Research on talent data analysis method based on Text Mining

Cao Jingjing[a], Wang Ye[b], Wang Ying[c], Zhang Xiaoxia[d], Zhang Xue[e]

Exchange, Development & Service Center for Science &Technology Talents, The Ministry of Science &Technology (MoST), Beijing, China
[a]caojj@sttc.net.cn, [b]wangye@sttc.net.cn, [c]wangy@sttc.net.cn, [d]zhangxx@sttc.net.cn, [e]zhangx@sttc.net.cn

Abstract: *in order to solve the problem of low clustering contour coefficient caused by inaccurate keyword extraction of talent data information, a talent data analysis method based on text mining is proposed. Through the preprocessing of word segmentation and stop words, the text set of talent data is established, and the keyword graph is constructed by text mining, and the information keywords are obtained according to the weight iteration results. The keywords in this paper are transformed into the form of multi-dimensional vector, and the similarity is calculated to get the results of text analysis. The experimental results show that the contour coefficient of the proposed method is 0.736, which is 0.267 and 0.221 higher than that of the K-means and single pass methods, respectively. The design method of this paper has a reasonable clustering performance, which is suitable for talent big data analysis.*

Keywords: *text mining; Talent data analysis; Talent demand; Text representation; Text segmentation; Clustering effect*

1. Introduction

The imbalance of talent supply and demand has always been a key problem to be solved in social development. From the aspect of talent supply, college graduates have great pressure on employment every year, and face employment problems. From the perspective of talent demand, a large number of posts can not recruit suitable personnel, so there is a large demand for talents. In the face of this problem, the use of talent data analysis method can deeply tap the potential content of talent data, making talent more in line with the actual needs of enterprises. Talent data analysis is of great significance to solve the imbalance of employment supply and demand. Due to the huge amount of talent data and the existence of a large number of unstructured data, it is difficult to analyze talents. Text mining is the application of data mining in the field of text. Through the recognition and analysis of text semantics and link structure, it can mine deeper information and provide more meaningful analysis value [1]. Using text mining can analyze the potential connection of talent data, which is convenient for job seekers and recruiters to grasp the key information and provide reference information for both sides of recruitment and application. Therefore, this paper applies the text mining technology to the talent data analysis method, and proposes a data analysis method, which can guide the talent to apply for a job, so as to alleviate the contradiction between talent supply and demand.

2. Talent data analysis method based on Text Mining

2.1. Preprocessing of text segmentation

Before analyzing the talent data, the data is preprocessed. The collected talent data information contains a large number of non text parts, mainly including web tags and symbols. Before text mining, this part of data is first removed. Word segmentation technology divides long sentences into independent words, which is the basis of text mining. Firstly, Jieba segmentation is used to segment the information in this paper, and the stop words are removed. The thesaurus contains more than 20000 words, which can realize Chinese text search and segmentation. After inputting the talent data information, the sentences containing special characters are separated, the DAG word graph is generated, and the global probability is calculated. According to the calculation results, find the maximum word frequency, mark the word frequency combination and output it [2]. Because the talent data contains a large number of professional vocabulary, Jieba's own thesaurus can not accurately describe the job characteristics. On the basis of Jieba's own thesaurus, this paper establishes three user-defined dictionaries, which are respectively used to describe professional, technical and professional terms, so as to improve the accuracy of word frequency segmentation and combination. The talent data information processed by word segmentation is a set of words, but there are meaningless function words and notional words in this set. These words have no value for talent analysis, so we should establish stop words list, delete stop words and improve calculation efficiency.

2.2. Extracting keywords of talent data information based on Text Mining

After word segmentation and de stop word preprocessing, the talent data information text is divided into independent words. At this time, the number of words is still large, so it is difficult to analyze directly. In order to improve the accuracy of talent analysis, it is necessary to further extract information keywords. Based on text mining, this paper extracts keywords of talent data information to provide the basis for data analysis. Key words are usually the words that appear most frequently in a text, that is, the more important the word is, the more important it is in the text. Considering the specific situation of talent data information text, this paper selects the product of word frequency and inverse document frequency as the parameter to measure the importance of words. This parameter can be expressed as:

$$\begin{cases} \gamma = w\varphi \\ w = \dfrac{n_1}{n_2} \\ \varphi = \log \dfrac{n_3}{n_4 + 1} \end{cases} \tag{1}$$

In formula (1), γ is the keyword parameter; w is word frequency; φ represents inverse document frequency; n_1 and n_2 denote the number of entries of a certain class and the number of all entries; n_3 and n_4 represent the total number of texts and the number of texts containing entries. According to the relationship between words, the ranking relationship is established, and talent information keywords are extracted according to the weight [3]. The specific steps of keyword extraction are as follows: firstly, the preprocessed text is collected into a document and divided into sentences; Secondly, the independent sentences are segmented and the words are extracted. According to the relationship between the words, the candidate keyword map is constructed; Then, the initial weight is given in the word graph and iterated; Finally, according to the iterative results, the weight value of the word is obtained, and the keyword extraction result is obtained after ranking according to the weight order.

