互联网职位数据可视化系统分析与实现

王玮 杨将天* 梁国际

南宁学院人工智能学院，南宁，530200

摘 要 在大数据环境下，了解每个互联网岗位的基本情况以及薪资待遇等方面是每个求职者最为关心的问题。在当前就业严峻的形势下，本文研究 python 网络爬虫技术获取互联网岗位的招聘信息，利用该数据开发一套完整的职位数据可视化系统给求职者提供参考依据。实现过程中，采取 python 网络爬虫技术获取当前主流的 51job、BOSS 直聘、拉勾网三个招聘网站上互联网岗位的招聘信息，对招聘数据经清洗和标准化之后对学历、经验等数据与薪资做相关性分析，选取相关性较好的因素构建多元线性回归模型进行薪资预测。接着使用 jieba 技术对招聘数据存在的结构化文本以及非结构化文本进行中文分词，提取描述中的英文单词作为岗位的技能要求，使用 Word2Vec 对其进行文本向量化，利用 NLP 的 CNN（卷积神经网络）模型，分析技能数据进行多层卷积、池化等操作，最后将岗位最重要的特征进行映射输出预测岗位类别。最后，使用 Flask 框架结合 Bootstrap 框架搭建前后端分离的 web 应用提供用户使用。

关键字 Python，多元线性回归，NLP

Analysis And Implementation Of the Internet Job Data Visualization System

WangWei Yang Jiangtian Liang Guoji

School of Artificial Intelligence Nanning University
Nanning 530200, China;
2508053272@qq.com jtyang2@iflytek.com gjliang@iflytek.com

Abstract—In the big data environment, understanding the basic information of each Internet job and the salary is the most important concern of every job seeker. In the current grim situation of employment, this paper studies the python network crawler technology to obtain the Internet job recruitment information, and uses the data to develop a complete set of job data visualization system to provide reference for job seekers. Implementation process, take python network crawler technology for the current mainstream 51job, BOSS straight hire, hook net three Internet recruitment site recruitment information, after cleaning and standardization of recruitment data of education, experience and salary correlation analysis, select good correlation factors to build multiple linear regression model for salary forecast. Then use jieba technology to exist in the recruitment data of structured text and unstructured text, extract the English words in the description as job skills, use Word2Vec for text vectorization, using NLP CNN (convolutional neural network) model, analysis skills data for convolution, pooling operations, and finally the most important features of job mapping output forecast job category. Finally, using the Flask framework combined with the Bootstrap framework to build the front and rear separated web applications to provide user use.

Key words—Python ,Multiple Linear Regression, NLP

1 引 言

随着招聘网站的出现，招聘信息过于繁杂，面对招聘网站丰富的招聘信息，求职者如何 快速、高效地从海量的招聘信息中找到有价值的信息成为当前研究的热点[1]，了解互联网行业的发展现状、薪资待遇等，从而规划、选择一个好的就业方向成为用户急需解决的问题。据教育部显示，今年毕业生已达到 1076 万人，同比增长 167 万人，增量均创历史新高[2]，在疫情的冲击以及各大互联网公司裁员的热潮下，拿到心仪的 offer，还是有一定的难度，就业形势非常严峻。在疫情后和高校毕业生首破千万人的就业压力之下，很多毕业生在招聘网站上获取岗位信息时，现有的数据分析系统得到的数据很碎片化，对大量的数据信息难以迅速的筛选出心仪的职位信息，不能通过招聘网站了解心仪岗位的薪资待遇等方面的情况，，这无疑加大了毕业生求职的难度因此开发一套完整的互联网职位数据可视化系统，通过系统简洁直观展示互联网岗位的招聘信息，帮助求职者精准预测岗位薪资，找到合适自己的岗位。

文献[3]采用 Hadoop 技术进行数据处理，利用 SSM 框架搭建大数据职位分析系统帮助求职者进行系统性的指导，提供学习地图以及学习博客获取专业指导，文献[4]通过网络爬虫技术获取 Boss 直聘网的招聘数据，并用 PHP 设计实现相应的数据可视化系统，文献[5]基于 python 和 Django 框架设计招聘数据可视

