· 科技查新与文献检索

基于 Lucene 的医学文献检索系统

李 焱,路 莹

[摘要]文章在介绍了 Lucene 的基本功能和特点,分析了其与通用数据库各自的优势与不足,结合医学文献检索系统的应用,设计了以 Lucene 为底层检索接口,与数据库相结合的体系结构,并就其中的关键技术,特别是汉语分词技术的优缺点进行了讨论。

[关键词]全文检索;Lucene;检索系统;分词

[中图分类号]R-5;G252.7 [文献标志码]A [文章编号]1671-3982(2010)09-0052-03

Lucence-based medical literature retrieval system

LI Yan, LU Ying

(Medical Library of Chinese PLA, Beijing 100039, China)

[**Abstract**] After a brief description of the basic functions and characteristics of Lucence, the advantages and disadvantages of Lucence and other general databases were analyzed. A Lucence-based medical literature retrieval system was designed with Lucence as its bottom layer retrieval interface combined with databases concerning the application of medical literature retrieval system. The key techniques used in developing this system, especially the advantages and disadvantages of Chinese standard analyzer techniques, were discussed.

[**Key words**] full-text retrieval; Lucence; retrieval system; word division

1 Lucene 简介

Lucene 是目前已经被广泛应用于全文检索的项目。值得注意的是,Lucene 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。它可以方便地引入项目中,在目标系统中实现完整的全文检索功能[1]。Lucene 最初是 Apache 软件基金会 Jakarta 项目组的一个子项目,提供了一个开放源代码的全文检索引擎工具包,最早由 Java 语言编写,目前已推出许多其他语言的版本,如 PHP、.NET 等。

Lucene 具有如下突出优点[2]。第一,索引文件格式独立于应用平台。Lucene 定义了一套以 8 位字节为基础的索引文件格式,使得兼容系统或者不同平台的应用能够共享建立的索引文件。第二,在传统倒排索引的基础上,实现了分块索引,能够针对新的文件建立小文件索引,提升索引速度。然后通过与原有索引的合并,达到优化的目的。第三,面向对象的优秀系统架构,降低了 Lucene 扩展的学习难度,方便扩充新功能。

笔者以 Lucene. Net 为例,对其结构模块进行介绍。Lucene. Net 提供了十分全面的索引创建、分析、查询等模块, 各模块的功能如表 1 所示。

Lucene 功能强大。但从根本上说,一是对需要索引的内容进行分词后建立索引文件;二是查询功能,即对索引进行检索,选出符合条件的记录。相关的 Lucene 功能库主要有分词、索引管理和检索管理,对应的程序集为 Lucene. Net. Analysis、Lucene. Net. Index、Lucene. Net. Search。由于代码是开源的,也可以对其功能进行扩展,开发适用的搜索引擎。

2 Lucene 的特点和系统设计

针对网络医学文献资源,采用 Lucene 和数据库相结合的方法,可以开发高效的文献检索系统。

[基金项目]国家科技支撑计划项目保密项目。

[作者单位]解放军医学图书馆,北京 100039

[作者简介]李 焱(1968-),男,北京市人,硕士,助理研究员,发表论文 8 篇,获军队科技进步奖 6 项。

2.1 Lucene 和数据库的特点

Lucene 和数据库虽然都可以通过索引对数据进行检索，但普通的数据库并非为全文检索而设计。如果采用普通的数据库，如 SQL Server 进行全文检索，通过 like 进行模糊检索，它对于系统的损耗相当大。除非有更好的硬件支撑，否则无法支持更多的用户。在索引机制的建立上，Lucene 与数据库有很大的不同，Lucene 对传统的倒排索引进行了改进，实现了分块索引[3]。它不是维护一个索引文件，而是在扩展索引时不断创建新的索引文件，然后定期把这些新的小索引文件合并到原先的大索引中，通过与原有索引的合并，达到优化的目的。合并参数可以设定，对批次的大小进行调整，实现不同的策略。可以说，Lucene 最核心的特征是通过特殊的索引结构实现了传统数据库不擅长的全文索引机制，并提供了扩展接口，以方便针对不同应用的定制。

