文章编号：2095-2163(2020)01-0223-05 中图分类号：TP311.13 文献标志码：A

基于网络爬虫的就业数据分析

项博良，唐淳淳，钱 前，曹健东
（上海工程技术大学 机械与汽车工程学院，上海 201620）

摘 要：随着网络信息量的爆炸式增长，大数据时代的来临，利用网络爬虫对大数据进行分析处理有非常重要的意义。本文以 BOSS 直聘网站为例，在 Python3.7 和 MySQL Server8.0 的基础上，设计并实现了一个关于就业信息的数据采集存储系统。并且通过对采集到的就业数据信息做出多个方面的分析，利用这些数据分析结果为大多数人在就业选择以及未来规划的时候提供一个有据可依的参照，起到一个指导就业的作用。

关键词：网络爬虫；就业信息；数据分析；就业指导

Analysis of employment data based on Web crawler

XIANG Boliang, TANG Chunchun, QIAN Qian, CAO Jiandong
(School of Mechanical and Automotive Engineering, Shanghai Unviersity of Engineering Science, Shanghai 201620, China)

【Abstract】With the explosive growth of the amount of network information, and the advent of the era of big data, it is important to use Web crawlers analyzing and processing big data. This paper takes the BOSS direct recruitment website as an example. Based on Python3.7 and MySQL Server8.0, a data acquisition and storage system for employment information is designed and implemented. And through the analysis of the collected employment data information, the use of these data analysis results could provide a evidence-based reference for most people in employment selection and future planning, and achieve a role in guiding employment.

【Key words】Web crawler; employment information; data analysis; career guidance

0 引 言

随着人工智能的概念逐步的深入展开，人工智能因其高效性和实用性受到越来越多的重视。作为人工智能的重要组成部分，大数据也开始在社会生产中发挥巨大作用，同时还带动了社会生活质量的全面提升，并提供了以往不曾有过的便利性。在国内对高等教育改革正迈向更深层次的时候，各校的毕业生规模也逐年增加。临近毕业时，或多或少都会存在许多迷茫。而在招聘、应聘的过程中，互联网作为当下承载海量招聘信息的重要载体，则给毕业生的择业提供了一条便捷途径。只是互联网的信息检索中却会面临许多用户并不需要的信息，只有通过人工筛选、再经总结对比后，才能得到最终想要的信息。

为了帮助高校毕业生在择业时能够快速获取特定的需求信息，并且通过快速数据分析得到自身择业的准确定位，从而做出更好的选择，为此本文设计研发了一套针对于招聘就业的专用爬虫。这里即以 BOSS 直聘作为实例，对如何开发爬虫获取信息，及对获取的信息快速分析进行了深入探讨与研究。对此拟展开剖析论述如下。

1 爬虫的设计

1.1 系统需求及分析

网络爬虫系统的开发是否成功取决于确保系统能够实现用户定制功能，达到预期设计目的。因此，在网络爬虫系统开发之前，就需要对该系统需求加以详尽分析，从而对整体的设计有一个清晰的思路。时下，普遍适用的爬虫系统都是模块化的，模块化的程序设计有利于代码块的测试与维护，而且也进一步增加了代码的适用性。在此基础上，只要对各个模块进行组合，就能够构建出一个完整的爬虫系统。本次研究即以 BOSS 直聘为例，开展模块化的编程设计。因为研究旨在通过爬虫系统对当前就业做出科学分析，故而针对此需求就要从 BOSS 直聘网站中获取全部的岗位信息，以及从每个岗位中获得包括各岗位名称、工作地点、薪水、公司规模性质、工作要求在内的各种关键信息。至此，在接下来的功能、模块设计中，就具备了较强的针对性。

