Software Engineering and Applications 软件工程与应用, 2020, 9(4), 317-325
Published Online August 2020 in Hans. http://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2020.94036

Hans汉斯

Job Information Data Analysis and Visualization System Implementation Based on Python Crawler

Juan Liu, Xidong Guan

School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou Jiangxi
Email: guanxidong@jxust.edu.cn

Received: Aug. 7th, 2020; accepted: Aug. 20th, 2020; published: Aug. 27th, 2020

Abstract

In order to have a more intuitive understanding of the domestic big data-related professions specific requirements for education and work experience and the regional distribution of different kinds of enterprises and so on, using Python's data analysis and processing functions, we crawl a large number of position information from the 51 Job network through the Python crawler technology. We delete the null value information, the irrelevant job information and the mismatched information according to the data cleaning method to preprocess the data, and save the clean data to the database, then use Pyecharts for visualization of data analysis, with Flask as Web framework for Web application development, display the visual data on the web page. It improves the speed of users to query information and facilitates job seekers to find suitable and satisfying professions [1].

Keywords

Python Crawler, Position Information, Data Cleaning, Visualization, Flask

基于Python爬虫的职位信息数据分析和可视化系统实现

刘　娟，管希东

江西理工大学信息工程学院，江西　赣州
Email: guanxidong@jxust.edu.cn

收稿日期：2020年8月7日；录用日期：2020年8月20日；发布日期：2020年8月27日

文章引用: 刘娟, 管希东. 基于 Python 爬虫的职位信息数据分析和可视化系统实现[J]. 软件工程与应用, 2020, 9(4): 317-325. DOI: 10.12677/sea.2020.94036

摘　要

为了能更加直观地了解到国内大数据有关的职业对学历和工作经验的具体要求以及不同性质企业地区分布等情况，采用Python的数据分析和处理功能，通过Python爬虫技术爬取前程无忧网大量职位信息。按照删除有空值的信息、与大数据无关的职业、信息错位的数据清洗方法，对数据进行预处理，然后将清洗后的数据存入数据库，再利用Pyecharts对数据进行可视化分析，用Flask作为Web框架开发Web应用程序，将可视化的数据展示在网页，提高了用户查询信息的速度，方便求职者找到适合且满意的职位[1]。

关键词

Python爬虫，职位信息，数据清洗，可视化，Flask

1. 引言

在这个新时代，人们根据现有的职位信息数据分析系统得到的职位信息越来越碎片化，面对收集到的大量的职位信息数据难以迅速地筛选出对自己最有帮助的职位信息，又或者筛选出信息后不能直观地看到数据的特征、一般规律、变化的趋势或者数据之间潜在联系。本文致力于解决将获取到的数据进行有效的筛选和从多个角度可视化分析，借助 Python 爬虫技术模拟浏览器访问职业信息网站，爬取大量数据，用 Pandas 实现数据清洗，用 Pyecharts 实现数据可视化，用其提供的较为轻量级的 Flask 框架将可视化结果呈现在 Web 页面。前端实现采用了 Html、CSS、JavaScripts，完成了用户和系统之间的交互[2]。

2. 总体设计

基于 Python + Pyecharts + Flask 的职位信息可视化系统设计与实现，要求实现数据爬取功能，数据清洗功能，数据可视化功能。实现对前程无忧职位信息的数据采集、清洗后存入 Excel 表中，再推送至 MySQL 数据库中，结合 Pyecharts 组件，实现数据到可视化图表的转换，后台采用 Flask 框架实现接口功能，将可视化的图表推送至前端。用户登录注册后在首页面上可查看 Excel 表中数据详情以及可视化后的图表信息。将近期发布的招聘信息存入 MySQL 数据库中，显示在首页，可供用户简单搜索，查看详情。简单系统总体架构设计如图 1 所示。

