Journal of Biomedical Informatics 58 (2015) 70–79

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Cluster-based query expansion using external collections in medical information retrieval

Heung-Seon Oh, Yuchul Jung *

Korea Institute of Science and Technology Information, 245 Daehak-ro, Yuseong-gu, Daejeon, South Korea

ARTICLE INFO

Article history:
Received 26 February 2015
Revised 22 September 2015
Accepted 23 September 2015
Available online 30 September 2015

Keywords:
Query expansion
External collections
Language models

ABSTRACT

Utilizing external collections to improve retrieval performance is challenging research because various test collections are created for different purposes. Improving medical information retrieval has also gained much attention as various types of medical documents have become available to researchers ever since they started storing them in machine processable formats. In this paper, we propose an effective method of utilizing external collections based on the pseudo relevance feedback approach. Our method incorporates the structure of external collections in estimating individual components in the final feedback model. Extensive experiments on three medical collections (TREC CDS, CLEF eHealth, and OHSUMED) were performed, and the results were compared with a representative expansion approach utilizing the external collections to show the superiority of our method.

1. Introduction

With the increasing amount of medical information available through the Internet, health-related contents have become one of the most searched-for topics on the Internet. Nowadays, people use Web search engines to acquire medical information or browse online health communities. Health professionals themselves frequently utilize Web search engines to get related medical knowledge or to facilitate a diagnosis.

Those searched contents include health related web pages, biomedical literature, clinical records, health Q/As, etc. [8]. Medical information retrieval (IR) can be explained as the activity of people seeking health information across diverse health information sources. Previously, medical literature retrieval (or medical document retrieval) was defined as "an activity that uses professional methods for medical research papers retrieval, report and other data to improve medicine research and practice".[1] Meanwhile, the people involved in medical IR are patients, laypeople, and healthcare professionals (doctors, nurses, therapists, etc.).

Laypeople have difficulties in forming appropriate queries that fit their information needs due to a lack of medical knowledge. In the case of clinicians, they have problems in interpreting the jargon of other professional groups (or subjects) as well as finding relevant medical cases from a large volume of biomedical literatures and prior clinical decision data. The fast increasing medical information from different types of resources makes medical IR even more challenging.

Most of the recent medical IR research has focused on developing knowledge-based (or concept-based) retrieval models depending on the medical resources such as the unified medical language system (UMLS) thesaurus [3,9,26,27,31,41].

The main reason for using the UMLS thesaurus is to solve the vocabulary mismatch problem (synonymy and polysemy) between a query and documents [7]. This problem is quite important in medical IR because a user may not have enough medical knowledge to sufficiently express their information needs in a query. An effective way of solving the problem is to expand a query that expresses the information needs by including more useful words. Pseudo relevance feedback (PRF) has been a promising approach to expand a query with the assumption that the top-ranked documents are relevant to a query. A mixture of relevance models (MoRM) [13] is an effective PRF method which utilizes available external collections. The key idea of the MoRM is to generate a feedback model by combining relevance models (RMs) built from individual collections. Recently, the external expansion model (EEM) was introduced based on the MoRM [42]. It is an extension of the MoRM to determine the mixture weights of external collections for a given query in constructing a feedback model. However, it was observed that the EEM doesn't exploit external collections to estimate document models in a feedback model.

In order to overcome the limitation of the EEM, in this study, we propose a cluster-based external expansion model (CBEEM) that

* Corresponding author.
E-mail addresses: ohs@kisti.re.kr (H.-S. Oh), jyc77@kisti.re.kr (Y. Jung).

[1] http://en.wikipedia.org/wiki/Medical_literature_retrieval.

http://dx.doi.org/10.1016/j.jbi.2015.09.017
1532-0464/

incorporates the structure of external collections to estimate each RM in a feedback model. In our CBEEM, we assumed that a collection corresponds to a cluster among all external collections partitioned by an explicit structure rather than implicit topics. Therefore, a RM is constructed by considering the structure of all external collections.

Our proposed CBEEM has strengths such as extending the original queries with related context information from the external collections in a more efficient manner. It doesn't require any additional computations, such as performing K-means clustering.

Most of the recent research on medical IR has been with the TREC Medical Record[2] collection [40] with 93,551 clinical reports. However, to concretely validate our method, we performed exhaustive experiments on collections for medical IR (TREC CDS, CLEF eHealth, and OHSUMED) which were developed for different purposes.

In addition to the experiments on the collections for medical IR, we examined the feasibility of Wikipedia as an external knowledge base. Wikipedia, an online free encyclopedia containing a tremendous number of concepts in the real world, is widely used as a text resource because of its free availability. The use of Wikipedia has shown moderate performance improvements in various IR studies [12,15,44]. However, we identified that Wikipedia is not always helpful in medical IR especially with MoRM, against our expectations.

The key contributions of this paper are summarized as follows:

(1) We propose a cluster-based external expansion model (CBEEM) that incorporates the structure of external collections at estimating document models with an assumption that a collection corresponds to an explicit cluster over all the external collections.
(2) We performed exhaustive experiments on three different medical test collections (TREC CDS, CLEF eHealth, and OHSUMED) to concretely evaluate the effectiveness of the proposed CBEEM. The experimental results show that our CBEEM significantly outperforms previous methods in all three test collections. Furthermore, we provide the feasibility test results for using Wikipedia as an additional external knowledge base.

The rest of this paper is organized as follows. Section 2 reviews important previous literature related to medical IR. Section 3 covers the basic foundation of language modeling for IR. Our proposed method is explained in Section 4. The results of our exhaustive experiments are reported with an in-depth analysis in Section 5. We conclude in Section 6 with a discussion and talk about future work in this area.

2. Related work

Vocabulary mismatch is a critical problem in IR. The use of concept mapping and external collections have been investigated extensively to try to solve the problem in medical IR.

Use of concept mapping: Concept-based IR [15] presents both a query and a document as a set of semantic concepts and performs retrieval in the concept space. In the medical domain, concept-mapping is performed with MetaMap [2] which is a tool that provides various NLP functions to process biomedical texts. As a biomedical concept resource, it uses the Unified Medical Language System (UMLS) [5] including about 900,000 biomedical concepts from several resources such as the NCBI taxonomy, Gene Ontology, Medical Subject Headings (MeSH), OMIM[3] and the Digital Anatomist Symbolic Knowledge Base. Many researchers have attempted to devise retrieval methods in concept-based IR. Considering word dependencies as a ranking function is useful in improving performance [20,33]. Similarly, [9] addresses the importance of concept dependencies in ranking and proposes semantic concept-enriched dependency features on Markov random fields (MRFs). Concept-weighting regularization methods [41] were developed based on an axiomatic IR framework to overcome the limitations of imperfect concept mapping. A query expansion method based on the PageRank algorithm was proposed in [31]. Simply, a set of highly-ranked concepts with high PageRank scores are selected as expansion terms.

Employment of external collections: A method utilizing external collections [13] was introduced to improve the estimation of accurate relevance models, and the authors showed the superiority of their method through experiments. In the medical domain, the effectiveness of using in-domain biomedical collections utilizing a mixture of relevance models for cohort identification was investigated in [47]. If we follow the result analysis, we can expect performance improvements by using diverse external collections in conjunction with a targeted collection. In the biomedical domain, terminology-based query expansion approaches mostly utilize concept-mapping based on the lexical knowledge resources such as MetaMap, UMLS, and MeSH. Queries can be expanded by identifying concepts in the resources. Thus, words in different or similar meanings can be mapped to concepts in various meanings. Although it is helpful in expanding various terminologies, deep understanding about the knowledge resources is essential to effectively use it. Meanwhile, the methods of utilizing external collections can discriminate contextual words corresponding to a query by constructing a relevance feedback model from the target collection.

