NAUIR at the NTCIR-15 WWW-3 Task

Zhu Liang
Nanjing Agricultural University
2019114015@njau.edu.cn

Jinling Shang
Nanjing Agricultural University
19117228@njau.edu.cn

Xuwen Song
Nanjing Agricultural University
2019814040@njau.edu.cn

Si Shen
Nanjing University Of Science And Technology
Shensi@njust.edu.cn

ABSTRACT

The NAUIR team participate in English subtasks of the NTCIR-15 We Want Web-3 (WWW-3) task. This paper describes our methods and results in the English subtask of WWW-3. For the English subtask in this task, we use the modified DRMM model and the BERT model for query and document matching. In the pre-training, we only use the documents of the WWW-3 task for word embedding training. The BERT model uses the bert-base-uncased pre-training model officially provided by Google. The experiment shows that the results of our modified DRMM model and the BERT model are better than those of BASELINE.

TEAM NAME

NAUIR

SUBTASKS

English Subtasks

1 INTRODUCTION

The NTCIR-15 WWW-3 task [7] is an ad-hoc web search task, mainly to solve the problem of poor representation of the query in the web search task. In the NTCIR-15 WWW-3 Task, a large number of open queries are provided to test the retrieval results of the problem. In such information retrieval tasks, the advantages of deep learning methods in information retrieval have been brought into full play.

At present, many deep learning retrieval models have good effects on the sorting of documents, such as DSSM [4], DRMM [2], etc. Guo et al. [3]divided deep learning models into (1) The first type are defined as the Representation-focused model. This type of model represented by the DSSM model focuses on how to vectorize the query and text content separately, and how to better use the representation results to calculate text similarity. (2) The second category can be defined as the Interaction-focused model. This method mainly constructs an interaction-matrix between input texts and calculates the matrix with a deep neural network to obtain the potential similarity between the texts. For example, the DRMM [2] model implements interaction-matrix by using the term-level. In addition, many studies have begun to use pre-training models such as BERT [1] to achieve semantic similarity calculations between long texts.

In the English subtasks achieved in this work, we use both the DRMM model and the BERT model to calculate the similarity between queries and documents. When using the DRMM model, we first use the nearly 160,000 documents to train word embedding and use word embedding for the subsequent construction of Interaction-matrix. Subsequently, we change the traditional training method of constructing positive and negative document pairs to the method of evaluating the similarity score, and find that the modified model is not only more stable in training, but also can obtain better training effects than traditional BM25 and other classic methods. In the process of using the BERT model, we build a query and document pair based on the BERT model to evaluate the relevance between query and document. Experimental results also show that the model can achieve better results than BM25 on this task.

2 METHOD AND FREAMWORK

2.1 Data Set

We use the NTCIR-15 WWW-3 English data set officially provided by NTCIR, which contains 80 WWW-2 queries and 80 WWW-3 queries, and 160,000 documents. We use 80 queries from WWW-2 competitions and their related documents to construct a training set. At the same time, we use the Word2vec [5, 6] model to train word embedding from 160,000 documents. The parameters of the model are the vector dimension is 100 dimensions, the min_count is 2, and the window size is 5. Word2vec can be trained to obtain a better semantic vector representation based on a limited data set.

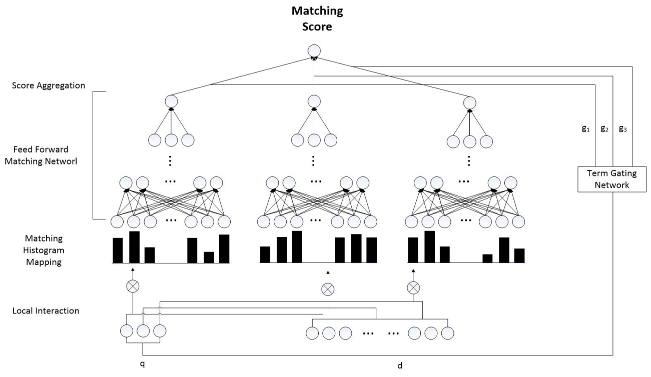


Figure 1: The model structure of DRMM

2.2 Our methods

2.2.1 DRMM with MSE loss function. The DRMM model is a typical Interaction-focused model. In the English subtask, we use Word2vec in the model to vectorize the query and the text of the document, calculate the relevance between the query and the words in the document, and construct an interaction-matrix based on matching