Lucene 内置了根据相关度排序的功能，其排序的参数可以动态指定。如我们可以对各个网站进行评分，把从专业网站搜索来的文章排到前面，从而提高检出文献的用户满意度，这是普通数据库所不具备的。此外，Lucene 可以通过 setSlop( ) 设置一个称为“坡度”的变量，以确定关键字之间是否允许和允许多少个无关词汇的出现。如当坡度为 2 时，检索“张军”时，可以同时检索出“张军”、“张海军”。这对于人名、地名的检索有特殊意义，可以提高检全率。

数据库在事物管理、数据存储、数据安全、用户管理等方面有成熟可靠的机制，而这些是 Lucene 所不具备的。如果把所有数据都保存在 Lucene 中，会带来索引膨胀过快问题。

2.2 系统结构设计

将 Lucene 和数据库结合起来，借助 Lucene 强大的全文检索功能，一方面可以减轻对数据库的压力，同时支持更多的用户；另一方面，发挥数据库在数据存储和管理上的优势，以弥补 Lucene 在这方面的不足。通过优势互补，可以使系统开发的效果更佳。系统结构如图 1 所示。

网上存在大量免费医学文献，对其加以利用可有效弥补图书馆经费的不足。网络爬虫根据一定的策略从网上采集信息，并对各种文档加以筛选和整理，将符合要求的数据存储于数据库中。索引模块则将要进行检索的数据项，如文章的题目、作者、文摘、关键词等取出，调用 Lucene 的索引创建模块，建立 Lucene 索引文件。索引建立时，将数据库中的记录 ID 号一并保存于 Lucene 的索引文件中，通过记录 ID 号，实现 Lucene 和数据库中记录的关联。

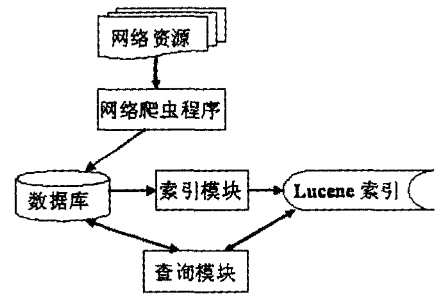

图 1　系统结构图

检索时先使用 Lucene 的查询接口在其索引文件中检出符合条件的记录 ID 号，然后通过 ID 号把记录的详细信息从数据库中调出。这样既可获得较快的查询速度，又不会因为把所有信息都存储于 Lucene 的文件中而造成索引文件膨胀过快的问题。

3.3 检索界面的设计

方便用户检索是每一个查询系统需要考虑的问题。因此，我们把传统的文献检索和现今流行的网络搜索引擎相结合，设计简单易用的检索界面（如图 2 所示），符合文献检索的要求和用户使用习惯。

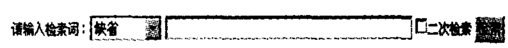

图 2　用户查询输入简图

在缺省情况下，与搜索引擎的检索方式完全一致。当用户输入“现代医学”进行检索时，可以不必确定是在题目还是在期刊名或是在文摘中查询，系统会自动在所有的检索字段，如题名、作者、文摘等多个字段中进行查询，并把查询的结果进行合并、过滤后，返回给用户。同时，以空格表示“或”操作，如输入为“现代 医学”则查出包含“现代”或“医学”的文献。这与网上搜索引擎的使用完全一致。

当用户勾选“二次检索”时，则在上一次检索的结果中进行再次检索，从而逐步得到用户想要的精确结果。当用户明确其检索的内容对应的检索项时，可以从前面的下拉框中选择要检索的字段。

这样一个简单的查询界面，可以满足用户的各种检索需求。

3.4 检索功能的实现

检索功能是文献检索的中心环节。使用 Lucne

提供的检索接口，充分发挥 Lucene 全文检索的优势，是检索设计和实现的最大难点。检索模块的数据流图 3 所示。

图 3 检索模块数据流图

QueryParsers 和 Search 是 Lucene 提供的检索接口。用户输入要查询的关键词后，系统对关键词进行切分，根据 QueryParser 的语法，调用其 API 设计具体的检索器，包括默认域、索引库位置的指定，以及将关键词通过布尔逻辑运算符连接起来形成复杂的查询语句。随后将正确解析的检索表达式传递给 Search，系统通过检索器对索引文件执行查询操作，然后进行去重、合并检索结果集，排序，最后将检索结果集提交给用户。

