Advances in Intelligent Systems Research, volume 134
2nd International Conference on Control, Automation, and Artificial Intelligence (CAAI 2017)

Question Classification Based on Improved TFIDF Algorithm

Jing Gao[1], Cuixiao Zhang[1*], Zhiqiang Wang[23], Guangzhen Zhao[1] and Xuan Li[1]
[1]School of Infomatics Science and Technology, Shijiazhuang Tiedao University, Shijiazhuang, 050043, China
[2]Institute of Applied Mathematics, Hebei Academy of Sciences, Shijiazhuang 050081, China
[3]Hebei Authentication Technology Engineering Research Center, Shijiazhuang 050081, China

***Abstract*—The feature weight calculation is an important part of question classification, feature weights directly affect the accuracy of the classification results, the traditional TFIDF algorithm is widely applied in the field, but it ignores the relationship between features and classes. Therefore, this paper introduces information entropy and mutual information to improve the traditional TFIDF, I convert the computed weight results into a space vector model format, and use the support vector machine (SVM) for question classification. Finally, I use TFIDF, TFIDF-MI and TFIDF-MI-E these three methods to calculate the feature weight, and compare the results of the classification experiments. The experimental results show that the accuracy, recall and F value of TFIDF-MI-E algorithm are higher than those of TFIDF and TFIDF-MI two algorithm.**

Keywords-TFIDF; mutual information; information entropy; space vector model; questioning classification

I. INTRODUCTION

The automatic question answering system is composed of three parts: preconditioning, information retrieval and answer extraction[1-2]. The question analysis is the basis of the latter two parts. After a series of semantic analysis and questioning classification, the information retrieval is narrowed and the answer is extracted more accurately. The question classification is the main part of the question analysis, which effectively reduces the number of candidate answers[1]

There are two general methods of classification: one is based on the rules of the method, the second is based on statistical methods, that is, machine learning methods, and now more is based on statistical methods. For the rule-based method is to manually analyze the structure of the question, and extract the characteristics from it, then according to the rules to determine the type of question, subjective factors have great influence, flexibility is poor[3-4]. For the statistic-based approach, it is through statistical learning to extract accurate features from the question, and then use the classifier to classification of questions through learning[4], this classification method is more accurate.

For the question classification section, the basic part includes Chinese word segmentation, feature extraction, weight calculation and question classification. The calculation of the characteristic weight affects the effect of the following classification. The choice of classifier is also an important factor in the classification of questions. Although the traditional TFIDF algorithm is applied to the calculation of feature weight, it has many shortcomings. In[5], it is mentioned that TFIDF ignores the relationship between the feature word and the category, And adjust the IDF by combining the information gain and the dispersion, suppress the high dependence of the IDF value on the document frequency, The author uses a feature fusion to improve the traditional TFIDF to achieve the extraction of keywords. In the literature[6], the mutual information is applied to the feature weighting method, and combined with some word frequency information and document frequency to achieve good classification results. In[8], the information gain is added to the TFIDF algorithm to obtain a certain classification effect. On this basis, the information entropy is added, and the accuracy of the classification is further improved. Therefore, this paper uses an improved TFIDF algorithm for questioning classification experiments, the classification effect has improved significantly. It is proved that this improved method can optimize the traditional TFIDF algorithm.

II. METHOD OF FEATURE WEIGHT CALCULATION

Before question classification, by quantifying the characteristic words to indicate the importance of his classification of questions, At present, the main method of calculating the weight of the characteristic word is Boolean weight method, square root weight method, logarithmic weight method, entropy-based weighting method and TFIDF (Term Frequency and Inverse Documentation Frequency) weight method [10].In this paper, the following three weight calculation methods are briefly introduced.

A. TFIDF

TFIDF is a classical feature weight function in the vector space model. TFIDF calculate the weight of a feature word as formula (1)

$$w(d) = tf * idf = f_d * log\frac{N}{n_d} \quad (1)$$

Where,

a) $w(d)$ *-- the weight of feature word* d*;*

b) f_d *-- the frequency which feature* d *appears in a document;*

c) N *-- the total number of files;*

d) n_d *-- the number of documents that feature* d *appears.*

Where *tf* is the entry frequency, which refers to the number of words d appears in the text. *IDF* is the document frequency,

which refers to the number of documents that words *d* appear in the entire document set, the fewer frequent words appear in the text, the more likely they are to express the content of the document; In addition, the more the word appear in the document set, the worse the division of the word and the smaller contribution to the classification. Therefore, in the traditional TFIDF feature selection method, the weight of the entry is proportional to the frequency of the entry, and inversely proportional to the document frequency.

