Fuzzy Systems and Data Mining IX
A.J. Tallón-Ballesteros and R. Beltrán-Barba (Eds.)

doi:10.3233/FAIA231020

An Efficient Matching Algorithm for Question Answering System

Jing ZHANG [a], XinYue ZHAO [a], Jianing HUANG [a], and Yunsheng SONG [a,1]
[a] *School of Information Science and Engineering, Shandong Agricultural University, Taian 271018, China*

Abstract. Wheat and corn are the two most important grain crops in northern China, and at this stage, it is difficult to obtain professional information in the field of wheat and corn, the acquisition efficiency is low, and the accuracy is poor, which seriously restricts the production efficiency. The intelligent question answering system can efficiently and accurately automatically screen out professional information, to effectively solve the above problems. However, the existing intelligent question answering system for wheat and maize has too low matching accuracy and slow retrieval speed to be widely promoted. Therefore, an efficient and accurate two-stage matching algorithm is designed, which uses the BM25 algorithm to recall the candidate set, and then uses the BERT model to screen out the optimal solution based on the candidate set. Experimental results show that the algorithm has high retrieval accuracy and retrieval efficiency.

Keywords. Question Answering System, Pre-trained models, Information Retrieval, Agriculture

1. Introduction

In the past, the acquisition of expertise in the agricultural field mainly relied on expert consultation, but with the spread of the Internet, farmers are more likely to use search engines to find the information they need from the Internet. However, the results returned by search engines are often very redundant and complex, containing a lot of inaccurate information, and users need to sift through this information themselves, which takes a lot of time and effort. And many farmers don't know enough about the Internet, making it difficult for them to use the Internet to get the knowledge they need. The emergence of an intelligent question answering system can effectively solve this problem, users only need to enter the questions they want to query through voice or text [1], intelligent question answering system can efficiently screen out the most accurate answer for users, simplifying the process of users to obtain knowledge [2]. The existing intelligent question answering system for wheat and corn has problems such as slow retrieval speed and low retrieval efficiency, and this paper constructs an efficient and accurate two-stage matching algorithm to help users solve problems in the planting process [3].

TF-IDF[4], BM25, LDA[5] and other traditional information retrieval algorithms to match text similarity, and the matching speed is fast. However, the semantic similarity of

[1]Corresponding Author: Yunsheng Song ; E-mail: songys@sdau.edu.cn

the text cannot be fully captured, and the matching accuracy rate is low. BERT[6] model can capture the semantic similarity of text, matching accuracy is high, but the matching speed of these models is very slow, can not be well applied in practical problems, so this paper designs a two-step retrieval algorithm, first with a faster traditional information retrieval method to screen out a candidate set containing the target question and answer, and then on this candidate set with slow but high accuracy deep matching method for accurate matching, to ensure the speed and accuracy of retrieval at the same time.

2. Main Content

To meet the needs of users for retrieval speed and retrieval accuracy, this paper designs a two-stage retrieval-style question answering algorithm based on BM25 algorithm and BERT model, which is described in detail in this section.

2.1. The overall architecture of the searchable question answering algorithm

Figure 1 shows the overall architecture of the search-style algorithm. The retrieval algorithm is generally divided into two stages, the first stage is recalled, using the BM25 algorithm to quickly recall several questions most similar to the user query problem from the database as a candidate set, and the second stage is rearrangement, training the BERT model to judge the semantic similarity of the two texts, scoring the questions in the candidate set, and returning the question with the highest score in the candidate set and its reply. The algorithm combines the advantages of the fast retrieval speed of BM25 algorithm and the high matching accuracy of BERT model, and can efficiently and accurately retrieve the target question and answer.

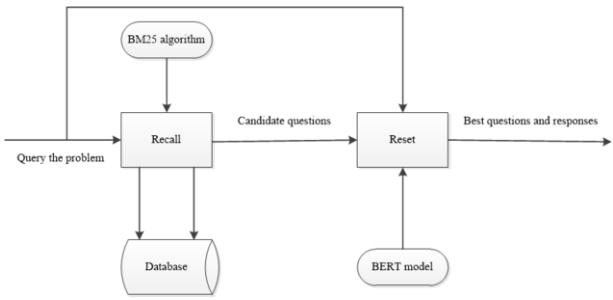


Figure 1. The overall architecture of the search-based algorithm

2.2. Recall module based on BM25 algorithm

The purpose of the recall module is to quickly screen out a candidate set that contains the target problem. The algorithm for selecting the recall module needs to consider two factors at the same time, the accuracy of retrieval and the speed of retrieval, and this paper uses the BM25 algorithm for recall .

