Physica A 459 (2016) 42–54

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Clinical-decision support based on medical literature: A complex network approach

Jingchi Jiang [a], Jichuan Zheng [b], Chao Zhao [a], Jia Su [a], Yi Guan [a,*], Qiubin Yu [c]

[a] School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China
[b] RICOH Software Research Center (Beijing) CO., LTD, Beijing 100044, China
[c] Medical Record Room, The 2nd Affiliated Hospital of Harbin Medical University, Harbin 150086, China

HIGHLIGHTS

- We constructed a medical literature network (MLN) based on retrieved literature.
- The MLN improves the relevance retrieval result for clinical-decision support.
- We also proposed a re-ranking model to sort all retrieved literature by relevance.
- Our clinical-decision method based on the MLN yields higher scores in TREC 2015.
- Our study results confirmed that the MLN can facilitate the investigation of CDS.

ARTICLE INFO

Article history:
Received 28 December 2015
Received in revised form 7 April 2016
Available online 28 April 2016

Keywords:
Small-world
Scale-free
Complex network
Medical literature network
Clinical decision support

ABSTRACT

In making clinical decisions, clinicians often review medical literature to ensure the reliability of diagnosis, test, and treatment because the medical literature can answer clinical questions and assist clinicians making clinical decisions. Therefore, finding the appropriate literature is a critical problem for clinical-decision support (CDS). First, the present study employs search engines to retrieve relevant literature about patient records. However, the result of the traditional method is usually unsatisfactory. To improve the relevance of the retrieval result, a medical literature network (MLN) based on these retrieved papers is constructed. Then, we show that this MLN has small-world and scale-free properties of a complex network. According to the structural characteristics of the MLN, we adopt two methods to further identify the potential relevant literature in addition to the retrieved literature. By integrating these potential papers into the MLN, a more comprehensive MLN is built to answer the question of actual patient records. Furthermore, we propose a re-ranking model to sort all papers by relevance. We experimentally find that the re-ranking model can improve the normalized discounted cumulative gain of the results. As participants of the Text Retrieval Conference 2015, our clinical-decision method based on the MLN also yields higher scores than the medians in most topics and achieves the best scores for topics: #11 and #12. These research results indicate that our study can be used to effectively assist clinicians in making clinical decisions, and the MLN can facilitate the investigation of CDS.

* Correspondence to: School of Computer Science and Technology, Harbin Institute of Technology, Comprehensive Building 803, China. Tel.: +86 186 8674 8550.
E-mail address: guanyi@hit.edu.cn (Y. Guan).

http://dx.doi.org/10.1016/j.physa.2016.04.026
0378-4371/

1. Introduction

With the advancing age of populations worldwide, people have paid more attention to health problems. Each year, the numbers of deaths caused by cardiovascular diseases (CVDs) and hypertension are estimated to be 17.5 million and 7.1 million, respectively [1]. A World Health Organization (WHO) report on CVDs shows that 80% of all CVD deaths are due to heart attacks and strokes, which represent approximately 31% of all deaths globally [2]. Moreover, yearly figures from the WHO revealed that a person dies of diabetes and diabetic complications about every 10s. Diabetes is directly responsible for 1.5 million deaths in 2012 and 89 million disability adjusted life years [3]. The population of diabetic adults is expected to reach 300 million by 2025 [4]. Finding an effective auxiliary means to assist clinicians making a more correct clinical decision has become a critical problem to upgrade the clinician knowledge and reduce mortality.

Clinical-decision support (CDS) [5–9] which is a health information technology, provides physicians, patients, and other health professionals with knowledge and person-specific information, that is intelligently filtered and retrieved at appropriate times, to enhance patient health and health care [10]. In making clinical decisions, clinicians often seek out medical literature on how to best care for their patients. Medical literature can answer the three most common generic clinical questions faced by clinicians on a daily basis: "What is the patient's diagnosis?", "What tests should the patient receive?" and "How should the patient be treated?". However, given the volume of existing literature and the rapid pace at which new research is published, locating the most relevant and timely information for a particular clinical need can be a daunting and time-consuming task [11].

The Text Retrieval Conference (TREC) 2015 CDS track [12], which is similar to the goal of TREC 2014 [13–15], is designed to retrieve relevant medical literature to answer generic clinical questions on 30 actual patient records. The patient record typically describes three types of challenging medical cases and consists of 10 records per type, including diagnosis, test and treatment. Each record mainly contains two sections: description (detailing the patient condition) and summary (extracting meaningful information from the 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 literature [11]. According to the summary of a patient record, its description, or both, participants are challenged to retrieve a ranked set of 1000 papers at most, which are likely to support the decision of a physician on appropriate patient care.

The important aspects of the CDS according to medical literature have been discussed, and many valuable research ideas have been proposed. Garcia-Gathrighta et al. [16] adopted the vector space model using term frequency-inverse document frequency similarity and a unigram language model with Jelinek–Mercer smoothing. Mourao et al. [17] proposed multiple information retrieval techniques: retrieval functions, re-ranking, query expansion and classification of medical articles. Xu et al. [18] demonstrated the efficiency of the Johns Hopkins University HAIRCUT retrieval engine using character n-grams as an indexing term. Choi et al. [19] proposed an external tagged knowledge-based query expansion method for relevance ranking. Moreover, a machine-learning classifier-based text categorization method was used for the task-specific ranking.

Although many experts and scholars have explored the issues of clinical decision in different perspectives and realized a series of achievements, the use of complex network approach to solve the clinical decision problem has not yet been studied. In the present study, we focus on helping clinicians make better clinical decisions by retrieving relevant medical literature. In summary, we conducted our investigation in the following manner:

(1) We proposed a method of building a medical literature network (MLN). Then, we further analyzed the topological structure and characteristics of the MLN, which refers to the features of a complex network.
(2) According to the MLN and some analytical methods of complex networks, we adopted two strategies to mine potential literature, which can also assist clinicians making clinical decisions in addition to the basic literature retrieval.
(3) Combining the relevance factor of a search engine with the structural factor of the MLN, we further proposed a re-ranking model to sort all retrieved literature.
(4) From the comparison with those of other participants in TREC 2015, we numerically found that our approach can better improve the normalized discounted cumulative gain (NDCG) indicator than the median scores in most topics.

The rest of this paper is organized as follows: in Section 2, we introduce the structural characteristics of a complex network and the MLN. In Section 3, the potential-literature-mining algorithm and re-ranking model are proposed on the basis of a complex network approach. In Section 4, we further evaluate the validity of the potential-literature-mining and the accuracy of the re-ranking model. Finally, we conclude this paper and discuss directions for future work in Section 5.

2. Construction of the MLN

2.1. Process of relevant literature retrieval

To search some indicative literature to help clinicians make clinical decisions, we first need to retrieve some relevant literature using a search engine. Some classical retrieval techniques are adopted, including index building, query construction and literature retrieval.

The corpus from the PMC is given as a set of XML files. Therefore, an XML parser is employed to extract the PMC IDs, keywords, titles, abstracts, full texts and references. If an abstract is not available, the conclusion section will be used as a

substitute for the abstract. On the basis of the above-described work, the index files are created using the search engine called Indri.

