

Selective Source for Query Expansion in Health Information Retrieval

Ornuma Thesprasith* and Chuleerat Jaruskulchai*

Received : May 31, 2018

Revised : September 7, 2018

Accepted : September 21, 2018

Abstract

Query expansion aims to solve the vocabulary mismatch problem by adding new terms to the original query or reweighting existing query terms. Since there are external sources available, it challenges to select the right source for each query expansion. Useful terms of the right expansion source result in increasing the rank of relevance documents. We propose the selective source expansion in health information retrieval framework. Our source selection method is based on the relative entropy of two probability distributions. These probability distributions estimate from pairs of terms in the query and pairs of terms in the collection, respectively, instead of the individual query term. The proposed framework improves retrieval performance from baseline retrieval, traditional query expansion, and existing selective query framework as well.

Keywords: Selective Query Expansion, Clarity, Query Performance Prediction, Health Information Retrieval.

1. Introduction

This paper aims to improve traditional query expansion approach by selecting the most effective source to expand the health-related query. The original of this work was presented in ICSEC2016 [1]. This paper extends the original work by modifying the formula of source selection method and explores the results in more details.

In the beginning, query expansion (QE) [2] aims to solve the vocabulary mismatch problem in information retrieval. Typically, the original query is expanded by related terms and possibly re-assigned query terms weight. Sources of terms

and re-weighting method are keys important to query expansion.

The first-round retrieval result is the pseudo-relevance feedback (PRF) which is a commonly used as a source of related terms. Additionally, there are many sources available. If the external source relates to the user's query and collection, therefore new terms may be introduced from the external source. Applying PRF method to the external collection is called external-PRF expansion in this paper. Typically, when using the external sources for expansion [3], [4], large collections are often recommended such as Wikipedia (<https://www.wikipedia.org>) and the Genomics (<http://trec.nist.gov/data/genomics.html>).

Although traditional query expansion yields better retrieval performance, some queries are worse and known as the drifting problem. The selective query expansion (SQE) proposes to use different expansion methods for each query. There are various approaches [5], [6], [7] to the SQE framework. The first SQE framework determined which query needs to expand by using query prediction performance [5]. On the other hand, a different setting on the number of documents in the first-round retrieval result has been examined [6]. Moreover, using a different source for each query expansion has been reported [7].

The query performance prediction (QPP) [8-11] is commonly used in the selective query expansion frameworks. In the beginning, QPP aims to predict retrieval performance of retrieval systems based on characteristics of the queries. There are two prediction types of query performance prediction; pre- and post-retrieval. The pre-retrieval QPP methods use information of collection such as term frequency and the number of documents whereas the

* Department of Computer Science, Faculty of Science, Kasetsart University, Bangkok, Thailand.

post-retrieval QPP methods use more information such as the first-round retrieval. Therefore the post-retrieval QPP methods are more robustness.

The above SQE frameworks [5], [6] based on unambiguity of terms in query and in pseudo-relevance feedback known as the query clarity [10] and another framework [7] based on the query specificity [11]. Since the query clarity takes more information to process than the query specificity thus it predicts more robustness.

Although those SQE frameworks [5], [6] predicted with more robustness method, their expansion source based on one target collection. Applying these frameworks to the external collections are challenged. On the other hand, the SQE framework [7] used the less robustness prediction method to select between target and external collection. From these mentioned challenges, we propose the SQE framework that selects between a target and two external collections with the more robust prediction method.

We evaluate our proposed SQE framework on two query sets of the CLEF eHealth collection (<http://clefehealth2.14.dcu.ie/task-3/dataset>). Each query set reflects different scenarios of query formulation. For instance, lay people may have a discharge summary that reports the health condition [12]. They may use medical terms from the discharge summary as keywords to search for more detail easily. On the other hand, lay people formulate the query by themselves may use general words to explain their health condition [13]. These general terms may appear in different kind of related health documents. Thus, queries with more general terms tend to be ambiguous. The query clarity method predicts the unambiguity based on individual term.

However, in general, lay people usually use collocations to talk about general health-related problems such as “bad back”, “bad leg”, and “sore throat” (<https://www.ecenglish.com/learnenglish/lessons/how-talk-about-health-problems>). These collocations are examples of the pair of terms in the query. We hypothesize that a pair of general terms is less ambiguity than the individual. Therefore existing methods

that consider individual terms should be revised.