2.2.1. Inverted indexing technique

Before using the BM25 algorithm to recall, this paper first uses the inverted index technique to improve the speed and accuracy of the recall. An inverted index is a vocabulary-document index that can quickly retrieve all documents containing that term based on a term. First, build an inverted table of all issues in the database according to the steps in algorithm 1. Break the user query question into words, and for each word in the query question, you can get a list of questions containing the word through the inverted table. Put the problem list corresponding to all the words in the user query problem into a collection, and then use this collection to recall, which can effectively improve the speed and accuracy of the recall stage.

Algorithm 1 Inverted index building process

Input: A list of all issues in the Q database.

Output: inverted table, dictionary form, the key is a word, value is a list of documents containing the word .

1: OInitialize the inverted table as a dictionary table;
2: Iterate through the list Q, taking out each issue q and its index id:
3: Use the jieba participle to turn the question q into a list of words q_L;
4: Traverse the q_L and take out each word w:

$$table[w] = table.get(w, id) + [id]$$

5: **return** Inverted table.

2.2.2. Inverted indexing technique

BM25 algorithm is a method to calculate the similarity between a query and a document, the specific steps of using the BM25 algorithm to calculate the similarity score between the query problem and the problem in the database are shown in algorithm 2, and the obtained similarity score is sorted from largest to smallest, and the top k questions with the highest score are returned, and the recall process is completed. For convenience, this article uses the BM25 module packaged in the Gensim library to complete this process.

2.3. BERT model design and implementation

At the heart of the rearrangement module is the matching model, which should accurately determine how similar two sentences are. The BERT model is the most powerful text-matching model at present, and this paper fine-tunes the trained BERT model on the constructed problem-problem similarity matching dataset, predicts whether two sentences are similar, and uses the trained model to score the recalled candidate question set to return the highest scoring questions and replies.

2.3.1. BERT matches the overall architecture of the model

In this paper, a linear layer is superimposed on the BERT model[7], which maps the BERT-encoded information to the output probability distribution, and uses the softmax function to normalize the output probability to a value between 0 and 1. The dimension

Algorithm 2 The BM25 algorithm calculates similarity

Input: Document set D consisting of all questions in the database, with each problem represented by d_i, The user queries Q.

Output: A list of scores containing the similarity scores of Q to all questions d_i.

1: Calculate the inverse document frequency idf for each word in D according to Equations , saved in the dictionary idf, $idf[w]$ represents the idf value of the word w;
2: Calculate the frequency of occurrence of each word in each question d_i, saved in list f,$f[i][w]$ indicates the number of occurrences of the word w in the ith question;
3: Calculate the average length $avgdl$ for all problems in D;
4: Use jieba to query the user for q word segmentation, each word is represented by q_i, and $TF(q_i)$ and $IDF(q_i)$ are obtained according to steps 1, 2;
5: Calculate the similarity score of Q to each question d_i according to Equations and save it to the list scores;
6: **return** Returns a list of similarity score scores.

of the output probability distribution is 2, which represents the probability of similarity and the probability of dissimilarity between two texts. The training phase uses this distribution to predict the similarity labels of the two texts, and the inference stage uses this distribution to score the similarity of the two texts and complete the rearrangement.

2.3.2. Pre-training and fine-tuning of BERT matching models

This paper uses the Bert-Chinese model officially released by Google as a pre-training model, retrains on the dataset built by itself, fine-tunes the parameters of the model, and predicts whether the two texts are similar, with a similar label of 1 and a non-similar label of 0, which can be regarded as a text binary classification problem. Each piece of data in the training set is converted into the format required by the fine-tuning task [8], fed into the BERT model, the output probability distribution is obtained, and the loss value of the predicted label and the real label is calculated using the cross-entropy function. Use the AdamW optimizer to optimize the loss value, and use the warm-up strategy to adjust the learning rate. After each training iteration is completed, the validation set is used to evaluate the model effect, and the accuracy rate is used as the evaluation index of the validation set, the accuracy is defined as the probability that the prediction result is correct, and the model with the highest accuracy is saved.

