Combining Q&A Pair Quality and Question Relevance Features on Community-based Question Retrieval

Dong Li  
School of Computer Science and Technology  
Wuhan University of Technology  
Wuhan, China  
cathylilin@whut.edu.cn

Lin Li  
School of Computer Science and Technology  
Wuhan University of Technology  
Wuhan, China  
cathylilin@whut.edu.cn

Dong Zhou  
School of Computer Science and Engineering  
Hunan University of Science and Technology  
Hunan, China  
dongzhou@hnust.edu.cn

Abstract—The Q&A community has become an important way for people to access knowledge and information from the Internet. However, existing translation based models do not consider term weights when assigning weights to query terms in question retrieval. We improve the term weighting model based on the traditional topic translation model and further considering the quality characteristics of question and answer pairs, this paper proposes a community-based question retrieval method that combines question and answer on quality and question relevance (T^2LM+). We have also proposed a question retrieval method based on convolutional neural networks. The results show that compared with the relatively advanced methods, the two methods proposed in this paper increase MAP by 4.91% and 6.31%.

Keywords—Question retrieval, Translation model, Topic model, Learn to rank, Convolutional neural network

I. INTRODUCTION

The main research direction of Community-based Question retrieval is question relevance, which is how to measure the correlation between query and historical question and then optimize the search results. Relevance ranking is the core issue in question retrieval. It specifically refers to sort questions and answers according to their relevance with user queries. Text relevance models in the traditional information retrieval field include vector space model (VSM) [1], BM25 model [2], and language model [3].

The question retrieval model based on the statistical machine topic translation model [4-7] is a popular model for question retrieval systems currently. Clustering [8-11] is another popular method. This kind of model uses a linear combination method to integrate the language model and has achieved good results. In addition to traditional machine learning methods, deep learning based retrieval methods [12-13] have good results recent years. Cross-modal retrieval [14-15] is to find the relationship between different modal samples, and to use some modal samples to search for other modal samples of approximate semantics.

Learning to rank is a sorting method based on supervised learning, which idea is to regard the relative order relationship between documents as training data and automatically obtain the sorting model through machine learning. LambdaMART is an algorithm in learning to rank that works for many sorting scenarios and works well, companies like Bing and Facebook are using this model [16]. The current topic translation model has the same weight for the topic in the process of summing the relevance of all topics. There is no way to respond differently to lexical relevance for different query questions. For example, for a given user query "How much is a pack of white soft Chungwa cigarettes?", The top 5 search results returned by the current advanced topic translation model (T^2LM) [7] are shown in Fig 1. From the table we can see that although the number of words ranked in the first candidate question is more than the others, the core word "Chungwa" in the query does not matched. In addition to question relevance, another recognized indicator of the impact of ranking results is the quality of questions and answers., we use the term weight and subject translation as the question relevance feature, and integrate the quality of question and answer pairs, and propose an optimized community-type question retrieval method. The method takes five features as input and learns training using the existing advanced model LambdaMART. The results show that compared with the relatively advanced method, the proposed method has a 4.91% increase in MAP@10. We propose a convolutional neural network method based on attention mechanism (TextCNN-Features) and the method has a 6.31% increase.

![Fig. 1. A search example of T^2LM](image)

II. RELATED WORK

In response to the shortcomings of the word-based exact matching question relevance model, the researchers introduced the statistical machine translation model into the field of information retrieval, and used the translation model to model the correlation between different words. Jeon [17] proposed a language model based on translation, the model combines the strengths of both the language model and the translation model, and introduces the translation probabilities into the language model, completing the matching between different words. Xue [18] improved the model proposed by...
Jeon. He integrated the maximum likelihood estimation of the question corresponding to the answer into the matching process with the query words, enriched the matching text, and proved to further the effect of question retrieval. Lee [19] used the more important feature words in the Q&A pair to enhance the translation model to optimize the translation probabilities between vocabularies. Bernhard and Gurevych [20] combined multiple high-quality monolingual parallel corpora as training data, which further improved the original retrieval model. Gao [21] integrated the interdependence of words in the question into the language model, and proposed a new language model based on word dependence. Cao [22] establishes a more complex language model that adds the classification information of the problem to the probability of generation of the keyword in the document to smooth the probability of generation of the language model on the classification.

