2021 2nd International Conference on Electronics, Communications and Information Technology (CECIT) | 978-1-6654-3757-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/CECIT53797.2021.00219

The research of text retrieval based on DeepCT and Conv-KNRM

Bing Ai
Big Data Center of State Grid Corporation of China, Beijing, 100052, China
bing-ai@sgcc.com.cn

Yibing Wang
Big Data Center of State Grid Corporation of China, Beijing, 100052, China
yibing-wang@sgcc.com.cn

Xiaoyu Zhang
Big Data Center of State Grid Corporation of China, Beijing, 100052, China
13366006598@163.com

Liang Ji
Big Data Center of State Grid Corporation of China, Beijing, 100052, China
liang-ji@sgcc.com.cn

Jia Yin
Big Data Center of State Grid Corporation of China, Beijing, 100052, China
jia-yin@sgcc.com.cn

Ting Wang
Big Data Center of State Grid Corporation of China, Beijing, 100052, China
ting-wang@sgcc.com.cn

Wentao Liu
Big Data Center of State Grid Corporation of China, Beijing, 100052, China
wentao-liu@sgcc.com.cn

Junhua Liu*
Department of Computer, North China Electric Power University, Baoding, Hebei, 071000, China
220192221061@ncepu.edu.cn

***Abstract*—Text retrieval is the main means for people to obtain information from a huge text database. Current document retrieval methods do not consider the association between words and context, and cannot effectively retrieve information based on semantics. Aiming at the semantic feature extraction of text retrieval, this paper uses Bert, DeepCT and KNRM and its derivative model to improve the construction and feature extraction of similarity matrix, and proposes a text retrieval model based on DeepCT and Conv-KNRM. The goal is to give appropriate word weights to different keywords and improve the accuracy of text retrieval. In this paper, through experimental comparison, it is proved that this model improves the accuracy of text retrieval.**

Keywords-KNRM; DeepCT; Bert; Similarity matrix

I. INTRODUCTION

Information retrieval (IR) is an important research direction in the field of natural language processing (NLP). Text retrieval is the most important basis of information retrieval, and it is the support of other complex content retrieval. More than 80% of the data on the Internet is retrieved in the form of text. Text retrieval is that the retrieval system finds the set of documents related to the query in the document library according to the query entered by the user, sorts the set according to the relevance sorting algorithm and returns it to the user. The text retrieval model usually maps user queries and candidate documents to the same vector space, and then calculates the correlation between them [1].

With the development of technology, the correlation between query requests and documents is becoming more and more difficult to measure. The traditional retrieval methods based on key string matching can not meet the increasingly complex query requirements because they can not analyze the semantics of words and sentences. With the rise of deep learning technology, people have proposed various methods to calculate semantic relevance. This paper improves on the basis of KNRM and Bert, and proposes the DeepCT+Conv-KNRM model.

II. BACKGROUND

A. *Text retrieval model based on statistics*

The Boolean model is an ancient method based on Boolean expressions. This model constructs a 0/1 binary matrix of query keywords and documents requested by the query. 0 means that the document does not have the query term, and 1 means that the document appears the search term. The query keyword in the Boolean model has the same weight as the word in the document. Its advantage is that the retrieval is efficient and fast, so it is widely used in large-scale retrieval systems that do not require strict word meanings, such as book retrieval systems. Obviously, its shortcomings are that it is unable to retrieve similar documents based on word meanings, and there are only two discrete results for query keywords and related readings of documents, which is difficult to meet complex retrieval requirements[2].

B. *Text matching model based on deep learning*

Microsoft Corporation proposes Deep Structured Semantic Models (DSSM). The principle is that queries and documents are mapped to a vector semantic space of

* corresponding author

978-1-6654-3757-8/21/$31.00 ©2021 IEEE
DOI 10.1109/CECIT53797.2021.00219

the same dimension, using cosine similarity to measure relevance, and maximizing it to train the retrieval model to obtain the final ranking result.

Google proposed the Bert model in 2018, which uses two-way Transform to extract text features, which can be combined with contextual semantic information to solve the problem of long-distance dependence of text sequences [3]. People often use it as a pre-training model. The Bert-based word vector representation method is more accurate than the traditional word2Vec, so it is widely used in the NLP field [4].

