ISSN: 2664-9640

Quantitative Analysis of Repeated Question Retrieval Based on TF-IDF Model

Cheng Yang, Zhixuan Chan, Jiaming Zhu

School of Statistics and Applied Mathematics, Anhui University of Finance and Economics, Bengbu 233041, China

Abstract

After the knowledge community has accumulated a large dataset of frequently asked questions, finding answers to similar questions becomes the key to retrieval. Based on the above problems, this paper establishes a classification model of question sentence similarity to recall and re-rank relevant answers. First, construct a dataset of similar problem pairs, and perform data cleaning by removing stop words, case conversion, etc. Then, the TF-IDF model is used to convert the unstructured text problem into a feature vector, and the feature of chi-square statistic is used for feature dimension reduction. To avoid the problem of poor generalization caused by imbalanced datasets, samples are balanced using under sampling and oversampling techniques. Finally, a logistic classification model is established to identify similar questions, so as to help users return similar questions with the highest correlation.

Keywords

Duplicate question retrieval; TF-IDF; chi-square statistic; Logistic; oversampling; under sampling.

1. Introduction

With the in-depth development of the Internet, various community forums have developed rapidly. Nowadays, social forums in all walks of life play a role in Q&A, puzzle solving and other functions, providing an effective way for relevant professionals to learn and improve. After the community has accumulated a large set of frequently asked questions, how to quickly find and return is a key step. Duplicate question retrieval is to retrieve similar questions from the data set accumulated in the history of the forum and return them to the user. In general, the search process for similar issues is divided into two phases: recall and reorder. The first stage is to carry out multiple similar statements through the traditional retrieval method, and finally to identify multiple binary problems based on the returned problem data, and sort according to the similarity[6]. Considering the computational efficiency and model complexity, this paper adopts the TF-IDF method based on the document corpus probability to reduce the retrieval time of the problem and obtain the vector expression of the problem statement. The downstream task uses Logistic to identify duplicate problems, and the overall processing flow of the problems in this paper is shown in Figure 1.

2. Repeat the Construction of The Problem Detection Model

2.1. Research Ideas

Similar question recognition, also known as question sentence repetition recognition, is designed to determine whether two natural questions given are semantically similar. The similar question recognition task is a fundamental and challenging task in the field of natural language processing [4]. This is because there are different kinds of expressions in natural questions, for the same thing, different people may not express it differently but the semantics

ISSN: 2664-9640

are consistent, or it may be that different people express it very close but the semantics are far apart. Therefore, this makes it necessary to recognize not only the differences in the expression of the sentence, but also the differences in the deep semantic information such as the subject, intention, and goal of the question [8].

In view of the above problems, this paper defines the retrieval task of the duplicate problem, assuming that the two forum problem statements M and N, $M = (m_1 \cdots m_j \cdots m_p)$, $N = (n_1 \cdots n_i \cdots n_q)$, q and represent the statement length of the problem M and N respectively, m_j and n_i represent the jth word and the i word in the problem M and N, respectively, the task is to test whether the two problems are the same problem type [5]. Taking into account the timeliness requirements of forum searches and the particularity of the corpus. This paper uses the traditional TF-IDF model to express statements characteristically.