Our novel prediction method is the Pair Clarity score where a pair is any two terms occurred within the text window. This score is used to select the one source for query expansion in our SQE framework.

The contributions of this work are as follows.

- An automatic query expansion framework to select a source of expansion between a target and two external related health collections.
- The Pair clarity, a novel method to predict expansion source and derived from the pair of query terms.
- An investigation of the Specificity method on SQE framework in health-related retrieval.
- An evaluation of the Clarity method on the source selection for query expansion in the health-related retrieval.

2. Related Work

2.1 Retrieval Models

There are three classical retrieval models [14]; the Boolean model, the vector space model, and the probabilistic model. The unigram language model (LM) [15] is a well-known probabilistic retrieval model.

Three main components of a retrieval process are a collection of documents, a query set, and a ranking method. In the LM model [15], the probability distribution of terms in a document represents the document model. The probability distribution of terms in a query is called query model. The likelihood of the query model according to the document model is using as a ranking method. It is known as the query likelihood scoring method [16]. With the document smoothing [17], the probability distribution of missing terms in the document interpolates with the probability distribution of that terms in the collection. This result gives a better retrieval performance.

2.2 Query Expansion Models

The relevance model (RM) [18] and the simple mixture model (SMM) [19] are commonly used in expansion. These two methods are grounded from the pseudo-relevance feedback

(PRF) paradigm [20] that takes the first-round results to the query expansion process.

2.3 External Source for Query Expansion

The RM [18] method has been further improved by many researchers, especially by external expansion approaches [3], [21], [22]. The objective of these works is to find expansion terms from many sources. The mixture of relevance model (MoRM) [3] applied the traditional relevance model to the external collections. The probability distribution of all collections is balanced weight. Furthermore, The external expansion model (EEM) [21] was weighted collections differently. Finally, the cluster-based external expansion (CBEEM) model [22] grouped PRF documents according to their similarities and then used these clustered documents to estimate the model.

2.4 Selective Query Expansion

The objective of selective query expansion is to take different actions when doing an expansion [5], [6], [7]. Retrieval effectiveness of SQE methods is better than the traditional query expansion. We introduce three approaches of the selective query expansion frameworks which each framework proposed for different situations.

The first SQE framework [5] solved to prevent the query drifting problem. The first PRF model estimated from the original query retrieval. The second PRF model derived from the expanded query retrieval. The original query will be selected to prevent the query drifting problem when the probability distribution of significant terms of the second model are lower.

The second SQE framework [6] proposed to select the most effectiveness expanded query. The candidate query models are derived from the various settings of PRF; the number of documents in PRF and number of expansion terms. The chosen query model should be similar to the PRF model of the original query retrieval and dis-similar to the collection model. These first two SQE frameworks [5], [6] employed the Kullback-Leibler divergence or relative entropy to measure the dissimilarity between the two models.

The last SQE framework [7] proposed to select the sources for the expansion grounded on the specificity of the query terms.

2.5 Query Performance Prediction

In the beginning, the query performance prediction aims to predict the performance of retrieval system when processing the query [8-11]. Interestingly, all previous SQE frameworks [5], [6], [7] applied the query performance prediction in the selection process.

The third SQE framework [7] employed the Average of Inverse Collection Term Frequency (AvICTF) [11] as a prediction method whereas the first two SQE frameworks [5], [6] employed the Clarity score as the grounded formula.

The AvICTF [11] is derived from the statistical information of a collection and a query without the retrieval process. It is a kind of pre-retrieval QPP method [23]. With an assumption that if terms occur not very often in a collection, it implies that these terms focus on some aspect. The pre-retrieval QPP method makes it easy to retrieval system to retrieve relevance documents. Therefore the query with high specificity is predicted as a well-performing query.

On the other hand, the post-retrieval QPPs use the first-round retrieval in the prediction process [9]. For example, the Clarity score [10] is the relative entropy of two distributions. The first distribution is the query relevance model which estimated from the first-round retrieval documents. The second distribution is the collection model; the probability of a term in the collection. The Clarity score is high if query terms are more density in the query relevance model than in the collection model.

3. Selective Source Expansion Framework

The proposed SQE framework consists of three main processes; retrieval, source selection and query expansion as shown in Figure 1. This framework processes with three collections that are target collection and two external collections. The workflow of our framework is as follows.

