基于Lucene 的搜索引擎设计与实现

0 前言

在过去几年里, Internet 的资源迅速增长, 使Web 发展成为包含多种信息资源、站点遍布全球的海量信息服务网络。同时, 也有越来越多的机构、团体和个人在Internet用搜索引擎查询信息。作为一个门户网站来说, 提供给用户搜索服务, 是吸引用户访问网站的重要手段。目前许多网站建立搜索引擎服务通常通过调用诸如Google 或百度的搜索服务, 来实现对本站点的搜索。但是对于一个有很多子网站的企业门户网站来说, 通用搜索引擎存在着很多缺陷, 满足不了这种搜索服务要求, 如: 尽管Google 等搜索引擎提供对指定站点内的查询, 但是不能同时对多个站点同时查询; 通用搜索引擎不能及时更新索引, 会导致搜索结果不全和出现?? 坏链接??; 调用通用搜索引擎的响应速度慢。因此研究一个由企业自主定制的搜索引擎, 具有重要的意义。文中采用Lucene 的开发工具包, 实现了一个全文搜索引擎。

1 搜索引擎的结构描述

通常, 一个搜索引擎由搜索器、索引器、检索器和用户接口等四个部分组成。

a. 搜索器的功能是在互联网中漫游, 发现和搜集信息。它常常是一个计算机程序, 日夜不停地运行。它要尽可能多、尽可能快地搜集各种类型的新信息, 同时因为互联网上的信息更新很快, 所以还要定期更新已经搜集过的旧信息, 以避免死连接和无效连接。

b. 索引器的功能是理解搜索器所搜索的信息, 从中抽取出索引项, 用于表示文档以及生成文档库的索引表。

c. 检索器的功能是根据用户的查询在索引库中快速检出文档, 进行文档与查询的相关度评价, 对将要输出的结果进行排序, 并实现某种用户相关性反馈机制。

d. 用户接口的作用是输入用户查询、显示查询结果、提供用户相关性反馈机制。主要的目的是方便用户使用搜索引擎, 高效率、多方式地从搜索引擎中得到有效、及时的信息。

2 Lucene 简介

Lucene[2, 3] 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, 它提供了多个API 函数与灵活的数据存储结构(可以定制) , 可以方便地嵌入到各种应用中实现针对应用的全文索引/ 检索功能。它是APACHE 基金会jakarta 的一个子项目。

2. 1 Lucene 的全文索引存储文件结构

Lucene 存放索引信息的是文件。它的索引存储文件设计的比较通用, 输入输出结构都很像数据库的表→记录→字段, 所以很多传统应用的文件、数据库等都可以比较方便地映射到Lucene 的存储结构/ 接口中。它的索引存储文件结构描述如下:

1) 基本定义。

Lucene 的基本概念为: 索引( Index) 、文档( Document) 、字段( Field) 、术语( term) 。索引由一个文档序列组成; 文档由一个个字段序列组成; 字段由一个命名的术语序列组成; 术语由字符串构成。

2) 字段类型。

Lucene 中的字段有三个重要的属性:

a. 可分词性: 表示字段内容被分成多个可被索引的词( token) 。每个词是一个术语。当某个字段不分词时, 整个字段内容作为一个术语。

b. 可存储性: 表示字段内容是直接按字词存放, 而不是以到排形式存放。

c. 可索引性: 表示字段内容以到排形式存放, 即: 记录了字段的每个术语在某一文档的出现频率。这三个属性不具有互斥性, 字段可以同时被到排、索引和分词。

3) 段。

Lucene 索引可以由多个子索引( 即: 段) 组成。每段是个完全独立的索引, 可以被独立搜索。索引可以通过两种方式演化: 为新增加的文档创建新的段; 合并已存在的段。每个段索引包含有: 字段名( Filed names) : 在索引中所用到的字段名的集合; 存储字段值: 存储字段值的集合( 存储字段值是文档的一些辅助存储属性, 如: title, url) ;术语字典( Term Dictionary) : 存放用在所有文档中的所有索引字段的术语, 也包含了拥有该术语的文档数, 和指向该术语频率数据与位置数据的指针; 术语的频率数据: 包含每个术语的文档数和该术语在文档中的频率; 术语位置数据: 对在字典中的每个术语, 记录了术语在每个文档中的位置; 标准化因子: 对每个文档的每个字段, 该值与它的得分值相乘得到最后的分值; 删除表: 记录了已被逻辑删除但没有物理删除的文档。段都存放在一个索引目录下。

2. 2 Lucene 的组成结构

Lucene 有七个程序包构成:

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 / : 用JavaCC[ 4] 实现的查询分析器。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。对于外部应用来说, 索引模块( index) 和检索模块( search) 是主要的外部应用入口。

2. 3 优点与缺点

由于Lucene 存放索引信息的不是一般数据库, 而是文件, 这使得Lucene 访问索引的时间快, 同时也使得Lucene 可以跨平台使用。Lucene 与大部分的搜索( 数据库) 引擎不同, 不是采用导致索引的更新会需要大量IO 操作的B 树结构来维护索引, 而是在扩展索引的时候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( 针对不同的更新策略, 批次的大小可以调整, 策略可以定制) 。这样在不影响检索的效率的前提下, 提高了索引的效率。

Lucene 工具包中大量地使用Strategy 设计模式, 使得应用接口设计灵活。用户可以利用这些接口, 定制出适合自己需要的语言分析器、查询分析器, 甚至检索器。