2.3.3. Rearrange the process

Use the trained BERT model to rearrange the questions in the recalled candidate set. Firstly, the user query problem is spliced with each question in the candidate set into the required input format, the text is converted into a text embedding vector according to the vocabulary, and the output probability distribution is obtained by feeding the trained BERT model, that is, the similar probability and the dissimilar probability of the two texts. The candidate questions are reordered according to the size of the similarity probability, and the candidate questions with the highest probability of similarity to the user query question and their corresponding replies are returned.

3. Experimental Result

To better illustrate the effect of each stage, this paper tests the recall stage, rearrangement stage, and the entire algorithm separately.

3.1. Experiment setup

This paper needs to use two datasets: the question-answer dataset and the question-question dataset[9], which use crawler technology to capture the data of wheat and corn on agricultural forums such as the Agricultural Science and Technology Online Book House and the China Agricultural Technology Extension Information Platform.

This paper uses the parameter settings of BERT-BASE, the constructed BERT model has a total of 12 layers, the number of heads of multi-head self-attention is also 12, the dimension of the coding vector is 768, and the parameters that the model can train have a total of 110,000. The model uses words as input units, and the size of the vocabulary is 21128. The maximum length of the model input vector is 512, and vectors smaller than this length are filled and vectors larger than this length are truncated. The rearrangement stage uses the trained BERT model to rearrange the candidate set. Using Sia-GRU, ESIM [10] models as comparison models. The hidden size of both models is set to 300, and the training data, training process, and training parameters are consistent with the settings of the BERT model.

3.2. Experimental testing during the recall phase

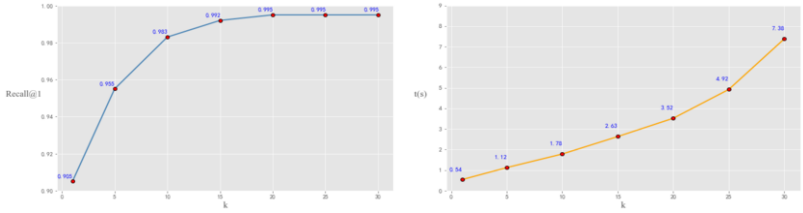
Use the TOP-K recall accuracy Recall@k as an evaluation metric. For each query question, the first k questions are retrieved using the above algorithm, and if the target question is included in the k, the query question is successfully recalled, Recall@k represents the ratio of the number of successfully recalled query questions to the total number of questions. Table 1shows the results of the experimental tests.

Table 1. Recall algorithm test results

	Recall@1	Recall@5	Recall@10	Recall@20
TF-IDF algorithm	88.30%	94.00%	97.50%	98.8%
LDA algorithm	68.30%	72.80%	76.50%	81.80%
BM25 algorithm	90.50%	95.50%	98.30%	99.50%

From the experimental results, the BM25 algorithm has a higher accuracy rate than the other two algorithms for the dataset constructed in this paper, so the BM25 algorithm is used to recall the set of candidate questions. At the same time, it can be seen that when the value of k is small, the recall accuracy is poor, and as the value of k increases, the recall accuracy continues to rise, but the increase of k slows down the speed of the subsequent matching stage.

The speed of recall of the BM25 algorithm is fast, and through the recall step, a candidate set containing the target problem can be quickly obtained, but the target problem cannot be accurately obtained, so the candidate problem needs to be further rearranged using the deep matching model.



(a) The retrieval accuracy varies with k (b) The retrieval time varies with k

Figure 2. The performance of the proposed algorithm with k

3.3. Rearrange phase experimental testing

Table 2 shows the results of the text-matching experiment test, it can be seen that the performance of the BERT model in the text matching task is much better than that of the commonly used deep text-matching models such as Sia-GRU and ESIM, so this paper uses the BERT model to rearrange the candidate set problem. The matching speed of the BERT model is very slow, and it is not realistic to directly use the BERT model to match all the problems in the database, and the recall step must be performed first to narrow the matching scope.

