Research on Recruitment Information Based on Text Mining Technology under the Digitization Perspective

Changhong Dong[1,a], Daoyuan Zheng[2,b], Junlong Wang[3,c], Zhangli Dong[4,d]

{xuptme@163.com[a], Victoriasuzdy123@163.com[b], a1978718441@163.com[c], 2623125596@qq.com[d]}

School of Modern Post, Xi'an University of Posts and Telecommunications, Xi'an, China [1]
School of Information Technology, Shanghai Ocean University, Shanghai, China [2]
School of Economics and Management, Northeast Agricultural University, Harbin, China [3]
Finance Section, 7th Division Hospital, Xinjiang Corps, Xinjiang, China [4]

Abstract. The development of the Internet has led to rapid progress in all walks of life, and domestic enterprises are no exception. There has been a sea change in talent recruitment, which has changed from traditional offline recruitment to online recruitment relying on recruitment apps or recruitment websites, such as Mileageplus, Wisdomlink Recruitment, Lagoo Recruitment, boss Direct Hire and so on. In this paper, we use a selenium module to crawl all the recruitment information released after June 2022 by Boss Direct Hire in a targeted way after data preprocessing and then analyze it using text mining methods. First of all, from the salary, work location, welfare benefits, and other aspects of concern to job seekers for visual analysis, followed by the use of TF-IDF algorithm for similarity analysis of popular jobs between similar jobs to consider other favorable factors, so that in the employment situation is grim, and the market competition is fierce, to provide reference and reference for job seekers, to help them find a more satisfactory job.

Keywords: text mining; python; Crawler; TF-IDF algorithm; Visual analysis

1 Introduction

During graduation season, which is the turning season of life, most graduates are facing employment problems. They are eager to find a satisfactory job. However, when logging on to the job boards, they are faced with countless job postings, which is not only time-consuming but also puts every graduate in a difficult situation if they deal with the job seekers manually[1]. However, the new global coronavirus swept through the world with overwhelming force in early 2020, making the employment problem even worse. According to statistics, the number of graduates in 2021 will be 9.09 million, an increase of 350,000 over 2019. The number of graduates in 2022 will reach 10.76 million, while the employment rate is only 23.6 percent, which allows graduates to face unprecedented difficulties and challenges[2]. Each graduate is in a very embarrassing "double superimposed" situation. Each graduating class is in a very awkward "double-stacked" situation. Academics are also aware of the seriousness of this problem and have made suggestions for graduate employment from different dimensions, such as recruitment data analysis[3][4][5], recruitment website analysis[6], recruitment position analysis[7], recruitment advertisement analysis[8][9], recruitment requirement analysis[10] and so on. However, previous studies have been limited

PMBDA 2023, December 15-17, Nanjing, People's Republic of China

DOI 10.4108/eai.15-12-2023.2345352

to one aspect of recruitment information and have not developed analyses it comprehensively. In view of this, this paper the use of Python language to write crawler code, the analysis of boss direct recruitment information, and the advantages of using the network recruitment information are mainly reflected in the following two aspects: first, the source of information is open, the information is convenient to obtain, the job seeker will be the depth of understanding of the different positions; the second is the content of the information true and reliable, the job seeker is interested in the position of the post-delivery and so on, to increase the employment opportunities.

2 Data crawl and analysis

The target website crawled in this paper is Boss Direct Hire (https://www.zhipin.com/); compared with other job boards, the advantages of this website are: (1)comprehensive and reliable job postings, and fast job updating; (2)direct can communicate with the boss or the person in charge; (3)fast processing of CVs; (4)suitable for a wide range of people. Crawling content includes eight fields of job positions, company name, salary, work location, experience requirements, educational requirements, benefits, and recruitment links. Crawling and writing at the same time are used, i.e., the crawled data is written to local Excel[11].

Because the amount of job information crawled is very large, if you use the get method in the requests module to request, even if you use the anti-crawl mechanism, such as in the headers inside the Host, user-agent, and cookies to identify uniquely, the recruitment site is also easy to identify, pop-up validation box in the page, and finally lead to the crawl failure, serious may be blocked up. To successfully crawl to the job information data, this paper uses selenium automation tools to simulate the browser crawl; not only can we get the page source code data, but also JS rendered data can also be obtained[12]. At the same time, the time module is imported for hibernation to prevent the IP from being blocked. To parse the data, regular expressions or XPath expressions are used to obtain the specified information in the web page source code, and the crawled data are written to the local Excel using the pandas module and xlwt module. A total of 14,485 pieces of data are written[13]. The data acquisition steps are initiating the request, responding to the data, parsing the data, and storing the data[14].

