NKU at TREC 2016: Clinical Decision Support Track

Hualong Zhang and Liting Liu
Intelligent Information Processing Lab,
College of Computer and Control Engineering，NanKai University
nankaizhl@gmail.com, 2120160451@mail.nankai.edu.cn

Abstract

This paper describes the participation of the NKU team at TREC2016 Clinical Decision Support track (CDS2016). The core problem is to find the most relevant literatures from the massive biomedical literatures according to the patient's condition and the needs of doctors. Unlike previous years' games, CDS2016 adds the note type querys[1], which are the original records from real clinical environment, apart from the summary and description topics.

Our work involves three aspects: the expansion of the query, medical literature preprocessing and weight model selection. We use Terrier as the search engine to test the query expansion methods such as pseudo relevance feedback(PRF), MeSH synonym expansion, query type vocabulary expansion, and weighting models such as TF_IDF, BB2, In_expB2 and In_expC2. In the experiment, we build the model based on the CDS2015 data set and do performance evaluation. For both summary and description, we get NDCG values over 0.3.

Keywords

Biological information retrieval, query expansion, search optimization

1 Instruction

The TREC Clinical Decision Support Track 2016 (CDS2016) requires selected articles from the PubMed Central (PMC)[2][3] Biomedical Literature to contain the most relevant articles for the corresponding symptom and clinical purpose for 30 Topics with summary, description and note. Unlike previous years' games, CDS2016 adds the note type querys[1], which are the original records from real clinical environment, apart from the summary and description topics. Each submission of the results can only use one type in summary, description or note as query.

Table 1: Three type of case description[1]

PatientInfo	Content
Summary	A 78 year old male presents with frequent stools and melena.
Description	78 M transferred to nursing home for rehab after CABG. Reportedly readmitted with a small NQWMI. Yesterday, he was noted to have a melanotic stool……
Note	78 M w/ pmh of CABG in early [**Month (only) 3**] at [**Hospital6 4406**] (transferred to nursing home for rehab on [**12-8**] after several falls out of bed.) He was then readmitted to [**Hospital6 1749**] on [**3120-12-11**] ……

There are three query types or the clinical purposes of the search, namely Diagnosis, Test and

Treatment.

Table 2: Query Types[3]

Type	Clinical purposes
Diagnosis	The case should be diagnosed into what
Test	What should be done for the patient examination and testing
Treatment	How to treat the case

2 Method

Our work involves three aspects: the expansion of the query, medical literature preprocessing and weighting model selection.

Model was trained and tuned with the help of the CDS2015 data set. Since the topics in 2015 dose not include the note part, we need to do more analysis and assumptions about the content of note.

Three types of case descriptions were observed. Combined with the results of TREC_CDS in previous years, we found that the content of summary was manually simplified and summarized. Summary is generally short, the number of words is small but the quality is high. The problem with retrieving with summary is the lack of information that results in a lower recall, so we do more query expansion for summary.

Descriptions are generally longer, in previous games the search using description was often not as good as summary. The main problem with using description to retrieve is that a large amount of noise leads to the result with irrelevant documents. The content of note part is often longer than description's, using note to retrieve also faces the problem of noise removal. [4][5][6][7][8]

2.1 Medical literature preprocessing

It is easy to find that a biomedical article often contains a lot of contents which are not related to research, for example, a variety of author information, Publishing information, organizational information and reference information. If these contents occupy a large space, and we also build index for them, then the search results will cause great interference. So we need to extract the core content of the literatures and store them separately.

In addition, the project uses the Terrier platform to index documents, requiring documents to meet certain formats, and for the subsequent establishment of different regions of the index pre-set labels. At the same time, we also want to know how much each part of the literature contributes to the search accuracy.

We write python scripts that extract the pmcid, title, keyword, abstract, and body parts of each document from the original article and organize them into acceptable XML formats for Terrier.