4 分词系统的比较

分词是全文检索的前提和核心。Lucene 中分词的工作由 Analyzer 的扩展类来实现。Lucene 自带了 StandardAnalyzer 类，可以支持中文，我们也可参照该类的实现写出自己的切分词分析器。

英文各词之间有空格分隔，分词不是大问题。而汉字之间没有分隔符，词与词之间的关系完全靠上下文判断。一个词可能就是一个字，也可能由两个或多个字组成。汉语分词技术虽有很大的发展，但仍有很多有待克服的问题，如专有名词与复合词的切分，同形异义字的区分等。

如何在切分紧密相连的词时最大限度地保持其原意，Lucene 大体提供两类方法。第一，通过词表法进行切分，即根据语言的习惯，建立一个完备的词库，依据词库中的词对文本进行切分。其优点是关键词清晰，索引效率高，数据膨胀率较低。但词表的维护成本较高，适用于需要复杂检索规则、较多高级检索要求的大型特殊系统。第二，通过算法进行切分。Lucene 采用一元切分或二元切分，即以单字或二个汉字为一组进行切分。也可以采用三元切分或多元切分，但其精细度不如一元或二元切分。采用此方法不需要对词表进行维护，成本较低，适合于一般的中小型系统。

本系统采用单汉字切分的一元切分算法，主要是考虑其用于全文检索具有如下优点。一是单字的组配非常灵活，任何新词都可以通过字的组配获得，这是一般词典法所不及的。单汉字标引全文检索又被称为“无标引检索”。由于无须建立词典，打破了不同学科领域词典的分割，用一个单字索引库即可快速完成全文检索，适用的学科领域比较宽广。二是采用单汉字索引的检索命中率较高这也许是最重要的，因为准确性、相关性都是以命中率为前提的。这方面比较成功的著名网站如“百度”网，在查询中有单字分词的明显特点。三是单汉字分词相比二元分词，实现容易，索引效率较高，并且其数据的膨胀率较低，索引文件为原文件的 50% 左右。而二元分词会造成很大的冗余，切出很多无意义的词，索引文件膨胀率较大[4]。

5 结束语

Lucene 是一款优秀的开源软件，适用于各种需要全文检索的系统，许多优秀的商业软件也采用其进行系统搜索。同时 Lucene 丰富的 API 接口和开源特性，为程序的扩展提供了广大的空间，极大地推动了全文检索技术在各行业或领域中的应用。

【参考文献】

[1] 赵 汀，孟祥武. 基于 LUCENE API 的中文全文数据库设计与实现[J]. 计算机工程与应用，2003，39(20)：179-181.

[2] 开放源代码的全文检索引擎 Lucene：介绍、系统结构与源码实现分析[EB/OL]. [2009-11-30]. http://www.lucene.com.cn/about.htm.

[3] Lucene 倒排索引原理[EB/OL]. [2009-11-30]. http://www.lucene.com.cn/yanli.htm.

[4] 周 祥，王丽芳，蒋泽军，等. 基于 Lucene 的企业信息门户搜索引擎设计[J]. 微处理机，2009，30(4)：62-64.

[收稿日期：2010-01-21]
[本文编辑：吕 婷]

(上接第 44 页)

【参考文献】

[1] 国际图书馆协会联合会. ISBD(M)专著出版物国际标准书目著录. 第 2 版[M]. 北京：书目文献出版社，1989.

[2] 中国图书馆学会.《西文文献著录条例》修订组. 西文文献著录条例(扩大修订版)[M]. 北京：科学技术文献出版社，2003.

[3] 王 斌. 基于 USMARC 格式著录西文授权重印版图书[J]. 图书馆杂志，2006，25(12)：40-42.

[4] 张佩仪. 授权影印版西文图书编目探讨[J]. 图书馆论坛，2006，26(2)：169-172.

[5] 国家图书馆 MARC21 格式使用手册课题组. MARC21 书目数据格式使用手册[M]. 北京：北京图书馆出版社，2005.

[收稿日期：2010-01-27]
[本文编辑：杜海洲]