III. COMBINING Q&A PAIR QUALITY AND QUESTION RELEVANCE FEATURES ON COMMUNITY-BASED QUESTION RETRIEVAL

A. Framework of the model

This paper uses the current state-of-the-art learning to rank model to train the quality and question relevance model features through fusion question and answer. The framework is shown in Fig 2. Learning to rank is a supervised machine learning method that can easily fuse multiple features with fewer artificial parameters. From the current research methods, there are three strategies for learning sorting, namely, pointwise, pairwise and listwise. The pointwise method converts the sorting problem into a multi-class classification or regression problem. The pairwise method converts the sorting problem into a document -to-classification problem. The listwise method learns the correct ordering of the entire candidate set under the query. The model used in this paper is LambdaMART. The advantage of LambdaMART is that it can convert unrecognizable information retrieval evaluation indicators, such as NDCG, into functions that can be derived.

B. Feature calculation

1) Relevance feature

Zhang's proposed topic translation model T2LM [7] uses topic information to control translation noise in the translation model.

\[
P_{c2lm}(w|q,a) = \mu_1 \cdot P_{ml}(w|q) + \mu_2 \cdot \sum_{t \in q} P_{ml}(w|t)P_{ml}(t|q)
\]

\[
+ \mu_3 \cdot \sum_{i=1}^{K} \left( \frac{K}{\sum_{i=1}^{K} P_{ml}(w|z_i)P_{ml}(t|z_i)} P_{ml}(t|q) \right)
\]

\[
+ \mu_4 \cdot P_{ml}(w|a)
\]

\[
P_c(w|z_i) \text{ is the distribution probability of the word } w \text{ under the topic } z_i. K \text{ is the number of topics in the topic model. } \mu_1, \mu_2, \mu_3, \text{ and } \mu_4 \text{ are the parameters used to control the impact of each part and } \mu_1 + \mu_2 + \mu_3 + \mu_4 = 1.
\]

As can be seen from the formula, they actually use \(P_{ml}(w|z_i) \cdot P_{ml}(t|z_i)\) to calculate the relevance of the word \(w\) and the word \(t\) under the topic \(z_i\). Then obtain the total relevance of the word \(w\) and the word \(t\) under the topic model by summing the results under all the topics. The problem is that the weight of each topic is same in the process of summarizing the relevance results under all the topics, which makes it possible to determine the relevance of the vocabulary once the topic model is trained to complete the topic model changed.

Therefore, we have improved the model of Zhang. It first uses the topic model to analyze the topic distribution of the user query, then uses the topic information of the query to determine the specific semantics of the word in the retrieval process, and then guides the correlation calculation between the words in the topic model to get more accurate question relevance.

T2LM+ first obtains the weight \(w_{query,w}\) by querying the word \(w\) in the query through the term weighting model, and then integrates the weight \(w_{query,w}\) into the correlation calculation. Specifically, for the word \(w\) in the query. We split the correlation of the four parts of T2LM+ into four correlation features. They are the \(F_1(query,q)\) query-question language model, \(F_2(query,q)\) query-question translation model, \(F_3(query,q)\) query-question topic model, and \(F_4(query,a)\) query-answer language model. Specifically, the four correlation features are calculated according to formulas (4) to (7) respectively. Here for the smoothing parameter \(\lambda\), we set \(\lambda\) to \(1/(|q|+1)\) or \(1/(|a|+1)\).