In recent research in the medical IR field, most experiments have been performed with the collections of the TREC Medical Record Track 2011 and 2012 [40]. The collections consist of various types of clinical reports such as Radiology Reports, Histories and Physicals, Consultations Reports, and Emergency Department Reports. Based on the needs of the medical IR field, the TREC CDS and CLEF eHealth [21] collections were developed to encourage medical IR research using different types of documents such as medical literature and medical web documents.

3. Background and preliminaries

In this section, we briefly introduce the scoring methods of the query-likelihood [35] and Kullback–Leibler divergence [23] as basic background information for IR. Then, we review the mixture of relevance models [13] to show an example of utilizing external collections using pseudo relevance feedback.

The query-likelihood (QL) method introduced in [35] is the foundation of language modeling approach to IR. Using this method, we compute a probability of generating a query Q from a unigram language model θ_D of a document D and rank documents with the probabilities:

$$Score(Q,D) = P(Q|\theta_D) = \prod_w P(w|\theta_D)^{c(w,Q)} \quad (1)$$

where $c(w,Q)$ is the count of a word w in a query Q.

It is difficult to incorporate relevance or pseudo relevance feedback into query Q using the QL method. To overcome this limitation, the Kullback–Leibler divergence (KLD) method was proposed in [23]. In the KLD method, a query Q is represented as a unigram language model θ_Q not the counts of the words. Therefore, it is easy to perform feedback by re-estimating θ_Q. The

[2] TREC Medical Record collection is not available since University of Pittsburgh is no longer distributing it.
[3] http://www.ncbi.nlm.nih.gov/omim.

KLD method computes the divergence between a query model θ_Q and a document model θ_D:

$$Score(Q,D) = -KLD(\theta_Q||\theta_D) = -\sum_w P(w|\theta_Q)\log\frac{P(w|\theta_Q)}{P(w|\theta_D)} \quad (2)$$

Due to its benefits, the KLD method is widely used in a variety of IR tasks with remarkable performance improvement. In document re-ranking [22], the KLD method is utilized to estimate an approximate probability of generating a document D_1 from D_2:

$$Score(D_1,D_2) = \exp(-KLD(\theta_{D_1}||\theta_{D_2}) \quad (3)$$

In the KLD method, the retrieval performance depends on how to estimate the query model θ_Q and document model θ_D. In general, a query model is estimated using the maximum likelihood estimate (MLE) without feedback:

$$P(w|\theta_Q) = \frac{c(w,Q)}{|Q|} \quad (4)$$

where $|Q|$ is the number of words in a query Q.

Estimating θ_D is a challenging research problem. A simple way of estimating θ_D is to utilize the MLE:

$$P_{ML}(w|\theta_D) = \frac{c(w,D)}{|D|} \quad (5)$$

where $|D|$ is the number of words in a document D.

However, the major drawback of the MLE is the assignment of a zero probability to unseen words in documents. Various smoothing methods have been developed not only to overcome this drawback but also to improve the retrieval performance in discriminating topical words in documents. One of the most promising methods is two-stage smoothing introduced in [46]:

$$P_{TS}(w|\theta_D) = (1-\lambda)\cdot\frac{c(w,D)+\mu\cdot P(w|C)}{|D|+\mu} + \lambda\cdot P(w|U) \quad (6)$$

where μ and λ are the Dirichlet prior and Jelinek–Mercer (JM) smoothing parameters, respectively, and C represents a document collection. The second term $P(w|U)$ is the user's query background language model. When $\lambda = 0$, two-stage smoothing is the same as Dirichlet smoothing while it becomes the same as JM smoothing when $\mu = 0$. In general, it is approximated by $P(w|C)$ with insufficient data to estimate $P(w|U)$ even though it is different from $P(w|C)$.

Query expansion aims at dealing with the vocabulary mismatch problem between a query Q and a document D [7]. In language modeling framework, the strategy of the PRF is a dominant way of expanding a query with promising performance improvement [24,25,29,30,32,44]. The basic idea of the PRF using language models is to re-estimate a query model θ_Q using a small number of top-ranked documents $R = \{D_1, D_2, \ldots, D_{|R|}\}$. It is assumed that the documents in R are relevant to a query and have useful words to re-estimate a more accurate query model θ_Q. A new query model is defined as

$$P(w|\theta'_Q) = (1-\lambda)\cdot P(w|\theta_Q) + \lambda\cdot P(w|Q) \quad (7)$$

Conventional IR tasks focused on the development of a retrieval model in a given collection. Due to the diversity of information needs, various IR tasks have been developed with a test collection and test queries for different purposes. Therefore, there is a need to develop retrieval models by exploiting existing collections.

In [13], the mixture of relevance models (MoRM) was introduced to utilize external collections in the context of the PRF. The key idea of the MoRM is to generate a feedback model by combining relevance models (RMs) constructed from the most relevant documents with respect to a given query for collections of interest. The formal definition of the MoRM is as follows:

$$P_{MoRM}(w|Q) \propto \sum_{C\in E}\frac{P(\theta_C)}{|R_c|}\sum_{D\in R_c}P(w|\theta_D)\cdot P(Q|\theta_D)\cdot P(D|\theta_C) \quad (8)$$

where R_C is a set of ranked documents in a collection C; $P(\theta_c)$ is a collection prior; $P(w|\theta_D)$ is a word probability; $P(Q|\theta_D)$ is a query-likelihood in a document D; $P(D|\theta_C)$ is a document-likelihood in a collection C, and E is the entire collection that includes the target and external collections.

The MoRM is a simple feedback model focusing on two collections: target and external collections (i.e., $E = \{C_{target}, C_{external}\}$). It is difficult to determine the proper contribution of an external collection for different queries because a prior fixed collection $P(\theta_C)$ controls the contributions of the two collections in the feedback model.

4. Our proposed method

In this section, we present a method, the cluster-based external expansion model (CBEEM) that effectively utilizes external collections in building a feedback model. For this aim, we combine the expand external expansion model (EEM) [42] and the concept of the cluster-based document model (CBDM) [28].

In [42], the external expansion model (EEM) was proposed to overcome the previously mentioned limitations of the MoRM. The key idea of the EEM is to generate a feedback model by determining the proper contributions of multiple collections for a given query. Formally, the EEM is defined as follows:

$$P_{EEM}(w|Q) \propto \sum_{C\in E}P(Q|\theta_C)\cdot P(\theta_C)\sum_{D\in C}P(w|\theta_D)\cdot P(Q|\theta_D)\cdot P(D|\theta_C) \quad (9)$$

Specifically, the EEM consists of five components: prior collection probability, document relevance, collection relevance, document importance, and word probability.

(1) Prior collection probability $P(\theta_C)$ is the prior importance of a collection among all the collections in use. A collection containing documents written by medical experts such as doctors and pharmacists may be more important than the one containing documents written by less experts such as patients or medical students. Without the prior knowledge of collections, it can be ignored by setting a uniform probability $P(\theta_C) = \frac{1}{|E|}$.

(2) Document relevance $P(Q|\theta_D)$ is the relevance of a document D to a given query Q. Precisely, it is a query-likelihood score given to a document. But various methods such as BM25 [36], DFR [1], and Markov random fields [33] can be utilized to estimate this component.