另外, Lucene 能够支持多用户的使用。但是它不支持中文分词, 且不提供在Internet 上抓网页与文件内容分析功能。

3 系统的设计与实现

在湖南铁通IDC 数据平台中, 采用Lucene 为核心开发了基于公司门户网站以及所有公司内子网站的搜索引擎[ 5] 。整个搜索引擎由七大模块生成: 系统维护接口, 文件内容分析器, 搜索器, 索引器, 检索器, 查询分析器, 用户界面( 见图1) 。

1) 系统维护接口: 本系统对索引内容的更新采用定时器触发后台启动与手工启动两种。每个子站点建立一个索引目录, 指定不同的索引目录的路径。这不仅可以灵活地放置索引数据, 而且也可以使得对湖南铁通下的每个子站点单独进行更新操作和访问操作。系统对每个站点都提供两种搜索策略与索引策略。搜索策略是宽度优先搜索与深度优先搜索两种。索引策略是批量索引与增量索引两种。批量索引是建立一个空索引, 然后将被索引的文件以Document 格式放在内存中, 当索引文件达到参数mergeFactor 时, 索引器一次将其写入索引; 增量索引是在原有索引的基础上删除增加被索引文件。每个站点这种的搜索与索引策略及其相关参数都存在. xml 的配置文件中, 可由系统维护人员通过该接口进行修改。

2) 文件内容分析器: 分析HTML, PDF 等多种格式文件, 从中提取链接和文件各字段内容。文件的字段由开发人员定义, 这里定义了url, contentT ype( 内容类型)、last Modified( 最后修改日期) 、contents ( 内容) 、title( 标题) 、keyWord( 关键字) 等字段。

3) 搜索器: 根据系统维护人员的定义的根URL, 按宽度优先搜索原则在Internet 抓取网页, 调用文件内容分析器分析文件内容, 将其转换为Document 格式输出到索引器。每个Document 对象包含多个字段Field 对象, 针对不同的字段属性和数据输出的需求, 对字段还可以选择不同的索引/ 存储字段规则, 如表1 所示。

搜索器对每个站点内容的更新是由定时器控制的。它采用配置文件中的搜索策略多线程地对站点进行搜索,抓取网页。

4) 索引器: 使用的是Lucene 的索引器IndexWrite 类。修改该索引器所使用的语言分析器, 使其支持中文分词。这里采用了简单的单汉字切分, 即将单个汉字作为一个术语, 利用JavaCC 编程实现[ 4, 6] 。并且在生成索引器将支持中文单汉字切分的语言分析器作为参数传递给索引器成为索引器的成员, 使其在索引过程中使用该语言分析器进行分词。索引器采用的索引策略来自于. xml 配置文件。批量索引器会删除原来的索引段, 建立新的索引段,然后将所有的被索引文件加入到索引段中。而增量索引器则是读取原索引段的内容, 调用SegementReader 类的delete( term) 方法将包含某ur l 地址的被索引文件从索引中删除, 同时将被删除文件的新内容( 或新文件) 加入索引段中。无论批量索引器还是增量索引器, 最后都是调用IndexWrite 类的addDocument( ) , 将搜索器抓取且转换为Document 格式的文件, 根据字段类型进行分别处理, 加入到索引段中。

5) 查询分析器: 使用的是Lucene 核心中的QueryParser 类, 对于用户提交的查询关键字组成的逻辑表达式进行分析, 生成系统内部的查询条件类Query。

6) 检索器: 调用Lucene 的搜索入口对索引进行查询,返回查询结果。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。可以通过它再访问Document= > Field 中的内容。本系统允许用户对指定站点范围查询。对要查询的每个子站点的索引目录定义一个子检索器( IndexSearcher 类) , 然后将其作为参数输出到Lucene 的MultiSearcher 类中, 调用MultiSearcher 类的search 方法对多个索引目录进行检索。Mult isearch 的search 方法不仅能够调用子检索器检索索引, 而且能将所有子检索器检索到的结果合并重新排名。

7) 用户界面: 输入用户查询、调用检索器, 将检索器返回的查询结果( 多个Hits) 分页显示在页面。

图2 是该搜索引擎实际运行图。

4 搜索引擎的性能分析

为了测试基于Lucene 开发的搜索引擎的性能, 向Google 提交了六个查询关键字, 并且通过对Google 高级查询中的选项进行配置使得搜索范围是在湖南铁通站点内。同时对湖南铁通搜索引擎提交相同的关键字。最后将两者的搜索结果进行比较, 如表2 所示。

从表1 来看, 总体上湖南铁通搜索引擎查询的结果比Google 的查询结果更全。这是因为在湖南铁通搜索引擎的结果中出现了很多实时更新的网页( 如: 新闻网页等) ,同时还包括了其他网站的结果( 主要为地方铁通网站, 如:湖南株洲) , 而这些结果在Google 中没有反映。但是有些查询结果在Google 中出现, 而在湖南铁通搜索引擎中没有出现。笔者分析一下, 发现是没有对动态页面做索引,这是未来要改进的一点。整个铁通搜索引擎的查询响应速度快, 响应时间在4 秒以内。

5 结论

对于一个拥有很多子网站的企业门户网站来说, 为满足对所有子网站内容搜索而定制一个适合需要的搜索引擎是克服调用通用搜索引擎而引起的响应速度慢、索引范围不全的缺点的最佳方法。Lucene 是一个基于Java的全文检索工具包。它强大的全文索引引擎工具包使人们能够利用它快速地开发一个全文检索系统/ 搜索引擎。同时它的可定制性使得人们可以灵活嵌入到各种应用中, 不断地扩展其功能。利用Lucene, 开发了基于公司门户网站以及所有公司内子网站的湖南铁通IDC 数据平台中的搜索引擎——一个由企业自主定制的搜索引擎实例。未来, 将以此为基础,对搜索引擎增加对动态页面的索引, 增加语义分析, 提高搜索精度。