**广 西 科 技 大 学**

**普通本科毕业设计（论文）开题报告**

课题名称 基于vue框架的生鲜商城的设计与实现

学 院 启迪数字学院

专 业 软件工程

班 级 软件Q183

学 号 201809601137

姓 名 唐迎霜

指导教师 刘永胜

2022年 3 月 3 日

一、毕业设计（论文）选题的目的和意义

**（一）研究意义：**

1.深入学习Python和Scrapy开源框架，自己动手实现有良好拓展性的网络爬虫软件，该系统针对需要登录的网站进行数据抓取，因此可以很好地弥补了通用搜索引擎的盲区。

2.该网站追求对scrapy框架的探索和完善，探究深层网搜索解决方案。针对不同用户的搜索历史提日志分析服务。

**（二）研究目的：**

1、数据抓取旨在使用基于python的开源技术，串联网络编程的知识，实现一个操作方便的拓展能力强的爬虫软件。体会深度优先原理；理解scrapy源码和所有组件的使用。

2、充分应用elasticsearch功能进行离线存储和分析，学习搭建起性能强大的搜索引擎网站。比较elasticsearch和NoSQL的区别。对从后端数据抓取至前端可视化的完整路径有全面认识，促进今后的进一步学习和研究。

3、本文基于开源技术，以IT行业为目标，获取文章、社区问答、岗位三种形式的信息，用selenium自动测试框架突破身份验证。

二、设计或研究主要内容和重点，预期达到的目标及拟解决的主 要问题和技术关键，有何创新之处

**（一）研究主要内容内容（解决的问题）**

随着新一代通信技术的普及和多媒体的快速发展，网络非结构化数据爆炸增长，获取信息的渠道大大拓宽，但是相比较表层网(Surface Web)，更纯粹的、有更大价值的数据仍然存放在深层网(Deep Web)中；如何从浩瀚的深层网络信息中抓取目标信息是大数据时代的必修课，这一需求也催生了搜索引擎的改进和爬虫技术的拓展。 搜索引擎是用户获得互联网资源的重要途径，在信息时代初期，搜索引擎给人们使用互联网获得知识和促进商业繁荣提供极大支持。随着各种开源框架的完善，很多特定领域的计算机工作者不再满足于不可控的传统搜索引擎，而追求自定义的垂直型搜索。现在主要的商业搜索引擎模块都应用了不同的爬虫算法，在金融、科研等领域爬虫技术也在发挥着重大作用。

爬虫技术经过长期的发展变化，已经贡献了无数的开源框架和开源库。其中以python为基础的scrapy框架是当今最流行的一种，本文将根据scrapy实现一个爬虫，利用elasticsearch集成爬虫算法，搭建一个搜索网站。在诸多设计中，爬虫是否能及时监测到网页内容的更新并采取对应策略成为判定该系统是否优秀的重要指标之一。

**（二）预期成果**

1、本文基于elasticsearch、scrapy和django三大开源框架，实现了一个简约易拓展的搜索引擎。该搜索引擎以博客园、知乎、拉勾网为目标，检索IT行业的文章、社区、职位三种形式信息。

**（三）研究重点和技术关键**

研究重点体现两个方面：自动登录机制和突破反爬机制，很多网站尤其是社交类型网站只有登陆后才能访问内容。Cookie中保存了用户名和密码等记录。本文考虑设计Cookie池，在爬虫运行后可以自动登录。针对网页的各种反爬措施，该项目也尝试了应对策略。经过实验，系统采用三种有效的方法：改变请求频率、设置随机User-Agent及设置IP代理池。

**（四）独创或新颖之处**

（1）、探索如何突破反爬，解决依靠登录才能访问的网站爬取

（2）、解决了URL重复问题。在高速运行的爬虫工作过程中，通过鉴别即将爬取的网页是否已经爬取过，提高了运行效率和资源的利用程度。

（3）、利用redis和cookie来保存热门搜索和搜索记录关键词，当用户发起相同的检索请求时，可以得到更快得响应。

三、研究方案

1、实验法:改进爬虫框架，并嵌入elasticsearch中，通过实验体会区别

2、文献研究法: 依据研究目的和课题，通过调查文献、论文全面地、客观地了解研究课题。研究文献包括垂直型搜索引擎发展历史和经验、爬取深层网络数据方案。从理论发展历史和现状；创新研究角度 ；前瞻性实践方面进行总结和分析。

3、实证研究法: 通过与python开发人员交流，了解如今网络爬虫技术特征，学习开发经验。

四、主要参考文献目录

[1]安子健.《基于Scrapy框架的网络爬虫实现与数据抓取分析》.[D].吉林大学,2017.

[2]马联帅.《基于Scrapy的分布式网络新闻抓取系统设计与实现》.[D].西安电子科技大学,2015.

[3]陈竹敏.《面向垂直搜索引擎的主题爬行技术研究》.[D].山东大学,2008.

[4]曾亚飞.《基于Elasticsearch的分布式智能搜索引擎的研究与实现》[D].重庆大学，2016.

[5]张高伟.《基于ElasticSearch的分布式视频垂直搜索引擎的设计与实现》[D].哈尔滨工业大学,2014.

[6] 孙晓，叶嘉麒，唐陈意.基于多策略的新浪微博大数据抓取及应用[J].合肥工业大学学报,2014,10:1211-1215.

[7] 刘宇，郑成焕. 基于Scrapy的深层网络爬虫研究[J].软件，2017, 38 (7):111一114.

[8] 王伟，魏乐，刘文清，舒红平. 基于Elas ticSearch的分布式全文搜索系统[J].电子科技，2018,8:56-60.

[9] 薛丽敏，吴琦，李骏.面向专用信息获取的用户定制主题网络爬虫技术研究[J].信息网络安全, 2017(2):12-21.

[10] 李培.基于Python的网络爬虫与反爬虫技术研究[J].计算机与数字工程，2018(6):1415-1496.

[11] 马联帅.《基于Scrapy的分布式网络新闻抓取系统设计与实现》.[D].西安电子科技大学,2015.

[12]韩洪光. 《搜索引擎分析》[D]. 北京交通大学, 2008.

[13] 周雪,刘乃文. 基于用户影响力的微博数据采集技术[J]. 山东师范大学学报(自然科学版), 2016(02):34-39.

五、毕业设计（论文）工作进度计划

|  |  |  |
| --- | --- | --- |
| 起迄日期 | 主要工作 | 预期结果 |
| 2022年2月1日-2022年3月1日 | 收集材料，搭建项目环境 | 整理出文献，配置环境 |
| 2022年3月1日-2022年3月11日 | 撰写开题报告 | 开题答辩 |
| 2022年3月11日-2022年4月15日 | 撰写论文，做完网站后台搭建 | 完成论文初稿 |
| 2022年4月16日-2022年5月22日 | 实现功能，制作PPT、修改论文 | 项目完工，考核答辩 |
| 2022年5月22日-2022年6月1日 | 总结答辩、修改论文 | 完成论文终稿 |

六、指导教师审查意见

指导教师签字：

年 月 日