基于 Lucene 的 MYSearch 全文搜索引擎

刘亚姝,黄 岳
(北京建筑工程学院 电信学院计算机系,北京 100044)

摘 要:基于 Lucene 开源框架设计实现了 MYSearch 全文搜索引擎。给出了 MYSearch 实现的基本原理和设计流程,以及实验结果,并针对 Lucene 在中文分词方面的不足展开了讨论,给出了改进方法。

关键词:全文搜索引擎;Lucence;分词;索引
中图分类号:TP391.72;TP393.03 **文献标识码**:A **文章编号**:1674-7720(2010)18-0086-03

MYSearch full text search engine based on Lucene

LIU Ya Shu, HUANG Yue
(Beijing University of Civil Engineering and Architecture, Beijing 100044, China)

Abstract: Reallizes MYSearch full text search engine based on Lucene and gives the principle and process of designing MYSearch and the results, then takes into accout the shortcoming of Lucene on the aspect of Chinese word segmentation, and gives an improved method.

Key words: full text search engine; Lucence; Chinese word segmentation; index

随着互联网的飞速发展,网络资源量也在迅速增加。最大的搜索引擎 Google 从 2002 年的 10 亿网页增加到现在近 40 亿网页;最近雅虎搜索引擎号称收录了 45 亿个网页;国内的中文搜索引擎百度的中文页面从两年前的 7 000 万页增加到了现在的 2 亿多。据估计,当前整个互联网的网页数达到 100 多亿,而且还在快速增长。用户要在如此浩瀚的信息海洋里寻找信息,犹如“大海捞针”,往往无功而返[1]。如何从资源的海洋里找到自己需要的内容就成了关键问题,搜索引擎的出现和研究,使网络上的资源变得有序,使用户能更加方便快捷地找到所需资源。目前被大家广泛使用的搜索引擎如 Google、百度等,其实现技术非常复杂,后台数据库也非常庞大,更新速度也很快。然而,若想搭建一个全文搜索引擎也并不是遥不可及的事情。本文主要针对 Lucene 介绍全文搜索引擎实现的各项技术,并给出改进方法。

1 Lucene 基本技术原理

目前网络上有许多全文搜索引擎的开源代码,若想构建自己的全文搜索引擎,可以在这些开源代码的基础上进行。其中,Lucene 是比较突出的一款。Lucene 不是一个完整的全文索引应用,而是一个用 Java 写的全文索引引擎工具包,因此它并不像“百度”或者“Google”那样,可以直接作为查询工具使用,而只是为全文搜索引擎的构建提供了基本的工具和设计方法。Lucene 提供了一系列 API,能够对文档进行预处理、过滤、分析、索引和检索排序。本文就是在 Lucene 基础上构建了一个全文搜索引擎 MYSearch。

2 MYSearch 工作流程

2.1 搜索引擎的基本构成

搜索引擎系统一般由蜘蛛(也叫网页爬行器)、切词器、索引器、查询器几部分组成。蜘蛛负责网页信息的抓取工作;一般情况下切词器和索引器一起使用,它们负责将抓取的网页内容进行切词处理并自动进行标引,建立索引数据库;查询器根据用户查询条件检索索引数据库并对检索结果进行排序和集合运算,再提取网页简单摘要信息反馈给查询用户。

2.2 MYSearch 工作流程

MYSearch 首先使用网络蜘蛛抓取网络上的可用网页链接,然后把抓取到的网页资源下载到本地计算机,对下载到本地计算机的网页进行初步的处理,去掉对搜索没有意义的信息和词汇。然后使用 Lucene 提供的索引功能,对处理后的信息资源建立索引,并且保存到索引数据库中。之后,根据用户提供的搜索信息,在索引中

