吉林大学本科毕业设计（论文）文献综述

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 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 | | | | |
| 文献综述（主要包括国内外现状、研究方向、进展情况、存在问题、参考文献等）（3000字以上）  *（说明：文献综述是通过系统地查阅与所选课题相关的国内外文献，进行搜集、整理、加工，从而撰写的综合性叙述和评价的文章。要体现“综合性”、“描述性”、“评价性”的特征。主体部分的结构包括该课题的“研究历史”的回顾，“研究现状”的对比，以及研究的“发展趋势”）*  **1.前言**  本文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》课题做出国内外研究现状的说明，以及对该课题的必要性与研究方向做出简要的概述和说明。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。   1. **正文**   通过查阅相关资料和新闻，可以得出以下结论：   1. 国内外对于租房的需求通常不同，显然国内的租房平台更多，相比之下，国外的房源更加充足，租房方式也更偏向于从房主发布的租房信息直接接触房主，而非租房平台。 2. 对于国内几家主流的租房平台（58、链家、自如等），网络上评论均褒贬不一，综合考虑，我们选择自如平台上面的信息作为信息源进行该课题的设计和实现。 3. 对于房源信息的采集，可以使用常见的Web爬虫进行采集。   **2.1.课题必要性**  每年都有大量的外地毕业生涌入北上广深这四座一线城市寻求就业机会，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。  **2.2.房源信息的选择**  对于房源信息的选择，我们在几个主流平台之间做出比较，分别是：58、链家、107间、自如。对于以上这些平台，我们在互联网上寻找了一些评价和相关信息。  58同城是一个中介形式的平台，由房主在该平台上发布房源信息，租客可以通过联系方式联系房主，但经过调查发现，该平台上的房源普遍不可信，很多房源是一些小中介发布到平台上的，而且网友评论普遍都不建议在这个平台上进行租房，因此排除该平台。  107间与链家这两个平台与58同城类似，但区别是他们的审查制度更加完善，房源的质量较58同城优秀很多，但是仍然存在着个人与房主之间合同不统一，产生各种各样纠纷的问题。  自如与上面提到的几个平台的区别是，他将房主房屋的使用权租到自己的手中，统一管理和装修，分离了房主和租客，房主将房屋代理到自如平台，租客租房时只需要与自如之间进行沟通和签订合同。而且它有如下优点：   1. 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的，不会出现小中介通过一个假的廉价房源骗租客去看别的房源的问题。 2. 有四个风格可以供人选择，装修质量不错，并且很统一，不用担心质量问题。 3. 租房过程中出现问题有人管理，有问题网上报修，不会出现租客交完钱中介不见了的问题。 4. 租房合同统一且透明，不存在合同中暗藏玄机的问题。交租金很方便，支付宝、微信、网银很轻松就能交付。   基于这些，本课题在设计与实现过程中决定选择自如的房源进行房源信息的采集。  **2.3.Web爬虫的选择**  使用Web爬虫对房源信息进行采集。对于爬虫的选择，常见的爬虫通常使用Python来实现，它有很多爬虫的库来很容易的实现数据的爬取功能。但是经过反复的考虑我们决定使用Node.js实现爬虫，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）。   **2.4.推荐算法的设计**  由于租客人群几乎都是进京工作，因此房屋距离工作地点的距离是很重要的，因此在推荐算法的设计过程中，房屋到工作地点的路程是很重要的因素之一，另外价格也是很重要的因素。因此，可以根据租客输入的条件为租客进行房源信息的推荐。租客可以提供理想的价位区间，可接受的通勤时间，房屋的类型等相关信息，推荐系统通过匹配这些信息为租客推荐符合条件的房源信息，通过相应的方式为租客进行推送。  目前国内的租房平台并没有这样的一个推荐系统，因此该课题对于弥补这方面的空白是很有意义的。   1. **总结**   本文结合相关文献、新闻、互联网资料，论述了国内外租房平台的现状、国内租房需求和形式以及采集房源信息爬虫的选取。其中，国内租房情况以北京为研究对象，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困扰毕业生的一个难题；获取房源信息平台的选择，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。  通过上述几个方面，结合现状，分别论述和证明了该课题研究的必要性、意义和价值，该课题可以弥补当前租房平台没有定制推荐功能的缺陷，能够解决大批进京应届毕业生刚毕业时面临的找房难租房难问题。   1. **参考文献** 2. 许笑.广域网分布式Web爬虫 3. 胡戎.基于CheerIO的MEANStack气象数据网络爬虫研究 4. 赵本本.基于Scrapy的GitHub数据爬虫 5. 刘鹏飞.基于神经网络的数据统计研究 6. 季松.基于数据仓库的数据挖掘技术在房地产中介信息分析系统的研究和应用 7. 方皓彦.基于数学建模法的武汉市买房或租房分析 8. 郑凯梅.基于统计机器学习的网络入侵检测分类研究 9. 周立柱.聚焦爬虫技术研究综述 10. 范并思.社会科学信息分析中的文本挖掘 11. 谈蕾.数据仓库和数据统计在研究生数据中的应用与展现 12. 孙立伟.网络爬虫技术的研究 13. 黄晓斌.文本挖掘在网络舆情信息分析中的应用 14. 谭东宁.小样本机器学习理论\_统计学习理论 15. 罗兵.支持AJAX的互联网搜索引擎爬虫设计与实现 16. 谷瑞.云环境下MongoDB海量数据存储模型的研究 17. 张路路.基于MongoDB的大数据存储方法研究与应用 18. 谢华成,马学文.MongoDB数据库下文件型数据存储研究 19. 姚立.IBM云计算平台下NodeJS应用支持环境的设计与实现 20. 许会元,何利力.NodeJS的异步非阻塞I/O研究 21. 林瑶.web实时数据同步研究 22. 高原.服务器端javascript技术研究 23. 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 24. 李迪迪.浅谈房产中介智能推荐系统的实现 25. 岳雨俭.基于Hadoop分布式网络爬虫技术的研究 26. 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 27. 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 28. 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 | | | | | | |