Original Article

Healthc Inform Res. 2011 June;17(2):120-130.
doi: 10.4258/hir.2011.17.2.120
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Evaluation of Term Ranking Algorithms for Pseudo-Relevance Feedback in MEDLINE Retrieval

Sooyoung Yoo, PhD[1], Jinwook Choi, MD, PhD[2]

[1]Medical Information Center, Seoul National University Bundang Hospital, Seongnam; [2]Department of Biomedical Engineering, Seoul National University College of Medicine, Seoul, Korea

Objectives: The purpose of this study was to investigate the effects of query expansion algorithms for MEDLINE retrieval within a pseudo-relevance feedback framework. **Methods:** A number of query expansion algorithms were tested using various term ranking formulas, focusing on query expansion based on pseudo-relevance feedback. The OHSUMED test collection, which is a subset of the MEDLINE database, was used as a test corpus. Various ranking algorithms were tested in combination with different term re-weighting algorithms. **Results:** Our comprehensive evaluation showed that the local context analysis ranking algorithm, when used in combination with one of the reweighting algorithms – Rocchio, the probabilistic model, and our variants – significantly outperformed other algorithm combinations by up to 12% (paired *t*-test; $p < 0.05$). In a pseudo-relevance feedback framework, effective query expansion would be achieved by the careful consideration of term ranking and re-weighting algorithm pairs, at least in the context of the OHSUMED corpus. **Conclusions:** Comparative experiments on term ranking algorithms were performed in the context of a subset of MEDLINE documents. With medical documents, local context analysis, which uses co-occurrence with all query terms, significantly outperformed various term ranking methods based on both frequency and distribution analyses. Furthermore, the results of the experiments demonstrated that the term rank-based re-weighting method contributed to a remarkable improvement in mean average precision.

Keywords: MEDLINE, Information Storage and Retrieval, Evaluation Studies

I. Introduction

In the medical domain, the most common use of an information retrieval system is to retrieve bibliographic information [1]. Previous studies on the information sources used by physicians reported that MEDLINE was by far the most commonly used resource [2,3]. However, sometimes MEDLINE retrieval is unsuccessful, especially for queries issued by inexperienced users [4].

Query expansion is the process of reformulating a seed query to improve retrieval performance [5] and is indispensable for solving ambiguous queries. Many types of query expansion methods have been developed. These methods can be classified into two groups: query expansion using a thesaurus and query expansion using pseudo-relevant feedback. Researchers have traditionally tried to adopt a thesaurus, such as WordNet, or the Unified Medical Language System (UMLS); however, the effectiveness of thesaurus adoption is still debated. Some research showed a meaningful improvement in query expansion, but other studies did not. The more widely accepted query expansion methods are those where user relevance feedback (RF) or pseudo-relevance

Received for review: January 17, 2011
Accepted for publication: April 29, 2011

Corresponding Author
Jinwook Choi, MD, PhD
Department of Biomedical Engineering, Seoul National University College of Medicine, 28 Yeongeon-dong, Jongno-gu, Seoul 110-799, Korea. Tel: +82-2-2072-3421, Fax: +82-2-745-7870, E-mail: jinchoi@snu.ac.kr

feedback (PRF) is adopted. Previous studies evaluating the performance of PRF showed average precision improvements of about 14-16% over the unexpanded queries when tested on a small MEDLINE test collection [6,7].

Another issue exists when using the MEDLINE corpus; most information retrieval algorithms are evaluated on the Text Retrieval Conference (TREC) test collections. The characteristics of MEDLINE and TREC documents are very different in terms of co-occurrence, document length, and term distributions. Therefore, the effects of the algorithms, which have been proven effective on the TREC collection, also need to be verified on the MEDLINE collection.

We designed a set of comparative experiments in which various combinations of term ranking and term reweighting algorithms could be evaluated on the OHSUMED data set, a larger collection of MEDLINE documents than the previous collection, which was developed for testing information retrieval algorithms against physicians' information needs. The purpose of this study is to investigate the effects of query expansion algorithms within PRF on MEDLINE retrieval. In this paper, we describe the results of an evaluation that was performed using methods that range from classical to state-of-the-art term ranking algorithms.

