CSKU GPRF-QE for Medical Topic Web Retrieval

Ornuma Thesprasith and Chuleerat Jaruskulchai

Department of Computer Science, Faculty of Science, Kasetsart University, Bangkok, Thailand
ornuma.thesprasith@gmail.com and fscichj@ku.ac.th

Abstract. Patients and their relatives have more chances to access their health-information in a form of discharge summary. Most of them do not totally understand contents in the discharge summary. The ShARe/CLEF eHealth Evaluation Lab organized a shared task for improving retrieval medical information from the web. Queries of this task are formulated based on information in discharge summaries. This paper investigates efficiency of query expansion using external collection. Co-occur terms in pseudo-relevance feedback of Genomics collection are selected and re-weighted based on Rocchio's formula with dynamic tunable parameters of pseudo-relevance part. LUCENE, vector space model, is baseline retrieval tool. The proposed expansion method improves from baseline in all level cut of nDCG and best perform in P@10 of 3 topics. Using biomedical related collection such as Genomics is useful for medical topics retrieval.

Keywords: Genomics Track 2004, pseudo-relevance feedback, re-weighting scheme, medical terminology retrieval.

1 Introduction

Most patients or their relatives may be questionably when reading their discharge summary because medical terminology is very specific domain and is un-easy to understand by laypeople. The ShARe/CLEF eHealth evaluation lab is established to help these users more comprehend the health information [1]. Especially in the Task 3: User-centred health information retrieval [2] focuses on web collection. Since search engines are usually used to retrieve more explanation about the medical-specific subject. The expected results about health information should be understood by general users and come from reliable resources. This means that the relevant web pages contents are consisted of the medical terminology along with general terms or common words that explain the medical term in more detail.

Query expansion techniques are widely used to improve retrieval performance. There are several factors effect to expansion results; source of expansion, term selection, and re-weighting method. Source to expand query may come from many sources such as local collection, external-standard collection, and ontology. External collections such as English Wikipedia and TREC (disk 1-5) are used for expansion as reported in [3]. Reliable and most often used biomedical ontologies are UMLS Metathesaurus [4], MeSH ontology [5], and SNOMED-CT [6].

Research work [7] proposed method for selection the most effective expansion source based on query performance prediction technique. The objective of this technique is to estimate performance of retrieval system without relevant judgment [8]. This technique is either analysis collection without retrieval or focus returned results [9]. However, query performance prediction can estimate degree of relation between difference collections also. We follow this idea to select source for expansion such as med [3] , OHSUMED [10] and Genomics collection [11] .

Expansion terms from an external collection should be similar to indexing terms of the local collection. In our previous work [12] used internal MeSH (Medical Subject Headings) terms of local collection (OHSUMED) for expansion based on pseudo-relevance feedback (PRF) method. Since users need information to describe disease and treatment in MEDLINE collection, expansion query with medical vocabulary may be beneficial method. On the other hand, the ShARe/CLEF Task 3a [2] queries are specific medical terms in discharge summaries whereas collection contains health information web pages for laypeople. We believe that there is a gap between specific medical terms in user's queries and general words used in relevant web pages. To expand medical terms in query, we select terms in title and abstract part instead of medical controlled vocabulary part (MeSH terms) for expansion. We expect that these candidate terms derived from this method should appear more in relevant web pages and effect to boost up retrieval scores.

Research works [13] expanded query based on pseudo-relevance feedback (PRF) method and adapted Rocchio's formula for re-weighting terms. We adapt PRF method in different way by using results of external PRF instead of local PRF as used in traditional PRF paradigm. We adapt re-weighting formula for appropriately expansion. The details of our expansion method and results are described in the next sections.

2 Method

2.1 Vector Space Model Method and Tool

The collection is represented by matrix of terms-documents and each raw is a representation of document and is consisted of weighted term values. The query is represented by vector of weighted term values as the document vector. The similarity between query and each document vector is used to rank the returned results. The classical vector similarity measure is cosine similarity defined as following.

$$CosineSim(\vec{d},\vec{q})= \frac{\vec{d}\cdot\vec{q}}{\|\vec{d}\|\cdot\|\vec{q}\|} \tag{1}$$

Lucene is vector space model retrieval tool [14]. This tool is implementing cosine similarity measure in sophisticate way. The Lucene similarity measure is define as following.

$$LuceneScore(d,q) = \sum_{t\ in\ q}(tf(t\ in\ d) \times IDF(t^2) \times Boost(\ t\ field\ in\ d) \times LengthNorm(t\ field\ in\ d) \times Coord(q,d) \times QueryNorm(q) \tag{2}$$

Lucene allows user to boost some query terms to have specific weight via "^" (the caret mark), for example, "hepatic^3.0 encephalopathy^4.6 liver". These boosting query weight will be used in Coord(q,d) function and normalized by QueryNorm(q) function.

