



Construction of Resume Recommendation Model for College Students Based on Machine Learning

Yingjie Bu^{a*}, Yaodong Jin^b, Xiaohua Zhang^c, Zhiyuan Zhan^d

Information Engineering School, Wenzhou Business College, Wenzhou, China

^a1978255193@qq.com*, ^b1835811581@qq.com
^c 20219049@wzbc.edu.cn, ^d1732140139@qq.com

Abstract. With the rapid growth of graduates and by the negative impact from the COVID-19, graduates have encountered many difficulties in employment. Therefore, it is particularly necessary to analyze the resume data of college graduates. Based on this need, this paper establishes a resume data extraction model, extracts college students' resume data into the database, saves it as structured data, and then conducts data pre-processing operations. Finally, through using machine learning algorithms to analyze the resume data, the job recommendation model established in this paper can greatly help college students find jobs.

Keywords: machine learning; college graduates; resume data; job recommendation model

1 Introduction

College graduates are a key talent resource for the country, and promoting high-quality employment for graduates is always an important task for colleges and universities[1]. According to statistics, the number of graduates in 2022 reached 10.67 million, which is 167 0000 more than the number of graduates in 2021[2]. Obviously, the employment situation for college students is becoming increasingly severe, and so it is necessary to improve the employment rate of college students in a targeted manner[3]. Through daily work and research on various literature, it is found that there is a short time interval between the production of resumes and employment for college students. That is to say, some graduates have already found jobs and signed tripartite agreements without waiting for guidance from teachers, and this will lead to some students taking a detour and signing jobs that are not suitable for them, which is not conducive to high-quality employment[4]. Therefore, this paper proposes a new approach to assist teachers in providing employment guidance to graduates as well as improving work efficiency.

2 Overall framework for resume data analysis based on machine learning

The overall analysis of resume data based on machine learning includes three stages: extraction of electronic resume information, data mining of resume information, and evaluation of job recommendation models. The detailed process is shown in Figure 1. Following collecting electronic resumes, a resume information extraction model is established. The basic personal information (name, gender, phone number, date of birth, graduation time, etc.), job intention, internship experience, organizational activity experience, award certificate, and other information from a single resume are extracted and stored in the SQL database. After storing the student's resume data into the database, a Python program is written to sequentially read the information of each resume, perform feature selection and preprocessing, and implement a content based reciprocal employment recommendation algorithm. The similarity calculation is performed between the information in the resume and the recruitment needs of the enterprise, and then matching recommendations are made for individuals and enterprises. The internship information in a student's resume can demonstrate their work ability and the competence of relevant enterprises. The part of displaying job content in enterprise recruitment information is the "job description" in the recruitment position information. When selecting talents, enterprises usually consider whether students' internship experiences match their job positions. If they are under the same conditions, a highly matched resume will definitely be valued by the human resources department of the enterprise. After establishing a job recommendation model, it is necessary to evaluate the model and divide the dataset into two parts. One part accounts for 80% of the total number as the training set, and the other part accounts for 20% of the total number as the test set. The model is trained and learned through the data from the training set, and then the prediction accuracy of the model is tested through the test set. Following the completion of the evaluation, the construction of the model is finished and then the model can be applied to college student employment guidance.

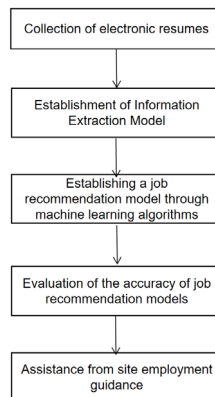


Fig. 1. Overall flowchart of resume data analysis

3 Establish a data extraction model to extract electronic resume data

There are two main methods for extracting data from electronic resumes. One is rule-based extraction, which uses regular expressions to determine rules for extraction. This method has high accuracy in extracting information like basic personal information[5]; Another method is based on statistics, which combines the word relationships of the entire text and has a high accuracy in extracting information such as internship experience and activity experience. For different data sections in electronic resumes, different extraction methods can be selected[6]. The paper combined the two natural language extraction methods.

3.1 Feature Analysis of Electronic Resume Information

In order to establish a comprehensive resume extraction model, it is firstly necessary to analyze the types of resume data for college students. Through reviewing relevant materials, it can be found that the content of college students' resumes mainly includes eight modules: personal basic information, educational experience, job intention, internship experience, project experience, organization and activity experience, award certificates, and other information. Analysis shows that among the eight resume modules, there are three modules: personal basic information, educational experience, and award certificates that have strong regularity. It is more appropriate to use rule-based data extraction methods, while others are more suitable to use statistical data extraction methods.