This paper is organized as follows. Section 2 describes related works about automatic query expansion. Section 3 explains the baseline information retrieval system and the term-ranking algorithms that were chosen for the experiment, as well as the detailed term reweighting methods considered in the experimental design. Section 4 shows our experimental results. Section 5 discusses the common features of expanded terms found by different term ranking algorithms and the experimental results.

II. Related Work

A variety of approaches to automatic query expansion for improving the performance of MEDLINE retrieval queries have been previously studied. Hersh et al. [8] assessed the retrieval effectiveness of RF by using the Ide method on the OHSUMED test collection; this method gave better precision than that of the original queries at the level of fewer retrieved documents (i.e., 15, but not 100). For a small collection of 2,334 MEDLINE documents, Srinivasan [7] investigated PRF using different expansion strategies, including expansion on the MeSH query field, expansion on the free-text field alone, and expansion on both the MeSH and the free text fields. This author achieved significant improvement on retrieval effectiveness for all three expansion strategies over the original queries independently of the availability of relevant documents for feedback information. Recently, Yu et al. [9] suggested a multi-level RF system for PubMed, which let the user make three levels of relevance judgments on initial retrieved documents and then induced a relevance function from the feedback using the RankSVM. The system accuracy evaluation showed higher accuracy with less feedback than others in their study. States et al. [10] proposed an adaptive literature search tool based on an implicit RF that used information on citations that a user has viewed during search and browsing. In [11], using the OHSUMED test collection, authors studied a PRF technique based on a new variant of the standard Rocchio's feedback formula, which utilized a group-based term reweighting scheme.

Query expansion using the UMLS metathesaurus has produced mixed results. Aronson et al. [12] reported a 4% improvement in average precision over unexpanded queries on a small collection of 3,000 MEDLINE documents by mapping the text of both queries and documents to terms in the UMLS metathesaurus. Yang and Chute [13,14] and Yang [15] investigated a linear least square technique and expert network to map query terms to MeSH terms in the UMLS metathesaurus. The authors reported a 32.2% improvement of average precision on a small collection. Hersh et al. [16] assessed query expansion using synonym, hierarchical, and related term information, as well as term definition from the UMLS metathesaurus. All types of query expansion caused a decline in aggregated retrieval effectiveness on the OHSUMED test collection. Chu et al. [17] suggested a knowledge-based query expansion technique that only appended the original query with terms related to the scenario of the query, such as treatment and diagnosis, by using the UMLS metathesaurus and the UMLS semantic network. This achieved a 33% improvement in the average 11-point precision-recall over unexpanded queries for a subset of forty OHSUMED queries that belonged to five scenarios.

Because of increasing interest in automatic query expansion, which is based on the top retrieved documents, various approaches to select the best terms for query expansion have been suggested [18]. In previous studies [19-22], various types of comparisons of selected term ranking algorithms were made using the TREC test collections. When Rocchio's formula was used for reweighting expanded queries, the term ranking methods evaluated had no significant effect on the result [21]. However, performance improvement tended to be dependent on the test collection selected. Improvement was also dependent on the number of additional terms and the number of top retrieved feedback documents chosen for the experiment.

This study differs from previous work in three primary aspects. First, PRF is evaluated on the relatively large collection of 348,566 MEDLINE documents and 101 clinical queries from the OHSUMED. Because smaller sets of test documents and queries produce higher performance improvements, this study is valuable for verifying the effects of different algorithms of PRF on the larger test set of MEDLINE documents. Second, a comparison of different term ranking algorithms is performed. The spectrum of tests includes state-of-the-art term ranking algorithms, such as local context analysis (LCA) [23], as well as classic algorithms, including the Rocchio and Robertson selection value (RSV). Third, to identify the effect of the scores or ranks of expansion terms on the retrieval performance, term ranking algorithms are evaluated using varying term reweighting frameworks.

