手机产品垂直搜索引擎的研究与实现

苏永红 张玉蓉

(武汉理工大学华夏学院 武汉 430223)

摘 要 随着网络技术的快速发展,通用搜索引擎已经不能满足用户的一些需求,特别是当用户需要搜索某一领域内的信息时,垂直搜索引擎就正好符合这种需求。以手机资源为背景,通过运用扩展 Heritrix 和 Lucene,构建了一个检索结果比较精准的垂直搜索引擎。研究了通过定制和扩展 Heritrix 从互联网上爬取相关的信息资源,利用 HtmlParser 工具对爬取的信息进行分析和抽取,运用 Lucene 建立全文索引和提供检索服务,并设计了 MVC 的查询接口。通过响应时间、查全率和查准率的测试实验表明,系统达到了设计目标。

关键词 垂直搜索,Heritrix,抽取,索引

中图法分类号 TP391 **文献标识码** A

Research and Implementation of Mobile Phone Vertical Search Engine

SU Yong-hong ZHANG Yu-rong

(Wuhan University of Technology Huaxia College, Wuhan 430223, China)

Abstract With the fast development of network technology, universal search engine always can not meet many user demands, especially when user needs to search some information in a field, vertical search engine accords with user demands. Cell phone resource search was discussed. It initially comes up with a vertical search with fairly precise outcome through expanding the use of Heritrix and Lucene. The major research work of this paper is divided into four parts. Firstly, by customizing and extending the Heritrix, it crawled some information from Internet. Secondly, the crawled information was analyzed and cramped out, some of that with the tool of HtmlParser. Thirdly, Lucene used to build a full-text index and retrieval service for the system. Finally, the system design a MVC connector. The system achieves design goals through the tests of response time, recall ratio and precision ratio.

Keywords Vertical search, Heritrix, Extraction, Index

搜索引擎的产生是为了让用户能够方便地从网络空间获得其所需要的信息,它根据一定的策略,运用特定的计算机程序搜集互联网上的信息,在对信息进行组织和处理后,将处理后的信息显示给用户,其是为用户提供检索服务的系统。但是随着网络的快速发展,通用搜索引擎已经不能满足用户的一些需求,特别是当用户需要搜索某一领域内的信息时,垂直搜索引擎就正好符合这种需求,它是针对于某一领域的专业搜索引擎,它将跟该领域相关的网页下载之后,从中提取出用户需要的信息,经过进一步的处理之后再呈现给用户。Leixian 等人研究了一种利用用户兴趣模型来优化主题爬虫的性能的系统[1],该系统利用蜘蛛爬行器分析计算下载的目标与对应主题的相关度,并确定哪些 URL 链接指向相关领域的页面;Yubo jia 等设计了一种垂直搜索引擎[2],该应用系统通过优化算法来提高用户的搜索效率,并将使用优化算法的垂直搜索引擎与传统的搜索引擎比较,证明了该模型可以提高搜索效率;Chuan Wang 等提出了基于通用图形处理单元的垂直搜索引擎[3],这个系统通过并行计算减少 CPU 的访问来提高搜索效率。在国内,垂直搜索引擎正处在一个蓬勃发展的时期,各种专业搜索引擎层出不穷,如以百度地图为代表的地图搜索(http://map.baidu.com/)、以 360 问答为代表的论坛搜索(http://wenda.so.com/)、以去哪儿为代表的旅游搜索(http://www.qunar.com)、以百度招聘搜索为代表的招聘搜索(http://opendata.baidu.com/zhaopin/),但以手机产品为主题的搜索引擎还有待研发,当今手机的更新换代比较频繁,如果开发一个可以对手机产品信息进行快速搜索的搜索引擎,将给人们的生活提供方便,本系统的目的就是给用户提供一个快速搜索手机产品信息的平台。

1 手机产品垂直搜索引擎系统设计

本系统是通过 Heritrix 框架和 Lucene 框架来构建一个手机垂直检索系统,本系统是基于 SSH(struts+Spring+Hiberate)框架在 Eclipse 集成开发环境下设计的,可以检索出手机产品的所有相关信息[4,5]。本系统的系统结构图如图 1 所示。

本文受武汉理工大学华夏学院院级科研基金项目(11030)资助。

苏永红(1980—),女,硕士生,讲师,主要研究方向为信息检索、分布式系统,E-mail:461782640@qq.com;**张玉蓉**(1976—),女,硕士生,讲师,主要研究方向为多媒体通信技术。