The first-round retrieval process uses the original query

to retrieve in three collections. The first-round retrieval results are three PRF sets from all collections.

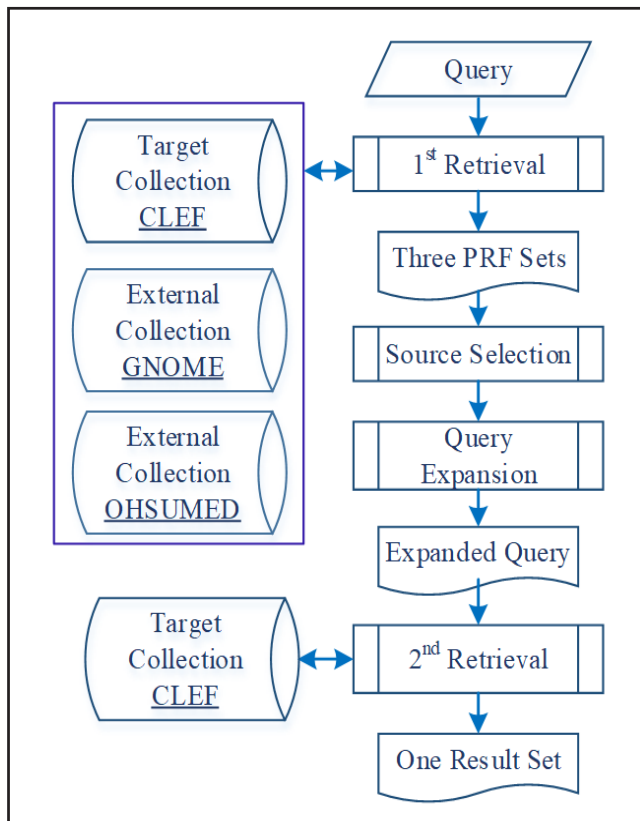


Figure 1. Selective source for query expansion framework in the health-related retrieval.

Then all PRF sets are used in the source selection process. Only one PRF with the highest predicting score is selected.

The query expansion process finds candidate terms in the selected source. The expanded query is the original query along with the candidate terms.

Finally, the second-round retrieval retrieves in the target collection with the expanded query.

3.1 Retrieval Processes

Given original query $Q = q_1, q_2, \dots, q_n$, document D is ranked with the query-likelihood score [16] that defined as follows

$$QL(Q, D) = \sum_{i=1}^n \log \frac{c(q_i, D) + \frac{c(q_i, C)}{|C|}}{|D| + \mu} \quad (1)$$

where q_i is the i th query term. $c(q_i, D)$ and $c(q_i, C)$ are the count of query term q_i in document D and collection C , respectively. $|D|$ and $|C|$ are the total number of terms in

document D and collection C respectively. μ is the Dirichlet smoothing parameter [17].

For the second-round retrieval, the target collection retrieves with the expanded query, using the Kullback-Leibler divergence approach [24]. This approach hypothesized that the relevant documents should be similar to the query relevance model and dissimilar to the collection model. The document ranking score is defined as follows

$$KL(Q_R, D) = \sum_{w \in D} p(w|\theta_{QR}) \log \frac{p_s(w|\theta_D)}{p(w|\theta_C)} \quad (2)$$

where $p(w|\theta_{QR})$ is the query relevance model. $p_s(w|\theta_D)$ is the smoothed document model and $p(w|\theta_C)$ is the collection model estimated as $p(w|\theta_C) = c(w, C)/|C|$ where $c(w, C)$ is the frequency of the word w in the collection C . The number of word occurrences in the collection denoted as $|C|$.

The query relevance model as defined in Equation (8) and will be explained in the Query Expansion Process section. The smoothed document model [17] is defined as follows

$$p_s(w|\theta_D) = \frac{c(w, D) + \mu p(w|\theta_C)}{|D| + \mu} \quad (3)$$

where $c(w, D)$ is the term frequency of the word w in the document D , μ is the Dirichlet smoothing parameter [17] and $p(w|\theta_C)$ is the collection model.

3.2 Source Selection Method with Pair Clarity

In this section, we explain our proposed method for predicting the quality of expansion source based on pairs of terms in the first-round retrieval results.