3.2 Preprocessing of Electronic Resume Data Information

After analyzing the features of resume information, it is also necessary to preprocess the electronic resume data information. The first step is to use Python to write a program that converts the data information into text information and saves it in a CSV file, thereby reducing the impact of different formats on different resumes; The second step is to use Python's Pandas library function to read the CSV file information, imports the re library to write regular expressions to remove punctuation mark; The third step is to use Python's segment method to further process the text data without punctuation mark. Here NLP word segmentation will synchronously perform named entity recognition. Named entity recognition can identify the named entities in the text data, while part of speech tagging can mark the words in the sentence; The fourth step is to remove the meaningless words such as adjectives and auxiliary words.

3.3 Establishment of Extraction Model

Before establishing the resume information extraction model, it is also necessary to block the text information and classify the attribute information of the same module

into one category, and the SVM algorithm is used. After dividing the blocks using SVM algorithm, based on the results of feature analysis, a combination of rule-based and statistical model-based methods was used to establish a resume information extraction model. Firstly, regular expression was used to construct extraction rules to extract attribute information from the four modules of personal basic information, educational experience, award certificates, and job search intention in the resume module, such as name, major, school, and GPA. Following extracting the modules with strong regularity, the HMM algorithm model was used to extract attribute information of modules such as internship experience, activity experience, and project experience.

After the establishment of the resume information extraction model for college students, 1000 resumes of senior graduates were collected to test the accuracy of the model, as shown in Figure 2. The accuracy test of resume data extraction showed that the average accuracy, average recall rate, and average F1 value of personal basic information and educational experience modules were significantly higher than the three values of internship experience, indicating that the information extracted by regular expressions was more accurate. However, text information such as internship experience, organization and activity experience contains complex information such as internship content, activity content, organization and achievements, resulting in low accuracy in extracting information. Therefore, the HMM algorithm model needs to be improved. The average accuracy, average recall, and average F1 value of each resume module in the resume data extraction model are above 80%, indicating that the overall accuracy of the model is relatively high.

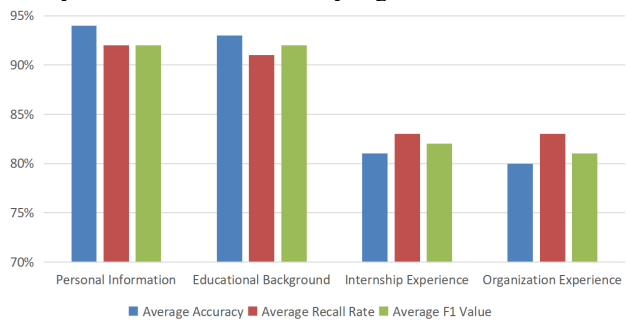


Fig. 2. Resume data extraction accuracy testing

4 Establishment of job recommendation model

The establishment of a job recommendation model for college students can provide auxiliary decision-making for teachers related to university employment guidance, which is beneficial for teachers to better guide according to their aptitude, propose career development paths suitable for each student's characteristics, and provide comprehensive employment guidance[7]. A resume data extraction model is used to extract student resume data into the SQL database, providing a data source for subsequent model establishment.

4.1 Establishment of recruitment information database

In order to match each resume data with the employer, it is necessary to establish a recruitment information database for the employer. The web crawler is used to crawl the enterprise recruitment information from the school employment website, Liepin recruitment and other platforms[8]. The main fields are company name, position name, salary, work content and location. It is found through research that the recruitment information on most websites is saved in the json string of the web page. In this way, the paper first calls Python's requirements and beautiful soup4 library, writes a web crawler program to extract data, and then uses the loads method of the json module to parse the json string. Finally, the crawled data is saved to the SQL database.

4.2 Content based reciprocal employment recommendation algorithm

Concerning the content of the algorithm, the random forest algorithm is used to build a reciprocal employment recommendation model which combines the needs of matching the students and enterprises' need. Set student satisfaction as X_1 , select four characteristic attributes of enterprise location, position name, enterprise type and position salary level as student satisfaction attributes, use Python to call Random Forest Classifier method of sklearn machine learning database, use random forest algorithm to calculate student satisfaction, and set the weight as W_1 . Similarly, selects political outlook, education background, profession and salary level as the enterprise satisfaction attribute, sets it as X_2 , calculates the X_2 value using the random forest algorithm, and sets the weight as W_2 . The calculation formula for total satisfaction X is:

$$X = W_1X_1 + W_2X_2 \quad (1)$$

After calculating the X value of a resume data and a recruitment information, a for loop is written to traverse each recruitment information, so as to calculate the X value of the resume and each recruitment information, and then arrange the X value from top to bottom to output the top five values. These five recruitment information are more suitable for the students of this resume. Through observation, it was found that several recruitment information with high matching degree all have common characteristics, such as job intention for electrical engineers, intention to work in Guangdong, intention to work in a private enterprise, and intention to income in the range of 4K-7K.