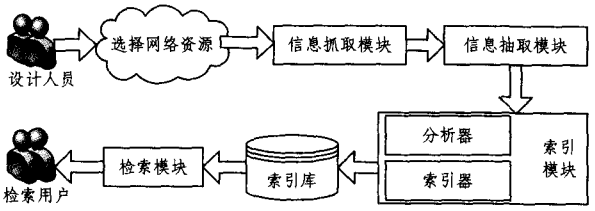

图 1　系统结构图

1.1　信息抓取模块

信息抓取模块主要是针对网络上的信息进行抓取，这一过程是通过网络爬虫来实现的。网络爬虫通过分析网页URL，并自动采集网页上的相关内容，然后在采集的过程中通过原始分析的网页 URL 来发现新的 URL，之后再对新的URL 内容进行采集，一直不断重复这一过程直到不会产生新的 URL 为止。而在这一采集过程中，垂直搜索引擎需要进行相应的优化，才能使得整个采集更有效率。

本项目使用网络爬虫 Heritrix 来进行信息采集，并对其功能进行相应的扩展，将基于手机知识库的信息采集策略扩展进 Heritrix，实现了只抓取相关领域的信息。

1.2　信息抽取模块

信息抽取模块主要是对下载下来的网页 URL 的信息进行分析和抽取，提取出用户需要的信息，并将其形成结构化的信息进行存储，以方便之后呈现给用户。

本项目通过分析抓取网页的页面结构，使用页面解析工具 HtmlParser 来对页面的信息进行抽取，最终提取出结构化的文本信息。

1.3　索引模块

索引模块主要由两个子模块组成，一个是分析器，分析器的作用是对结构化的文本信息进行分析处理，如去除违法字、特殊字符等，然后再对该文本信息进行分词，以便建立索引。另一个是索引器，索引器的作用就是用来为已经分析处理的信息建立索引。当信息量巨大时，通过索引来进行检索能提高系统的性能。

常用的信息存储方式是数据库和索引文件方式，数据库方式简单灵活，便于按照产品关键字检索和组织，但是不利于实现全文检索。索引文件方式速度快，能够组织海量的信息，但不利于格式化存储。本项目将结合两种方法，将所有的信息存储在数据库，而在索引文件中只存储需要检索的字段和跟数据库一一对应的字段，其中通过 Lucene 来建立索引，并通过引入 JE 分词工具来取代 Lucene 自带的分词工具。

1.4　检索模块

这一模块是系统与用户交互的模块，系统对用户输入的查询关联字进行分析处理，然后通过索引数据库搜索出相关的数据，去除掉不相关的数据，然后对结果按照一定的算法排序，并最终呈现给用户。

当用户在查询界面输入查询关键词时，后台配置文件会根据该关键词分发到对应的处理方法，然后通过对应的业务逻辑在索引库中检索关键词，当在索引库中查询到该关键词时，系统会到数据库中查询到对应产品的详细信息，最后以特定的排序方式返回给用户。系统由于采用 SSH 开发模型，不仅实现了 MVC 的分离，而且还实现了业务逻辑层与持久层的分离，大大提高了系统的可复用性和查询效率。下面分别介绍每个模块的设计与实现。

2　信息抓取模块设计

在本垂直搜索引擎系统中，展示的信息实际上是通过爬虫将这些商品信息抓取下来的。为了获取这些信息，需要构造针对手机产品的下载系统。垂直搜索引擎的爬虫结构与通用搜索引擎爬虫类似，但增加了链接限制和来源获取的部分内容[6]。

网络爬虫在进行任何抓取前，都需要对所抓取的内容进行详细的分析，了解网站内容的基本结构，以确定种子链接，即抓取的起始页。没有一个网络爬虫软件能够自动判断网站的哪些内容是需要抓取的，网站起始链分析的工作需要人工参与完成，这里所要开发的搜索引擎，并不是传统意义上的搜索引擎，而是要对所抓取的网站的内容进行深度分析和再加工，进而提取出有用信息的搜索引擎。系统中的网络爬虫利用改造的 Heritrix 来实现，完成的系统功能强大，支持特定类型网页的下载和业务管理。图 2 示出了垂直搜索引擎网络爬虫结构。