We use v to denote a pair of terms; $v = (t_s, t_w)$. Any two terms that occur within some text window size is pared. The window size in this paper is 7; the estimation of the sentence length after removing stop-words. Any pair that occurred within a document more than twice is useful.

The probability distribution of pairs of terms in a collection is called the pair collection model. We estimate the pair collection model using the maximum likelihood; $p(v|\theta_{cv}) = c(v, C_v)/|C_v|$ where $c(v, C_v)$ is the frequency of the pair v in collection C and $|C_v|$ is the number of all pairs occurred in the collection.

We also estimate the pair feedback model using the maximum likelihood; $p_F(v|\theta_{Fv}) = c(v; F_v) / F_v$ where $c(v; F_v)$ is the frequency of the pair v in the PRF documents and $|F_v|$ is the number of all pairs occurred in the PRF documents.

Any two terms in the original query within window size are paired together. For original query $Q = q_1, q_2, \dots, q_n$, a set of pairs is $Q_v = v_1, v_2, \dots, v_m$, where $v_x = (q_i, q_k)$

With an assumption that if the pair of two general terms has often occurred in a small set of documents, then the clarity score of this pair should be more than considering individually.

The Pair Clarity score is a summation of the KL divergence between the pair feedback model and the pair collection model and formally defined as in Equation (4)

$$PairCS(Q, F) = \sum_{v \in Q_v} p_F(v|\theta_F) \log \frac{p_F(v|\theta_F)}{p(v|\theta_C)} \quad (4)$$

Since each collection has the different aspect, we define the free parameter β_x called the collection-dependent parameter. The collection with the maximum value of the Pair Clarity score along with the collection dependent parameter is selected as follows

$$ExpSource = \max(\beta_C PairCS_C, \beta_G PairCS_G, \beta_O PairCS_O) \quad (5)$$

where β_C, β_G and β_O are the collection dependent parameters of the CLEF collection [12], the GNOME collection [25], and the OHSUMED collection [26], respectively. The Pair Clarity score for each collection has denoted in the formula as the *PairCS* with the subscript of a collection first letter.

3.3 Source Selection Method with Existing Query Performance Predictions

We evaluate our proposed method in the effectiveness of source selection by comparing with two existing methods; the Specificity [11] and the Clarity [10]. The Specificity can be presented in the form of the KL divergence between the original query model and the collection model as follows

$$Specificity(Q, C) = \sum_{w \in Q} p(w|Q) \log \frac{p(w|Q)}{p(w|\theta_C)} \quad (6)$$

where $p(w|Q)$ is the original query model estimated by the maximum likelihood method and $p(w|\theta_C)$ is the collection

model.

The Clarity [10] derives from the KL divergence of the query relevance model and the collection model defined as follows

$$Clarity(Q, C) = \sum_{w \in Q} p(w|\theta_R) \log \frac{p(w|\theta_R)}{p(w|\theta_C)} \quad (7)$$

where $p(w|\theta_R)$ is the query relevance model as defined in Equation (9) and $p(w|\theta_C)$ is the collection model.

3.4 Query Expansion Process

Our query expansion process is the same process as the relevance model (RM3) [18] that first estimate query relevance model from the PRF documents and then we interpolate the model with the original query model.

The RM3 model [18] is defined as follows

$$p_{RM3}(w|\theta_R) = (1-\alpha)p(w|Q) + \alpha p(w|\theta_R) \quad (8)$$

where $p(w|Q)$ is the original query model estimated by the maximum likelihood method. α is the interpolated feedback coefficient. $p(w|\theta_R)$ is the first relevance model and defined as

$$p(w|\theta_R) = \sum_{D \in F} p(w|D) p(D|Q) \quad (9)$$

where $p(D|Q)$ is the score of the document D given a query Q derived from Equation (1) and F is the PRF documents which are results from the first-round retrieval. $p(w|D)$ is the probability of the word in document D which derives as $p_S(w|\theta_D)$ in Equation (3).

If the selected source based on our prediction method is the external collection, then the RM3 will be applied to use the external instead.

4. Experiments

We conduct the experiments to evaluate our proposed method with two query sets of one target collection. In this section, we first present three health-related collections and then introduce the retrieval tool which we have employed. Finally, we describe all retrieval settings; baseline, traditional query expansion and SQEs retrieval.

