

<<解密搜索引擎技术实战>>

图书基本信息

书名：<<解密搜索引擎技术实战>>

13位ISBN编号：9787621096407

10位ISBN编号：7621096403

出版时间：2011-5-13

出版时间：电子工业出版社

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>>

内容概要

本书是猎兔搜索开发团队的软件研发和教学实践的经验汇总。

本书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。

主要包括总体介绍部分、爬虫部分、自然语言处理部分、全文检索部分以及相关案例分析。

爬虫部分介绍了网页遍历方法和如何实现增量抓取。
并介绍了从网页等各种格式的文档中提取主要内容的方法。

自然语言处理部分从统计机器学习的原理出发，包括了中文分词与词性标注的理论与实现以及在搜索引擎中的实用等细节。

同时对文档排重、文本分类、自动聚类、句法分析树、拼写检查等自然语言处理领域的经典问题做出了深入浅出的介绍并总结了实现方法。

在全文检索部分，结合Lucene3.0介绍了搜索引擎的原理与进展。
用简单的例子介绍了Lucene的最新应用方法。
包括完整的搜索实现过程：从完成索引到搜索用户界面的实现。
本书还进一步介绍了实现准实时搜索的方法。

展示了Solr1.4版本的用法以及实现分布式搜索服务集群的方法。

最后介绍了在地理信息系统领域和户外活动搜索领域的应用。

<<解密搜索引擎技术实战>>

作者简介

猎兔搜索(<http://www.lietu.com>)创始人，当前猎兔搜索在北京和上海均设有研发部。带领猎兔搜索技术开发团队先后开发出猎兔中文分词系统、猎兔信息提取系统、猎兔智能垂直搜索系统以及网络信息监测系统等，实现互联网信息的采集、过滤、搜索和实时监测。

<<解密搜索引擎技术实战>>

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	3
1.3 搜索引擎工作原理	4
1.3.1 网络爬虫	5
1.3.2 全文索引结构与Lucene实现	5
1.3.3 搜索用户界面	10
1.3.4 计算框架	10
1.3.5 文本挖掘	11
1.4 本章小结	12
第2章 网络爬虫的原理与应用	13
2.1 爬虫的基本原理	13
2.2 爬虫架构	16
2.2.1 基本架构	16
2.2.2 分布式爬虫架构	19
2.2.3 垂直爬虫架构	20
2.3 抓取网页	20
2.3.1 下载网页的基本方法	21
2.3.2 网页更新	25
2.3.3 抓取限制应对方法	27
2.3.4 URL地址提取	29
2.3.5 抓取JavaScript动态页面	29
2.3.6 抓取即时信息	32
2.3.7 抓取暗网	33
2.3.8 信息过滤	34
2.3.9 最好优先遍历	41
2.4 存储URL地址	42
2.4.1 BerkeleyDB	43
2.4.2 布隆过滤器	45
2.5 并行抓取	47
2.5.1 多线程爬虫	48
2.5.2 垂直搜索的多线程爬虫	50
2.5.3 异步IO	52
2.6 RSS抓取	56
2.7 抓取FTP	58
2.8 下载图片	59
2.9 图像的OCR识别	59
2.9.1 图像二值化	60
2.9.2 切分图像	63
2.9.3 SVM分类	67
2.10 Web结构挖掘	71
2.10.1 存储Web图	72
2.10.2 PageRank算法	76
2.10.3 HITS算法	84
2.10.4 主题相关的PageRank	88