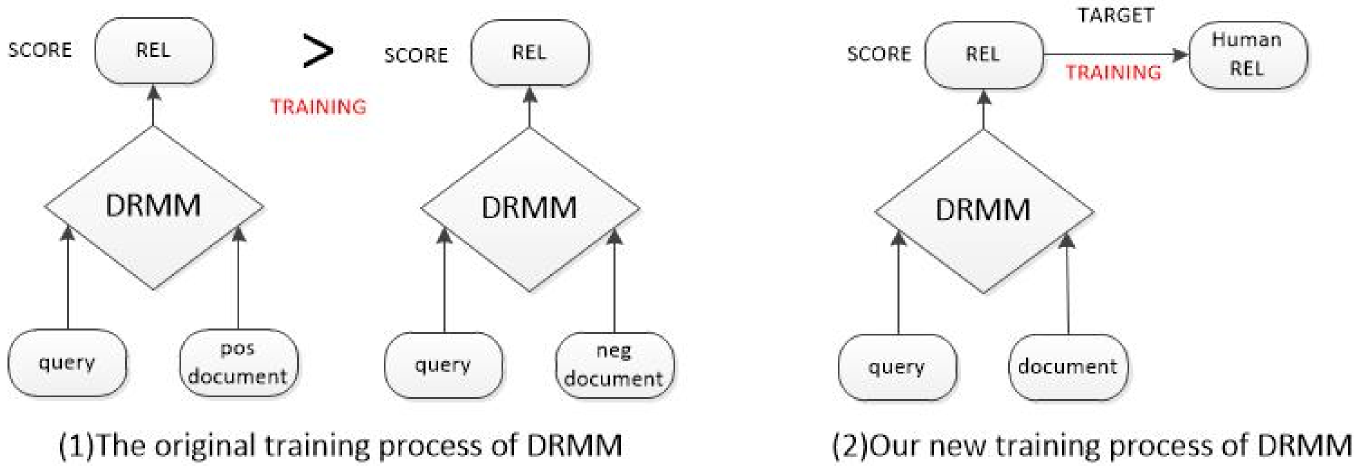


Figure 2: The training method of DRMM model

histogram. This matrix represents the semantic similarity between query and document from the term-level perspective. The specific content of the DRMM model is shown in Figure 1.

In the task, we set the similarity interval of the two words as {[-1,-0.5), [-0.5,-0), [0, 0.5), [0.5,1),[1,1]}.The Word2vec model is used to train word embedding for semantic representation of text. This method can be used to calculate the similarity between the query and the document at the term level. At the same time, we use the log function to reduce the value range of the matrix to facilitate model convergence.

In the training process of the model, we mainly use the two methods shown in Figure 2. (1) The original loss function of the DRMM model. For triples (q, d^+, d^-), when the query is q and the order of the document d^+ is higher than d^-, the loss function is defined as $L(q, d^+, d^-; \theta) = max(0, 1 - s(q, d^+) + s(q, d^-))$. $s(q, d)$ represents the relevance of query q and the document d. (2) MSE loss function. This function refers to the design idea of the loss function of THUIR [8] in the NTCIR-14 WWW task. Subsequent results show that the training method using the MSE loss function can get better results.

