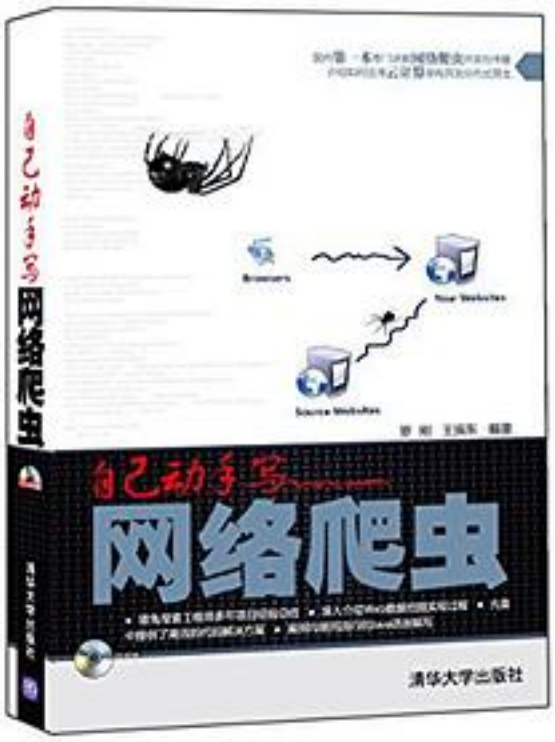
自己动手写网络爬虫



自己动手写网络爬虫_下载链接1_

著者:罗刚

出版者:清华大学出版社

出版时间:2010-10-1

装帧:平装

isbn:9787302236474

本书介绍了网络爬虫开发中的关键问题与Java实现。主要包括从互联网获取信息与提取信息和对Web信息挖掘等内容。本书在介绍基本原理的同时注重辅以具体代码实现来帮

助读者加深理解，书中部分代码甚至可以直接使用。
本书适用于有Java程序设计基础的开发人员。同时也可以作为计算机相关专业本科生或研究生的参考教材。

作者介绍:

