# HIT-WI at TREC 2015 Clinical Decision Support Track

Jingchi Jiang[1], Yi Guan[1*], Jia Su[1], Chao Zhao[1], Jinfeng Yang[2]

[1]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
[2]School of Computer Science and Technology, Harbin University of Science and Technology, Harbin, 150080, China

jiangjingchi0118@163.com, guanyi@hit.edu.cn, sjd163mail@163.com, hitsa.zc@gmail.com, fondofbeyond@163.com

**Abstract.** The TREC 2015 Clinical Decision Support track is composed of two subtasks, task A and task B. Similar to 2014 [1], the participants need to answer 30 clinical questions from patient cases for each task. According to the three types of clinical question: diagnosis, test and treatment, these tasks are to retrieve relevant literatures for helping clinicians to make clinical decision.

This paper describes how the clinical decision support system is developed for completing the task A and B by the HIT-WI group. For the automatic runs, some classical retrieval strategies are adopted, including query extraction, query expansion and the process of retrieval. Moreover, we propose two novel re-ranking methods: the one uses SVM model with 10-dimensional feature to re-rank the retrieved list, and the other is based on word co-occurrence network.

The 178 runs are submitted from 36 different groups. Our evaluation results show that 1) The Indri performs better than Lucene's for artificially-constructed queries. 2) Compare to the basic retrieval method, two re-ranking methods show the effectiveness in some topics. 3) Our results are higher than the median scores in most topics of task B. Furthermore, the system achieves the best scores for topics: #11 and #12.

## 1 Introduction

As a hot spot of academic frontier, Clinical decision support (CDS) provides clinicians and health professionals with knowledge and personalized information at appropriate times, to enhance the health level of patients. In making clinical decisions, clinicians often review the medical literature to further ensure the reliability for diagnosis and treatment. Medical literature can answer the three most common generic clinical questions faced by clinicians everyday [2]:"what is the patient's diagnosis?", "what tests should the patient receive?", "how should the patient be treated?". However, the problem of retrieving the relevant literatures can be time-consuming and difficult under the circumstance of massive literatures.

Similar to the goal of 2014, the TREC 2015 Clinical Decision Support (CDS) track is designed to retrieve relevant medical articles for answering generic clinical questions, according to actual patient records [3]. A patient record typically describes a

* Corresponding author. Tel:+86 18686748550. E-mail: guanyi@hit.edu.cn

challenging medical case, and mainly contains two sections: description which describes patients' condition in detail and summary, which synthesizes meaningful information from description based on the experience of doctors. The corpus for the retrieval task is the Open Access Subset of PubMed Central (PMC) on January 21, 2014, which contains a total of 733,138 articles [4].

In this paper, traditional retrieval techniques are adopted [5], including medical terms extraction, query expansion and literature retrieval. Then, we propose two re-ranking methods to enhance the relevance of retrieved results.

The rest of this paper is arranged as follows. In Sec. 2, we discuss the materials and methods in detail, and also focus on the construction of re-ranking models. Moreover, we conduct the experiments to testify the effectiveness of clinical decision system in Sec. 3. In Sec. 4, we conclude this paper and discuss the directions for further work.

## 2 Methods

### 2.1 Query construction

The query construction consists of query extraction, query expansion and query set generation. In the process of query's auto-construction, Metamap (a tool to map bio-medical text to the UMLS Metathesaurus) is used for extracting the medical concepts from the summary section of patient records. In addition, some rules are established, according to whether the concept's semantic type belongs to what we summarize, such as Neoplastic Process, Sign or Symptom et al. Then we regard these filtered medical concepts as the basic query set.

However, the basic queries which are only extracted from the given patient record, cannot completely retrieve relevant literatures for answering the clinical questions. Therefore, we adopt the UMLS Metathesaurus to expand the concepts. In the process of expanding, we avoid the same words presented in query as much as possible and add the type words (diagnosis, test and treatment) for improving the accuracy.

After a series of steps, the query sets are generated automatically in a different formats, to fit the different search engines.

### 2.2 The process of retrieval

The PubMed Central articles are published in the form of XML, one file per article. Therefore, an XML parser is employed to extract PMC ID, keyword, title, abstract, body and reference from each article.

In order to compare the retrieval performance of search engine, we adopt two kinds of toolkits: Intri and Apache Lucene, respectively. The former provides state-of-the-art text search and a rich structured query language. The latter is based on language model approach with Jelinek-Mercer smoothing for retrieving articles.