2.2 Pseudo relevance feedback

Pseudo relevance feedback is the most commonly used query expansion. The principle is that after a regular search, assuming that the results of the first few documents are more accurate. We select a number of representative vocabularies from the first few selected documents as an extension to add to the original query. Then a second round retrieve can be performed to obtain more relevant literatures[9].

There are two parameters that we need to consider in the search process:

TopDocuments

TopTerms

2.3 Synonym expansion

Synonym expansion refers to the use of NCBI MeSH terms for synonym expansion. MeSH contains a large number of biomedical terms, with the same medical concept of the different words organized with same entry[10]. With MeSH we can expand the Topics with the synonyms of medical terms in it. A Python script was used to crawl medical synonyms with the same entry from MeSH for the original query.

2.4 Query type vocabulary expansion

Because the task involves three kinds of query types: Diagnosis, Test and Treatment. It is difficult to only select articles belonging to the specified Query Type and shielding other types of articles related to the Topics. A Python script is used to crawl the terms related to diagnosis, testing, and treatment from MeSH to add words related to the specific Query Type to the original Topic content in a proportion.

2.5 Word processing

It is not possible to use the original words in a document or query when indexing a document set or retrieving it from a Topic, for the simple reason that a word will appear in different forms in the text and some words have no meaning to the retrieve but also produce interference. So the word preprocessing is an indispensable step. There is much work on word processing, here we only mentioned the removal of stop words and stemming.

2.5.1 Remove stop words

Stop words refer to words that should be removed in natural language processing[11]. We can define our own list of disabled words according to different purposes, such as is, at, which, on. We use Terrier's own stop words list, words in which will be removed during indexing and retrieval.

2.5.2 Stemmer

In linguistics and information retrieve, stemming is the process of cutting complex words into their stem, base, or root[12]. For example, in English, we will process happiness into happy, happy is the stem of happiness. The most notable stemming parser is the Porter stemmer[13] which was used in our lab.

2.6 Weighting model

Terrier platform implements a batch of commonly used weight models, we can easily configure and switch them. By doing experiment on different weighting models we can rank and compare algorithms by the search results. The weighting models involved in the experiment are as follows:[14]

BB2(DFR)	BM25	DFR_BM25
DLH	DLH13	DPH（DFR）
Hiemstra_LM	IFB2	In_expB2
In_expC2	InL2	LemurTF_IDF

LGD PL2 TF_IDF

2.7 Selection of search area

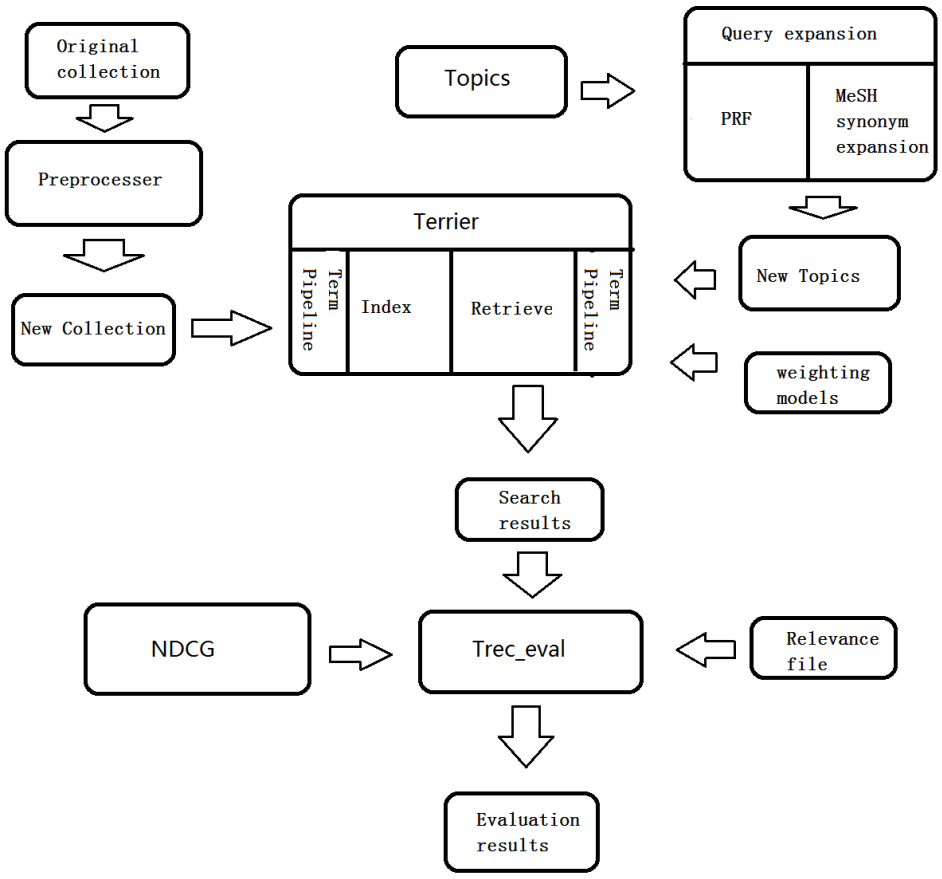
As we have already said, not all of the texts in a document are relevant to the core meaning of the article, so some texts should be removed from the medical literatures before we build index, such as author information, publisher information, organization information, and so on.