III. Methods

To compare several term ranking algorithms based on PRF, we developed a test-bed information retrieval system in which retrieved documents were sorted by the degree of relevance to queries. Based on the assumption that the top R documents initially retrieved are relevant, the automatic query expansion selects the top-ranked E terms from the pseudo-relevant documents. In this study, we evaluated the retrieval effectiveness of term ranking methods ranging from the classics to the state-of-the-art on the OHSUMED test collection. Because the importance of the terms in the expanded query is usually recalculated before submitting the query to the system for a second-pass retrieval, we evaluated the effectiveness of the combination of term ranking algorithms and term reweighting formulas in our experimental design.

1. OHSUMED Test Collection

We used OHSUMED [8] as a test collection. This collection is a subset of the MEDLINE database, which is itself a bibliographic database of important, peer-reviewed medical literature maintained by the National Library of Medicine (NLM). OHSUMED contains 348,566 MEDLINE references from 1987 to 1991 and 106 topics (queries) that were generated by actual physicians in the course of patient care. The references contain human-assigned subject headings from the Medical Subject Headings (MeSH), as well as titles and abstracts. Each query consists of an information need of physicians and a brief description of a patient. Relevance judgments corresponding to each query are provided using the scale of 'definitely relevant,' 'possibly relevant,' and 'not relevant.'

In our experiments, we limited the relevant documents to documents judged as definitely relevant; thus, only the 101 queries with at least one definitely relevant document were used. Each document is represented by combining the title, abstract, and MeSH fields. A query is generated from only the information needed field in the OHSUMED query because this query is the most similar to user queries issued in information retrieval systems.

2. Test-Bed Information Retrieval System

In the test-bed information retrieval system that we developed, text processing was performed through tokenizing, through removing stopwords using SMART [24] stopwords, and by applying Lovins' stemmer [25]. After the stemming process, each stemmed word was used as an index term for the inverted file. Once the inverted file was created with all of the indexed terms from the test collection, the query terms were matched against the indexing terms in a document to retrieve relevant documents. For our baseline retrieval model, we implemented the well-known Okapi BM25 weighting scheme [26]. In the Okapi BM25 formula, the initial top-ranked documents are retrieved by computing the similarity measure between a query q and a document d, as follows:

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

$$\text{with } w_{d,j} = \frac{(k_1 + 1) \cdot f_{d,f}}{K + f_{d,f}} \text{ and } w_{q,f} = \frac{(k_3 + 1) \cdot f_{q,t}}{k_3 + f_{q,t}} \cdot \log \frac{N - n + 0.5}{n + 0.5}$$

where t is a term of query q, n is the number of documents containing the term t across a document collection that contains N documents, and $f_{d,t}$ is the frequency of the term t in document d. K is $k_1((1\text{-}b) + b \times dl/avdl)$. The parameters k_1, b, and k_3 are set by default to 1.2, 0.75, and 1,000, respectively. The parameters dl and $avdl$ are the document length and the average document length, respectively, measured in some suitable unit (in this study, we used the byte length).

Once an ordered set of documents that match a given query the best is retrieved in the first run, the query is then automatically expanded using the top-ranked R documents (i.e., pseudo-relevant documents) in a PRF run. Finally, the reranked documents retrieved by the expanded query in the second run are shown to the user.

3. Term Ranking Algorithms

The process of PRF consists of two steps: query expansion (the addition of new terms selected from pseudo-relevant documents) and term reweighting (the modification of term weights) [27]. In the query expansion step, the term ranking algorithm is used to select the most useful terms from pseudo-relevant documents. Given a term occurring in pseudo-relevant documents, the term ranking algorithm returns a score reflecting the degree to which the term is meaningful.

To investigate the effect of the term ranking algorithm on retrieval effectiveness, we evaluated a wide range of algorithms classified as vector space models, probabilistic feedback models, or statistical models. Specifically, we performed a series of comparative analyses on eleven term ranking algorithms: total frequency (total freq), inverse document frequency (IDF), r_lohi [20], Rocchio, F4MODIFIED [18,28], expected mutual information measure (EMIM) [20], Robertson selection value (RSV) [29], Knullback-Leibler divergence (KLD) [21], CHI-squared (CHI2) [21], Doszkocs' variant of CHI-squared (CHI1) [21], and local context analysis (LCA) [23]. Table 1 describes the term scoring formulas of term *t* in which algorithms are grouped according to the categories of their underlying common features.