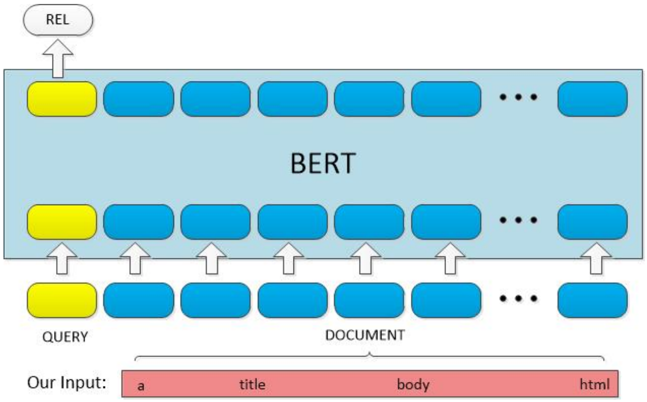


Figure 3: The model structure of BERT for web search

2.2.2 Retrieval model with BERT for web search. Given the neural network model proposed by Google in 2018 that uses the bidirectional Transformer as the basic network structure, this model can model the text based on the self-attention mechanism on the basis of discarding the recurrent neural network structure. Compared with the recurrent neural network model, the BERT model has a natural advantage. We choose to use the BERT model for information retrieval, as shown in Figure 3. First, we construct the document representation by arranging the tag of the document according to the order of the <a> tag, the <title> tag, the <body> tag, and the <html> tag. Therefore,the important content in the <html> tag is located in front of the document, and then the useless content is cut accordingly. We choose bert-base-uncased as the pre-training model, AdamW as the optimizer for model training, and max_seqlenth is 512.

3 EVALUATION AND RESULT

In the training process of the DRMM model and the BERT model, we use the WWW-2 qrels to build the training set and the validation set to complete the NTCIR-15 WWW-3 task. Then, we retain the model that can get better results according to the result based on the validation set. And we select the appropriate re-ranking range based on the results of the closed test. Generally speaking, the <body> tag of the document only includes the browseable content of the website, and the <html> tag of the document also includes website self-indexing information, such as keywords and titles. This self-indexing information is generally considered to have the meaning of being retrieved. Therefore, in the process of DRMM model parameter selection, we use the content of different document structures for evaluation: mainly including the text content of the <body> tag and the text content of the <html> tag. At the same time, due to the lack of the global IDF value in the real data in this task, we use all the document sets provided by the organizer to calculate the IDF value. Although there may be some differences in the results, we believe that this IDF value can still judge the importance of each word in the query, and the method has little effect on the results.

Table 1: The basic information comparison of our model

run	Model	Loss	Document Structure	Re-rank Range
NAUIR-E-CO-NEW-1	DRMM	MSE	Html	40
NAUIR-E-CO-NEW-2	DRMM	MSE	Body	60
NAUIR-E-CO-NEW-3	DRMM	Original	Html	70
NAUIR-E-CO-NEW-4	DRMM	Original	Body	70
NAUIR-E-CO-NEW-5	BERT	Original	A, Title, Body, Html	50

The loss function, document content, and Re-rank Range used by each model are shown in Table 1.

In the process of training the DRMM model, we use word embedding and IDF values to generate an interaction matrix from all documents and queries before training the neural network, which can reduce the computer's memory loss and increase the iteration speed of the model. The experimental results show that after using the above training method, it only takes about 1 hour to complete the model training. As shown in the third column of Table 1, in the result of submission using the DRMM model, we use two loss functions described above: the original loss function of the model and the MSE function. In terms of the document structure used in the fourth column of Table 1, we extract the text content in the document in the order of tags. However, under the limitations of the BERT model and hardware conditions, the length of the document is limited to 512 tokens. Therefore, according to the characteristics of web documents, we combine the content of the <a> tag, the content of the <tilte> tag, the content of the <body> tag, and the content of the <html> tag to generate a document set containing only text and no tags.

Table 2: The results of our model on the open test

run	nDCG	Q	ERR
baselineEng	0.5748	0.5850	0.6757
NAUIR-E-CO-NEW-1	0.5989	0.6089	0.7144
NAUIR-E-CO-NEW-2	0.5980	0.6095	0.7190
NAUIR-E-CO-NEW-3	0.5851	0.5977	0.6915
NAUIR-E-CO-NEW-4	0.5557	0.5712	0.6786
NAUIR-E-CO-NEW-5	0.5982	0.6083	0.7124

The experimental results of the model under the open test are shown in Table 2. The results show that the document structure has no significant effect on the results of the model. Whether using the content of the <html> tag or the content of the <body> tag, not only NAUIR-E-CO-NEW-1 and NAUIR-E-CO-NEW-2 have no significant difference in the Randomised Tukey HSD test, but NAUIR-E-CO-NEW-3 and NAUIR-E-CO-NEW-4 are also not significantly different. For the DRMM model, using the MSE loss function to evaluate the relevance between the query and the document can get a better result, rather than using the original loss function to predict the order of the document pair. The above conclusion is reflected in the NDCG indicators in Table 2. It can be seen that NAUIR-E-CO-NEW-1 and NAUIR-E-CO-NEW-2 are significantly different from NAUIR-E-CO-NEW-3 and NAUIR-E-CO-NEW-4 in the Randomised Tukey HSD test. This result can show that the training method that directly evaluates the relevance between the query and the document benefits the convergence of the model training process to get better results. But for the Bert model, due to the powerful feature extraction capabilities of Transformer during the pre-training process, the model can represent strong semantics between the query and the document. This method is very helpful to evaluate the relevance between query and document, so the result of NAUIR-E-CO-NEW-1, NAUIR-E-CO-NEW-2, and NAUIR-E-CO-NEW-5 in Table 2 do not show a significant gap.