Similarly, even the effective content of one article, the contribution of different parts is different. This paper will compare the worth of abstract, keywords, body and medical literature in clinical literature retrieving by experiment.

In the experiment, we built index for Title, Keywords, Abstract, Body and Fulltext of articles separately. Then we retrieved from each part with BM25 weighting model. And the results were evaluated by the trec_eval program for comparison.

3. Experiment and result

Fig.1 Diagram of the retrieve system designed to optimize the experimental.

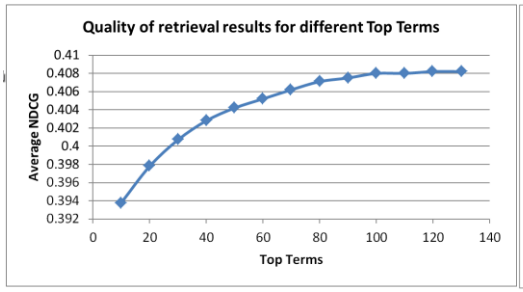


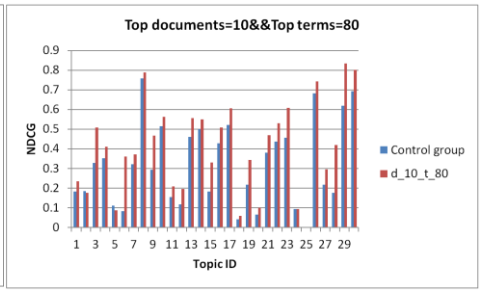
The original collections are pre-processed by scripts, organized into new collections and indexed by Terrier. The original query Topics are expanded to new Topics, together with the selected weight algorithm to be the input of Terrier search function, then output search results. The retrieve results will be input to the Trec_eval program together with the relevance document, and the NDCG is used as the evaluation algorithm. The evaluation value of the retrieval is obtained.

3.1 Pseudo relevance feedback

3.1.1 Top terms parameter

Usually the use of pseudo relevance feedback Top document number in about 10 to 50, we choose 10 as the Top document to find the best number of vocabulary expansion.