Once terms are sorted by one of the term ranking algorithms described above, either a fixed number of terms or terms above a certain threshold value are used as expansion terms. From the sorted list of terms, we selected the E highest-ranked new terms that were above a threshold of zero and added them to the original query.

4. Experimental Design

The process of PRF includes a term reweighting stage, as mentioned previously. During term reweighting, terms in the expanded query can be reweighted with or without consideration of the results of the term ranking algorithm used. Traditional methods, such as the standard Rocchio feedback formula [30] and the Robertson/Sparck-Jones weight method [26], reweight terms according to both the uniqueness of

Table 1. Term ranking algorithms and their formulas

Term ranking algorithm	Formula
Algorithms based on frequency heuristics	
total_freq	$\text{Score}(t) = f_{R,t}$, where $f_{R,t}$ is the total frequency of term *t* within the set of pseudo-relevant documents
IDF	$\text{Score}(t) = \log \frac{N}{n_t}$
r_lohi [1]	$\text{Score}(t) = r_t$ for ties, n_t in ascending order, where r_t is the number of pseudo-relevant documents containing term *t*
Algorithm based on vector space model	
Rocchio	$\text{Score}(t) = \sum_{\forall d \in \text{R}} w_{d,t}$
Algorithms based on distribution analysis	
F4MODIFIED [2,3]	$\text{Score}(t) = \log \frac{p_t}{1 - p_t} - \log \frac{q_t}{1 - q_t}$
EMIM [1]	$\text{Score}(t) = \sum_{i \in \{t_k, \bar{t}_k\}, j \in \{R, \bar{R}\}} P(i,j) \log \frac{P(i,j)}{P(i)P(j)}$
RSV [4]	$\text{Score}(t) = w_t \, (p_t - q_t)$
KLD [5]	$\text{Score}(t) = p_t \cdot \log \frac{p_t}{c_t}$
CHI2 [5]	$\text{Score}(t) = \frac{(p_t - c_t)^2}{c_t}$
CHI1 [5]	$\text{Score}(t) = \frac{(p_t - c_t)}{c_t}$
Algorithm based co-occurrence analysis	
LCA [6]	Measures the degree of co-occurrence of a given term t with all query terms.

p_t: the probability of occurrence of term *t* in the set of pseudo-relevant documents, q_t: the probability of occurrence of term *t* in the set of non-relevant documents, c_t: the probability of occurrence of term *t* in the whole document collection, w_t: the weight to be assigned to term *t*, EMIM: expected mutual information measure, F4: F4MODIFIED, IDF: inverse document frequency, KLD: Knullback-Leibler divergence, LCA: local context analysis, RSV: Robertson selection value.

terms in the pseudo-relevance documents and the probability of terms occurring in the pseudo-relevant documents.

In this study, we evaluated the retrieval effectiveness of term ranking algorithms with different formulas for term reweighting in order to compare the algorithms fairly, considering the scores and the ranks of the terms they produced.

For reweighting terms in the expanded query, we applied five different methods, including three popular approaches and two variants. In the following definitions, the formulas for calculating the new weight $w'_{q,t}$ of a query term t are described, in which $w_{q,t}$ is the weight of the term t in the unexpanded query, and the tuning parameters are set to a default value (i.e., $\alpha = \beta = 1$).

1) Standard Rocchio formula

A positive feedback strategy for a Rocchio formula with modified feedback [30] was applied, as follows:

$$w'_{q,t} = \alpha \cdot w_{q,t} + \frac{\beta}{R} \cdot \sum_{k=1}^{R} W_{k,t} \cdot \qquad (2)$$

where $w_{k,t}$ is the weight of the term t in a pseudo-relevant document k (which equals the $w_{d,t}$ component of the Okapi BM25 formula).