On the other hand, although the BERT model generally gets a good result in tasks, the difference between the model and the DRMM model using the MSE loss function is not significant in indicators such as NDCG, Q, and ERR. The reason behind the preliminary speculation may be due to the small training data set and the inability to take advantage of the BERT model in a large-scale corpus. Therefore, the word embedding trained through 160,000 documents can obtain good semantic word vectors, which can meet the requirements of the DRMM model design. And the performance of the model is not different from the BERT model.

Based on the above analysis, we analyzed various query topics in the open test (WWW-3 task), as shown in Table 3. After we manually classify query topics, we find that the effects of retrieval models under different topics are different. The query provided in the WWW-3 task can be preliminarily divided into six categories: entity(name, company, etc.), daily, commodity, terminology, website address, and knowledge inquiry.

Table 3: The topic distribution of query in WWW-3 task

query topics	Num	Example
Entity	22	Peppa Pig,FedEx
Daily	14	greeting cards
Commodity	3	honda
Terminology	21	Frozen shoulder
Website address	4	bank of america
Knowledge inquiry	16	learn spanish

On the basis of the above analysis, we continue to compare and analyze the performance of our BASELINE and our model under each query, as shown in Table 4. We compare the results of NAUIR-E-CO-NEW-1, NAUIR-E-CO-NEW-2, and NAUIR-E-CO-NEW-5 with that of BASELINE in different topics under the evaluation indicators of NDCG, such as Table 4 shows. The results show that the results of our proposed DRMM model (NAUIR-E-CO-NEW-1, NAUIR-E-CO-NEW-2) and BERT model (NAUIR-E-CO-NEW-5) are better than that of the traditional BM25 algorithm (BASELINE) in the "daily" topic and the "knowledge inquiry" topic; there is no obvious difference in the "entity" topic and the "terminology" topic.

Table 4: Differences in model results under different topics

query topics	NAUIR-E-CO-NEW-1	NAUIR-E-CO-NEW-2	NAUIR-E-CO-NEW-5	baselineEng
Entity	0.526790909	0.499354545	0.555804545	0.528236364
Daily	0.628428571	0.633471429	0.644942857	0.585692857
Commodity	0.7015	0.7016	0.673233333	0.6962
Terminology	0.607171429	0.633442857	0.573333333	0.593809524
Website address	0.63695	0.629225	0.591975	0.496
Knowledge inquiry	0.63251875	0.62864375	0.6355375	0.60123125

Regrettably, there are fewer queries involved in the "commodity" topic and the "website address" topic, which are not meant for discussion. It is worth noting that the model proposed by us contains rich semantic information, and the queries of the "daily" topic and the "knowledge inquiry" topic usually contain deep semantic information. Therefore, we believe that the advantage of our model is that it contains semantic information, which makes up for the shortcomings of the traditional retrieval model (BM25) in information retrieval.

Subsequently, we sort according to the evaluation results of each query of BASELINE. The first 40 queries represent queries with poor BASELINE performance, and the last 40 queries represent queries with better BASELINE performance. Under the different performance of BASELINE, the differences between our model(NAUIR-E-CO-NEW-1, NAUIR-E-CO-NEW-2, NAUIR-E-CO-NEW-5) and BASELINE are shown in Table 5. The results show that in the open test, our model has a significant improvement in the results of the first 40 queries. And the difference between our model and BASELINE has been significantly reduced in the results of the last 40 queries. Therefore, we believe that our model performs well in the query with the poor performance of the BASELINE.