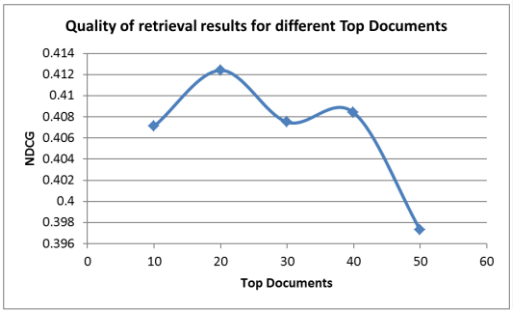


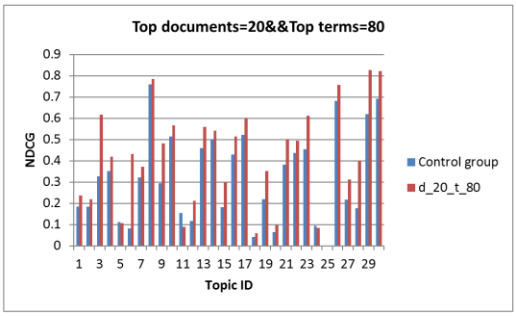
It can be seen that when the number of extended words exceeds 80, the increase of the number of extended words is no longer obvious for the optimization of the search results, since the IDF of the newly introduced extended words is very low when the Top terms exceeds 80. When the Top terms is 80, the 30 search results compared with the control group, there is a substantial increase in the score.

3.1.2 Top documents parameter

In this paper, we choose 80 as the best extended vocabulary number for pseudo-correlation feedback, and find the effect of Top document number on the search results.

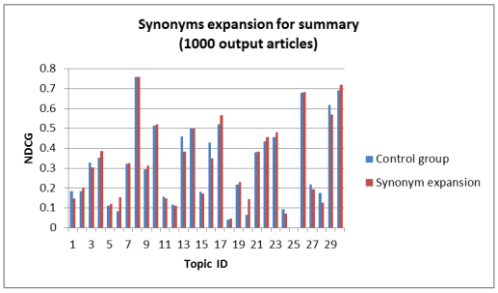
It can be seen that when the Top document number is around 20, the NDCG value of the search results reaches the highest value compared with Top documents = 10 group and no PRF group. The results of the 30 queries are as follows. The average NDCG value was over 0.41, and only 5 out of the 30 results was less than 0.2.





3.2 Synonym expansion

We did not find a way to make the results better by expanding the topics with synonyms. The noise generated when the query is extended using the synonym term, makes the effect more unpredictable.



3.3 Query type vocabulary expansion

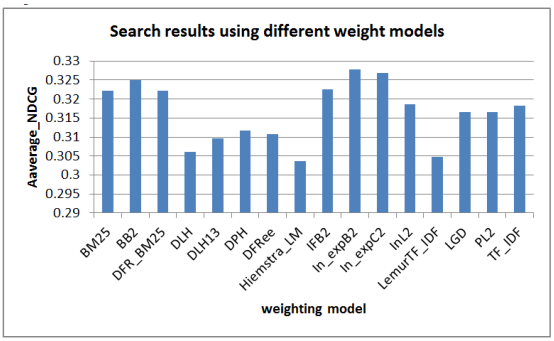
It can be seen that the inclusion of type-related information into the query will help to make the medical literature more precise for different clinical purposes.

Table 3: Query type vocabulary expansion & Original search

Topic type	Average NDCG for 30 Topics retrieve
Summary	0.3186
Summary+Type	0.3221
Description	0.2962
Description+Type	0.2982