2) Q&A on quality feature

User information in the Q&A community is very useful and we try to incorporate this information into the model. The model is as follows:
The listwise method learns the correct regression problem. The pairwise method converts the problem to the probability of generation of the language model, and proposed a new language model between vocabularies. Bernhard and Gurevych \cite{1} establishes a more effective method for determining the relevance of the vocabulary once the translation model to optimize the translation probabilities is shown in Fig 3.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. Experimental data

We used the data set from NDBC CUP 2016 as experimental data. There are 578608 questions and 1,729,263 answers in the data set. Zhang uses manual markup to obtain experimental data, so we adopt the same method. We first get 1500 questions from the data set as a standby query by random sampling method, and use Lucene as a search tool to obtain the top 20 related problem pairs for each query as its candidate questions. We select the available queries based on the content and length of the query and the top20 candidate questions that are related or similar to the query. Finally, 990 queries are randomly divided into training data sets and test data sets in a 1:1 ratio. The training set is mainly used to adjust the model parameters, and the test set is mainly used to test the effect of the model. In the experiment, we set the number of trees in LambdaMART to 50, the number of leaves to 4, the learning rate to be 0.2, and the minimum number of instances of a single leaf to be 30. Finally, we let the caller manually label the relevance of the query to its corresponding candidate question. Specifically, two labelers are invited first, each of which puts one of the following three labels for each candidate question according to the labeling specification: "good (1)" and "poor (0)". If the two callers cannot agree on a candidate, the third caller is invited to make the final decision (he can only choose one). In order to reduce the workload of manual markup, we only mark the top10 candidate questions for the running results of each model. The other unmarked default tags are 0.

B. Experimental comparison system

In order to verify the effect of our proposed learning ranking model, we also use the classic model of question retrieval and the current state-of-the-art model as a comparison system. As follows:

1) Vector space model (VSM) \cite{1}.
2) BM25 model \cite{2}.
3) Unary language model (LM) \cite{4}.
4) Translation-based language model (TLM) \cite{5}.
5) Intent-based language model (IBLM) \cite{6}.
6) Language model based on topic translation model (TFLM) \cite{7}.

Alg. 1. Q&A based on user information for quality assessment algorithm

Alg 1 scores its authority based on the user's best answer number first, and then based on the assumption that the quality of the user's published information is positively related to its authority, the question and answer of the authoritative evaluation of the questioner and the respondent is used as a question and answer pair quality feature. Alg 2 uses the learning to rank to combine the question relevance feature $F_1$ and the question and answer quality feature $F_5$ to form the community-based question retrieval method for the fusion question and answer on the quality and question relevance features.

Alg 2. Community-based question retrieval method based on fusion question and answer on quality and question relevance

\begin{equation}
S_u = \frac{\text{Mot} (K_{u,20})}{20}
\end{equation}

$A_u$ is the best answer for user $u$. We set an upper limit for the user's authoritative score to eliminate the possible adverse effects of outliers.

\begin{equation}
P_{\text{listwise}}((w, \text{query})|(q, a)) = \mu W_{\text{query}, w} P_m(w | q)
+ \mu_2 \sum_{t \notin q} \sum_{t \neq t} \left( P_z(q | \text{query}) P_{\text{query}}(w | z) P_m(t | z) \right) P_m(t | q)
+ \mu_4 W_{\text{query}, w} P_m(w | a)
\end{equation}

\begin{equation}
W_{\text{query}, w} = \frac{- \sum_{t \notin q} \sum_{t \neq t} \left( P_z(q | \text{query}) P(w | z) \right) \ln P(w | z)}{- \sum_{t \notin q} \sum_{t \neq t} \left( P_z(q | \text{query}) P(t | z) \right) \ln P(t | z)}
\end{equation}

\begin{equation}
F_1(q, a) = \prod_{w \notin q} \left( 1 - \lambda \right) W_{\text{query}, w} P_m(w | q) + \lambda P_m(w | C)
\end{equation}

\begin{equation}
F_2(q, a) = \prod_{w \notin q} \left( 1 - \lambda \right) \sum_{t \notin q} \sum_{t \neq t} \left( P_z(q | \text{query}) \ln P(w | z) \right) P(t | z) + \lambda P_m(w | C)
\end{equation}

\begin{equation}
F_3(q, a) = \prod_{w \notin q} \left( 1 - \lambda \right) \sum_{t \notin q} \sum_{t \neq t} \left( P_z(q | \text{query}) \ln P(w | z) \right) P(t | z) + \lambda P_m(w | C)
\end{equation}

