、添加删除记录以及读取记录等；search 主要提供检索接口，通过该包可以输入条件并得到查询结果集，与 queryPaser 包配合还可以自定义查询规则。

1.2 Lucene 检索原理

Lucene 的检索算法属于索引检索，即用空间来换取时间，对需要检索的文件或字符流进行全文索引，在检索的时候对索引进行快速搜索，得到检索位置，该位置记录了检索词出现的文件路径或者某个关键词。实际上，Lucene 的检索过程是将模糊查询变成多个利用索引进行精确查询的逻辑组合过程。并且，Lucene 的 API 接口设计比较通用，很多传统应用的文件、数据库等都可以方便地映射到 Lucene 的存储结构/接口中。因此，在一定程度上可以把 Lucene 看着一个支持全文索引的数据库系统，其检索原理如图 2 所示[3]。

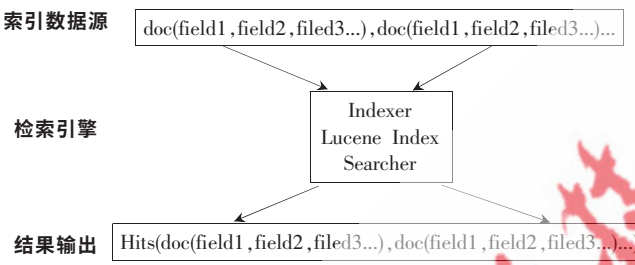

图 2 Lucene 检索原理图

2 系统模型

根据开源 Lucene 的相关理论知识（如检索原理），构建了一个通用、易于扩展、基于 Lucene 的全文检索系统模型，具体模型如图 3 所示。该模型主要包括的功能模块有文档抽取模块、全文检索核心处理模块和输入输出模块。其中，全文检索核心处理模块是模型中最重要的模块[4]。

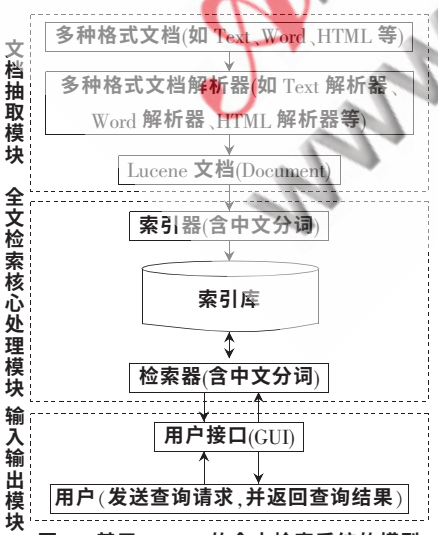

图 3 基于 Lucene 的全文检索系统的模型

由图 3 可知，基于 Lucene 的全文检索系统处理的数据源可以是一些常见的文档格式，如 Text、Word 以及 HTML 等，系统首先通过文档解析器提取出这些文档的原始数据信息并保存为 Lucene 能够处理的文档类型（Document），然后 Lucene 的索引器将接收这些文档，并对其内容进行分析，最后提取索引项并生成索引库。然后以索引库为基础，可以设计满足各种查询需求的检索器和符合用户使用特点的可视化个性操作界面，并将查询结果展现给用户。

3 模型应用

在远程网络教育中，若能对以前所提的问题和已经回答过的答案进行全文检索，用户便可以在系统中找到自己需要的答案，实现自动答疑的目的。

3.1 系统设计

根据前面设计的 Lucene 全文检索系统模型，采用 Struts 框架作为系统的整体基础架构，并运用 Ajax 对 Struts 在表示层上的补充[5]，设计并开发一个自动答疑系统，该自动答疑系统的总体设计如图 4 所示。

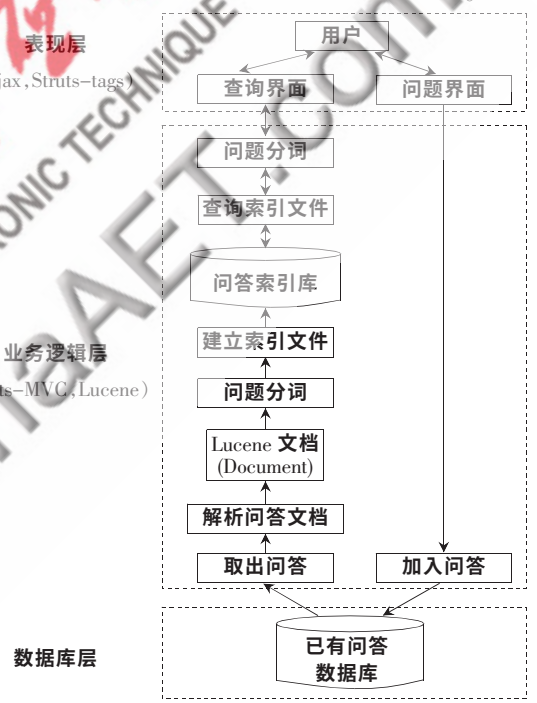

图 4 自动答疑系统

从图 4 可以看出，该自动答疑系统采用表现层、业务逻辑层和数据库层的三层结构设计模式。其中，表现层为用户界面，它为用户提供交互接口，主要包括查询界面和问答添加界面，并采用 Struts-tags 和 Ajax 技术来实现，通过 Ajax 技术可以实现按需取数据、局部更新页面的功能，从而增强了用户体验[6]。业务逻辑层主要完成问答的全文检索功能，具体包括在查询问题时检索索引库和从关系数据库取出问答并向索引库中添加索引文件，主要采用 Struts 和 Lucene 技术来完成。其中，Struts 是一个优秀的基于 MVC（Model View Controller）模式的开源框架，它通过控制器将表示逻辑和业务逻辑解耦。数据库层完成问题收集功能，对提出的问题和相应的答

