

《Lucene+nutch搜索引擎开发》

图书基本信息

书名：《Lucene+nutch搜索引擎开发》

13位ISBN编号：9787115182166

10位ISBN编号：7115182167

出版时间：2008-8

出版社：人民邮电出版社

作者：王学松

页数：452

版权说明：本站所提供下载的PDF图书仅提供预览和简介以及在线试读，请支持正版图书。

更多资源请访问：www.tushu000.com

《Lucene+nutch搜索引擎开发》

内容概要

《Lucene+nutch搜索引擎开发》以Lucene构建搜索引擎的开发过程为主线，由浅入深，循序渐进，为读者展示如何使用Lucene开发自己的搜索引擎系统。全书内容包括搜索引擎概述和原理、Lucene部署安装、Nutch网络蜘蛛与数据获取、Lucene索引建立、Lucene检索与查询、搜索结果排序、文档分析器与中文分词、格式化文本分析、分布式搜索与缓存等。为便于读者理解搜索引擎快速开发过程，《Lucene+nutch搜索引擎开发》最后几章进行了应用实例的讲解，包括Nutch构建专题搜索、Lucene构建企业级搜索实例以及相关的整体工程性能测试。

书籍目录

- 第1篇 入门篇第1章 搜索引擎概述1.1 什么是搜索引擎1.1.1 搜索引擎与信息检索1.1.2 搜索引擎的概念1.1.3 搜索引擎的使用1.1.4 搜索引擎发展历史1.2 搜索引擎分类1.2.1 按照工作方式分类1.2.2 按照领域范围分类1.2.3 信息类型分类1.3 主流搜索引擎1.3.1 全球著名搜索引擎1.3.2 中文搜索引擎的发展历史1.3.3 著名中文搜索引擎1.3.4 其他细化搜索引擎1.4 搜索引擎评价原则1.4.1 评价指标体系1.4.2 其他评测因素1.5 搜索引擎相关资源1.5.1 搜索引擎开源项目1.5.2 搜索引擎研究网站1.5.3 搜索论坛和厂商黑板报1.6 系统运行环境准备1.6.1 Java环境安装设置1.6.2 Tomcat服务器安装1.6.3 Eclipse开发环境准备1.7 未来搜索技术前瞻1.7.1 现状存在问题1.7.2 未来发展趋势1.8 小结第2章 搜索引擎原理探秘2.1 解密搜索引擎原理2.1.1 搜索引擎技术框架2.1.2 网页信息抓取技术2.1.3 网页内容分析技术2.1.4 网页索引建立技术2.1.5 用户检索与结果排序2.1.6 网页检索工具与接口2.2 网络爬虫简单实现2.2.1 网络蜘蛛功能需求2.2.2 网络蜘蛛实现原理2.2.3 网络爬虫系统结构2.2.4 网页采集程序设计2.2.5 网页采集程序实现2.2.6 程序实现存储扩展2.3 网页分析程序实现2.3.1 网页分析功能需求2.3.2 网页分析实现原理2.3.3 网页分析系统结构2.3.4 网页分析程序设计2.3.5 文本语素分割与过滤2.4 网页索引程序实现2.4.1 网页索引功能需求2.4.2 网页索引实现原理2.4.3 网页索引程序设计2.4.4 网页索引程序实现2.5 检索程序实现2.5.1 检索功能需求2.5.2 检索实现原理2.5.3 检索程序设计2.5.4 网页检索程序实现2.6 简单搜索引擎系统2.7 小结第3章 开源搜索引擎入门3.1 开源搜索引擎简介3.1.1 Lucene系统概述3.1.2 Nutch概述3.2 Lucene全文检索系统部署3.2.1 下载Lucene系统3.2.2 Lucene部署配置3.2.3 Lucene测试运行3.3 Lucene开发实例入门3.3.1 Lucene实例功能3.3.2 Lucene开发实例3.3.3 代码实例解析3.4 Nutch开源搜索引擎部署3.4.1 Cygwin软件安装3.4.2 Nutch下载与安装3.4.3 Nutch系统环境测试3.4.4 Nutch搜索页面部署3.5 Nutch系统调试与开发3.5.1 Eclipse中加载Nutch3.5.2 Nutch工程编译与发布3.6 小结第2篇 内核揭秘篇第4章 