The current retrieval model is widely used based on interaction structure. It constructs a similarity matrix by querying keywords and document keywords, and extracting features for training. The most popular interaction-based models are the DRMM and KNRM models. In 2016, JieFeng Guo proposed Deep Relevance Matching Model (DRMM), which uses the histogram mapping method to extract the characteristics of the similarity matrix. This idea has greatly promoted the development of information retrieval.In 2017, YanXiong Chen proposed a kernel based neural model for document ranking (KNRM)[9]. It uses Gaussian kernel mapping to extract features. Soon after the improved Convolutional Neural Networks for Soft-Matching N-Grams (Conv-KNRM) came out, It added a convolutional layer on the basis of KNRM to extract similar features more effectively[2].

However, due to the error of the word vector representation itself, such as Word2Vec cannot consider the ambiguity of a word, it will cause the final retrieval result and its sorting sequence to be inaccurate. In this paper, a comparative experiment has been carried out, and the results show that this method is beneficial to improve the retrieval accuracy.

III. METHODS

How to represent words in vectors is extremely important. The accuracy of word vectors directly affects the feature extraction of the similarity matrix, and ultimately affects the retrieval results. This paper mainly verifies the impact of three models of Word2Vec, Bert and DeepCT on the retrieval results, and combines the Conv-KNRM model to propose an improved retrieval model, which is mainly composed of three parts: keyword semantic representation, similarity matrix and ranking learning. The overall structure is shown in Fig.1.

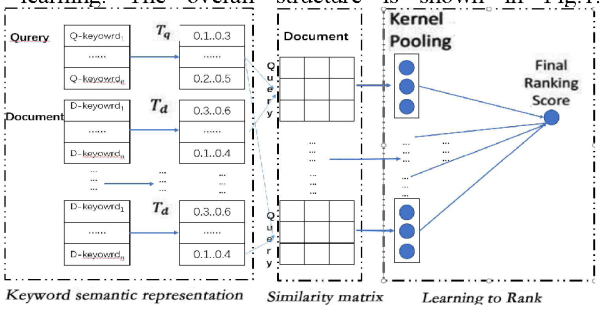

Figure 1. The overall structure of the model

A. Keyword semantic representation

First, perform keyword extraction on all documents in the document library. In recent years, DeepCT (Deep Contextualized Term Weighting framework) is very popular and is a way to set word weights for retrieval. The word frequency is closely integrated with the word vector and is an important influencing factor of the word vector, and the word frequency is an important basis for keyword extraction. However, in long texts, the word frequency distribution curve in the document is relatively flat, difficult to distinguish, and not obvious. DeepCT uses BERT's text to represent the word weights mapped to the context of sentences and paragraphs, and assigns different weights to the same words in different texts. It uses BERT to encode the document and outputs the weight score of each document keyword, which can well represent the characteristics of the semantic vector of the word, and can more accurately measure the importance of the keyword in the document. The main content of DeepCT has two points: (1) The BERT model generates a word embedding vector fused with context information; (2) The weight of the word is output according to the word vector [5].

The DeepCT model converts the word vector of a word into the importance of the word relative to the document. In different documents, the importance of the same word is different, which conforms to the rules of daily life. The most common and simplest mapping method is through linear transformation [6], the formula is as follows.

$$\tilde{y}_{t,c} = \omega T_{t,c} + b \tag{1}$$

$T_{t,c}$ is the semantic vector of the word t of document c, and ω and b are the weight and bias of the linear function, respectively.