案添加到数据库中，使用 MySQL 来保存所有问答，或直接存放在硬盘目录上。

3.2 关键技术

实现该自动答疑系统的关键技术主要集中在中文分词、索引管理和结果排序三方面。

(1)中文分词

在 Lucene 中执行分词任务的是 Analyzer 对象，该对象中最关键的方法是 TokenStream 方法，通过执行该方法可以返回一个包含 token 的集合，也即 TokenStream 对象。TokenStream 本身是一个有着类似迭代器接口的抽象类，其具体类有两种：一种是以 Reader 对象作为输入的 Tokenizer 对象，另一种是以另一个 TokenStream 对象作为输入的 TokenFilter。因此在该自动答疑系统中，问题分词采用 Lucene Analyzer 分词器来实现，即首先创建自己的 Analyzer 对象，以及与其相关的 Tokenizer 和 TokenFilter，然后通过这几个类的有机配合进而实现问题的中文分词。

(2)索引管理

从根本上说，索引管理主要包括两方面内容：建立索引和基于索引进行的检索。

建立索引是对文本内容切分词后索引入库。切分后的 token 通过 Lucene.index 索引器的处理最终添加到索引库中，Lucene.store 存储器负责数据存储管理，主要包括一些底层的 I/O 操作。核心代码如下：

```
Analyzer luceneAnalyzer=new StandardAnalyzer();
IndexWriter indexWriter=new IndexWriter(indexDir,
luceneAnalyzer,true);
indexWriter.addDocument(document);
```

检索是在建立好的索引上进行的搜索，并根据查询条件返回结果。其核心代码如下：

```
QueryParser queryParser=new QueryParser(field,new
StandardAnalyzer());
Query luceneQuery=queryParser.parse(QueryString);
IndexSearcher indexSearcher=new IndexSearcher
(IndexReader.open(indexDir));
Hits hits=indexSearcher.search(luceneQuery);(hits 用来保存检索结果集的对象)
```

(3)结果排序

排序的基本原则是尽可能地把对用户更有价值的问题排在前面而不影响性能。Lucene 是按照自己的相关度算法(score)对结果进行排序的，除了匹配 score 外，还可以用索引记录的 ID 来进行排序，所以较为高效的排序方法是：在索引时，让进入 Lucene 全文的顺序对应着一定的规则(如用户对问答的评价值越高反映该问题的价值越高)，而在检索时让检索结果按照索引记录的 ID 进行倒排文档(invert document)。

3.3 实验结果

本实验的硬件环境为处理器：Intel Pentium 4 CPU 3.20 GHz(2 CPUs)，内存：三星 DDR2 1 GB，硬盘：酷睿 120 GB 等；软件环境为 JKD 1.6，Eclipse 3.2，MyEclipse 6.5，Tomcat 6.0，MySQL 5.0 等。实验时使用所设计的自动答疑系统，不仅对存储在关系数据库和存放在硬盘目录上不同大小的问答文本文件建立索引所花费时间进行对比，而且还对关系数据库和索引库中的问答文件进行检索所需的时间进行比较，具体实验结果如表 1 所示。

表 1 自动答疑系统实验结果

文件(.txt)大小/KB	建立索引所需时间/ms		问答查询所需时间/ms	
	硬盘	关系数据库	索引库	关系数据库
21	1 082	1 324	109	113
63	1 123	1 329	134	172
165	1 240	1 556	197	247
268	1 308	1 671	211	294
450	1 327	1 697	280	368
852	1 356	1 895	396	415
1 509	1 469	2 013	465	592
3 534	2 015	3 116	633	826

从表 1 可以看出，同样采用 Lucene 作为检索系统的核心，数据源同为文本文件时，对硬盘文件目录建立的索引的效率要比对关系数据库建立索引的效率要高，对索引库中的文件检索速度比对关系数据库中文件检索速度要快，并且随着数据源的增大，效果越来越明显，实验结果基本符合理论要求。

Lucene 是当前比较成熟的检索技术，利用它可以方便地实现全文检索。本文在剖析 Lucene 相关技术的基础上，构建了一个基于 Lucene 的全文检索系统模型，并将该模型应用到一个具体实例——自动答疑系统中进行实验，实验结果表明，以 Lucene 为核心的检索系统不仅建立索引的效率高，而且检索速度也较快。

参考文献

[1] 郎小伟，王申康.基于 Lucene 的全文检索系统的研究与开发[J].计算机工程，2006，4(2)：95-99.

[2] 李晶，文登敏.基于 Lucene 的全文检索引擎的研究与应用[J].淮阴工学院学报，2008(2)：57-59.

[3] 刘建湘，杨文涛.基于 Lucene 的搜索引擎在 Struts 中的应用[J].软件导刊，2007(2)：53-54.

[4] 吴青，夏红霞.基于 Lucene 的全文检索引擎的应用与改进[J].武汉理工大学学报，2008，7(30)：145-147.

[5] 孙晓峰.基于轻量级框架的互动问答平台的设计与实现[D].北京：中国地质大学，2008(5)：13-23.

[6] 谌湘倩，狄文辉.基于 SSH 框架与 AJAX 技术的 Java Web 应用开发[J].计算机工程与设计，2009，30(10)：2590-2592.

(收稿日期：2010-09-05)

作者简介：

梁弼，男，1982 年生，硕士，助教，主要研究方向：智能信息处理与应用软件。

王光琼，女，1965 年生，硕士，副教授，主要研究方向：软件工程。

邓小清，女，1982 年生，硕士，助教，主要研究方向：计算机网络。