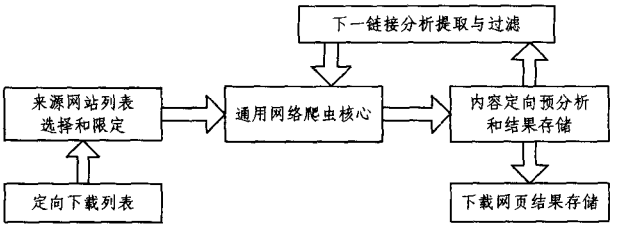

图 2　垂直搜索网络爬虫结构

2.1　网站内容与链接分析

垂直搜索引起的邮电在于对来源信息的精确定向抓取，如果采集信息的质量较高，则搜索的结果会更好，那么，究竟该怎样准备抓取清单？抓取起始清单准备的最重要的一个原则就是，在清单上的 URL 能够很方便地链接到所有具体的产品网页，选择网站内通常所称的目录或者频道页非常合适。本系统选择网易手机频道导航页和太平洋电脑网的手机品牌目录页。

以网易手机频道为例，通过分析网易手机频道的结构，网易手机频道的上部是一个手机品牌列表，有完整的手机品牌，当用户单击品牌时，就会切换到相应的品牌产品的页面，此时再单击相应手机链接，便会进入手机详细信息页面，所以将手机频道目录作为下载的起始页：http://product. mobile. 163. com/ 。

2.2　设计网易抓取的 Extractor 扩展

首先继承扩展了 Extractor 类，实现 extract 方法，实现的功能包括：分析下载到的网页，按照每一行分析内容。如果发现包含特定的标记，说明本行含手机品牌地址。并建新类文件 Mobil63Extractor，这个类是专门用来解析网易手机品牌的列表信息的，分析页面内容，该类将其中的字符信息按行读入，碰到包含特定标识的行，就认为找到一个品牌的 URL，然后对其截取，并在前部加入前缀 http://product. mobile. 163. com/，这样就构成了一个完整的 URL，然后将得到的完整 URL 加入到待处理队列中，以等待 FrontierScheduler 的处理。

2.3　设计网易抓取的 Frontier 扩展

除了手机品牌的入口列表，还需要对得到的列表首页进行进一步分析，得到具体手机型号产品页面链接。新建类文

件 FrontierSchedulerFor163Mobile，该类所定制的处理逻辑只允许下面的信息被加入到等待队列中：包含有“product. mobile. 163. com/product/”的 URL（产品详细信息）、包含有“product. mobile. 163. com/brand/”的 URL（产品手机品牌页）、所有图片、DNS 和 Robots 请求，同时，在连接中它还不允许带有“#”的 URL 通过，原因是此类链接往往是特殊内容或页面内部跳转，因此要拒绝这类 URL 被放入等待队列。

2.4 执行网易手机频道网页抓取任务

在完成上述网易手机频道的定制扩展后，还需要将这两个定制扩展类加入 processor. options 配置文件中，具体内容如下：my. extractor. Mobile163Extractor | Mobile163Extractor my. postprocesor. FrontierSchedulerFor163Mobile | FrontierSchedulerFor163Mobile，配置完成后就可以运行抓取网易手机频道的数据了。具体执行步骤为：第一步，在 Eclipse 中启动 Heritrix，打开浏览器输入 herix. properties 配置文件中设置的 admin 的用户名和密码，进入管理界面，选择 Job 连接，添加新的任务，在 seeds 项输入：http://product. mobile. 163. com/，选择 Modules，在 Extractors 项中加入 Mobile63Extractor 和 FrontierSchedulerFor163Mobile 两个类，将 Writer 项设置为 MirrorWriterProcessor，表示采用镜像方式写入，另外，为了提高抓取速度，在运行前，将 Java 虚拟机的内存设置成较大的值，如 1G，就可以增大 Heritrix 的最大可用内存，提高抓取速度。还可以通过减少睡眠线程的等待时间来提高抓取速度。接下来就可以开始抓取了，在管理员页面中选择控制台 Console，启动新添加的任务，控制台中打印出从 http://product. mobile. 163. com/页面中解析出来的手机品牌主页面 URL，说明解析成功，只需要等待网络爬虫抓取完毕，就可以对抓取的结果进行处理了。图 3 是抓取完毕后的镜像目录。

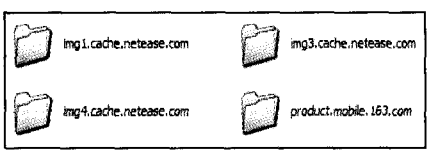

图 3 抓取完毕后的镜像目录

3 信息抽取模块的设计

由于采用 Lucene 建立索引并没有规定的信息格式，只需要将信息转化为文本信息并且构造一个 Document 对象，就可以建立索引，因此本模块主要是对 HTML 形式的网页信息进行抽取，利用抽取工具从 HTML 网页中提取出相应的文本内容，然后将之建立索引并且存入数据库。