2.2 Genomics Pseudo-Relevance Feedback Expansion Method

Indexing Process.

This current work, the documents are web pages collected from many medical-related resources [2] . Queries of this collection are formulated by using medical terminologies in discharge summaries. We use 5 train queries to determine indexing method; a) all document (web pages with raw data), b) non-html tags documents (some pages are missing), and c) non-html tags documents compensate missing pages with original pages. In our preliminarily experiment, we evaluate MAP performance using train relevant judgments. The results are mixed and inconsistency. Therefore we select compensate indexing method to avoid losing under-estimate webpages.

Expansion Source Selection.

Original purpose of query performance prediction (QPP) is to estimate retrieval system without relevant judgment [8]. This technique is either analysis collection without retrieval or focus returned results [9]. Research work [7] used QPP to select appropriate expansion sources by comparing average term frequency of query with local and external collection. Our work uses simplest method by comparing number of documents returned from retrieval in three TREC standard sub-collections such as med [3], OHSUMED [10], and Genomics 2004 [11]. Results of 5 train queries from Genomic collection are larger than OHSUMED and med. In this current work, we believe that more documents returned provide more useful expansion terms. Even the Genomic collection based on genomics information, we expect that biology terms in genomics-based documents have relationship with medical terminologies.

Term Selection.

We hypothesize that relevant documents should contain more general terms that easy to understand by laypeople. Therefore expansion terms could be binding specific terms in queries and general terms in web pages. We select terms co-occur more often in Genomics-PRF set for expansion. Procedures for term selection describe as following steps. First, we retrieve in Genomics collection (uses title and abstract for indexing process). Second, top-k documents that contain any query terms are included in Genomic-PRF set. Third, terms in title and abstract part of this set are selected based on term frequency as candidate set.

Since candidate terms derive only from Genomics collection (Genomic-PRF set), these terms can be redundant with query terms or new added terms. Each candidate terms should be assigned with different weight based on its appearance.

Re-weighting Method.

Rocchio's formula is widely used for PRF-based query expansion re-weighting schemes. The formula composes of three part; original weight, relevance-based weight, and non-relevance-based weight.

$$W'_Q = \alpha\ W_Q + \beta / |D_R| \times (\sum d_r) + \gamma / |D_N| \times (\sum d_n) \tag{3}$$

where α is tunable weight of initial query,
W_Q is weight of term in initial query,
β is tunable weight of relevant documents (d_r),
$|D_R|$ is number of relevant documents,
γ is tunable weight of non-relevant documents (d_n),
$|D_N|$ is number of non-relevant documents.

The traditional pseudo-relevance re-weighting formula replaces the relevant part with pseudo-relevance part and ignores non-relevant part by setting γ to 0. It defined as follow,

$$W'_Q = \alpha\ W_Q + \beta \times W_{PRF} \tag{4}$$

where W_{PRF} is weight of term in pseudo-relevance documents.

Our method divides original query into two parts according to appearance in candidate set, non-candidate terms (W_{NQ}) and candidate term (W_{CQ}). These two parts of query terms have corresponding tunable parameters are $\alpha_1$and α_2 respectively. Our pseudo-relevant part uses top k documents that returned from Genomics collection and define as Genomics-PRF terms (W_{GPRF}). We are not using pseudo-relevant feedback from local collection. The re-weighting formula is defined as follow,

$$W'_Q = \alpha_1 W_{NQ} + \alpha_2 W_{CQ} + \beta W_{GPRF} \tag{5}$$

where W_{NQ} is weight of initial query term that not appear in PRF set,
α_1 is tunable parameter for initial query term that not appear in PRF set,
W_{CQ} is weight of initial query term that appear in PRF set,
α_2 is tunable parameter for initial query term that appear in PRF set,
W_{GPRF} is weight of Genomics expansion term in PRF set,
β is dynamic tunable parameter for new expansion term in PRF set.

We set more value to original query terms that are not appear in external-based expansion source to prevent "query drifting". We use external resource for finding new terms for increase recall. If these terms are original query terms we set the offset value of the W_{CQ} less than the W_{NQ} and use frequency for boosting up from the offset. This approach reduces effect of over-weighting terms.

