Personalized employment recommendation method based on semantic matching of requirements

Weihong Wang\textsuperscript{a} Qian Feng\textsuperscript{b} Fuxiang Zhou\textsuperscript{c} Yuhui Cao\textsuperscript{d} Hongyan LV\textsuperscript{e}

Hebei University of Economics and Business

\textsuperscript{a}wangwhs@163.com 18131112669 Hebei, China
\textsuperscript{b}15733100624@163.com 15733100624 Hebei, China
\textsuperscript{c}1257058169@qq.com 18330167149 Hebei, China
\textsuperscript{d}412503238@qq.com 18131112668 Hebei, China
\textsuperscript{e}981220135@qq.com 1522719319 Hebei, China

ABSTRACT: Aiming at the lack of personalized recommendation method in existing employment application system, a personalized employment recommendation method based on semantic matching of requirements (PERM-SMR) was proposed in the paper. Firstly, the web crawler technology is used to collect the recruitment information and clean the data. Secondly, word2vec method was selected for corpus training to obtain the word vector model. Thirdly, we performed semantic analysis on the demand information of both sides, calculated the weight of the feature words in the text and generated text vector, and then carried out the requirement information matching and generated the recommendation list. Finally, it is verified by experiments that the PERM-SMR is more accurate and effective, and provides better employment guidance for graduates.

CCS Concepts

- Information systems → Information retrieval → Retrieval tasks and goals → Recommender systems

1. INTRODUCTION

According to the "Statistical Bulletin on the Development of National Education in 2017" issued by the Ministry of Education, the number of college graduates nationwide reached 7,792,300, of which the number of graduates of all college was as high as 7,358,300, an increase of 316,500 over the previous year, an increase of 4.49%. There are 600,000 masters and doctors.

Graduate employment is closely related to family interests and social construction, and has become a key issue of the Ministry of Education and the state. According to the survey, graduates mainly seek employment through the campus recruitment seminar, campus employment information network and recruitment website. However, the campus recruitment information conference and the campus employment information network involve a narrow range of employment information, more and more graduates are used to applying for jobs on the recruitment websites. Recruitment websites provide employment platforms for graduates. There are many kinds of recruitment information, so graduates are easy to lose themselves when they are not clear about their employment direction. It is difficult to accurately obtain suitable positions from a large number of recruitment information. Recommended technology can effectively solve the problem of information overload and provide personalized employment guidance and recommendation for graduates by establishing a graduate preference model. At present, the employment recommendation research still has some shortcomings in the
semantic analysis of the demand between graduates and enterprises. Based on the combination of data collection, cleaning, semantic analysis and recommendation techniques, this paper studies the whole process of employment recommendation and proposes personalized employment recommendation method based on demand semantic matching.

2. RELATED RESEARCH
Deep learning has shown strong learning ability in the field of image recognition and speech recognition. Researchers have developed a strong interest in the application of deep learning to the field of natural language processing. Based on the NLNM (Neural Network Language Model) proposed by Bengio [1] and the Log Linear model of Hinton [2], Mikolov et al. [3] proposed the word2vec model for calculating word vectors in 2013, which has strong semantic expression ability. And successfully applied to emotion recognition, text classification, text recommendation, and so on. Hongyang et al. [4] used the word2vec word model to represent the text in a vector, and combined SVM and RNN to analyze the text emotion; Chengzhang et al. [5] proposed a text digest algorithm based on word2vec; Gao Mingxia et al. [6] proposed a Chinese short text classification method based on the word2vec word model, which is significantly better than the traditional word bag model classification; Zhu Xuemei [7] proposed a micro blog recommendation framework based on word2vec topic extraction for users.

In the field of employment recommendation, scholars also have a series of research results. Singh et al. [8] matched the graduate resume information with the recruitment information content to generate the recommendation results; Wu Di [9] used PageRank algorithm to obtain the enterprise "job search index" and used SimRank algorithm to calculate the similarity as the recommended employment information for graduates; Chen Yu [10] used ontology reasoning to model the employment information, and compares six common similarity calculation methods. Yin Chuancheng [11] comprehensive employment characteristics of graduates and school history information propose a reciprocal employment recommendation model based on global preference. Most of the research content is based on the matching of the job search information attribute value and the recruitment information attribute value, and lacks the semantic matching of the demand information of both parties.