This stage is a typical regression task. The weight of each word in document c is the degree of importance, denoted $y_{1,c} \ldots\ldots y_{N,c}$, the label value is denoted as y, the predicted value is denoted as $\tilde{y}$ and the loss function is the mean square error.

$$\text{loss}_{MSE} = \sum_{t,c} \left(y_{t,c} - \tilde{y}_{t,c}\right)^2 \tag{2}$$

The label value and predicted value of the training data set are in the range of [0, 1], negative values are not accepted, and the segmentation word Token of Bert is not counted. The label value of word weight is difficult to measure, and manual labeling is not realistic. However, the keywords used in the query request largely represent the general theme of the document, so the query keywords in the data set can be used as an important basis for the weight of the document keywords, that is, the non-query keys appearing in the documents retrieved based on the query keywords Words are not that important. Based on this idea, this article uses the QTR (query term recall)

method to generate the true weight of the word, the formula is as follows.

$$QTR_{t,d} = \frac{|Q_{d,t}|}{|Q_d|} \tag{3}$$

Q_d is multiple queries associated with document d;$Q_{d,t}$ is some queries of design words in the package collection Q_d; $QTR_{t,d}$ is the recall weight of words in the document, $QTR_{t,d} \in [0, 1]$.

In this way, the weight of the query keywords is obtained. The non-query keywords in the keywords in the document are processed using TF-IDF combined with the BM25 method. In the end, the weights of the query keywords and the document keywords are unified.

This article is aimed at the retrieval of long text documents, using TextRank, LDA-based topic model and DeepCT as the basis for keyword extraction.

Then perform semantic vector representation on the extracted keywords. There are many methods for vectorized representation of words. At present, the mainstream methods include Word2Vec, Glove and Bert. The word vector generated by the Bert method contains the context information of the word and is widely used in the fields of keyword extraction and text summarization. For long texts, the text composition is very complicated, and the word vector generated by Bert is very suitable.

B. *Similarity matrix*

The similarity matrix is one of the most important links of KNRM and its derivative models, and it directly affects the quality of retrieval. The similarity matrix calculates the similarity between the query keywords and document keywords of each query request. This article uses the query keywords as the vertical axis and the document keywords as the horizontal axis, and uses the cosine similarity to calculate the similarity of each pair of words.

$$\cos(q_i, d_i) = \frac{\sum_{i=1}^{n} q_i * d_i}{\sqrt{\sum_{i=1}^{n}(q_i)^2}\sqrt{\sum_{i=1}^{n}(d_i)^2}} \tag{4}$$

q_i is the word meaning vector of the i-th query keyword in the query, d_i is the word meaning vector of the C-th document keyword in the document, n is the total dimension of the word vector, and n=256 is specified, that is, a 256-dimensional word vector is used.

C. *Learning to Rank*

In recent years, KNRM is very popular. It uses the Kernel-pooling method to extract the features of the similarity matrix. Many scholars have improved on this model [7]. This paper uses Gaussian kernel to extract the features of the similarity matrix.

Learning to rank is to quantify the search results and calculate the matching score through regression or classification. The Pairwise method is widely used. It outputs scores for the same query document and converts the sorting problem into a binary classification problem.

Conv-KNRM adds a shared convolution mechanism on the basis of KNRM, which not only strengthens the relationship between query keywords and document keywords, but also considers the relationship between query keywords, and the search results are more accurate.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. *Evaluation standard*

In the field of information retrieval, there are many measurement indicators for the accuracy of information retrieval, and the commonly used ones are NDCG and MRR [8].

The full name of NDCG is Normalized Discounted Cumulative Gain. The basic principle is that the result of the query and the document is highly relevant to the indicator, and when the result of the query and the document is highly relevant, the indicator will increase significantly. NDCG is related to DCG and IDCG, @n is the first n documents, the formula is as follows.

$$NDCG@n = \frac{DCG@n}{IDCG@n} \tag{5}$$

The full name of DCG Discounted Cumulative Gain includes factors related to relevance and location of search results. The formula is as follows.

$$\mathrm{DCG@n} = \sum_{i=1}^{n} \frac{2^{\mathrm{rel_{i-1}}}}{\log_2(i+1)} \tag{6}$$

i represents the position of the current document in the search result, rel_i represents the relevance of the i-th search result, this article sets it to 0-4, a total of 5 levels, and |REL| represents the document sorted by relevance. The formula is as follows.

$$IDCG@n = \sum_{i=1}^{|REL|} \frac{2^{rel_{i-1}}}{\log_2(i+1)} \tag{7}$$

In the end, NDCG$\in$[0,1] can measure different search results.