Then, based on the relationship between the question-questioner and the answer-responder, the user's authoritative scoring model is transformed into a question-and-answer quality feature:

\begin{equation}
F_S((\text{query}, a)) \leftarrow (S_{\text{qu}}, S_{\text{au}})
\end{equation}

The user $qu$ is the proposer of the question $q$, and the user $au$ is the respondent of the answer $a$.

C. Algorithm Description

In summary, we propose a user-based Q&A quality assessment algorithm as shown in Alg 1., and a community-based question retrieval for the quality and question relevance features as shown in Alg 2. method.

Alg 1. Q&A based on user information for quality assessment algorithm
C. Experimental results and analysis

As shown in Table I, we first made a comparison of the TextCNN model (TextCNN+) with the existing advanced model, and based on TextCNN+ we use TextCNN+ as four features (TextCNN+5) to conduct a fusion of Q&A on community-based question retrieval methods for quality and question relevance features. Finally, we compare the neural network model (TextCNN+ Features) with them. From Table II, we can see that the improved model (TextCNN+) is more effective than the topic-based translation query model (TextCNN), indicating that the improved model has a positive impact on question retrieval. At the same time, we can also see that the retrieval model based on learning sorting is stronger than TextCNN+, which is mainly because the learning sorting LambdaMAR adopts the tree model, so it can learn the different combinations of features. We can also find that TextCNN has improved performance compared to traditional machine learning methods.

V. CONCLUSION

In view of the shortcomings of the traditional term weighting model, we propose a new topic-based model of term weighting model. It scores the importance of the term according to the amount of information, thereby highlighting the role of important words in the search and further optimizing the results of the question search. The quality of questions and answers is considered to be an important indicator of the ranking results in addition to relevance. Just like PageRank is not based on web content, the quality of questions and answers can usually not be judged by question and answer text. In related research, researchers make full use of the diversity of data in the Q&A community, such as the number of support and anti-objectives, to assess the quality of questions and answers. In this paper, we propose a community-based question retrieval method that combines question and answer on quality and question relevance. We also proposed a deep learning question retrieval method. The experimental results validate the effects of our proposed model.

TABLE I. COMPARISON OF T\textsuperscript{2}LM+ WITH EXISTING ADVANCED MODELS

| MODEL | MAP       |
|-------|-----------|
| VSM   | 0.3475    |
| BM25  | 0.3506    |
| LM    | 0.3583    |
| TLM   | 0.3746    |
| IBLM  | 0.3916    |
| T\textsuperscript{2}LM | 0.4361 |
| T\textsuperscript{2}LM+ | 0.4695 |

TABLE II. EXPERIMENTAL RESULTS OF COMMUNITY-BASED QUESTION RETRIEVAL METHODS THAT COMBINE QUESTION AND ANSWER ON QUALITY AND QUESTION RELEVANCE

| MODEL               | MAP       |
|---------------------|-----------|
| T\textsuperscript{2}LM+ | 0.4992    |
| TextCNN+5           | 0.4882    |

As shown in Table I, we first made a comparison of the T\textsuperscript{2}LM model (T\textsuperscript{2}LM+) with the existing advanced model, and based on T\textsuperscript{2}LM+ we use T\textsuperscript{2}LM+ as four features (T\textsuperscript{2}LM+5) to conduct a fusion of Q&A on community-based question retrieval methods for quality and question relevance features. Finally, we compare the neural network model (TextCNN + Features) with them. From Table II, we can see that the improved model (TextCNN+) is more effective than the topic-based translation query model (TextCNN), indicating that the improved model has a positive impact on question retrieval. At the same time, we can also see that the retrieval model based on learning sorting is stronger than TextCNN+, which is mainly because the learning sorting LambdaMAR adopts the tree model, so it can learn the different combinations of features. We can also find that TextCNN has improved performance compared to traditional machine learning methods.