常用的抽取方法是通过正则表达式去提取 HTML 文本，而在使用正则表达式的时候需要考虑很多细节，比如一些大小写和空白符都必须考虑，不然得不到想要的结果，同时正则表达式的复用性太差，针对每个特定的网页都需要单独地写正则表达式去提取。针对这一情况，本系统采用 HTMLParser 工具来抽取 HTML 文本的信息。

HTMLParser 是一个纯 Java 写的 HTML 解析的库，主要用于改造和提取 HTML，HTMLParser 能超高速解析 HTML，而且不会出错，使得搜索引擎开发者摆脱了繁琐的正则匹配过程，只需要将 HTMLParser 提供的 Jar 包：htmllexer. jar 和 htmlparser. jar 加入 classpath 中，就可以使用 HTMLParser 提供的 API。通过 HTMLParser 提供的 API，可以很方便地提取特定文本，大大提高开发效率，从而方便了开发者对特定网页信息的提取。

3.1 Lexer 模式功能

Org. htmlparser. lexer 包作为 HTMLParser 底层 I/O 的子系统，负责从 HTML 资源读取字符，解析 Node 节点，HTML 有 3 种类型的 Node 节点：RemarkNode，表示 Html 中的注释；TagNode，表示标签节点；TextNode，表示文本节点。Lexer 从页面读取字符串，用 Cursor 记录当前字符所在的位置，通过状态机来生成 Nodes 节点，然后通过 nextNode 函数查找并返回下一个 Node 节点，直到 Page 的结尾。

3.2 HTMLParser 功能

Lexer 解析 HTML 的方式更底层一些，只能返回一个线性的 Node 节点序列，不能产生树形层次结构的 Node 节点集合，HTMLParser 通过封装 Lexer 对 I/O 节点解析的处理，提供了 2 种访问节点的方法，即 Visitor 模式和 Filter 模式。Filter 模式通过设置一定的过滤条件，对每个节点进行过滤，返回一个符合规则的节点的列表，Visitor 模式定义了访问一个节点时进行的操作接口，只需要基层 NodeVisitor，然后实现相应的接口就可以了。

3.3 提取网页内容

使用解析产品网页信息的基类 Extractor 来提取网页内容，在基类中，实现了大部分的公用方法，比如递归遍历一个目录、复制图片、正则匹配等。Extractor 基类有一个抽象方法 extract()，它主要是提供子类实现对不同格式页面的解析，并指明当前处理器处理后，将文件写到哪个目录下，将产品的相关图片复制到哪个目录下，抓取网页后的镜像目录，通过调用 HTMLParser 的相关 API 对网页进行解析，并使用内容过滤器，分别获取产品名、型号、产品属性信息、图片地址等内容，这些内容依照顺序写入一个 BufferWriter 中，这个 BufferedWriter 在从基类获取的 outpath 目录内创建一个新的文本文件。然后递归遍历镜像目录下的所有文件，提取出网页内容，将处理结果输出到指定目录。图 4 示出解析后的文件。

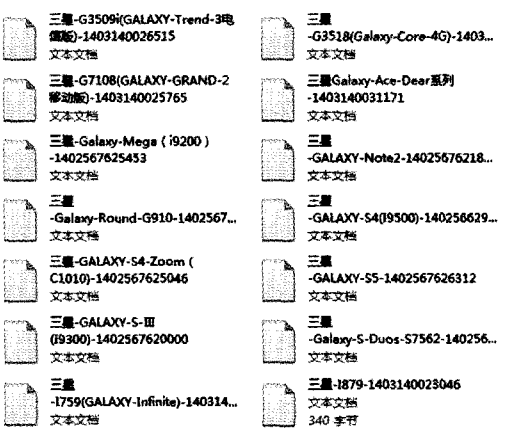

图 4 解析后的文件

4 索引模块的设计

前面下载并提取了网页内容，这里针对提取的数据作进

一步分析整理，建立产品词库，设计数据内容的数据库和索引存储结构，构造数据库存储的程序，设计文本索引形式的搜索引擎结构。

4.1 构建产品检索名称信息词库

通过网络爬虫可以下载大量的手机产品，对于大型垂直搜索引擎会形成海量的信息来源。垂直搜索引擎需要提供精确的信息，仅仅根据分词后的文本索引无法满足，需要为结果建立一个标准的语义词库，通过产品名称词汇选择来实现。