The query construction consists of query extraction, query expansion and query set generation. In the process of query auto-construction, MetaMap (a tool to map biomedical text to the Unified Medical Language System (UMLS) Metathesaurus) is used to extract the medical concepts from the summary section of a patient record. 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 exactly describe the topic of each record. To compensate for the insufficiency of the basic queries, we further adopt the UMLS Metathesaurus to expand these concepts. After a series of steps, the query sets are generated in a format that conforms to Indri.

The organizers of TREC 2015 allow at most 1000 retrieved literature to be submitted for each patient record. We select the top 1000 papers as the result, which are ranked as the default score by the search engine.

2.2. Characteristics of the complex network

A complex network is composed of a large number of nodes and their intricate relationships. Complex network theories [20–23] have been employed to analyze many aspects of natural language processing, including language translation of a semantic network [24], investigation of knowledge graph [25], and construction of a literature citation network [26]. These complex networks frequently possess small-world [27] and scale-free [28] features. Small-world networks have a large clustering coefficient and a small average path length, which means that strong connectivity and high correlation exist between the nodes in the complex network. Scale-free networks are characterized by a power-law decay of the degree distribution. When some nodes fail, the topology of a complex network is more likely to be influenced than a random network [29].

According to the characteristics of a complex network, Tachimori et al. [30] constructed a hospital network and a medical knowledge network. An experimental study proved that the structure of a clinical practice may emerge from the mutual influence of medical knowledge and clinical practice. Therefore, we assume that an MLN can help clinicians in clinical decisions.

2.3. MLN

According to the analysis of the results of TREC 2014, we find that words often simultaneously occur in the annotated relevant literature. We believe that these simultaneously occurring words can reveal the relevance of the literature. To validate this assumption, we build an intuitive co-occurrence network based on 1000 retrieved literature. From the title, abstract, keyword and reference to literature, we empirically extract the co-occurring words using the top level of the MeSH hierarchy [17] as follows:

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

When a common medical word from the above list appears on two papers, an edge will be created to connect them. Moreover, the edge weight gradually grows, along with increase in the number of common words. After the 1000 retrieved literature are iterated, we build an MLN based on the co-occurring words. This MLN is composed of the literature as the node and the co-occurring words as the edge. The topology of this network is shown in Fig. 1.

Furthermore, we use Gephi [31], which is an open-source graph visualization and manipulation software, to visualize the MLN. Using the modularity-class [32] algorithm to detect communities, the MLN is divided into several communities marking different colors. The community structure plays a vital role in mining the relevant literature. The features of this network include the following: (1) the literature nodes within the same community are strongly attached to one another [33,34], and (2) the nodes from different communities represent a "weaker" relationship [35].

2.4. Network analysis

To analyze the MLN structure, we present the degree and the graph distance distributions. First, we randomly select three different types of patient records from 30 medical cases. Because the number of nodes is fixed at 1000, the difference between these MLNs lies in the number of edges. In the diagnosis, test and treatment MLNs, the numbers of edges are 25,601, 42,413 and 61,113, respectively. Second, we show the degree distribution of the three MLNs in log–log plot. To some extent, the degree distribution of the three MLNs generally follows the truncated power-law distribution [36] in Fig. 2. Thus, the scale-free property of the MLN is proved.

Meanwhile, we show the graph distance distribution in Fig. 3, including the betweenness centrality, closeness centrality and eccentricity distributions.

Furthermore, the average path length and the average clustering coefficient for the three MLNs are calculated. Table 1 lists the diagnosis, test and treatment MLNs. C_{mln} and C_{random} are defined as the average clustering coefficients of the MLN and the corresponding random network, respectively. Similarly, L_{mln} is defined as the average path length. Because $L_{mln}/L_{random} \sim 1$

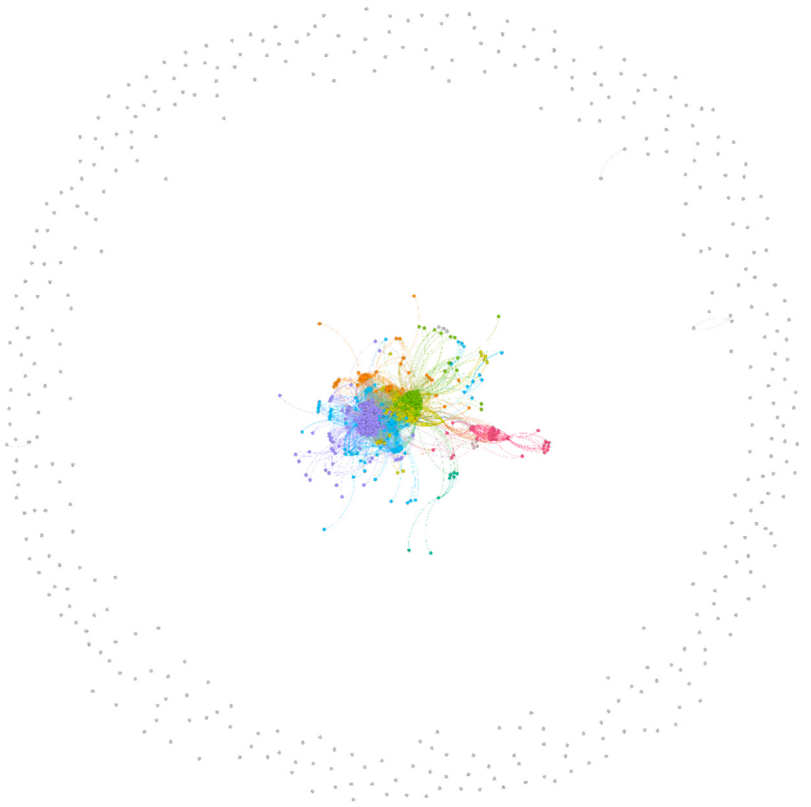

Fig. 1. The topology of MLN. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

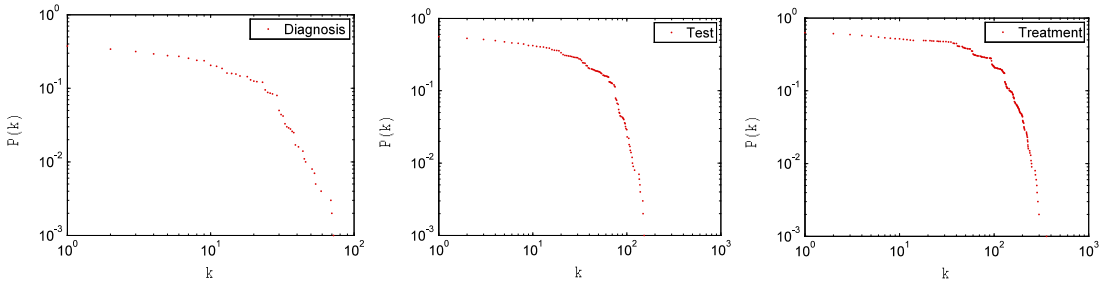

Fig. 2. Degree distribution of the diagnosis, test and treatment MLNs.

Table 1
Comparison of the MLN statistics with the corresponding random network.

Network	Nodes	Edges	C_{mln}	C_{random}	L_{mln}	L_{random}
$MLN_{diagnosis}$	1000	25,601	0.879	4.99×10^{-3}	2.29	2.02
MLN_{test}	1000	42,413	0.887	8.14×10^{-3}	2.17	1.92
$MLN_{treatment}$	1000	61,113	0.885	9.36×10^{-3}	1.97	1.89

and $C_{mln} \gg C_{random}$, the MLNs have almost the same average path length and a far larger average clustering coefficient compared with the corresponding random network. According to the small-world characteristics, the MLNs have small-world features.