*通讯作者：杨将天，男，工程师，jtyang2@iflytek.com

化平台能够把拉勾网、前程无忧、智联招聘网的计算机岗位数据爬取下来存储到数据库中，利用 jieba 分词技术提高了文本分析效率，帮助求职者做出更好的职业规划，文献[6]采用 Hadoop 以及采用协同过滤数据分析技术设计了一套针对企业招聘数据和求职者的智能分析可视化平台，能够根据用户填写的信息个性化推荐岗位，但是平台目前只面向求职者，随着越来越多平台的出现，单一的可视化系统只能对招聘数据进行可视化分析，而不能利用数据特征来给求职者更多的便利。互联网职位数据可视化系统在数据可视化的基础上还能帮助求职者根据自身条件预测薪资并预知能否在当地买的起房子，还能根据具备的技术条件推荐适合自身的岗位，给求职者一个合理、科学的参考依据。

2 系统需求分析

求职者在浏览招聘网站时，只能通过岗位关键字去特定的招聘网查询岗位的招聘信息，而不能更好的通过单一的招聘网了解该岗位在全国范围内的情况，也不能明确某一个岗位或者行业对他们的要求，且效率以及时间上都十分浪费。而开发一套互联网职位数据可视化系统可以帮助求职者结合自身情况了解互联网岗位在全国范围内的岗位详情，并且能够结合自身情况预测大致薪资以及合理推荐适合的岗位。帮助求职者能够在激烈的岗位竞争中脱颖而出。

互联网职位数据可视化系统主要功能为数据采集、数据可视化、薪资预测、岗位推荐，而薪资预测以及岗位推荐的前提和资料来源于用户，针对不同用户自身不同的学历、经验等完成预测及推荐，并能够结合当地的房价水平情况供求职者参考。例如，系统需要用户填写学历、经验、职位、具备技能等信息来计算薪资及匹配岗位。

3 互联网职位数据可视化系统的设计与实现

3．1 系统总体架构设计

互联网职位数据可视化系统利用 Flask 框架结合 Bootstrap 框架搭建 PythonWeb 网站，利用 pyEcharts 支撑数据可视化的使用。其架构图如下图 1 所示。

（1）系统逻辑架构设计。本系统主要针对想要从事互联网岗位的求职者，用户在访问网站时，可视化模块展示三大招聘网各个互联网岗位的平均薪资对比情况，以及岗位的需求量和需求量最多的城市；点击想要从事的岗位可以跳转到岗位数量、平均薪资、房价水平的城市分布图，还有学历、经验对该岗位薪资的影响情况以及岗位所需的各个技术能力占比情况。薪资预测模块，前端通过输入岗位、学历、工作经验、目标城市这四个因素，可以预测大致的薪资情况。岗位推荐模块前端通过输入学历、经验以及具备是技术能力可以给用户推荐四个岗位。

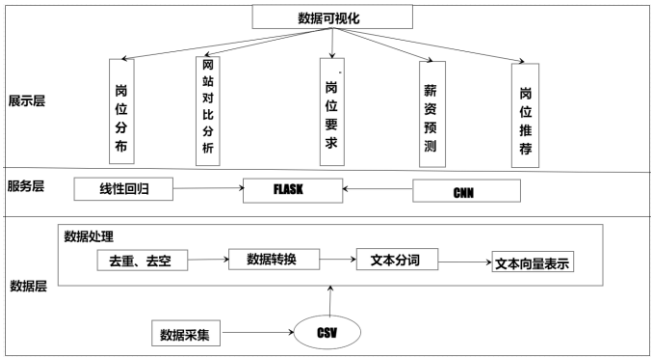


图 1 系统总体架构图

（2）系统功能模块划分。互联网职位数据可视化系统按照需求分析可划分为数据采集模块、数据分析模块、数据可视化模块、薪资预测模块、岗位推荐模块。数据采集模块又可以划分为数据爬虫以及数据预处理两个模块，为下一步数据分析做准备。