3 Data cleaning

To improve the quality of the data, data cleansing is done after storing the crawled data locally[15][16].

3.1 Format content cleaning

The content of the document will be cleaned into a unified format. For example, the recruitment position has inverted commas, experience, and educational requirements of some companies with double quotes bracketed by Excel's MID function to obtain the format of the table to make uniform adjustments. At the same time, the workplace, educational requirements, and educational requirements are organized into a separate Excel table for subsequent visualization.

3.2 Eliminating Duplicate Data

Due to the possibility of duplicate job postings from some companies, duplicate data was not filtered during the web scraping process. After reading the data, a deduplication operation was performed using the specific code df.drop_ duplicates(inplace=True), resulting in a final set of 14,314 valid entries.

3.3 Missing value processing

After deduplicating the data in the previous step, missing value processing was carried out using Excel tools. By applying conditional filtering, all empty values were located, revealing that the missing values were present only in the 'salary' and 'job_welf' columns, while the remaining columns were complete. The main methods for handling these missing values in the two columns include direct deletion and imputation. Due to the association between salary and education, experience, as well as job location, the direct deletion approach was employed for handling missing values in the 'salary' column. On the other hand, the imputation method was used for addressing the missing values in the 'job_welf' column, with the vacant positions being filled with "five insurances and one fund" for convenience. As a result, the final valid dataset comprised 14,284 entries.

3.4 Consistent treatment

To achieve uniformity in the preliminary processing of the data from the previous step for visualization, the following steps were undertaken. Firstly, the 'work_city' column was processed. While a few companies listed their location at the city level, the majority provided specific districts (or counties) and, in some cases, even precise locations such as "Baiyun District, Guangzhou" or "Tianning District, Qinglong Neighborhood Center, Changzhou." For the convenience of subsequent visualization, this study standardized the work locations at the city level. This was achieved through batch processing using the MID() function in Excel, ensuring the uniform treatment of work locations. Next, for the 'salary' column, on the one hand, there was a lack of uniformity in the measurement of salaries, with some being calculated on a daily, weekly, or hourly basis, such as "60-120 yuan/day," "2500-4000 yuan/week," and "120-150 yuan/hour." These were standardized to a monthly calculation to achieve uniform treatment of salaries.

On the other hand, the salary data was presented in ranges, and some entries needed to be standardized. For ease of subsequent processing, the MID() function in Excel was utilized to batch process the 'salary' column, splitting it into 'max_salary' (maximum salary) and 'min_salary' (minimum salary), and an additional 'average_salary' column was created.

4 Visualization of job information

Based on the previous data processing, this paper uses various text mining technologies of python to comprehensively mine the job information and extract the core information to support the subsequent analysis and application.

4.1 Visualization of job word cloud

Different enterprises for the recruitment of job naming are common. This paper on the job_name column to read and summarize through the jilbab module will be a string cut into a list of words by the word frequency of its generated word cloud. The results are shown in **Figure 1**. It is easy to find through the word cloud diagram, development engineer, Python, sales specialist, sales manager, data analyst, and other appearances in the forefront. In recent years, the IT industry for Python, data analysis and other job recruitment information is larger, especially in the Python development engineer position.

Fig. 1. Position word cloud map

4.2 Visualization of workplace

For the vast majority of job seekers, there is a high demand for work locations, using the Pyecharts module in Python to aggregate the work_city column and generate a map. The results are shown in **Figure 2**. The distribution of jobs in various cities is also a large difference. Beijing is clearly a one-trick pony, with 1004 job opportunities, followed by Shanghai, with 925 job opportunities, ranked third in Shenzhen, with 677 job opportunities, ranked fourth in Guangzhou, with 524 job opportunities, which is enough to illustrate the first-tier cities with high demand and a higher concentration of talent. Other new first-tier cities, second-tier cities, third-tier cities and other cities also have job distribution; compared with fewer job opportunities, the number of recruiters is also less.