搜索引擎数据获取4.1 网络蜘蛛原理4.1.1 体系结构设计4.1.2 访问策略与算法4.1.3 效率优化与更新4.1.4 蜘蛛访问规范4.1.5 开源蜘蛛简介4.2 Nutch网络蜘蛛4.2.1 Nutch网络蜘蛛概述4.2.2 Nutch抓取模式分类4.2.3 抓取测试站点建立4.3 Nutch局域网抓取4.3.1 本地下载准备4.3.2 启动下载过程4.3.3 下载过程解析4.3.4 下载多个网站4.4 Nutch互联网抓取4.4.1 下载列表获取4.4.2 下载大量网站4.5 Nutch抓取比较4.6 Nutch结果检测4.6.1 网页内容检索4.6.2 使用Readdb获取摘要4.6.3 使用SegRead读取分段4.6.4 Luke工具使用4.7 Nutch配置文件解析4.8 Heritrix网络蜘蛛4.8.1 Heritrix概述4.8.2 Heritrix体系结构4.8.3 Heritrix安装与使用4.9 小结第5章 搜索引擎信息索引5.1 文档索引原理5.1.1 索引概述5.1.2 索引基本结构5.1.3 倒排索引原理5.1.4 索引分类5.1.5 高性能索引5.2 Lucene索引器5.2.1 Lucene索引介绍5.2.2 Lucene索引结构5.2.3 多文件索引结构5.2.4 复合索引结构5.3 Lucene索引实例5.3.1 索引创建代码解析5.3.2 索引创建器 (IndexWriter) 5.3.3 索引管理器 (IndexReader) 5.3.4 索引修改器 (IndexModifier) 5.3.5 索引分析器 (Analyzer) 5.4 Lucene索引操作5.4.1 添加文本文件索引5.4.2 创建Lucene增量索引5.4.3 使用索引项删除文档5.4.4 使用编号删除文档5.4.5 压缩文档编号5.4.6 索引文档更新5.5 Lucene索引高级特性5.5.1 选择索引域类型5.5.2 索引参数优化5.5.3 使用磁盘索引5.5.4 使用内存索引5.5.5 同步与锁机制5.6 Lucene高级应用实例5.6.1 创建本地搜索的索引5.6.2 索引数据库记录5.6.3 索引优化与合并5.7 Nutch中的Lucene索引5.8 小结第6章 搜索引擎查询处理6.1 信息查询原理6.1.1 信息查询概述6.1.2 查询基本流程6.1.3 查询结果显示6.1.4 高性能查询6.2 Lucene查询概述6.2.1 Lucene查询操作基础6.2.2 Lucene查询实例入门6.2.3 查询工具IndexSearcher类6.2.4 查询封装Query类6.2.5 查询分析器QueryParser类6.2.6 查询结果集Hits类6.3 Lucene基本查询6.3.1 Lucene查询Query对象6.3.2 最小项查询TermQuery6.3.3 区间范围搜索RangeQuery6.3.4 逻辑1/4组合搜索BooleanQuery6.3.5 字符串前缀搜索PrefixQuery6.3.6 短语搜索PhraseQuery6.3.7 模糊搜索FuzzyQuery6.3.8 通配符搜索WildcardQuery6.3.9 位置跨度搜索SpanQuery6.4 Lucene高级查询6.4.1 索引内存检索6.4.2 多关键字跨域检索6.4.3 多检索器跨索引检索6.5 Nutch中的Lucene查询6.6 小结第7章 搜索引擎结果排序7.1 搜索引擎文档排序原理7.1.1 传统检索排序技术7.1.2 向量模型排序局限7.1.3 搜索引擎相关性排序7.1.4 链接分析PageRank原理7.1.5 搜索引擎排序流程7.2 Lucene检索排序7.2.1 Lucene相关性因素7.2.2 Lucene相关排序流程7.2.3 Lucene排序计算体系7.2.4 Lucene排序控制方法7.3 文档Boost加权排序7.3.1 Lucene中Boost介绍7.3.2 Boost值全文档排序7.3.3 Boost值文档域排序7.3.4