3. 详细设计

3.1. 数据获取

爬取前程无忧网站大数据职位相关数据，防止不是通过浏览器正常访问会被网站禁止访问的问题，手动在 header 里加上 UA 属性，伪装成浏览器进行访问[3]。打开网页开发者模式，进入 Network 里的 Headers 找到自己浏览器的 UA 属性，构造 header 方法如下：

header={

(1)'Host':需要访问的网站信息，

(2) 加入访问请求，

(3) 'User-Agent':浏览器的 UA 属性} [4]。

接收到 URL 地址的 HTML 页面后采用正则表达式进行匹配字符串利用双层循环来实现换页爬取与换行输出[5] [6]，获取数据，保存到 Excel 表格中。

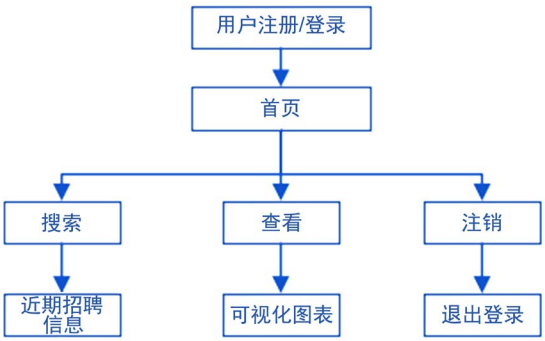

Figure 1. Overall architecture design of the system
图 1. 系统总体架构设计图

3.2. 数据清洗

数据清洗部分利用 pandas 从 Excel 表格中读取原始数据并对其进行清洗，操作过程如下：

step 1：读取文件；

step 2：删除所有空值行；

step 3：for 循环遍历：

3.1 判断职位是否为“大数据”相关职业，如果是，则保留，如果否，则整行删除；

3.2 判断信息是否匹配，如果是，则保留，如果否，则整行删除；

3.3 判断薪资单位是否为万/月，如果是，则保留，如果否，则转化为万/月。

清洗算法设计流程图如下图 2 所示。

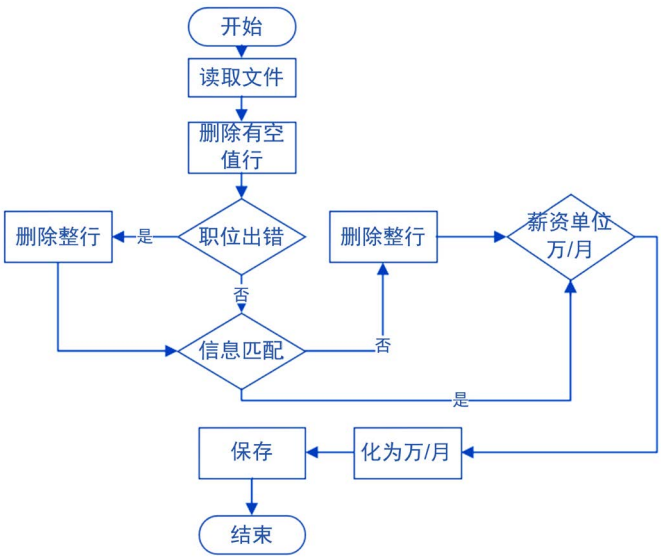

Figure 2. Flow chart of cleaning algorithm
图 2. 清洗算法流程图

数据清洗完毕后将其保存到新的 Excel 表格中，导入 MySQL 数据库中[7] [8]。爬取了超过 10,000 条数据，清洗后有近 4000 条数据，整理后的部分数据如下表 1 所示。

Table 1. Summarizes some of the data
表 1. 整理后部分数据

序号	职位	公司名称	公司地点	公司性质	工作经验	学历要求	公司福利	薪资	发布时间
1	大数据工程师	字节跳动	北京	民营公司	在校生/应届生	本科	六险一金	2.5~5 万/月	07-13
4	大数据开发工程师	紫金展锐	上海	外资(非欧美)	3~4 年工作经验	本科	弹性工作	2~4 万/月	07-13
10	大数据架构师	广州新科佳都科技有限公司	广州	上市公司	5~7 年经验	大专	免费班车	1.5~3 万/月	07-13