4.1 Health Related Collections

The CLEF eHealth which provided by the Khresmoi

project (<http://clefehealth2.14.dcu.ie/task-3/dataset>) is the target collection. This dataset contains one million web pages from reliable sources about health-related information; denoted as CLEF in this paper.

We evaluate two query sets of the CLEF eHealth dataset; denoted as Q2014 and Q2015. The Q2014 query set [12] aims to help lay people to find relevance documents in case that they use medical terms from discharge summary to search for more information. On the other hand, the Q2015 query set [13] is assumed that lay people formulate query without discharge summary by using general words instead. These two query sets are different in the aspect of query terms significantly.

External collections are bio-literature collections from the online-medical information database; known as the MEDLINE database (<https://www.nlm.nih.gov>); GNOME [25] and OHSUMED [26] collections. The GNOME collection [25] is a test collection in the genomics domain that contains a ten-year subset of 4.5 million MEDLINE records. The OHSUMED collection obtained by William Hersh [26] which contains a five-year subset of 348,566 medical journals.

4.2 Indexing and Retrieval Tool

The Indri Lemur toolkit is the tool for indexing and retrieval process (<https://www.lemurproject.org/>). This tool supports the query-likelihood ranking method and the KL divergence as described in Equation (1) and (2).

In the indexing process, the CLEF's web pages are cleaned by removing HTML tags and then removing the stop words. Finally, each word is stemmed using the Krovetz stemming algorithm [27].

For two external collections, the title, abstract, and medical subject heading (MeSH) are represented as a content of the documents. Then, the same process of indexing of the target collection is repeated.

4.3 Run Objectives and Parameter Settings

When all collections are ready, we start experiments with the baseline retrieval. Next, we examine traditional query expansion which uses the target collection as expansion source

for all queries; Target-PRF. Then three selective query expansion frameworks are evaluated; SQE-PairCS, SQE-Specificity, and SQE-Clarity. The description of each run are explained in Table 1.

Table 1. Retrieval run and description.

Run	Description
Baseline	Original query retrieval
Target-PRF	CLEF expansion for all queries
SQE-PairCS	Selected source expansion for each queries using the Pair Clarity score
SQE-Specificity	Selected source expansion for each queries using the Specificity score
SQE-Clarity	Selected source expansion for each queries using the Clarity score

Table 2. Parameter settings.

Process	Setting
Retrieval	Dirichlet smoothing (μ) = 1000
Query Expansion	PRF in traditional QE = 10 PRF in selective QE = 100 Top terms = 15 Feedback coefficient (α) = 0.8
Source Selection	PRF in Pair Clarity = 1000 PRF in Specificity = 0 PRF in Clarity = 200 CLEF dependent () = 2.0 GNOME dependent () = 1.0 OHSUMED dependent () = 1.0

All parameters are shown in Table 2. This setting is based on the best results of our preliminary studies. In external expansion process, the number of documents in PRF is more than the traditional expansion. The collection-dependent parameter of the CLEF is higher than others because of a lower number of specific terms.

5. Results and Discussion

There are four parts of the results that are discussed in this section. The first part is the retrieval performance of two query sets that evaluated by the standard tool. To demonstrate the different performance of baseline retrievals of two characteristic query.

The second part is the comparison of prediction performance.

We also uncover the underlying process of each prediction method in the third part. Finally, we focus on our pair clarity method on the right source selection along with expanding terms from all collections.

5.1 Retrieval Performance

The trec_eval (http://trec.nist.gov/trec_eval/) is a standard tool used for evaluating the retrieval performance of the query set. We report three main retrieval performances; the number of relevance return (num_rel_ret), the Mean Average Precision (MAP), and the precision at ten documents (p@10). The right-most two columns in Table 3 represent baseline retrieval with the original query; Baseline-Q2014 and Baseline-Q2015. Each record in Table 3 is describing as follows. The num_rel_ret is the total number of relevant documents of all queries in the set. The MAP is the average value of the mean average precision of all queries in the query set. The p@10 is the average of the precision at top ten documents of all queries.

Table 3. Baseline retrieval performance.

	Baseline-Q2014	Baseline-Q2015
num_rel_ret	2429	1300
Map	0.3297	0.1391
p@10	0.7460	0.2894

From Table 3, three retrieval performances of the first query set are better than the second query set. Since the characteristic of the first query set assumed that users have known the medical terms and formulate the query using these terms, the query set tends to specific than the second set. With the more specific query, the retrieval system can retrieve relevant documents in a higher number.

