MEDINFO 2010
C. Safran et al. (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-588-4-1174

Clinical Task-Specific Query Expansion for the Retrieval of Scientifically Rigorous Research Documents

Sooyoung Yoo[a], Jinwook Choi[b], Sungbin Choi[b]

[a] *Medical Information Center, Seoul National University Bundang Hospital, Gyeonggi-do, South Korea*
[b] *Medical Informatics Lab, Dept. of Biomedical Engineering, Seoul National University, Seoul, South Korea*

Abstract

To support the practice of evidence-based medicine (EBM), clinically relevant and scientifically sound articles should be easily accessible. Due to the huge volume of medical literature and the low performance of present retrieval models, clinicians could only get relevant documents in the order of publication time. This study propose a new clinical task-specific retrieval technique that improves retrieval accuracy by exploiting clinical task-specific EBM terms to query expansion using co-occurrence analysis technique. The idea is aimed at selecting query expansion terms that are relevant to a specific clinical-task using task-specific EBM terms. Focusing on treatment and diagnosis tasks, the new method which was performed on the OHSUMED collection showed a further improved result than the existing method.

Keywords:
Evidence-based medicine, Clinical task-specific query expansion, Local context analysis

Introduction

Evidence-based medicine (EBM) is the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients[1]. The exponential increase of the volume of medical literature requires the development of effective information retrieval strategies. Clinically relevant and scientifically sound articles (i.e., high-quality) should be accessible in a fast and easy manner, especially to support the practice of EBM.

In terms of clinical activity there are four main clinical task categories: treatment, diagnosis, etiology, and prognosis. EBM defines methodological criteria for the task categories in order to identify high-quality articles. For example, methodological criteria for treatment task is defined as "*random allocation of participants to comparison groups; Outcome assessment of at least 80% of those entering the investigation; analysis consistent with study design*" [2]. Based on the definition the main task of the information retrieval strategy is to search articles that meet the criteria for each clinical task.

Aiming at retrieving ranked relevant articles of high quality for a given clinical query, this paper describes a new information retrieval technique that uses EBM-related terms to improve retrieval accuracy in a clinical task-specific ranking system. Focusing on a treatment and diagnosis task, we suggest a new query expansion strategy based on co-occurrence analysis with clinical task-specific EBM terms as well as all query terms to restrict query expansion to terms that are relevant to a given clinical task in the principle of EBM.

Related Work

There have been different approaches to clinical task-specific retrieval.

Haynes and colleagues [2-7] developed optimal MEDLINE search strategies, called clinical query filters targetting four clinical task areas: treatment, diagnosis, etiology, and prognosis by validating diagnostic tests using proposed search terms and their manually-constructed gold standard. The filters were Boolean query strategies optimized for sensitivity and specificity that were added to the original user query. They were adopted by the U.S. NLM strategies for use in the Clinical Queries feature [8] in PubMed. However, since the query filter is Boolean form, it still retrieves thousands of articles from MEDLINE when a common clinical term is given as a user query ("*breast cancer*" for treatment task category, for example). It displays too many or too small number of articles depending on the user query. It also does not support ranking of the retrieved articles according to relevance to the user query.

Chu and colleagues [9, 10] proposed a knowledge-based query expansion to support clinical task-specific retrieval. In their retrieval system, the authors tried to expand the original user query with additional terms that are specifically relevant to the query's task using domain knowledge source such as UMLS Metathesaurus and semantic structure. Focusing on five types of tasks such as treatment, diagnosis, prevention, cause, and indication, they evaluated their approach on the OHSUMED test collection for a subset of 40 queries explicitly mentioning the tasks in the OHSUMED. Comparison of their approach with no-expansion and statistical expansion approach based on a co-occurrence thesaurus showed the effectiveness of their approach over the traditional approaches. Rather than using a knowledge source, we exploit clinical task-specific EBM terms to restrict query expansion to task-specific terms in the principle of EBM.

Recently, some researchers dealt with clinical task-specific retrieval using a text classification technique.

Aphinyanaphongs et al. [11] applied machine learning techniques to identify high-quality articles in internal medicine for the areas of treatment, diagnosis, etiology, and prognosis. Using inclusion or citations by the *ACP Journal Club*[12] for one specific time period as a gold standard, the authors constructed test corpus and automatically built machine learning models. They found support vector machine (SVM) classifier shows the best performance and machine learning techniques have better or comparable performance than the 1994 PubMed clinical query filters [13]. The results obtained by Aphinyanaphongs et al. [11] were tested for the generalizability to other gold standard used in the development of PubMed clinical query filters by the work of Kilicoglu et al. [14]. The authors confirmed that machine learning approaches can be used to recognize high-quality articles.