In conclusion, these data demonstrate that the MLN possesses complex network properties: scale free and small world. Therefore, we can regard the MLN as a complex network.

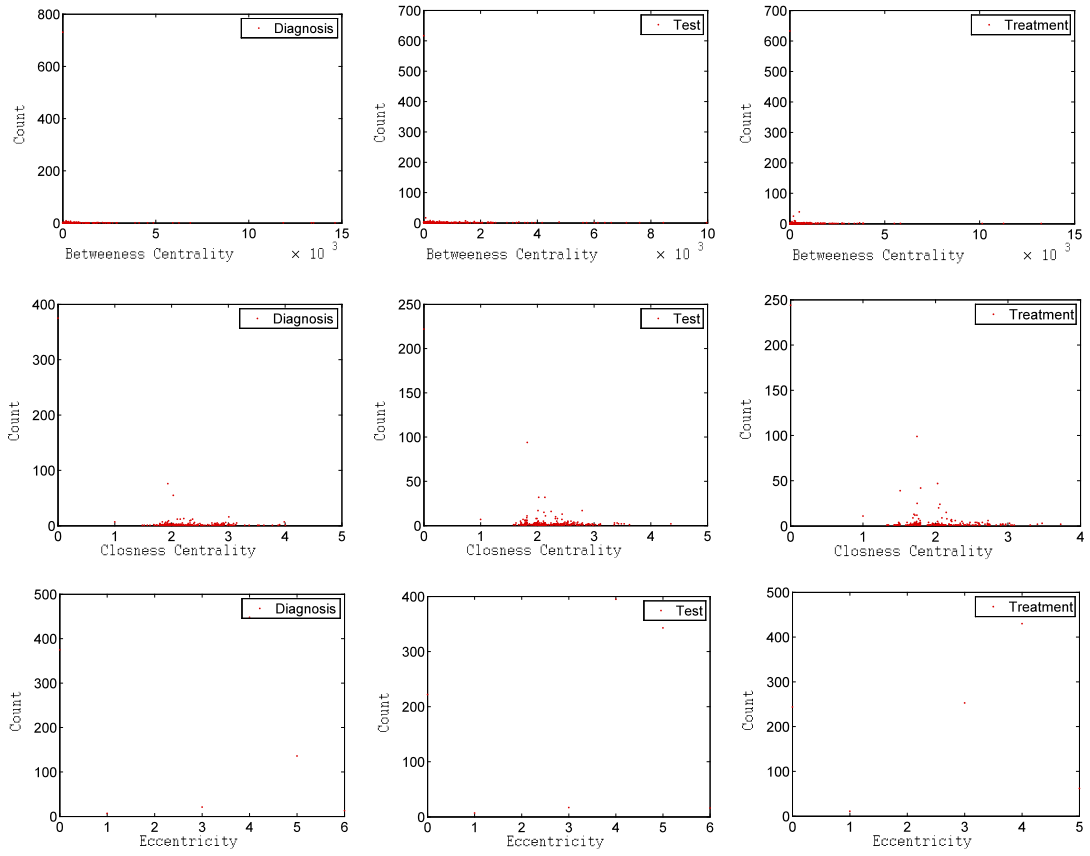

Fig. 3. Graph distance distribution of the diagnosis, tests and treatment MLNs.

3. Methodology

3.1. Mining of potential literature

Automatic extraction and expansion suffer from limitations and uncertainties, which lead to incomplete and non-credibility of the query set. Therefore, some relevant papers, which can also answer the clinical question, might be missed except for the 1000 retrieved literature. To solve this problem, we propose two different methods based on the MLN, to mine potential relevant literature from the rest of the corpus. The first method employs the characteristic of clustering coefficient [37–39] to identify whether a node belongs to potential relevant literature. The other method calculates the connectivity [40] between the specific community and node.

3.1.1. Mining of potential literature based on clustering coefficient

In some fields of a complex network, the related research of mining of potential nodes has been profoundly studied. In social networks, the node-mining technique is always applied in the community-detection algorithm and important-node perception algorithms.

In the decision file of TREC 2014, the relevant papers are labeled by medical librarians and physicians. Then, we find that most of the relevant papers are located inside the community in the MLN, whereas the discrete nodes always play the non or minimally relevant roles. Furthermore, we determine that every community has dissimilar emphases for a given patient record. The co-occurring medical words from a community are more suitable as a topic of the patient record, and the literature within this community is more relevant.

According to the above-mentioned analysis, we propose a node-mining method to identify the potential relevant literature on the basis of a clustering coefficient. Let $G = (V, E)$ denote an MLN, where $V = \{v_1, v_2, \ldots, v_n\}$ is a set of n relevant papers, and $E \subseteq V \times V$ is a set of undirected co-occurring word relationships. The clustering coefficient can be divided into two categories: local and network-average clustering coefficients. The local clustering coefficient is defined as the proportion of connections among its neighbors, which are actually realized compared with the number of all possible connections. The local clustering coefficient of v_i in an undirected network is defined as follows:

$$L(v_i) = \{v_j : e_{ij} \in E\} \tag{1}$$

$$C(v_i) = \frac{2\left|\left\{e_{jk} : v_j, v_k \in L(v_i), e_{jk} \in E\right\}\right|}{k_i(k_i - 1)}. \tag{2}$$

The neighborhood $L(v_i)$ of node v_i represents its immediately connected neighbors. Parameter k_i is defined as the number of connections between node v_i and community ζ. The network average clustering coefficient is defined as the mean of the entire node coefficient within the community. $C(\zeta)$ represents the average clustering coefficient of community ζ:

$$C(\zeta) = \frac{\sum_{i=1}^{n} C(v_i)}{n}. \tag{3}$$

If the local clustering coefficient of potential node v_i is greater than the average clustering coefficient of specific community ζ, we can conclude that node v_i has high similarity to some nodes within ζ. Then, we locate node v_i into ζ as a potential relevant literature. Along with the continuous increase in the potential literature, some new MeSH terms will be obtained from the MLN that can precisely describe the topic. After all of the papers are traversed, a more comprehensive MLN $G' = (V', E')$ is built.

3.1.2. Mining of potential literature based on connectivity

In MLN, not all communities are suitable for the topic of a patient record. Choosing the relevant communities and identifying the potential nodes using the structure of these communities are very important. Considering the importance of the community structure in identifying potential nodes, we propose a node-mining method based on the community connectivity. First, we quantify the relevance of the community by calculating the MeSH lexical density, and choose a community with the highest lexical density.

$$\zeta = \underset{\varphi}{\operatorname{argmax}} \left\{ \frac{\left|\left\{term_\varphi : \varphi \subseteq G, term \in MeSH\right\}\right|}{\log\left|\{v_i : v_i \in \varphi\}\right|} \right\} \tag{4}$$

where $term_\varphi$ denotes a MeSH medical term in the community φ and ζ denotes the densest community. We assume that a community with high lexical density characteristics covers more healthcare areas and is likely to identify more relevant nodes. To avoid selecting some incorrectly small communities as ζ, we consider a logarithmic function to represent the community scale, and set $|v_i : v_i \in \varphi| > 1$. After the densest community is determined, we identify the potential nodes by centering on the structure of this community.

$$\frac{\left|\{v_m : v_m \in \zeta\}\right|}{\left|\left\{e_{ij} : v_i, v_j \in \zeta\right\}\right|} \leq \left|\left\{e_{jk} : v_j \in \zeta, v_k \notin \zeta, e_{jk} \in E\right\}\right|. \tag{5}$$

When the number of connections between v_k and ζ is greater than the community connectivity, we can indicate that node v_k is similar to the topic of community ζ. In conclusion, the two node-mining methods have all presented the importance of the network structure of the MLN. The effectiveness of these methods is verified through a series of experiments.

