Automatic Text Classification Method of Personnel Electronic Archives Based on Word Segmentation Algorithm

Jiangjing Lin[1*], Ming Guo[2], Linhua Gong[3], Jiafa Hu[4]

[1*]Corresponding author: linjiangjing@whhwtech.com

[2]guoming@whhwtech.com

[3]gonglinhua@whhwtech.com

[4]290722034@qq.com

[1]Wuhan Second Ship Design Institute Wuhan, Hubei 430205, China

[2]Wuhan Second Ship Design Institute Wuhan, Hubei 430205, China

[3]Wuhan Second Ship Design Institute Wuhan, Hubei 430205, China

[4]Wuhan Second Ship Design Institute Wuhan, Hubei 430205, China

Abstract—To improve the level of personnel management, we need to learn from advanced management theory and advanced science and technology. Thus, establishing an intelligent personnel information management system to assist daily management is helpful to improve the efficiency of personnel management and the quality of service. Aiming at the lack of intelligent analysis and decision-making function in personnel management informatization, this paper proposes a method based on machine learning and deep learning to transform relationship extraction into classification task, and realizes the automatic classification method of personnel electronic archives text by combining entity context information. It can also integrate dependency, part of speech and other multiple features. Corpus data sets are selected in the experiment, and the experimental results show that the proposed method has better performance in convergence speed and model accuracy. It realizes the purpose of intelligent classification of personnel file information. The proposed method can provide promotion and reference for related work of other industries and departments.

Keywords-Personnel files; Electronic information; Intelligent classification; Word segmentation algorithm.

1 INTRODUCTION

After years of development, the traditional personnel management information system has been upgraded and its functions have been gradually improved, which plays an important role in personnel management. Compared with the traditional manual operation, these systems have outstanding advantages in the management of personnel information, such as convenient search,

ICIDC 2022, June 17-19, Qingdao, People's Republic of China

DOI 10.4108/eai.17-6-2022.2322692

rapid retrieval, large storage capacity, high reliability, good confidentiality, low cost, and so on. The author has worked in the personnel management post for many years. He has used different time, and different version personnel management information systems [1]. Besides the database routine maintenance, the inquiry module is deservedly arranged in the first place. The people are getting higher and higher to the system inquiry function expectation. How to make the query module intelligent and create a more friendly query interface has always been a concern of the author [2].

With the development of Natural Language Processing (NLP) and information retrieval technology, NLP plays an important role in information retrieval and information processing [3]. The powerful retrieval function of NLP, if integrated into MIS, will surely produce transformative results. Based on the above considerations, exploring the construction of a Restricted Chinese Natural Language Processing system and using it in the personnel management information system will produce good technical innovation and practical effect [4].

The Restricted Chinese Natural Language Processing System (RCNLP) is put forward in view of the fact that most personnel management workers do not have the professional knowledge of computers. Chinese natural language interface and SQL have very close expression ability, but different from SQL, it is natural for users and easier to master [5]. Based on the "Divide and Conquer" strategy, RCNLP adopts a non-procedural language structure, so it presents the characteristics of multiple statements in query description. This paper explores the system structure and interface design of RCNLP Chinese database system supporting natural language query, not only at the theoretical level of the NLP, but also in the development of the experimental system.

2 CONSTRUCTION OF MULTIPLE DECISION MODEL

In this section, the ensemble learning model is applied to solve the multi-decision problem. The web page resource attribute vector to be classified is horizontally split into multiple sub-vectors, and multiple classifiers are used to classify the sub-vectors. Finally, the classification results are input into the decision function to obtain the final results. Assume that a feature vector to be classified is $(x_1, x_2, \cdots, x_n)$ and the number of classifiers is k, if the vector is divided equally, the sub-vector of the ith group is as shown in Formula 1:

$$C_i = \begin{cases} x\frac{x}{k}(i-1), x\frac{x}{k}(i-2), \cdots, x\frac{x}{k}(i-n), \ i < \left(\frac{n}{k}\right) \\ x\frac{n}{k}(i-1), x\frac{n}{k}(i-2), \cdots, x\frac{n}{k}(i-n), \ i = \left(\frac{n}{k}\right) \end{cases} \tag{1}$$

Then X is equivalent to $(c_1, c_2, \cdots, c_k)$ defining k classifiers $Classfier_i, i \in [l, k]$ [6], respectively extracting sub-vectors G in the training vector to train the classifier Classfieri. Then, the trained classifier group is used to predict the rating of the resource. For the vector X to be classified, k trained classifiers are used to classify its sub-vector $(c_1, c_2, \cdots, c_k)$ to obtain

the result set $R=(r_1,r_2,\cdots,r_k)$, where r_i is the resource quality label [7]. Finally, the f (R) function is used to output the final result. The f (R) function is the key step to realize multiple decisions. The decision function has different definitions of input and output according to the actual needs. For example, in the classification mode, the decision function outputs the result of the predicted label after voting for the model group; while in the classification mode, the decision function outputs the dot product of the attribute weight vector and the predicted result [8].