Materials and Methods

Text Corpus

We used OHSUMED [15] as a test collection. It is a subset of the MEDLINE database. It consists of 348,566 MEDLINE references from 1987 to 1991, and 106 topics (queries) generated by actual physicians in the course of patient care. Relevance judgments to each query are provided, with the scale of 'definitely relevant', 'possibly relevant', and 'not relevant'. In this study, we limit relevant documents to those judged as 'definitely relevant' to retrieve high-relevant documents.

For the new clinical task-specific retrieval, we reviewed the OHSUMED queries and selected 60 treatment-specific queries and 26 diagnosis-specific queries according to the definition of treatment and diagnosis task [2]. Among the 60 treatment-specific queries, we use 57 queries with at least one definitely relevant document for our treatment-specific experiments (see Table 1).

Table 1 - Queries used for our experiments

Task	Query IDs
Treatment	1,2,5,10,13,15,16,18,19,22,24,27,29,30,31,32,33,35,37,38,39,40,42,43,45,52,53,56,57,58,60,61,62,63,64,67,69,71,72,73,74,75,76,77,78,79,81,84,85,88,89,91,94,98,100,102,104
Diagnosis	14,15,21,23,37,41,43,47,51,53,57,58,65,69,70,72,74,76,80,81,82,92,97,99,101,103

Text Representation

For the document representation, MeSH, title and abstract fields of each MEDLINE reference are used. For the index generation, we parse the three fields from each MEDLINE document as follows. First, all the texts in each field are tokenized into single words. Each word is then processed by the removal of stopwords identified by SMART stopwords. It is further stemmed by the Lovins' stemmer [16] and is case-folded. Finally, all single-stemmed terms are indexed in the form of inverted file.

A user query is represented by using *information need* field from OHSUMED queries since it is the most likely user queries issued in the information retrieval system, and is processed by the same text processing method mentioned above.

Document Ranking Model

As our baseline retrieval model for ranking retrieved documents according to relevance to the query, we implemented the well-known Okapi BM25 weighting scheme [17], which is a highly effective retrieval formula that represents the classic probabilistic retrieval model [18, 19].

In the Okapi BM25 formula, the top-ranked documents are retrieved by computing a measurement of similarity between a query, q, and a document, d, as follows:

$$sim(q,d) = \sum_{t \in q \wedge d} w_{d,t} \cdot w_{q,t} \quad (1)$$

$$w_{d,t} = \frac{(k_1+1)\cdot f_{d,t}}{K + f_{d,t}} \quad (2)$$

$$w_{q,t} = \frac{(k_3+1)\cdot f_{q,t}}{k_3 + f_{q,t}} \cdot \log\frac{N-n+0.5}{n+0.5} \quad (3)$$

where t is a term of the query q, $w_{d,t}$ is the weight of the term t in the document d, $w_{q,t}$ is the weight of the term t in the query q, n is the number of documents containing the term t across the document collection that contains N documents, $f_{d,t}$ is the frequency of the term t in the document d and $f_{q,t}$ is the frequency of the term t in the query q. K is $k_1((1-b) + b \cdot dl/avdl)$. k_1, b, and k_3 are tuning parameters set to 1.2, 0.75, and 1,000, respectively, by default. We use the default setting in this study. Document length and average document length, *dl* and *avdl* respectively, are measured in suitable units such as the number of terms or the number of bytes (in this study byte length is used).

Clinical Task-Specific Query Expansion by Exploiting EBM Terms

To retrieve topically relevant documents that pertain to a specific clinical-task and contain clinical evidence of high-quality, we use clinical task-specific EBM terms.

We focus on treatment and diagnosis task for this study. By employing and evaluating each task-specific EBM terms used in the PubMed Clinical Queries and their combinations in our preliminary experimentation, we define task-specific EBM terms *TASK-EBM* as follows.

Treatment: "clinical trials therapeutic"

Diagnosis: "sensitivity specificity diagnosis diagnostic"

The EBM terms are utilized following two ways in this study.

Method 1. Simple expansion with EBM terms (*SE-EBM*): The simplest way to use the EBM terms is to add these terms to the

original query before the query is submitted to the system. We refer this approach to *SE-EBM*.