3.3. 数据分析与可视化

3.3.1. 数据分析

清洗后的表格职位信息应全是大数据相关的职业，表格信息还包括公司名称，公司地点，公司性质，薪资，学历要求，公司福利以及公司招聘发布时间且信息全部对应正确。求职者想要找到合适的职位，需要对各招聘公司的学历要求，工作经验要求，公司性质、所处地区进行可视化分析。了解哪些性质的公司在招聘人才，何种性质的公司对人才需求最多，分析公司招聘何种学历和工作经验的人才更为普遍，也就是求职者应达到的最基本的要求。此外，对招聘公司所处地区经济是否发达进行分析，是对公司的发展和经济实力的预测，方便求职者更准确的做出选择[9] [10]。

3.3.2. 学历要求玫瑰图

对各个企业学历要求生成玫瑰图。具体操作如下：

(1) 读取文件；

(2) 创建列表存储“学历要求”列信息；

(3) for 循环遍历列表获取不同学历要求值及.count()方法计算其出现次数；

(4) .key()方法获取各学历属性，.value()方法获取对应的值，pie.add()，半径设为 rosetype 绘制玫瑰图。

可视化结果如图 3 所示。玫瑰图的优势在于能十分直观的看出不同学历的占比情况，从图中不难看出接近七成的企业对学历的要求都是本科学历，不超过 25%的企业对学历基本要求是大专，约 5%的企业学历要求是硕士以上，极少数企业学历要求是高中，和中专学历，几乎没有企业要求中技和初中及以下的学历。求职者可以根据学历要求情况去加强自身学习努力达到大多数企业的要求。

3.3.3. 工作经验漏斗图

对各个企业工作经验要求生成漏斗图。与生成玫瑰图操作类似，具体操作如下：

(1) 读取文件；

(2) 创建列表存储“工作经验”列信息；

(3) for 循环遍历列表获取不同工作经验值及.count()方法计算其出现次数；

(4) .key()方法获取各工作经验属性，.value()方法获取对应的值，funnel.add()绘制漏斗图。

可视化结果如下图 4 所示。漏斗图的优势在于能十分清晰的看出招聘企业对于工作经验要求情况，自上而下，逐层减少。从图中不难看出企业对于 3~4 年工作经验是最普遍的要求，对 5~7 年工作经验要

求的企业数量排第二，绝大多数企业对于工作经验要求至少是一年，对于在校生和没有工作经验的求职者企业也会相应给一些机会，但是不多，较少企业要求求职者有 8 年以上的经验。

