**吉林工商学院工学院**

**2018届本科毕业论文(设计)开题报告**

**专业：网络工程**

**班级：14404**

**姓名：陈巍瑜**

**学号：421414404004**

**表一: 2018届本科毕业论文(设计)开题报告表**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 学生姓名 | 陈巍瑜 | 专业班级 | 网络工程14404 | 学号 | 421414404004 |
| 论文(设计)  题 目 | 基于网络爬虫的数据分析中的应用与实现 | | | | |
| **一、选题意义：**  随着以互联网为代表的各种网络技术的飞速发展，网络上存储的各种信息已经可以用海量来形容, 并呈现出指数增长的趋势，成为21世纪最为重要的经济资源之一。如何在海量的信息中快速检索到所需要的信息，成为了网络技术研究领域的一大热点。最初始的搜索引擎采用人工方式收集信息，由编辑员查看信息之后，人工形成信息摘要，并将信息置于事先确定的分类框架中。这种方法虽然准确可靠，但随着网络的飞速发展，在信息收集的数量和更新速度上已经远远无法满足要求，因此目前绝大多数搜索引擎都采用自动化的信息收集方式，由一个称为网络爬虫的机器人程序以某种策略自动地在互联网中搜集和发现信息，由索引器为搜集到的信息建立索引。该类搜索引擎的优点是信息量大、更新及时、毋需人工干预。可以看到，网络爬虫程序的质量直接关系到搜索引擎的质量，因此是目前网络大数据技术研究领域的一个热点。  网络爬虫，又称网页蜘蛛、网络机器人。随着计算机技术的高速发展，互联网中的信息量越来越大，搜索引擎应运而生。传统的搜索引擎会有返回结果不精确等局限性。为了解决传统搜索引擎的局限性，专用型网络爬虫在互联网中越来越常见。同时，专用型网络爬虫具有专用性，可以根据制定的规则和特征，最后只体现和筛选出有用的信息，并且可以对爬取得数据进行有效的数据分析实现快速的获取数据获取加快大数据在现实世界的应用。   1. **写作提纲：**   本课题研究的内容是重点如何设计实现高性能网络爬虫并且对爬取的数据进行简单的数据分析   1. 网页的抓取与HTML分析。 2. URL正则化与消重 3. 网络HTTP，Robots协议 4. Cookie处理，登录 5. 如何具备更强的抓取能力。 6. 对网页数学建模建立数据结构：图。   7 如何防止图环路的出现。  8. 数据获取的规则：正则表达式  9. 数据库MongoDB与redis缓存  10 爬虫架构Scrapy的设计与实现。  11. 对于缓存和并发请求的处理  12. 什么是反扒并且对反扒机制的应对  13. 对于数据的存储格式与形式  14. 对爬取的数据进行分析  15. 得出结论   1. **参考文献**   [1]李勇，韩亮.主题搜索引擎中网络爬虫的搜索策略研究[J].计算机与数字工程，2008, 228(10): 50-53.  [2]罗刚，王振东.自己动手写网络爬虫[M].北京:清华大学出版社，2010.  [3](美)Miguel Grinberg. Flask Web开发[M].安道，译.北京:人民邮电出版社，2015.  [4] Magnus Lie Hetland. Python基础教程(第二版)[M].司维，曾军岚，谭颖华，等，译.北京:人民邮电出版社，2010.  [5」叶允明，于水，马范援，等.分布式Web Crawler的研究:结构、算法和策略[J].电子学报，2002, 30(121:2008-2011.  [6] 网易博客 http://blog. 163. com  [7] ECMA http: //www.ecma-international.org  [8] W3C http://www.w3c.org  [9]Python爬虫开发与项目实践 范传辉 机械工业出版,2017年6月  [10]数学之美(第二版)作者吴军博士 人民邮电出版社 2014年  [11]TCP/IP详解 卷1：协议（第2版） （美）Kevin R. Fall，W. Richard Stevens 著 /2012-05-01 /机械工业出版社  [12]TCP/IP详解 卷3：TCP事务协议、HTTP、NNTP和UNIX域协议 [美]史蒂文斯（Stevens,W.R.） 著；胡谷雨等 译 /2000-09-01 /机械工业出版社  [13]计算机网络:自顶向下方法（原书第6版）（美）库罗斯，（美）罗斯 著，陈鸣 译 /2014-10-01 /机械工业出版社  [14]MongoDB权威指南（第2版）作者:(美)霍多罗夫 2014年01月 人民邮电出版社  [ 15]陈竹敏.面向垂直搜索引擎的主题爬虫技术研究[ D][博士论文].山东大学，2008: 10.  [16] Kleinberg J. Authoritative sources in a hyperlinked environment[J].Journal of the ACM ，1999, 46 (5). 604-632  [ 17]杨贞，等.基于本体的主题爬虫的设计与实现[J].科技情报开发与经济，2008, (2): 73-75.  [ 18]姚树宇，赵少东.一种使用分布式技术的搜索引擎[ J].计算机应用与软件，2005, 22 (10 ) : 127-129.   1. **写作进度：**   2017年11月20日——2017年11月26日：  完成开题报告二稿，筛选使用资料与工具  2017年11月27日——2017年12月3日：  完成开题报告终稿，期间与导师沟通，重新整理资料并提炼重点  2018年3月1日——2018年3月30日：  经过导师的集中指导，进行论文的初稿撰写  2018年3月31日——2018年4月11日：  与导师进行对接，请求进行技术性指导，修改并完善初稿，形成二稿  2018年4月12日——2018年4月26日：  期间进行最终集中指导，与导师进行最后阶段的沟通与请教，完善并丰富论文的内容，使论文更加具有实用性意义，更加有可读性，学习性，形成毕业论文终稿      学生签字：  年 月 日 | | | | | |
| **指导教师意见：**  指导教师签字：  年 月 日 | | | | | |