Specifically, given a query q for a specific clinical- task such as treatment or diagnosis, input query to the system is built by appending the query q with the *TASK-EBM* terms of the clinical task that are not found in the query, and then submitted to the system. For the input query, the ranked retrieved documents are returned according to the score of Okapi BM25.

Method 2. Co-occurrence analysis with EBM terms (*CO-EBM*): By extending the local context analysis (*LCA*) [20], we propose a new query expansion approach to expand query with additional task-specific terms using task-specific EBM terms.

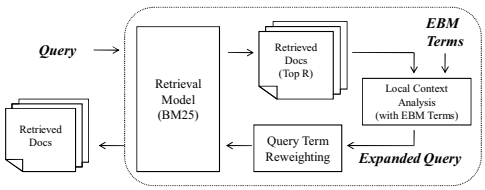

Figure 1 - Co-occurrence analysis with EBM terms

Given a query q for a specific clinical-task, our new query expansion is performed as follows. Let R be the number of pseudo-relevant document, and let E be the number of expansion terms that are appended to the original queries. First, query q is submitted to the system and documents are retrieved based on the Okapi BM25 for the given query. Then, the top-ranked R documents retrieved for the query q are assumed to be relevant and are used as a source of expansion terms. After extracting and merging all indexing terms from the top-ranked R documents, each term t is scored by:

$$score(t) = \prod_{k_i \in q \cup TASK-EBM} \left(\delta + \frac{\log_{10}\left(f(t,k_i)+1\right) \times idf_t}{\log_{10} R} \right)^{idf_i} \quad (4)$$

where the function $f(t,k_i)$ measures the degree of co-occurrence of term t with term k_i in the query q and the *TASK-EBM* terms of the clinical task, and idf_t and idf_i are inverse document frequency of term t and term k_i , respectively. Each factor is calculated by:

$$f(t,k_i) = \sum_{j=1}^{R} tf_{t,j} \cdot tf_{i,j} \quad (5)$$

$$idf_t = \min\left(1, \frac{\log_{10} N/n_t}{5}\right) \quad (6)$$

$$idf_i = \min\left(1, \frac{\log_{10} N/n_i}{5}\right) \quad (7)$$

where $tf_{t,j}$ is the frequency of term t in document j, $tf_{i,j}$ is the frequency of term k_i in document j, N is the number of documents in the collection, n_t is the number of documents in the collection containing term t, and n_i is the number of documents in the collection containing term k_i. The tuning parameter δ is set to default 0.1. The *TASK-EBM* terms are employed to reflect co-occurrence degree with those terms. All terms are then sorted in descending order by their scores. Lastly, the highly scored E terms are selected and added to the query q. The expanded query is automatically submitted to the system to get the final results. We refer this approach to *CO-EBM*. It is compared with the existing *LCA* method based on co-occurrence analysis with only query terms.

Query Term Reweighting

The expanded query produced by *CO-EBM* method is reweighted in the second-pass retrieval.

The standard Rocchio's feedback formula is a commonly used for (pseudo-) relevance feedback and term reweighting. It reweights terms in the expanded query by adding the weights from the actual occurrence of those query terms in the relevant documents and subtracting the weights of those terms occurring in the non-relevant documents [21]. In this study, we modify the formula so that all expansion terms are given the same weight for fair comparison of different query expansion approaches. Based on the positive feedback form of the standard Rocchio feedback formula, the new weight $w'_{q,t}$ of term t in the expanded query is assigned as:

$$w'_{q,t} = w_{q,t} + c \quad (8)$$

where $w_{q,t}$ is the weight of term t in the unexpanded original query q submitted to the system initially and c is a constant to give the same weight to all expansion terms (c is set to 1 in this study).

Results

We evaluate our experimental results using mean average precision (MAP), precision at given document cutoff value X (P@X). MAP is an average overall precision measurement for each relevant document in the ranking. It serves as a good measurement of the overall ranking accuracy. We measure the performance for the top-ranked 100 documents retrieved in our experiments. P@X is the percent of retrieved documents that are relevant after X documents have been retrieved. Since most users are interested in a few top-ranked documents, it is a good measurement in terms of users' perspective.

Table 2 - Performance of SE-EBM compared with no expansion for each treatment and diagnosis task.