BoostingTermQuery排序7.4 Sort对象检索排序7.4.1 Sort对象概述7.4.2 Sort对象相关性排序7.4.3 Sort对象文档编号排序7.4.4 Sort对象独立域排序7.4.5 Sort对象联合域排序7.4.6 Sort对象逆向排序7.5 Lucene相关性公式7.5.1 Lucene评分结果分析7.5.2 Lucene排序公式7.5.3 其他动态排序因子7.6 Lucene自定义排序7.6.1 自定义排序比较接口7.6.2 自定义排序接口类实例7.6.3 自定义排序结果测试实例7.6.4 自定义排序测试结果7.7 Nutch中的结果排序7.7.1 Nutch排序因素7.7.2 Nutch链接分析7.7.3 Nutch相关度计算7.8 小结第8章 文档分析器与中文分词8.1 文档分析与中文分词原理8.1.1 文档分析预处理概述8.1.2 文档分析基本流程8.1.3 中文分析处理中的分词8.2 Lucene分析器内核原理8.2.1 Lucene分析器原理8.2.2 Analysis包简介8.2.3 Analyzer类的组合结构8.2.4 JavaCC构造分析器8.2.5 StopAnalyzer内核代码分析8.2.6 StandardAnalyzer内核代码分析8.3 Lucene分析器应用模式8.3.1 使用默认分析器建立索引8.3.2 使用多种分析器建立索引8.3.3 使用分析器检索查询8.4 Lucene主要分析器应用实例8.4.1 停用词分析器StopAnalyzer8.4.2 标准分析器StandardAnalyzer8.4.3 简单分析器SimpleAnalyzer8.4.4 空格分析器WhitespaceAnalyzer8.4.5 关键字分析器KeywordAnalyzer8.5 TokenStream分词器内核分析8.5.1 Tokenizer分词器8.5.2 标准分词器StandardTokenizer8.5.3 字符分词器CharTokenizer8.5.4 空格分词器WhiteSpaceTokenizer8.5.5 字母分词器LetterTokenizer8.5.6 小写分词器LowerCaseTokenizer8.6 TokenStream过滤器内核分析8.6.1 TokenFilter过滤器8.6.2 标准过滤器StandardFilter8.6.3 停用词过滤器StopFilter8.6.4 小写过滤器LowerCaseFilter8.6.5 长度过滤器LengthFilter8.6.6 词干过滤器PorterStemFilter8.7 Lucene中文分词8.7.1 中文分词基本原理方法8.7.2 StandardAnalyzer分析器中文处理8.7.3 CJKAnalyzer中文分析器8.7.4 ChineseAnalyzer中文分析器8.7.5 IK_CAnalyzer中文分析器8.7.6 中科院ICTCLAS中文分词8.7.7 JE中文分词8.7.8 中文分词问题8.8 Nutch分词和预处理8.8.1 Nutch分析器8.8.2 Nutch中文分词8.9 小结第9章 搜索引擎文本分析9.1 非结构化文本简介9.1.1 非结构化文本概述9.1.2 非结构化文本检索9.2 HTML文档分析9.2.1 主流HTML文档分析器9.2.2 HTMLParser安装配置9.2.3 HTMLParser的框架结构9.3 HTMLParser应用实例9.3.1 HTMLParser功能模式9.3.2 HTMLParser内容解析方式9.3.3 Visitor模式正文解析9.3.4 Filter模式简单链接提取9.3.5 Filter模式搜索链接提取9.3.6 Lexer模式遍历文档9.4 PDF文档分析9.4.1 常用的PDF处理包9.4.2 PDFBox安装配置9.5 PDFBox应用实例9.5.1 PDFBox提取文档内容9.5.2 PDFBox文档内容索引9.6 Office文档分析9.6.1 常用Office文档处理包9.6.2 使用POI安装与配置9.6.3 POI原理与接口介绍9.7 POI分析Office文档实例9.7.1 POI处理Excel文档9.7.2 POI处理Word文档9.8 XML文档分析9.8.1 主流XML文档分析器9.8.2 JDOM分析器安装配置9.8.3 xerces分析器安装配置9.9 XML解析应用实例9.9.1 使用JDOM分析XML文档9.9.2 使用xerces分析XML文档9.10 Nutch文档处理9.11 小结第10章 分布式搜索与缓存10.1 分布式检索与缓存10.1.1 分布式搜索引擎现状10.1.2 分布式搜索引擎原理10.1.3 搜索引擎缓存现状10.1.4 搜索引擎缓存原理10.2 Nutch与分布式检索10.2.1 Google分布式文件系统10.2.2 MapReduce系统介绍10.2.3 Hadoop分布式文件系统10.2.4 Nutch分布式文件系统10.2.5 Nutch分布式检索概述10.2.6 Nutch分布式检索器10.3 Lucene分布式检索10.3.1 Socket通信基础10.3.2 Lucene索引服务器10.4 Nutch与搜索缓存10.5 开源系统缓存系统10.6 小结第3篇 实战篇第11章 Nutch专题搜索引擎实例11.1 专题搜索需求分析11.1.1 专题搜索功能需求11.1.2 专题搜索用例分析11.2 构建Nutch基础搜索引擎11.2.1 Nutch搜索功能分析11.2.2 信息下载功能测试11.2.3 Nutch基础Web检索11.2.4 Web用户页面修改11.3 专题搜索系统设计11.3.1 系统框架设计11.3.2 选择开发工具组件11.4 专题关键词管理11.4.1 专题关键词策略11.4.2 关键词存储设计11.4.3 关键词管理程序11.5 专题资源发现11.5.1 专题网页链接发现11.5.2 专题资源网站提取11.6 专题信息下载11.6.1 批量信息下载11.6.2 信息自动下载11.7 专题信息分析与索引11.7.1 网页信息分析11.7.2 创建索引11.8 检索辅助功能11.8.1 相关词推荐11.8.2 检索词高亮显示11.8.3 检索结果翻页11.9 小结第12章 Lucene实现企业搜索实例12.1 企业搜索需求分析12.1.1 企业搜索需求概述12.1.2 企业搜索用例分析12.2 企业级搜索系统设计12.2.1 系统框架设计12.2.2 Lucene检索框架12.3 企业级搜索系统设计12.3.1 创建Lucene工程12.3.2 全文检索索引生成12.3.3 全文检索检索页面12.4 数据引擎设计12.4.1 数据库数据管理12.4.2 非结构化文档12.5 企业信息索引12.5.1 数据索引建立12.5.2 信息检索代码12.5.3 检索Web代码12.5.4 检索结果测试12.6 小结