3.2. Re-ranking model

In this section, we will discuss how to re-rank the relevant literature on the basis of the MLN. After identifying the potential literature, the list of relevant literature contains more than 1000 papers. Therefore, we should uniformly re-rank the retrieved and potential literature and select the top 1000 papers to answer the clinical question.

In the process of determining the relevance of the literature, the average clustering coefficient of a community and the importance of a node are considered as a structural factor in the re-ranking model. The calculation of the average clustering coefficient of a community is similar to that of Eq. (3). In a complex network, the measures of the node importance are varied, such as PageRank, betweenness centrality and degree. In this study, we use these measures as the node importance and analyze the effectiveness of each measure in the experimental part.

In addition to the structural factor, relevance factor is also considered in the re-ranking model. The retrieved literature by the search engine has a default sort order with a relevance score. However, the absence of the relevance score for the potential literature makes it impossible to calculate the re-rank score in a uniform formula. Therefore, a computational method for calculating the relevance score of the potential literature is proposed, which is defined as follows:

$$Score(v_i) = \begin{cases} Score_{se}(v_i) & v_i \in \textit{Retrieved Set} \\ \dfrac{C_\zeta(v_i) \cdot \sum_j Score_{se}(v_j)}{n} & v_i \in \textit{Potential Set},\ v_j \in \zeta,\ e_{ij} \in E \end{cases} \tag{6}$$

$Score_{se}(v_i)$ represents the default relevance score of retrieved literature v_i, determined by a search engine. $C_\zeta(v_i)$ is the local clustering coefficient between node v_i and community ζ. $\sum_j Score_{se}(v_j)$ denotes that the sum of the relevance scores of all the retrieved nodes connected to v_i. Because $Score_{se}(v) \in [0, 1]$ and $C_\zeta(v) \in [0, 1]$, we can deduce that $Score(v) \in [0, 1]$.

Following the previous ideas of the re-ranking model, we combine the structural factor with the relevance factor to calculate the relevance. Let $ReRank(v_i, G) \in [0, 1]$ be the relevance of node v_i, which can be calculated as follows:

$$ReRank(v_i, G') = \alpha \cdot C(loc(v_i)) \cdot I(v_i) + \beta \cdot Score(v_i) \tag{7}$$

where α and β are two tunable parameters. $loc(v_i)$ denotes the community where node v_i is located. The functions C and I represent the average clustering coefficients of the community and the importance of the node, respectively. To direct the value range of α and β, we suppose that G' exists such that $\alpha + \beta > 1$. We consider that only three nodes v_1, v_2 and v_3 in G', and every node contains all queries. Thus, we obtain $I(v_1) = I(v_2) = I(v_3) = 1$, $v_1, v_2, v_3 \in \zeta = G'$, $e_{12}, e_{13}, e_{23} \in E'$ and $Score(v) = 1$. In addition, we can prove that $C(G') = 1$ according to the fully connected network. Using Eq. (7), we obtain $ReRank(v_i, G') = \alpha \cdot C(loc(v_i)) \cdot I(v_i) + \beta \cdot Score(v_i) = \alpha + \beta > 1$. However, the result goes against $ReRank(v_1, G) \in [0, 1]$. Therefore, we have $\alpha + \beta \leq 1$. To avoid being trapped in the local optimum of the parameters, we choose a special case $\alpha + \beta = 1$, which can be transformed to $\beta = 1 - \alpha$. Using the method of potential-literature mining and re-ranking the literature, we design an algorithm to calculate the relevance of each literature in the MLN. The detailed algorithm is listed in Table 2.

According to the description of the mining of the potential-literature method, we must walk through the full corpus. To reduce the time complexity, the clinical-decision algorithm extracts the MeSH medical terms from the basic MLN and retrieves the c-top literature on the basis of each term by Indri. Then, these retrieved literature will displace the full corpus as the candidate of potential literature. The reasons are that the potential literature must be connected to the existing literature within the MLN and the edges are formed by the MeSH medical terms. Thus, the optimization algorithm only needs to traverse the occurring terms in the MLN. The algorithm results in time complexity $O(c \cdot k)$, where k is the number of MeSH terms.

4. Experiments and discussion

As participants to the TREC 2015 CDS track, we download 30 topics from the TREC official website [12], which consist of ten "diagnosis" topics, ten "test" topics and ten "treatment" topics. Meanwhile, the TREC official provides the same corpus with TREC 2014, which contains a total of 733,138 literature. To train our model, we regulate the values of some parameters and compare the effectiveness of two mining methods using the 30 topics of TREC 2014. After the optimal model is constructed, we use this model to complete the TREC 2015 CDS track. From the comparison with those of the other participants, our model achieves better result in TREC 2015. The three main modules of the model are shown as follows:

4.1. Mining method analysis

In this section, we focus on the effectiveness of mining of the potential literature. We adopt two methods, namely, based on clustering coefficient (CC-based) and based on connectivity (CO-based) mining methods, to analyze the corpus of TREC 2014. Fig. 4 shows the x-axis representing the serial numbers of topics from 1 to 30. The first 10 topics belong to the "diagnosis" type. The middle ten topics belong to the "test" type. The remaining ten topics belong to the "treatment" type. The y-axis represents the amount of relevant literature. By referring to the decision file of TREC 2014, we can determine which papers are relevant. We choose the number of mined potential literature as 50, 100, and 200 respectively. The experimental results of the CC-based and CO-based mining methods are shown in Fig. 5.

Fig. 5 shows the basic results representing the number of relevant literature, which were retrieved by a search engine. We can observe that the basic results are unstable. The highest result reaches 211 and the lowest is zero. Following the increase in the number of basic results, the number of potential relevant literature increases under the CC-based method. When the effectiveness of the basic results is maintained at a low level, the performance of the CC-based method is unsatisfactory. A similar situation is observed for the CO-based method. These results illustrate that the effect of the mining method depends on the established MLN. The reason for this is that when a irrelevant MLN is built, the potential literature which are identified based on the structure of the MLN, will inevitably result in an off topic. In addition, the sequence of the mining effect for different types of topic from high to low is from treatment to test and to diagnosis. Finally, we compare the effect of the two methods. Fig. 5 shows that the CC-based method is superior to the CO-based method.

To further verify whether the proportion of relevant paper position in all literature will decrease with the increase in the number of mined literature, we calculate the ratio of the different scales of mined literature. We choose the number of mined literature (NUM) as 50, 100, and 200 respectively. The results are shown in Fig. 6.

Under the CC-based method, the proportion of the relevant paper remains steady when NUM $=$ 50, 100 and 200. In particular, when NUM $=$ 200, the ratio is higher than the basic results in some topics, which illustrates that node addition by the CC-based method is effective. In contrast, we can observe that a greater NUM will result in a smaller relevant ratio under the CO-based method.