REFERENCES

[1] Salton G, Wong A, Yang C S. A vector s-space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11):613-620.
[2] Robertson S E, Walker S Okapi\Karim\Keenbowa-TREC-8[J]. TREC, 1999, 8: 151-162.
[3] Song F, Croft W B. A general language model for information retrieval [C]//Proceedings of the eighth international conference on Information and knowledge management ACM,1999: 316-321.
[4] Ponte J M, Croft W B. A language modeling approach to information retrieval[C]//Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1998:275-281.
[5] Xue X, Jeon J, Croft W B. Retrieval models for question and answer archives[C]//Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2008:475-482.
[6] Haocheng W. Research on Ranking Method in Community Question and Answer Search [D]. Anhui: University of Science and Technology of China, 2017.
[7] Weinan Z, Yu Z, Ting L. A topic translation model for community-based question retrieval[J]. Journal of Computer,2015, 38(2):313-321.
[8] Y. Wang, X. Lin, L. Wu, et al, Robust subspace clustering for multi-view data by exploiting correlation consensus. IEEE Transactions on Image Processing, 24(11):3939-3949, 2015.
[9] Y. Wang, L. Wu, X. Lin, J. Gao. Multiview Spectral Clustering via Structured Low-Rank Matrix Factorization. IEEE Transactions on Neural Networks and Learning Systems 29 (10), 4833-4843, 2018.
[10] Y. Wang, W. Zhang, L. Wu et al., Iterative Views Agreement: An Iterative Low-Rank Based Structured Optimization Method to Multi-View Spectral Clustering. UCAL 2016.
[11] L. Wu, Y. Wang. Beyond Low-Rank Representations: Orthogonal Clustering Basis Reconstruction with Optimized Graph Structure for Multi-view Spectral Clustering. Neural Networks, 103:1-8, 2018.
[12] Y. Wang, X. Lin, L. Wu, W. Zhang. Effective Multi-Query Expansion via Collaborative Deep Networks for Robust Landmark Retrieval. IEEE Transactions on Image Processing 26 (3), 1393-1404.
[13] L. Wu, Y. Wang, X. Li, J. Gao. Deep Attention-based Spatially Recurrent Networks for Fine-Grained Visual Recognition. IEEE Transactions on Cybernetics 49 (5), 1791-1802, 2019.
[14] L. Wu, Y. Wang, L. Shao. Cycle-Consistent Deep Generative Hashing for Cross-Modal Retrieval. IEEE Transactions on Image Processing 28 (4), 1602-1612, 2019.
[15] Y. Wang, X. Lin, L. Wu et al., LBMCH: Learning Bridging Mapping for Cross-modal Hashing. ACM SIGIR 2015.
[16] huagong_ada. Learning To Rank. https://blog.csdn.net/huagong_ada/article/details/40710305.
[17] Ponte J M, Croft W B. A language modeling approach to information retrieval[C]//Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1998:275-281.
[18] Xue X, Jeon J, Croft W B. Retrieval models for question and answer archives[C]//Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2008:475-482.
[19] Lee J T, Kim S B, Song Y. I et al. Bridging lexical gaps between queries and questions on large online Q&A collections with compact translation models[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2008:410-418.
[20] Bernhard D, Gurevych I. Combining lexical semantic resources with question & answer archives for translation-based answer finding[C]//Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2. Association for Computational Linguistics,2009:728-736.
[21] Gao J, Nie J Y, Wu G, et al. Dependence language model for information retrieval[C]//Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval.20 04:170-177.
[22] Cao X, Cong G, Cui B, et al. Approaches to exploring category information for question retrieval in community question-answer archives[J]. A CM Transactions on Information Systems,2012,30(2):