4.3 Evaluation of Job Recommendation Model

The accuracy of the recommended model is evaluated by calculating the average absolute error of MAE and the root mean square error of RMSE. For comparison, the paper implemented a content based recommendation employment model using traditional recommendation algorithms. In terms of data, as there is currently no standard resume recommendation dataset, the paper introduced a questionnaire survey on 1000 senior students, collecting electronic versions of resumes and job information

submitted. These data information were used as experimental datasets, of which 800 were used as training sets and 200 were used as testing sets.

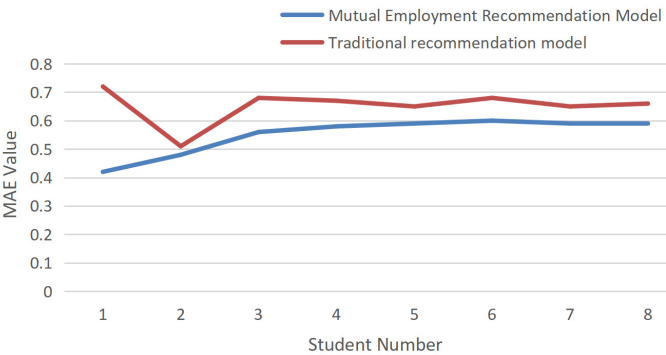


Fig. 3. Comparison of MAE values with line chart

Run the resume data extraction model to extract resume data into the SQL database Data1, write a program to read the electronic resume training set data from the Data1 database, and then run the reciprocal employment recommendation model based on the random forest algorithm and the traditional recommendation algorithm model. After the training of the two models is completed, use the test set data to test them respectively, and calculate the MAE average absolute error value and RMSE root mean square error value. As shown in Figure 3, it is evident from the graph that the MAE values of the reciprocal employment recommendation model and the traditional recommendation model both decrease first, then increase, and finally tend to stabilize. After experiments, the RMSE values of the two models also show the similar trend as the MAE values, and the RMSE values of the reciprocal employment recommendation model are lower, which has certain advantages compared to the traditional recommendation model.

Concerning the recommendation diversity testing, ILS indicator is used. ILS is a proprietary indicator for measuring the diversity of push recruitment information, and the method used in this article calculates the variation of ILS in the diversity of push recruitment information lists when the amount of push recruitment information varies[9]. Following the analysis of the recommendation diversity effect, the results are shown in Figure 4.

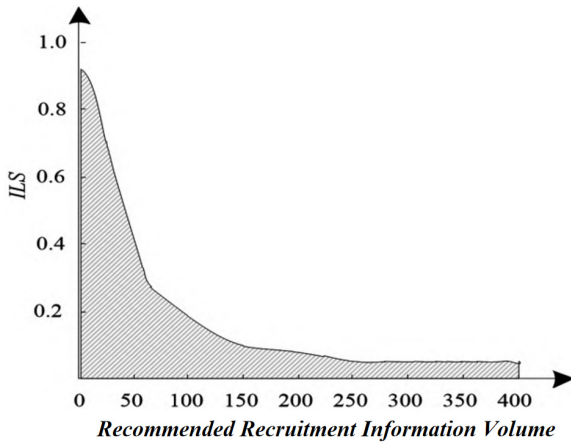


Fig. 4. Recommendation Diversity Test Results

Through analyzing Figure 4, it can be seen that the ILS value is inversely proportional to the amount of recruitment information pushed. Before pushing about 60 recruitment information, the ILS value rapidly decreased from around 0.9 to around 0.25, and its decline rate was relatively fast. As the amount of recruitment information pushed increases, although the ILS value continues to show a downward trend, the overall ILS curve remains flat. Between 250 and 400 recruitment information pushed, the curve remains in a straight line state, with an ILS value of around 0.25. The above results indicate that the ILS value of the recruitment information pushed by this method is small, and the diversity of the recruitment information pushed is good.