In conclusion, according to the MLN, the effectiveness of the CC-based method is better than that of the CO-based method. The number of relevant papers obviously increase under the CC-based method. In addition, the proportion of relevant papers remains steady even when NUM $=$ 200. Therefore, the CC-based method with NUM $=$ 200 is very significant for mining more relevant papers.

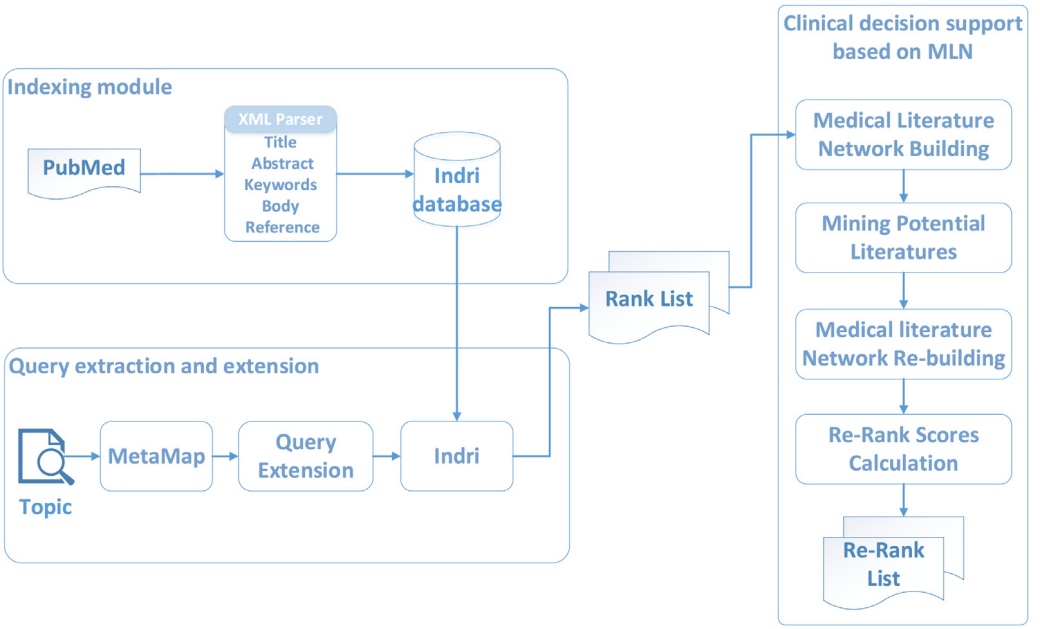

Fig. 4. Flow diagram of the CDS based on MLN.

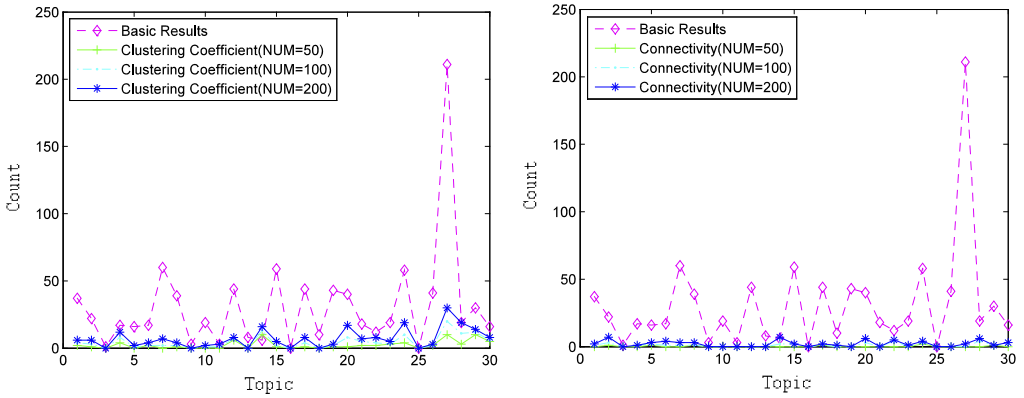

Fig. 5. Number of relevant literature under the CC-based and CO-based methods.

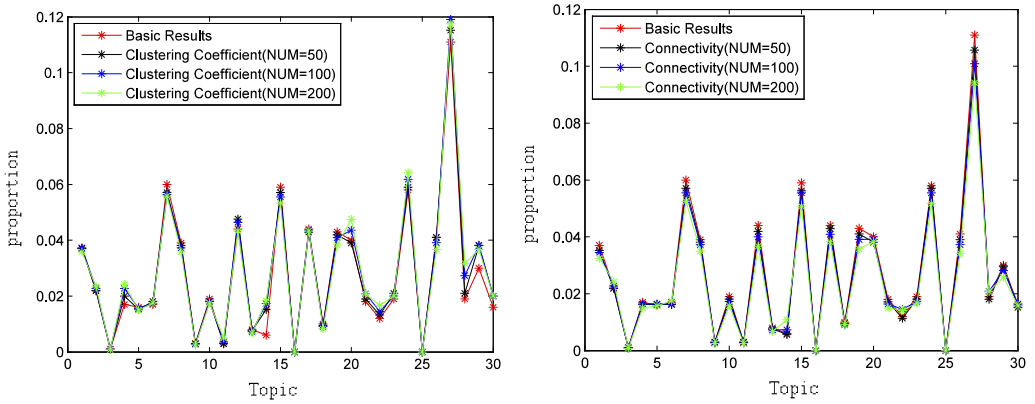

Fig. 6. Proportion of relevant literature under the CC-based and CO-based methods.

Table 2
CDS algorithm based on medical literature.

Algorithm 1: CDS algorithm based on medical literature

Input: R_i: the content of the i-th patient record.

Output: $List_{fin}$: the list of relevant literature after re-ranking.

Begin

1: Initialize the MeSH medical terms $List_{mesh}$ and the number of retrieved literature by Indri, c.

2: Initialize parameters α, β.

3: Extract the queries from the summary section of R_i by MetaMap → $Query = \{Query_1, Query_2, ..., Query_n\}$

4: Extend the basic $Query$ by UMLS, $Query$ → $Query' = \{Query_1', Query_2', ..., Query_m'\}$

5: ***Function*** $Indri(Query', 1{,}000)$ ***do*** → $List_{ini}$

6: Initialize the MLN $G_{ini} = \{V_{ini}, E_{ini}\}$ based on $List_{ini}$ in accordance with $List_{mesh}$.

7: Extract the MeSH medical terms from G_{ini}, $List_{mesh}'$.

8: ***for*** $mesh_i \in List_{mesh}'$ ***do***

9: $List_{temp} = Indri(mesh_i, c)$

10: ***for*** $node_{temp} \in List_{temp}$ ***do***

11: ***if*** the clustering coefficient method is adopted ***then***

12: Calculate the clustering coefficient of communities, C_{com}.

13: Calculate the clustering coefficient of $node_{temp}$, C_{node}.

14: ***if*** $C_{node} > C_{com}$ ***then*** $node_{temp}$ → G_{ini}

15: ***else if*** the connectivity method is adopted ***then***

16: Calculate the lexical density of communities, and select the most densely community $Com_{max.}$

17: Calculate the connectivity of Com_{max} and $node_{temp}$ → C_{com}, C_{node}.