Query log is useful information for relevant judgment [15]. We assume that results from train queries act as query log. We use train queries and their relevant judgments to set tunable parameter values. Term frequency in Genomics-PRF set is used to set these parameters. We derive optimized values of each weight is 1.0 and tunable parameter values ($\alpha_1, \alpha_2, \beta$) are 3.0, (2.0 + $log_2(tf)$) and (0.5+$log_2(tf)$) , respectively. These setting are done quite well in training set in heuristic manner. We expect that these values will work well on test query set also.

3 Experiments

3.1 Experimental Setup

The collection contains 8 parts of .zip files [2]. The html content of a web page is within "#UID" and "#EOR" tag. The html pages total is 875,486 files. Jsoup [16] is html parser tool used for extract major content from web page such as title, description, keywords, header, bold, and strong text. From this content extraction process, some files are very small (size less than 200 bytes). These are qualified files that contain 460,279 files. In our preliminary experimental, we indexed collection three types: a) raw html (whole collection), b) only major content that without html tags (460,279 files) and c) compensate missing major content file with raw html (whole collection).

Lucene version 4 is indexing and retrieval tool. We use Lucene's Standard Analyzer for indexing three collection types [14]. We retrieved 5 training topics and evaluated with train relevant judgment. Since MAP results are mixed, we avoid losing the under-estimate web pages by indexing with compensate method (type c).

Our research work focuses on finding a suitable source for query expansion. In preliminary, we compare results returned from retrieval 5 train queries. The preliminary results demonstrated that Genomics 2004 collection returns maximum number of documents in all train queries. This collection contains more biomedical terms and gene information thus we believe that returned documents are likely to have more related medical terms.

Since we assume that keywords or information need of users are similar to keywords used in the train queries. This paper investigates efficiency of using Genomics collection to expand medical topics queries. With the preliminary experiment, we retrieve 5 train queries and vary number documents (top k) in Genomic-PRF set and number of expansion terms (top m) according to equation (5). By considering MAP results from our variations, we found that the optimized values for top-k and top-m are 19 documents and 8 terms, respectively. We expect that the test queries are not different from the train queries that we used to setting these parameter. In expansion process, candidate terms are terms that co-occur in the same document of query terms in pseudo-relevance feedback (PRF) set.

3.2 Remarks before Discussion

Since our official baseline is missing result of topic no.50 because of program error. This error result to the evaluation of baseline is lower than usual. Therefore we re-examine the correction baseline (with returned result of topic no. 50) and re-evaluate the retrieval performance. The MAP values for correction baseline run and expansion run are 0.1820 and 0.2076, respectively.

3.3 Results and Discussion

The results from all runs are shown in this section. We demonstrate nDCG comparison as detail in Table 1. All nDCG cut level of expansions are higher than two baselines (both official and correction version). This means terms in Genomics documents occur in relevant web pages. Re-weighting these terms are effect to result ranking. Detail of other metrics in trec evaluation of our runs shown in Table 2.

Precision at 10 (P@10), our baseline-run above median 10 topics whereas expansion-run above median 14 topics. Our expansion proposed is best performance in 3 topics (4, 9, and 17) of 50 topics. Fortunately, terms in pseudo-relevance feedback of these topics more relate to main keyword such as “anoxic” vs. “anoxia”, “pneumonia” vs. “lung”, and “duodenal” vs. “gastric”. These expansion terms are very helpful.

The expansion results improve from official baseline 8 topics whereas official baseline outperforms expansion 4 topics. As shown in the following figures.

Table 1. nDCG comparison with baseline runs

nDCG	Official Baseline	Correction Baseline	Expansion
ndcg_cut_5	0.4896	0.5096	**0.5601**
ndcg_cut_10	0.4688	0.4855	**0.5471**
ndcg_cut_15	0.4392	0.4557	**0.4861**
ndcg_cut_20	0.4005	0.4170	**0.4476**
ndcg_cut_30	0.3568	0.3708	**0.3943**
ndcg_cut_100	0.3224	0.3338	**0.3560**
ndcg_cut_200	0.3678	0.3793	**0.4072**
ndcg_cut_500	0.4118	0.4250	**0.4574**
ndcg_cut_1000	0.4412	0.4557	**0.4895**

Table 2. Results of official baseline, correction baseline and expansion run

Metric	Official Baseline	Correction Baseline	Expansion
Num_q	50	50	50
Num_ret	47538	48538	49014
Num_rel	3132	3209	3209
Num_rel_ret	1665	1725	1828
MAP	0.1740	0.1820	0.2076
P@10	0.4680	0.4840	0.5540