数据分析模块主要是对经过预处理之后的数据进行提取所需要的岗位数据进行数据分析;薪资预测就是将所提取的岗位数据进行薪资相关性分析，利用相关性较大的因素作为影响因素构建预测模型预测薪资;岗位推荐模块就是对职位描述中的技术关键词经过神经网络算法模型进行分类推荐；数据可视化模块就是将数据分析的结果展示在 Web 上，让用户可以直观的查看各个岗位的薪资待遇、技术能力要求等;最后将各个模块的功能通过搭建前后端分离的 Web 应用提供用户使用。

3．2 系统模块详细设计

互联网职位数据可视化系统详细设计主要是对系统的各个功能模块进行分析实现，主要分为数据采集模块、数据分析模块、数据可视化模块、薪资预测模块、岗位推荐模块。

（1）数据采集模块。数据采集模块主要是利用python网络爬虫对招聘网的互联网岗位数据进行抓取。通过开发者工具观察网站网页结构及其响应方式，利用xpath进行数据定位，通过设计爬虫程序并根据网页反爬特征设计相应的反爬措施进行爬虫，数据存储在CSV中。爬取的岗位数据应包括：职位名称,职位名称关键字,职位薪资,公司名称,地区,工作经验,学历要求,职位描述要求。数据预处理模块主要是利用split()等函数对工作经验、学历要求、职位薪资、地区数据作格式统一，方便后续进行数据分析。

（2）数据采集模块。数据分析模块就是对上一步预处理之后的数据中提取所需的岗位信息进行分析，比如每个岗位所需的技术要点，这时需要通过分词对职位描述的技术要点进行提取并进行词频统计、词向

量表示等。将三个网站的平均薪资一一计算利用图表进行对比得出薪资最高的岗位，以及高待遇所在的城市、数量等作一个总和，方便用户查看岗位的趋势以及更快的找到合适自己的岗位。

（3）数据可视化模块就是对数据分析模块的数据作一个图表展示，采用 Web 的方式简洁、直观的展示岗位数据的特征。示例图如图 2-9 所示。

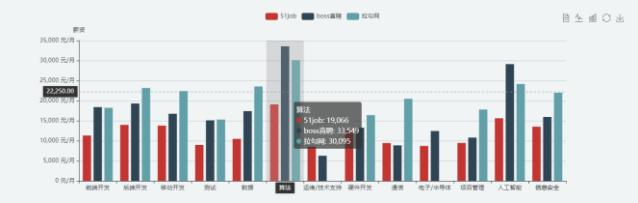
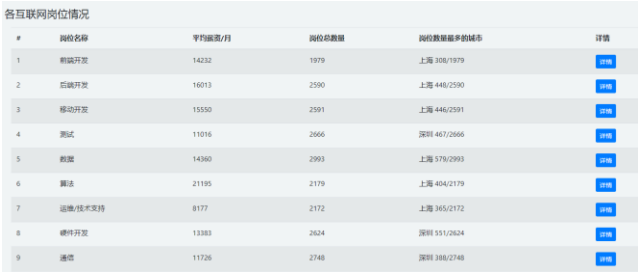


图 2 各大互联网岗位薪资对比图



各互联网岗位情况

#	岗位名称	平均薪资/月	岗位总数量	岗位数量最多的城市	详情
1	前端开发	14232	1979	上海 308/1979	详情
2	后端开发	16013	2590	上海 448/2590	详情
3	移动开发	15550	2591	上海 446/2591	详情
4	测试	11016	2666	深圳 467/2666	详情
5	数据	14360	2993	上海 579/2993	详情
6	算法	21195	2179	上海 404/2179	详情
7	运维/技术支持	8177	2172	上海 365/2172	详情
8	硬件开发	13383	2624	深圳 551/2624	详情
9	通信	11726	2748	深圳 388/2748	详情