下载的手机产品信息，在搜索的过程中需要提供对应的检索词汇，由于 Lucene 在开发之初，仅仅只有对英文分词的支持，后来的版本虽然加入了对中文分词的支持，但支持的方法仅仅是简单地以空格等来对中文进行分词，而由于中文词义的特殊性，Lucene 的中文分词在很多情况下并不能得到很好的效果，本系统通过采用第三方中文分词系统 JE 分词工具来取代 Lucene 自身的 Analyzer 对中文文本进行分词。

JE 中文分析器是一套由 Java 编写的中文分词系统，使用该分词器的时候无需安装任何插件，可以直接取代 Lucene 自身的分词器来使用[7]。JE 中文分析器内部主要是采用正向最大匹配算法，支持英文、数字、中文等混合分词，通过使用 JE 中文分析器可以很好地处理中文分词问题。表 1 列出了 Lucene 分词器与 JE 分析器对“诺基亚摩托罗拉三星”进行分词的结果。

表 1 不同分词器的分词结果

分词器	分词结果
Whitespaceanalyzer	诺基亚摩托罗拉三星
Stopanalyzer	诺基亚摩托罗拉三星
SimpleFilter	诺基亚摩托罗拉三星
StandardFilter	诺|基|亚|摩|托|罗|拉|三|星
JE 分词器	诺基亚|摩托罗拉|三星

通过表 1 可以看出，JE 分词系统在对中文分词中无论成词率还是准确率都高于 Lucene 自身的分词系统，因此本系统采用 JE 分词器作为分词系统。

标准词库为所有的产品构建一个产品信息的词库，这个词库包括抓取下来的所有产品的品牌和型号，以便用户输入关键字时能够检索到。现在已经有了所有产品的具体信息文件，而且文件名就是由“品牌”和“型号”加一个时间构成，因此可以很方便地解析出所需要的词库，将产品名称词汇提取保存在文件中，可以进一步存储到数据库中或索引文件中，提高检索精确度。

产品名称词库提取代码需要循环检索所有的产品信息文件，并将从中解析出产品的品牌和型号放入一个集合中暂存，最后调用排序方法对其进行排序，然后将排序后的结果写入指定的产品词库文件中去，完成词库提取。

4.2 手机产品数据库与文件索引结构

当手机产品词库建立好后，需要将几部分内容进行存储管理，同时需要提供方便的查询和访问方法。常用的信息存储方式是数据库和索引文件方式。数据库方式简单灵活，便于按照产品关键字检索和组织，但是不利于实现全文检索。索引文件方式速度快，能够组织海量的信息，但不利于格式化存储。本系统将结合这两种方法，分别将信息存储到数据库和索引文件中，形成高效的垂直搜索引擎[8]。

具体需要完成的工作包括将产品详细信息插入数据库，并建立 Lucene 索引。为实现该功能，需要完成以下内容：

(1)定义一个 Product 类，作为装载数据的值对象。

(2)确定数据库与索引的结构和数据类型。

本系统需要将手机产品的名称、型号、内容等存入数据库，依照这些数据来确定数据的所有字段，其格式如表 2 所列。

表 2 手机产品存储的属性信息

属性	说明
Name	表示品牌名称
Type	表示型号
Content	表示详细信息
Summary	显示在搜索结果的信息摘要
originalURL	原始 URL
imageURL	图像 URL
Updatetime	最后更新时间

数据库结构很简单，只有一张 priduct 表，主要就是用来存储产品的各种信息。为了实现全文检索的搜索功能，还要定义 Lucene 的索引格式，这里设计了 ProductDocument 类，该类中构建了 7 个域，其中，前 6 个域与数据库中的内容有直接的对应关系，而最后一个 Field 则是将 name、type 和 content 这 3 个域拼接起来进行保存，免去了利用多域搜索带来的性能损失。

4.3 产品信息数据库存储与处理

使用数据库便于进行数据统计和分析，这里采用 MySQL 数据库，通过可视化管理工具 Navicate for MySQL 创建数据库和数据表。通过 JDBC 方式连接数据库，手机产品的数据库操作主要完成了信息的插入操作，另外提供了获取当前产品最大编号的方法，避免产品编号重复，核心方法是 addProduct 方法，调用参数是一个 Product 类型的对象，在该方法中，从 Product 对象去除相应的值，然后构建一个 PreparedStatement 来执行 JDBC 的写入，在写入前，首先调用 getNextId 方法取得所要插入的这条记录的 ID 号，并随着方法返回给调用者，以便调用者可以将这个 ID 传到 Lucene 的索引中，以此将数据库记录和 Lucene 的索引对应起来[9]。程序的流程图如图 5 所示。