We start to retrieve the relevant literatures, including query extraction and expansion, building index. Each participant can only submit 1000 literatures at most for each topic. Therefore, we select the top 1000 literatures as the final result, which is ranked as the given score by the search engine.

## 2.3 Re-ranking model

### 2.3.1 Re-ranking based on machine learning

According to the relevant results of TREC 2014, it becomes possible for us to use machine learning method to re-rank the retrieved list. To judge whether a literature is relevant to the clinical decision, we think empirically that the appearing position of a query is a significant feature. Because, it is reasonable that the query appears in title of literature is more important than the same query appears in body. In addition, the position of type words, such as diagnosis, test or treatment, is also a strong feature to judge the relevance.

Due to a literature contains five fields, including title, abstract, keywords, body and reference, we extract the query feature and the type word feature from each fields, respectively. Therefore, a total of 10 features will be extracted for a certain literature.

We construct queries from TREC 2014 topics and obtain the retrieved results. Every retrieved literature is labeled as 0, 1 and 2, which represent the non-relevance, possible relevance and completely relevance. However, considering the quantity of relevant literatures is far less than the quantity of irrelevant ones, we regard the label 1 and 2 as relevant. Using the SVM classifier with a linear kernel to classify relevant literatures from irrelevant ones, the SVM model is trained. Then this model is applied to retrieved results of TREC 2015. Every result would be labeled either 1 if relevant, or 0 if not relevant. Adding this score with a 0.25 gain to the original indri score, we obtain the new score, which is used for our re-ranking.

### 2.3.2 Re-ranking based on co-occurrence network

In order to improve the performance of relevance ranking, we propose a novel method to re-rank the retrieved results. The idea of this method is based on co-occurrence words. We build a co-occurrence network to mine the potential literatures. For improving the recall rate, the re-ranking formula is constructed based on some network features.

#### 2.3.2.1 The construction of co-occurrence network

In the process of analysis for the TREC 2014, we find that these relevant literatures have a lot of co-occurrence words. We assume that these co-occurrence words can reveal the relevance of literature. In order to validate this assumption, an intuitive co-occurrence network based on 1000 retrieved literatures is needed. Firstly, we empirically extract the co-occurrence words from literature by the top level of MeSH hierarchy [6]:

- Diagnosis: B03, B04, C
- Test: E01
- Treatment: D02, D04, D06, D26, D27, E02, E04

When a common medical word from MeSH appears on two literatures, an edge will be created to connect them. Moreover, the edge weight gradually grow, along with the number of common words increasing. After 1000 retrieved literatures are iterated, a co-occurrence network is built, which is composed of the literature as the node and the co-occurrence medical word as the edge. The topology of the co-occurrence network is shown in Fig. 1.

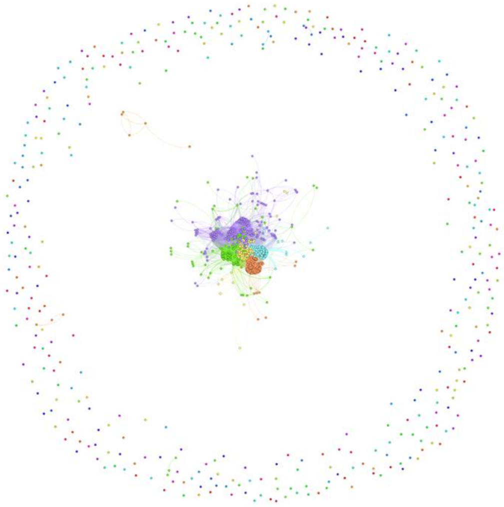

Fig. 1 The topology of co-occurrence network

As shown in Fig.1, the network consists of several communities in different color. The features of the co-occurrence network include the followings: 1. The literature nodes within the same community are strongly attached to each other. 2. Instead, the nodes from different communities represent a "weaker" relation. Through observing and analyzing the judgment file of TREC 2014, we find that the most of relevant literature locate the inside of community, while the discrete nodes always play the non-relevance or low relevance roles for each topic. Furthermore, we can summarize that every community have dissimilar emphases for the given patient record. The subject of some communities are appropriate to answer the clinical question, while the literatures from the other communities are irrelevant. Therefore, how to choose the appropriate communities is an important work.

#### 2.3.2.2 Mining potential literatures

Because the automatic extraction and expansion have its limitations and uncertainties, lead to the incomplete and non-credibility of query set. Therefore, some relevant literatures might be missed except 1000 retrieved literatures. To solve this problem, we propose a method based on the co-occurrence network, to mining potential relevant literatures from the rest of the corpus. This method uses the indicator of clustering coefficient to determine whether a literature is associated with the topic.