目录: 第1篇 自己动手抓取数据第1章 全面剖析网络爬虫 1.1 抓取网页 1.1.1 深入理解URL 1.1.2 通过指定的URL抓取网页内容 1.1.3 Java网页抓取示例 1.1.4 处理HTTP状态码 1.2 宽度优先爬虫和带偏好的爬虫 1.2.1 图的宽度优先遍历 1.2.2 宽度优先遍历互联网 1.2.3 Java宽度优先爬虫示例 1.2.4 带偏好的爬虫 1.2.5 Java带偏好的爬虫示例 1.3 设计爬虫队列 1.3.1 爬虫队列 1.3.2 使用Berkeley DB构建爬虫队列 1.3.3 使用Berkeley DB构建爬虫队列示例 1.3.4 使用布隆过滤器构建Visited表 1.3.5 详解Heritrix爬虫队列 1.4 设计爬虫架构 1.4.1 爬虫架构 1.4.2 设计并行爬虫架构 1.4.3 详解Heritrix爬虫架构 1.5 使用多线程技术提升爬虫性能 1.5.1 详解Java多线程 1.5.2 爬虫中的多线程 1.5.3 一个简单的多线程爬虫实现 1.5.4 详解Heritrix多线程结构 1.6 本章小结第2章 分布式爬虫 2.1 设计分布式爬虫 2.1.1 分布式与云计算 2.1.2 分布式与云计算技术在爬虫中的应用——浅析Google的云计算架构 2.2 分布式存储 2.2.1 从Ralation_DB到keyvalue存储 2.2.2 Consistent Hash算法 2.2.3 Consistent Hash代码实现 2.3 Google的成功之道——GFS 2.3.1 GFS详解 2.3.2 开源GFS——HDFS 2.4 Google网页存储秘诀——BigTable 2.4.1 详解BigTable 2.4.2 开源BigTable——HBase 2.5 Google的成功之道——MapReduce算法 2.5.1 详解MapReduce算法 2.5.2 MapReduce容错处理 2.5.3 MapReduce实现架构 2.5.4 Hadoop中的MapReduce简介 2.5.5 wordCount例子的实现 2.6 Nutch中的分布式 2.6.1 Nutch爬虫详解 2.6.2 Nutch中的分布式 2.7 本章小结第3章 爬虫的“方方面面” 3.1 爬虫中的“黑洞” 3.2 限定爬虫和主题爬虫 3.2.1 理解主题爬虫 3.2.2 Java主题爬虫 3.2.3 理解限定爬虫 3.2.4 Java限定爬虫示例 3.3 有“道德”的爬虫 3.4 本章小结 第2篇 自己动手抽取Web内容第4章 “处理”HTML页面 4.1 征服正则表达式 4.1.1 学习正则表达式 4.1.2 Java正则表达式 4.2 抽取HTML正文 4.2.1 了解HtmlParser 4.2.2 使用正则表达式抽取示例 4.3 抽取正文 4.4 从JavaScript中抽取信息 4.4.1 JavaScript抽取方法 4.4.2 JavaScript抽取示例 4.5 本章小结第5章 非HTML正文抽取 5.1 抽取PDF文件 5.1.1 学习PDFBox 5.1.2 使用PDFBox抽取示例 5.1.3 提取PDF文件标题 5.1.4 处理PDF格式的公文 5.2 抽取Office文档 5.2.1 学习POI 5.2.2 使用POI抽取Word示例 5.2.3 使用POI抽取PPT示例 5.2.4 使用POI抽取Excel示例 5.3 抽取RTF 5.3.1 开源RTF文件解析器 5.3.2 实现一个RTF文件解析器 5.3.3 解析RTF示例 5.4 本章小结第6章 多媒体抽取 6.1 抽取视频 6.1.1 抽取视频关键帧 6.1.2 Java视频处理框架 6.1.3 Java视频抽取示例 6.2 音频抽取 6.2.1 抽取音频 6.2.2 学习Java音频抽取技术 6.3 本章小结第7章 去掉网页中的“噪声” 7.1 “噪声”对网页的影响 7.2 利用“统计学”消除“噪声” 7.2.1 网站风格树 7.2.2 “统计学去噪”Java实现 7.3 利用“视觉”消除“噪声” 7.3.1 “视觉”与“噪声” 7.3.2 “视觉去噪”Java实现 7.4 本章小结 第3篇 自己动手挖掘Web数据第8章 分析Web图 8.1 存储Web“图” 8.2 利用Web“图”分析链接 8.3 Google的秘密——PageRank 8.3.1 深入理解PageRank算法 8.3.2 PageRank算法的Java实现 8.3.3 应用PageRank进行链接分析 8.4 PageRank的兄弟HITS 8.4.1 深入理解HITS算法 8.4.2 HITS算法的Java实现 8.4.3 应用HITS进行链接分析 8.5 PageRank与HITS的比较 8.6 本章小结第9章 去掉重复的“文档” 9.1 何为“重复”的文档 9.2 去除“重复”文档——排重 9.3 利用“语义指纹”排重 9.3.1 理解“语义指纹” 9.3.2 “语义指纹”排重的Java实现 9.4 SimHash排重 9.4.1 理解SimHash 9.4.2 SimHash排重的Java实现 9.5 分布式文档排重 9.6 本章小结第10章 分类与聚类的应用 10.1 网页分类 10.1.1 收集语料库 10.1.2 选取网页的“特征” 10.1.3 使用支持向量机进行网页分类 10.1.4 利用URL地址进行网页分类 10.1.5 使用AdaBoost进行网页分类 10.2 网页聚类 10.2.1 深入理解DBScan算法 10.2.2 使用DBScan算法聚类实例 10.3 本章小结
・・・・・・(收起)

自己动手写网络爬虫_下载链接1_

标签

网络爬虫

搜索引擎

编程

爬虫

互联网

信息检索

计算机

搜索

评论

只读了第一和第二章，实在看不下去了，相关背景知识占用了太多太多的篇幅，真正我关心的只有2%-3%的几段话，刚开始讲就收了场。

翻完了, Java实现, 不是俺的菜,大概了解爬虫

不怎么的。感觉作者有点坑爹。

跑题了不少。大量代码堆砌，甚至有的章节上来什么都没说，先印刷了十来页代码。

废话连篇，各种东西都写上凑篇幅

没学会。

白开水一般，居然有些章节是网上的文章拼凑成的...