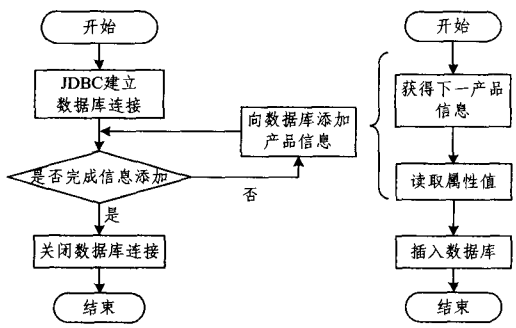

图 5 添加产品信息流程图

4.4 产品信息文件存储与 Lucene 索引

索引文件存数具有更高的效率和速度，为了便于进行全文检索，采用 Lucene 进行文本的索引和处理。创建索引的基本流程步骤如下：

(1)初始化一个 JE 分词的 MMAnalyzer 的实例，用户对文本进行中文分词和文本隔离处理。

(2)读取磁盘词库文件，将前面所处理的词库进行加载，相关的文件存储为孩子通过前面的配置管理类实例得到。

(3)实现 addProduct 方法,以 Product 类型的对象和一个 ID 值为参数,调用构建产品文档方法处理,生成 Lucene 的 Document 文档对象。

(4)调用 IndexWriter 的 addDocument 方法将处理后的结果加入索引中,完成索引的创建和追加。索引的创建流程如图 6 所示。

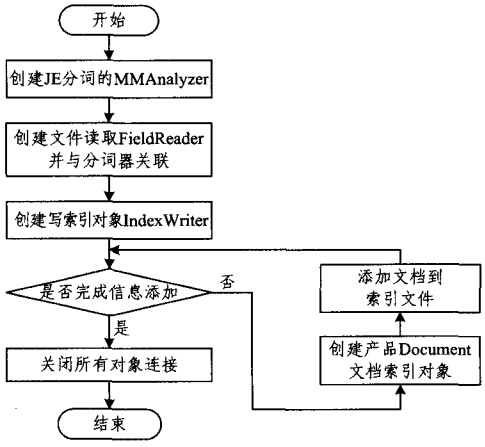

图 6 索引创建程序流程图

4.5 产品信息综合处理

现在创建一个类,把数据库操作、索引操作都集中起来。首先它能从文本文档中提取出需要的内容,来构建 Product 对象,然后调用 ProductJDBC,向数据库中写入 Product 相关信息。另外,再调用 ProductIndexer,把 Product 对象加入到索引中。具体的操作步骤如下:

(1)调用前面所建立的 ProductJDBC 类和 ProductIndexer 类,来对具体的产品信息进行处理。

(2)在构造函数被调用并生成一个 ProductTextFileProcessor 的实例后,需要调用其 initialize 方法来初始化 ProductJDBC 类的实例和 ProductIndexer 的实例。

(3)通过调用 setDirectories 方法,为其注入所要处理的产品详细信息文件所在的目录,接着调用 process 方法就可以开始处理产品信息了。程序的流程图如图 7 所示。

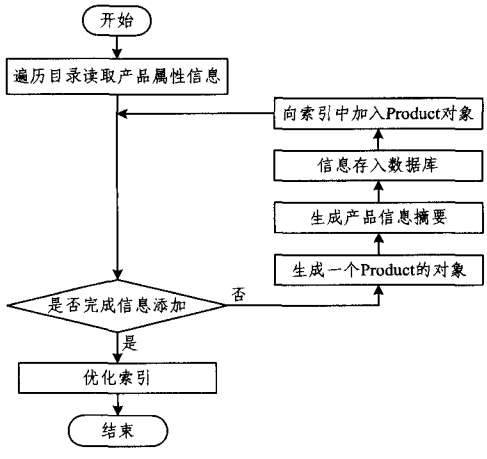

图 7 调用数据库处理类和索引处理类流程图

5 检索模块的设计

整个系统的后台是由 Spring 框架来管理的,首先开发 Searchservice 类与 DAO 类接口,SearchService 类向用户提供了检索的接口,而 DAO 类主要负责从数据库中取出详细的信息返回给用户,然后将前面介绍的索引和数据库进行整合,调用相应的搜索 bean 来完成后台搜索,最后通过视图显示的 JSP 文件在前端显示输出[10]。检索模块主要是查询用户输入的关键词,然后将查询的结果返回到视图层,因此本系统采用 MVC(模型—视图—控制)架构开发,整个设计结构如图 8 所示。