章节摘录

第1篇 入门篇 第1章 搜索引擎概述 1.1 什么是搜索引擎 搜索引擎是一款特别的软件系统，能够从互联网上自动搜集信息，并为用户提供查询服务。搜索引擎对原始文档进行了一系列的整理和处理。用户的查询结果是搜索引擎按照某种规则计算获得的。搜索引擎为网民提供了资源查找和导航的有效手段。

1.1.1 搜索引擎与信息检索 搜索引擎并不是一个完全创新的系统，而是借鉴了以往全文检索系统和网络软件系统开发而成的。搜索引擎采用了以往产品的很多技术和思路，尤其是继承了很多信息检索系统的技术和方法。互联网搜索引擎在继承历史技术的同时，针对互联网信息处理的特点，开发出了互联网信息查找工具。

编辑推荐

《Lucene+nutch搜索引擎开发》适合对搜索引擎开发有兴趣的读者阅读，包括搜索引擎开发的初学者、高等院校、信息专业学生、从事搜索开发的程序设计人员等。入门：引导读者快速掌握（Lucene和nutch的使用方法）；揭秘：深度剖析搜索引擎内核；实战：手把手带您构建企业级搜索引擎；推荐：Web开发专家强烈推荐。

互联网搜索的使用水平可以反映全民的信息处理能力，几年前有研究发现美国用户比欧洲用户的互联网使用水平领先半年左右，主要是根据谁搜索时平均使用的关键词的个数多。中文用户的搜索使用水平相对于西文用户目前仍然处于比较初级的阶段，而中文网站搜索功能的缺失也是一个重要的因素。