**Node coefficient** is defined as the proportion of connections among its neighbors which are actually realized compared with the number of all possible connections. The parameter *k* is defined as the number of the common terms from MeSH between literature *i* and community *ζ*. *T*(*i*) represents the number of all possible connections among the *k* vertices.

$$T(i) = k(k-1)/2 \tag{1}$$

*E*(*i*) represents the actual number of edges among the *k* vertices. *c*(*i*) is the clustering coefficient of node *i* and can be computed as follows.

$$c(i) = E(i)/T(i) \tag{2}$$

Community coefficient is defined as the mean of the entire node coefficient within the community. *c*(*ζ*) is defined as the clustering coefficient of community *ζ* :

$$c(\zeta) = \frac{\sum_{i-1}^{n} c(i)}{n} \tag{3}$$

If the node coefficient is greater than the community coefficient of a specific community, we can conclude that this node has similarity to this community, and put the node as a potential literature to the existing community. The bigger node coefficient means that the higher connectivity with the community. Along with the continuous increase of the potential literatures, some new MeSH terms will be found from the co-occurrence network, which can describe the topic better. After all of the literatures are traversed, a richer co-occurrence network is built.

#### 2.3.2.3 Re-ranking calculation

Based on the richer co-occurrence network, we need to re-rank all the nodes. To calculate the score of re-ranking, some measures should be introduced, including the measure of node importance and the medical terms density of community where the node locate.

Because the potential literatures are different than the retrieved literatures which have a relevance score by search engine. Therefore, a computational method for calculating the relevance score of potential literature is also proposed, which is defined as follows:

$$Score(i) = \begin{cases} Rscore(i) & i \in RetrievedSet \\ NC(i) * \sum_j Rscore(j)/n & i \in PotentialSet \end{cases} \tag{4}$$

*Rscore*(*i*) represents the relevance score of retrieved literature by search engine. *NC*(*i*) is the node coefficient of literature *i*. *j* is defined as a literature within retrieved set, while is connected to the literature *i*. n is the number of literature *j*.

The terms density of community is defined as the ratio of the number of terms to the number of relationships. In addition, we adopt the value of pagerank to regard as the measure of node importance. After the preparation of the theory, we propose the formula of re-ranking model:

$$ReRankScore(i) = \alpha \cdot CD_i \cdot PR(i) + \beta \cdot Score(i) \tag{5}$$

*CD*$_i$ represents the density of community where is the literature *i* location. *PR*(*i*) is the importance of literature *i*. *α* and *β* are both the weight parameters for regulating the factor proportion between the co-occurrence network and the search engine.

## 3 Experiments

### 3.1 Clinical decision support system design

Our clinical decision support system consists of four main modules.

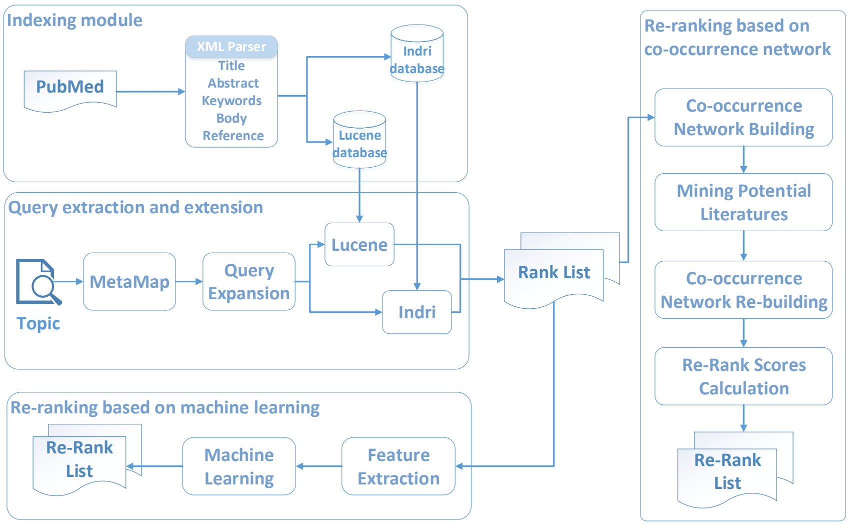
Indexing module
PubMed
XML Parser
Title
Abstract
Keywords
Body
Reference
Indri database
Lucene database
Query extraction and extension
Topic
MetaMap
Query Expansion
Lucene
Indri
Rank List
Re-ranking based on machine learning
Re-Rank List
Machine Learning
Feature Extraction
Re-ranking based on co-occurrence network
Co-occurrence Network Building
Mining Potential Literatures
Co-occurrence Network Re-building
Re-Rank Scores Calculation
Re-Rank List