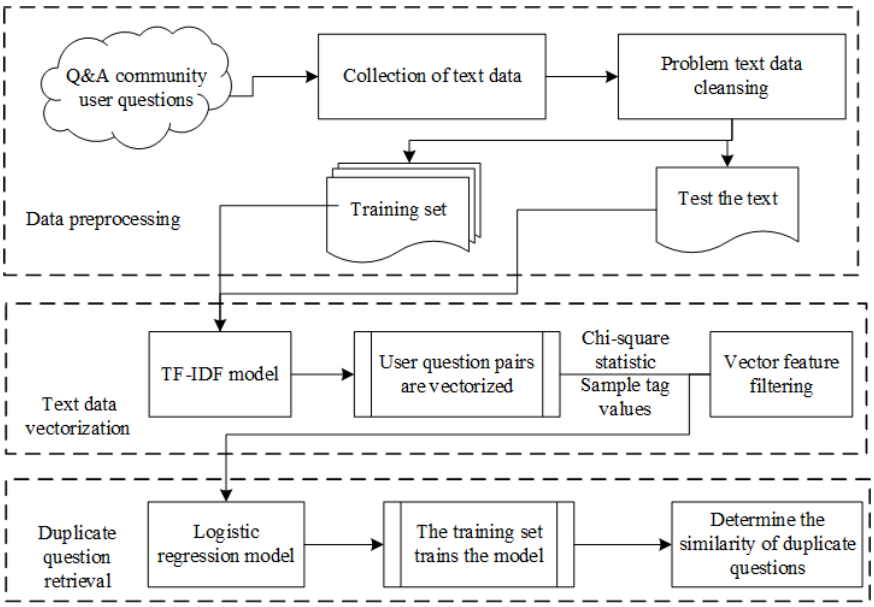

Figure 1. Overall Problem Processing Flowchart

2.2. Tf-Idf Text Vectorization

TF-IDF (Term Frequency-Inverse Document Frequency) is a word-weighting technique commonly used for information retrieval and data mining. TF-IDF is a method of word frequency statistics that evaluates the importance of a word relative to a text corpus. The importance of a word increases with the frequency with which it appears in the corpus.

TF-IDF can be divided into TF and IDF, TF indicates the frequency of words appearing in the corresponding corpus, and also carries out corresponding normalization processing. Its formula is expressed as follows:

$$TF_{i,j} = \frac{w_{i,j}}{\sum_k m_{k,j}} \tag{1}$$

where $w_{i,j}$ represents the frequency with which the word t_i appears in the corpus c_j, and $\sum_k w_{k,j}$ represents the frequency of all words in the corpus c_j. The IDF indicates the prevalence of certain keywords. If the corpus c_j fewer documents containing the word i, the larger the IDF is, because it has the ability to distinguish itself well from other types. The specific calculation method is as follows.

$$IDF_i = log\frac{|N|}{|\{j:t_i \in d_j\}|} \tag{2}$$

Of these, |N| represents the number of documents contained in the corresponding corpus, $|\{j: t_i \in d_j|$ represents the corresponding number of documents that contain the word t_i. The

ISSN: 2664-9640

TF and IDF values are calculated separately, and the above two are multiplied. That is, the TF-IDF value of the word is obtained, and the higher the TF-IDF value calculated by the word corresponding to it, it indicates that the word has a stronger importance than other texts, and it is very likely to become the keyword of the corpus.

$$TF - IDF_{i,j} = TF_{i,j} \times IDF_i \tag{3}$$

Because community Q&A sentences have longer question lengths than common texts, they tend to have words specific to them in longer sentences[1]. Therefore, the keyword extraction algorithm based on TF-IDF is used to characterize long question sentences.

2.3. An Expression of Similarity in Text

Define the two questions obtained M, N features expressed as A, B. To measure the similarity of text statements [8], the absolute value of the difference between the two is used to measure the similarity of text S. The more similar the meaning of the problem is expressed, the smaller the gap between the feature vectors of the two. The bigger the opposite.

$$S = |A - B| \tag{4}$$

2.4. Text Feature Reduction

The TF-IDF model is taken to obtain a question vector dimension that is too high and too sparse. In order to speed up the calculation and improve the accuracy of question sentence detection, it is necessary to reduce the characteristic dimension of the obtained question vector. In this paper, chi-square test is used to filter out the attribute columns of the sample label value and the text similarity S height correlation, calculate the deviation between the observed value and the theoretical value by the chi-square statistic, and use the significance level α to filter the highly correlated attributes.

