<<解密搜索引擎技术实战：Lucene >>

图书基本信息

书名：<<解密搜索引擎技术实战：Lucene & Java精华版(第2版)>>

13位ISBN编号：9787121217326

10位ISBN编号：7121217325

出版时间：2013-11-29

作者：罗刚

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>

内容概要

从实用的角度出发，全面介绍了搜索引擎相关技术。

作者简介

猎兔搜索创始人

书籍目录

Lucene开发实践	1
第1章 搜索引擎总体结构	2
1.1 为什么要做搜索引擎	2
1.1.1 比价搜索	3
1.2 搜索引擎基本模块	3
1.3 开发环境	4
1.4 搜索引擎工作原理	5
1.4.1 网络爬虫	6
1.4.2 全文索引	6
1.4.3 搜索用户界面	8
1.4.4 计算框架	9
1.4.5 文本挖掘	10
1.5 算法基础	11
1.5.1 折半查找	11
1.5.2 排序	11
1.6 软件工具	14
1.7 本章小结	14
1.8 术语表	14
第2章 自己动手写全文检索	18
2.1 构建索引	18
2.2 生成索引文件	19
2.3 读入索引文件	19
2.4 查询	19
2.4.1 按相关度排序	21
2.5 有限状态机	23
2.5.1 运算	23
2.5.2 编辑距离有限状态机	24
2.6 本章小结	25
第3章 Lucene原理与应用	26
3.1 Lucene快速入门	26
3.1.1 创建索引	26
3.1.2 查询索引库	27
3.1.3 创建文档索引	29
3.1.4 查询文档索引	29
3.2 创建和维护索引库	30
3.2.1 设计索引库结构	30
3.2.2 创建索引库	31
3.2.3 向索引库中添加索引文档	33
3.2.4 删除索引库中的索引文档	36
3.2.5 更新索引库中的索引文档	37
3.2.6 关闭索引库	38
3.2.7 索引的优化与合并	38
3.2.8 灵活索引	39
3.2.9 索引文件格式	40
3.2.10 定制索引存储结构	43

- 3.2.11 写索引集成到爬虫 48
- 3.2.12 多线程写索引 51
- 3.2.13 分发索引 54
- 3.2.14 修复索引 57
- 3.3 查找索引库 57
 - 3.3.1 查询过程 57
 - 3.3.2 常用查询 60
 - 3.3.3 基本词查询 61
 - 3.3.4 模糊匹配 62
 - 3.3.5 布尔查询 63
 - 3.3.6 短语查询 65
 - 3.3.7 跨度查询 66
 - 3.3.8 FieldScoreQuery 70
 - 3.3.9 排序 74
 - 3.3.10 使用Filter筛选搜索结果 79
 - 3.3.11 使用Collector筛选搜索结果 80
 - 3.3.12 遍历索引库 82
 - 3.3.13 关键词高亮显示 86
 - 3.3.14 列合并 88
 - 3.3.15 关联内容(BlockJoinQuery) 90
 - 3.3.16 查询大容量索引 95
- 3.4 读写并发 96
- 3.5 Lucene深入介绍 97
 - 3.5.1 整体结构 97
 - 3.5.2 索引原理 98
 - 3.5.3 文档值 103
- 3.6 查询语法与解析 106
 - 3.6.1 JavaCC 107
 - 3.6.2 简单的查询解析器 119
 - 3.6.3 灵活的查询解析器 120
- 3.7 查询原理 126
 - 3.7.1 布尔匹配 126
 - 3.7.2 相关性 127
- 3.8 分析文本 130
 - 3.8.1 Analyzer 130
 - 3.8.2 TokenStream 137
 - 3.8.3 定制Tokenizer 139
 - 3.8.4 重用Tokenizer 141
 - 3.8.5 有限状态转换 141
 - 3.8.6 索引数值列 142
 - 3.8.7 检索结果排序 145
 - 3.8.8 处理价格 146
- 3.9 Lucene中的压缩算法 146
 - 3.9.1 变长压缩 147
 - 3.9.2 PForDelta 149
 - 3.9.3 VSEncoding 152
 - 3.9.4 前缀压缩 153