(3) Collection relevance $P(Q|\theta_C)$ is the relevance of a query Q with respect to a collection C. This component determines the query-dependent contribution of a collection in constructing the EEM. We should compute relevance scores against a query Q by iterating all documents in a collection C. To avoid this time-consuming iteration, it can be estimated using the most relevant documents with an assumption that documents are equally important in a given collection C as follows:

$$P(Q|\theta_C) = \sum_{D\in C}P(Q|\theta_D)\cdot P(\theta_D|C) \propto \frac{1}{|R_C|}\cdot\sum_{D\in R_c}P(Q|\theta_D) \quad (10)$$

This is the average score of the documents in R_C.

(4) Document importance $P(D|\theta_C)$ is the importance of a document D in a collection C. Similar to the prior collection probability, documents are not equally important in a given collection. There are various options to estimate this component including PageRank [34], recency [14], document

length [4,18], and URL length [43]. This is also ignored by setting to a uniform probability $P(D|\theta_C) = \frac{1}{|C|}$ without the prior knowledge of documents in a collection C.

(5) Word probability $P(w|\theta_D)$ is a probability of observing a word w in a document D. In [42], the MLE is utilized to estimate this component. Their experiments showed that the EEM obtained improvements over the MoRM on blog search task utilizing multiple external collections (i.e., news, Web, blogs, and Wikipedia). Specifically, in the EEM, the query-dependent collection relevance $P(Q|\theta_C)$ significantly improves performance while the prior collection probability $P(\theta_C)$ hardly affects performance.

In the EEM, information from external collections is not employed in parameter estimation to construct the RM for individual collections. Our initial idea to derive an advanced version of the EEM was that utilizing external collections in constructing the RMs can generate a more accurate EEM. In the cluster-based document model [28], a document model is smoothed with cluster and collection models in which clusters are generated with the K-means algorithm. Therefore, we can obtain more accurate document models because the probabilities of words which occur frequently in a cluster or a collection are decreased. Similarly, we can assume that each collection corresponds to a cluster explicitly partitioned over E. This assumption makes us use the cluster-based document model without any additional computations performing K-means clustering because K is determined as $|E|$, and each collection is a cluster. All that is required is to utilize the statistics of a collection C for a cluster. Then, a document model is defined as follows:

$$\begin{aligned} P(w|\theta_D) &= (1-\lambda_E)\cdot\frac{c(w,D)+\mu\cdot P(w|C)}{|D|+\mu}+\lambda_E\cdot P(w|E) \\ &= (1-\lambda_E)\cdot\left[\frac{|D|}{|D|+\mu}P(w|D)+\frac{\mu}{|D|+\mu}P(w|C)\right]+\lambda_E \\ &\quad\cdot P(w|E) \end{aligned} \tag{11}$$

where λ_E is a control parameter for all collections in E.

Our proposed CBEEM is defined by revising $P(w|\theta_D)$ in Eq. (9) and replacing it with that of Eq. (11). Based on the revision, the CBEEM is expected to be a probability distribution over topical words because it is combined with individual RMs because of the decrease in probability of common words in the feedback documents.

Time complexity for clustering is not needed because the explicit structure among external collections is used to define clusters. Therefore, the time complexity of generating the CBEEM is $O(|E|\times|R_C|)$ while it is $O(|R_C|)$ for a feedback model using a single collection. It was not a major concern in our experiments because R_C was constrained to small numbers, such as 5.

5. Experiments

5.1. Data

Tables 1 and 2 show a summary of the collections used in our experiments and examples of test queries for each collection, respectively. Three medical collections with test queries were exploited for a concrete evaluation: TREC CDS 2014,[4] CLEF eHealth 2014 [16], and OHSUMED [17]. TREC CDS consists of biomedical literature, specifically a subset of PubMed Central, with 30 test queries. A document is a full-text XML of a journal article. Test queries are classified into one of three classes: diagnosis, treatment, and test. The characteristics of the TREC CDS collection are that the average length of both a document and query is very long compared with other collections. The relevance is judged from 0 to 2 where 0 is not relevant; 1 is relevant, and 2 is strongly relevant. Only the description is used as query text for retrieving documents. CLEF eHealth consists of medical-related documents from several online sources with 50 test queries. Title and description are used as query text. Similar to TREC CDS, the relevance is judged from 0 to 3. OHSUMED consists of biomedical literature which is a subset of the clinically-oriented MEDLINE with 105 test queries. Similar to TREC CDS, a document corresponds to a journal article while the body text of an article is not provided in OHSUMED. The average length of both a query and document is very short compared to the others. Unlike TREC CDS and CLEF eHealth, the relevance is judged as one of the following: definitely relevant (d), possibly relevant (p), and not relevant (n). We converted them to 2, 1, and 0, respectively, for evaluation. Query text is generated with patient information denoted as .B. In addition to the three biomedical collections, we exploited Wikipedia as an external collection only to provide useful information. Clearly, $E = \{C_{CDS}, C_{eHealth}, C_{OHSUMED}, C_{WIKI}\}$. Except for Wikipedia, each collection can be a target collection for a retrieval task in our experiments.

Table 1
Summary of collection statistics (queries without relevance judgments are omitted, and the lengths are counted after stop-word removal).

	TREC CDS	CLEF eHealth	OHSUMED	Wikipedia
#Docs	732,451	1,102,289	348,566	7,214,991
Voc. Size	6,931,356	2,647,062	122,512	14,008,271
Avg. Doc. Len	1779.0	540.0	68.0	178.0
#Queries	30	50	105	N/A
Avg. Query Len.	39.6	7.2	3.8	N/A
#Query-Doc	3356	3756	4527	N/A

5.2. Experimental setup

The experimental procedure is as follows. For a query, a set of documents for each $C \in E$ are retrieved using a search engine. They are the initial search results for E. Then, a CBEEM is constructed using Eqs. (9)–(11) with a small number of top-ranked documents from the initial search result for C. The query is updated using the CBEEM, specifically by Eq. (7). Finally, documents are scored with the expanded query using Eq. (3).

Lucene,[5] an open source search engine, was chosen to index and retrieve documents. Stopwords were removed using 419 stopwords[6] in INQUERY. Numbers were normalized to NU<# of DIGITS>. For example, "19-years-old" was normalized to "NU2-years-old". The QL method with Dirichlet smoothing ($\mu = 1000$) was applied to obtain N documents from each collection as the initial search results. As a default value, N was set to 100. The QL method was chosen as a baseline, and it was compared with the EEM and CBEEM.

5.3. Evaluation metrics

Two popular metrics in IR were used for the evaluation: mean average precision (MAP) and normalized discounted cumulative gain (NDCG).

MAP is the mean value of average precision (AP) over all test queries. To understand MAP, we first have to explain precision at rank K ($P@k$). It computes the fraction of the top-k documents that are relevant:

[4] http://www.trec-cds.org/2014.html.

[5] http://lucene.apache.org/.

[6] http://sourceforge.net/p/lemur/galago/ci/default/tree/core/src/main/resources/stopwords/inquery.

Table 2
Examples of TREC CDS, CLEF eHealth, and OHSUMED queries.

