第一图书网, tushu007.com

<<开发自己的搜索引擎--Lucen>>

图书基本信息

书名：<<开发自己的搜索引擎--Lucene 2.0+Heriterx>>

13位ISBN编号：9787115160003

10位ISBN编号：7115160007

出版时间：2007-6

出版时间：人民邮电出版社

作者：邱哲,符滔滔

页数：521

字数：662000

版权说明：本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问：http://www.tushu007.com

第一图书网, tushu007.com

<<开发自己的搜索引擎--Lucen>>

内容概要

本书是一本针对搜索引擎开发的书籍。
通过学习本书，读者可以独立构建出一个企业级的搜索引擎网站。
本书详细讲解了搜索引擎与信息检索基础，Lucene入门实例，Lucene索引的建立，使用Lucene进行搜索，排序，过滤和分页，Lucene的分析器，对Word、Excel和PDF格式文档的处理，Compass搜索引擎框架，Lucene分布式和Google Search API，爬虫Heritrix，HTMLParser，DWR等内容。
最后综合使用所讲述的技术，构建了一个典型的垂直搜索系统，该系统具有很强的商业实用价值。

本书是一本介绍如何使用Lucene 2.0和Heritrix来构建搜索引擎的书。
通过对相关API和源代码的分析，力求使读者在掌握应用的基础上能够深入其核心，自行扩展和开发相应组件，开发出更有创意的搜索引擎产品。

本书适合从事计算机软件开发的人员阅读，同时也可以作为搜索引擎爱好者的入门书籍。
阅读本书需要具备Java语言基础。

第一图书网, tushu007.com

作者简介

邱哲，北京理工大学硕士，现为某公司技术经理，主要从事欧美软件外包开发。
在J2EE方面有4年的开发经验，在搜索引擎与“爬虫”方面有3年的开发经验，著有《征服Ajax+Lucene构建搜索引擎》一书。