Treatment (Average over 57 Queries)		
	No expansion	SE-EBM
MAP	0.2269	0.2175 (-4.14%)
P@5	0.3439	0.3333 (-3.08%)
P@10	0.2982	0.2772 (-7.04%)
Diagnosis (Average over 26 Queries)		
	No expansion	SE-EBM
MAP	0.2155	0.1877 (-12.9%)
P@5	0.3385	0.3462 (+2.27%)
P@10	0.3462	0.2885 (-16.67%)

Table 2 shows the performance of *SE-EBM* method compared with the one of unexpanded run for two clinical tasks of treat-

ment and diagnosis. As can be seen in the table, *SE-EBM* approach generally makes the performance decreases for both treatment and diagnosis tasks. It indicates that clinical task-specific EBM terms are not useful as additional expansion terms to retrieve task-specific relevant articles. Rather, these terms make relevant articles be retrieved in a lower rank. EBM terms would be not used as expansion terms in a ranking system.

On the other hand, the effectiveness of clinical task-specific EBM terms on improving retrieval accuracy is evaluated for selecting expansion terms. The performance of *CO-EBM* method is evaluated for a wide range of *R* (5 to100 by 5) and *E* (5 to 100 by 5) settings to see the sensitivity of the parameter settings, and is compared with *LCA* method based on co-occurrence with only query terms. Since 15 expansion terms generally provided a good performance on OHSUMED collection in our previous study [22], we present the performance of *CO-EBM* and *LCA* over a different parameter of *R* at a fixed E parameter of 15 in this paper. Figure 2 and Figure 3 display MAP percentage change of *CO-EBM* and *LCA* methods over unexpanded run for treatment and diagnosis task, respectively.

As can be seen, the maximum performance improvement is achieved using our *CO-EBM* method for both treatment and diagnosis tasks. On treatment-task experiments (Figure 2), our *CO-EBM* shows better performance than *LCA* when more than 50 documents retrieved are used for co-occurrence analysis. On diagnosis-task experiments (Figure 3), our *CO-EBM* shows better performance than LCA regardless of the number of pseudo-relevant documents used. It indicates that task-specific EBM terms can be effectively used for restricting query expansion to terms that are relevant to a given clinical task.

Conclusion

In order to support the practice of EBM, we have proposed a new information retrieval technique that exploits clinical task-specific EBM terms for the query expansion using co-occurrence analysis. Focusing on treatment and diagnosis tasks, our experimental results on the OHSUMED collection showed that our proposed method was performed best. The co-occurrence analysis with clinical task-specific EBM terms can select expansion terms more specific to a given clinical task. We believe that our method can be effectively used for clinical task-specific ranking system.

We plan to evaluate our approach for other clinical tasks including etiology and prognosis by preparing for new test collections since OHSUMED does not provide sufficient test queries for evaluation of other clinical tasks.

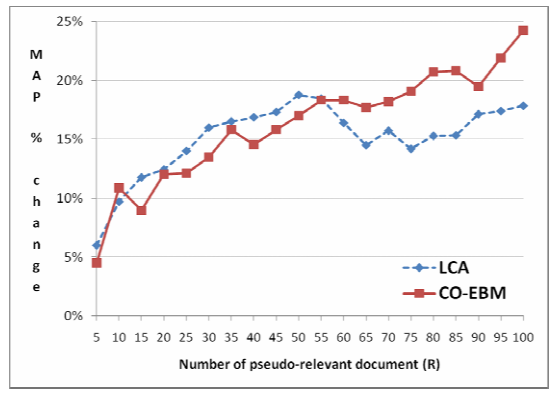

Figure 2 - Performance change of CO-EBM from no expansion compared with LCA when 15 terms are expanded using different number of pseudo-relevant documents for a treatment task.

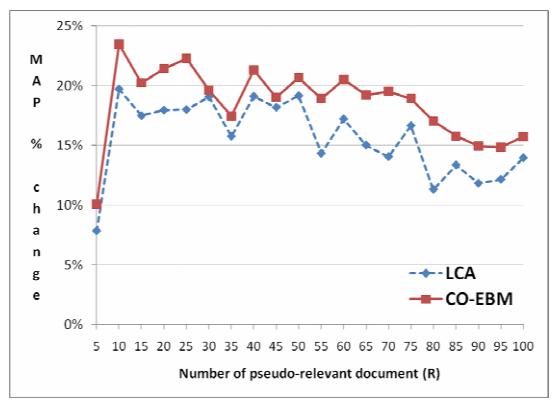

Figure 3 - Performance change of CO-EBM from no expansion compared with LCA when 15 terms are expanded using different number of pseudo-relevant documents for a diagnosis task.

Acknowledgments

This study was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MEST) (No. 2009-0075089), and in part by a grant of the Korea Healthcare technology R&D Project, Ministry for Health, Welfare & Family Affairs, Republic of Korea (A070001).