Collection	Query
TREC CDS	<topic number = "17" type = "test"> <description>A 48-year-old white male with history of common variable immunodeficiency (CVID) with acute abdominal pain, fever, dehydration, HR of 132 bpm, BP 80/40. The physical examination is remarkable for tenderness and positive Murphy sign. Abdominal ultrasound shows hepatomegaly and abundant free intraperitoneal fluid. Exploratory laparotomy reveals a ruptured liver abscess, which is then surgically drained. After surgery, the patient is taken to the ICU.</description> <summary>48-year-old man with common variable immunodeficiency presents with abdominal pain and fever. Ultrasound reveals hepatomegaly and free intraperitoneal fluid. A ruptured liver abscess is found and drained during exploratory laparotomy.</summary> </topic>
CLEF eHealth	<topic><id>qtest2014.1</id> <discharge_summary>00211-027889-DISCHARGE_SUMMARY.txt</discharge_summary> <title>Coronary artery disease.</title> <desc>What does coronary artery disease mean? </desc> <narr>The documents should contain basic information about coronary artery disease and its care.</narr> <profile>This positive 83 year old woman has had problems with her heart with increased shortness of breath for a while. She has now received a diagnosis for these problems having visited a doctor. She and her daughter are seeking information from the internet related to the condition she has been diagnosed with. They have no knowledge about the disease.</profile> </topic>
OHSUMED	.I 15 .B 68 y.o. m. with adult-onset diabetes mellitus noted to have thrombocytosis .W thrombocytosis, treatment and diagnosis

$$P@k = \frac{1}{k} \cdot \sum_{i=1}^{k} br(i)$$

where $br(i)$ is a binary relevance function returning 1 if a document ranked at i is relevant otherwise 0.

AP is defined using the average of $P@k$:

$$AP@k = \frac{1}{R} \cdot \sum_{i=1}^{k} br(i) \cdot P@i$$

where R is the number of relevant documents at rank k.

MAP is defined as the mean of the AP over a set of queries:

$$MAP@k = \frac{1}{|TQ|} \cdot \sum_{Q \in TQ} AP(Q)@k$$

where TQ is a set of test queries.

Compared with MAP which considers binary relevance, NDCG is a measure of incorporating different degrees of relevance penalizing highly relevant documents presented at lower ranks:

$$NDCG@k = Z(k) \cdot \sum_{i=1}^{k} \frac{(2^{r(i)} - 1)}{\log_2(i+1)}$$

where $0 \leqslant r(i) \leqslant m$ is a relevance function. It is strongly relevant as $r(i)$ becomes high, and $Z(k)$ is a normalization term implying the ideal NDCG value at rank k.

Similar to MAP, the average of NDCG over all test queries is used. In our evaluation, k was set to 20. Performances marked with * and ** are statistically significant on two-sided paired t-test with $p \leqslant 0.05$ and $p \leqslant 0.01$.

5.4. Results

There are several parameters involved in both the EEM and CBEEM. D_{FB} and W_{FB} are the numbers of feedback documents and expansion words in each collection set to 5 and 25, respectively. λ_{FB} and λ_E are the mixture parameter of the CBEEM in a new query model and the smoothing parameter with external collections in the RM. We set them as $\lambda_{FB} = 0.5$ and $\lambda_E = 0.5$.

Table 3 shows the performances of the QL, EEM, and CBEEM on the three collections with the default settings. Improvements over QL are denoted in the parentheses. Among the three tasks, OHSUMED is the most difficult retrieval task because its aim is to find specialized biomedical literature with short queries and abstracts of journal articles. The performances of the QL were 0.1634 and 0.2432 in MAP and NDCG, respectively. They are the lowest scores compared with TREC CDS (0.3936 in MAP and 05342 in NDCG) and CLEF eHealth (0.8453 in MAP and 0.8629 in NDCG). CBEEM had the best performance with 0.1934 (15.51%) in MAP and 0.2774 (12.33%) in NDCG while the EEM is not guaranteed to achieve improvements over the QL because it is superior with 0.1662 (1.68%) in MAP but inferior with 0.2420 (−0.5%) in NDCG. On TREC CDS, the CBEEM achieved the best performance with 0.4641 (15.19%) in MAP and 0.5957 (10.32%) in NDCG. Similarly, it is difficult to conclude that the EEM works well because the two metrics behave differently with 0.4026 (2.24%) in MAP and 0.5270 (−1.37%) in NDCG. We can assume that TREC CDS is a more obvious task because it utilizes long queries which express information needs in detail and the full-text of articles shown in Table 2. CLEF eHealth supporting laypeople who do not have medical expertise is regarded as a much easier task compared with the others because the baseline performances, 0.8453 in MAP and 0.8629 in NDCG, are relatively high. It shows that the EEM is not effective with CLEF eHealth due to performance degradation in

Table 3
Performance comparison among the QL, EEM, and CBEEM for the three biomedical collections. Each bold number indicates the best performance in a metric (column). ($D_{FB} = 5, W_{FB} = 25, \lambda_{FB} = 0.5, \lambda_E = 0.5$).

Model	TREC CDS		OHSUMED		CLEF eHealth	
	MAP	NDCG	MAP	NDCG	MAP	NDCG
QL	0.3936	0.5342	0.1634	0.2432	0.8453	0.8629
EEM	0.4026 (2.24%)	0.5270 (−1.37%)	0.1662 (1.68%)	0.2420 (−0.5%)	0.8276 (−2.14%)	0.8597 (−0.37%)
CBEEM	***0.4641 (15.19%)**	***0.5957 (10.32%)**	****0.1934 (15.51%)**	****0.2774 (12.33%)**	**0.8478 (0.29%)**	**0.8776 (1.68%)**

Performances marked with * and ** are statistically significant on two-sided paired t-test with $p \leqslant 0.05$ and $p \leqslant 0.01$.

Table 4
Performance comparison by varying D_{FB} and W_{FB} with the CBEEM. Each bold number indicates the best performance in a metric (column) ($\lambda_{FB} = 0.5$ and $\lambda_E = 0.5$).

D_{FB}	W_{FB}	TREC CDS		OHSUMED		CLEF eHealth	
		MAP	NDCG	MAP	NDCG	MAP	NDCG
5	5	0.4009	0.5454	0.1854	0.2658	0.8383	0.8641
	10	0.4222	0.5605	0.1829	0.2688	0.8446	0.8753
	25	**0.4641**	**0.5957**	0.1934	0.2774	0.8478	0.8776
	50	0.4347	0.5762	0.1850	0.2705	0.8507	0.8751
	100	0.4234	0.5667	0.1869	0.2714	0.8540	0.8778
10	5	0.3979	0.5438	0.1911	0.2742	0.8355	0.8647
	10	0.4141	0.5510	0.1845	0.2699	0.8487	0.8758
	25	0.4028	0.5436	0.1927	0.2784	0.8509	0.8803
	50	0.4234	0.5592	0.1875	0.2733	0.8590	0.8822
	100	0.4090	0.5567	0.1873	0.2761	0.8620	0.8839
25	5	0.3965	0.5435	0.1928	0.2759	0.8308	0.8641
	10	0.3922	0.5273	0.1902	0.2745	0.8454	0.8772
	25	0.4039	0.5443	0.1901	0.2768	0.8626	0.8880
	50	0.4118	0.5593	0.1871	0.2758	0.8614	0.8875
	100	0.4092	0.5499	0.1865	0.2749	**0.8655**	0.8893
50	5	0.3981	0.5435	**0.1937**	0.2783	0.8317	0.8689
	10	0.3935	0.5338	0.1880	0.2722	0.8434	0.8766
	25	0.4030	0.5492	0.1932	0.2799	0.8547	0.8817
	50	0.4027	0.5517	0.1926	**0.2826**	0.8609	**0.8897**
	100	0.4127	0.5547	0.1893	0.2775	0.8642	0.8894
Max		*0.4641 (15.19%)	*0.5957 (10.32%)	**0.1937 (15.64%)	**0.2826 (13.94%)	*0.8655 (2.33%)	*0.8897 (3.01%)

Performances marked with * and ** are statistically significant on two-sided paired t-test with $p \leqslant 0.05$ and $p \leqslant 0.01$.

the two metrics while the CBEEM improves both performance metrics. From the results, the CBEEM consistently outperforms the QL and EEM for the three collections. Significant improvements are found for TREC CDS and OHSUMED but not for CLEF eHealth. Interestingly, the EEM is not guaranteed to improve performance for biomedical collections.