TFIDF is a more traditional method of calculating the weight of the feature, the application is also more extensive. However, the traditional TFIDF algorithm is a problem, it can only reflect when a word appears more documents, the discriminative power is worse, and it does not recognize whether the document containing the word belongs to the same category, that is, the traditional TFIDF algorithm ignores the relationship between the feature word and the category.

B. Mutual Information

Mutual information in the sentence classification is the relationship between the characteristics and categories. We get the mutual information between t_i and C_j as formula (2)

$$I(t_i, C_j) = log \frac{P(t_i, C_j)}{P(t_i)P(C_j)} \quad (2)$$

Where,

a) i *-- the number of features;;*

b) j *-- the number of categories;*

c) $I(t_i, C_j)$ *-- the mutual information between* t_i *and* C_j*;*

d) $P(t_i)$ *-- the probability of all sentences with the characteristic* t_i*;*

e) $P(t_i, C_j)$ *-- the probability that the sentence contains the feature* t_i*and belongs to class* C_j.

For multiple questionnaires, you can calculate the mutual information of the characteristics of each category , and then seek the average, We can get multi-category mutual information formula (3)

$$I_{ave}(t_i) = \sum_{j=1}^{m} P(C_j)\ I(t_i, C_j) \quad (3)$$

Mutual information is mainly applied to feature extraction, but it is also applied to the feature weight calculation, through the analysis of its formula that it fully expresses the relationship between the characteristic word and the category. But there is also a drawback that the frequency of occurrence of the feature word in the document is not taken into account.

C. Information Entropy

Information entropy is the source of the uncertainty described[11]. Assuming the events *X (x1, x2 ... xn)*, they occur at the probability of *P (p1, p2 ... pn)*, then the corresponding information entropy formula is:

$$H(x) = -\sum_{i=1}^{n} P_i log(P_i) \quad (4)$$

Where,

a) $H(x)$ *-- The entropy of event X;*

b) P_i *-- The probability of occurrence of the i th event;*

c) n *-- The number of events.*

D. TFIDF-MI-E

For the shortcomings of the traditional TFIDF algorithm, the mutual information and TFIDF are combined to form the TFIDF-MI algorithm to improve the traditional TFIDF, but considering the shortcomings of mutual information, the concept of information entropy is introduced.

With the concept of information entropy, the relationship between the characteristic word and the document and the relationship between the characteristic word and the category are shown. The relationship is represented by the formula (5)

$$E = -\sum_{i=1}^{m} \frac{tf(t,d_i)}{tf(t,C_j)} log \frac{tf(t,d_i)}{tf(t,C_j)} \quad (5)$$

Where,

a) E *-- The entropy of the ratio of the frequency of the feature word appears in the document and the frequency of the feature word appears in the category;*

b) $tf(t, d_i)$ *-- the word frequency of the characteristic word t in the document;*

c) $tf(t, C_j)$ *-- The number of events.*

Integrated information entropy and mutual information to improve the traditional TFIDF, the formation of improved TFIDF-MI-E algorithm, The new weight calculation formula is (6)

$$W_{ij} = w(t_i) * Iave(t_i, C_j) * E(t_i) \quad (6)$$

Where,

a) W_{ij} *-- the weight of the feature word* t_i *in text* C_j*;*

b) $w(t_i)$ *-- the traditional tfidf calculates the weight of the feature word* t_i *in the document;*

c) $Iave(t_i, C_j)$*-- The mutual information between the feature* t_i *and the category* C_j*;*

d) $E_{C_j}(t_i)$ *-- The relationship between the word frequency in the class and the word frequency in the document.*

The formula optimizes the traditional tfidf, which not only links the feature words to the category, but also does not forget to express the relationship between the characteristic word and the document, so that the characteristic words, documents and categories are cleverly combined.

III. QUESTIONING

A. Questioning System

Classification system is a simple question into the characters, places, numbers, time, entities, description 6 categories.

B. Space Vector Model

The vector space model, also known as VSM (Vector Space Model), is a classic text representation model proposed by Gerard Salton [12]. Without considering the order, the space vector model convert all the feature words into the space vector, so that the same word in the same column appears to facilitate the comparison of the question. This can improve the classification effect. It is similar to the Boolean model, but instead of 0,1 replaced by their own specific weight value [13].