3. PERM-SMR
Employment of graduates is a two-way matching problem. It is necessary to consider the employment intention of graduates and the needs of recruiting units. At present, most of the current research is based on attribute values and lacks expression on the semantic information of the requirements. Aiming at this problem, this paper proposes an employment recommendation method based on demand semantic matching. The process of this method is shown in Figure 1.

![Flow chart of employment recommendation method](image-url)
3.1 Data Collection
In this paper, the Nutch open source crawler framework based on the Java platform is used to collect recruitment information data from major mainstream websites and store the data in the database. Recruitment information data includes work location, ability level, work experience, position information, salary, etc.

3.2 Data Preprocessing
The data preprocessing module includes data cleaning, remove stopping words and participle.

1. Data cleaning. The recruitment information data obtained on mainstream recruitment websites through web crawling technology has high repetition rate, so carry out the weight reduction operation on data. Recruitment information are in text format, and we need to pay attention to the encoding format of the text and convert it into a coding format that computer can recognize. We need to extract the content of the recruitment information, including academic requirements, professional requirements, and competency requirements, and the standardized processing and manual marking of these information.

2. Remove stopping words. The stop word means that in the process of text processing, words that are useless to the experimental results need to be eliminated, which can reduce the storage space of the computer and improve the retrieval efficiency of the text [12]. Reducing the number of stop words in graduate employment demand information and recruitment information can effectively help us to increase the keyword density[7]. In this paper, the data processing is carried out by using the stop word list of Harbin Institute University.

3. Participle. Word segmentation is the process of splitting a text statement into words according to a certain rule method. The current Chinese word segmentation method are based on strings, rules, and statistics. Through comparative analysis, this paper adopts the precise pattern in the dictionary-based jieba word segmentation tool as the word segmentation engine, and carries out part-of-speech tagging, which has achieved a good segmentation effect.

3.3 Requirement Information Vector Representation
Some words have different proportions in different types of texts, and the demand information indicates that the feature words should be highlighted to fully express the semantic information. Based on this idea, in this paper, the TF-IDF algorithm is used to calculate the weight of feature words, and the importance of feature words to different demand information is highlighted. TF-IDF combines the frequency of words and the frequency of reverse files to evaluate the importance of feature words in a text.

The requirement information text set is represented as $D=\{d_1,d_2,\ldots,d_i\}$, $i$ is the number of text, each requirement information is represented as $d_i=\{t_1,t_2,\ldots,t_n\}$, $n$ is the number of feature words in the requirement information, $m$-dimensional vector corresponding to each word is obtained by training the word2vec model, represented as $t_i=\{v_1,v_2,\ldots,v_m\}$. The TF-IDF calculation formula as follows.

\[
idf(i) = \log \left( \frac{i}{n_i + 0.01} \right) \tag{1}
\]

\[
TF-IDF = \frac{tf(j,d) \times idf(j)}{\sqrt[\sum t_f(j,d) \times idf(j)}] \tag{2}
\]

In this paper, the TF-IDF value is recorded as $\omega$, and each job information text vector is expressed as follows.

\[
d_i = \sum_n t_i \omega_i \tag{3}
\]

The calculation method of the graduate job search information text vector is consistent with the recruitment information text, and the text vector is expressed as follows.

\[
S_c = \sum_h t_c \omega_c \tag{4}
\]
b is the number of feature words in the text, \( t_c \) is the word vector of the feature word \( c \) in the text, \( \omega_c \) is the weight of the feature word.

### 3.4 Matching Degree Calculation

The matching degree calculation is a very important part in the recommendation system, and the recommendation result is a list of recommendations generated based on the matching degree calculation result. Currently, there are many similarity calculation methods, through comparative analysis, this paper adopts the distance-based cosine similarity method to calculate the matching degree, which focuses on the application in vector space. The calculation formula is as follows.

\[
sim(d,s) = \frac{d \cdot s}{\|d\| \times \|s\|} = \frac{\sum d_i s_i}{\sqrt{\sum d_i^2} \sqrt{\sum s_i^2}}
\]  

(5)

The cosine similarity value ranges from \([-1, 1]\). The closer the value is to 1, the higher the similarity between the recruitment information and the student information. The TOP-K recruitment information is selected to generate the recommendation list.

### 4. EXPERIMENTAL RESULTS AND EVALUATION

In this paper, 1100 pieces of recruitment information after data preprocessing were selected, and the job search information of 100 graduates of Hebei University of Economics and Business 2016 was randomly selected as a test set. In addition, the recruitment information has industry categories.

#### 4.1 Evaluation Index

In this paper, Precision, Recall and F1 values are used as evaluation criteria.