**Fig. 1.** The flow diagram of the clinical decision support system.

### 3.2 Comparing Indri with Lucene

In the TREC 2015 Clinical Decision Support track, the task consists of two parts: the task A and task B which adds the “diagnosis” section from the last twenty topics. For the task A, we submit the retrieved results including artificial Indri result, automatic Indri result and automatic Indri result with the re-ranking model based on machine learning. Similar with task A, the artificial list of Lucene, automatic list of Indri and the automatic result with the co-occurrence network are submitted for the task B.

Because the topics of diagnosis type have exactly the same contents and structures in task A and B. So we can compare the result of artificial Indri and artificial Lucene based on the same query set. Figure 2 shows the difference between Indri and Lucene using four different measurement indicators. We can see that the former outperforms the latter in most of topic. It follows that the search engine of Indri is more effective than Lucene for the retrieval task.

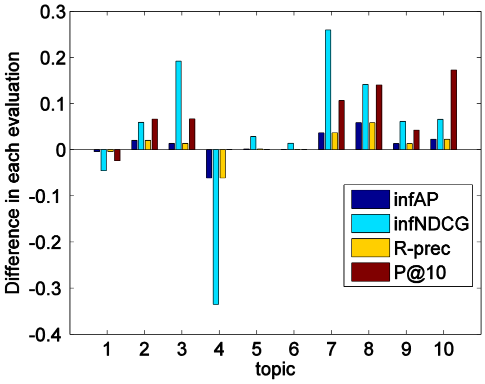

**Fig. 2.** Comparing Indri with Lucene.

### 3.3 Comparing submitted runs to each other

In order to testify the effectiveness of our methods, we compare the infAP and infNDCG of each method. For the task A as shown as Figure. 3, the artificial results are much higher than other automatic results. It is also find that the re-ranking model based on machine learning has less effective than the expectations.

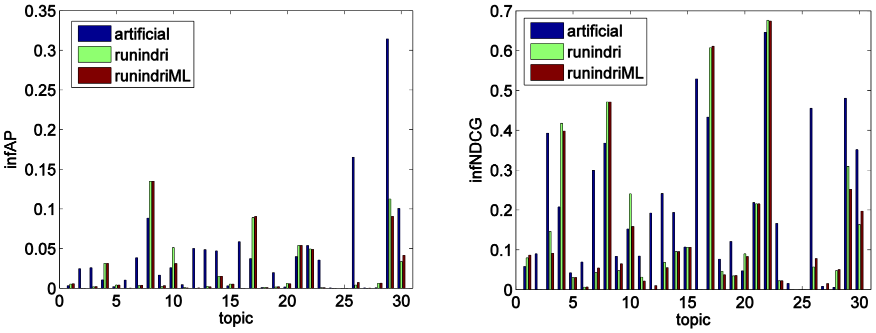

**Fig. 3.** Retrieval results for task A in two different indicators.

From the statistical results of Figure. 4, the re-ranking model based on co-occurrence network does not perform well enough. The performance of most topics is not improved except a small rise in the topic 9 and 27. For the possible reasons of unsatisfactory result, we analyze that it could be caused by the weight parameters of $\alpha$ and $\beta$, which are not adjusted to the optimal values.