1.2 爬虫模块设计

1.2.1 爬虫整体设计思路

爬虫系统的设计思路为：首先，需要获得所有包

作者简介：项博良(1994-)，男，硕士研究生，主要研究方向：机器视觉、图像处理；唐淳淳(1994-)，女，硕士研究生，主要研究方向：电池管理系统。

收稿日期：2019-09-16

括岗位信息网页的源码;其次,在每一页的网页源码中寻找出与需求相匹配的信息,此时就需要连接爬虫系统和数据库,将每次成功匹配到的信息均存入数据库中,直至所有网页检索完毕。在数据爬取的整个过程中,针对 BOSS 直聘的高度反爬,还要在各个模块中引入适当的反扒策略,以此保证数据爬取的连续性。研究可得整体设计框架如图 1 所示。

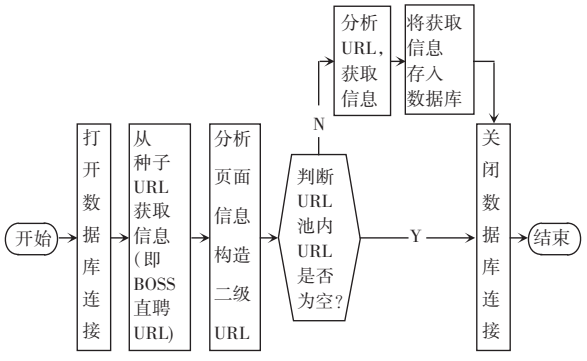

图 1 整体设计框图

Fig. 1 Overall design diagram

1.2.2 爬虫的网页抓取模块

网页抓取模块作为爬虫系统中最重要的部分,也是起始的模块。但是从实际爬取的情况来看,针对同一个 IP 在短时间内的多次爬取,会被网站屏蔽 IP 地址,因此在这里采用代理 IP 池的技术去访问。为了避免被对方发现,还需要加入 User-Agent 将自己伪装成代理服务器。通过构造代理 IP 池以及由众多用户代理组成的代理池,每次随机选择访问 IP 与用户代理的搭配,据此而将自己伪装成来自不同 IP 的用户访问,大大降低了被反爬虫的概率。接下来采用 Requsets 库的 API 去解析当前第一层的 URL。如:

resp = requests.get(url, headers = headers, proxies = proxies, timeout = 5)

1.2.3 网页源码分析模块

在提取好第一层 URL 的源码后,分析当前文本,寻找用户需要的关键信息,根据用户的需求,还需要了解每一类工作的名称与对应网页链接,通过对 ELEMENTS 的寻找,发现在标签 a-href 下存在着用户需要的信息,将所有的工作名称存入 JOB 列表,将所有的工作链接构造成完整的 URL 存入与 JOB 列表对应的 JOBURL 列表。

1.2.4 信息获取模块

由于 BOSS 直聘网站每一类工作的链接数最多不超过 10 页,在构造具体到每一页链接的时候,page 的数不应超过 10,且当链接无效,即已经检测超出最后一页的时候,便自动退出了。构造规则如下:

urlbase = link + '? page =' + str(i) + '&ka = page -' + str(i)

接下来便用 requests 库去实现当前网页解析,同样也可以运用代理 IP 池加上用户代理池随机选择与搭配的方法以便能够更加流畅地爬取信息。一个工作岗位对于求职人员最关心的应为岗位、薪水、公司信息,工作要求这些关键信息。用 Beautifulsoup 库去解析好的网页提取这些信息,此时将用到如下设计代码:

soupxbl = BeautifulSoup(resp1.text, 'lxml')

jobkinds = soupxb1.select('div.info - primary > h3 > a > div.job - title')

salarys = soupxb1.find_all('span', class = 'red')

yaoqius = soupxb1.find_all('div.info - primary > p')

names = soupxb1.select('div.company - text > h3 > a')

situations = soupxb1.select('div.info - company > div > p')

