

<<开发自己的搜索引擎>>

图书基本信息

书名：<<开发自己的搜索引擎>>

13位ISBN编号：9787115215291

10位ISBN编号：7115215294

出版时间：2010-1

出版时间：人民邮电出版社

作者：邱哲，符滔滔，王学松 著

页数：562

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>

<<开发自己的搜索引擎>>

前言

2007年初，我们编写了本书第1版，该书曾经连续数周占据互动出版网计算机类畅销书排行榜首。转眼间两年过去了，我收集了很多读者的建议和问题，针对书中的细节进行了调整，推出了本书的第2版。

第2版主要在以下方面进行了改进。

(1) 对第1版中语言进行了优化，以使行文更加流畅，便于阅读，同时对一些表达模糊的地方进行了改写。

(2) 针对读者提出的一些问题，进行了勘误。

(3) 对书中的大部分Visio图片进行了重绘，看起来更美观。

(4) 对书中涉及的软件进行全面梳理，对版本进行了更新，主要包括以下两部分。

· Lucene采用了2.0稳定版，正文中的范例都是使用它来实现的。

目前Lucene已经发布到了2.4版，但是这个版本还没有大规模的商业化应用，存在很多不稳定因素，因此笔者只在附录中介绍了其核心功能。

· 网络爬虫采用了Heritrix1.14.0版本，包括后面的大案例都是使用这个版本重新实现的。

(5) 去除了第9章中已经不再使用的Google的searchAPI部分。

<<开发自己的搜索引擎>>

内容概要

《开发自己的搜索引擎：Lucene+Heritrix(第2版)》是一本介绍搜索引擎开发的书籍，通过《开发自己的搜索引擎：Lucene+Heritrix(第2版)》，读者可以独立构建一个企业级的搜索引擎网站。

《开发自己的搜索引擎：Lucene+Heritrix(第2版)》讲解了搜索引擎与信息检索基础，Lucene入门实例，索引的建立，使用Lucene来搜索，排序，分析器，对Word、Excel和PDF格式文档的解析，Compass搜索引擎框架，Lucene分布式，爬虫Heritrix，HTMLParser，DWR等内容。

最后综合这些技术，构建了一个典型的垂直搜索系统，具有很强的商业实用价值。

《开发自己的搜索引擎：Lucene+Heritrix(第2版)》是一本使用Lucene和Heritrix来讲解搜索引擎构建的书，通过对API和源代码的分析，力求使读者在应用的基础上，能够深入其核心，自行扩展和开发相应组件，发挥想象力，开发出更具有创意的搜索引擎产品。

