郑州工商学院

本 科 生 毕 业 设 计 开 题 报 告

院（部）：信息工程学院　　专　业：计算机科学与技术	
学　　号：180508010832　　姓　名：张永远	
毕业设计题目：基于 Python 爬虫的关键词与观看量分析的设计与实现	

（开题报告包含以下几个方面的内容：一、研究的背景和意义；二、文献综述；三、研究的主要内容和方法；四、主要参考文献；五、研究进度。）

一、研究的背景和意义

（一）研究背景

网络爬虫（web crawler），简单来说就是一种运用电脑来代替人工进行“自动化浏览网络”的程序，就其实质来讲也可以算作是一个小型的智能网络爬虫机器人。它们能够自动采集你想要知道的，并且能够访问到的页面内容，方便我们设计程序来进行下一步的信息处理及分析，并将处理结果进行存储。实践中常用的网络爬虫技术是使用者指定目标网页的 URL 为起点进行数据的爬取，其爬取过程的描述大致如下：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、保存所需要的文本数据。[1]

数据分析是指用适当的统计分析方法对收集来的大量数据进行分析，将它们加以汇总和理解并消化，以求最大化地开发数据的功能，发挥数据的作用。数据分析是为了提取有用信息和形成结论而对数据加以详细研究和概括总结的过程。[5]有目的的收集数据，是确保数据分析过程有效的基础。组织需要对收集数据的内容、渠道、方法进行策划。[6]

知乎网站是目前国内互联网最受欢迎的知识型问答社区，用户通过知乎分享信息资源或者汲取自己所需的资源。截至 2020 年 12 月，知乎上的总问题数超过 4400 万条，总回答数超过 2.4 亿条。在付费内容领域，知乎月活跃付费用户数已超过 250 万，总内容数超过 300 万，年访问人次超过 30 亿。目前知乎网站的话题标签是用户自己根据所提问题自行设置并用来标注问题所属类别。然而，由于用户自行标注的标签可能不准确而造成知乎网站无法及时有效地向用户推荐适当的答案。此外，对于知乎网站的海量文本数据，这种方法会产生大量的人力成本。[4]大数据时代人们需要更强的整理信息和多维度分析信息的能力，人没有交流讨论就没有思想碰撞就不能有进步，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。

（二）研究意义

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，数据采集以及分析能力已成为驱动业务决策的关键技能。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。现如今大数据时代已经到来，网络爬虫技术成为这个时代不可或缺的一部分，企业需要数据来分析用户行为、自己产品的不足之处以及竞争对手的信息等，而这一切的首要条件就是数据的采集。网络爬虫的价值其实就是数据的价值，在互联网社会中，数据是无价之宝，一切皆为数据，谁拥有了大量有用的数据，谁就拥有了决策的主动权。爬虫可以代替一些重复性工作，抓取互联网上的数据，为我所用，有了大量的数据，就如同有了一个数据银行一样，下一步做的就是如何将这些爬取的数据产品化，商业化。搜索引擎百度就是一个巨大的爬虫，天眼查的数据也是通过爬虫企业信息整理进而通过信息盈利。

数据分析可以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。在大数据时代中，数据处理技术得到了突飞猛进的发展，我们终于拥有了发现及挖掘隐者在海量数据背后的信息，并且将这些信息转化为知识及智慧的能力，数据开始了从量变到质变的转化过程。不管你从事什么行业，掌握了数据分析能力，往往在岗位上更有竞争力。数据分析可以提升民众对数据敏感程度，数据通过数据可视化的程度展示成信息，可以直观且清晰的加大对数据的了解程度。

二、文献综述

网络爬虫，英文名为 Network Spider，故可以简单理解为网络中爬行的爬虫，它的本质是一组计算机程序，要完成对网页的搜索要按照网页链接中的下载网页。[8]相关数据，例如文本、标记信息、到其他网站的 URL 链接信息等。网络爬虫在搜索引擎上扮演着重要的角色，尤其是在提取网页时。互联网爬虫通过请求页面上的 HTML 文档来识别对特定页面的访问。检查现有站点，不断在站点之间移动，自动创建文件并在内部存储它们网络数据库。当网络爬虫进入超文本时，它会搜索信息并接收其他超链接。该 URL 主要基于 HTML 结构，不依赖于用户干预。[12]