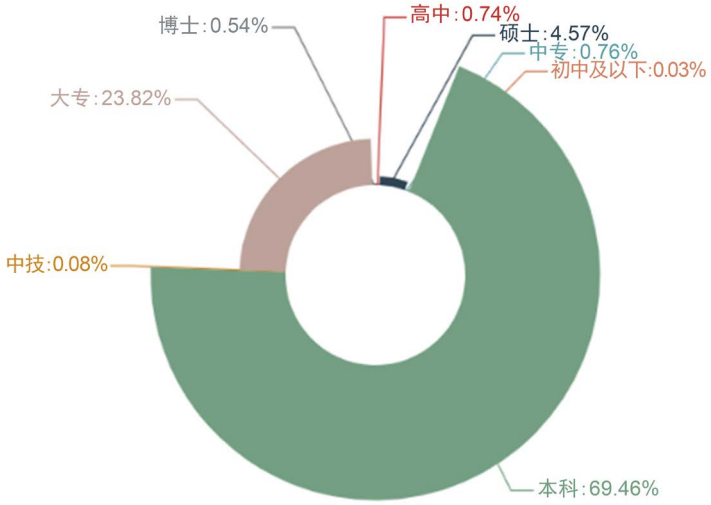

Figure 3. Rose chart for education requirements
图 3. 学历要求玫瑰图

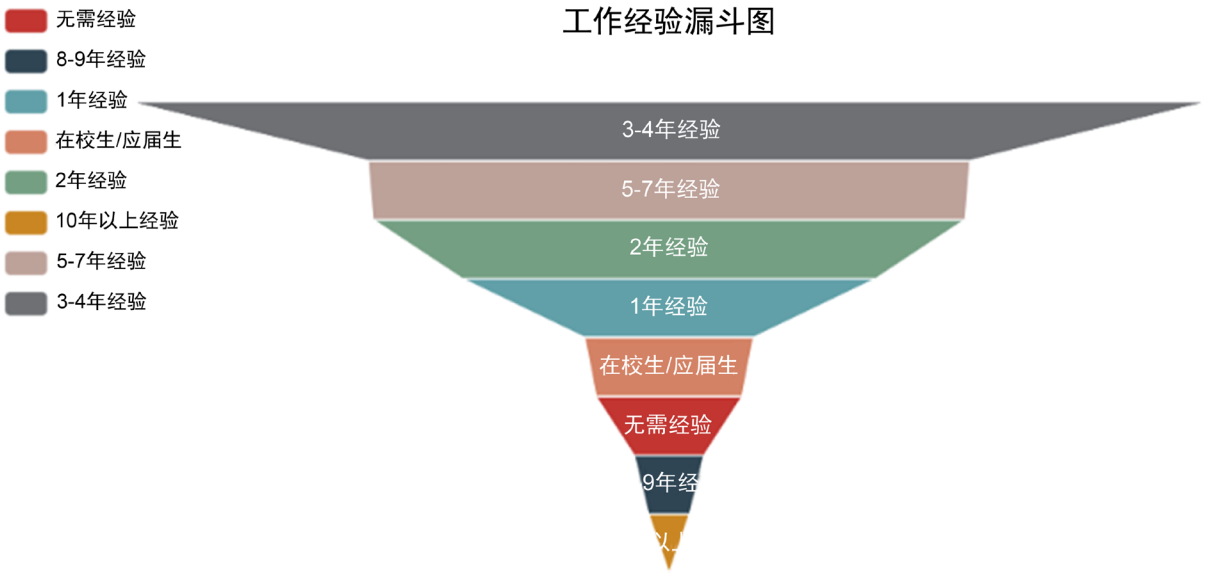

Figure 4. Funnel chart for work experience
图 4. 工作经验漏斗图

3.3.4. 公司性质饼图

分析所有招聘公司的性质。对所有公司性质情况生成了饼图[7]。操作方法如下：

(1) 读取文件；

(2) 创建列表存储“公司性质”列信息；

(3) for 循环遍历列表获取不同公司性质值及其出现次数；

(4) .key()方法获取各公司性质属性，.value()方法获取对应的值，pie.add()绘制饼图。

可视化结果如下图 5 所示。饼图的优势就是可以清晰的看出在整个系统中各种规模的公司占比的权重。由饼图可以分析出最主流的招聘公司是合资公司，外资公司和民营公司招聘较多，然后是事业单位，上市公司和创业公司数量差不多，但均不足 10%，国企占比最少。了解各个公司规模占比可以让求职者有一个更加清楚的认识，去公司规模占比大的公司求职，成功的可能性更大。