1) **Precision**: this paper collects and randomly selects a list of job information that graduates are interested in in the test set, recorded as \( w(t) \). The accuracy rate is defined as follows.

\[
P = \frac{z(t) \cap w(t)}{z(t)}
\]  

(6)

\( z(t) \) is a list of recommendations recommended for graduates using the algorithm of this paper.

2) **Recall**: the recall rate is defined as follows:

\[
R = \frac{z(t) \cap w(t)}{w(t)}
\]  

(7)

3) **F1**: In order to fully evaluate the results, the F1 value takes into account the accuracy and recall rate, and the calculation formula is as follows.

\[
F1 = \frac{2PR}{P + R}
\]  

(8)

#### 4.2 Experimental Results and Analysis

This paper selects the different number of recommended recruitment information, records the accuracy and recall rate of the recommendation results, and calculates the F1 value. In addition, compared with the mean text vector representation, the results are shown in Figure 2, Figure 3, and Figure 4. It can be seen from the figure that the accuracy, recall rate and F1 value of the TF-TDF weighting method are higher than the average method. In Figure 2, the trend of accuracy is to rise first and then decreases, basically reaching the maximum when the recommended number of recruitment information is 30. In Figure 3, the recall rate increases as the number of recommendations increases. Considering the accuracy rate and the recall rate, the trend of F1 changes. As can be seen from Figure 4, the F1 value increases with the number of recommendations for recruitment information. When the recommended list length is 30, the effect is optimal. It can be seen that when the recommended number of recruitment information is 30, the overall recommendation effect is the best, which can provide a certain employment reference for graduates.
5. CONCLUSION
Based on the data employment, cleaning, semantic analysis and recommendation techniques, this paper studies the whole process of employment recommendation and proposes personalized employment recommendation method based on demand semantic matching (PERM-SMR). The method performs two-way matching through semantic analysis. The experimental results show that this method can provide employment guidance for graduates and alleviate certain employment pressure. This paper believes that combining students' comprehensive ability indicators, analyzing students' potential needs and improving recommendation quality is the next research direction.

ACKNOWLEDGMENTS
This work was financially supported by Returned Personnel to Study Abroad of Preferred funding project (C2015003042) fund, Hebei Provincial Natural Science Foundation Youth Project (F2015207009) fund, Hebei University of Economics and Business Important Research Project (2014KYY09, 2017KYZ11) fund, Hebei University of Economics and Business Important Education Research Project (2018JYZ06) fund.

REFERENCES
[1] Bengio Y, Ducharme, R éjean, Vincent P. A Neural Probabilistic Language Model. [M]// Innovations in Machine Learning. Springer Berlin Heidelberg, 2006.
[2] Mnih A . Three New Graphical Models for Statistical Language Modelling[C]// International Conference on Machine Learning. ACM, 2007.

[3] Mikolov T , Sutskever I , Chen K , et al. Distributed Representations of Words and Phrases and their Compositionality[J]. Advances in Neural Information Processing Systems, 2013, 26:3111-3119.

[4] Hongyang Xu, Hui Lu, Guowei Yang and Cong Zhang. Sentiment Analysis of Chinese Version Using SVM & RNN. 2017. 6th International Conference on Information Engineering Proceedings.

[5] Chengzhang X , Dan L . Chinese Text Summarization Algorithm Based on Word2vec[C]// Journal of Physics Conference Series. Journal of Physics Conference Series, 2018.

[6] GAO Mingxia,LI Jingwei. Chinese shot text classification method based on word2vec embedding[J]. Journal of Shandong University(Engineering Science), 2018, 48(6):0-00.

[7] ZHU Xuemei, Micro-blog Recommendation Based on word2vec Topic Extraction[D], Beijing Institute of Technology, 2014.

[8] Leskovec J , Singh A , Kleinberg J . Patterns of Influence in a Recommendation Network[J]. 2006.

[9] WU Di, The Design and Implement of Graduate Occupation Recommending System[D], Dalian University of Technology, 2010.

[10] Chen Shanshan, Research of Graduates’ Employment Recommendation System Based on Semantic[D], Wuhan University of Science and Technology, 2014.

[11] YIN Chuancheng, Research on Employment Recommendations and Algorithms for College Graduates[D], Shandong Normal University, 2016.

[12] Yin Rui, Research on Paper Clustering Algorithm Based on Semantic Similarity[D], Dalian University of Technology, 2009.