网站拥有了较多内容后，最先会考虑基于目录的内容分类，以解决信息快速定位的问题，随着容量的进一步增加，很多内容在发表之后就很快被湮没，成为“信息孤岛”，而不断加深的目录结构也会让用户逐渐失去耐心，这时，关键词检索的优势就体现出来了：关键词检索可以让处于“信息孤岛”状态的内容以一种更直接的方法提供给用户；和基于目录/分类的树形结构不同，基于关键词检索还可以让内容之间实现网状的关联结构，从而大大提高信息的引用密度。

基于传统数据库的关键词检索由于性能问题让很多网站放弃了搜索功能，问题的解决归根结底还是需要一个全文引擎。而Lucene开源引擎的出现让这种原来被少数公司掌握的技术得到了迅速的普及，这里应该再次感谢引擎的核心贡献者Doug Cutting先生，同时也希望有更多的中文开发人员能积极投入：到Lucene的相关项目开发中去，尤其在中文和其他亚洲双字节语言处理方面的问题。

Lucene也是我学习的第一个Java程序，当初是通过jdb一行行debug了解其中的原理和机制的，非常高兴有这样一本专门的参考书出现，它无疑会为开发人员了解并更快掌握全文检索技术节省大量的时间。

精彩短评

- 1、大量的重复无用代码....
- 2、选择性的看了部分。对Lucene中的信息索引，查询处理，结果排序及中文分词有了大致的认识。版本太老了，基本上就是API的介绍。而且有的函数已经被取消。适合入门。
<http://www.xmind.net/m/Dec6/>
- 3、讲的比较系统，但附带的代码有点问题
- 4、写的不是太好
- 5、国内垃圾书几大技术特点：1、某教授带领多人合著的；2、代码采用中文注释的；3、不注重代码排版的；4、采用Windows系统，图片都是Windows弹出窗口和对话框的；5、官话连篇的
- 6、作为入门书，对搜索引擎的了解。
- 7、Lucene太老了。Nutch开发讲的太少
- 8、还没看完。。。
- 9、倒排索引结构，boolean模型，向量模型。
- 10、由于搜索引擎开发门槛在那里，作者的实践经验在那里，虽然写的还不如百度百科，虽然在08年的前沿知识在现在已经陈旧，但还是给3星安慰一把。关注了推荐的博客、网站，其他没有帮助。耗时：40min
- 11、不能与时俱进啊，很多API都过时了唉。看来还得自己多动手才是。
- 12、读研时从图书馆借来粗粗翻过，建议还不如直接看官网的帮助文档更有用。
- 13、对我等菜鸟帮助挺大的，如果是高手的话没必要看直接源代码走起

精彩书评

- 1、最近做的东西有相关nutch和lucene的内容，其实这本书貌似nutch的东西没有讲很多，版本也比较老了，还不如直接网上搜索来的快，lucene倒是讲了很多，不过基本都是api的介绍，可能这样看起来比直接看文档舒服点。原理的方面也是基本的介绍了下，说的不多也不太深入。感觉如果要用lucene和nutch做东西就直接网上找答案吧，毕竟已经更新好几代了，这书里面的有些函数都被取消了。买回来当本入门书籍就可以了。
- 2、买了有段时间了，最近刚读完，觉得还好吧，挺系统的。没有具体调试过上面的代码，不过看书主要看原理，代码也不那么重要。
- 3、买了这本书，直接看这几天一直困惑自己的中文分词~~前面介绍了一大段中文分词的基本概要，和lucene的分析器后面nutch的分析器只是简单的介绍了几个类，nutch中文分词只用了200字左右。书中也没用很系统的介绍nutch如何实现中文分词，~~后面的案例也只是简单的单字切分。另外nutch最出彩的插件也没提到过~~

章节试读

1、《Lucene+nutch搜索引擎开发》的笔记-第65页

这书写的还不错，就是到目前为止版本太老了，没有更新。期待作者的更新版本。

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:www.tushu000.com