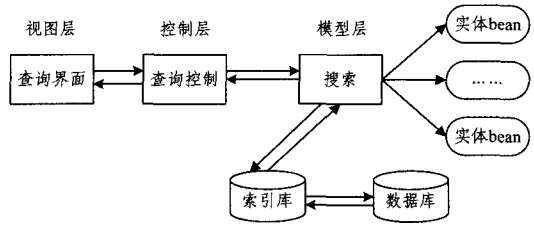

图 8 查询模块结构图

视图层是通过 main.jsp 文件来实现的,主要功能是提供了用户的查询界面,当用户输入关键词查询的时候,调用搜索方法,从后台取得搜索结果集,查询出的结果会在该页面分页显示给用户。

当控制层接受到用户的请求时,控制层会根据用户的请求在 Struts 的配置文件 Struts-config.xml 中找到相应的处理方式,本系统通过 Struts 会找到 Spring 配置好的实体 bean 来进行处理。跟踪配置文件,系统会找到检索类来处理用户的请求,系统检索功能通过 Lucene 中提供的检索接口实现,对于用户输入的关键词,首先会将该关键词构造成 Query 对象,然后将该对象传递给 IndexSearcher 类,IndexSearcher 类会到对应的索引库进行检索。由于本系统中用户使用的是多维短语检索,因此先将查询词进行分词,构造多个查询词进行查询,返回的结果都在 Hits 类中,Hits 类主要是获取查询的结果,排序后返回,与查询关键词相关度越大的结果排在越前面,例如查询"华为 T8850",那么型号为"华为 T8850"的手机信息排在前面,而其他类型的华为手机信息排在后面。

本系统的模型层中定义了 3 个实体类,分别为 SearchResult、SearchResults 和 SearchRequest。SearchResult 代表一个查询结果的 Bean 类,主要对结果属性进行封装,首先通过 DAO 类从数据库中查询出详细的产品信息后,对应 SearchResult 类的每个属性分别填入相应的值。SearchResults 是保存所有查询结果的类,封装了相关的 Bean 形成列表存储,内部的结果集合列表用来存储结果集类型的数据,另外通过 startindex、minpage 和 maxpage 这些属性控制分页。当用户在页面单击一次搜索按钮或选择分页时,都会构造一个 SearchRequest 对象,该对象通过调用查询方法获得查询结果集并返回给用户。

SearchService 类提供了对 Lucene 索引和数据库内容进行搜索的功能,其中的 getSearchResult 方法接收一个 SearchRequest 的对象,并返回一个 SearchResults 的对象、getSearchResultById 方法从数据库中取出一个产品的详细信息。

最后,还需要修改后台服务器 Tomcat 的配置文件 Web.xml,使得 Tomcat 能够找到 Struts 的中央控制器,以便分发到对应的 Action 中进行处理,配置完之后整个系统就完成了。

6 系统运行结果及性能测试

6.1 系统运行结果

图 9 列出了以“三星”为关键词进行搜索得到的搜索结果显示页面。

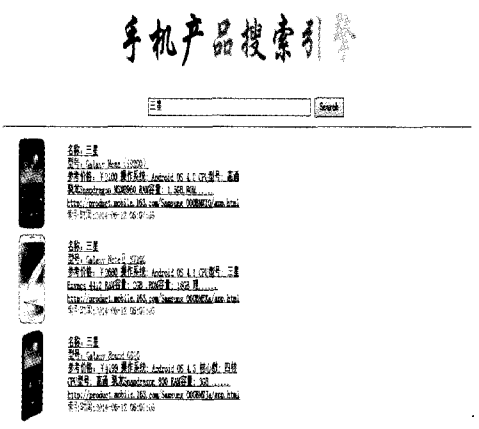

图 9 搜索结果显示页面

6.2 响应时间

为了测试系统，首先从网易手机频道采集了相关手机产品网页共 8893 个，使用关键词“三星”作为查询词，检索出最终结果 184 个，耗时约 0.196 秒，通过人工分析，发现检索结果符合用户期望，能够在系统期望时间内将最重要的网页显示在搜索结果首页上。