《开发自己的搜索引擎：Lucene+Heritrix(第2版)》适合Java程序员和从事计算机软件开发的编程人员阅读，同时也可以作为搜索引擎爱好者的入门书籍。

<<开发自己的搜索引擎>>

作者简介

邱哲，北京理工大学软件工程硕士。

现为Eskalate.com公司技术经理，同时负责开发人员招聘工作—主要从事欧美软件外包开发，曾承接多家美国本土公司项目，在J2EE方面有7年的开发经验。

曾经编写了《souts Web设计与开发大全》、《开发自己的搜索引擎——Lucerie 2.0+Heritrix》。

王学松，博士。

曾任职知名互联网搜索引擎公司，担任高级软件工程师、研发经理等职位，参与大型搜索引擎开发多年。

开发完成亿级网页的互联网科技类信息垂直搜索引擎系统，完成中文搜索引擎的页面下载与分析、大规模索引建立、分类聚类技术、高并发检索和Web高速访问技术开发。

目前从事海量信息挖掘、语义网搜索引擎和基于内容图像检索的研究和开发。

<<开发自己的搜索引擎>>

书籍目录

第1章 搜索引擎与信息检索 11.1 搜索引擎的历史 11.1.1 萌芽：Archie、Gopher 11.1.2 起步：Robot(网络机器人)的出现与Spider(网络爬虫) 31.1.3 发展：Excite、Galaxy、Yahoo等 41.1.4 繁荣：Infoseek、AltaVista、Google和Baidu 61.2 信息检索系统的基本知识 91.2.1 信息检索系统 91.2.2 信息检索的过程 111.2.3 传统查找的优点和不足 121.2.4 使用索引提高检索速度 121.2.5 倒排索引 131.2.6 评价信息检索系统的标准 141.3 Lucene的简介 141.4 小结 15第2章 Lucene入门实例 162.1 实例介绍 162.1.1 实例说明 162.1.2 开发过程 162.2 准备工作 172.2.1 将文档的全角标点转成半角标点 172.2.2 将大文档切分成多个小文档 202.2.3 预处理源文件的统一接口 212.3 创建Eclipse工程 222.3.1 准备工作 222.3.2 创建工程并引入Lucene的JAR包 242.3.3 运行文档预处理类 312.3.4 创建处理文档的索引类：IndexProcessor 322.3.5 创建检索索引的搜索类 342.4 运行效果 382.5 小结 41第3章 索引的建立 423.1 Document逻辑文件 423.1.1 Lucene的Document 423.1.2 为Document添加多种Field 433.1.3 Document的内部实现 453.2 Field的内部实现 463.2.1 Field包含的类 473.2.2 Field类的构造方法 483.3 Lucene的索引工具IndexWriter 493.3.1 IndexWriter的初始化 503.3.2 向索引添加文档 523.3.3 限制每个Field中的词条的数量 533.4 Lucene索引过程详解 543.4.1 Lucene索引建立过程概览 543.4.2 使用addDocument方法向索引添加文档 553.4.3 DocumentWriter的addDocument方法 573.4.4 文档的倒排 623.4.5 对postingTable进行排序 663.4.6 将Posting信息写入索引 683.5 索引文件格式 683.5.1 索引的segment 693.5.2 .fnm格式 693.5.3 .fdx与.fdt格式 703.5.4 .tii与.tis格式 713.5.5 deletable格式 713.5.6 复合索引格式.cfs 713.6 索引过程的调优 723.6.1 合并因子mergeFactor 723.6.2 maxMergeDocs 733.6.3 minMergeDocs 733.7 索引的合并与索引的优化 743.7.1 FSDirectory与RAMDirectory 743.7.2 使用IndexWriter来合并索引 753.7.3 索引的优化 763.8 从索引中删除文档 783.8.1 索引的读取工具Index-Reader 783.8.2 使用文档ID号来删除特定文档 813.8.3 使用Field信息来删除批量文档 843.9 Lucene的同步问题 853.9.1 为什么要进行同步以及Lucene的同步法则 853.9.2 commit.lock与write.lock 853.10 Lucene 