To concretely evaluate the effects of the CBEEM, we performed extensive experiments by varying several parameters (i.e., $\lambda_{FB}, \lambda_E, D_{FB}$, and W_{FB}) in the query expansion.

Table 4 shows the results of the CBEEM with $\lambda_{FB} = 0.5$ and $\lambda_E = 0.5$ by changing D_{FB} and W_{FB} from 5 to 10 and 5 to 100, respectively. The last row of the table shows the best performance with improvement over the QL. In the case of TREC CDS, the CBEEM with a few highly relevant documents and approximately 25 relevant words is the best configuration. Meanwhile, in the case of OHSUMED and CLEF eHealth, fruitful words from additional documents seem to make the CBEEM more powerful. Compared with the performance for CLEF eHealth in Table 3, we achieved significant improvements by changing D_{FB} and W_{FB}.

Table 5 shows the results of varying λ_E, a smoothing parameter for estimating RM, for individual collections with the CBEEM. The best performances for TREC CDS and OHSUMED are obtained with $\lambda_E = 0.5$. Similarly, they are obtained with $\lambda_E = 0.75$ for CLEF eHealth. It indicates that the balanced use of external collections can contributed to performance improvements while relying too much on the external collections results in poor performance.

Table 5
Performance comparison when varying λ_E with the CBEEM ($D_{FB} = 5, W_{FB} = 25$, $\lambda_{FB} = 0.5$).

λ_E	TREC CDS		OHSUMED		CLEF eHealth	
	MAP	NDCG	MAP	NDCG	MAP	NDCG
0	0.4026	0.5270	0.1662	0.2420	0.8276	0.8597
0.25	0.4191	0.5535	0.1765	0.2599	0.8402	0.8720
0.50	**0.4641**	**0.5957**	**0.1934**	**0.2774**	0.8478	0.8776
0.75	0.4312	0.5591	0.1873	0.2707	**0.8621**	**0.8843**
1	0.1927	0.3229	0.1275	0.2148	0.4807	0.6488
Max	*0.4641 (15.19%)	*0.5957 (10.32%)	**0.1934 (15.51%)	**0.2774 (12.33%)	0.8621 (1.95%)	*0.8843 (2.42%)

Performances marked with * and ** are statistically significant on two-sided paired t-test with $p \leqslant 0.05$ and $p \leqslant 0.01$.

Table 6
Performance comparison by varying λ_{FB} wih CBEEM ($D_{FB} = 5, W_{FB} = 25, \lambda_E = 0.5$).

λ_{FB}	TREC CDS		OHSUMED		CLEF eHealth	
	MAP	NDCG	MAP	NDCG	MAP	NDCG
0	0.3985	0.5363	0.1860	0.2692	**0.8563**	0.8761
0.25	0.4082	0.5468	0.1838	0.2687	0.8536	0.8726
0.50	**0.4641**	**0.5957**	**0.1934**	**0.2774**	0.8478	**0.8776**
0.75	0.3915	0.5458	0.1881	0.2674	0.8082	0.8517
1	0.3471	0.4947	0.1605	0.2339	0.6204	0.6948
Max	*0.4641 (15.19%)	*0.5957 (10.32%)	**0.1934 (15.51%)	**0.2774 (12.33%)	0.8563 (1.28%)	0.8776 (1.68%)

Performances marked with * and ** are statistically significant on two-sided paired t-test with $p \leqslant 0.05$ and $p \leqslant 0.01$.

Table 7
Performance comparison by varying N with CBEEM ($D_{FB} = 5, W_{FB} = 25$, $\lambda_{FB} = 0.5, \lambda_E = 0.5$).

N	TREC CDS		OHSUMED		CLEF eHealth	
	MAP	NDCG	MAP	NDCG	MAP	NDCG
100	**0.4641**	**0.5957**	0.1934	**0.2774**	0.8478	0.8776
200	0.4623	0.5949	**0.1935**	0.2773	0.8470	0.8773
300	0.4611	0.5941	0.1934	0.2772	0.8470	0.8773
400	0.4612	0.5942	0.1921	0.2770	**0.8489**	**0.8782**
500	0.4612	0.5942	0.1921	0.2771	**0.8489**	**0.8782**
Max	*0.4641 (15.19%)	*0.5957 (10.32%)	**0.1935 (15.56%)	**0.2774 (12.33%)	0.8489 (0.42%)	0.8782 (1.74%)

Performances marked with * and ** are statistically significant on two-sided paired t-test with $p \leqslant 0.05$ and $p \leqslant 0.01$.

In Table 6, we can see the impacts of λ_{FB} which determines the amount of CBEEM in a new query model. CBEEM is dominantly used in a new query model as λ_{FB} goes to 1. The best performances for TREC CDS and OHSUMED are obtained with $\lambda_{FB} = 0.5$. For CLEF eHealth, they are achieved with $\lambda_{FB} = 0$ in MAP and $\lambda_{FB} = 0.5$ in NDCG, respectively. Interestingly, the best performance in MAP is found without query expansion ($\lambda_{FB} = 0$). According to Tables 5 and 6, we can infer that modifying λ_E with a balanced λ_{FB} could be a good strategy to obtain reliable performances.

Table 7 presents the effects of increasing the number of initial documents (i.e., N). The experiments were done under the

Table 8
Performance comparison using Wikipedia as an external collection in CBEEM ($D_{FB} = 5, W_{FB} = 25, \lambda_{FB} = 0.5, \lambda_E = 0.5, N = 100$).

	TREC CDS		OHSUMED		CLEF eHealth	
	MAP	NDCG	MAP	NDCG	MAP	NDCG
CBEEM W/O Wikipedia	**0.4641**	**0.5957**	0.1934	0.2774	0.8478	0.8776
CBEEM W/Wikipedia	0.4294 (−8.08%)	0.5712 (−4.29%)	**0.1965 (1.58%)**	**0.2852 (2.73%)**	**0.8499 (0.25%)**	**0.8807 (0.35%)**

Table 9
Best performances for $N = 100$ in NDCG using CBEEM.

	TREC CDS		OHSUMED		CLEF eHealth	
	MAP	NDCG	MAP	NDCG	MAP	NDCG
CBEEM	*0.4641 (15.19%)	*0.5957 (10.32%)	**0.1965 (16.84%)	**0.2852 (14.73%)	0.8609 (1.81%)	*0.8897 (3.01%)
D_{FB}	5				50	
W_{FB}	25				50	
λ_{FB}	0.5					
λ_E.	0.5					
Wikipedia	not used		Used		Not used	

Performances marked with * and ** are statistically significant on two-sided paired t-test with $p \leqslant 0.05$ and $p \leqslant 0.01$.

parameter setting of $D_{FB} = 5, W_{FB} = 25, \lambda_{FB} = 0.5$, and $\lambda_E = 0.5$. The best performances are obtained for TREC CDS and OHSUMED with $100 \leqslant N \leqslant 200$ while they are obtained for CLEF eHealth with $400 \leqslant N \leqslant 500$. This result indicates that the performances using CBEEM is not sensitive to the size of the initial search results. According to our analysis, increasing N includes more numbers of relevant documents for all target collections. However, re-ranking with CBEEM cannot correctly assign higher ranks for those documents.