To compare all retrieval approaches, we show the Mean Average Precision (MAP) of all runs in Table 4. Each record in Table 4 represents retrieval approach which evaluates on each query set. There are five retrieval approaches; the baseline retrieval, the traditional query expansion, and the three selective query expansions. In Baseline record, the values of Q2014 and Q2015 are taken from Table 3. The last column presents

the total value of MAP from two query sets. of two query sets. The Target-PRF record retrieves with target source expansion for all queries. Each SQE record retrieves with different source expansion for each query.

Table 4. Mean average precision of all retrievals.

Run	Q2014	Q2015	Total
Baseline	0.3297	0.1391	0.4688
Target-PRF	0.3305	0.1558	0.4868
SQE-PairCS	0.3378	0.1523	0.4901
SQE-Specificity	0.3347	0.1493	0.4840
SQE-Clarity	0.3370	0.1436	0.4805

From Table 4, the overall performance of traditional query expansion; Target-PRF, of two query sets improves from baseline retrieval. Similarly, the overall performance of selective query expansions of two query sets also improves from baseline retrieval. Among three selective frameworks, the SQE-PairCS performance is maximum in two query sets. In the total, our SQE-PairCS performance is the best retrieval approach.

5.2 Prediction Performance

All queries right prediction by each method is showing in Table 5. There are three rows for each prediction method. The last column also presents the total of two query sets.

Table 5. Number of right source predicted queries.

Run	Q2014	Q2015	Total
SQE-PairCS	17	24	41
SQE-Specificity	21	16	37
SQE-Clarity	18	23	41

From Table 5, the SQE-PairCS and the SQE-Clarity have the most number of queries right prediction. The SQE-Specificity is best only on the first query set which is the more specific query.

5.3 Exploring Prediction Process

Now we examine how prediction methods are working. The example query is “dry feel with irritation.” The best source for expansion of the example query is the CLEF collection.

We start with a novel method; the Pair Clarity, and then two existing methods; the Specificity and the Clarity as shown in Table 6-8.

In Table 6, the Pair column is the pair of query terms from the original query. Other three columns are the Pair Clarity for each pair of the collection. The last row is the Pair Clarity score of the collection. The high score the more effective source.

From the results in Table 6, two external collections do not contain these pairs. Therefore the score is zero. Because this query consists of more general terms, they occur not often in the more specific collections.

Table 6. Pair clarity scoring.

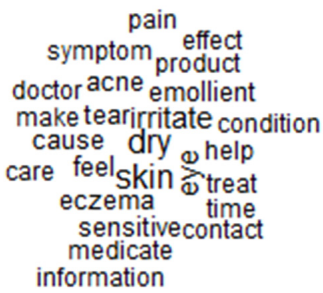
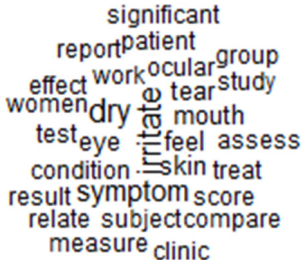
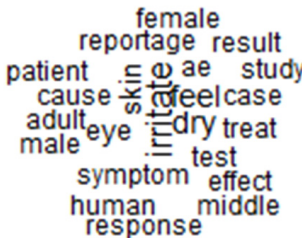
Pair	CLEF	GNOME	OHSUMED
dry feel	0.00022	0.00	0.00
dry irritation	0.00	0.00	0.00
feel dry	0.00022	0.00	0.00
feel irritation	0.00011	0.00	0.00
irritation dry	0.00	0.00	0.00
irritation feel	0.00011	0.00	0.00
Score	0.00066	0.00	0.00

Table 7. Specificity scoring.

Term	CLEF	GNOME	OHSUMED
feel	10.7280	14.400	15.3200
dry	11.8251	13.200	14.0900
irritation	13.3256	15.480	16.0029
Score	11.9595	14.420	15.1392

In Table 7, the Term column is the query term while other left columns are the specificity of each collection. The Score in the last record is the average of term specificity as defined in Equation (6). Since the specificity method is the pre-retrieval QPP method, it considers only query terms in the collection. Typically, the specificity of the general terms is low in the general domain collection. On the other hand, the specificity of the general terms is high in the specific domain collection.