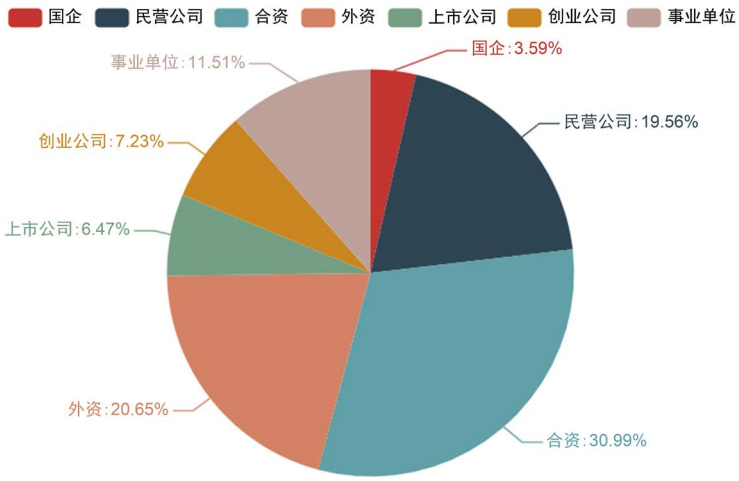

Figure 5. Pie chart for company nature
图 5. 公司性质饼图

3.3.5. 公司地区分布叠柱状图

对于公司地点分布情况生成叠柱状图[10]，分析在经济较一般的地区和经济较为发达的地区各种规模公司的数量。根据各地区经济发展状况自划分为较一般和较发达，同样方法绘制两个柱状图，使用 is_stack 标签堆叠将两个柱状图在一起，mark_point 标签标记经济较一般地区各类型企业数量的平均值，mark_line 标签标记经济较发达地区的最小/最大值。

可视化结果如下图 6 所示。叠柱状图的优势就是可以直观的看到各种规模的公司在两种地区的数量并形成鲜明的对比。由叠柱状图可以明显看出除创业公司和非营利公司外，其他类型的公司均是分布在经济较为发达的地区较多。给了求职者明确的地区方向。

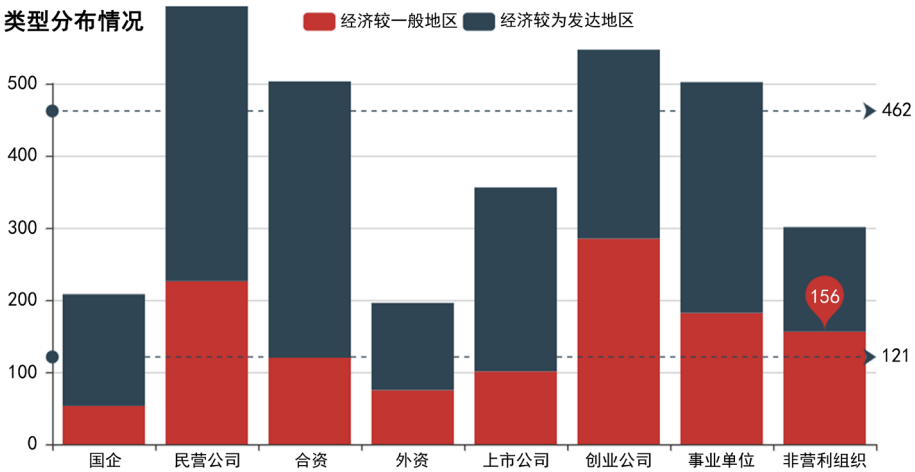

Figure 6. Double bar charts for regional distribution of company
图 6. 公司地区分布叠柱状图

3.3.6. 人才需求双折线图

对各类型的公司人才录取要求同样方法生成双折线图[10]。统计各类企业对学历有要求和工作经验有要求的数量。

可视化结果如下图 7 所示。双折线图的优势在于可以清晰的看到各种类型的公司对工作经验要求和学历要求的变化趋势。从图中可以清晰的看到总体上对于学历有要求的公司数量明显要高于工作经验的要求。