C. Support Vector Machines

Support Vector Machine (SVM) is the first proposed by Corinna Cortes and Vapnik in 1995. It presents many unique advantages in solving small sample, nonlinear and high dimensional pattern recognition, and can be applied to other machine learning problems such as function fitting. The basic principle of SVM is to map the vector to high dimension space and establish the maximum interval hyperplane. We analyze the classification result by calculating the distance of parallel hyperplane. The bigger the distance, the better the classification effect[14]. By comparing the time efficiency and classification accuracy of several common classification algorithms in[1], the support vector machine achieves higher classification effect without taking up more time. In[9], SVM is combined with distance-based pattern recognition to achieve a good classification effect. The method has achieved good classification effect. Therefore, this article uses SVM for classification.

IV. EXPERIMENTS

A. Experimental Data

The experimental data of this paper is to use some of the questions in the language technology platform of HIT Information Retrieval Research Center, which is divided into six categories: character, location, number, time, entity and description. The data set is shown in Table 1.

TABLE I. CATEGORIES AND DATASETS

category	training set	test set
characters	120	30
places	120	30
numbers	120	30
time	120	30
entities	120	30
description	120	30
characters	720	180

B. Experimental Steps

The experimental steps in this paper are as follows:

1) Segmentation word with the word segmentation tool ikanalyzer, remove the stop word after the word segmentation.

2) The word segmentation tool itself has the function of disambiguation, can directly eliminate the ambiguous word output, as the question is relatively short, no longer deal with the word after the disambiguation, directly to all words as a characteristic word.

3) This paper calculate the corresponding weight of each word though the improved feature weight calculation formula , according to VSM into a spatial vector model.

4) The data in the space vector model are converted into convert the data into a format that the classifier SVM can recognize by the format converter FormatDatalibsvm.

5) Each category with 120 questions to do training, 30 questions to do the test, the questioner is classified by the classifier SVM. The question is classified by SVM classifier.

6) Evaluate the classification results using recall rate, precision rate and F1 value respectively:

a) the recall rate (R): (The number of relevant questions retrieved / search system in the total number of related questions) * *100%.*

b) the precision rate (P): (The number of relevant questions retrieved / the total number of questions retrieved) * *100%.*

c) F1=2PR/((P+R)): It is used to comprehensively evaluate recall and precision.

C. Comparison of Experimental Results

In this paper, we use the traditional TFIDF algorithm, the algorithm of TFIDF and mutual information, and the algorithm of TFIDF and mutual information based on information entropy, and then calculate the corresponding weight of the feature word. The weight of the obtained feature word is transformed into a space vector model, which is classified by the classifier SVM. The classification results of the three methods are analyzed by comparing recall, precision and F1.

TABLE II. THE RECALL RATE OF THREE METHODS (%)

category	TFIDF	TFIDF-MI	TFIDF-MI-E
characters	13.3	63.3	96.7
places	70.0	73.3	76.7
numbers	90.0	96.7	96.7
time	50.0	80.0	90.0
entities	90.0	80.0	90.0
description	100.0	83.3	96.7
average value	68.9	79.4	91.1

As can be seen from Table 2, the TFIDF-MI algorithm has a significant improvement in character and time categories, increasing by 50% and 30% respectively, and a slight decrease in the entity and description category, but the average is 10.5% higher than that of the traditional TFIDF. The TFIDF-MI-E algorithm has a slight decrease in the description category, and the average value is significantly higher than that of the former two, which is 22.2% higher than that of the TFIDF algorithm and 11.7% higher than the TFIDF-MI algorithm.

TABLE III. THE ACCURACY OF THE THREE METHODS (%)

category	TFIDF	TFIDF-MI	TFIDF-MI-E
characters	100.0	79.2	82.9
places	91.3	84.6	92.0
numbers	96.4	85.3	96.7
time	100.0	80.0	100.0
entities	93.1	64.9	84.4
description	37	86.2	93.5
average value	86.3	80.0	91.6

It can be seen from Table 3 that TFIDF-MI-E has been reduced in the category of characters and entities, and has improved in other categories. The final average is 5.3% higher than that of TFIDF algorithm and 11.6% higher than TFIDF-MI algorithm. It can also be seen that TFIDF-MI-E not

only improves the recall rate, but also improves the accuracy rate.