1.2.5 MySQL 数据库的联合使用

研究遍历完 BOSS 直聘网站上每一个工作岗位获得的信息相对来说是一个比较大的数据,在这里选择 MySQL 数据库对爬取的数据进行存储,因为 MySQL 数据库开源,易操作、并且速度、可靠性以及适应性都适宜。使用 MySQL Server8.0,并通过 pymysql 库去对数据库进行操作,在程序开端,利用 API 建立数据库的链接。设计研发代码参见如下:

conn = pymysql.connect(host = '127.0.0.1', user = 'root', password = 'xnxbl123@qq.com', db = 'bossapply', charset = 'utf8')

接下来,将基于用户需要保存的信息建立数据表格。设计研发代码见如下:

cur.execute (" DROP TABLE IF EXISTS bossapply")

sql_c = " create table bossapply (*job* char(50), *salary* char (50), *requirements* varchar (265), *company_name* char (100), *situation* varchar (265));"

此后,在网页的分析模块中提取信息后,将这些数据导入所创建的数据库中的表里面。设计研发代码见如下:

sql_insert = "insert into bossapply(*job*, *salary*, *requirements*, *company_name*, *situation*) values (%s,

%s,%s,%s,%s);"

cur.execute(*sql_insert*,(*s*1,*s*2,*s*3,*s*4,*s*5))

这样就能实现对数据库的操作，将研究中爬取到信息成功存入数据库，为下一步的就业数据分析奠定了基础。文中，利用数据库可视化工具 MySQL WorkBench 展示的部分爬取数据如图 2 所示。

图 2　部分爬取数据

Fig. 2　Partially crawling data

2　数据分析

2.1　数据处理

通过设计好的网络爬虫系统，从 BOSS 直聘网站上爬取了上海地区 13 万多的岗位招聘信息数据，从招聘岗位、工资待遇、工作地点、工作要求、公司性质这几方面的信息，对上海地区的就业数据做出研究与分析，对广大择业人员可起到一个初步指导的作用。

通过 Navicat Premium 将数据库导出成 Excel 文件，在 Python 中通过 pandas 库对数据进行处理，首先将所有的数据通过 read_excel 的 API 读取到处理环境下，将每一列的数据分别提取出来构造出 job、salary、requirements、situation 四个列表，通过遍历整个 requirements，检索每一个元素的字段，可以统计出上海市每个地区大约能够提供多少个工作岗位；同理，用上述的方法，可以统计出上海地区提供的工作岗位对学历的要求，以及公司规模的情况。对于就业数据分析来说，至关重要的就是薪资分析，将提取出来的 salary 列表，对每一个元素采用正则表达式匹配前两个数字，也就是这份工作的薪水上下限，求一个平均值，遍历整个列表，对薪水分布进行统计。同时，通过定位以及包含字符段的方法，可以将每个地区的工作以及相对应的薪水提取出来，再通过前文对全上海各地区的工作岗位统计，对上海各地区的平均薪资做出分析。在此基础上，各行各业的薪资水平也能够根据各行业的岗位数以及对应的平均薪资计算得出。

2.2　数据分析结果

随着应届毕业生的人数每年不断上升，带给社会的就业压力也随即增大，在这种就业形势竞争激烈的就业市场里面如何做出最佳的选择即已成为研究的热点与焦点。

研究可得，上海地区提供岗位图如图 3 所示。从图 3 可以看出上海每个地区提供的岗位数还是有很大差别的，其中以浦东地区提供的岗位数最多，且从图 3 可以看出金山、宝山、青浦、奉贤提供的就业岗位相对来说较少，大多数的就业岗位还是集中在市区。同时，也可得到，上海各教育程度提供岗位图如图 4 所示。从图 4 结果可以看出招聘当前的需求主要还是本科以上，大专以上，对于部分应届生，则可选择考研考博，凭此来提升在未来就业市场上的竞争实力。