4.4 Application Testing

Using the employment rate of students as a measurement indicator, the test the employment rate of fresh graduates in different majors of the university is carried out and the resulat is compared with that at the same period last year[10]. Last year, the traditional campus on-site recruitment method was used, while this year, the personalized recommendation method based on this article and the traditional campus on-site recruitment method are both used.

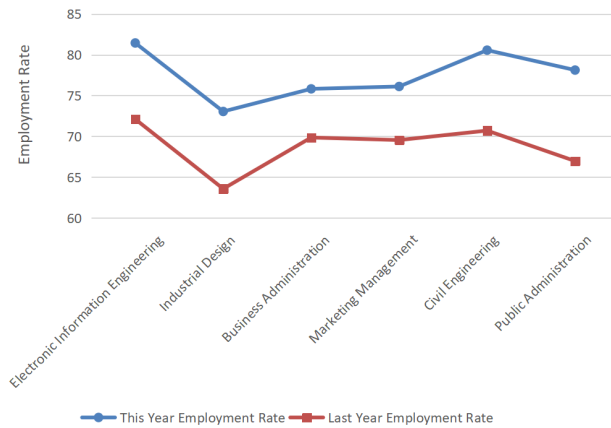


Fig. 5. Comparison of This-year Employment Rate and Last-year Employment Rate

From figure 5, it can be seen that after using the method presented in this paper, the employment rate of various majors in the university has improved. This result indicates that the application of this model can effectively improve the employment rate of fresh graduates in universities and has good practical application effects.

5 Conclusion

This paper introduced a resume data extraction model that can quickly extract students' electronic resume data and store it in a structured form in a database, saving teachers a lot of time in reviewing resumes. Teachers can clearly see the key points of each student's resume. The reciprocal employment recommendation model based on random forest algorithm established can effectively provide students with appropriate employment directions. The combination of the two models can enable teachers to have a more comprehensive understanding of students' employment needs and provide targeted employment guidance. The method has also been applied to a certain university. The experimental results show that the method has good information extraction ability and the effectiveness of recruitment information pushed for students is strong. After application, the employment rate of fresh graduates in the university has significantly increased, and good practical application results have been achieved.

References

1. Li Ping, Jiang Junyi, "Research on Improving the Efficiency of Employment Services for College Students Driven by Big Data," in *Journal of Science and Technology Economics*. Vol. 27, pp161-162, April, 2019.
2. Chen Xuehui, Chen Shaozhen, Wang Peibin, et al., "Design and Implementation of a Content Based Information Recommendation System," in *Computer Knowledge and Technology*. Vol.15, pp.14-16, September, 2019.

3. Huang Chao, "Extracting Word Segmentation Dictionaries from Text Based on Statistical Methods," in *Computer Knowledge and Technology*. Vol.16, pp.213-214, April,2020.
4. Zheng Liang, "Exploration of Employment Guidance for College Students in Vocational Colleges from the Perspective of Mobile Internet," in *Journal of Zhengde Vocational and Technical College*. Vol.15, pp.50-52, February, 2018.
5. Liu Jinyan, Wang Dongqing, Lin Guocong, "Classification of Screw Locking Results Based on SVM Algorithm," in *Journal of Qingdao University (Engineering Technology Edition)*. Vol.34, pp.21-26, March,2019.
6. Xiao Xiangfeng, Lu Na, Yu Lingna, Wu Jingyi, Zhong Wenbin, "Analysis of the Current Situation and Improvement Strategies of College Student Internship Recruitment Platform," in *China's Collective Economy*. Vol.15,pp.111-112, November,2018.
7. Zhu Xianjun, Hong Yu, Huang Yalin, et al., "Application of HMM based algorithm optimization in Chinese word segmentation," in *Journal of Jinling University of Science and Technology*. Vol.35,pp.1-7, March,2019.
8. Xu Jinyang, Zhang Gaoyu, Wang Manxi, Lou Huanyu, Xue Weicheng, Mao Xiaoyu, "The Similarity Algorithm for Bidirectional Matching between Positions and Resumes on Recruitment Websites," in *Information Technology*. Vol.51, pp.43-46, August,2016.
9. Yao Jianbin, Zhao Longwei and Li Hairui, "An interpretable hybrid employment recommendation algorithm," in *Information Systems Engineering*. Vol.6,pp.142-144, June, 2019.
10. Gu Nannan, Feng Jun, Sun Xia, Zhao Yan and Zhang Lei, "Chinese resume automatic parsing and recommendation algorithm," in *Computer Engineering and Applications*. Vol.53,pp.141-148, December 2017.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