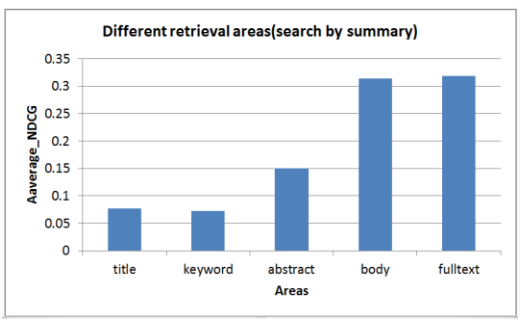
3.4 Weighting model

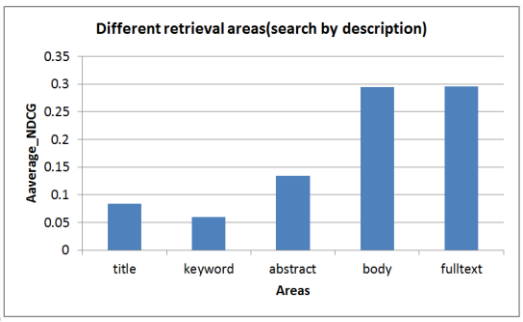
In the experiment, we compare the search results of different weighting models and ranking algorithms. The result of BM25 weighting model was used as control group. The average NDCG of BM25 group was 0.3221. Each group of experiments was retrieved using the contents of summary + type in 30 Topics.



The experimental results show that BB2, IFB2, In_expB2 and In_expC2 are the best. And these four models are the weight model of DFR framework. The Divergence from Randomness (DFR) paradigm is a generalization of one of the very first models of Information Retrieval, Harter's 2-Poisson indexing-model[15][16]. It provides a 1 + 2 search model framework, that is, a basic model (Basic Randomness Models) plus two normalization methods (First Normalization and Term Frequency Normalization). Terrier supports a variety of retrieval models based on DFR framework, such as BB2, IFB2, In_expB2, In_expC2, etc.[15]

3.5 Selection of search area





We can see the body part of the content has the best performance similar to Fulltext. We think title, abstract and keywords and other areas are not dominant in the relevance of the search due to

their shorter text length, less vocabulary, the description of the relevant medical problems is not clear enough,.

3.6 Final submission

The final five submissions of our team CDS2016 and the corresponding treatment are shown in the table 4.

Table 4: final five submissions

RunID	Topics type	Method
nkuRun1	notes	Pretreat + TypeWords + BM25 + n_expB2
nkuRun2	notes	Pretreat + TypeWords + BM25 + PRF
nkuRun3	summaries	Pretreat + TypeWords + ln_expB2 + PRF
nkuRun4	descriptions	Pretreat + TypeWords + BM25 + PRF
nkuRun5	notes	Pretreat + ln_expB2 + PRF

Table 5 Evaluation results of five submission

RunID	infNDCG	P@10
nkuRun1	0.1959	0.2900
nkuRun2	0.1570	0.2233
nkuRun3	0.1984	0.2700
nkuRun4	0.1516	0.2100
nkuRun5	0.1976	0.2767

Submissions nkuRun1, nkuRun4 and nkuRun5 were all listed in Top8 in the corresponding groups[17].

3.7 Conclusion

This project analyzes the project of medical literature retrieval in TREC in previous years. The emphasis is on the optimization of query expansion and weighting model of retrieve, which basically achieves the expected results.

From the view of query expansion, the experiments show that the setting of PRF parameters and the choice of weight model have a great impact on the optimization of search results. The query type-related vocabulary is added into the query to help more accurate classification and retrieval. But the use of synonymous MeSH terms to query content did not achieve the desired effect, to be further improved.

From the perspective of the retrieved documents, filtering and preprocessing of the original text can also play a role. Through the experiment of different parts in articles, the most valuable part of the medical literature is the main part of the literature, that is, the Body part.

At the same time, I admit that we have over-fitted the CDS2015 data which leads to differences between experimental and real results.

Reference

[1] Text REtrieval Conference (TREC). http://trec.nist.gov/act_part/tracks/clinical/topics2016.xml [DB/OL]

[2] Wikipedia contributors. PubMed Central [Internet]. Wikipedia, The Free Encyclopedia; 2016 Apr 7, 20:10 UTC [cited 2016 May 11]. Available from: https://en.wikipedia.org/w/index.php?title=PubMed_Central&oldid=714126105.