References

[1] Sackett D, Rosenberg W, Gray J, Haynes R, Richardson W. Evidence based medicine: what it is and what it isn't. BMJ. 1996;312(7023):71-2.

[2] Wilczynski N, Morgan D, Hynes R. An overview of the design and methods for retrieving high-quality studies for

clinical care. BMC Medical Informatics and Decision Making. 2005 June 2005;5(20).

[3] Haynes R, Mckibbon K, Wilczynski N, Walter S, Werre S. Optimal search strategies for retrieving Medline: analytical survey scientifically strong studies of treatment from. BMJ. 2005 May 13.

[4] Haynes RB, Wilczynski NL. Optimal search strategies for retrieving scientifically strong studies of diagnosis from Medline: analytical survey. BMJ. 2004 May 1;328(7447):1040.

[5] Wilczynski NL, Haynes RB. Developing optimal search strategies for detecting clinically sound causation studies in MEDLINE. AMIA Annu Symp Proc. 2003:719-23.

[6] Wilczynski NL, Haynes RB. Developing optimal search strategies for detecting clinically sound prognostic studies in MEDLINE: an analytic survey. BMC Med. 2004 Jun 9;2:23.

[7] Haynes RB, Wilczynski N. Finding the gold in Medline: clinical queries. Evidence-Based Medicine. 2005;10(4):101-2.

[8] PubMed Clinical Queries. Available from: http://www.ncbi.nlm.nih.gov/entrez/query/static/clinical.shtml

[9] Chu W, Liu Z, Mao W, Zou Q. A knowledge-based approach for retrieving scenario-specific medical text documents. Control Engineering Practice. 2005;13(9):1105-21.

[10] Liu Z, Chu W. Knowledge-based query expansion to support scenario-specific retrieval of medical free text. Proceedings of the 2005 ACM symposium on Applied computing. Santa Fe, New Mexico: ACM; 2005. p. 1076-83.

[11] Aphinyanaphongs Y, Tsamardinos I, Statnikov A, Hardin D, Aliferis CF. Text categorization models for high-quality article retrieval in internal medicine. J Am Med Inform Assoc. 2005 Mar-Apr;12(2):207-16.

[12] ACP Journal Club. Available from: http://www.acpjc.org/shared/purpose_and_procedure.htm

[13] Haynes B, Wilczynski N, McKibbon KA, Walker CJ, Sinclair JC. Developing Optimal Search Strategies for Detecting Sound Clinical Studies in MEDLINE. JAMIA. 1994;1(6):447-58.

[14] Kilicoglu H, Demner-Fushman D, Rindflesch TC, Wilczynski NL, Haynes RB. Towards Automatic Recognition of Scientifically Rigorous Clinical Research Evidence. Journal of the American Medical Informatics Association. 2009;16(1):25-31.

[15] Hersh W, Buckley C, Leone TJ, Hickam D. OHSUMED: an interactive retrieval evaluation and new large test collection for research. Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval. Dublin, Ireland: Springer-Verlag New York, Inc.; 1994. p. 192-201.

[16] Lovins JB. Development of a stemming algorithm. Mechanical Translation and Computational Linguistics. 1968;11:22-31.

[17] Robertson SE, Walker S. Okapi/Keenbow at TREC-8. Proceedings of the eighth Text REtrieval Conference (TREC-8). Gaithersburg, Maryland: NIST; 1999. p. 151-61.

[18] Fang H, Tao T, Zhai C. A formal study of information retrieval heuristics. Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. Sheffield, United Kingdom: ACM; 2004. p. 49-56.

[19] Savoy J. Data Fusion for Effective European Monolingual Information Retrieval. Multilingual Information Access for Text, Speech and Images Lecture Notes in Computer Science.3491:233-44.

[20] Xu J, Croft WB. Improving the effectiveness of information retrieval with local context analysis. ACM Trans Inf Syst. 2000;18(1):79-112.

[21] Harman D. Relevance feedback revisited. Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval. Copenhagen, Denmark: ACM; 1992. p. 1-10.

[22] Yoo S, Choi J. Improving MEDLINE Document Retrieval Using Automatic Query Expansion. LNCS Asian Digital Libraries. 2007 2007/12/10;4822:241-9.

Address for correspondence

Jinwook Choi, Medical Informatics Lab, Dept. of Biomedical Engineering, Seoul National University College of Medicine, 28 Yeongeon-dong, Jongno-gu, Seoul 110-799, Korea ; email:<jinchoi@snu.ac.kr>