The average reciprocal ranking of MRR is also a key indicator [9]. The principle of this method is to use the reciprocal of the search result position of a certain query in the data set as the evaluation score of this retrieval, which is biased towards the optimal query.

In addition, in the field of deep learning, there are some general indicators, such as accuracy, recall, F1 indicators, and average accuracy MAP.

B. *Environment*

In order to verify the effectiveness of the proposed model, we conduct experiments on the data set released by

YanXiong Chen [10] and the MQ2007 data set, the environment configuration is shown as Table I.

TABLE I. ENVIRONMENT CONFIGURATION

software and hardware	
CPU	Intel(R) Core(TM) i7-10700 CPU @2.90GHz 2.90 GHz
GPU	NVIDIA GeForce RTX 3070Ti/16G
OS	Ubuntu 16.04
Software	Python3.6、pytorch、pandas

This article compares some mainstream information retrieval algorithms with the above-mentioned hybrid models, using the default parameters of the model and the Adam optimization algorithm, with a learning rate of 0.0001. The construction basis of the matrix is verified by the 10-fold cross-validation method.

C. *Analyze*

Use Word2vec, Bert, DeepCT for keyword semantic representation, KNRM and Conv-KNRM for ranking learning, and experiment results using the data set released by YanXiong Chen as shown in Table II and Table III.

TABLE II. NON-QUERY KEYWORDS IN THE DOCUMENT ARE PROCESSED BY TEXTRANK IN NON-DEEPCT

Model	**NDCG@5**	**NDCG@10**	**NDCG@15**	**NDCG@20**	**MRR**
Word2vec+KNRM	0.341	0.377	0.407	0.436	0.229
Bert + KNRM	0.362	0.407	0.478	0.510	0.241
DeepCT+KNRM	0.367	0.415	0.489	0.522	0.249
Bert+ Conv-KNRM	0.372	0.441	**0.496**	**0.538**	0.258
DeepCT+Conv-KNRM(Ours)	**0.375**	**0.449**	0.493	0.532	**0.269**

TABLE III. NON-QUERY KEYWORDS IN THE DOCUMENT ARE PROCESSED IN NON-DEEPCT USING THE TOPIC MODEL BASED ON LDA

Model	**NDCG@5**	**NDCG@10**	**NDCG@15**	**NDCG@20**	**MRR**
Word2vec+KNRM	0.344	0.375	0.401	0.442	0.231
Bert + KNRM	0.365	0.404	0.469	0.511	0.243
DeepCT+KNRM	0.369	0.418	0.487	0.518	0.253
Bert+ Conv-KNRM	0.371	0.451	0.495	0.528	0.257
DeepCT+Conv-KNRM(Ours)	**0.377**	**0.452**	**0.499**	**0.531**	**0.270**

After using the Bet model, the retrieval accuracy is significantly improved. DeepCT has a slight improvement over Bert, but with the Conv-KNRM model, there is a decline in NDGG@20 and MRR indicators. This is caused by the irregular distribution of the text length of the data set, which is improved by using the topic model based on LDA to extract.

Experiments on MQ2007 data sets, the experimental parameters and model structure are the same as above, and the experimental results are as shown in Table IV and Table V.

TABLE IV. NON-QUERY KEYWORDS IN THE DOCUMENT ARE PROCESSED BY TEXTRANK IN NON-DEEPCT (MQ2007 DATA SET)

Model(MQ2007 data set)	**NDCG@5**	**NDCG@10**	**NDCG@15**	**NDCG@20**	**MRR**
Word2vec+KNRM	0.438	0.449	0.463	0.461	0.318
Bert + KNRM	0.469	0.498	0.509	0.541	0.377
DeepCT+KNRM	0.471	0.499	0.508	0.536	0.371
Bert+ Conv-KNRM	0.477	0.506	0.528	0.568	0.395
DeepCT+Conv-KNRM(Ours)	**0.481**	**0.517**	**0.536**	**0.577**	**0.398**

TABLE V. NON-QUERY KEYWORDS IN THE DOCUMENT ARE PROCESSED IN NON-DEEPCT USING THE TOPIC MODEL BASED ON LDA (MQ2007 DATA SET)

