吉林大学本科毕业设计（论文）开题报告

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 学 院 | 软件学院 | | | | 专业 | 软件工程 |
| 学生姓名 | 宋治佳 | | | | 学号 | 55130511 |
| 指导教师 | 韩冬冰 | | | | 职称 | 讲师 |
| 合作导师 |  | | 职称 |  | 单位 |  |
| 设计（论文）题目 | | 基于Node的北京租房信息推荐系统的设计与实现 | | | | |
| 1. 课题研究的背景和意义（综述国内外相关研究现状，阐述课题的研究目的、意义）   每年都有大量的外地毕业生涌入北京寻求就业机会，而其中几乎所有人都有租房的需求。房源紧张造成的供求关系不平等，使一些不法黑中介有机可乘，初来乍到的毕业生面临着钱房两空的风险。随着近年来O2O模式的兴起，互联网在线租房平台也如雨后春笋般纷纷出现在人们的视野中，各个平台形式不一，房源质量也不尽相同。但很重要的一点是，租客在寻找房源的过程中往往都需要自己一点一点的去发掘适合自己的房子，而没有一个针对租客自身情况进行推荐的租房推荐系统。因此，本课题致力于设计并实现这样一个可以根据租客的需要进行房源信息推荐的系统，以解决大批进京毕业生刚毕业时面临的找房难租房难问题。  对于Web爬虫方面，国内外已有较多的文献和实践。本课题将在目前已有研究成果的基础上，进一步挑选和改进Web爬虫部分的技术，使其能与该系统所需的应用场景更加适合。 | | | | | | |
| 1. 课题研究已有的工作基础（总结归纳本人的学习、科研、实习等成果，以及已掌握的前人资料，简述自己初步的学术见解，附证书、报告、外文文献）   通过查阅数篇现有论文与互联网资料、新闻，针对《基于Node的北京租房信息推荐系统的设计与实现》课题相关已了解国内外研究现状，并对该课题的必要性与研究方向给出了概述和说明。相关文献主要来自Web爬虫和住房信息统计方面的论文，并调查了国内几个较大的互联网租房平台。  目前已经确定相关技术的选择，对于课题的可行性、意义和价值也已明确，相关外文文献的翻译工作和文献综述也已经完成；另外对于相关技术也有一定的基础，Web爬虫方面同样已经进行了一系列尝试和实践，取得了一些成果。 | | | | | | |

|  |
| --- |
| 三、研究的内容及可行性分析  本课题主要针对北京租房房源信息进行采集、整理、分析和推荐。其中房源信息房源的获取，综合考虑各个平台，自如由于房源较为真实，价格明晰，官网所查到的房源都是真实存在并且固定价格，并且租客不需要与房主打交道，直接与自如签订合同，合同形式和内容全部统一透明，不会有陷阱条款等优势，我们选择了自如作为房源信息的采集对象；爬虫方面考虑到Node.js非阻塞模式的IO处理给Node.js带来在相对低系统资源耗用下的高性能与出众的负载能力，且Node.js采用事件驱动、异步编程，为网络服务而设计，因此我们选用了Node.js来实现。后台的数据存储由于MongoDB是文档型数据库，一个单一的对象，结构相对清晰，没有复杂的连接，并且使用JSON风格文件的形式，面向文档存储，因此我们选择MongoDB作为数据库；推荐主要根据用户的具体需求（如房屋价位，地理位置等）进行房屋的筛选和推送。  考虑到每年有大批的应届毕业生进京寻求工作机会，因此租房问题将会是持续困扰应届毕业生甚至大部分外来进京工作人员的一个难题，而且目前国内的主流互联网租房平台均无该功能供用户使用，因此该课题的研究是具有其自身的意义和价值的；另外从技术方面，主要应用了爬虫技术、MongoDB数据库以及基于邮件的房屋推荐，其中爬虫和MongoDB的部分已经做出了一些基本尝试，基于邮件的推送方式以前也有相关经验，因此该课题在技术方面同样可行。 |