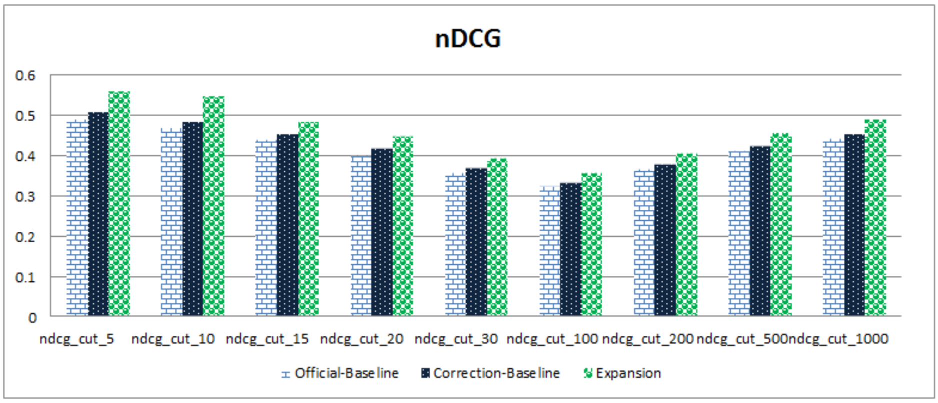

Fig. 1. nDCG baseline runs compare with expansion run

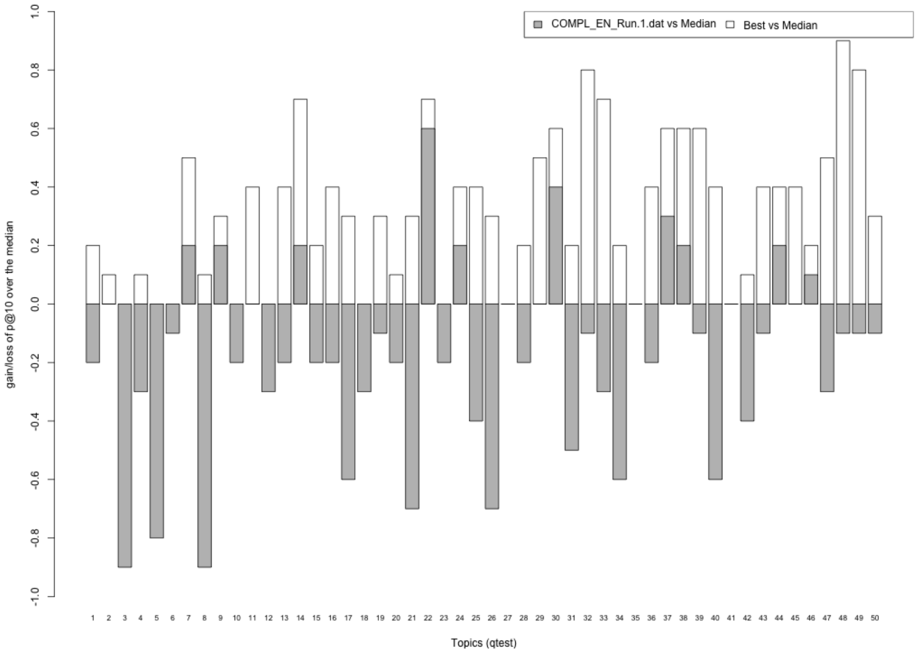

Fig. 2. P@10 baseline run compare with median and best performance

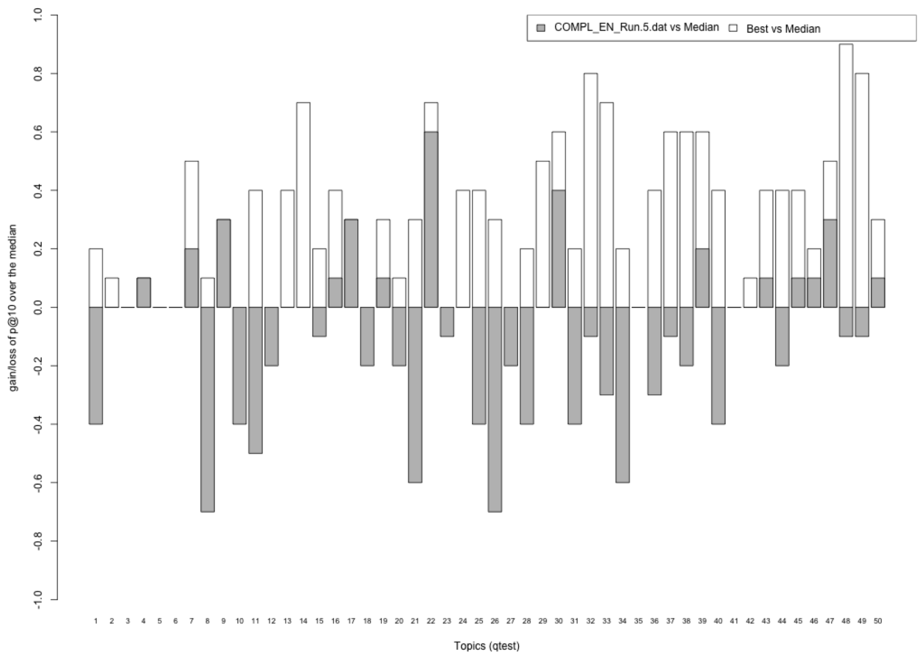

Fig. 3. P@10 Gnomic expansion-run compare with median and best performance

4 Conclusion

Medical terms in discharge summary are difficult for laypeople because these terms are very specific domain terminology. Retrieval by using queries constructed from discharge summary will be returned too specific web pages and users still need more explanation and information about the subject.

We believe that relevant web page contain both medical terminology and general terms. We use query expansion technique to explore useful terms and increase possibility of retrieval more relevant documents. Our query expansion approach is based on pseudo-relevance feedback using external biological (genomics literature) collection. We use train queries and train relevant judgments to set the optimized parameters for our proposed expansion method.

The importance issues for query expansion are source of terms, type of term for expansion, and re-weighting scheme. We determine expansion source based on query performance prediction technique. We estimate usefulness of external collection base on size of returned set. Since biomedical references in Genomics collection has disease and related-gene information. Terms in these references are selected and re-weighted based on frequency in PRF set. Although we use only statistical information in pseudo-relevance feedback set, this proposed method shows MAP improvement from baseline.

In future work, we will keep going on more sophisticated criteria to select external collection to expand query and experiment on various external collections.

References

1. Kelly, L.,Goeuriot, L.,Suominen, H.,Schrek, T.,Leroy, G.,Mowery, D. L.,Velupillai, S.,Chapman, W. W.,Martinez, D.,Zuccon, G., and Palotti, J.: *Overview of the ShARe/CLEF eHealth Evaluation Lab 2014*. Springer (2014)