<<解密搜索引擎技术实战>>

- 2.11 部署爬虫 90
- 2.12 本章小结 90
- 第3章 索引内容提取 93
 - 3.1 从HTML文件中提取文本 93
 - 3.1.1 字符集编码 93
 - 3.1.2 识别网页的编码 96
 - 3.1.3 网页编码转换为字符串编码 98
 - 3.1.4 使用HTMLParser实现定向抓取 99
 - 3.1.5 使用正则表达式提取数据 104
 - 3.1.6 结构化信息提取 106
 - 3.1.7 网页的DOM结构 109
 - 3.1.8 使用NekoHTML提取信息 111
 - 3.1.9 网页去噪 117
 - 3.1.10 网页结构相似度计算 122
 - 3.1.11 提取标题 124
 - 3.1.12 提取日期 126
 - 3.2 从非HTML文件中提取文本 126
 - 3.2.1 提取标题的一般方法 127
 - 3.2.2 PDF文件 132
 - 3.2.3 Word文件 135
 - 3.2.4 Rtf文件 137
 - 3.2.5 Excel文件 149
 - 3.2.6 PowerPoint文件 152
 - 3.3 提取垂直行业信息 153
 - 3.3.1 医疗行业 153
 - 3.3.2 旅游行业 153
 - 3.4 流媒体内容提取 154
 - 3.4.1 音频流内容提取 154
 - 3.4.2 视频流内容提取 158
 - 3.5 存储提取内容 159
 - 3.6 本章小结 160
- 第4章 中文分词原理与实现 162
 - 4.1 Lucene中的中文分词 162
 - 4.1.1 Lucene切分原理 162
 - 4.1.2 Lucene中的Analyzer 164
 - 4.1.3 自己写Analyzer 167
 - 4.1.4 Lietu中文分词 170
 - 4.2 查找词典算法 170
 - 4.2.1 标准Trie树 171
 - 4.2.2 三叉Trie树 174
 - 4.3 中文分词的原理 179
 - 4.4 中文分词流程与结构 183
 - 4.5 形成切分词图 184
 - 4.6 概率语言模型的分词方法 191
 - 4.7 N元分词方法 195
 - 4.8 新词发现 198
 - 4.9 未登录词识别 199

<<解密搜索引擎技术实战>>

- 4.10 词性标注 200
 - 4.10.1 隐马尔可夫模型 204
 - 4.10.2 基于转换的错误学习方法 214
- 4.11 平滑算法 216
- 4.12 机器学习的方法 220
 - 4.12.1 最大熵 221
 - 4.12.2 条件随机场 224
- 4.13 有限状态机 224
- 4.14 本章小结 232
- 第5章 让搜索引擎理解自然语言 233
 - 5.1 停用词表 233
 - 5.2 句法分析树 235
 - 5.3 相似度计算 240
 - 5.4 文档排重 244
 - 5.4.1 语义指纹 245
 - 5.4.2 SimHash 248
 - 5.4.3 分布式文档排重 259
 - 5.5 中文关键词提取 260
 - 5.5.1 关键词提取的基本方法 260
 - 5.5.2 HITS算法应用于关键词提取 262
 - 5.5.3 从网页中提取关键词 265
 - 5.6 相关搜索词 265
 - 5.6.1 挖掘相关搜索词 265
 - 5.6.2 使用多线程计算相关搜索词 268
 - 5.7 信息提取 269
 - 5.8 拼写检查与建议 274
 - 5.8.1 模糊匹配问题 276
 - 5.8.2 英文拼写检查 279
 - 5.8.3 中文拼写检查 281
 - 5.9 自动摘要 284
 - 5.9.1 自动摘要技术 284
 - 5.9.2 自动摘要的设计 285
 - 5.9.3 基于篇章结构的自动摘要 291
 - 5.9.4 Lucene中的动态摘要 291
 - 5.10 文本分类 295
 - 5.10.1 特征提取 297
 - 5.10.2 中心向量法 300
 - 5.10.3 朴素贝叶斯 303
 - 5.10.4 支持向量机 313
 - 5.10.5 多级分类 322
 - 5.10.6 规则方法 324
 - 5.10.7 网页分类 327
 - 5.11 自动聚类 327
 - 5.11.1 聚类的定义 327
 - 5.11.2 K均值聚类方法 327
 - 5.11.3 K均值实现 329
 - 5.11.4 深入理解DBScan算法 335