18: ***if*** $C_{node} > C_{com}$ ***then*** $node_{temp}$ → G_{ini}

19: ***end if***

20: ***end for***

21: ***end for***

22: After identifying the potential literature, the more comprehensive MLN is built, G_{ini} → G_{fin}

23: ***Function*** $Re\text{-}rank(G_{fin}, \alpha, \beta)$ → $List_{fin}$ ***do***

24: return $List_{fin}$

4.2. Re-ranking effect analysis

After the potential-literature mining, we adopt the re-ranking model to identify 1000 relevant literature from the 1200 literature. However, the re-ranking model contains some uncertain factors that may influence the accuracy, such as weight parameter α and the importance of node $I(v_i)$. To choose appropriate α and the proper importance index of the node, we calculate the NDCG of each topic to evaluate the effects under different α values and importance indexes. The NDCG formula [41] for measuring the relevance of the retrieval result is expressed as follows:

$$NDCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i} \tag{8}$$

where rel_i represents the score of the ith literature; relevant is one, whereas irrelevance is zero. p is the number of the retrieved result, which is 1000 in this paper. We randomly select six topics from "diagnosis", "test" and "treatment". The

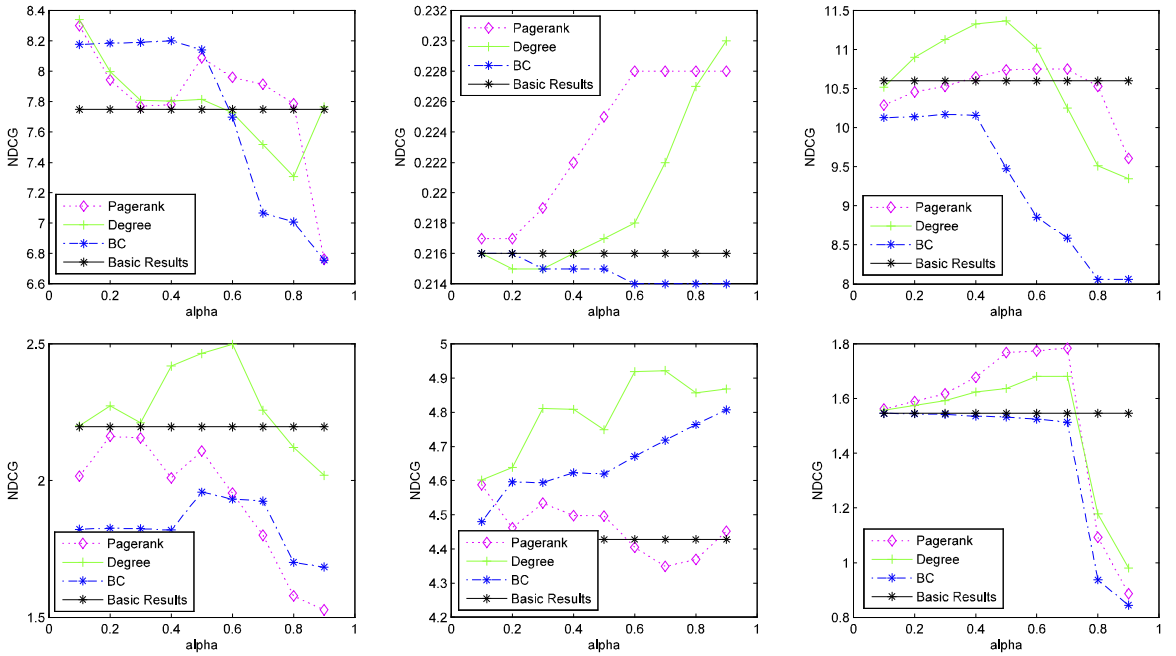

Fig. 7. Re-ranking effect analysis of the "diagnosis" topics.

x-axis represents weight parameter α from 0.1 to 0.9. The y-axis represents the NDCG score. In addition, we employ PageRank, degree and betweenness centrality as the importance indexes. The NDCG of the basic results without re-ranking is also calculated for comparison of the effects. The experimental results are shown in Figs. 7–9.

Fig. 7 shows that the curve of the NDCG score is irregular. When the value of α is between 0.4 and 0.6, the NDCG generally shows an increase, which illustrates that the re-ranking model exhibits the best performance when the structural and relevance factors coexist and play the same important roles. We can also observe that the re-ranking results of the degree distribution is better than the two other importance indexes. Finally, we numerically select 0.6 as weight parameter α and the degree distribution as the importance index of the re-ranking model for the "diagnosis" topics.

Following the same analysis of the mean of the "diagnosis" topic, we choose the weight parameter and importance index for the "test" topics and "treatment" topics. However, some singular curves appear in the abovementioned topics. The fourth topic in Fig. 9 shows that the curve does not change with α and remains at zero because our method does not find any relevant literature in this topic. Furthermore, we find that some topics exhibit continual decrease with increasing α, which illustrates that the accuracy of the re-ranking model decreases with the increasing percentage of the structural factor. Therefore, we doubt that a less comprehensive MLN has been built, and the structural factor of the re-ranking model suffers from a negative effect. Considering these singular curves and all the cases, we use 0.3 and 0.7 as the weight parameter of the "test" topics and "treatment" topics, respectively. PageRank and betweenness centrality are employed as the importance indexes, respectively.

In addition, the distribution of the relevant literature is analyzed. We perform statistical analysis of the number of relevant literature in the top 100, 200, 500 and 1000 retrieved results. Fig. 10 shows that the number of relevant literature is evenly distributed. This phenomenon does not show that all the papers are concentrated in a specific range, such as distributed between 500 and 1000. Therefore, this experiment proves the rationality of the re-ranking model.

Given the above results, we can conclude that the re-ranking model can improve the performance of relevance ranking. Through many experiments and analyses, we choose an appropriate α value and a proper importance index for the different types of topics. Finally, we reveal the distribution of relevant literature in the retrieval results and verify the rationality of the re-ranking model.

4.3. Comparing the submitted runs with those of the other participants

As participants in TREC 2015, our retrieval results, which are generated by the mining method and re-ranking model, are submitted to the official website. After the evaluation results of our runs are announced, a set of experiments for comparison with those of the other participants is given, as shown in Fig. 11.

Fig. 11 shows that the performance of our method based on the MLN is superior to the median scores in most topics. In addition, the sequence of the mining effect for the different types of topics from high to low is as follows: treatment, test and diagnosis. The result is similar to the experimental results using the topic of TREC 2014. Surprisingly, we find that