2.5. Logistic Regression Classification Model

Through the above operations, the characteristic expression of the forum repeat question pair is realized. The next step is to construct a binary classification model of problem pairs, which is easy to solve and less resource-intensive logistic regression [9]. Let the label value be 1, that is, the probability of the occurrence of the repeated problem pair is $P(Y = 1|x) = p(x)$, and the probability of the problem pair with the label value of 0 is $P(Y = 0|x) = 1 - p(x)$.

$$L(w) = \prod [p(x_i)]^{y_i} [1 - p(x_i)]^{1-y_i} \tag{5}$$

The probability of all sample occurrences is multiplied, the likelihood function $L(w)$ is constructed, and the logarithm is taken to solve it in the way of maximizing the likelihood function to obtain the loss function. For the loss function, the parameter is biased, and the parameter value is calculated using a random gradient descent method [10].

$$J(w) = -\frac{1}{n}(\sum_{i=1}^{n}(y_i ln\, p(x_i) + (1 - y_i) ln(1 - p(x_i))) \tag{6}$$

For the above loss function, the deviation derivative of the parameters is obtained, and the direction of the parameter descent is obtained, and the parameter value is evaluated by using the method of random gradient descent. where k is the number of iterations. Each time the parameters are updated in the direction of g_i gradient descent and the α step length, you can stop iterations by comparing $\| J(w^{k+1}) - J(w^k) \|$ less than the threshold or reaching the maximum number of iterations.

$$g_i = \frac{\partial J(w)}{\partial w_i} = (p(x_i) - y_i)x_i \tag{7}$$

$$w_i^{k+1} = w_i^k - \alpha g_i \tag{8}$$

Based on the above method, we can train a model about repeating problem pairs, and perhaps use test sets to test the effect of model training. In the end, we get a text classification model based on TF-IDF, which can use the above method to retrieve the duplicate questions in the

ISSN: 2664-9640

forum, speed up the user to obtain answers to similar questions, and solve the problems encountered.

3. Preprocessing of Model Data

3.1. The Process of Processing

Faced with the obtained forum text data, where non-semantic text or other non-text-type symbols exist. To reduce the interference of these characters with the classification effect, basic pre-processing of text data is required. The specific process of processing is shown in Figure 2.

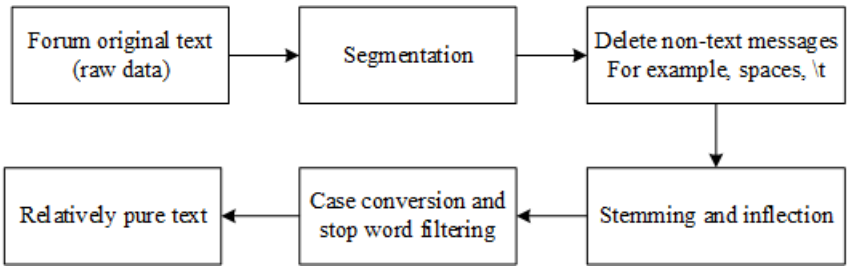

Figure 2. Basic flowchart of text preprocessing

3.2. Lowercase Conversion and Stopword Filtering

In order to prevent the statement from being retrieved, the case of the same word is treated as a different word. Therefore, the problem statement needs to be capitalized. And because there are usually a large number of words in the text that have no actual semantics, such as "of" "did" "a" "an", they do not have a disturbing effect on semantics [2].

In order to reduce the dimension size of the text vector, the english deactivated vocabulary is used for removal. This improves the efficiency of subsequent learning and calculations [3]. The results of the pretreatment are shown in Table 1.

Table 1. Comparison table before and after text preprocessing

Question ID	The question text
97939	I have an ArcGis Desktop Standard license and I would like to perform an .....
5041	Is it possible to create a text field in a database table with a length that i……
75289	I'm searching for the way to use the new QGis non blocking notification……
Question ID	The problem text after text preprocessing
97939	arcgi desktop standard licens would like perform oper find differ 2 .....
50415	possibl creat text field databas tabl length greater 254 charact use arcgi 10 1……
75289	search way use new qgi non block notif system qgi python wrapper found ……

3.3. Balance of Samples

Based on the duplicate problem IDs in the data, after the paired problem groups are constructed, the overall dataset has 13105520 bars, as shown in Table 2. However, at the same time, it faces extremely unbalanced samples, and the sample size difference between sample labels 0 and 1 is thousands of times different. Since the non-repeat problem is too many pairs, when performing a binary classification problem, the loss function is greatly affected by the label 0, which will give us the illusion that the model is excellent on the test set.

