**郑州工商学院**

本科毕业设计开题答辩记录

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 毕业设计题目 | | 基于Python的知乎话题数据可视化系统设计与实现 | | | |
| 学 院 | 信息工程 | 指导教师 | 李敏 | 职 称 | 副教授 |
| 学 号 | 180508010832 | 学生姓名 | 张永远 | 专业班级 | 计算机科学与技术本科1909班 |
| 答辩地点 | 腾讯视频会议 | 答辩时间 | 2023年 01 月 06 日 | | |
| 答辩组长 | 吕菊慧 | 答辩组成员 | 吕菊慧、陈腾飞、屈天舒 | | |
| 1. **课题研究自述（选题依据、目的和意义、研究内容、研究方法等）**   选题依据：知乎网站是目前国内互联网最受欢迎的知识型问答社区，知乎可以说部分代替了旺盛时期的贴吧、论坛。有一个自动收集整理信息并将信息进行可视化分析的工具具有非常重要的意义。网络爬虫完成对网页的搜索要按照网页链接中的下载网页的相关数据，它会搜索信息并接收信息和超链接。信息提取是指从特定的信息流中将人们感兴趣的信息过滤出来，在本文中的信息可以提取转化为具体的问题。可视化图表可以通过直接分析数据来揭示难以察觉的联系。  目的和意义：（目的）互联网成了海量信息的载体，互联网目前是分析市场趋势、监视竞争对手或者获取销售线索的最佳场所，如何有效地提取并利用这些信息成了一个巨大的挑战，而网络爬虫是一种很好的自动采集数据的通用手段。从海量数据中获得潜藏的有价值的信息，帮助企业或个人预测未来的趋势和行为，使得商务和生产活动具有前瞻性。通过数据可视化可以直观且清晰的加大对数据的了解程度。  （意义）热搜、热榜会带来巨大流量，这些都是后台统计数据分析的结果，商人赚的就是信息差，信息差的来源之一就是数据，分析数据可以获得更多的信息。分析热点可以分析社会发展、了解时事，还有是可以分析舆论走向。数据分析通过可视化可以直观且清晰的发现规律。爬虫可以代替一些重复性工作，解放生产力。搜索引擎谷歌、百度是一种网络爬虫机器人，天眼查、企查查等通过爬虫获取工商数据、企业信息整理进而通过信息盈利。所以研究爬虫可视化具有巨大的社会和商业价值。  研究内容：通过网络爬虫技术抓取了一定量知乎网站文本数据作为研究对象，然后对信息进行初步处理，其次展现数据，对数据进行解读。网络爬虫首先向待爬取的网站发起请求，如果目标网站的服务器响应正常，会得到响应，然后再通过合理技术手段解析目标网页内容，最后选择爬取、保存所需要的数据。爬取知乎问题的自带话题、题目和被浏览量等信息，在此基础上对文本数据进行处理，其中包括对文本进行异常字符的删除修改工作等。处理信息以分析反应各种话题热度，做回答量与浏览量分析，以及题目关键词的热度词云图。  研究方法：先多浏览网站确定要爬取的网站和网站信息，爬虫阶段使用bs4用于网页解析、获取数据，re正则表达式进行文字匹配、urllib制定URL、获取网页数据和获取网页错误信息。数据可视化基于Python爬虫，多种第三方工具实现数据的可视化处理，具体有网站快速搭建工具flask，Echarts数据可视化，jieba中文分词工具，wordcloud词云。  **二、开题答辩记录**（答辩中提出的主要问题及回答情况记录）  问题1： 只有技术描述，你的研究对象是谁？  答：分析知乎数据的关系和热点。知乎网站的问题和话题与浏览量回答数量的关系，根据问题做出词云分析出热点。  问题2： jieba使用过程  答：jieba是在做词云的过程中用到的，把问题的句子分成词，使用了cut精准模式不够理想分词有把长词分散的情况，use\_paddle=True的 paddle模式：利用PaddlePaddle深度学习框架实现分词。再返回列表后排除连接词、停顿词。  问题3： 叙述一下你的研究方法  答：观察法和数量研究法的结合，观察法是指研究者根据一定的研究目的、研究提纲或观察表，用自己的感官和辅助工具去直接观察被研究对象，从而获得资料的一种方法。数量研究法通过对研究对象的规模、速度、范围、程度等数量关系的分析研究，认识和揭示事物间的相互关系、变化规律和发展趋势，借以达到对事物的正确解释和预测的一种研究方法。  **三、开题答辩综合意见**  该生陈述开题报告主要内容陈述完整、语言表达基本流畅,对提出的问题阐述基本正确。该生对自己的设计思路清晰、研究方法可行，对毕业设计进度安排合理，工作量适中。建议课题要结合实际应用。  经答辩小组评议，该同学开题修改后通过。  **答辩组长（签字）：**  **年 月 日** | | | | | |