<<自己动手写网络爬虫>>

图书基本信息

书名：<<自己动手写网络爬虫>>

13位ISBN编号：9787302236474

10位ISBN编号：730223647X

出版时间：2010-10

出版时间：清华大学

作者：罗刚//王振东

页数：346

前言

当你在网上冲浪时，你是否知道还有一类特殊的网络用户也在互联网上默默地工作着，它们就是网络爬虫。
这些网络爬虫按照设计者预定的方式，在网络中穿梭，同时自动收集有用的信息，进行分类和整理，将整理结果提供给用户，以方便用户查找他们感兴趣的内容。
由于网络爬虫的实用性，引起了很多程序员，特别是Web程序员的兴趣。
但是大多数网络爬虫的开发原理与技巧在专业的公司内部都秘而不宣，至今仍然缺少理论与实践相结合的专门介绍网络爬虫的书籍。
本书将弥补这个问题，尝试理论与实践相结合，深入透彻地讲解网络爬虫的原理，并且辅以相关代码作为参考。
本书相关的代码在附带光盘中可以找到。
本书的两位主要作者在搜索引擎领域都有丰富的理论和实践经验。
同时，还有多个程序员帮忙开发或编写了代码实现，例如Java实现异步I/O或对PDF文件的处理等。
由于作者的日常工作繁忙，做得不够的地方敬请谅解。
作者罗刚在参加编写本书之前，还独立撰写过《自己动手写搜索引擎》一书，但存在讲解不够细致、知识点不够深入等问题。
此次与王振东合著本书，相对于上一本书而言，对读者反馈有更高的预期。
因为作者相信如下的假设：如果能够与更多的人更好地合作，事情往往能做得更好。
本书从基本的爬虫原理开始讲解，通过介绍优先级队列、宽度优先搜索等内容引领读者入门；之后根据当前风起云涌的云计算热潮，重点讲述了云计算的相关内容及其在爬虫中的应用，以及带偏好的爬虫、信息抽取、链接分析等内容；为了能够让读者更深入地了解爬虫，本书在最后两章还介绍了有关爬虫的数据挖掘的内容。
由于搜索引擎相关领域也正在快速发展中，而且由于篇幅的限制，有些不成熟的内容，没有能够在本书体现，例如有关"暗网"的内容。
随着技术的不断发展，我们将在今后的版本中加入这些内容。

<<自己动手写网络爬虫>>

内容概要

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。

<<自己动手写网络爬虫>>

书籍目录

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key／value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 木章小结第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

章节摘录

插图：代表主机ftp.yoyodyne.com的根目录。
爬虫最主要的处理对象就是URL，它根据URL地址取得所需要的文件内容，然后对它进行进一步的处理。
因此，准确地理解URL对理解网络爬虫至关重要。
从下一节开始，我们将详细地讲述如何根据URL地址来获得网页内容。
1.1.2 通过指定的URL抓取网页内容上一节详细介绍了URL的构成，这一节主要阐述如何根据给定的URL来抓取网页。
所谓网页抓取，就是把URL地址中指定的网络资源从网络流中读取出来，保存到本地。
类似于使用程序模拟IE浏览器的功能，把URL作为HTTP请求的内容发送到服务器端，然后读取服务器端的响应资源。
Java语言是为网络而生的编程语言，它把网络资源看成是一种文件，它对网络资源的访问和对本地文件的访问一样方便。
它把请求和响应封装为流。
因此我们可以根据相应内容，获得响应流，之后从流中按字节读取数据。
例如，java.net.URL类可以对相应的Web服务器发出请求并且获得响应文档。

<<自己动手写网络爬虫>>

编辑推荐

《自己动手写网络爬虫》是国内第一本专门讲解网络爬虫开发的书籍，介绍如何应用云计算架构开发分布式爬虫。
猎兔搜索工程师多年项目经验总结深入介绍Web数据挖掘实现过程光盘中提供了高效的代码解决方案案例均使用流行的Java语言编写

版权说明