TABLE IV. F1 VALUES OF THE THREE METHODS (%)

category	TFIDF	TFIDF-MI	TFIDF-MI-E
characters	23.5	70.4	89.3
places	79.2	78.5	83.7
numbers	93.1	90.6	96.7
time	66.7	80.0	94.7
entities	91.5	71.7	87.1
description	54.0	84.7	95.1
average value	68.0	79.3	91.1

Finally, through the comparison of the geometric mean of the two, as shown in Table 4. although the TFIDF-MI algorithm is lower than the TFIDF algorithm in the accuracy rate, the accuracy rate is improved when the precision and recall rate are taken into account. The average value is 11.3% higher than the TFIDF algorithm. TFIDF-MI-E algorithm is significantly improved, 23.1% higher than TFIDF, 11.8% higher than TFIDF-MI. From this we can see that the improved algorithm TFIDF-MI-E in this paper is very helpful to the question classification, which improves the effect of question classification, and paves the way for the future work of the intelligent question answering system.

V. CONCLUSION AND PROSPECT

In this paper, the traditional TFIDF algorithm is improved by introducing mutual information and information entropy, and the relationship between the characteristic word and the document and the relationship between the characteristic word and the category are expressed by entropy. Through the comparison test to analyze and evaluate the experimental results, it is found that the improved algorithm greatly improves the accuracy of classification, the accuracy rate of 91.1%. At the same time, it is found that accuracy rate of TFIDF is higher than that of TFIDF-MI, and then the mutual information can be improved and then analyzed by experiment. After the VSM conversion of the weight calculation results, the number of 0 appears particularly large, resulting in a particularly large dimension, which may also affect the classification results. The main work of the next step is to study a dimension reduction algorithm, and then combine the improved algorithm to test the question classification.

ACKNOWLEDGMENT

This paper is supported by graduate base practice funded projects，and supported Hebei Academy of Sciences Program for Cooperation with Chinese Academy of Sciences （171401）.The author would like to thank the teachers and the students for their help, and also would like to express appreciation to the anonymous reviewers for their helpful comments on improving the paper.

REFERENCES

[1] Zhen Lihua, Wang Xiaolin, Yang Sichun.A Summary of the Study on the Classification of Questions in Automatic Question Answering System[J]. Journal of Anhui University of Technology (Natural Science Edition),2015, 32(1), 48-54

[2] Zheng Shifu, Liu Ting, Qin Bing, Li Sheng.Automatic question and answer review[J]. Chinese Journal of Information,2002,16(6), 46-52

[3] Zhang Ning, Zhu Lijun. A Review of Chinese Q & A System Questions[J]. Information engineering,2016,2(1), 32-42

[4] Yu Zhengtao, Fan Xiaozhong, Guo Jianyi. Classification of Chinese Questions Based on Support Vector Machine[J]. Journal of South China University of Technology (Natural Science Edition).2005,33(9), 25-29

[5] Yang Kaiyan. Research on Automatic Extraction Algorithm Based on Improved TFIDF Keyword [D]. Xiangtan University,2015

[6] Fan Xiaochao, Zhang Chongyang, Dengxiongwei. Text Feature Weighting Method Based on Mutual Information[J]. Computer Engineering and Applications,2015, 51(13), 145-148

[7] Wanxiang Che, Zhenghua Li, Ting Liu. LTP: A Chinese Language Technology Platform. In Proceedings of the Coling 2010: Demonstrations. 2010.08, pp13-16, Beijing, China.

[8] Li Xueming, Li Hairui, Xue Liang, He Guangjun. TFIDF Algorithm Based on Information Gain and Information Entropy[J]. computer engineering,2012,38(8), 37-40

[9] An xu, zhang shudong. Research on Fuzzy Feature Classification Algorithm Based on Support Vector Machine[J]. computer engineering,2017,43(1), 237-240

[10] Aas K, Eikvil L. Text Categorization: A Survey[R]. Oslo, Norway: Norwegian Computing Center, Tech. Rep.: NR941, 1999.

[11] YI Linlin, Zhang Renjun. Information Entropy Theory and Behavioral Finance[J]. Financial science,2004, 6(207), 29-33

[12] SALTON G. Automatic processing of foreign language documents[C]. Proceedings of the 1969 Conference on Computational Linguistics. Association for Computational Linguistics, 1969: 1-28

[13] Heyuan. Research on hot topic discovery based on microblogging[D]. Shijiazhuang Tiedao University, 2016

[14] Chen Jiaxi. Text Classification Based on Support Vector Machine[J]. Electronic world,2007, 64