第一图书网, tushu007.com

<<开发自己的搜索引擎--Lucen>>

书籍目录

第一篇 搜索引擎入门 第1章 搜索引擎与信息检索基础 1.1 搜索引擎的历史 1.1.1 萌芽：Archie、Gopher 1.1.2 起步：Robot(网络机器人)的出现与Spider(网络爬虫) 1.1.3 发展：Excite、Galaxy、Yahoo等 1.1.4 繁荣：Infoseek、AltaVista、Google和Baidu 1.2 信息检索系统的基本知识 1.2.1 什么是信息检索系统 1.2.2 信息检索的过程 1.2.3 传统查找的优点和不足 1.2.4 使用索引提高检索速度 1.2.5 倒排索引 1.2.6 评价信息检索系统的标准 1.3 Lucene简介 1.4 小结 第二篇 Lucene开发详解 第2章 Lucene入门实例 2.1 实例介绍 2.1.1 实例说明 2.1.2 开发过程 2.2 准备工作 2.2.1 将文档的全角标点转换成半角标点 2.2.2 将大文档切分成多个小文档 2.2.3 预处理源文件的统一接口 2.3 创建Eclipse工程 2.3.1 准备工作 2.3.2 创建工程并引入Lucene的JAR包 2.3.3 运行文档预处理类 2.3.4 创建处理文档的索引类：IndexProcessor 2.3.5 创建检索索引的搜索类 2.4 运行效果 2.5 小结 第3章 Lucene索引的建立 3.1 Document逻辑文件 3.1.1 Lucene的Document 3.1.2 为Document添加多种Field 3.1.3 Document的内部实现 3.2 Field的内部实现 3.2.1 Field包含的类 3.2.2 Field类的构造方法 3.3 Lucene的索引工具IndexWriter 3.3.1 IndexWriter的初始化 3.3.2 向索引添加文档 3.3.3 限制每个Field中的词条的数量 3.4 Lucene索引过程详解 3.4.1 Lucene索引建立过程概述 3.4.2 使用addDocument方法向索引添加文档 3.4.3 DocumentWriter的addDocument方法 3.4.4 文档的倒排 3.4.5 对postingTable进行排序 3.4.6 将Posting信息写入索引 3.5 索引文件格式 3.5.1 索引的segment 3.5.2 .fnm格式 3.5.3 .fdx与.fdt格式 3.5.4 .tii与.tis格式 3.5.5 deletable格式 3.5.6 复合索引格式.cfs 3.6 索引过程的优化 3.6.1 合并因子mergeFactor 3.6.2 maxMergeDocs 3.6.3 minMergeDocs 3.7 索引的合并与索引的优化 3.7.1 FSDirectory与RAMDirectory 3.7.2 使用IndexWriter来合并索引 3.7.3 索引的优化 3.8 从索引中删除文档 3.8.1 索引的读取工具IndexReader 3.8.2 使用文档ID号来删除特定文档 3.8.3 使用Field信息来删除批量文档 3.9 Lucene的同步问题 3.9.1 为什么要进行同步以及Lucene的同步法则 3.9.2 commit.lock与write.lock 3.10 Lucene 2.0的新类：IndexModifier类 3.11 小结 第4章 Lucene的搜索 4.1 使用IndexSearcher进行搜索 4.1.1 初始化IndexSearcher 4.1.2 IndexSearcher最简单的使用 4.1.3 IndexSearcher的多种search方法 4.2 Hits类详解 4.2.1 Hits类的公有接口 4.2.2 效率分析 4.2.3 Hits内部的缓存 4.2.4 Hits类的工作原理 4.3 对搜索结果的评分 4.3.1 文档与词条的向量空间 4.3.2 Lucene的文档得分算法 4.4 构建各种Lucene内建的Query对象 4.4.1 toString查看原子查询 4.4.2 查询重写与权重 4.4.3 TermQuery词条搜索 4.4.4 BooleanQuery布尔搜索 4.4.5 RangeQuery范围搜索 4.4.6 PrefixQuery前缀搜索 4.4.7 PhraseQuery短语搜索 4.4.8 MultiPhraseQuery多短语搜索 4.4.9 FuzzyQuery模糊搜索 4.4.10 WildcardQuery通配符搜索 4.4.11 SpanQuery跨度搜索 4.5 第三方提供的Query对象：RegexQuery 4.6 通过QueryParser转换用户关键字 4.6.1 词条的定义 4.6.2 QueryParser初始化 4.6.3 改变QueryParser默认的布尔逻辑 4.6.4 短语和QueryParser 4.6.5 FuzzyQuery和QueryParser 4.6.6 通配符与QueryParser 4.6.7 查找指定的Field 4.6.8 RangeQuery与QueryParser 4.6.9 QueryParser和SpanQuery 4.7 多Field搜索与多索引搜索 4.7.1 多域搜索MultiFieldQueryParser 4.7.2 MultiSearcher在多个索引上搜索 4.7.3 ParalellMultiSearcher：多线程搜索 4.7.4 Searchable和RMI 4.8 小结 第5章 排序、过滤和分页 5.1 相关度排序 5.1.1 使用Score进行自然排序 5.1.2 Searcher的explain方法 5.1.3 通过改变boost值来改变文档的得分 5.2 使用Sort来排序 5.2.1 Sort简介 5.2.2 SortField 5.2.3 按文档得分进行排序 5.2.4 按文档的内部ID号来排序 5.2.5 按一个或多个Field来排序 5.2.6 改变SortField中的Locale信息 5.3 搜索的过滤器 5.3.1 过滤器的基本结构 5.3.2 一个简单的Filter：建立索引 5.3.3 一个简单的Filter：打印索引文档信息 5.3.4 一个简单的Filter