In the multi-decision model, the quality of resources is determined by multiple items to be evaluated. The multiple classifiers are used to classify different items to be evaluated to obtain a set of classification results, which are called multi-decision vectors. If the dimension of the resource feature vector is m, the number of weak classifiers is m. Assuming that the feature vector X is P, a classifier group D is trained, including iG [1, m] [9]. For each vector X, there is a unique vector $x_1,x_2,\cdots,x_m$ corresponding to it, as shown in Equation 2:

$$R=(r_1,r_2,\cdots,r_m),\quad r_i\in\{0,1,\cdots,k\} \tag{2}$$

The R vector is the result obtained by classifying the resource attributes by multiple classifiers respectively. The k represents the number of final quality classifications, and its value is a natural number from 0 to k. The higher the quality level is, the larger the value is .

It is assumed that each attribute corresponds to a weight, which reflects the impact of the attribute on quality. On this basis, the calculation formula of quality classification is obtained by drawing lessons from the manual review model [10], as shown in Formula 3:

$$\text{score}=\sum_{i=0}^{m} r_i w_i \tag{3}$$

Among them, $score$ is the quality classification, and w_i is the corresponding weight.

3 WEIGHTS ARE ESTABLISHED

The greater the information gain of a resource attribute, the greater the impact of this attribute on its quality, and the greater the corresponding w_i should be. For the resource vector set D to be evaluated, if there are n quality classifications, and the corresponding probability distribution is $P=(P_1,P_2,\cdots,P_n)$, then there is Formula 4:

$$E(D)=-\sum_{n}^{i=1} p_i\log_2(p_i) \tag{4}$$

In Formula 4, $E(x)$ is the total information content of D. For a continuous attribute x_i, the continuous values need to be divided into n discrete values. The commonly used discretization strategy is the dichotomy, assuming that there are m different values of x_i on D, the values are arranged from small to $\left(x_i^1, x_i^2, \cdots, x_i^m\right)$. There is a partition point k to divide D into D- and D +, where D- is the set of values not greater than k, and D + is the set of values greater than k. The log function formula is to return the logarithm of a number according to the specified base.

Using the system information gain as the weight reference can clearly reflect the weight of the feature attributes in the current data set, but it can not be applied to the data set that is too different from the training set. When the data size of the set to be evaluated is much larger than D, the error will be too large, resulting in inaccurate quality classification. However, the resource quality evaluation is applied to the same data set in most scenarios, and the newly added resources are in the minority, so the weight determination method is feasible enough in theory.

4 EXPERIMENTAL ANALYSIS

4.1 Experimental environment

This paper constructs a web crawler based on the python, uses the MySQL as data storage, and randomly crawls the entry attributes and quality tags of Baidu Encyclopedia, including featured entries and ordinary entries. Then, for the historical version of each entry, the attribute data of all editors of each historical version are crawled, including grade, experience, pass rate, etc. Then, the average value of the editor attributes is spliced with the entry attributes, and finally 2008 characteristic entries and 5989 common entries are obtained.

In this paper, Numpy is used as a data processing tool to divide the original data into feature data and result in label data. The feature data consists of 20 entries and editor attribute values, and the labels are divided into 0 (ordinary entries) and 1 (high-quality entries). Due to the large span of feature values, it is necessary to normalize the data, otherwise, the results of the model will be affected by the features with a large distribution. Therefore, the max-min normalization method is used to process the data to improve the convergence speed and accuracy of the model.

4.2 Experimental results and analysis

In this paper, 3200 entries to be tested are scored by the multi-decision model, including 810 featured entries and 2390 ordinary entries. The classification is sorted from high to low, and the classification statistics are carried out according to the labels of the test data.

There is a significant positive correlation between entry attributes and entry quality, such as the length of the entry text, the number of words in the entry summary, the number of reference pages, and the number of pictures. The more words in the entry, the more content it contains, which is considered to be a high-quality entry. Insufficient conditions are ordinary entries. Number of entries in each category before different statistical ranking points as shown in Table 1.

Table 1 Number of entries in each category before different statistical ranking points

Position	300	600	900	1500	2100		2700
High quality	295	542	696		783	805	816
Normal quality	5	58	204		717	1295	1884

Figure 1 and Figure 2 show the classification distribution of featured terms and common terms respectively.