2) Ide regular formula

The positive feedback strategy of Ide [31] was applied, as follows:

$$w'_{q,t} = \alpha \cdot w_{q,t} + \beta \cdot \sum_{k=1}^{R} W_{k,t} \cdot \qquad (3)$$

3) Variant 1

As a variant of the standard Rocchio formula, the formula to employ the sorted result of term ranking algorithms in reweighting terms was suggested by Carpineto et al. [32] because the Rocchio considers the usefulness of a term with respect to the entire collection rather than its importance with respect to the user query. For the OHSUMED test collection, we applied the Rocchio formula, as follows:

$$w'_{q,t} = \alpha \cdot w_{q,t} + \beta \cdot max\_norm\_score_t \cdot \qquad (4)$$

The $max\_norm\_score_t$ is the normalized value assigned to term t by dividing the score of term t by the maximum score of the term ranking algorithm used. We called this a *max_norm* term reweighting.

4) Variant 2

The *max_norm* term reweighting does not accurately reflect the importance of the relative ranking orders of terms sorted by the term ranking algorithm. To understand the importance of their orders and to provide comparisons to the *max_norm* term reweighting, a formula to emphasize how well terms were ordered by the term ranking method used was devised by the authors, as follows:

$$w'_{q,t} = \alpha \cdot w_{q,t} + \beta \cdot rank\_norm\_score_t \cdot \qquad (5)$$

The $rank\_norm\_score_t$ is the evenly decreasing, normalized value assigned to term t according to the rank position of term t in the sorted term list and calculated by $1 - (rank_t - 1) / |\text{term\_list}|$, where $rank_t$ is the rank position of term t and $|\text{term\_list}|$ is the number of terms in the expanded query. We call this a *rank_norm* term reweighting.

5) Probabilistic term reweighting

In the probabilistic feedback model, we applied a modified Robertson/Sparck-Jones weight [26], which downgrades the weight of the expansion terms to 1/3, as follows:

$$\frac{1}{3} \times \log\left[\frac{(r + 0.5)/(R - r + 0.5)}{(n - r + 0.5)/(N - n - R + r + 0.5)}\right] \qquad (6)$$

In this model, the IDF component (i.e., $log((N - n + 0.5)/(n + 0.5))$) of the Okapi BM25 formula is replaced with the new weight.

The performance of paired combinations of the term ranking algorithms and the term reweighting methods described above was evaluated for both a different number of pseudo-relevant documents (R parameter) and a different number of expansion terms (E parameter). We varied the values for the R parameter from 5 to 50 with a step-size of 5. For the E parameter, we varied the value from 5 to 80, also with a step-size of 5, to see how the retrieval effectiveness varied for different parameter values.

5. Evaluation Measurements

To evaluate our experimental results, we primarily used the mean average precision (MAP) as the evaluation metric for retrieval effectiveness. MAP is the mean value of the average precisions computed to multiple queries where the average precision of each query is calculated by the average of the precision values over all of the retrieved relevant documents. MAP serves as a good measure of the overall ranking accuracy, and MAP favors systems that retrieve relevant documents early in the ranking [27]. In addition, we reported the precision of the top few retrieved documents, which reflects the user's perspective.

In all of the experiments, the measures were evaluated for the 100 top-ranked retrieved documents. The significance test was performed across multiple experiments using a paired *t*-test, which is one of the recommended methods for evaluating retrieval experiments [33].

IV. Results

In this paper, we primarily report the experimental results for two fixed parameter values. The experimental results for the standard values of the R and E parameters (R = 10 and E = 25) and the maximum performance values of R = 50 and E = 15, where the best MAP was found for the OHSUMED test collection, were chosen for detailed analysis. In our system, the unexpanded baseline MAP was 0.2163. This baseline was used as the reference for calculating performance improvement.