第一图书网, tushu007.com

：安全级别与过滤器代码 5.3.5 一个简单的Filter：在搜索时应用过滤器 5.3.6 一个简单的Filter：总结 5.3.7 按范围过滤RangeFilter 5.3.8 在结果中查询QueryFilter 5.3.9 缓存结果：CachingWrapperFilter 5.4 翻页问题 5.4.1 依赖于session的翻页 5.4.2 多次查询 5.4.3 缓存+多次查询 5.4.4 缓存+多次查询+数据库 5.5 小结 第6章 Lucene的分析器 6.1 分析 6.1.1 分词 6.1.2 Lucene的分析器的结构 6.1.3 Lucene的分析器的实现 6.2 Lucene与JavaCC 6.2.1 JavaCC简介 6.2.2 JavaCC为Lucene提供的分析器脚本 6.2.3 Lucene的标准分析器 6.2.4 标准过滤器：StandardFilter 6.2.5 大小写转换器：LowerCaseFilter 6.2.6 忽略词过滤器：StopFilter 6.3 分析器的进阶 6.3.1 再看StandardAnalyzer中的管道过滤器结构 6.3.2 长度过滤器：LengthFilter 6.3.3 PerFieldAnalyzerWrapper 6.3.4 其他 6.4 对中文的分析 6.4.1 现有的中文分词方式简介 6.4.2 中科院的分词软件和JE分词 6.5 小结 第三篇 Lucene相关话题 第7章 对Word、Excel 和PDF的处理 7.1 使用PDFBox处理PDF文档 7.1.1 PDFBox的下载 7.1.2 在Eclipse中配置 7.1.3 使用PDFBox解析PDF内容 7.1.4 运行效果 7.1.5 与Lucene的集成 7.2 使用xpdf来处理中文PDF文档 7.2.1 xpdf的下载 7.2.2 配置 7.2.3 提取中文 7.2.4 运行效果 7.3 使用POI来处理Excel和Word文件格式 7.3.1 对Excel的处理类 7.3.2 ExcelReader的运行效果 7.3.3 POI中Excel文件Cell的类型 7.3.4 对Word的处理类 7.4 使用Jacob来处理Word文档 7.4.1 Jacob的下载 7.4.2 在Eclipse中配置 7.5 小结 第8章 Compass：封装了Lucene的框架 8.1 Compass简介 8.1.1 Compass的下载 8.1.2 Compass的代码片断 8.2 Compass的初始配置 8.2.1 Compass的配置文件 8.2.2 将索引存放于内存中 8.2.3 使用JDBC来存储索引 8.2.4 使用连接池来存储索引 8.2.5 加载compass.cfg.xml文件 8.3 域模型的配置 8.3.1 实体代码 8.3.2 实体关系 8.3.3 实体Book的配置文件 8.3.4 通用元数据定义文件(.cmd.xml) 8.3.5 Author和Article的配置文件 8.4 使用Compass来建立索引 8.4.1 索引代码 8.4.2 对象关系图和运行结果 8.5 使用Compass来搜索 8.5.1 使用find()方法搜索 8.5.2 CompassHits类型 8.5.3 CompassHit类型 8.5.4 使用Lucene语法来查找 8.6 配置Analyzer和Optimizer 8.7 小结 第9章 Lucene分布式和Google Search API 9.1 Lucene与分布式 9.1.1 什么是GFS 9.1.2 为Lucene提供分布式的几点设想 9.2 Google的Search API 9.2.1 搭建环境 9.2.2 构建搜索类 9.2.3 设置查询时的参数和查询语法 9.2.4 运行测试 9.3 小结 第四篇 网络爬虫Heritrix 第10章 无比强大的网络爬虫Heritrix 10.1 Heritrix使用入门 10.1.1 下载和运行Heritrix 10.1.2 在Eclipse里配置Heritrix的开发环境 10.1.3 创建一个新的抓取任务 10.1.4 设置抓取时的处理链 10.1.5 设置运行时的参数 10.1.6 运行抓取任务 10.1.7 Heritrix的镜像存储结构 10.1.8 终止抓取或终止Heritrix的运行 10.2 Heritrix的架构 10.2.1 抓取任务CrawlOrder 10.2.2 中央控制器CrawlController 10.2.3 Frontier链接制造工厂 10.2.4 用Berkeley DB实现的BdbFrontier 10.2.5 Heritrix的多线程ToeThread和ToePool 10.2.6 处理链和Processor 10.3 扩展和定制Heritrix 10.3.1 向Heritrix中添加自己的Extractor 10.3.2 定制Queue-assignment-policy的两个问题 10.3.3 定制Queue-assignment-policy继承QueueAssignmentPolicy类 10.3.4 扩展FrontierScheduler来抓取特定的内容 10.3.5 在Prefetcher中取消robots.txt的限制 10.4 小结 第五篇 构建垂直搜索引擎 第11章 搜索引擎综合实例：准备篇 11.1 实例简介以及实现途径 11.1.1 选择网站 11.1.2 太平洋电脑网和网易手机频道 11.1.3 分析网站内容并准备抓取清单 11.1.4 从下拉列表获得手机品牌首页 11.1.5 解析手机品牌页面 11.2 在Heritrix中为pconline开发抓取所需的定制类 11.2.1 保存所有产品的页面和图片 11.2.2 不保存其他无关页面 11.2.3 开始抓取 11.3 在Heritrix中为网易手机频道开发抓取所需的定制类 11.3.1 分析网易手机频道 11.3.2 设计抓取代码 11.4 在Eclipse中创建工程结构 11.4.1 下载插件 11.4.2 在Eclipse中配置插件 11.4.3 创建工程 11.4.4 设置工程的Context 11.4.5 设定源代码存放和输出路径 11.4.6 添加Java代码 11.4.7 添加Jar包 11.4.8 创建JSP文件 11.4.9 工程整体结构一览 11.5 设定配置文件及其相关类 11.5.1 系统属性配置文件 11.5.2 封装配置文件 11.6 产品详细信息