ISSN: 2664-9640

But in the real business, the main target is to detect duplicate problems. As a result, it may result in poor generalization and failure to successfully detect duplicate problems. Therefore, when building model training data, you need to balance the sample data. A hierarchical sampling is used to balance the dataset by sampling a sample with a label of 0. Then, oversampled and undersampled are used to solve the situation that the number of different data labels is too different.

4. Characterize Repeat Problem Pairs

Using TfidfVectorizer, the above obtained data is converted into vector form [7]. After that, the absolute value and the label value of the difference of the vector are selected according to the problem. In order to further reduce the problem that the expression of question sentence features is too high dimensional, calculate the degree of correlation between column features and label values, and filter the key feature columns in the dataset based on 20% correlation.

After the above operation, the characteristics of the sample dataset are reduced to 1238 dimensions. Reducing the difficulty of the operation speeds up the calculation and facilitates the rapid retrieval of duplicate problem pairs. Finally, the table is merged with the above balanced sample ID, and finally the eigenvector representation of each problem pair is obtained, and some of the results are as shown in Table 2.

Table 2. Table of vectors for problem representation of features

number	...	yield	york	zero	zip	zipcod	...	zoomcontrol
0	...	0.0	0.1	0.0	0.0	0.0	...	0.0
1	...	0.0	0.0	0.0	0.0	0.0	...	0.0
2	...	0.0	0.0	0.5	0.0	0.6	...	0.0
...	...	...	...	...	...	...	...	...
130938	...	0.0	0.4	0.0	0.0	0.0	...	0.0
130939	...	0.0	0.0	0.0	0.0	0.0	...	0.0

The equilibrium vectorization problem obtained above is paired and the dataset is divided into training sets and test sets at a fixed scale. This document uses a logistic regression model to reorder recalls of duplicate problem pairs. The model parameters are obtained, and the test set is used for the evaluation and analysis of the model. Finally, the common evaluation indicators of the classification model are used to evaluate the efficiency of the model. The process of model construction and calculation of duplicate problem detection model is shown in Figure 3.

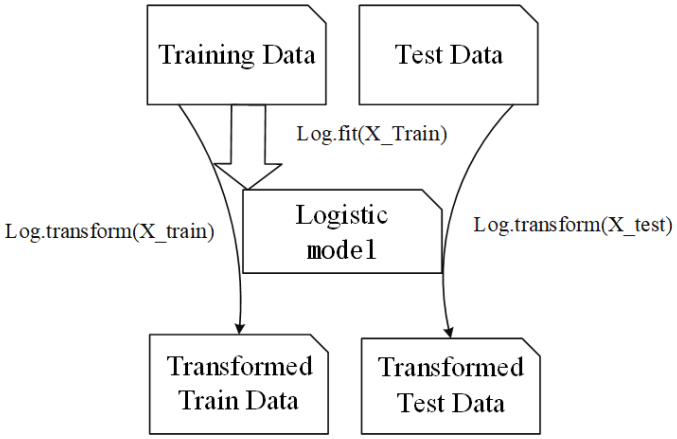

Figure 3 Logistic Regression model calculation flowchart

After the model is trained completely, it is necessary to evaluate the classification model for its good or bad degree. Because this paper translates the retrieval of duplicate questions into the

ISSN: 2664-9640

traditional binary classification model. Evaluation metrics commonly used in classification questions are therefore used: Accuracy, Precision, Recall, and F1-score[1]. Based on the above indicators of the evaluation classification problem, this paper calculates the scores of various evaluation indexes for the two classification problems of the duplicate problem. The results are shown in Table 3.

Table 3. Model evaluation score table

Evaluation index	accuracy	precise	recall	F1（sample）	F1（test）
score	98.8%	99.5%	98.1%	98.9%	98.8%

From Table 3, it can be seen that the repeated problem retrieval based on TF-IDF in this paper has extremely high accuracy, and the accuracy is more than 98% in both the training set and the sample set. On other evaluation indicators of the classifier, the accuracy rate was 98.8%, the accuracy rate was 99.5%, and the recall rate was 98.1%, respectively. Overall, almost all duplicate problems were successfully detected when classifying text. On the ROC curve, as shown in Figure 4, close to the (0,1) point, it is possible to recall duplicate problem groups very much at different thresholds. It is very robust. It has met the expected requirements and obtained better results.