2.0的新类：IndexModifier类 863.11 小结 87第4章 Lucene搜索 884.1 使用IndexSearcher进行搜索 884.1.1 初始化IndexSearcher 884.1.2 IndexSearcher的最简单使用 894.1.3 IndexSearcher的多种search方法 904.2 Hits类详解 924.2.1 Hits类的公有接口 924.2.2 效率分析 934.2.3 Hits内部的缓存 954.2.4 Hits类的工作原理 984.3 对搜索结果的评分 984.3.1 文档与词条的向量空间 984.3.2 Lucene的文档得分算法 994.4 构建各种Lucene内建的Query对象 1034.4.1 toString：查看原子查询 1034.4.2 查询重写与权重 1034.4.3 TermQuery词条搜索 1044.4.4 BooleanQuery布尔搜索 1054.4.5 RangeQuery范围搜索 1134.4.6 PrefixQuery前缀搜索 1174.4.7 PhraseQuery：短语搜索 1194.4.8 MultiPhraseQuery：多短语搜索 1234.4.9 FuzzyQuery模糊搜索 1284.4.10 WildcardQuery通配符搜索 1314.4.11 SpanQuery跨度搜索 1324.5 第三方提供的Query对象：RegexQuery 1404.6 通过QueryParser转换用户关键字 1424.6.1 词条的定义 1434.6.2 QueryParser初始化 1434.6.3 改变QueryParser默认的布尔逻辑 1444.6.4 短语和QueryParser 1454.6.5 FuzzyQuery和QueryParser 1474.6.6 通配符与QueryParser 1474.6.7 查找指定的Field 1484.6.8 RangeQuery与QueryParser 1514.6.9 QueryParser和SpanQuery 1524.7 多Field搜索与多索引搜索 1534.7.1 多域搜索MultiFieldQuery-Parser 1534.7.2 MultiSearcher在多个索引上搜索 1554.7.3 ParalellMultiSearcher：多线程搜索 1584.7.4 Searchable和RMI 1614.8 小结 162第5章 排序、过滤和分页 1635.1 相关度排序 1635.1.1 使用Score进行自然排序 1635.1.2 Searcher的explain方法 1655.1.3 通过改变boost值来改变文档的得分 1665.2 使用Sort来排序 1705.2.1 Sort简介 1705.2.2 SortField 1715.2.3 按文档得分进行排序 1725.2.4 按文档的内部ID号来排序 1755.2.5 按一个或多个Field来排序 1755.2.6 改变SortField中的Locale信息 1825.3 搜索的过滤器 1835.3.1 过滤器的基本结构 1835.3.2 一个简单的Filter：建立索引 1845.3.3 一个简单的Filter：打印索引文档信息 1865.3.4 一个简单的Filter：安全级别与过滤器代码 1875.3.5 一个简单的Filter：在搜索时应用过滤器 1885.3.6 一个简单的Filter：总结 1905.3.7 按范围过滤RangeFilter 1905.3.8 在结果中查询QueryFilter 1945.3.9 缓存结果：Caching-WrapperFilter 1975.4 翻页问题 1985.4.1 依赖于session的翻页 1985.4.2 多次查询 1985.4.3 缓存+多次查询 1995.4.4 缓存+多次查询+数据库 1995.5 小结 200第6章 Lucene的分析器 2016.1 分析 2016.1.1 分词 2016.1.2 Lucene的分析器结构 2026.1.3 Lucene的分析器实现 2046.2 Lucene与JavaCC 2056.2.1 JavaCC简介 2056.2.2 JavaCC为Lucene提供的分析器脚本 2066.2.3 Lucene的标准分析器 2106.2.4 标准过滤器：Standard-Filter 2116.2.5 大小写转换器：Lower-CaseFilter 2126.2.6 忽略词过滤器：StopFilter