Table 8. Clarity scoring.

Word Cloud	Collection/Score
	CLEF
	2.656
	GNOME
	3.246
	OHSUMED
	1.917

The more general terms such as “feel”, “dry”, and “irritation” in the target collection are less specificity than two external collections. With the specificity method, the selected source is the OHSUMED collection.

Table 8 shows word cloud of terms with the different size according to clarity score of each term. The right-most column shows both collection and its Clarity score.

The Clarity score as defined in Equation (7) is deriving from original query and the PRF set. Some related terms in the 200 documents are used to indicate the cohesion of topics in the PRF set. If PRF set is focusing on some topic, the clarity score of related terms is high. However, when comparing clarity score across collections, the more general terms tend to have a higher clarity than usual. Therefore, with the Clarity method, the GNOME collection is selected.

5.4 Exploring Terms of Selected Source

Finally, we demonstrate how the Pair Clarity score relates to the useful terms for expansion by examining three queries in Table 9 and Table 10 together.

By presenting the results of source prediction and terms derived from the source, we organize tables into three rows. Table 9 shows three queries along with the Pair Clarity method for all collections. The Query column is the original query that in the form of a pair. The next three columns represent the Pair Clarity score in the collection.

For each query in Table 9, only one of three sources has the highest score which becomes the selected source. In Table 10 shows expanded query with the selected source. The size of the term is according to its weight in the query relevance model. Typically, the size of the original terms is larger than the expanded terms.

Table 9. Pair Clarity scoring example.

Query	CLEF	GNOME	OHSUMED
rosacea symptom	0.01310	0.0008	0.0000
cerebral aneurysm	0.000231	0.0690	0.0146
black tooth	0.0000	0.0000	0.0030

From Table 9, there are three cases of the Pair Clarity scores. One source is zero on the Pair Clarity score; two sources are zero; and all sources are more than zero.

The first row is “rosacea symptom” query. There are two collections that the Pair Clarity score is not zero, this is the second case. Both CLEF and GNOME collections provide common useful terms for expansion such as “ocular” and “skin.” However, the unique term provided by the CLEF collection is “dermatol.”

The second row is “cerebral aneurysm” query. This is a specific query because it consists of more specific query terms meanwhile the pair occurs more often in all collections. This is the last case which all collections provide useful terms for expansion.

The last row of Table 9 is “black tooth” query. The only one collection that the Pair Clarity score is not zero is the

Table 10. Word cloud of expansion terms from the right source.

Word Cloud	Collection/Score
	CLEF rosacea symptom
	GNOME cerebral aneurysm
	OHSUMED black tooth

OHSUMED collection. The useful terms acquired from the OHSUMED collection are “molar”, “teeth”, and “dental” as shown in Table 8.

6. Conclusion

The retrieval performance of query expansion is affected by expanding terms and re-weighting method. Different sources provide different expanding terms for each query. Our proposed method is a novel source selection method; called the Pair Clarity score, that extends from the query prediction method. From the prediction performance results, we can confirm our hypothesis that considering the query using the pair terms is more effective than using individual terms.

Factors of source selection are characteristic of the query, collection, and the PRF set. The specificity method dominates

in the query set that more specific. However, lay people form queries without knowledge of medical terminology. Therefore our proposed method works best for the more general query.

Our selective query expansion framework is reporting on health-related retrieval. Therefore we choose the more specific collections as external sources. With the limitation of the existing methods, our proposed method overcome with boosting the target collection. Therefore the proposed framework outperforms the others.

As reported, the overall retrieval performance of the traditional query expansion method is improved from the baseline retrieval with the small number of documents in the PRF set; 10 documents. When using external collections that more specific contents than the target collection, the number of documents should be more than ten documents; 100 documents.

However, each collection has different characteristics. Thus selecting a source for expansion is still challenging, especially in health-related retrieval.