进行查询，并将搜索结果显示到用户搜索的界面上。其流程框图如图 1 所示。

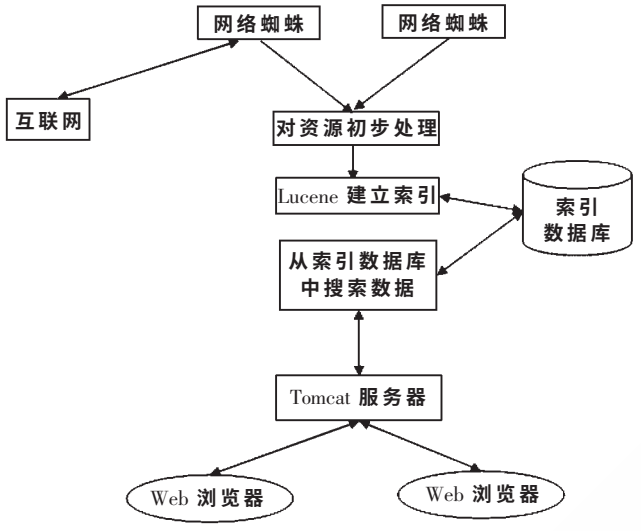


图 1 系统工作流程图

3 MYSearch 实现

3.1 系统功能模块的划分

MYSearch 全文搜索系统主要分为网络蜘蛛抓取、资源初步处理、建立索引、搜索以及显示等功能模块。

(1)网络蜘蛛抓取功能模块：首先根据事先设定好的网络入口地址和设置的搜索条件，读取网页的内容，分析网页中其他的链接地址，然后垂直链接到下一个网页，这样一直循环，直到网站的所有网页都抓取完成或者满足了搜索的条件为止。

(2)资源初步处理功能模块：将搜索来的网页中的信息进行相关处理，去掉没有用的格式内容和其他对搜索结果没有实际意义的信息。

(3)建立索引功能模块：将处理后的网页资源写入数据库，并使用倒排索引算法实现网页资源索引的建立。

(4)搜索功能模块：根据用户的搜索关键词，在已建好索引的数据库中，根据语素向量的匹配度和相似度进行相关的匹配，然后按照一定的排列顺序把搜索结果返回给用户。

(5)显示功能模块：将搜索结果按照一定的显示方式显示在页面中，供用户选择和浏览。

3.2 MYSearch 全文搜索引擎的实现

3.2.1 网络蜘蛛

网络蜘蛛是指某个能以人类无法达到的速度不断重复执行某项任务的自动程序[1]。本系统中使用的蜘蛛程序是 Nutch，核心是 Crawl 工具。它可以根据之前设定好的入口 URL 列表不断地自动下载页面，直到满足系统预设的停止条件。图 2 所示是 Nutch 的工作机制。

3.2.2 网页初步处理

网页刚刚被抓取下来的时候，存在很多格式化的信息(如 html 的网页标记)，还有很多多余的信息(比如“is，the，an”)。这些信息都是噪音，如果想要使搜索引擎更高效、更准确地运行，就要去除这些信息，留下有效的信息。

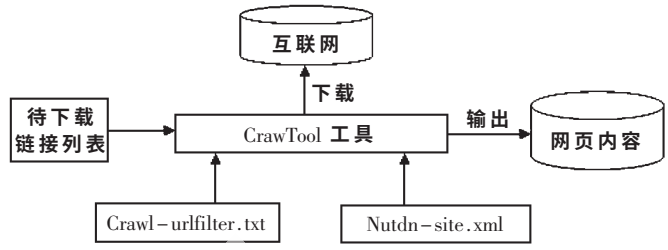


图 2 Nutch 工作机制

对于 html 标记的处理，首先就是准备一个空字符串，然后判断网页的文字中是否存在 html 的“<>”符号，如果是 html“<>”的符号，就继续判断网页中的下一个字符，如果不是就把该字符保存到这个空字符串中；如果判断完成，就结束；否则就继续判断。对于多余信息，在 Lucene 中提供了相关的包进行处理。

通过上面的处理之后，下载的文件在建立索引的时候，就会更加便捷。

3.2.3 索引的建立

在日常的生活中，往往需要快速地从海量页面信息中定位页面资源。这样的需求就需要用索引技术来实现。索引建立的好坏直接影响搜索效果和用户的体验感觉，所以索引的建立方法十分重要。Lucene 采用倒排索引算法建立索引[2]，主要包括索引类(IndexWriter)、文档对象类(Document)和信息字段对象类(Field)。索引建立的过程为：

(1)建立索引器 IndexWriter；

(2)建立文档对象 Document；

(3)建立信息字段对象 Field；

(4)将 Field 添加到 Document；

(5)将 Document 添加到 IndexWriter 里面；

(6)关闭索引器 IndexWriter。

Lucene 将建好的索引信息存储在“_0.cfs”、“segments.gen”以及“segments_s”文件中。

3.2.4 信息搜索

用户提交的查询请求通常是一个词语或者短语，MYSearch 搜索引擎在接受用户访问后会进行一系列处理并最终向用户提交。当用户输入关键词搜索后，由搜索程序从索引数据库中找到符合该关键词的所有相关文档。因为所有文档针对该关键词的相关度早已算好，所以只需按照现成的相关度数值排序。排序规则是相关度越高，排名就越靠前。然后，就会把查询到的信息返回给用户，并进行显示。基本查询流程如图 3 所示。