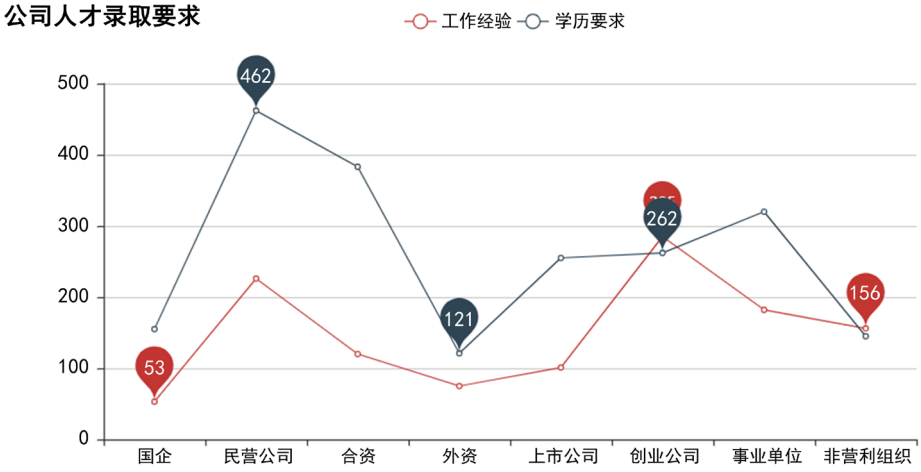

Figure 7. Double line charts for the talent demand of company
图 7. 公司人才录取要求图

3.3.7. 职位信息词云图

对于职业信息情况生成词云图[7]，由于爬取职位较多，选取 15 个职位出现频率最高的职位生成词云图给用户以视觉上的突出，操作如下：

(1) 读取文件；

(2) 创建列表存储“职位”列信息；

(3) for 循环遍历列表获取不同职位值及其出现次数；

(4) 选取出现次数最高的 15 个职位 wordcloud.add()绘制词云图。

可视化结果如下图 8 所示。从词云图可以看出企业对于大数据开发工程师需求是最多的。

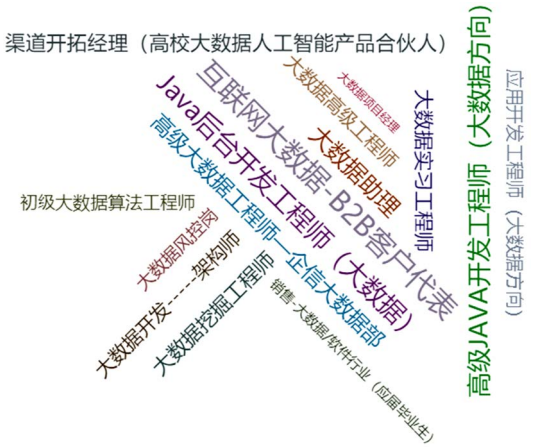

Figure 8. Wordcloud chart for job information
图 8. 职位信息词云图

4. 基于 Flask 的 web 功能实现

通过使用 Flask 的 route()装饰器用于把一个函数绑定到一个 URL，函数名称用于生成相关联的 URL，并返回需要在用户浏览器中显示的信息。在首页 HTML 文件中，引用 URL 实现将可视化呈现在 web 页面[11] [12]。

分别点击公司类型分布、公司规模、人才要求、学历要求、职位信息词云、工作经验要求按钮页面将跳转到相应的上述可视化图表页面，点击详情按钮，将显示数据清洗后的 Excel 表，将近期发布的招聘信息存入 MySQL 数据库中，显示在首页，可供用户简单搜索，查看详情。如下图 9 所示。与既有的研究成果比较，系统能较简洁直观的满足用户的基本需求。