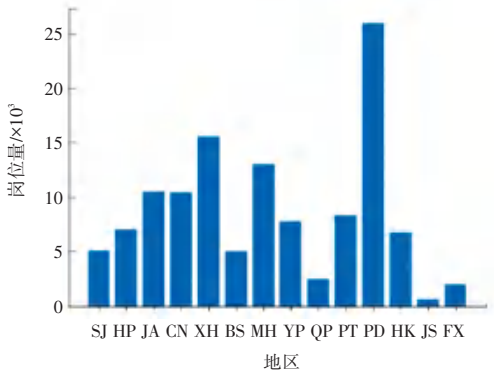

图 3　上海地区提供岗位图

Fig. 3　Provided job map in Shanghai

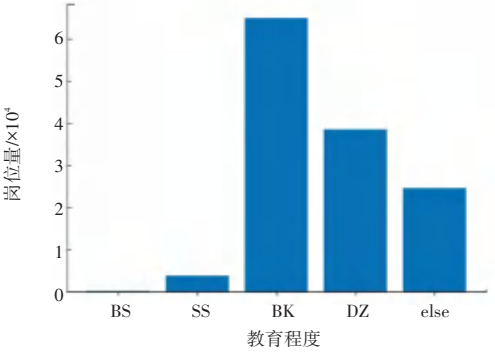

图 4　上海各教育程度提供岗位图

Fig. 4　Provided job map of education level in Shanghai

就业市场对各职业的需求也是各有不同。上海各就业种类招聘情况如图 5 所示。由图 5 分析可知，对技术岗的需求甚至超过了其它众多行业的需求总和，伴随着人工智能时代的来临，对人工智能相关的技术岗位缺口还是很大的，是一个前景可期的就业方向。在未来职业规划还未具备清晰认知时，可以作为一个参考方向。另外，市场营销与生产制造行业也能提供不错的岗位数。在薪资水平方面，

总体还是令人满意的,主要集中在月薪 6~10 K,以及 10~20 K 之间,月薪在 10 K 以上和以下的各占大约 50%,整体的收入水平保持在一个比较高的水准。