Through the above experiments, we believe that CBEEM works well on the three collections for medical IR. However, the question, "Will CBEEM perform well when Wikipedia is employed?" remains because Wikipedia is a well-known knowledge resource which showed moderate performance improvements in various IR tasks. To evaluate the usefulness of Wikipedia as an external collection, experiments were performed shown in Table 8. We downloaded English XML Wikipedia dump.[7] The text of articles was indexed after removing Wikipedia specific tags.

It shows the results with and without Wikipedia. Improvements over CBEEM without Wikipedia were calculated. Surprisingly, using Wikipedia does not guarantee improved performance. We obtained small improvements using Wikipedia compared to those results without Wikipedia for OHSUMED and CLEF eHealth while there were relatively large performance drops for TREC CDS. Thus, we believe that the information from Wikipedia articles is less useful compared with TREC CDS because articles in TREC CDS are written by biomedical experts.

Table 9 presents the best performances for $N = 100$ in NDCG using CBEEM. All the best performances are obtained with $\lambda_{FB} = 0.5$ and $\lambda_E = 0.5$ for the three collections. This result suggests that the balanced use of external collections in a new query model is important. Unlike Table 8, OHSUMED only benefits from Wikipedia even when considering other parameters.

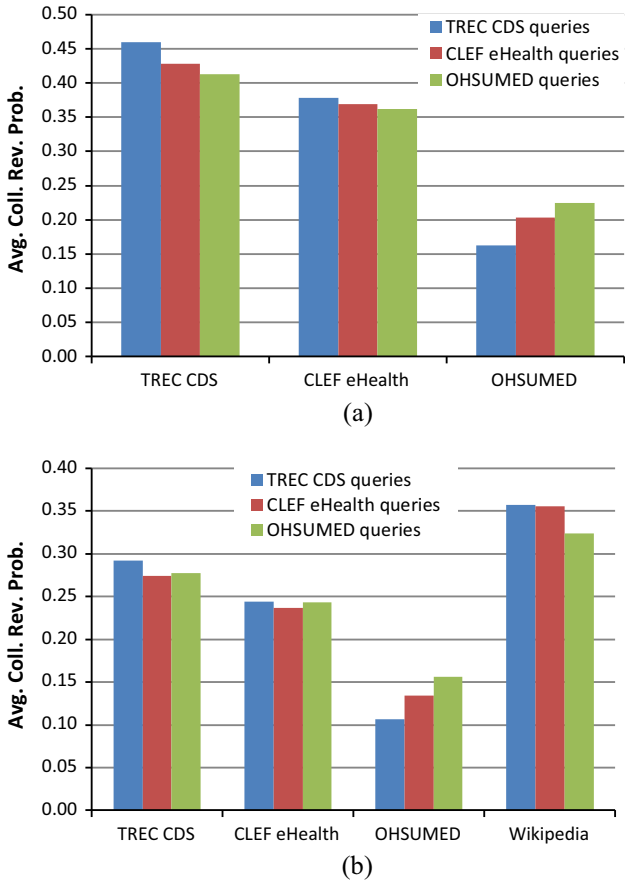

Fig. 1. Averages of $P(Q|\theta_C)$ for the target collections (a) three biomedical collections are used (b) four collections which included Wikipedia are used.

5.5. Discussion

From the results of Tables 4–7, we have confirmed that the CBEEM can be optimized further by varying key parameters (i.e., $\lambda_{FB}, \lambda_E, D_{FB}$, and W_{FB}) for query expansion according to external collections. In the cases of TREC CDS and OHSUMED, they had similar performance tendencies due to their innate homogeneity, consisting of biomedical literature, despite the different lengths of text. Based on the results, we noticed that the balanced use of external collections during query expansion is effective for three different test collections with a common basis. In addition, the strategy of increasing the number of documents in the initial search does not always lead to performance improvements.

The key idea of utilizing external collections based on the EEM is to determine the query dependent collection relevance $P(Q|\theta_C)$. Here, we report the role of $P(Q|\theta_C)$ in medical IR. Fig. 1 shows

[7] The dump file was enwiki-20140614-pages-articles.xml.bz2 from https://meta.wikimedia.org/wiki/Data_dump_torrents#enwiki.

the averages of the $P(Q|\theta_C)$ estimated in all the experiments. In Fig. 1a, the three biomedical collections are used to build the CBEEM. Generally, we can expect that $P(Q|\theta_C)$ for a target collection is higher than those for the other collections because a query is developed for the target collection. Interestingly, $P(Q|\theta_{C_{CDS}})$ was higher than $P(Q|\theta_{C_{eHealth}})$ and $P(Q|\theta_{C_{OHSUMED}})$ not only for the TREC CDS queries but also for the CLEF eHealth and OHSUMED queries. Namely, the CBEEM contains a large proportion of information from the TREC CDS collection regardless of the target collection.

It should be noted that TREC CDS contains reliable expertise compared with CLEF eHealth and OHUSMED because it consists of biomedical literature with full-text journal articles. Therefore, most queries produce documents with high relevance scores via retrieval. For test queries for OHSUMED, the CBEEM reflects a large amount of information from TREC CDS and CLEF eHealth rather than from OHSUMED which is a target collection because $(Q|\theta_{C_{OHSUMED}})$ is lower than the others.

In Fig. 1b, the averages of $P(Q|\theta_C)$ in our experiments are shown. Compared with Fig. 1a, $P(Q|\theta_{C_{WIKI}})$ has the highest value regardless of the target collection. This could be one reason which prevents the effective use of Wikipedia because the CBEEM can include much information from Wikipedia articles that are irrelevant to the biomedical domain.

Table 10 shows examples of expanding a query model with the CBEEM for TREC CDS, CLEF eHealth, and OHSUMED using the queries described in Table 2. Words in bold that appear in the CBEEM do not appear in the original query model. For TREC CDS, the original query model $P(w|\theta_Q)$ is generated from the 17th query. In $P_{CBEEM}(w|Q)$, there are several new words not presented in $P(w|\theta_Q)$. The new words such as *treatment*, *health*, and *patient* are relevant to general medical words. They help to express the information need of the query because medical queries commonly expect to find a treatment for a patient or disease even though it is not explicitly expressed using the general medical words in the query. As a result, $P(w|\theta'_Q)$ includes the new words with some probability mass while the probabilities of the existing words in $P(w|\theta_Q)$ are decreased. In the query text, there are a number of numbers and abbreviations. Numbers have an important and specific meaning. However, their importance and meaning are likely to be diminished in a new query model after normalization and expansion. To improve retrieval performance, we believe that it is essential to devise a proper normalization method and scoring function that can deal with numerical words effectively. Similarly, abbreviation resolution is important to understand the intent of the query. We can expect that relevant documents are highly ranked if abbreviations are resolved to the full form such as HR (heart rate), bpm (beats per minute), BP (Blood pressure), and ICU (Intensive Care Unit). In case of CLEF eHealth and OHSUMED, the query text is relatively short compared to TREC CDS. Meanwhile, the intent of the CLEF eHealth query is more general compared to the others because it was developed for laypeople.

In addition, to help readers' understanding, we provide our discussion on prior results using the three collections as follows.

(1) The OHSUMED collection has been widely used for evaluating various types of ranking models [6,19,45]. A recent learning-to-rank approach [6] with query-specific feedback model reported MAP@20(=0.37) and meanNDCG@20 (=0.42) using the OHSUMED. Although it is difficult to directly compare performances with the model due to the differences in evaluation settings, this suggests that the state-of-the-art approach has an advance in performance over ours where MAP@20(=0.1965) and NDCG@20 (=0.2852) with the CBEEM. We just confirmed the moderate effects of the CBEEM within our experimental framework.