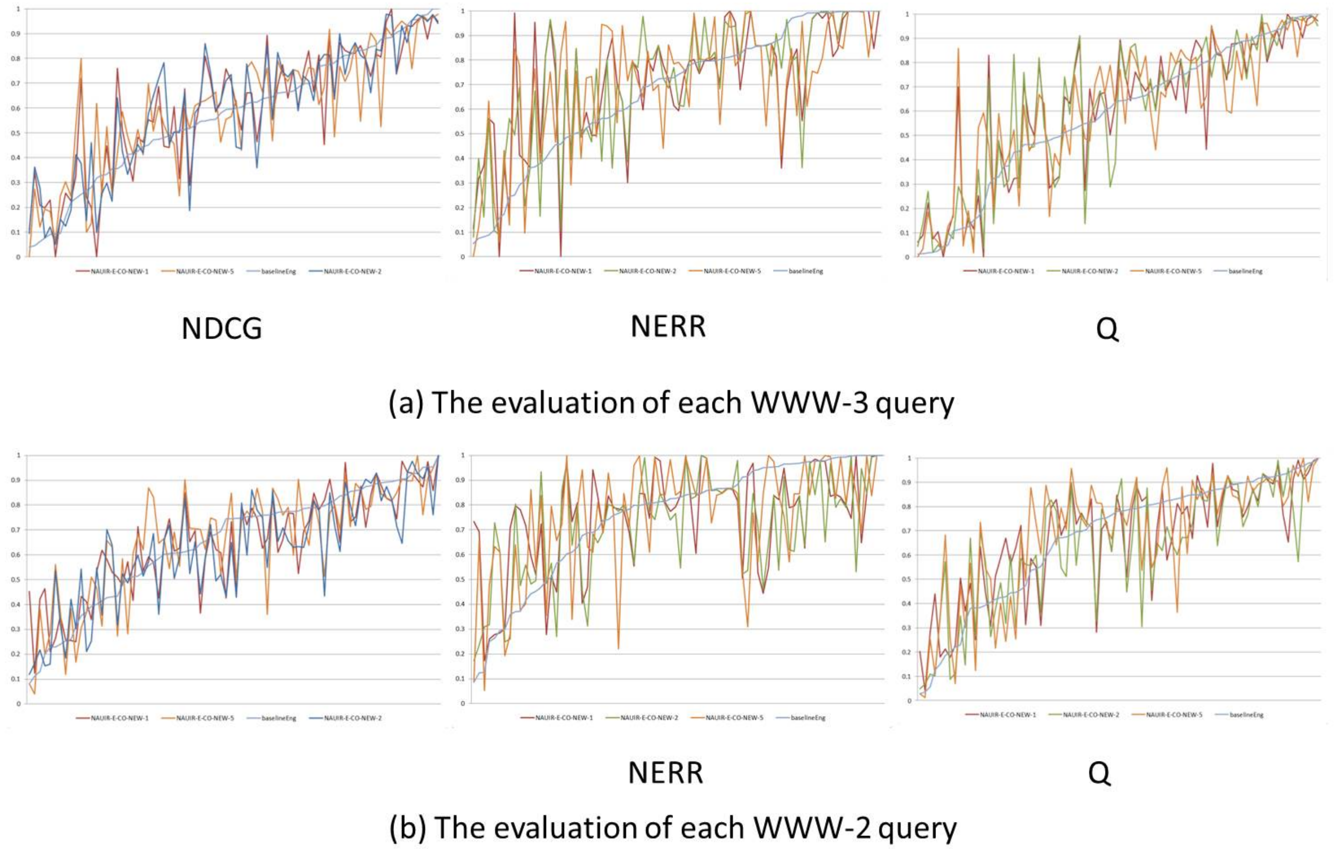
Finally, the performance of the model under the corresponding query in the closed test and the open test was compared and analyzed. We select three models with better performance: NAUIR-E-CO-NEW-1, NAUIR-E-CO-NEW-2, and NAUIR-E-CO-NEW-5. We arrange the evaluation results of each query in BASELINE from low to high, and map the results of each model corresponding to each query. The results are shown in Figure 4. As can be seen from Figure 4, the results of our model gradually increase with the results of BASELINE and float up or down. Overall, the results of our model are modified on the results of the BASELINE model. From Figure 4(b), it can be seen that our model differs greatly from BASELINE on each query in the closed test, and the overall performance is slightly inferior to the BASELINE result. Only NAUIR-E-CO-NEW-5 is not much different from BASELINE in overall performance. It is worth noting that, as shown in Figure 4(a), the model has a very good improvement when BASELINE performs poorly in the open test, and when BASELINE performs perfectly in the open test, the result of our model is slightly worse.