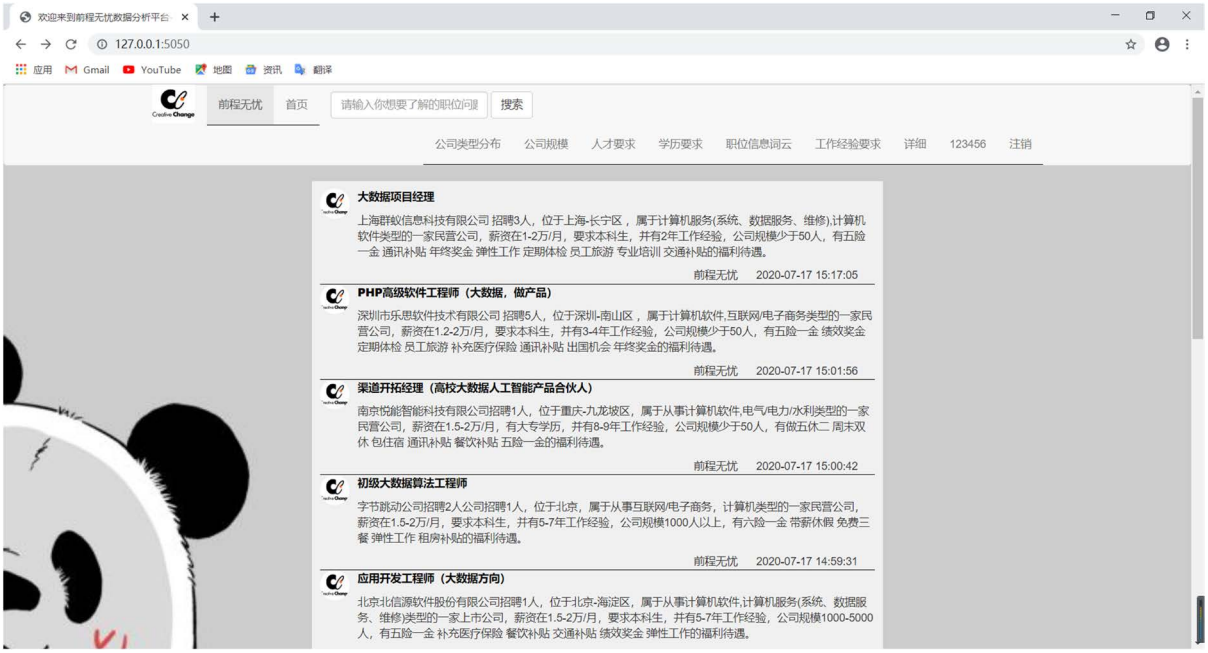

Figure 9. First page figure
图 9. 首页面图

5. 总结

利用 python 设计实现了职位信息数据分析和可视化系统，对前程无忧职位信息数据的爬取、清洗和可视化，实现了数据像图表的转化，使数据更有价值，并利用 Flask 框架实现将可视化结果呈现在 web 页面上，将各种数据以图表的形式呈现给用户，使繁多的数据变得直观，用户能够容易发掘隐藏的数据关联，筛选出对自己最有利的信息，做出最好的选择，此外，从多种角度进行分析，不同数据进行各不相同的可视化，利用不同可视化图表的优点，最大程度地展现数据的特征，能够给用户带来更清晰直观的体验。

参考文献

[1] 高巍, 孙盼盼, 李大舟. 基于 Python 爬虫的电影数据可视化分析[J]. 沈阳化工大学学报, 2020(1): 73-78.
[2] 曾诚. 基于 Python 的网络爬虫及数据可视化和预测分析[J]. 信息与电脑(理论版), 2020(9): 167-169.
[3] 黄岷昊, 丁浪, 张雪莲. 基于 Python 的网络爬虫及文本可视化[J]. 电脑编程技巧与维护, 2020(7): 24-25.
[4] 陈清. 基于 Python 的网站爬虫应用研究[J]. 通讯世界, 2020, 27(1): 202-203.

[5] 刘鑫. 网络爬虫在信息检索中的研究与应用[J]. 数字技术与应用, 2017(5): 95-97.

[6] 熊畅. 基于 Python 爬虫技术的网页数据抓取与分析研究[J]. 数字技术与应用, 2017(9): 35-36.

[7] 贾柠瑜. 基于 python 爬虫的岗位数据分析——以拉勾网为例[J]. 信息技术与信息化, 2019(4): 64-66.

[8] 詹静潇. 关于网络招聘信息的数据挖掘分析[D]: [硕士学位论文]. 天津: 天津财经大学, 2018.

[9] 何佳, 惠建忠, 王曙东, 洪晓媛, 王阔音. Python 在 CINRAD 风暴数据可视化中的应用[J]. 气象科技, 2020(3): 374-379.

[10] 刘艳玲, 姚建盛. Python 在数据可视化中的应用[J]. 福建电脑, 2020(3): 30-31, 34.

[11] 王瑞梅. 网络招聘数据可视化分析系统的设计与实现[D]: [硕士学位论文]. 石家庄: 河北师范大学, 2020.

[12] 谢勤政. 面向网络招聘系统的个性化推荐技术研究[D]: [硕士学位论文]. 长沙: 国防科技大学, 2018.