2.3. Establish talent data cluster analysis model

The key words of talent information extraction are still the form of text. Only by transforming them into multi-dimensional vector can they be recognized by computer and then cluster analysis can be carried out. The higher the similarity is, the closer the distance is. Each word can be expressed as two vectors of the center word and the background word. Given the probability of generating the background word, it can be obtained by the vector inner product operation. Through each word of text sequence, we can get the jump word model of training word vector. By adjusting the random gradient, the model is iterative and the word vector matrix is updated. When the loss function is the minimum, the word vector is the vector matrix needed for clustering analysis. The cosine similarity of vector matrix is calculated and the clustering analysis results are obtained. The formula of similarity is as follows:

$$\cos\left(\theta_1,\theta_2\right)=\frac{\sum_1^m \theta_1\theta_2}{\sqrt{\sum_1^m \theta_1^2}\sqrt{\sum_1^m \theta_2^2}} \quad (2)$$

In formula (2), θ_1 and θ_2 represent text vectors; m is the dimension of the vector. In the above formula, the greater the similarity, the more consistent the proportion of keywords in the text vector, the more matching the clustering, the better the effect. By clustering training the text vector, the analysis results of talent data can be obtained.

3. Experiment

3.1. Experimental preparation

This paper proposes a talent data analysis method, in order to verify the effect of the method, the following design experiment to test. The talent data collected in this experiment comes from a recruitment website, which has a large amount of information, more registered users and enterprises, and rich recruitment information. Firstly, Python is used to crawl the talent data information, and 17526 pieces of data are collected. Then, the collected data are cleaned, and the duplicate and wireless data are eliminated. After cleaning, there are 16472 pieces of data left. Based on the above data acquisition and preprocessing results, the application effect of the talent data analysis method proposed in this paper is tested.

3.2. Experimental result

In order to verify the clustering effect of the talent data analysis method based on text mining proposed in this paper, this experiment uses the contour coefficient to evaluate the clustering effect. The coefficient can evaluate the impact of different methods on the clustering results. The specific calculation formula is as follows:

$$\lambda=\frac{\alpha\left(x\right)-\beta\left(x\right)}{\max\left\{\alpha\left(x\right),\beta\left(x\right)\right\}} \quad (3)$$

In equation (3), λ is the contour coefficient; x is the sample; $\alpha\left(x\right)$ and $\beta\left(x\right)$ represent the dissimilarity within and between clusters. The closer the coefficient is to 1, the more reasonable the clustering is and the better the effect is. The talent data analysis methods based on K-means and single pass were selected as the control group, and the clustering effects of the three methods were compared. The experimental results are shown in Table 1.

Table 1: Comparison results of contour coefficients of different methods

Number of clusters	Analysis method of this paper	Analysis based on k-means	Analysis method based on single pass
2	0.689	0.452	0.523
3	0.724	0.462	0.524
4	0.736	0.469	0.515
5	0.745	0.458	0.518
6	0.762	0.452	0.506

According to the experimental test results in Table 1, under the condition of different number of clusters, the profile coefficients of each analysis method are significantly different. Under the condition of the same number of clusters, the contour coefficient of the analysis method based on text mining is higher than that based on K-means and single pass. Taking the case that the number of clusters is equal to 4 as an example, the contour coefficient of this design method is 0.736, which is 0.267 and 0.221 higher than the two control methods respectively. The above results show that the sample clustering method designed in this paper is more reasonable than the two control group methods, has better text clustering performance for talent data analysis, and better highlights the characteristics of talent data.

4. Conclusion

This paper proposes a method of talent data analysis based on text mining. The experimental results show that this method can improve the contour coefficient and has good clustering effect. However, this paper only considers the word frequency when constructing the text vector, and does not consider the influence of position on the text. In addition, the similarity and analogy of similar words need to be further studied. In the follow-up research, we can continue to construct text vectors to further optimize the keyword extraction results.

References

[1] LI Xinqin, MA Xiaoning, WANG Zhe, et al.Performance analysis of railway safety supervision personnel based on Text Mining Technology[J].Railway Computer Application, 2019,28(10):30-34.

[2] ZENG Li, CAI Yuxia, ZHANG Jiantao, et al. On the cultivation of information management professionals based on the text mining of employment market demand[J]. Modern computer,2019(21): 59-64.

[3] FU Xiao. Analysis of talent training cycle based on multi-source text similarity [J]. Electronic technique, 2020, 49(08): 114-115.