

<<解密搜索引擎技术实战(第二版)>>

图书基本信息

书名：<<解密搜索引擎技术实战(第二版)>>

13位ISBN编号：9788121217323

10位ISBN编号：8121217326

出版时间：2013-11-30

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>

<<解密搜索引擎技术实战(第二版)>>

内容概要

总结搜索引擎相关理论与实际解决方案，并给出了 java 实现，其中利用了流行的开源项目 lucene 和 solr，而且还包括原创的实现。

作者简介

猎兔搜索创始人

书籍目录

第1章 搜索引擎总体结构	1
1.1 搜索引擎基本模块	2
1.2 开发环境	2
1.3 搜索引擎工作原理	4
1.3.1 网络爬虫	4
1.3.2 全文索引结构与Lucene实现	4
1.3.3 搜索用户界面	7
1.3.4 计算框架	8
1.3.5 文本挖掘	9
1.4 本章小结	10
第2章 网络爬虫的原理与应用	11
2.1 爬虫的基本原理	12
2.2 爬虫架构	14
2.2.1 基本架构	15
2.2.2 分布式爬虫架构	17
2.2.3 垂直爬虫架构	18
2.3 抓取网页	19
2.3.1 下载网页的基本方法	20
2.3.2 网页更新	23
2.3.3 抓取限制应对方法	25
2.3.4 URL地址提取	27
2.3.5 抓取JavaScript动态页面	28
2.3.6 抓取即时信息	31
2.3.7 抓取暗网	32
2.3.8 信息过滤	33
2.3.9 最好优先遍历	38
2.4 存储URL地址	40
2.4.1 BerkeleyDB	40
2.4.2 布隆过滤器	42
2.5 并行抓取	45
2.5.1 多线程爬虫	45
2.5.2 垂直搜索的多线程爬虫	47
2.5.3 异步I/O	49
2.6 RSS抓取	52
2.7 抓取FTP	54
2.8 下载图片	55
2.9 图像的OCR识别	56
2.9.1 图像二值化	57
2.9.2 切分图像	59
2.9.3 SVM分类	62
2.10 Web结构挖掘	66
2.10.1 存储Web图	66
2.10.2 PageRank算法	70
2.10.3 HITs算法	77
2.10.4 主题相关的PageRank	81

- 2.11 部署爬虫 82
- 2.12 本章小结 82
- 第3章 索引内容提取 86
 - 3.1 从HTML文件中提取文本 87
 - 3.1.1 字符集编码 87
 - 3.1.2 识别网页的编码 90
 - 3.1.3 网页编码转换为字符串编码 93
 - 3.1.4 使用HTMLParser实现定向抓取 93
 - 3.1.5 使用正则表达式提取数据 98
 - 3.1.6 结构化信息提取 99
 - 3.1.7 网页的DOM结构 102
 - 3.1.8 使用NekoHTML提取信息 104
 - 3.1.9 网页去噪 109
 - 3.1.10 网页结构相似度计算 114
 - 3.1.11 提取标题 116
 - 3.1.12 提取日期 117
 - 3.2 从非HTML文件中提取文本 117
 - 3.2.1 提取标题的一般方法 118
 - 3.2.2 PDF文件 122
 - 3.2.3 Word文件 126
 - 3.2.4 Rtf文件 127
 - 3.2.5 Excel文件 138
 - 3.2.6 PowerPoint文件 141
 - 3.3 提取垂直行业信息 141
 - 3.3.1 医疗行业 141
 - 3.3.2 旅游行业 142
 - 3.4 流媒体内容提取 143
 - 3.4.1 音频流内容提取 143
 - 3.4.2 视频流内容提取 147
 - 3.5 存储提取内容 148
 - 3.6 本章小结 149
- 第4章 中文分词原理与实现 151
 - 4.1 Lucene中的中文分词 152
 - 4.1.1 Lucene切分原理 152
 - 4.1.2 Lucene中的Analyzer 154
 - 4.1.3 自己写Analyzer 155
 - 4.1.4 Lietu中文分词 158
 - 4.2 查找词典算法 158
 - 4.2.1 标准Trie树 159
 - 4.2.2 三叉Trie树 162
 - 4.3 中文分词的原理 166
 - 4.4 中文分词流程与结构 170
 - 4.5 形成切分词图 171
 - 4.6 概率语言模型的分词方法 177
 - 4.7 N元分词方法 181
 - 4.8 新词发现 183
 - 4.9 未登录词识别 185

- 4.10 词性标注 186
 - 4.10.1 隐马尔可夫模型 189
 - 4.10.2 基于转换的错误学习方法 197
- 4.11 平滑算法 199
- 4.12 机器学习的方法 203
 - 4.12.1 最大熵 204
 - 4.12.2 条件随机场 207
- 4.13 有限状态机 207
- 4.14 本章小结 214
- 第5章 让搜索引擎理解自然语言 216
 - 5.1 停用词表 217
 - 5.2 句法分析树 219
 - 5.3 相似度计算 223
 - 5.4 文档排重 226
 - 5.4.1 语义指纹 227
 - 5.4.2 SimHash 230
 - 5.4.3 分布式文档排重 240
 - 5.5 中文关键词提取 241
 - 5.5.1 关键词提取的基本方法 241
 - 5.5.2 HITS算法应用于关键词提取 243
 - 5.5.3 从网页中提取关键词 245
 - 5.6 相关搜索词 246
 - 5.6.1 挖掘相关搜索词 246
 - 5.6.2 使用多线程计算相关搜索词 248
 - 5.7 信息提取 249
 - 5.8 拼写检查与建议 254
 - 5.8.1 模糊匹配问题 257
 - 5.8.2 英文拼写检查 260
 - 5.8.3 中文拼写检查 261
 - 5.9 自动摘要 264
 - 5.9.1 自动摘要技术 264
 - 5.9.2 自动摘要的设计 265
 - 5.9.3 基于篇章结构的自动摘要 271
 - 5.9.4 Lucene中的动态摘要 271
 - 5.10 文本分类 274
 - 5.10.1 特征提取 276
 - 5.10.2 中心向量法 280
 - 5.10.3 朴素贝叶斯 282
 - 5.10.4 支持向量机 291
 - 5.10.5 多级分类 299
 - 5.10.6 规则方法 300
 - 5.10.7 网页分类 303
 - 5.11 自动聚类 304
 - 5.11.1 聚类的定义 304
 - 5.11.2 K均值聚类方法 304
 - 5.11.3 K均值实现 306
 - 5.11.4 深入理解DBScan算法 310