<<开发自己的搜索引擎>>

2136.3 分析器的进阶 2136.3.1 再看StandardAnalyzer中的管道过滤器结构 2146.3.2 长度过滤器
：LengthFilter 2146.3.3 PerFieldAnalyzerWrapper 2156.3.4 其他 2156.4 对中文的分析 2166.4.1 现有的中文分
词方式简介 2166.4.2 中科院的分词软件和JE分词 2186.5 小结 224第7章 Word、Excel和PDF的处理 2257.1
使用PDFBox处理PDF文档 2257.1.1 PDFBox的下载 2257.1.2 在Eclipse中配置 2267.1.3 使用PDFBox解析PDF
内容 2277.1.4 运行效果 2287.1.5 与Lucene的集成 2307.2 使用xpdf来处理中文PDF文档 2327.2.1 xpdf的下载
2327.2.2 配置 2327.2.3 提取中文 2337.2.4 运行效果 2367.3 使用POI来处理Excel和Word文件格式 2377.3.1
对Excel的处理类 2377.3.2 ExcelReader的运行效果 2417.3.3 POI中Excel文件Cell的类型 2437.3.4 对Word的
处理类 2457.4 使用Jacob来处理Word文档 2477.4.1 Jacob的下载 2477.4.2 在Eclipse中配置 2477.5 小结 249
第8章 Compass：封装了Lucene的框架 2508.1 Compass简介 2508.1.1 Compass的下载 2508.1.2 Compass的代
码片断 2508.2 Compass的初始配置 2528.2.1 Compass的配置文件 2528.2.2 将索引存放于内存中 2538.2.3 使
用JDBC来存储索引 2538.2.4 使用连接池来存储索引 2548.2.5 加载compass.cfg.xml文件 2558.3 域模型的配
置 2558.3.1 实体代码 2558.3.2 实体关系 2618.3.3 实体Book的配置文件 2628.3.4 通用元数据定义文
件(.cmd.xml) 2638.3.5 Author和Article的配置文件 2678.4 使用Compass来建立索引 2698.4.1 索引代码
2698.4.2 对象关系图和运行结果 2718.5 使用Compass来搜索 2728.5.1 使用find()方法搜索 2728.5.2
CompassHits类型 2738.5.3 CompassHit类型 2748.5.4 使用Lucene语法来查找 2758.6 配置Analyzer
和Optimizer 2768.7 小结 277第9章 Lucene分布式 2789.1 Lucene与分布式 2789.1.1 什么是GFS 2789.1.2
为Lucene提供分布式的几点设想 2799.2 小结 281第10章 无比强大的网络爬虫Heritrix 28210.1 Heritrix的使
用入门 28210.1.1 下载和运行Heritrix 28210.1.2 在Eclipse里配置heritrix的开发环境 28510.1.3 创建一个新的
抓取任务 29010.1.4 设置抓取时的处理链 29210.1.5 设置运行时的参数 29510.1.6 运行抓取任务 29710.1.7
Heritrix的镜像存储结构 30210.1.8 终止抓取或终止Heritrix的运行 30310.2 Heritrix的架构 30410.2.1 抓取任
务CrawlOrder 30410.2.2 中央控制器CrawlController 30510.2.3 Frontier链接制造工厂 30810.2.4
用Berkeley DB实现的BdbFrontier 31310.2.5 Heritrix的多线程ToeThread和ToePool 31610.2.6 处理链