7. References

- [1] O. Thesprasith, and C. Jaruskulchai. "Simple-pharse score for selective query expansion in health Information Retrieval." *Computer Science and Engineering Conference (ICSEC), 2016 International, IEEE*, 2016.
- [2] C. Carpineto, and G. Romano. "A survey of automatic query expansion in information retrieval." *ACM Computing Surveys (CSUR)*, Vol. 44, No. 1, pp. 1, 2012.
- [3] F. Diaz, and D. Metzler. "Improving the estimation of relevance models using large external corpora." *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM*, 2006.
- [4] D. Zhu, S. Wu, and B. Carterette et al. "Using large clinical corpora for query expansion in text-based cohort identification." *Journal of Biomedical Informatics*, Vol. 49, pp. 275-281, 2014.
- [5] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. "A framework for selective query expansion." *Proceedings of the thirteenth ACM International Conference on Information and Knowledge Management. ACM*, 2004.
- [6] M. Winaver, O. Kurland, and C. Domshlak. "Towards robust query expansion: model selection in the language modeling framework." *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM*, 2007.
- [7] B. He, and I. Ounis. "Combining fields for query expansion and adaptive query expansion." *Information Processing and Management*, Vol. 43, No. 5, pp. 1294-1307, 2007.
- [8] C. Hauff. "Predicting the effectiveness of queries and retrieval systems." *SIGIR Forum*. Vol. 44. No. 1. 2010.
- [9] C. Hauff, L. Azzopardi, and D. Hiemstra et al. "Query performance prediction: Evaluation contrasted with effectiveness." *European Conference on Information Retrieval*. Springer, Berlin, Heidelberg, 2010.
- [10] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. "Predicting query performance." *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM*, 2002.
- [11] B. He, and I. Ounis. "Query performance prediction." *Information Systems*, Vol. 31, No. 7, pp. 585-594, 2006.
- [12] L. Kelly, L. Goeuriot, and H. Suominen et al. "Overview of the share/clef ehealth evaluation lab 2014." *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, Cham, 2014.
- [13] L. Goeuriot, L. Kelly, and H. Suominen et al. "Overview of the CLEF eHealth evaluation lab 2015." *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, Cham, 2015.
- [14] W. B. Croft, D. Metzler, and T. Strohman. *Search engines: Information Retrieval in Practice*: Addison-Wesley Reading, 2010.

- [15] C. Zhai. "Statistical language models for information retrieval." *Synthesis Lectures on Human Language Technologies*, Vol. 1, No. 1, pp. 1-141, 2008.
- [16] F. Song, and W. B. Croft. "A general language model for information retrieval." *Proceedings of The Eighth International Conference on Information and Knowledge Management, ACM*, 1999.
- [17] C. Zhai, and J. Lafferty. "A study of smoothing methods for language models applied to ad hoc information retrieval." *ACM SIGIR Forum*. Vol. 51. No. 2. ACM, 2017.
- [18] V. Lavrenko, and W. B. Croft. "Relevance-based language models." *ACM SIGIR Forum*. Vol. 51. No. 2. ACM, 2017.
- [19] C. Zhai, and J. Lafferty. "Model-based feedback in the language modeling approach to information retrieval." *Proceedings of The Tenth International Conference on Information and Knowledge Management. ACM*, 2001.
- [20] H. Schütze, C. D. Manning, and P. Raghavan. *Introduction to Information Retrieval*. Cambridge University Press, 2008.
- [21] W. Weerkamp, K. Balog, and M. de Rijke. "Exploiting external collections for query expansion." *ACM Transactions on the Web (TWEB)*, Vol. 6, No. 4, pp. 18, 2012.
- [22] H.-S. Oh, and Y. Jung. "Cluster-based query expansion using external collections in medical information retrieval." *Journal of Biomedical Informatics*, Vol. 58, pp. 70-79, 2015.
- [23] C. Hauff, D. Hiemstra, and F. de Jong. "A survey of pre-retrieval query performance predictors." *Proceedings of The 17th ACM Conference on Information and Knowledge Management, ACM*, 2008.
- [24] P. Ogilvie, and J. P. Callan. "Experiments Using the Lemur Toolkit." *TREC*. Vol. 10. 2001.
- [25] W. R. Hersh, and R. T. Bhupatiraju. "TREC genomics track overview." *TREC*, Vol. 2003, 2003.
- [26] W. Hersh, C. Buckley, and T. Leone et al. "OHSUMED: an interactive retrieval evaluation and new large test collection for research." *SIGIR '94*. Springer, London, 1994.
- [27] A. G. Jivani. "A comparative study of stemming algorithms." *International Journal of Computer Applications in Technology, Appl*, Vol. 2, No. 6, pp. 1930-1938, 2011.