- 3.9.5 差分编码 155
- 3.9.6 静态索引裁剪 157
- 3.10 搜索中文 157
 - 3.10.1 Lucene切分原理 160
 - 3.10.2 Lucene中的Analyzer 161
 - 3.10.3 自己写Analyzer 164
 - 3.10.4 Lietu中文分词 167
 - 3.10.5 字词混合索引 167
- 3.11 索引数据库中的文本 172
- 3.12 优化使用Lucene 174
 - 3.12.1 系统优化 174
 - 3.12.2 查询优化 175
 - 3.12.3 实现时间加权排序 178
 - 3.12.4 词性标注 182
- 3.13 检索模型 185
 - 3.13.1 向量空间模型 186
 - 3.13.2 DFR 192
 - 3.13.3 BM25概率模型 199
 - 3.13.4 统计语言模型 205
 - 3.13.5 隐含语义索引 206
 - 3.13.6 学习评分 207
 - 3.13.7 查询与相关度 208
 - 3.13.8 提高相关度 208
 - 3.13.9 使用Payload调整相关性 209
 - 3.13.10 索引统计 214
- 3.14 实时搜索 216
- 3.15 概念搜索 218
 - 3.15.1 发现同义词 219
 - 3.15.2 垂直领域同义词 223
 - 3.15.3 同义词扩展 224
- 3.16 本章小结 228
- 3.17 术语表 228
- 第4章 搜索引擎用户界面 230
 - 4.1 实现Lucene搜索 230
 - 4.1.1 测试搜索功能 230
 - 4.1.2 加载索引 232
 - 4.2 手机搜索界面 233
 - 4.3 搜索页面设计 236
 - 4.3.1 Struts2实现的搜索界面 236
 - 4.3.2 实现翻页 239
 - 4.4 实现搜索接口 241
 - 4.4.1 编码识别 241
 - 4.4.2 布尔搜索 245
 - 4.4.3 指定范围搜索 245
 - 4.4.4 搜索结果排序 247
 - 4.4.5 索引缓存与更新 248
 - 4.5 实现分类统计视图 255

4.5.1 单值列分类统计	262
4.6 实现相似文档搜索	263
4.7 实现AJAX搜索联想词	265
4.7.1 估计查询词的文档频率	265
4.7.2 搜索联想词总体结构	266
4.7.3 服务器端处理	267
4.7.4 浏览器端处理	272
4.7.5 拼音提示	274
4.7.6 部署总结	275
4.8 推荐搜索词	276
4.8.1 挖掘相关搜索词	276
4.8.2 使用多线程计算相关搜索词	278
4.9 拼音搜索	280
4.10 集成其他功能	280
4.10.1 拼写检查	280
4.10.2 分类统计	285
4.10.3 相关搜索	292
4.10.4 再次查找	295
4.10.5 搜索日志	295
4.11 查询分析	297
4.11.1 历史搜索词记录	297
4.11.2 日志信息过滤	298
4.11.3 信息统计	299
4.11.4 挖掘日志信息	301
4.11.5 查询词意图分析	302
4.12 部署网站	302
4.12.1 部署到Web服务器	302
4.12.2 防止攻击	305
4.13 本章小结	309
第5章 使用Solr实现企业搜索	311
5.1 Solr简介	312
5.1.1 使用Solr	312
5.2 Solr基本用法	313
5.2.1 Solr服务器端的配置与中文支持	313
5.2.2 数据类型	319
5.2.3 解析器	320
5.2.4 把数据放进Solr	320
5.2.5 删除数据	325
5.2.6 查询语法	326
5.3 使用SolrJ	327
5.3.1 Solr客户端与搜索界面	327
5.3.2 Solr索引库的查找	329
5.3.3 分类统计	333
5.3.4 高亮	335
5.3.5 同义词	337
5.3.6 嵌入式Solr	337
5.3.7 索引分发	338

- 5.3.8 Solr搜索优化 341
- 5.4 SolrItas 344
- 5.5 从FAST Search移植到Solr 344
- 5.6 简单应用 346
- 5.7 Solr扩展与定制 346
 - 5.7.1 插件 346
 - 5.7.2 Solr中字词混合索引 346
 - 5.7.3 相关检索 348
 - 5.7.4 搜索结果去重 350
 - 5.7.5 定制输入输出 354
 - 5.7.6 聚类 359
 - 5.7.7 分布式搜索 360
 - 5.7.8 分布式索引 364
 - 5.7.9 SolrJ查询分析器 366
 - 5.7.10 扩展SolrJ 375
 - 5.7.11 扩展Solr 376
 - 5.7.12 日文搜索 379
 - 5.7.13 查询Web图 380
- 5.8 SolrNet 383
 - 5.8.1 使用SolrNet实现全文搜索 383
 - 5.8.2 实现原理 387
 - 5.8.3 扩展SolrNet 388
- 5.9 Solr的其它客户端 393
 - 5.9.1 Solr的PHP客户端 394
- 5.10 为网站增加搜索功能 397
- 5.11 手机客户端 397
- 5.12 Solr原理 398
 - 5.12.1 支持Solr的中文分词 398
 - 5.12.2 缓存技术 399
- 5.13 本章小结 399
- 第6章 地图搜索 401
 - 6.1 Solr 401
- 第7章 视频搜索 402
- 第8章 垂直搜索 403
 - 8.1 自动化网站 403
 - 8.2 招聘行业网站 403
 - 8.2.1 网络爬虫 403
 - 8.2.2 全文中文引擎 403
 - 8.2.3 Email地址人工添加简易工具 404
 - 8.2.4 职位推荐 404
 - 8.2.5 用户权限 404

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>