Table 10
Examples of query expansion using the CBEEM for (a) TREC CDS, (b) CLEF eHealth, (c) OHSUMED. (Queries are shown in Table 2).

(a) Collection	TREC CDS
Query text	A 48-year-old white male with history of common variable immunodeficiency (CVID) with acute abdominal pain, fever, dehydration, HR of 132 bpm, BP 80/40. The physical examination is remarkable for tenderness and positive Murphy sign. Abdominal ultrasound shows hepatomegaly and abundant free intraperitoneal fluid. Exploratory laparotomy reveals a ruptured liver abscess, which is then surgically drained. After surgery, the patient is taken to the ICU
$P(w\|\theta_Q)$	abdomin:0.04255 murphi:0.02128 exploratori:0.02128 reveal:0.02128 dehydr:0.02128 sign:0.02128 show:0.02128 fever:0.02128 hepatomegali:0.02128 ruptur:0.02128 hr:0.02128 remark:0.02128 bp:0.02128 ...
$P_{CBEEM}(w\|Q)$	**patient**:0.05639 **cell**:0.04221 **studi**:0.03795 **case**:0.02273 **treatment**:0.02262 **health**:0.02217 result:0.01996 effect:0.01932 activ:0.01919 diseas:0.01915 show:0.0187 increas:0.01818 group:0.01805 al:0.01761 clinic:0.01758 time:0.01728 differ:0.01711 inform:0.01706 report:0.017 ...
$P(w\|\theta'_Q)$	**patient**:0.03884 abdomin:0.02128 **cell**:0.02111 show:0.01999 **studi**:0.01898 **case**:0.01137 **treatment**:0.01131 **health**:0.01109 exploratori:0.01064 reveal:0.01064 dehydr:0.01064 sign:0.01064 hr:0.01064 surgic:0.01064...
(b) CLEF eHealth	
Query text	Coronary artery disease What does coronary artery disease mean?
$P(w\|\theta_Q)$	coronari:0.28571 arteri:0.28571 diseas:0.28571 mean:0.14286
$P_{CBEEM}(w\|Q)$	coronari:0.06603 arteri:0.06484 **patient**:0.05493 diseas:0.04634 **studi**:0.03413 **cell**:0.03217 **heart**:0.02238 **health**:0.01973 **group**:0.0187 **al**:0.01836 **treatment**:0.01771 **p**:0.01676 **result**:0.01651 **differ**:0.01638 **activ**:0.01588 **control**:0.01561 **effect**:0.01529 **increas**:0.01529 ...
$P(w\|\theta'_Q)$	coronari:0.17587 arteri:0.17528 diseas:0.16603 mean:0.07143 **patient**:0.02746 **studi**:0.01706 **cell**:0.01608 **heart**:0.01119 **health**:0.00986 **group**:0.00935 **al**:0.00918 **treatment**:0.00885 **p**:0.00838 **result**:0.00825 **differ**:0.00819 **activ**:0.00794 **control**:0.00781 **effect**:0.00765 **increas**:0.00765 ...
(c) OHSUMED	
Query text	68 y.o. m. with adult-onset diabetes mellitus noted to have thrombocytosis
$P(w\|\theta_Q)$	note:0.1 thrombocytosi:0.1 mellitu:0.1 y:0.1 adult:0.1 onset:0.1 diabet:0.1 m:0.1 o:0.1 ...
$P_{CBEEM}(w\|Q)$	diabet:0.05825 **patient**:0.05512 **cell**:0.03703 **studi**:0.03677 **al**:0.02766 **type**:0.02557 **group**:0.02218 **clinic**:0.02143 **diseas**:0.01978 **gene**:0.01955 **health**:0.0186 mellitu:0.01806 **treatment**:0.01749 **level**:0.01707 **p**:0.01686 **activ**:0.0167 **factor**:0.01663 **result**:0.01659 **associ**:0.01651 ...
$P(w\|\theta'_Q)$	diabet:0.07913 mellitu:0.05903 note:0.05 thrombocytosi:0.05 onset:0.05 m:0.05 o:0.05 y:0.05 adult:0.05 **patient**:0.02756 **cell**:0.01851 **studi**:0.01838 **al**:0.01383 **type**:0.01278 **group**:0.01109 **clinic**:0.01071 **diseas**:0.00989 **gene**:0.00978 **health**:0.0093 **treatment**:0.00875 **level**:0.00853 **p**:0.00843 **activ**:0.00835 **factor**:0.00832 **result**:0.0083 **associ**:0.00826 ...

Table 11
Performance comparison with the best performance in CLEF eHealth 2014.

	P@5	P@10	NDCG@5	NDCG@10
GRIUM [GRIUM EN Run.5]	0.7680	0.7560	0.7423	0.7445
SNUMedinfo [SNUMEDINFO EN Run.5]	0.8160	0.7520	0.7749	0.7426
CBEEM	**0.8320**	**0.8080**	**0.8693**	**0.8790**

(2) In case of CLEF 2014 eHealth, the results of 14 participants in Task 3 are summarized in [16]. Among them, two groups showed outstanding performance. GRIUM [37] achieved the best performance in the task, 0.7560 and 0.7445 in P@10 and NDCG@10, respectively. Interestingly, SNUMedinfo [10] produced the best performance, 0.8160 and 0.7749 in both P@5 and NDCG@5, respectively. Although the query expansion methods used by the two groups are different, they both utilized the UMLS and MetaMap as external resources. We performed additional experiments by adjusting K to compare the CBEEM with the prior results in CLEF eHealth. As in Table 11, the best performing CBEEM outperformed the prior best performance in four measures.

(3) The overall retrieval scores obtained through TREC CDS 2014 [38] were relatively low due to the difficulty of the task. The best results in two different measures were obtained by different participants. In terms of precision measure, NovaSearch [39] obtained the best performance (0.3900) in P@10 using query expansion with MeSH, ranking fusion with multiple retrieval models, and PRF. Meanwhile, the best result (0.2674) in infNDCG was achieved by SNUMedinfo [11] which used query expansion with MEDLINE and a type-specific classifier. Compared with the two teams's results, the CBEEM where MAP@20(=0.1965) and NDCG@20 (=0.2852) seems to outperform them by a large margin.

6. Conclusion

This paper presented a cluster-based expansion model (CBEEM) for leveraging the effectiveness of external collections. According to our exhaustive evaluation of three popular biomedical collections (TREC CDS, CLEF eHealth, and OHSUMED) in terms of re-ranking, we demonstrated the superiority of our method compared with a representative expansion approach aimed at using external collections. In addition, we showed that using Wikipedia as an external collection in medical IR does not guarantee performance improvements because it is possible to include non-relevant non-biomedical information in a feedback model.

To improve performance in medical IR, we plan to develop advanced ranking methods from two aspects. Dealing with number normalization and abbreviation resolution without degrading their importance is our first concern. Second, we plan to develop an effective way to use Wikipedia in medical IR because the straightforward adoption of Wikipedia without any concern for that knowledge to be in the medical domain was not so successful.