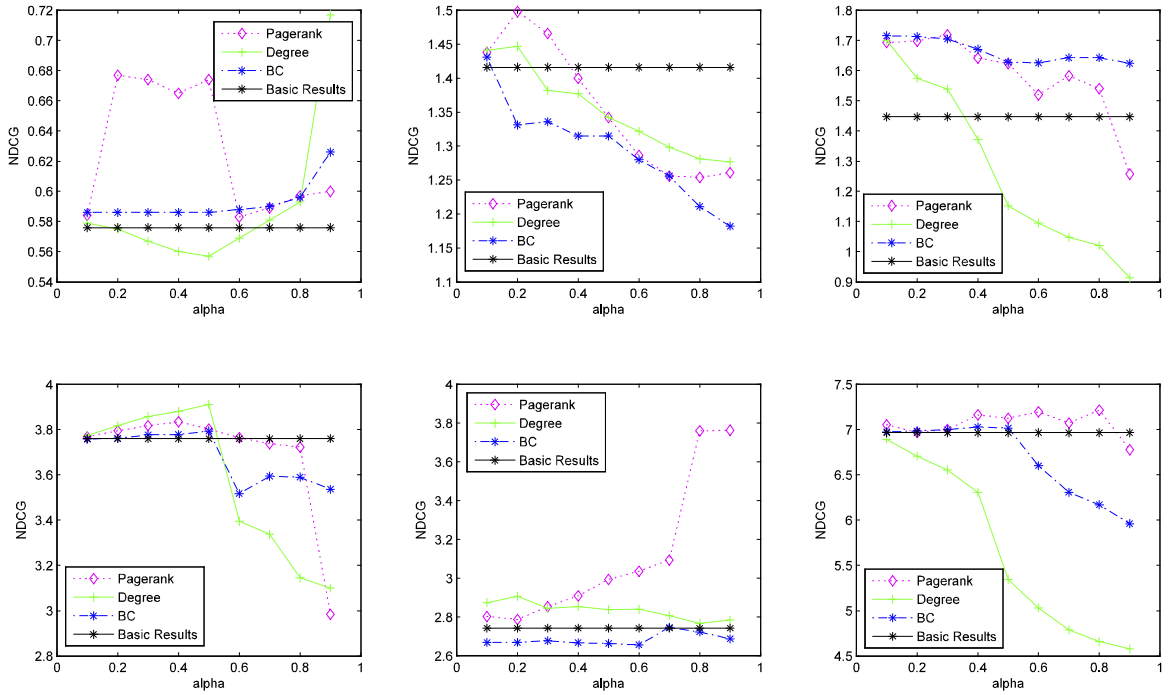

Fig. 8. Re-ranking effect analysis of the "test" topics.

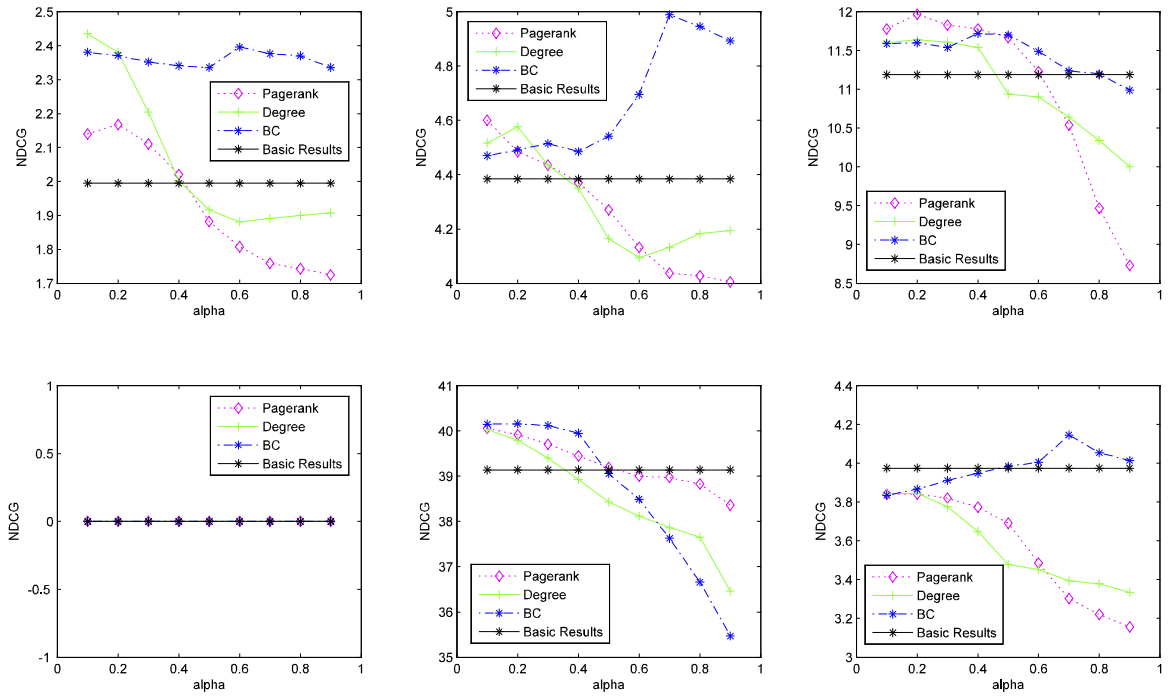

Fig. 9. Re-ranking effect analysis of the "treatment" topics.

our method achieves the best score in two topics: #11 and #12. These results further testify to the effectiveness of the CDS algorithm.

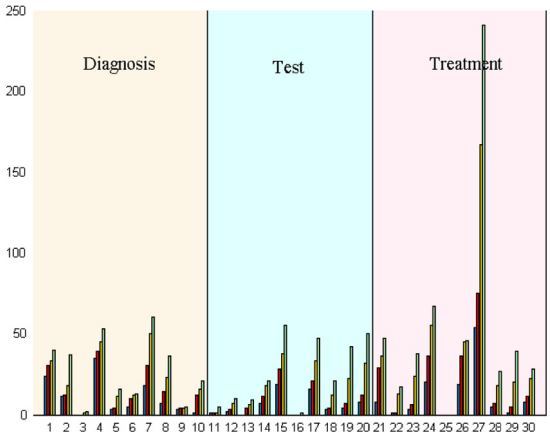

Fig. 10. Distribution of the relevant literature.

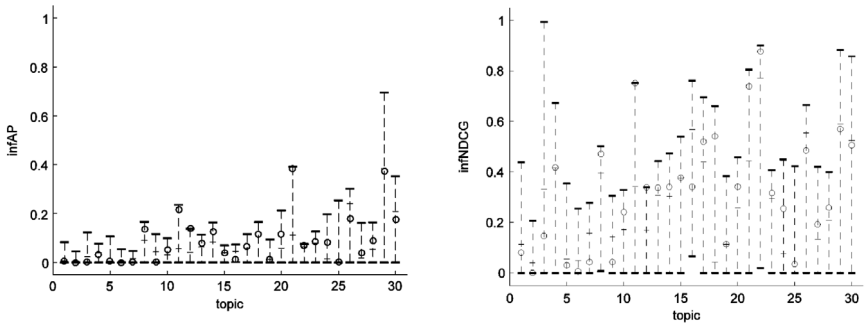

Fig. 11. Comparison of our submitted runs with those of the other participants.

5. Conclusion and future work

In this paper, we presented our study on CDS on the basis of medical literature. We built an MLN using the basic retrieved results. The small-world and scale-free properties were validated in the MLN. To help clinicians make the most appropriate diagnosis, test and treatment, we used the features of a complex network to mine more papers. We adopted two mining methods, namely, CC-based and CO-based methods, to identify the potential literature. We further proposed the re-ranking model, which synthesizes the structural factor of the MLN and the relevance factor of the search engine. We numerically determined that the re-ranking model can improve the relevance of the results. Compared with those of the other participants in TREC 2015, the effectiveness of our method was verified.

In the future, according to the different sources of medical knowledge, we will construct and study the massive commonsense knowledge network to more effectively help clinicians in making clinical decisions.

Acknowledgments

The Open Access Subset of PubMed Central used in this paper was provided by TREC 2015 Clinical Decision Support (CDS) track. We would like to thank the reviewers for their detailed reviews and insightful comments, which have helped to improve the quality of this paper.