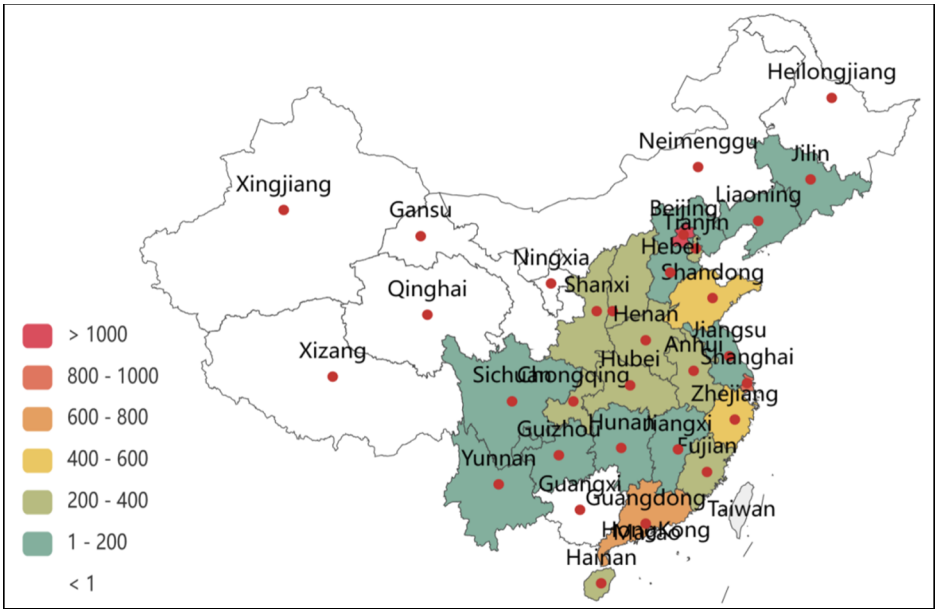

Fig. 2. Work site distribution diagram

4.3 Visualization of work experience

Work experience is also a measure of a candidate's professional competence for companies. By aggregating the EXPERIENCE column and visualizing it using the Pyecharts module, the results are shown in **Figure 3**. As can be seen from the figure, most companies prefer to recruit fresh graduates, accounting for more than 1/3 of all job openings, and with the increase in work experience, job opportunities are becoming less and less.

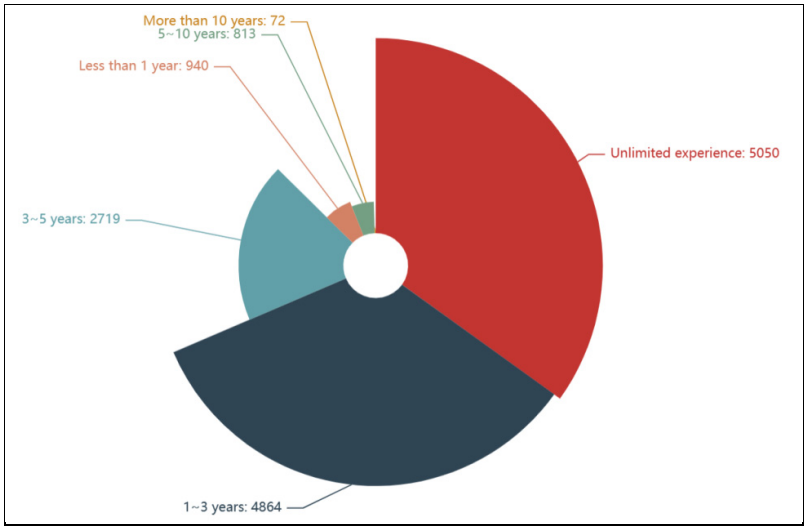

Fig. 3. Rose chart of work experience

5 Job Information Mining

Through the analysis of job information visualization, in order to provide more accurate and intuitive job search and employment information[10], we use the text mining method to mine the job information of popular positions and popular cities and take 18 companies with the position of "python development engineer" and the workplace of "Beijing" as the mining object. The 18 enterprises with the position of "python development engineer" and the workplace in "Beijing" are the mining objects. The positions are analyzed similarly by the IF-IDF algorithm (Inverse Text Word Frequency Index)[17-19], and the similarity of the positions is shown in **Table 1**. In order to show more intuitively the similarity of jobs between companies, this paper takes number 01 and number 02 as an example for visualization, as shown in **Figure 4**.

Table 1. Information on video and audio files that can accompany a manuscript submission.