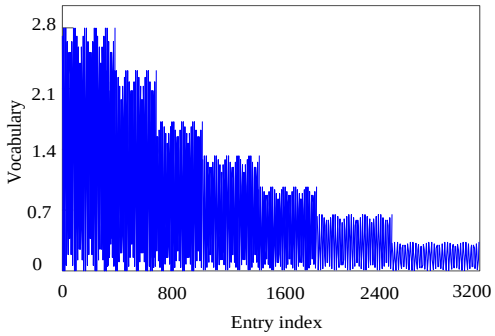

Fig. 1 Classification distribution of quality entries

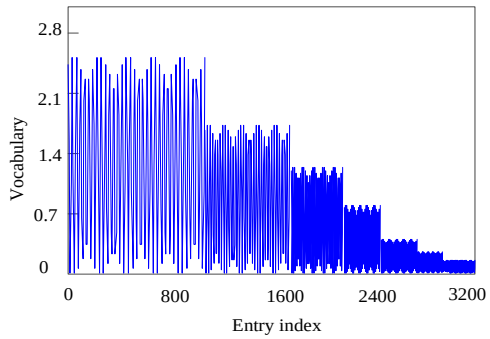

Fig. 2 Classification distribution of common terms

The scores of common terms in the multi-decision model are generally lower than those of featured terms. The data marked as featured terms are mostly distributed above 2 points, while the data marked as common terms are mostly below 2.5 points. It can be seen that the multi-decision model has a certain degree of discrimination in the quality of the entries.

By setting the threshold percentage of regression segmentation, classification regression can be carried out on the classified entries. In this paper, the entries whose classification is higher than the threshold are regressed to the high-quality classification (1 label). The entries whose classification is lower than the threshold are regressed to the low-quality classification (0

labels), and the evaluation index after classification regression is calculated. Table 2 shows the regression results for the multi-decision model under different regression segmentation thresholds, where Score represents the classification calculated by the regression segmentation threshold.

Table 2 Regression results of multi-decision model under different thresholds

Threshold	16%	20%	24%	28%	32%	36%
Score	2.43881	2.12098	1.77567	1.48520	1.35396	1.04994
Precise	0.90215	0.87323	0.81070	0.75754	0.72367	0.63075
Recall	0.58726	0.71082	0.79108	0.86369	0.88407	0.92484
F1-Score	0.71141	0.78370	0.80077	0.80714	0.79587	0.75

It can be seen from Table 2 that whether the information gain is used as the attribute weight has an impact on the classification effect. The classification mode corresponding to the former achieves a higher TPRate under the same FPRate, and the AUC value reaches 0.947. Under the same conditions, the ROC curvature of the classification model without weight is lower, and the calculated AUC value is 0.913, which is also lower than the former. It is not difficult to conclude that information gain will add to the attributes that have a greater impact on the quality of entries. In this case, the influence of attributes with low relevance on the classification of entries will be weakened, and the authenticity and credibility of the classification will be significantly improved.

5 CONCLUSION

In order to solve the problem of automatic classification of personnel electronic files, this paper uses intelligent analysis methods to study the personnel management system.

(1) Use a global feature extraction module to extract semantic features, and use the multiple convolutions to extract local features of terms.

(2) Classify the resources to be archived by using a plurality of machine learning classifiers.

(3) Experiments show that the model has high accuracy in the extraction of personnel electronic file segmentation.

Through the research of this paper, the intelligent personnel information management model can be realized, which has a high level of office automation for information retrieval and classification and promotes the information development of personnel management.

REFERENCES

[1] Kalgin A. Implementation of performance management in regional government in Russia: Evidence of data manipulation. Public Management Review, 2016, 18(1): 110-138.

[2] Vetterlein, Antje. Responsibility is more than accountability: from regulatory towards negotiated governance. Contemporary Politics, 2018, 24(5): 545-567.

[3] Shi B, Wang X, Lyu P, et al. Robust scene text recognition with automatic rectification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 4168-4176.
[4] Shi B, Bai X, Yao C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE transactions on pattern analysis and machine intelligence, 2016, 39(11): 2298-2304.
[5] Borisyuk F, Gordo A, Sivakumar V. Rosetta: Large scale system for text detection and recognition in images. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2018, pp. 71-79.
[6] Yu D, Li X, Zhang C, et al. Towards accurate scene text recognition with semantic reasoning networks[C]. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 12113-12122.
[7] Sun Y, Ni Z, Chng C K, et al. ICDAR 2019 competition on large-scale street view text with partial labeling-RRC-LSVT. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2019: 1557-1562.
[8] Chng, C.K., et al.: ICDAR 2019 robust reading challenge on arbitrary-shaped text-RRC-ArT. In: 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 1571–1576. IEEE (2019).
[9] Shi B, Yao C, Liao M, et al. ICDAR2017 competition on reading chinese text in the wild (RCTW-17) [C]. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2017, 1: 1429-1434.
[10] Pappas N, Popescu-Belis A. Multilingual Hierarchical Attention Networks for Document Classification[C]. In: Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2017: 1015-1025.