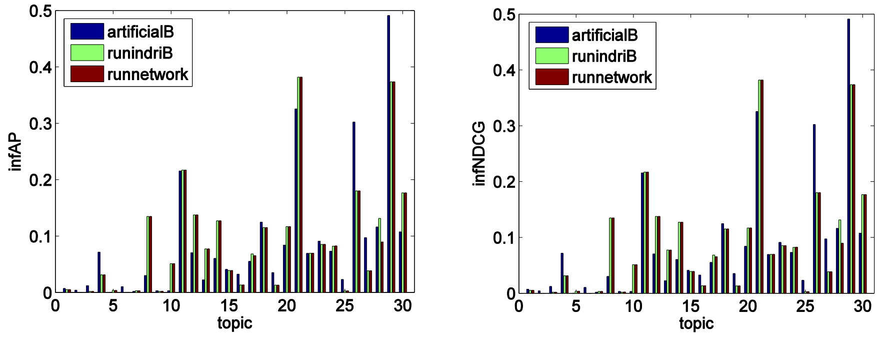

**Fig. 4.** Retrieval results for task B in two different indicators.

### 3.4 Comparing submitted runs to the median

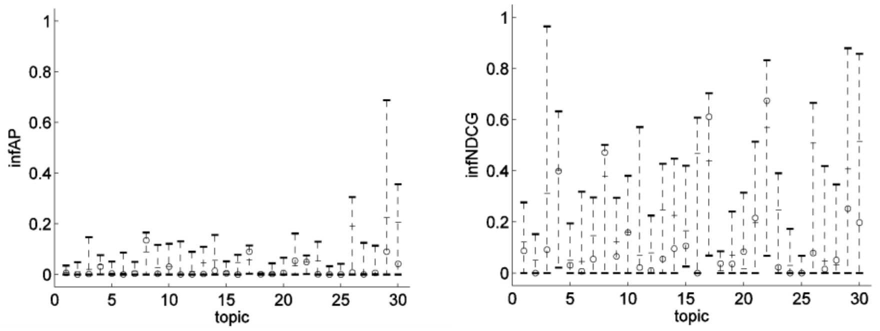

**Fig. 5.** Comparing the automatic runs based on SVM model with other participants.

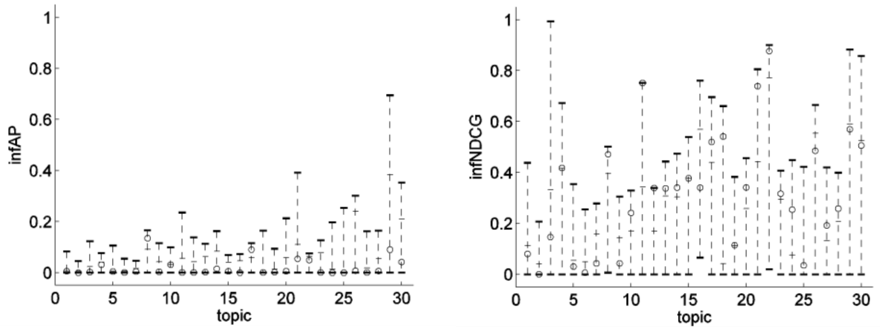

**Fig. 6.** Comparing the automatic runs based on co-occurrence network with other participants.

Finally, a set of experimental results is given. The automatic runs with re-ranking model based on SVM are below the median scores for the most topics, as shown in Figure. 5. From the Figure. 6, we can see that our re-ranking model based on co-occurrence network achieve the best score in two topics: #11 and #12. The model also performs much better than the median scores for the other topics. These results further testify the effectiveness of our clinical decision support system.

## 4 Conclusion

This paper described the clinical decision support task in the TREC 2015. To complete the task, a clinical decision support system based on literatures is designed and developed by the HIT-WI group. On the basis of traditional retrieval techniques, we propose two novel re-ranking methods to improve the retrieval results. The two methods use the models of the machine learning and the network. Moreover, the analysis of the experimental result demonstrates the effectiveness of our system. Our future work will focus on optimizing the re-ranking model and cutting down time consumption in the process of retrieval.

**Acknowledgments.** The Open Access Subset of PubMed Central used in this paper were provided by TREC 2015 Clinical Decision Support (CDS) track, and thanks to the organizing committee of TREC and the evaluators of the results.

## References

1. Simpson M S, Voorhees E, Hersh W. Overview of the TREC 2014 Clinical Decision Support Track[C]//Proc. 23rd Text Retrieval Conference (TREC 2014). National Institute of Standards and Technology (NIST). 2014.
2. Hasan S A, Zhu X, Dong Y, et al. A Hybrid Approach to Clinical Question Answering[J].
3. Gobeillab J, Gaudinata A, Paschec E, et al. Full-texts representation with Medical Subject Headings, and co-citations network rerank-ing strategies for TREC 2014 Clinical Decision Support Track[J].
4. Wei Y, Hsu C C, Thomas A, et al. Atigeo at TREC 2014 Clinical Decision Support Task[J].
5. Garcia-Gathrighta J I, Menga F, Hsua W. UCLA at TREC 2014 Clinical Decision Support Track: Exploring Language Models, Query Expansion, and Boosting[J].
6. Mourao A, Martins F, Magalhaes J. NovaSearch at TREC 2014 Clinical Decision Support Track[J].