No	01	02	03	04	05	06	07	08	09	10	11	02	13	14	15	16	17	18
01	1.00																	
02	0.00	1.00																
03	0.14	0.32	1.00															
04	0.05	0.14	0.08	1.00														
05	0.20	0.55	0.27	0.04	1.00													
06	0.18	0.21	0.21	0.03	0.30	1.00												
07	0.24	0.15	0.16	0.02	0.25	0.08	1.00											
08	0.29	0.26	0.19	0.01	0.26	0.15	0.05	1.00										
09	0.14	0.33	0.12	0.02	0.35	0.18	0.05	0.14	1.00									
10	0.17	0.27	0.15	0.03	0.38	0.14	0.18	0.18	0.33	1.00								
11	0.22	0.15	0.15	0.00	0.20	0.17	0.07	0.12	0.05	0.13	1.00							
12	0.13	0.12	0.04	0.01	0.11	0.03	0.17	0.17	0.15	0.13	0.02	1.00						
13	0.18	0.22	0.12	0.14	0.19	0.24	0.07	0.17	0.19	0.30	0.15	0.08	1.00					
14	0.21	0.22	0.15	0.08	0.19	0.25	0.09	0.18	0.22	0.30	0.12	0.10	0.18	1.00				
15	0.15	0.19	0.19	0.02	0.15	0.07	0.09	0.06	0.16	0.19	0.05	0.08	0.13	0.15	1.00			
16	0.04	0.12	0.12	0.01	0.05	0.18	0.08	0.02	0.18	0.05	0.03	0.03	0.07	0.13	0.03	1.00		
17	0.18	0.19	0.24	0.19	0.15	0.14	0.05	0.13	0.13	0.17	0.13	0.05	0.15	0.20	0.15	0.02	1.00	
18	0.27	0.19	0.21	0.03	0.15	0.24	0.06	0.19	0.09	0.18	0.20	0.03	0.11	0.19	0.13	0.03	0.16	1.00

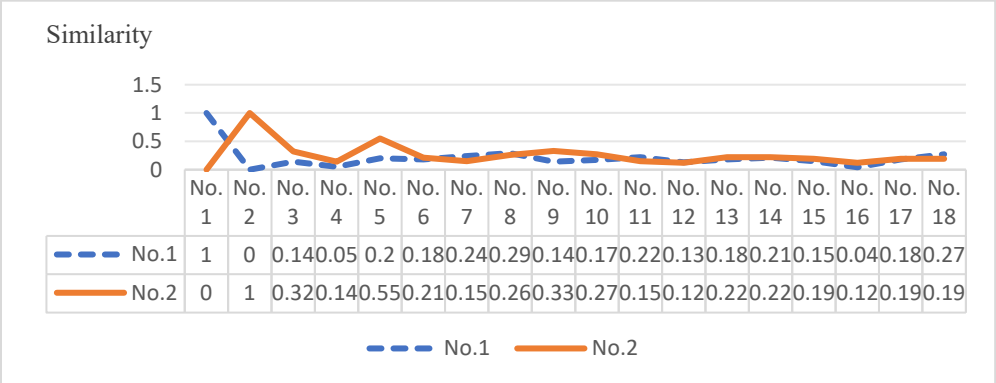

Fig. 4. Comparison chart of similarity between companies numbered 01 and 02

As can be seen from the table, numbers 01 and 08 company on python development engineer job similarity is higher, the similarity coefficient is 0.29, followed by 07 company job similarity is higher, the similarity coefficient is 0.24. It is not difficult to see from the Table that among these 18 companies, the job similarity between No. 02 company and No. 05 company is the highest, with a similarity coefficient of 0.55, and the similarity between No. 02 company and other companies is significantly higher than the similarity between No. 01 company and other companies. When job seekers face companies with high job similarity, they can make decisions by comparing salaries and benefits. For example, the number 02 company for Beijing byte jumping, 08 number company for Needle beacon digital, two enterprises, the educational requirements are a bachelor's degree, work experience are 3-5 years, welfare benefits are basically the same, but the first enterprise's salary is 20k-30k, the second enterprise's salary is 25k-35k, the difference in salary is the relatively flat case, you can through communication, integrated other personal favorable factors for decision-making, and finally find a satisfactory job.