[3] Text REtrieval Conference (TREC). TREC Clinical Decision Support Track[DB/OL]. http://www.trec-cds.org/2015.html

[4] Kirk Roberts, Matthew S. Simpson, Ellen Voorhees, William R. Hersh. Overview of the TREC 2015 Clinical Decision Support Track. The Twenty-Fourth Text REtrieval Conference (TREC 2015) , Nov 2015, Gaithersburg, United States. 2015, The Twenty-Fourth Text REtrieval Conference (TREC 2015) Proceedings.

[5] Asma Ben , Abacha, Saoussen Khelifi . LIST at TREC 2015 Clinical Decision Support Track:Question Analysis and Unsupervised Result Fusion. The Twenty-Fourth Text REtrieval Conference (TREC 2015) , Nov 2015, Gaithersburg, United States. 2015, The Twenty-Fourth Text REtrieval Conference (TREC 2015) Proceedings.

[6] Sihui Zhang, Bin He , Weiguo Fan . CBIA VT at TREC 2015 Clinical Decision Support Track-Exploring Relevance Feedback and Query Expansion in Biomedical Information Retrieval. The Twenty-Fourth Text REtrieval Conference (TREC 2015) , Nov 2015, Gaithersburg, United States. 2015, The Twenty-Fourth Text REtrieval Conference (TREC 2015) Proceedings.

[7] Seung-Hyeon Jo , Jae-Wook Seol , Kyung-Soon Lee. CBNU at TREC 2015 Clinical Decision Support Track. The Twenty-Fourth Text REtrieval Conference (TREC 2015) , Nov 2015, Gaithersburg, United States. 2015, The Twenty-Fourth Text REtrieval Conference (TREC 2015) Proceedings.

[8] Kirk Roberts, Matthew S. Simpson, Ellen Voorhees, William R. Hersh. Overview of the TREC 2015 Clinical Decision Support Track. The Twenty-Fourth Text REtrieval Conference (TREC 2015) , Nov 2015, Gaithersburg, United States. 2015, The Twenty-Fourth Text REtrieval Conference (TREC 2015) Proceedings.

[9] Wikipedia contributors. Relevance feedback [Internet]. Wikipedia, The Free Encyclopedia; 2016 Jul 1, 15:53 UTC [cited 2017 Jan 28]. Available from: https://en.wikipedia.org/w/index.php?title=Relevance_feedback&oldid=727842398.

[10] MeSH. NCBI. https://www.ncbi.nlm.nih.gov/mesh[DB/OL]

[11] Wikipedia contributors. Stop words [Internet]. Wikipedia, The Free Encyclopedia; 2016 Mar 8, 22:14 UTC [cited 2016 May 11]. Available from: https://en.wikipedia.org/w/index.php?title=Stop_words&oldid=709047952.

[12] Wikipedia contributors. Stemming [Internet]. Wikipedia, The Free Encyclopedia; 2016 May 9, 04:48 UTC [cited 2016 May 11]. Available from: https://en.wikipedia.org/w/index.php?title=Stemming&oldid=719349770.

[13] Martin Porter. The Porter Stemming Algorithm[DB/OL]. http://tartarus.org/~martin/PorterStemmer/

[14] School of Computing Science, University of Glasgow, Terrier Team. Models[DB/OL]. http://terrier.org/docs/v4.1/javadoc/org/terrier/matching/models/package-summary.html

[15] G. Amati. Probabilistic Models for Information Retrieval based on Divergence from Randomness. PhD Thesis, School of Computing Science, University of Glasgow, 2003.

[16] S.P. Harter. A probabilistic approach to automatic keyword indexing. PhD thesis, Graduate Library, The University of Chicago, 1974.

[17] Kirk Roberts, Matthew S. Simpson, Ellen Voorhees, William R. Hersh. Overview of the TREC 2016 Clinical Decision Support Track. The Twenty-Five Text REtrieval Conference (TREC 2016) , Nov 2016, Gaithersburg, United States. 2016, The Twenty-Five Text REtrieval Conference (TREC 2016) Proceedings.