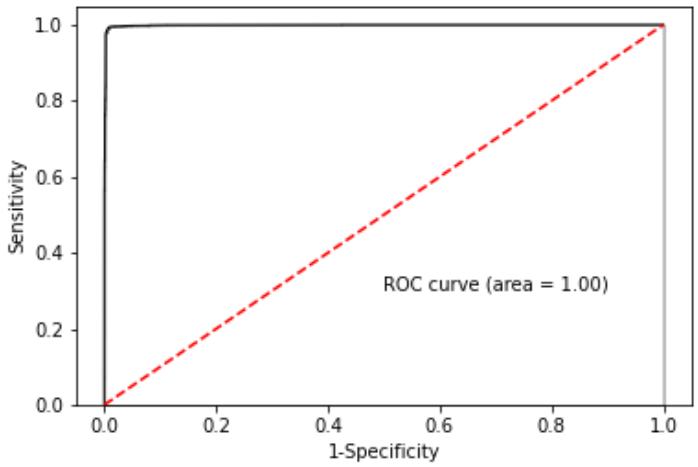

Figure 4 ROC curve

5. Conclusion

This paper focuses on the search for duplicate problems, which mainly converts the search for the similarity of two user problems into a binary classification problem, and measures the similarity of the two problems by using the absolute value of the difference between the two problem vectors. The vectorization of text is the key to the problem, and this article uses morpheme-based information to transform, ignoring the semantic information inherent in the statement itself.

Now with the development of deep learning and the maturity of the natural language of the transform system. Great strides have been made in the characteristic expression of textual sentences, such as bert models based on large-scale pre-training, which can express semantic information very well. But the time it takes is too expensive. Recently, SBERT, which is based on the BERT model, has made significant progress in time consumption, and at the same time, it is not inferior to the original BRET model in terms of accuracy. It has a huge advantage in the face of repeated questions in the forum. There are great prospects for development in the industrial world.

ISSN: 2664-9640

Acknowledgments

The related research in this paper was funded by the annual undergraduate scientific research innovation fund project of Anhui University of Finance and Economics (XSKY22236)

References

[1] TANG Xiao-bo;LIU Jiang-nan. Automatic Indexing of Questions in Q&A Community Based on BERT and TF-IDF--Taking the CNGOLD Q&A Community as an Example[J]. Information Science, 2021,39(03):3-10.

[2] LI Ke-yue;CHEN Yi;NIU Shao-zhang. Social E-commerce Text Classification Algorithm Based on [J]. Computer Science,2021,48(2):87-92.

[3] Duan Dandan. Research and Application of Feature Reduction Methods in Text Classification [D]. Nanjing: Nanjing University of Posts and Telecommunications,2020.

[4] ZHANG Yan-kun;CHEN Yu-zhong;LIU Zhang-hui;. Hybrid Neural Network Model for Community Question/Answer Matching [J]. Journal of Chinese Computer Systems,2020,41(9):1833-1838.

[5] Chen Xin. Research on Question Deep Semantic Matching for Community Question Answering[D]. Suzhou University,2020.

[6] Xu Zhuojia. Forum Duplicate Question Detection Based on the Word-to-sentence Interaction Mechanism and Multi-task Learning [D]. South China University of Technology,2020.

[7] SHAO Ming-rui;MA Deng-hao;CHEN Yue-guo;QIN Xiong-pai;DU Xiao-yong. Transfer learning based QA model of FAQ using CQA data [J]. Journal of East China Normal University(Natural Science),2019(05):74-84.

[8] Yan Manl. Similar Question and Answer Summary in Community Question Answering [D]. eijing University of Posts and Telecommunications,2018.

[9] DING Hao;ZHU Jia-ming. Independent Prediction of Gratification Delay of College Students based on Regression Analysis [J]. Journal of Huaiyin Teachers College(Natural Science Edition), 2022, 21(01):37-41.

[10] XU Xiang;JIANG Juan;ZHU Jia-ming. Measurement Analysis of Influencing Factors of Allergic Diseases Based on Logistic Regression [J]. Journal of Hebei North University(Natural Science Edition),2021,37(5):55-62.