第一图书网, tushu007.com

文件格式 11.7 解析网页信息的基类Extractor 11.8 太平洋电脑网手机产品页面Extractor 11.9 pconline产品信息运行效果测试 11.9.1 编写测试函数 11.9.2 执行测试 11.10 网易手机频道的产品信息运行效果 11.11 构建产品信息词库 11.12 数据库与索引结构 11.12.1 定义Product类 11.12.2 确定数据库与索引的结构 11.13 数据库处理和索引处理 11.13.1 对数据库进行操作 11.13.2 对索引进行操作 11.14 调用数据库处理类和索引处理类 11.15 运行 11.16 小结 第12章 使用正则表达式与HTMLParser提取网页内容 12.1 HTML的基本知识 12.2 JDK中的正则表达式提取网页内容 12.2.1 java.util.regex包 12.2.2 正则表达式提取网页内容实例 12.3 HTMLParser提取网页内容 12.3.1 HTMLParser的下载 12.3.2 HTMLParser概述 12.3.3 Lexer的功能及实现 12.3.4 HTMLParser的功能及实现 12.3.5 HTMLParser实例 12.4 小结 第13章 搜索引擎综合实例：DWR 13.1 DWR的下载 13.2 DWR入门与实例演示 13.2.1 创建工程结构 13.2.2 在web.xml中配置DWR 13.2.3 配置dwr.xml 13.2.4 页面代码 13.2.5 运行效果 13.2.6 DWR与直接使用XMLHttpRequest对象的比较 13.2.7 在DWR中操纵自定义的对象 13.2.8 查看DWR的输出日志 13.3 dwr.xml的配置 13.3.1 dwr.xml的标准结构 13.3.2 标签与DWR自带的converter和creator 13.3.3 标签 13.3.4 标签 13.3.5 另一个例子 13.4 util.js 13.4.1 调用util.js 13.4.2 使用useLoadingMessage方法显示提示图标 13.4.3 DWRUtil.setValue和DWRUtil.getValue 13.4.4 DWRUtil.getValues和DWRUtil.setValues 13.4.5 DWRUtil.addOptions和DWRUtil.removeAllOptions 13.4.6 DWRUtil.addRows和DWRUtil.removeAllRows 13.4.7 DWRUtil.toDescriptiveString方法 13.5 小结 第14章 搜索引擎综合实例：Web篇 14.1 配置文件 14.1.1 Spring配置文件 14.1.2 DWR配置文件 14.1.3 web.xml 14.2 各种Bean类 14.2.1 SearchResult 14.2.2 SearchResults 14.2.3 SearchRequest 14.3 SearchService的实现 14.4 SearchResultDao 14.5 前台部分 14.5.1 搜索主页面main.jsp 14.5.2 图片的显示 14.5.3 详细信息页面detail.jsp 14.6 问题 14.7 小结

第一图书网, tushu007.com

版权说明

本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问:http://www.tushu007.com