和Processor 31910.3 扩展和定制Heritrix 32210.3.1 向Heritrix中添加自己的Extractor 32310.3.2 定
制Queue-assignment-policy两个问题 32710.3.3 定制Queue-assignment-policy继承 QueueAssignmentPolicy类
32810.3.4 扩展FrontierScheduler来抓取特定的内容 32910.3.5 在Prefetcher中取消robots.txt的限制 33010.4 小
结 331第11章 搜索引擎综合实例：准备篇 33211.1 数码产品垂直搜索引擎实例简介 33211.1.1 垂直搜索引
擎实现流程 33211.1.2 数码垂直搜索引擎搜索功能 33311.1.3 信息来源网站的选择方法 33311.1.4 太平洋电
脑网和网易手机频道 33411.2 准备Eclipse的Web开发环境 33511.2.1 准备Eclipse的Web插件环境 33511.2.2
在Eclipse中配置插件 33611.3 准备垂直搜索引擎工程 33711.3.1 建立搜索引擎Eclipse工程 33811.3.2 设置搜
索引擎工程上下文信息 33911.3.3 设定源代码存放和输出路径 34011.3.4 添加自定义的Java代码 34111.3.5
添加工程中引用的Jar包 34311.3.6 创建工程JSP页面文件 34511.3.7 构造完成的工程整体结构 34711.4 搜索
引擎配置信息管理及相关类 34911.4.1 工程配置信息管理 34911.4.2 系统属性配置文件 35011.4.3 配置文件
管理封装类 35011.5 小结 352第12章 搜索引擎综合实例：下载篇 35312.1 数码产品网络爬虫 35312.1.1 垂
直搜索引擎网络爬虫设计 35312.1.2 来源网站内容与链接分析 35412.2 数码产品信息来源列表准备
35612.2.1 太平洋电脑网待抓取内容页面分析 35612.2.2 太平洋电脑网带抓取内容代码分析 35912.2.3 太平
洋电脑网手机品牌清单分析 36212.3 Eclipse中定制数码产品Heritrix爬虫 36712.3.1 数码产品Heritrix爬虫
的功能 36712.3.2 Eclipse中导入编译Heritrix工程 36812.3.3 Eclipse中运行Heritrix工程 37012.4 抓取pconline
网页的定制扩展类 37112.4.1 抓取pconline网页的Frontier扩展 37112.4.2 执行pconline手机网页抓取任务
37312.5 抓取网易手机频道的定制扩展类 37512.5.1 网易手机频道结构分析 37512.5.2 设计网易抓取
的Extractor扩展 37812.5.3 设计网易抓取的Frontier扩展 38112.5.4 执行网易手机频道网页抓取任务 38212.6
小结 383第13章 使用正则表达式与HTML Parser分析网页 38413.1 网页内容分析方法概述 38413.1.1 网
页HTML的基本知识 38413.1.2 JDK正则表达式简介 38513.1.3 HTMLParser开源库介绍 38713.2 正则表达
式精确提取网页内容 38813.2.1 正则表达式java.util.regex使用 38813.2.2 正则表达式提取tom星座内容实例
39013.2.3 正则表达式提取pconline手机品牌列表 39613.3 HTMLParser高效提取网页内容 39813.3.1
HTMLParser使用准备 39813.3.2 Lexer模式功能及实现 39913.3.3 HTMLParser功能及实现 40413.3.4
HTMLParser解析星座网页实例 41013.4 数码产品网页内容解析系统 41313.4.1 产品详细信息文件格式
41313.4.2 解析产品网页信息的基类Extractor 41413.5 pconline手机产品网页内容解析 41813.5.1 pconline手