只看了一点

初步了解了网络爬虫的概念，了解了Google的PageRank算法的实现，HITS的实现。学到了很多

感觉一般般。另外就是我的爬虫都是用python写的。。

因为要写这方面的毕业设计所以买了这本书，我觉得入门看挺好的，如果要深入研究就找个开源爬虫实现好好研究一下

2013-12-24 重复

入门不错.. 只看了一点.. 最近暂时不搞爬虫了 暂时就看到这

只找到第一章啦～～～ 2了吧，HttpClient的api改了，里面的代码都不能用～～～

我看的是16年9月出版的。写得比较差，看似什么都有，实际一点实质内容没有。到处都是大段的无注释代码，拼凑内容。分布式爬虫到底怎么实现，solr和es一字不提，看完这本书收获几乎没有。

泛泛而谈

一般了

果然看国人出的技术书要抱着很低的期望

大致看了下，具体做还是过断时间吧！对编程基础要求蛮高的

感觉只是整理一下网络上的资源，很少有作者自己的思考。

自己动手写网络爬虫_下载链接1_

书评

开始从Web开发转向了网络爬虫的方向，然后在书店一个特偶然的机会，就看到了这本书，由于这是国内唯一一本关于网络爬虫的书，所以想也没想就买下了。
其实我原本是很不信任中国人写的书的，不过最近看了一些比如《Javascript王者归来》，再追溯到很久以前的《你所知道的.NET》...

虽然是最近才出的一本书，里面涉及到知识大部分可以在网络上找到，如第一章后面列举的爬虫，就有同样的英文文档，而且是很早以前的，作者根本没有自己去做些分析。第二章的bigtable,consistent hash都是现有论文或文章。
书中大段的代码其实也是没必要的，光盘里都有，书的内容...

想了解一下nutch，然后买了这本书，但是作者大量的copy网络资料，而且例子举得也很烂，然后东一点，西一点拼凑了这本书，看了几章，实在看不下去了。。。...

----
http://www.topteam.cc/02-shop-detail.php?cid=47&pid=236
當你在飆網時，是否知道還有一類特殊的網路使用者也再網際網路上默默的工作著，他們就是網路爬蟲。這些網路爬蟲按照設計者預定的方式，在網路中穿梭，同時自動蒐集有用的資訊，進行分類和整理，將整理結果提供給使用...

书中介绍的HttpClient版本旧了，下面是新的4.0版本的例子： import java.io.BufferedInputStream; import java.io.IOException; import java.io.InputStream; import java.util.ArrayList; import java.util.List; import org.apache.http.HttpResponse; import org.apache....

国内唯一的专业的爬虫与搜索开发培训课程。 http://www.lietu.com/train/ 联系: luogang @ gmail.com

----

做技术，心态很重要。见贤思齐，见不贤而内自省
这本书又轻知识点有多，天天在路上看！挺不错的！
这是真实好评吧。你自己对着镜子检查下，肯定发现自己更烂。 在 2012年9月9日上午10:25，Min Sun 写道： >...

腾道数据(http://www.tendata.cn/)是一家创业型的外贸咨讯网站，上线于2007年左右。 网站目前年营业额在1000万左右，略有盈利。
目前因发展需要，寻求资金合作，资金量在300-1000万左右。
如果您有兴趣，请致电010-81727660，或联系QQ：270954928
gtalk:luogang@gmail.com 详谈。

猎兔搜索从事企业搜索，自然语言处理等软件开发。产品包括多种语言的自然语言处理和搜索系统，网站搜索和垂直搜索软件，网络信息监测软件等。服务于农业信息化，竞争情报分析等领域。 岗位要求： 1.熟悉数据结构及其实现； 2.熟悉Java或c#；
3.数学基础较好。 开发工作: 中...

《自己动手写网络爬虫》 作者亲自主讲。每年培训不超过3期。
随着智能软件的不断普及，搜索引擎开发成为一项极富含金量的工作，市场对搜索软件开发工程师的需求极其旺盛。大型搜索门户需要大量专门的搜索软件开发人才，而众多中小型网站及企业也需要垂直搜索，...

自己动手写网络爬虫_下载链接1_