References

[1] World Health Organization (WHO), Cardiovascular diseases. [Online] Available: http://www.who.int/cardiovascular_diseases/en/.
[2] American Heart Association, Make the Effort to Prevent Heart Disease. [Online] Available: http://www.heart.org/HEARTORG/GettingHealthy/Make-the-Effort-to-Prevent-Heart-Disease-with-Lifes-Simple-7_UCM_443750_Article.jsp.
[3] World Health Organization (WHO), Raised fasting blood glucose. [Online] Available: http://www.who.int/gho/ncd/risk_factors/blood_glucose_text/en/.
[4] C. Day, The rising tide of type 2 diabetes, Br. J. Diabetes Vasc. Dis. 1 (1) (2001) 37–43.
[5] M.A. Musen, B. Middleton, R.A. Greenes, Clinical decision-support systems, in: Biomedical Informatics, Springer, London, 2014, pp. 643–674.
[6] T.J. Bright, A. Wong, R. Dhurjati, et al., Effect of clinical decision-support systems: A systematic review, Ann. Intern. Med. 157 (1) (2012) 29–43.
[7] C.C. Tseng, P.J. Gmytrasiewicz, Real-time decision support and information gathering system for financial domain, Physica A 363 (2) (2006) 417–436.
[8] T.J. Carney, G.P. Morgan, J. Jones, et al., Using computational modeling to assess the impact of clinical decision support on cancer screening improvement strategies within the community health centers, J. Biomed. Inform. 51 (2014) 200–209.

[9] P.S. Roshanov, N. Fernandes, J.M. Wilczynski, et al., Features of effective computerised clinical decision support systems: Meta-regression of 162 randomised trials, BMJ 346 (2013) f657.
[10] J.A. Osheroff, J.M. Teich, B. Middleton, et al., A roadmap for national action on clinical decision support, J. Am. Med. Inform. Assoc. 14 (2) (2007) 141–145.
[11] M.S. Simpson, E. Voorhees, W. Hersh, Overview of the TREC 2014 clinical decision support track, in: Proc. 23rd Text Retrieval Conference (TREC 2014), National Institute of Standards and Technology(NIST), 2014.
[12] TREC Clinical Decision Support Track, 2015 Task: Generic Clinical Questions. [Online] Available: http://www.trec-cds.org/2015.html.
[13] TREC Clinical Decision Support Track, 2014 Task: Generic Clinical Questions. [Online] Available: http://www.trec-cds.org/2014.html.
[14] H.S. Oh, Y. Jung, KISTI at TREC 2014 Clinical Decision Support Track: Concept-based Document Re-ranking to Biomedical Information Retrieval, TREC 2014 Track, 2014.
[15] J. Gobeillab, A. Gaudinata, E. Paschec, et al. Full-texts representation with Medical Subject Headings, and co-citations network rerank-ing strategies for TREC 2014 Clinical Decision Support Track, TREC 2014 Track, 2014.
[16] J.I. Garcia-Gathrighta, F. Menga, W. Hsua, UCLA at TREC 2014 Clinical Decision Support Track: Exploring Language Models, Query Expansion, and Boosting, TREC 2014 Track, 2014.
[17] A. Mourao, F. Martins, J. Magalhaes, NovaSearch at TREC 2014 Clinical Decision Support Track, TREC 2014 Track, 2014.
[18] T. Xu, P. McNamee, D.W. Oard, HLTCOE at TREC 2014: Microblog and Clinical Decision Support, TREC 2014 Track, 2014.
[19] S. Choi, J. Choi, SNUMedinfo at TREC CDS track 2014: Medical case-based retrieval task, TREC 2014 Track, 2014.
[20] K. Yamamoto, Y. Yamazaki, Structure and modeling of the network of two-Chinese-character compound words in the Japanese language, Physica A 412 (2014) 84–91.
[21] D. Tomasi, N.D. Volkow, Resting functional connectivity of language networks: Characterization and reproducibility, Mol. Psychiatry 17 (8) (2012) 841–854.
[22] Y. Gao, W. Liang, Y. Shi, et al., Comparison of directed and weighted co-occurrence networks of six languages, Physica A 393 (2014) 579–589.
[23] C. Yi, Y. Bao, J. Jiang, et al., Modeling cascading failures with the crisis of trust in social networks, Physica A 436 (2015) 256–271.
[24] D.R. Amancio, M.G.V. Nunes, O.N. Oliveira, et al., Using metrics from complex networks to evaluate machine translation, Physica A 390 (1) (2011) 131–142.
[25] S. Guo, Q. Wang, B. Wang, et al. Semantically smooth knowledge graph embedding, in: Proceedings of ACL, 2015.
[26] D.R. Amancio, O.N. Oliveira, L. da Fontoura Costa, Three-feature model to reproduce the topology of citation networks and the effects from authors' visibility on their h-index, J. Informetr. 6 (3) (2012) 427–434.
[27] D.J. Watts, S.H. Strogatz, Collective dynamics of "small-world" networks, Nature 393 (6684) (1998) 440–442.
[28] H. Jeong, B. Tombor, R. Albert, et al., The large-scale organization of metabolic networks, Nature 407 (6804) (2000) 651–654.
[29] A.L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (5439) (1998) 509–512.
[30] Y. Tachimori, H. Iwanaga, T. Tahara, The networks from medical knowledge and clinical practice have small-world, scale-free, and hierarchical features, Physica A 392 (23) (2013) 6084–6089.
[31] M. Bastian, S. Heymann, M. Jacomy, Gephi: an open source software for exploring and manipulating networks, in: Proc. of 3rd International AAAI Conference on Weblogs and Social Media, ICWSM, 2009.
[32] M.E.J. Newman, Modularity and community structure in networks, Proc. Natl. Acad. Sci. 103 (23) (2006) 8577–8582.
[33] L. Šubelj, M. Bajec, Unfolding communities in large complex networks: Combining defensive and offensive label propagation for core extraction, Phys. Rev. E 83 (3) (2011) 036103.
[34] J. Shang, L. Liu, X. Li, et al., Epidemic spreading on complex networks with overlapping and non-overlapping community structure, Physica A 419 (2015) 171–182.
[35] J. Xie, S. Kelley, B.K. Szymanski, Overlapping community detection in networks: The state-of-the-art and comparative study, ACM Comput. Surv. (CSUR) 45 (4) (2013) 43.
[36] M.E.J. Newman, The structure and function of complex networks, SIAM Rev. 45 (2) (2003) 167–256.
[37] T. Zhou, G. Yan, B.H. Wang, Maximal planar networks with large clustering coefficient and power-law degree distribution, Phys. Rev. E 71 (4) (2005) 046141.
[38] J. Saramäki, M. Kivelä, J.P. Onnela, et al., Generalizations of the clustering coefficient to weighted complex networks, Phys. Rev. E 75 (2) (2007) 027105.
[39] Y. Cui, X. Wang, J. Li, Detecting overlapping communities in networks using the maximal sub-graph and the clustering coefficient, Physica A 405 (2014) 85–91.
[40] L. Wu, Q. Tan, Y. Zhang, Network connectivity entropy and its application on network connectivity reliability, Physica A 392 (21) (2013) 5536–5541.
[41] E. Yilmaz, E. Kanoulas, J.A. Aslam, A simple and efficient sampling method for estimating AP and NDCG, in: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2008.