<<解密搜索引擎技术实战>>

- 5.11.5 使用DBScan算法聚类实例 336
- 5.12 拼音转换 338
- 5.13 概念搜索 339
- 5.14 多语言搜索 348
- 5.15 跨语言搜索 349
- 5.16 情感识别 350
 - 5.16.1 确定词语的褒贬倾向 353
 - 5.16.2 实现情感识别 355
 - 5.16.3 用户协同过滤 356
- 5.17 本章小结 358
- 第6章 Lucene原理与应用 359
 - 6.1 Lucene深入介绍 359
 - 6.1.1 常用查询 359
 - 6.1.2 查询语法与解析 361
 - 6.1.3 查询原理 365
 - 6.1.4 使用Filter筛选搜索结果 366
 - 6.1.5 索引库中的高频成分 367
 - 6.1.6 索引数值列 367
 - 6.2 Lucene中的压缩算法 370
 - 6.2.1 变长压缩 370
 - 6.2.2 PForDelta 372
 - 6.2.3 前缀压缩 375
 - 6.2.4 差分编码 376
 - 6.2.5 设计索引库结构 379
 - 6.3 创建和维护索引库 380
 - 6.3.1 创建索引库 380
 - 6.3.2 向索引库中添加索引文档 380
 - 6.3.3 删除索引库中的索引文档 384
 - 6.3.4 更新索引库中的索引文档 384
 - 6.3.5 索引的合并 385
 - 6.3.6 索引文件格式 386
 - 6.3.7 分发索引 388
 - 6.3.8 修复索引 391
 - 6.4 查找索引库 391
 - 6.4.1 排序 392
 - 6.5 读写并发控制 392
 - 6.6 优化使用Lucene 393
 - 6.6.1 索引优化 393
 - 6.6.2 查询优化 394
 - 6.6.3 实现时间加权排序 397
 - 6.6.4 实现字词混合索引 400
 - 6.6.5 重用Tokenizer 405
 - 6.6.6 定制Tokenizer 405
 - 6.7 检索模型 407
 - 6.7.1 向量空间模型 408
 - 6.7.2 BM25概率模型 412
 - 6.7.3 统计语言模型 418

<<解密搜索引擎技术实战>>

- 6.8 查询大容量索引 419
- 6.9 实时搜索 420
- 6.10 本章小结 421
- 第7章 搜索引擎用户界面 422
 - 7.1 实现Lucene搜索 422
 - 7.2 搜索页面设计 423
 - 7.2.1 Struts2实现的搜索界面 424
 - 7.2.2 翻页组件 424
 - 7.3 实现搜索接口 425
 - 7.3.1 编码识别 425
 - 7.3.2 布尔搜索 429
 - 7.3.3 指定范围搜索 429
 - 7.3.4 搜索结果排序 430
 - 7.3.5 搜索页面的索引缓存与更新 431
 - 7.4 历史搜索词记录 432
 - 7.5 实现关键词高亮显示 433
 - 7.6 实现分类统计视图 435
 - 7.7 实现相似文档搜索 441
 - 7.8 实现AJAX搜索联想词 443
 - 7.8.1 估计查询词的文档频率 443
 - 7.8.2 搜索联想词总体结构 444
 - 7.8.3 服务器端处理 444
 - 7.8.4 浏览器端处理 446
 - 7.8.5 服务器端改进 451
 - 7.8.6 拼音提示 454
 - 7.8.7 部署总结 455
 - 7.9 集成其他功能 455
 - 7.9.1 拼写检查 455
 - 7.9.2 分类统计 456
 - 7.9.3 相关搜索 458
 - 7.9.4 再次查找 462
 - 7.9.5 搜索日志 462
 - 7.10 搜索日志分析 464
 - 7.10.1 日志信息过滤 464
 - 7.10.2 信息统计 465
 - 7.10.3 挖掘日志信息 468
 - 7.11 本章小结 469
- 第8章 使用Solr实现企业搜索 470
 - 8.1 Solr简介 470
 - 8.2 Solr基本用法 471
 - 8.2.1 Solr服务器端的配置与中文支持 472
 - 8.2.2 把数据放进Solr 477
 - 8.2.3 删除数据 479
 - 8.2.4 Solr客户端与搜索界面 480
 - 8.2.5 Solr索引库的查找 482
 - 8.2.6 索引分发 486
 - 8.2.7 Solr搜索优化 489

<<解密搜索引擎技术实战>>

8.3 从FAST Search移植到Solr	492
8.4 Solr扩展与定制	493
8.4.1 Solr中字词混合索引	493
8.4.2 相关检索	496
8.4.3 搜索结果去重	498
8.4.4 定制输入输出	501
8.4.5 分布式搜索	506
8.4.6 SolrJ查询分析器	508
8.4.7 扩展SolrJ	517
8.4.8 扩展Solr	518
8.4.9 查询Web图	522
8.5 Solr的.net客户端	525
8.6 Solr的PHP客户端	527
8.7 本章小结	530
第9章 地理信息系统案例分析	531
9.1 新闻提取	531
9.2 POI信息提取	536
9.2.1 提取主体	537
9.2.2 提取地区	538
9.2.3 指代消解	540
9.3 本章小结	542
第10章 户外活动搜索案例分析	543
10.1 爬虫	543
10.2 信息提取	544
10.3 搜索	547
10.4 本章小结	547
参考资源	548
书籍	548
网址	548

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>