5.11.5 使用DBScan算法聚类实例	312
5.12 拼音转换	314
5.13 概念搜索	315
5.14 多语言搜索	323
5.15 跨语言搜索	324
5.16 情感识别	325
5.16.1 确定词语的褒贬倾向	328
5.16.2 实现情感识别	329
5.16.3 用户协同过滤	331
5.17 本章小结	332
第6章 Lucene原理与应用	334
6.1 Lucene深入介绍	335
6.1.1 常用查询	335
6.1.2 查询语法与解析	336
6.1.3 查询原理	340
6.1.4 使用Filter筛选搜索结果	341
6.1.5 遍历索引库	341
6.1.6 索引数值列	343
6.2 Lucene中的压缩算法	346
6.2.1 变长压缩	346
6.2.2 PForDelta	348
6.2.3 前缀压缩	351
6.2.4 差分编码	352
6.2.5 设计索引库结构	354
6.3 创建和维护索引库	355
6.3.1 创建索引库	355
6.3.2 向索引库中添加索引文档	356
6.3.3 删除索引库中的索引文档	359
6.3.4 更新索引库中的索引文档	359
6.3.5 索引的合并	360
6.3.6 索引文件格式	360
6.3.7 分发索引	363
6.3.8 修复索引	366
6.4 查找索引库	366
6.5 读写并发控制	367
6.6 优化使用Lucene	368
6.6.1 索引优化	368
6.6.2 查询优化	369
6.6.3 实现时间加权排序	372
6.6.4 实现字词混合索引	375
6.6.5 重用Tokenizer	380
6.6.6 定制Tokenizer	380
6.7 检索模型	382
6.7.1 向量空间模型	382
6.7.2 BM25概率模型	387
6.7.3 统计语言模型	392
6.8 查询大容量索引	394

<<解密搜索引擎技术实战(第二版)>>

- 6.9 实时搜索 395
- 6.10 本章小结 396
- 第7章 搜索引擎用户界面 397
 - 7.1 实现Lucene搜索 398
 - 7.2 搜索页面设计 399
 - 7.2.1 Struts2实现的搜索界面 399
 - 7.2.2 翻页组件 400
 - 7.3 实现搜索接口 401
 - 7.3.1 编码识别 401
 - 7.3.2 布尔搜索 404
 - 7.3.3 指定范围搜索 405
 - 7.3.4 搜索结果排序 406
 - 7.3.5 搜索页面的索引缓存与更新 406
 - 7.4 历史搜索词记录 409
 - 7.5 实现关键词高亮显示 410
 - 7.6 实现分类统计视图 412
 - 7.7 实现相似文档搜索 417
 - 7.8 实现AJAX搜索联想词 419
 - 7.8.1 估计查询词的文档频率 419
 - 7.8.2 搜索联想词总体结构 420
 - 7.8.3 服务器端处理 420
 - 7.8.4 浏览器端处理 421
 - 7.8.5 服务器端改进 426
 - 7.8.6 拼音提示 429
 - 7.8.7 部署总结 430
 - 7.9 集成其他功能 430
 - 7.9.1 拼写检查 430
 - 7.9.2 分类统计 431
 - 7.9.3 相关搜索 433
 - 7.9.4 再次查找 436
 - 7.9.5 搜索日志 436
 - 7.10 搜索日志分析 438
 - 7.10.1 日志信息过滤 438
 - 7.10.2 信息统计 440
 - 7.10.3 挖掘日志信息 442
 - 7.11 本章小结 443
- 第8章 使用Solr实现企业搜索 444
 - 8.1 Solr简介 445
 - 8.2 Solr基本用法 446
 - 8.2.1 Solr服务器端的配置与中文支持 447
 - 8.2.2 把数据放进Solr 452
 - 8.2.3 删除数据 454
 - 8.2.4 Solr客户端与搜索界面 455
 - 8.2.5 Solr索引库的查找 457
 - 8.2.6 索引分发 461
 - 8.2.7 Solr搜索优化 464
 - 8.3 从FAST Search移植到Solr 467

8.4 Solr扩展与定制	468
8.4.1 Solr中字词混合索引	469
8.4.2 相关检索	470
8.4.3 搜索结果去重	472
8.4.4 定制输入输出	476
8.4.5 分布式搜索	480
8.4.6 SolrJ查询分析器	481
8.4.7 扩展SolrJ	489
8.4.8 扩展Solr	490
8.4.9 查询Web图	494
8.5 Solr的.NET客户端	496
8.6 Solr的PHP客户端	502
8.7 本章小结	505
第9章 地理信息系统案例分析	506
9.1 新闻提取	508
9.2 POI信息提取	512
9.2.1 提取主体	517
9.2.2 提取地区	519
9.2.3 指代消解	520
9.3 本章小结	522
第10章 户外活动搜索案例分析	523
10.1 爬虫	524
10.2 信息提取	525
10.3 活动分类	528
10.4 搜索	529
10.5 本章小结	530
参考资料	531

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>