6 Conclusions

This study is based on crawling boss direct job postings and analyzing them using text mining methods, which provides us with the opportunity to gain a deeper understanding of the opportunities and challenges of the current job market. By visualizing and analyzing key factors such as salary, work location, and benefits, we are able to better grasp the needs and preferences of job seekers and provide companies with more targeted recruitment strategies. At the same time, the similarity analysis using the TF-IDF algorithm helps job seekers find more suitable jobs among many positions, providing them with more targeted employment guidance. This study provides a new perspective and understanding of the current changes in the job market and provides useful references and insights for recruitment practices and job seekers. It is hoped that the study can provide a useful reference for scholars and practitioners in related fields and positively promote future research and practice. Future research can further explore how to combine AI technology to make more accurate recruitment recommendations, as well as how to predict the industry development trend through big data analysis to provide more comprehensive and in-depth support and guidance for both job seekers and enterprises.

Acknowledgments: Many thanks to the reviewers and editors for their comments.

References

[1] Wang Y.L. (2023) The current situation, causes and coping strategies of "slow employment" for college students. Ideological and Theoretical Education, 11:106-111.

[2] Xu, X.C., Tang, Y., and Hu Y.N. (2021) An Analysis of China's Economic Situation in 2020 and its Outlook in 2021. Economic Dynamics, 01: 35-48.

[3] Chen, J., Li, H.X., Xu, J.G. et al. (2022) City Size and Employment Shocks - An Analysis of Online Recruitment Data Based on the Post-New Crown Epidemic. Economics(Quarterly), 22(06):2125-2146.

[4] Geng, L., and Mao, Y.F. (2017) Construction, prediction and judgement of employment situation

of China's employment boom index--a study based on online recruitment big data[J]. Journal of Renmin University of China,31(06):24-35.
[5] Geng, L., and Mao, Y.Y. (2017) Construction, prediction and judgement of employment situation of China's employment boom index - a study based on online recruitment big data. Journal of Renmin University of China,31(06):24-35.
[6] Su, J., Dong, C., Su, K., & He, L. (2023). Research on the Construction of Digital Economy Index System Based on K-means-SA Algorithm. SAGE Open, 13(4), 21582440231216359.
[7] Mao, Y.F., and Zeng, X.Q. (2022) Impact of the New Crown Pneumonia Epidemic on the Employment of College Graduates-Empirical Evidence from Job Board Data. Academic Research, (01):104-110.
[8] Zhou W. (2008) A study of gender discrimination in urban employment in China--Taking the conditions of 300,000 newspaper job advertisements in Shanghai and Chengdu from 1995 to 2005 as an example. Politics and Law, 4:27-33.
[9] Liu, C.R.,Wang, H.J., and Wang, C. et al. (2020) Analysis of the demand for cyber security talents in the job market - based on the content analysis of corporate recruitment advertisements. Science and Technology Management Research,2020,40(03):182-187.
[10] Zhao, J.G. (2013) Employment Choice of College Students:Requirements and Implications of SME Recruitment. College Education Management,7(05):112-115+120.
[11] Li, Y.D., and Zhang S.Q. (2022) The ability of college students to cope with change in the era of artificial intelligence —— Research based on Internet recruitment information. Higher Engineering Education Research, 05: 93-98+110.
[12] Pei, L.L. (2022) To crawl and analyze Boss data based on Selenium framework. Shanxi Electronic Technology, 05: 66-68+76.
[13] Dong, C.H., Zhang, L., You, Y.R., and Chen, X.H. (2023) Research on Text Mining of JD Commodity Review Information Based on NLP. In: The 7th International Conference on Computer Science and Application Engineering (CSAE 2023), Wu Han, China.
[14] Gao, R., and Yang, X.P. (2020) See the establishment of academic exchange positions from the recruitment information. University Library work, 40: 57-60+90.
[15] Qian, M.H., Xu, Z.X., and Wang, Y.X. (2022) Research on enterprise competitiveness identification based on text mining of online recruitment information. Management Review. 34:150-156.
[16] Zhou, J.Y., and Feng, S.C. (2020) What kind of teachers are needed in the labor market?—— Mining and analysis of web-based recruitment information. Education and Economics, 36: 68-76.
[17] Qaiser, S., & Ali, R. (2018). Text mining: use of TF-IDF to examine the relevance of words to documents. International Journal of Computer Applications, 181: 25-29.
[18] Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. Information, 10(4), 150.
[19] Altınel, B., & Ganiz, M. C. (2018). Semantic text classification: A survey of past and recent advances. Information Processing & Management, 54(6), 1129-1153.