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] G. Amati, Probabilistic models of information retrieval based on measuring the divergence from randomness, ACM Trans. Inform. 20 (4) (2002) 357–389.
[2] A.R. Aronson, Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program, Proc. AMIA Sympos. (2001) 17–21.
[3] A. Babashzadeh, J. Huang, M. Daoud, Exploiting semantics for improving clinical information retrieval, in: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR '13, ACM Press, 2013, pp. 801–804.
[4] R. Blanco, A. Barreiro, Probabilistic document length priors for language models, in: 30th European Conference on Advances in Information Retrieval, 2008, pp. 394–405.
[5] O. Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology, Nucleic Acids Research 32, Database issue (2004), 267–270.
[6] E.F. Can, W.B. Croft, R. Manmatha, Incorporating query-specific feedback into learning-to-rank models, in: Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval – SIGIR '14, 2014, pp. 1035–1038.
[7] C. Carpineto, G.A. Romano, Survey of automatic query expansion in information retrieval, ACM Comput. Surv. 44 (1) (2012) 1–50.
[8] L. Casebeer, N. Bennett, R. Kristofco, A. Carillo, R. Centor, Physician Internet medical information seeking and on-line continuing education use patterns, J. Contin. Educ. Health Professions 22 (1) (2002) 33–42.
[9] S. Choi, J. Choi, S. Yoo, H. Kim, Y. Lee, Semantic concept-enriched dependence model for medical information retrieval, J. Biomed. Inform. 47 (2014) 18–27.
[10] S. Choi, J. Choi, Exploring effective information retrieval technique for the medical web documents: SNUMedinfo at. CLEFeHealth2014 Task 3, in: Proceedings of CLEF 2014, 2014, pp. 167–175.
[11] S. Choi, J. Choi, SNUMedinfo at TREC CDS track 2014: medical case-based retrieval task, in: Proceedings of The Twenty-Third Text REtrieval Conference, TREC, 2014.
[12] R. Deveaud, E. SanJuan, P. Bellot, Estimating topical context by diverging from external resources, in: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR '13, ACM Press, 2013, pp. 1001–1004.
[13] F. Diaz, D. Metzler, Improving the estimation of relevance models using large external corpora, in: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR '06, ACM Press, 2006, p. 154.
[14] A. Dong, Y. Chang, Z. Zheng, et al., Towards recency ranking in web search, Proceedings of the third ACM International Conference on Web Search and Data Mining – WSDM '10, ACM Press, 2010, p. 11.
[15] O. Egozi, S. Markovitch, E. Gabrilovich, Concept-based information retrieval using explicit semantic analysis, ACM Trans. Inform. Syst. 29 (2) (2011) 1–34.
[16] L. Goeuriot, L. Kelly, W. Li, et al., ShARe/CLEF eHealth evaluation lab 2014, Task 3: user-centred health information retrieval, Proc. CLEF 2014 (2014).
[17] W. Hersh, C. Buckley, T.J. Leone, D. Hickam, OHSUMED: an interactive retrieval evaluation and new large test collection for research, in: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR '94, 1994, pp. 192–201.
[18] J. Kamps, M. de Rijke, B. Sigurbjörnsson, Length normalization in XML retrieval, Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval – SIGIR '04, ACM Press, 2004, p. 80.
[19] M. Karimzadehgan, W. Li, R. Zhang, J. Mao, A stochastic learning-to-rank algorithm and its application to contextual advertising, in: Proceedings of the 20th International Conference on World Wide Web, 2011, pp. 377–386.
[20] M. Karimzadehgan, C. Zhai, Estimation of statistical translation models based on mutual information for ad hoc information retrieval, Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR '10, ACM Press, 2010, pp. 323–330.
[21] L. Kelly, L. Goeuriot, H. Suominen, et al., Overview of the ShARe/CLEF eHealth Evaluation Lab 2014, Proceedings of CLEF 2014, Springer, 2014.
[22] O. Kurland, L. Lee, PageRank without hyperlinks: structural re-ranking using links induced by language models, Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR '05, ACM Press, 2006, pp. 306–313.
[23] J. Lafferty, C. Zhai, Document language models, query models, and risk minimization for information retrieval, Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR '01, ACM Press, 2001, pp. 111–119.
[24] V. Lavrenko, W.B. Croft, Relevance based language models, Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR '01, ACM Press, 2001, pp. 120–127.
[25] K.S. Lee, W.B. Croft, J. Allan, A cluster-based resampling method for pseudo-relevance feedback, Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR '08, ACM Press, 235–242, 2008, pp. 235–242.
[26] N. Limsopatham, C. Macdonald, I. Ounis, Modelling relevance towards multiple inclusion criteria when ranking patients, Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management – CIKM '14, ACM Press, 2014, pp. 1639–1648.
[27] J. Lin, D. Demner-Fushman, The role of knowledge in conceptual retrieval, Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR '06, ACM Press, 2006, pp. 99–106.
[28] X. Liu, W.B. Croft, Cluster-based retrieval using language models, Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval – SIGIR '04, ACM Press, 2004, pp. 186–193.
[29] Y. Lv, C. Zhai, W. Chen, A boosting approach to improving pseudo-relevance feedback, Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information – SIGIR '11, ACM Press, 2011, pp. 165–174.

[30] Y. Lv, C. Zhai, Positional relevance model for pseudo-relevance feedback, Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR '10, ACM Press, 2010, pp. 579–586.
[31] D. Martinez, A. Otegi, A. Soroa, E. Agirre, Improving search over electronic health records using UMLS-based query expansion through random walks, J. Biomed. Inform. 51 (2014) 100–106.
[32] Q. Mei, D. Zhang, C. Zhai, a general optimization framework for smoothing langauge model on graph structures, Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR '08, ACM Press, 2008, p. 611.
[33] D. Metzler, W.B. Croft, A Markov random field model for term dependencies, Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR '05, ACM, 2005, pp. 472–479.
[34] L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank citation ranking: bringing order to the web, Technical Report (1999) 1–17.
[35] J.M. Ponte, W.B. Croft, A language modeling approach to information retrieval, Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR '98, ACM Press, 1998, pp. 275–281.
[36] S.E. Robertson, S. Walker, S. Jones, H.M. Hancock-Beaulieu, M. Gatford, Okapi at TREC-3, in: Proceedings of the Third Text REtrieval Conference, 1994, pp. 109–126.
[37] W. Shen, J. Nie, X. Liu, X. Liu, An investigation of the effectiveness of concept-based approach in medical information retrieval, Proc. CLEF 2014 (2014) 236–247.
[38] M.S. Simpson, E.M. Voorhees, W. Hersh, Overview of the TREC 2014 clinical decision support track, in: Proceedings of Text REtrieval Conference (TREC), 2014.
[39] U.N. De Lisboa, NovaSearch at TREC 2014 clinical decision support track, in: Proceedings of Text REtrieval Conference (TREC), 2014.
[40] E.M. Voorhees, W. Hersh, Overview of the TREC 2012 medical records, in: Proceedings of Text REtrieval Conference (TREC), 2012.
[41] Y. Wang, X. Liu, H.A. Fang, study of concept-based weighting regularization for medical records search, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014, pp. 603–612.
[42] W. Weerkamp, K. Balog, M. de Rijke, Exploiting external collections for query expansion, ACM Trans. Web 6 (4) (2012) 1–29.
[43] T. Westerveld, W. Kraaij, D. Hiemstra, Retrieving web pages using content, links, urls and anchors, in: Proceedings of Text REtrieval Conference (TREC), 2001.
[44] Y. Xu, G.J.F. Jones, B. Wang, Query dependent pseudo-relevance feedback based on wikipedia, Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR '09, ACM Press, 2009, pp. 59–66.
[45] S. Yoo, J. Choi, Evaluation of term ranking algorithms for pseudo-relevance feedback in MEDLINE retrieval, Healthcare Inform. Res. 17 (2) (2011) 120–130.
[46] C. Zhai, J. Lafferty, Two-stage language models for information retrieval, Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR '02, ACM Press, 2002, pp. 49–56.
[47] D. Zhu, S. Wu, B. Carterette, H. Liu, Using large clinical corpora for query expansion in text-based cohort identification, J. Biomed. Inform. 49 (2014) 275–281.