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，在本文中的信息提取可以转化为文本的分类问题。[13]数据分析的很大一部分是从样本到总体意义上的推断，但这些只是部分，而不是整体。数据分析的很大一部分是精辟的，通过对原始数据的简单而直接的提取，我们无法看出这些迹象，但这些也只是部分，而不是整体。数据分析可以处理大量数据并识别该数据中最有用的部分。近年来该领域的成功可能主要归功于制图技术的改进。这些图可以通过直接分析数据来揭示难以察觉的联系。数据分析的数学基础是在 20 世纪初建立的，但直到计算机的出

现，才使行动成为可能，数据分析才变得普遍。数据分析是数学和计算机科学相结合的结果。

三、研究的主要内容和方法

通过 python 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。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。

使用 Python 作为爬虫工具，Python 具有完整且强大的第三方工具，含有发起请求的 Request 模块，网页解析的 Beautiful Soup 库。基本思路是：先获取待爬取网页的链接:url，解析 html 网页源代码获得文本信息，对爬虫技术做伪装。使用目标站点发起请求（即 Request），如果服务器能正常响应则会得到一个含有网页资源的 Response，然后解析网页内容，一般使用正则表达式或 BeautifulSoup 等第三方模块。最后将规整的数据存到数据库或其他，即完成爬虫操作。将解析后得到的评论文本数据存储，若需要存储的数据量不是很大，我们可以将其存储在 excel 文档中，若数据量较为庞大，可将其存储在关系型数据库 sqlite 中，这种存储同样有利于文本分析时的数据访问。

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。

四、主要参考文献

[1]. 余洋. 豆瓣电影评论文本的情感分析及主题提取研究[D].云南财经大学,2018.

[2]. 郑鑫臻,吴韶波.基于网络爬虫技术的时令旅游信息获取[J].物联网技术,2018,8(5):83-87.

[3]. 孙磊,詹宏伟.国内认知语言学研究的可视化分析[J].外语与翻译,2020,27(3):60-66+98.

[4]. 张闯.基于深度学习的知乎标题的多标签文本分类[D].北京交通大学：电子与通信工程,2018.

[5]. 陶皖主编.云计算与大数据[M].西安电子科技大学出版社,2017.

[6]. 赵凯,李玮瑶著.大数据与云计算技术漫谈[M].光明日报出版,2016.

[7]. Daniel Liang.Python 语言程序设计[M].成都:机械工业出版社,2013.

[8]. 封俊.基于 Hadoop 的分布式搜索引擎研究与实现[D].太原:太原理工大学,2010.

[9]. 郭志杰,周世平,顾惊璞等.基于主题爬虫技术的三农舆情监测管理平台开发与应用[J].农业工程技术,2018,38(15):29-34.

[10]. 杨文刚,韩海涛.大数据背景下基于主题网络爬虫的档案信息采集[J].兰台世界,2015(7):20-21.
[11]. 董付国.Python 可以这样学[M].北京:清华大学出版社,2018:401-405.
[12]. 潘娜. 基于大数据技术的电信客户维系挽留的分析与研究[D]. 2017.
[13]. 张园园. 医疗贴吧中广告的提取系统[D]. 2016. DOI:10.7666/d.D01052856.

五、研究进度

1.设计选题阶段：2021 年 11 月 8 日—2021 年 11 月 25 日；
2.开题报告阶段：2021 年 11 月 26 日—2021 年 12 月 24 日；
3.初稿提交阶段：2021 年 12 月 25 日—2022 年 3 月 4 日；
4.中稿提交阶段：2022 年 3 月 5 日—2022 年 3 月 25 日；
5.定稿及查重阶段：2022 年 3 月 26 日—2022 年 4 月 15 日；
6.答辩阶段：2022 年 4 月 16 日—2022 年 5 月 7 日。

指导教师意见：

该生的开题报告格式基本符合开题报告的撰写规范，选题符合专业培养的目标，能够提升学生的综合运用知识的能力，有一定的难度。同意开题。

指导教师签名：

2022 年 12 月 24 日