Model(MQ2007 data set)	**NDCG@5**	**NDCG@10**	**NDCG@15**	**NDCG@20**	**MRR**
Word2vec+KNRM	0.429	0.452	0.474	0.477	0.321
Bert + KNRM	0.471	0.488	0.500	0.551	0.379
DeepCT+KNRM	0.476	0.498	0.511	0.554	0.375
Bert+ Conv-KNRM	**0.478**	**0.509**	**0.519**	**0.566**	**0.388**
DeepCT+Conv-KNRM(Ours)	0.476	0.523	0.533	0.579	0.396

After using DeepCT and Conv-KNRM, each indicator has improved. Because this data set is relatively standardized, there is no problem of some indicators falling in the previous data set.

It can be seen that the keyword semantic vector generated by Bert for long text retrieval can effectively improve the retrieval accuracy. DeepCT integrates the word weight information into the semantics, which has a significant impact on the retrieval results, but because the word frequency in the long text is not necessarily the most frequent It is a keyword, or it may be a discussion on the topic of the article, so it may cause instability, and it needs to be carefully considered when applied in the vertical field.

V. CONCLUSION

In the text retrieval task, the text representation is a very important part. The text representation based on deep learning technology can greatly improve the semantic conformity of the text representation and strengthen the computer's understanding of the text, which is very important to the accuracy of the text retrieval. For a word, its semantic vector is not only related to the word itself, but also related to the distribution of the word in the

document. In the field of text retrieval, it is also related to query, which increases the difficulty of text representation.

This paper proposes the BeepCT and Conv-KNRM model. This model combines word weights and word context. The generated keyword semantic vector can better represent its position in the semantic space, so the calculation of word and word similarity is more accurate. It is of great significance to retrieval tasks. The experimental results show that, compared with the typical benchmark algorithm model of text retrieval, the retrieval model proposed in this paper can improve the accuracy in the retrieval task of long text. However, the amount of calculation is relatively large and the retrieval time is relatively long. Consider building a distributed computing platform to solve this problem.

ACKNOWLEDGMENT

This work is supported by the Science and Technology Project of Big Data Center of State Grid Corporation of China "Research on unstructured recognition technology based on NLP natural language analysis and keyword extraction technology service data Think-Tank construction".

REFERENCES

[1] Yuanyuan Q. Research on Text Retrieval based on Feature Representations Learning[D]. Beijing University of Posts and Telecommunications, 2021.

[2] Lei B. Research on text retrieval method based on graph entity representation and ranking learning[D].Jilin University, 2020.

[3] Briskilal J,Subalalitha C.N. An ensemble model for classifying idioms and literal texts using BERT and RoBERTa[J]. Information Processing and Management, 2022, 59(1): 102756

[4] Yurong W,Ming L.BERT Mongolian word vector learning[J/OL].Computer Engineering and Applications:1-7 http://kns.cnki.net/kcms/detail/11.2127.tp.20210915.1133.004.html.

[5] Lai Jiang,Mai Xu,Shanyi Zhang,Leonid Sigal. DeepCT: A novel deep complex-valued network with learnable transform for video saliency prediction[J]. Pattern Recognition, 2020, 102: 107234

[6] Jiang, Lai, Xu, Mai, Zhang. Deepct: a Novel Deep Complex-valued Network With Learnable Transform for Video Saliency Prediction[J]. Pattern Recognition, v102, June 2020: 107234

[7] Xiong C, Dai Z, Callan J, et al. End-to-end neural ad-hoc ranking with kernelpooling[C]// Proceedings of the 40th International ACM SIGIR conference onresearch and development in information retrieval. 2017:55-64.

[8] Zhouling T. design and implementation of text retrieval system based on deep learning[D].Beijing University of Posts and Telecommunications, 2019.

[9] Biying Z.Research on Text Intelligent Retrieval Technology Based on Natural Language[D].Northwest University, 2013.

[10] Xiong C, Power R, Callan J. Explicit semantic ranking for academic search viaknowledge graph embedding[C]//Proceedings of the 26th international conferenceon world wide web. International World Wide Web Conferences SteeringCommittee, 2017: 1271-1279.