图 3 岗位详情

（4）薪资预测模块。薪资预测模块就采用回归分析方法研究学历、经验、工作地、岗位数量等自变量与因变量薪资做相关性分析，选取相关性较好的因素构建多元线性回归模型组合预测，结果表明学历、经验、工作地是影响薪资的最主要的因素，利用三者组合模型对薪资进行定量预测。多元线性回归模型具体公式如下所示。其示例图如图 10 所示。

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \cdots + \beta_n x_n + \varepsilon$$

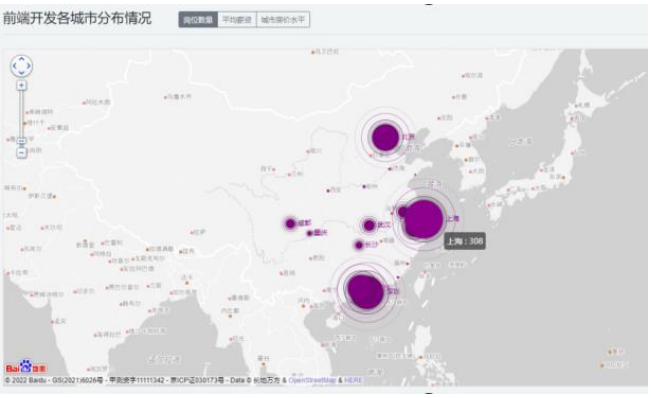


图 4 岗位数量城市分布图

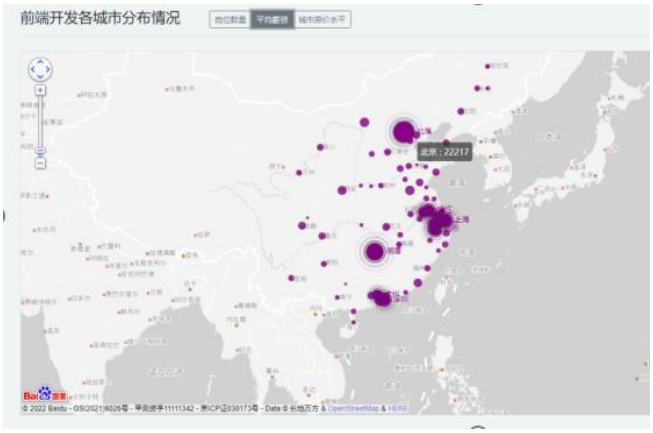


图 5 岗位平均薪资城市分布图

图 6 岗位所在城市房价水平

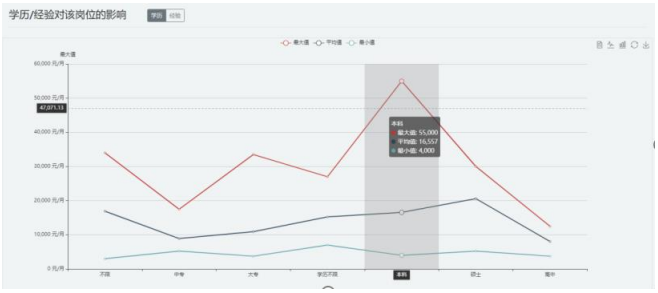


图 7 学历与薪资关系图

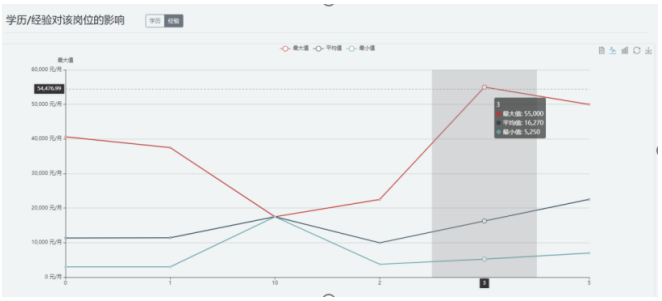


图 8 工作经验与薪资关系

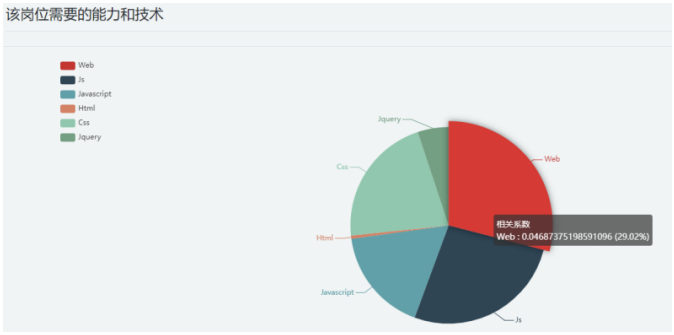


图 9 岗位所需技能分布图

图 10 薪资预测界面

（5）岗位推荐模块。岗位推荐模块采用NLP(自然语言处理)的CNN（卷积神经网络）算法[7]，其结构以及各层的实现如图11-12所示，将职位描述中提取的计算机专业技能词语对应的worddevctor依次排列的矩阵向量输入模型。

（6）该矩阵经过多层卷积层的卷积处理进行特征提取，卷积提取数据深度特征的的关键性技术，当提取的向量经过池化层进行数据降维，提取最重要的一个岗位技能输出到全连接层,全连接层组合这些特征来最终确定是哪一个分类并进行分类输出。当输出的结果与期望值相符时，输出结果。其示例图如图13所示。

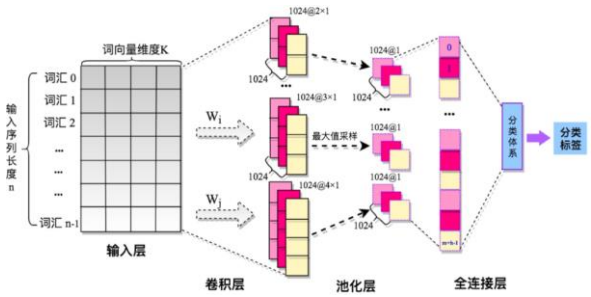


图 11 CNN 结构图