|  |
| --- |
| 1. 论文拟解决的关键问题及难点   本课题需要解决的主要问题有以下几点：   1. 房源数据的爬取：使用Node对自如租房平台网站的北京地区房源信息进行爬取。 2. 爬取到房源数据的存储：使用MongoDB对获取到的房源信息进行持久化的存储。 3. 对于房源信息的归类、针对租客不同需求的输入给出相关房源信息的推荐，并通过邮件不断推送给租客。 4. 推荐时相关筛选和推荐算法的准确性。   本课题主要的难点如下：   1. 现在大多数公司都有反爬虫的方案，房源数据进行爬取的过程中，如何巧妙的应对反爬虫方案，防止被服务器加入黑名单封IP。 2. 对于推荐算法的部分，准确的匹配用户的需求是不容易的。 3. 房屋距离工作地点之间的距离和所需时间的计算不容易得出，要充分考虑早晚高峰等情况。 |
| 五、研究方法与技术路线(重点论述技术方案)  研究方法主要采用调查法、文献研究法和实验研究法，主要通过有目的、有计划、有系统地搜集有关研究对象现实状况或历史状况的材料，通过调查文献来获得资料，从而全面地、正确地了解掌握所要研究问题，依据现有的环境和条件小部分的对提出的方案和构想进行实验，调查其可行性。  该系统通过爬虫抓取自如平台的房源数据，将数据持久化存储到MongoDB数据库中，然后根据用户的需求，匹配相关条件，并且通过邮件向用户推送推荐的房屋。整个系统使用Node.js实现，原因如下：   1. Node.js采用事件驱动、异步编程，为网络服务而设计。其中ECMAScript的匿名函数和闭包特性非常适合事件驱动、异步编程。而且从语言角度讲，ECMAScript更加熟悉，不需要额外的学习成本。 2. Node.js非阻塞模式的IO处理给Node.js带来在相对低系统资源耗用下的高性能与出众的负载能力，非常适合用作依赖其它IO资源的中间层服务。 3. Node.js轻量高效，可以认为是数据密集型分布式部署环境下的实时应用系统的完美解决方案。Node非常适合如下情况：在响应客户端之前，预计可能有很高的流量，但所需的服务器端逻辑和处理不一定很多。 4. Node.js的引擎是V8，V8引擎是Google用于其Chrome浏览器的底层JavaScript引擎。Google使用V8创建了一个用C++编写的超快解释器，基于这些，V8引擎的执行效率很高，已经超过了大多数动态类型的语言（如Python、Ruby）。 |
| 1. 论文的进度安排   论文和系统的具体时间进度安排如下表：   |  |  | | --- | --- | | **时间** | **工作进度安排** | | 2017年1-3月 | 撰写文献综述，翻译外文文献，填写开题报告，开题。 | | 2017年3月中旬 | 完成项目的初期工作，准备前期检查。 | | 2017年4月中旬 | 完成实现整个系统并进行相关测试，准备开始撰写论文，准备中期检查。 | | 2017年5月中旬 | 基本完成论文，进入论文的修改阶段，同辅导老师对论文进行修改，准备后期检查。 | | 2017年5月下旬 | 论文查重，提交论文。 | | 2017年6月 | 准备进行毕业设计论文答辩。 | |
| 七、毕业设计研制报告或毕业论文撰写提纲（初步）  **毕业论文撰写提纲**  标题：基于Node的北京租房信息推荐系统的设计与实现  宋治佳  指导老师：韩冬冰  [摘要]  [关键词]  [引言]  [正文]   1. 引言    1. 研究背景    2. 意义和价值 2. 相关问题的提出    1. 需要解决的问题 3. 设计方案    1. 基本方案的提出    2. 方案的优化和改进 4. 系统实现    1. 系统具体实现    2. 遇到的的问题及解决方案    3. 系统的优化 5. 结论 6. 参考文献 |
| 1. 主要参考文献 2. 许笑,张伟哲,张宏莉,方滨兴. 广域网分布式Web爬虫[J]. 软件学报,2010,(05):1067-1082. 3. 胡戎,冯仲科,蒋君志伟. 基于CheerIO的MEAN Stack气象数据网络爬虫研究[J]. 农业机械学报,2016,(06):275-282. 赵本本.基于Scrapy的GitHub数据爬虫 4. 刘鹏飞. 基于神经网络的数据统计研究[D].西安科技大学,2012. 5. 方皓彦,雷科阳,邱奕,郝志强,邱雅琪. 基于数学建模法的武汉市买房或租房分析[J]. 科技创业月刊,2016,(22):36-38. 6. 郑凯梅. 基于统计机器学习的网络入侵检测分类研究[D].中国矿业大学（北京）,2010. 8. 周立柱.聚焦爬虫技术研究综述 9. 范并思.社会科学信息分析中的文本挖掘 10. 谈蕾.数据仓库和数据统计在研究生数据中的应用与展现 11. 孙立伟.网络爬虫技术的研究 12. 黄晓斌.文本挖掘在网络舆情信息分析中的应用 13. 谭东宁.小样本机器学习理论\_统计学习理论 14. 罗兵.支持AJAX的互联网搜索引擎爬虫设计与实现 15. 谷瑞.云环境下MongoDB海量数据存储模型的研究 16. 张路路.基于MongoDB的大数据存储方法研究与应用 17. 谢华成,马学文.MongoDB数据库下文件型数据存储研究 18. 姚立.IBM云计算平台下NodeJS应用支持环境的设计与实现 19. 许会元,何利力.NodeJS的异步非阻塞I/O研究 20. 林瑶.web实时数据同步研究 21. 高原.服务器端javascript技术研究 22. 李博洋.基于Node.js的分布式数字资源开放服务系统的设计与实现 23. 李迪迪.浅谈房产中介智能推荐系统的实现 24. 岳雨俭.基于Hadoop分布式网络爬虫技术的研究 25. 陈欢.面向垂直搜索引擎的聚焦网络爬虫关键技术研究与实现 26. 马汉超.基于主题网络爬虫的汽车行业多元信息web系统设计与实现 27. 李春生.基于WEB信息采集的分布式网络爬虫搜索引擎的研究 |
| 九、指导教师意见  签字： 2017 年 月 日 |
| 十、开题审查小组意见  （要求具体意见，对前8项进行评价，结论：通过，不通过）  开题小组评分（百分制）：  开题小组组长签字： 2017 年 月 日 |