在配置为 Pentium(R) Dual-Core CPU E5400 @ 2.70GHz、内存为 2G 的机器上进行了多次测试。在 8746 个网页建立的索引上，以不同的关键词进行了测试，测试结果如图 10 所示。从点击“查询”按钮到显示查询结果，平均时间为 194.0ms，达到了系统设计目标。

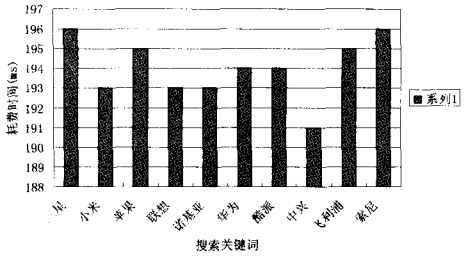

图 10 查询耗费时间

6.3 查全率和查准率

由于本搜索引擎采集的网页取决于初始的种子站点，故初始的种子站点的质量决定了采集到的网页的数量和质量。首先以网易手机频道作为种子站点，共采集了 8893 个网页，使用关键词“三星”作为查询词，检索出最终结果 184 个，经分析，含有“三星”关键词的网页都被检索出来，对于采集到的网页查全率为 100%。但对于整个互联网上的网页而言，查全率取决于种子站点的选择和机器的配置。

查准率：经由人工分析，在检索出的 184 个网页中，符合用户期望、有紧密联系的网页可能共计 67 个，查准率为(67/184)*100%=36.41%，通过与通用搜索引擎百度比较，本系统的查准率较高，排在前面的都是相关度高的手机网页信息，而百度搜索中的结果排在前面的是一些电商网站和排行榜信息，需要作进一步搜索。

综上所述，本文所实现的手机垂直搜索引擎系统达到了预期设计的目标。

结束语 本文通过引用 Lucene 和 Heritrix 构建了一个能够针对手机产品信息进行检索的垂直搜索系统。主要完成了以下工作：

1)通过扩展 Heritrix 相应模块实现了对网页的选择性下载。

2)通过 HtmlParser 工具从网页信息中抽取出机构化文本信息。

3)利用 Lucene 框架对相关的信息建立了索引。

4)设计并实现了基于 MVC 架构的查询接口。

尽管本文的研究取得了一定成果，但还需要在以下几个方面进行完善：

要想维护后台数据库的内容和索引文件，唯一的方法就是重新对网页进行处理，然后重新建索引，重新插入数据库。因此，需要对系统的后台进行部分改造，例如，加入守护线程、不间断地抓取新内容、更新索引。

对于本系统而言，由于只对单个网站的信息进行了搜集，因此产品的信息过于单一，此后将对多个网站的信息进行搜集并整理。

由于本文在使用 HtmlParser 工具对页面进行抽取的时候，对所下载的 URL 页面自身的机构有很大的依赖性，因此在对页面抽取的方面需要作进一步研究，希望能找到一个跟页面结构无关的信息提取算法。

参 考 文 献

[1] Lei Xiang, Xin Meng. A Data Mining Approach to Topic-Specific Web Resource Discovery[C]// Second International Conference on Intelligent Computation Technology and Automation. 2009, 2:595-599

[2] Jia Y, Fan H, et al. Design of an Application Model Based on Vertical Search Engine[C]// Second International Conference on Networking and Distributed Computing. 2011:57-60

[3] Wang Chuan, Chang Gui-ran, et al. An Architecture for Improving the Efficiency of Specialized Vertical Search Engine Based on GPGPUs[C]// Fourth International Conference on Genetic and Evolutionary Computing. 2010:67-70

[4] 王晔. 垂直搜索引擎若干问题研究[D]. 上海：复旦大学，2011

[5] 刘育莲. 手机产品垂直搜索引擎的设计与实现[D]. 西安：西安电子科技大学，2012

[6] 刘丽杰. 垂直搜索引擎中聚焦爬虫技术的研究[D]. 哈尔滨：哈尔滨工程大学，2012

[7] 奉国和，郑伟. 国内中文自动分词技术研究综述[J]. 图书情报工作，2011(2):43-47

[8] 刘琦. 垂直搜索引擎的设计与开发[D]. 广州：中山大学，2010

[9] 罗刚. 解密搜索引擎技术实战[M]. 北京：电子工业出版社，2011

[10] 邱哲，符涛涛，王学松. 开发自己的搜索引擎[M]. 北京. 人民邮电出版社，2010