图 12 模型各层实现算法

图 13 岗位推荐界面示例图

4 结束语

大数据时代背景下求职者通过互联网访问招聘网站获取相关的岗位数据信息不能直观的分析岗位所需的要求，在此背景下设计一个系统帮助求职者精准定位、直观展示数据特征的系统尤为重要。本文介绍了互联网职位数据可视化系统的分析与实现，利用大数据技术抓取海量数据处理并进行可视化展示，通过该系统帮助求职以及者直观的了解互联网岗位的一些基本情况和透析岗位的需求，预测就业薪资待遇并结合求职者情况给他们推荐合适他们的岗位；也为高校根据社会需求全面制定培养方案。但系统还有很多的不足,还有优化的空间。例如网站的反爬技术日渐增强，后期将根据反爬技术的提高更新爬虫代码，根据用户的预测结果给他们推荐提升技术能力的平台以及优劣势分析。

参 考 文 献

[1] 葛 琳 ， 杨 娜 .Python 招聘数据分析 [J]. 计算机与网络 ， 2020,46(16):62-65.

[2] 何仕.当代中国大学生就业的经济学研究[D].福建师范大学,2014.

[3] 刘海,王晓钰,王政为,乔昭源,王星祎.基于 Hadoop 的大数据职位分析系统的设计与实现[J].信息与电脑(理论

版),2022,34(01):110-112.

[4] 李艳,丁国强,张庆.网络招聘数据可视化系统的设计与实现[J].信息与电脑(理论版),2021,33(01):112-115.

[5] 王慧玲. 招聘网站数据可视化分析平台的设计与实现[D].曲阜师范大学,2020.DOI:10.27267/d.cnki.gqfsu.2020.001252.

[6] 季杰,陈强仁,朱东.基于互联网大数据的招聘智能分析平台的设计和实现[J].内江科技,2020,41(05):47-48.

[7] 季长清,高志勇,秦静,汪祖民.基于卷积神经网络的图像分类算法综述[J].计算机应用,2022,42(04):1044-1049.

[8] 蒲云鹏.大数据时代基于 Python 的数据可视化研究[J].信息与电脑(理论版),2021,33(23):179-182.