Table 5: The performance of the model is under the different performance of BASELINE

NDCG				
	WWW-2 query		WWW-3 query	
run	first 40	last 40	first 40	last 40
NAUIR-E-CO-NEW-1	0.0384725	-0.04454	0.060705	-0.01252
NAUIR-E-CO-NEW-2	0.003185	-0.06717	0.05742	-0.01108
NAUIR-E-CO-NEW-5	0.055625	-0.04324	0.06134	-0.01459
baselineEng	0	0	0	0
NERR				
	WWW-2 query		WWW-3 query	
run	first 40	last 40	first 40	last 40
NAUIR-E-CO-NEW-1	0.094215	-0.10312	0.130588	-0.05332
NAUIR-E-CO-NEW-2	0.05137	-0.12945	0.108048	-0.02142
NAUIR-E-CO-NEW-5	0.0847975	-0.08692	0.141965	-0.06871
baselineEng	0	0	0	0
Q				
	WWW-2 query		WWW-3 query	
run	first 40	last 40	first 40	last 40
NAUIR-E-CO-NEW-1	-0.002697	0.0945	0.05498	-0.00724
NAUIR-E-CO-NEW-2	-0.024398	-0.0425	0.039995	0.008933
NAUIR-E-CO-NEW-5	-0.000695	-0.2033	0.065893	-0.01939
baselineEng	0	0	0	0

4 CONCLUSION

In the NTCIR-15 WWW-3 task, we participate in the English subtask, and our model has achieved good results compared to BASELINE in the open test. In this English subtask, we modify the original training method of the DRMM model, and used a limited corpus to train word embedding, and achieved good results. In the training process of the DRMM model, we first use the original training method of the DRMM model for training, and it don't exceed BASELINE in the closed test. The results in the open test are also not different from BASELINE. And we use the modified training method, no matter in the closed test and the open test, the performance of the model surpassed BASELINE. In the experiment of the BERT model, because the document length of the document is inconsistent and the content of the document's some tags contains rich information, we arrang in order according to the importance of the tag content. Under the premise of being restricted by the hardware environment, we try our best to retain the useful information of the document during the operation of cutting the document length. Although due to the limited data of queries and documents in this task, the characteristics of the BERT model on large-scale corpus cannot be fully utilized.And the advantages of the model are still reflected in the experimental results. Through further analysis of the experimental results, it is found that the modified DRMM model and the BERT model can perform better than traditional retrieval models such as BM25 in the queries of the "daily" topics and the



(a) The evaluation of each WWW-3 query

(b) The evaluation of each WWW-2 query

Figure 4: The results of the model on the training set and test set under different queries

"knowledge inquiry" topic. In future research, we will pay more attention to the modification of the interaction-matrix under the word vector trained with a limited data set.

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[2] Jiafeng Guo, Yixing Fan, Qingyao Ai, and W Bruce Croft. 2016. A deep relevance matching model for ad-hoc retrieval. (2016), 55–64.

[3] Jiafeng Guo, Yixing Fan, Liang Pang, Liu Yang, Qingyao Ai, Hamed Zamani, Chen Wu, W Bruce Croft, and Xueqi Cheng. 2019. A deep look into neural ranking models for information retrieval. *Information Processing & Management* (2019), 102067.

[4] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 2333–2338.

[5] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[6] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[7] Tetsuya Sakai, Sijie Tao, Zhaohao Zeng, Yukun Zheng, Jiaxin Mao, Zhumin Chu, Yiqun Liu, Maria Maistro, Zhicheng Dou, Nicola Ferro, and Ian Soboroff. 2020. Overview of the NTCIR-15 We Want Web with CENTRE (WWW-3) Task. In *Proceedings of NTCIR-15*. to appear.

[8] Yukun Zheng, Zhumin Chu, Xiangsheng Li, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2019. THUIR at the NTCIR-14 WWW-2 Task. In *NII Conference on Testbeds and Community for Information Access Research*. Springer, 165–179.