<<开发自己的搜索引擎>>

机产品页面Extractor解析器 41813.5.2 pconline产品信息解析测试函数 42113.5.3 pconline产品信息解析代码执行结果 42213.6 网易手机频道产品信息解析 42513.6.1 网易手机频道产品信息的Extractor解析器 42513.6.2 网易手机频道的产品信息运行测试效果 42813.7 小结 429第14章 网页内容存储与索引 43014.1 构建产品检索名称信息词库 43014.1.1 产品名称词汇选择 43014.1.2 产品名称词库提取代码 43114.1.3 产品名称词库提取结果 43314.2 手机产品数据库与文件索引结构 43414.2.1 手机产品的存储方法 43414.2.2 手机产品信息Product类 43514.2.3 产品信息数据库存储结构 43714.2.4 产品信息Lucene索引结构 43814.3 产品信息数据库存储与处理 43914.3.1 数据库创建与准备 43914.3.2 Java数据库基本操作 44014.3.3 数码产品数据库记录操作 44114.4 产品信息文件存储与Lucene索引 44314.4.1 数码产品Lucene索引操作设计 44314.4.2 数码产品具体索引操作代码 44514.5 产品信息综合处理与运行 44614.5.1 调用数据库处理类和索引处理类 44614.5.2 数码产品数据处理类运行 45214.6 小结 454第15章 搜索引擎综合实例：交互篇 45515.1 DWR的技术介绍 45515.1.1 Ajax与DWR简介 45515.1.2 Ajax与传统模式搜索架构 45615.2 DWR安装与配置 45715.2.1 DWR的下载与安装 45715.2.2 创建工程结构 45815.2.3 配置web.xml内容 46015.2.4 建立配置dwr.xml内容 46115.3 DWR入门与实例演示 46115.3.1 简单Ajax页面代码 46115.3.2 运行效果与对比 46415.3.3 DWR与直接使用XMLHttpRequest对象的比较 46815.3.4 在DWR中操纵自定义的对象 47015.3.5 查看DWR的输出日志 47715.4 dwr.xml的配置进阶 47715.4.1 dwr.xml的标准结构 47815.4.2 init标签与DWR自带的converter和creator 47915.4.3 allow标签 48315.4.4 signature标签 48415.4.5 转换器converter 48515.5 使用DWR工具库util.js 48815.5.1 页面中调用util.js 48915.5.2 使用useLoadingMessage()方法显示提示图标 49015.5.3 DWRUtil.setValue()和DWRUtil.getValue() 49515.5.4 DWRUtil.getValues和DWRUtil.setValues 49815.5.5 DWRUtil.addOptions和DWRUtil.removeAll-Options 50315.5.6 DWRUtil.addRow和DWRUtil.removeAll-Rows 50815.5.7 DWRUtil.toDescriptive-String方法 51515.6 小结 516第16章 搜索引擎综合实例：Web篇 51716.1 Web配置文件 51716.1.1 配置文件及其作用 51716.1.2 Spring配置文件 51816.1.3 DWR配置文件 51916.1.4 web.xml配置文件 52016.2 各种搜索相关Bean类 52116.2.1 产品SearchResult结果记录类 52216.2.2 产品SearchResults结果集合类 52416.2.3 产品SearchRequest检索请求类 52616.3 数据库访问SearchResultDAO类实现 52716.3.1 数码库访问类接口定义 52716.3.2 数码库访问类实现 52816.4 Lucene索引检索SearchService类实现 53016.4.1 索引检索类接口定义 53016.4.2 索引检索类实现 53116.5 前台Web页面设计 53616.5.1 数码垂直搜索主页面main.jsp 53616.5.2 数码搜索手机产品图片的显示 54216.5.3 手机产品详细信息页面detail.jsp 54316.6 实例中的问题与功能扩展 54616.7 小结 548附录 Lucene 2.4更新内容 549F1 IndexWriter的构造函数 549F2 IndexWriter的init方法 550F3 IndexWriter中的flush、commit和close 552F4 Lucene 2.4中的Segment 553F5 IndexCommit和IndexDeletion-Policy 555F6 IndexWriter中的add-Document 558F7 DocumentsWriter类的add-Document方法 559F8 DocumentsWriter的索引链 562

<<开发自己的搜索引擎>>

编辑推荐

销书升级，原书是国内第一本讲解搜索引擎开发的畅销书 超值，提供了价值上万元的大型数码产品搜索引擎开发案侧，可直接应用于项目 版本最新，采用了最新的Heritrix-1.140版、HTMLParser1.6.0版、DWR2.0.5版 实践性强，用案例的方式讲解，便于读者实践 注重原理讲解，提供了结构框图和流程图，讲解搜索引擎的原理 《开发自己的搜索引擎：Lucene+Heritrix(第2版)》在第1版的基础上做了以下改变：

重新组织了实例中开发前期的准备内容，涉及信息来源准备、基本开发环境准备、工程总体框架和配置信息管理等。

升级了内核代码版本，使用Heritrix-1.14.0版本，并增加了网络爬虫Heritrix代码工程导人和配置

的详细步骤。

升级了实例代码，解决了因来源网站内容变更而导致的部分代码无法执行的问题。

增加了对网页内容分析的概述和基本说明，便于读者理解相关代码和内容。

HTMLParser升级到1.6.0版本。

为了适应网站代码的修改，变动了其中的正则分析代码和网页解析代码。

更新了实例中与文档组织和存储相关的内容，并根据所分析网站页面内容的变化，升级了其中的部分代码。

升级交互篇和用户Web界面中的核心代码为DWR2.0.5版本，并针对目前的主流技术，对搜索引擎交互方式进行分析和介绍，还增加了代码的部分图例说明。

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>