3.2.5 搜索结果显示

良好的交互设计可以使用户的操作更加简便，可以使用户能够更快更准确地找到自己想要的信息，同时能

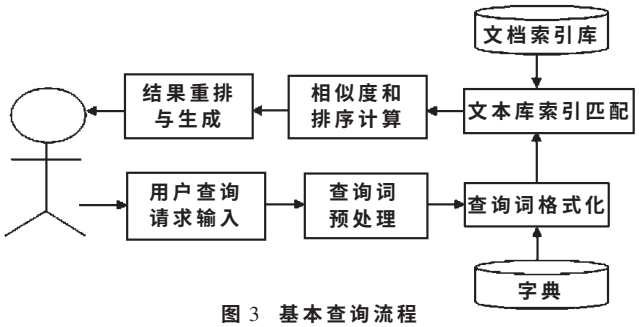


图 3 基本查询流程

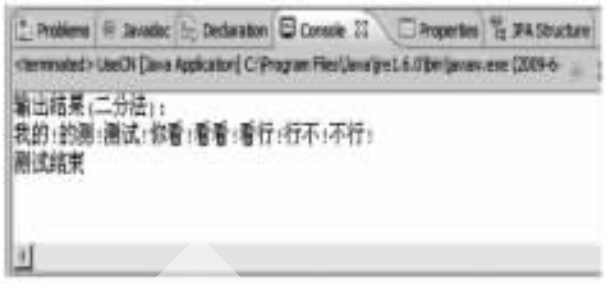
够增加用户的满意度。MYSearch 全文搜索引擎设计了一个简捷的搜索界面，用户在该界面中输入搜索条件，提交后就可以看到查询结果。

4 改进

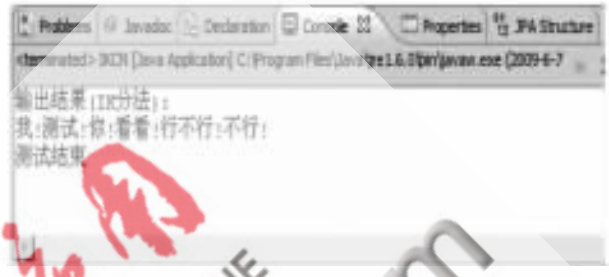
在一个搜索引擎中，搜索速度的快慢、搜索效率的高低、搜索准确度的高低，很大程度上取决于分词的优劣。分词就是为生成索引提供原材料，如果分词分得不明确，则生成的索引必然复杂，那些没有实际意义的分词被称为噪音，噪音多了搜索速度必然下降。Lucene 其实自身是带有中文分词功能的，主要采用“单字切分”和“二分法”，但是由于它没有做到确定最小索引项，因此无法去除噪音，搜索效率大大降低。

为了改进 Lucene 的中文分词的缺陷，MYSearch 全文搜索引擎采用了 IK_Canalyzer 中文分析器。IK_Canalyzer 中文分析器实现了以词典为基础的正反向全切分和一级正反向最大匹配切分两种方法。IK_Canalyzer 中文分析器是第三方实现的分析器，继承自 Lucene 的 Analyzer 类。图 4(a)和图 4(b)分别为采用 Lucene 与 IK_Canalyer 分词的显示结果，可明显看出后者优于前者。

MYSearch 是基于 Lucene 设计实现的一个全文搜索引擎，本文给出了设计过程以及实验结果，并针对 Lucene 在中文分词方面的不足给出了解决办法。此外目前可以获得的 Lucene 开源代码中并没有对 PDF、Word、Excel 等常用的文本格式进行搜索。要想克服上述问题，就要对不同格式的文本进行解析，把解析出来的文字提取出纯文本，然后就像建立网页的索引一样，对提出来的文字建立索引，以便查询。这将是进一步需要改进 MYSearch 全文搜索引擎的工作重点。



(a)Lucene 分词



(b)IK_Canalyer 分词

图 4 Lucene 与 IK_Canalyer 分词的比较

参考文献

[1] 吴卓斌.基于 Lucene 全文搜索引擎关键技术的研究[D].广州：暨南大学，2007.

[2] 隋丽萍，徐承韬.一个中文全文检索系统的设计与实现[J].科技资讯，2007，5(18)：244-245.

[3] 张校乾，金玉玲，侯丽波.一种基于 Lucene 检索引擎的全文数据库的研究与实现[J].现代图书情报技术，2005，21(2)：12-14

(收稿日期：2010-04-26)

作者简介：

刘亚姝，女，1977年生，讲师，硕士，主要研究方向：CSCW、网络数据库、CAD。