2. Goeuriot, L.,Kelly, L.,Li, W.,Palotti, J.,Pecina, P.,Zuccon, G.,Hanbury, A.,Jones, G., and Mueller, H.: *ShARe/CLEF eHealth Evaluation Lab 2014, Task 3: User-centred health information retrieval*, In *CLEF 2014*. (2014)
3. Voorhees, E. M., and Harman, D.: *Overview of the Fifth Text REtrieval Conference (TREC-5)*, In *TREC*. (1996)
4. Unified Medical Language Systems, http://www.nlm.nih.gov/research/umls
5. The Basics of Medical Subject Headings (MeSH®), http://www.nlm.nih.gov/bsd/disted/mesh/
6. SNOMED Clinical Terms® (SNOMED CT®), http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html
7. He, B., and Ounis, I.: *Combining fields for query expansion and adaptive query expansion.* **43**, 1294-1307 (2007)
8. Kurland, O.,Raiber, F., and Shtok, A.: *Query-performance prediction and cluster ranking: Two sides of the same coin*, In *Proceedings of the 21st ACM international conference on Information and knowledge management*. pp. 2459-2462. ACM (2012)
9. Cummins, R.,Jose, J., and O'riordan, C.: *Improved query performance prediction using standard deviation*. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. pp. 1089-1090. ACM (2011)
10. Hersh, W.,Buckley, C.,Leone, T. J., and Hickam, D.: *OHSUMED: An Interactive Retrieval Evaluation and New Large Test Collection for Research*, in *SIGIR '94*, B. Croft and C.J. Rijsbergen.(eds). p. 192-201. Springer London (1994)
11. William R. Hersh , R. T. B., Laura Ross , Phoebe Johnson , Aaron M. Cohen , Dale F. Kraemer. *TREC 2004 genomics track overview* In *The 13th Text REtrieval Conference*. (2004)
12. Thesprasith, O., and Jaruskulchai, C.: *Query Expansion Using Medical Subject Headings Terms in the Biomedical Documents*, in *Intelligent Information and Database Systems*. p. 93-102. Springer (2014)
13. Abdou, S., and Savoy, J.: *Searching in Medline: Query expansion and manual indexing evaluation.* **44**, 781-789 (2008)
14. Apache Lucene - Apache Lucene Core, http://lucene.apache.org/core/
15. Cui, H.,Wen, J.-R.,Nie, J.-Y., and Ma, W.-Y.: *Query expansion by mining user logs.* **15**, 829-839 (2003)
16. jsoup: Java HTML Parser, http://jsoup.org/