Table 2. Rearrange model test results

	Accuracy	Recall	F1
SiamGRU	82.60%	82.40%	82.40%
ESIM	83.80%	84.20%	83.90%
BERT	89.50%	90.40%	89.90%

3.4. Overall experimental test of the algorithm

The BM25 algorithm was used to recall the first k similar problems as the candidate set, and then the BERT model was used to select the most similar problems from the candidate set. Figures 2 show the relationship between the retrieval accuracy Recall@1 and the average retrieval time t of the algorithm and the number of candidate set questions k, respectively, Recall@1 indicates that the retrieved top-1 problem is equal to the probability of the target problem. It mainly verifies the retrieval accuracy and retrieval efficiency of the entire algorithm, verifies the impact of the number of recall questions k on the retrieval time and retrieval accuracy, and selects the most suitable k.

Comparing the results of Figure 2(a) with the results of Table 1, it is found that the rearrangement module based on the BERT model can accurately retrieve the target problem from the recalled candidate set, which verifies the necessity of the rearrangement stage and the effectiveness of the entire algorithm. At the same time, it can be found that the selection of k value will affect the retrieval accuracy and retrieval efficiency of the algorithm, the k value is too small, the retrieval accuracy is low, the k value is too large, and the retrieval speed is slow. This paper prioritizes the accuracy of the algorithm, and then considers the retrieval speed of the algorithm under the premise that the accuracy

rate is as high as possible. Therefore, this paper finally sets the number of recall questions k to 20, which has high search accuracy and relatively fast retrieval speed, which is within the acceptable range of users.

4. Conclusions

In this paper, a two-stage retrieval algorithm is designed, which uses BM25 algorithm to recall candidate problem sets and uses BERT model to rearrange candidate question sets, which combines the advantages of fast matching speed and high matching accuracy of BM25 algorithm. Experimental results show that the algorithm designed in this paper can obtain high accuracy while ensuring accuracy.

In addition, based on the retrieval algorithm designed in this paper, a web-based intelligent question answering system is developed by using Flask back-end framework and front-end technology. More features can be introduced, such as the introduction of login and registration modules to save user information, and it can also introduce users with various permissions, such as experts and administrators, to improve the scalability and reliability of the system.

Acknowledgement

This research was supported by Shandong Provincial Natural Science Foundation, China, grant number ZR2020MF146, and Open Project Foundation of Intelligent Information Processing Key Laboratory of Shanxi Province, grant number CICIP2021002.

References

[1] Fehri H, Dardour S, Haddar K. ARmed question answering system. CONCURRENCY AND COMPUTATION-PRACTICE & EXPERIENCE. 2022;34(21):1-13.
[2] Shum HY, He X, Li D. From Eliza to XiaoIce: challenges and opportunities with social chatbots. Frontiers of Information Technology & Electronic Engineering. 2018;19:10-26.
[3] Aithal SG, Rao AB, Singh S. Automatic question-answer pairs generation and question similarity mechanism in question answering system. Applied Intelligence. 2021;51(11):8484-97.
[4] Qaiser S, Ali R. Text mining: use of TF-IDF to examine the relevance of words to documents. International Journal of Computer Applications. 2018;181(1):25-9.
[5] Jelodar H, Wang Y, Yuan C, Feng X, Jiang X, Li Y, et al. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia Tools and Applications. 2019;78:15169-211.
[6] Elsadig M, Ibrahim AO, Basheer S, Alohali MA, Alshunaifi S, Alqahtani H, et al. Intelligent Deep Machine Learning Cyber Phishing URL Detection Based on BERT Features Extraction. Electronics. 2022;11(22):3647.
[7] Geetha M, Renuka DK. Improving the performance of aspect based sentiment analysis using fine-tuned Bert Base Uncased model. International Journal of Intelligent Networks. 2021;2:64-9.
[8] Zhang J, Chang WC, Yu HF, Dhillon I. Fast multi-resolution transformer fine-tuning for extreme multi-label text classification. Advances in Neural Information Processing Systems. 2021;34:7267-80.
[9] Zou Y, He Y, Liu Y. Research and implementation of intelligent question answering system based on knowledge Graph of traditional Chinese medicine. In: 2020 39th Chinese Control Conference (CCC). IEEE; 2020. p. 4266-72.
[10] Rebecq H, Gehrig D, Scaramuzza D. ESIM: an open event camera simulator. In: Conference on robot learning. PMLR; 2018. p. 969-82.