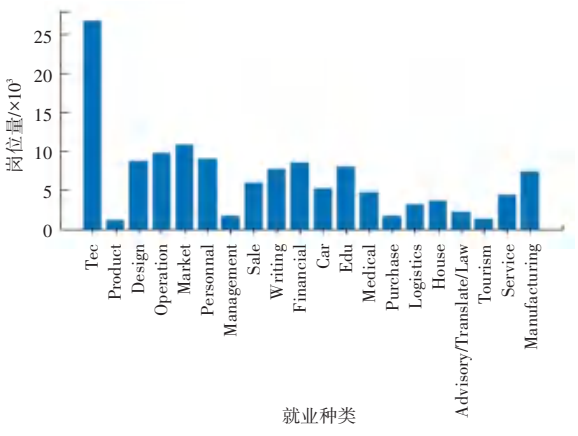

图 5 上海各就业种类招聘情况

Fig. 5 Each type of employment recruitment in Shanghai

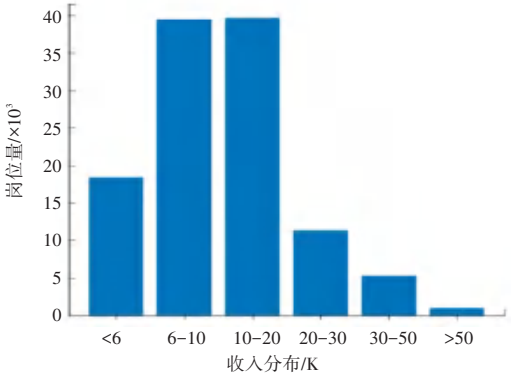

图 6 上海招聘收入情况

Fig. 6 Recruitment income in Shanghai

在前文对上海市总体的收入水平进行了直观判断基础上,继而得到上海各地区招聘收入情况如图 7 所示,上海各就业种类招聘收入如图 8 所示,以便能够对就业选择以及未来职业规划进行准确及有效判断。从图 7 中可以看出,上海各地区的收入情况差距不大,但是整体上来看,徐汇区还是略高一筹,这样在选择就业时可以根据地区消费的不同,以及未来规划选择工作区域。从图 8 中可以看出,薪资水平处于前三位的行业分别是产品行业、管理行业以及技术行业。而在数据分析后得出,在提供岗位数量最多的技术岗上,工资并不是最高,有些岗位虽然需求量不大,但是薪水很高。而且,从薪资分布来看,80%以上的行业的月收入都已经达到 10 K 或者 10 K 以上了,这样人们在选择就业的时候,可以更少地受到薪资影响,从而做出更适合自己的选择。

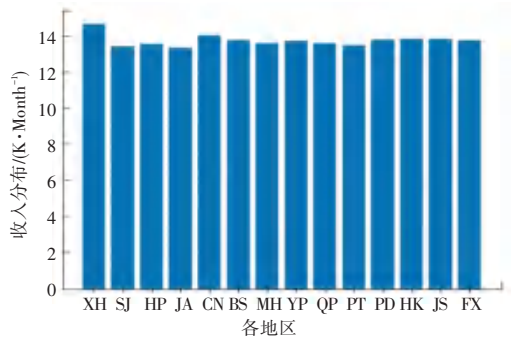

图 7 上海各地区招聘收入情况

Fig. 7 Recruitment income per region in Shanghai

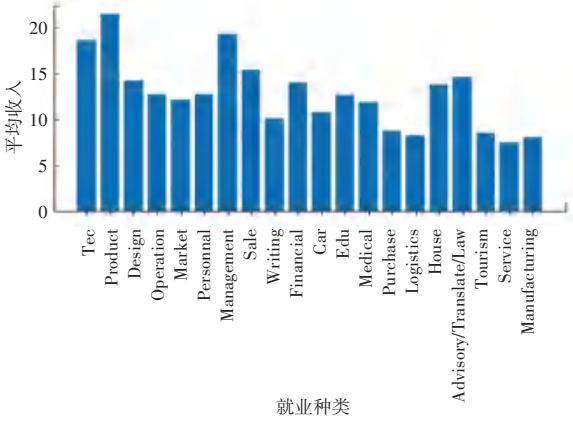

图 8 上海各就业种类招聘收入

Fig. 8 Income of each type of employment jobs in Shanghai

3 结束语

本文通过 Python 加上 MySQL Server 的配置,创建了一个基于 BOSS 直聘网站的网络爬虫数据收集分析系统,该系统能够登录到 BOSS 直聘,并获取页面信息,分析页面中的 URL,同时对筛选构造后的 URL 再一次进行数据筛选,将用户获取到的数据存储到数据库,在此基础上将对数据进行深层次的挖掘,也就是运用一系列的数据分析手段,获得关于上海各地区、各岗位的薪资待遇、招聘需求等一系列重要信息,为广大的就业人员提供有益的借鉴与参考。

参考文献

[1] 徐远超, 刘江华, 刘丽珍, 等. 基于 Web 的网络爬虫的设计与实现[J]. 微计算机信息, 2007, 23(21):119-121.

[2] 郭丽蓉. 基于 Python 的网络爬虫程序设计[J]. 电子技术与软件工程, 2017(23):248-249.

[3] 周中华, 张惠然, 谢江. 基于 Python 的新浪微博数据爬虫[J]. 计算机应用, 2014, 34(11):3131-3134.

[4] 陈琳, 任芳. 基于 Python 的新浪微博数据爬虫程序设计[J]. 信息系统工程, 2016(9):97-99.

[5] 左卫刚. 基于 Python 的新闻聚合系统网络爬虫研究[J]. 长春师范大学学报, 2018, 37(12):29-33.

[6] 张明杰. 基于网络爬虫技术的舆情数据采集系统设计与实现[J]. 现代计算机(专业版), 2015(18):72-75.

(下转第 230 页)