1. Performance at the Default Parameters

Using the standard parameter settings (R = 10, E = 25), the performance of selected term ranking algorithms over different reweighting methods was measured. The MAP and the percentage of improvement over the unexpanded queries are shown in Table 2. The significant differences in terms of the paired *t*-test are indicated by $p < 0.01$ and $p < 0.05$. As shown, using default parameter settings did not noticeably improve any of the term ranking methods for the

Table 2. Comparisons of different term ranking algorithms for different term reweighting methods (R = 10 and E = 25)

Rank	Reweight				
	Rocchio	Ide	Probabilistic	Max_norm	Rank_norm
CHI1	0.2226	0.2183	0.2087	0.2143	0.2227
	(+2.91)[a]	(+0.92)	(-3.51)	(-0.92)	(+2.96)
CHI2	0.2256	0.2227	0.2166	0.2083	0.2259
	(+4.30)[a]	(+2.96)	(+0.14%)	(-3.70)	(+4.44)
EMIM	0.2142	0.216	0.2205	0.2062	0.2204
	(-0.97)	(-0.14)	(+1.94)	(-4.67)	(+1.90)
F4MODIFIED	0.226	0.2225	0.2136	0.2259	0.2244
	(+4.48)[b]	(+2.87)	(-1.25)	**(+4.44)**[a]	(+3.74)
total_freq	0.2112	0.2115	0.2118	0.2005	0.2126
	(-2.36)	(-2.22)	(-2.08)	**(-7.30)**[a]	(-1.71)
IDF	0.2167	0.2055	0.2031	0.2186	0.2175
	(+0.18)	(-4.99)	(-6.10%)	(+1.06)	(+0.55)
KLD	0.2153	0.2154	0.2151	0.2006	0.2177
	(-0.46)	(-0.42)	(-0.55)	**(-7.26)**[a]	(+0.65)
LCA	0.2208	0.2225	0.2227	0.2205	0.2275
	(+2.08)	(+2.87)	(+2.96)	(+1.94)	(+5.18)
r_lohi	0.2102	0.2139	0.225	0.2065	0.2104
	(-2.82)	(-1.11)	(+4.02)	(-4.53)	(-2.73)
Rocchio	0.222	0.2219	0.2306	0.2112	0.2187
	(+2.64)	(+2.59)	(+6.61)	(-2.36)	(+1.11)
RSV	0.2254	0.226	0.2201	0.2149	0.2281
	(+4.21)	(+4.48)	(+1.76)	(-0.65)	(+5.46)

The mean average precision is presented for a combination of term ranking (rows) and term reweighting (columns) methods, including, in parentheses, the percent (%) improvement from the unexpanded queries.

EMIM: expected mutual information measure, F4: F4MODIFIED, IDF: inverse document frequency, KLD: Knullback-Leibler divergence, LCA: local context analysis, RSV: Robertson selection value.

[a]($p < 0.05$) and [b]($p < 0.01$) are in bold.

OHSUMED test collection. However, it was interesting that only the CHI1, CHI2, and F4MODIFIED term ranking algorithms, which favor infrequent terms, showed a statistically significant improvement when using Rocchio term reweighting. We analyzed the overlapping ratio of expansion terms between pair-wise term ranking algorithms. Figure 1 shows the top 15 overlapping ratios, where term ranking algorithms were linked according to the ratio of overlapping terms. As can be seen in the figure, although CHI1, CHI2, F4MODIFIED, and IDF found similar terms using their term ranking algorithms, IDF did not show a significant improvement. It appears that a few unique terms, expanded by a specific term ranking algorithm, may have a significant improving effect. In addition, because the performance of term ranking algorithms was differentiated by the term reweighting algorithms applied, both term ranking and reweighting methods should be taken into account when evaluating the performance of the PRF algorithms. With the default parameter setting, it is likely that, in conjunction with Rocchio term reweighting, the term ranking algorithms that favor infrequent terms can perform better in comparison to the other term ranking algorithms.

The *max_norm* and *rank_norm* term reweighting can explain the MAP differences in both the scores and the ranks of terms produced by different term ranking algorithms, as well as in the distinct terms selected by these algorithms. By comparing the *max_norm* and the *rank_norm* term reweighting methods from Table 2, we can see that rank-based normalizations are generally better than score-based ranking algorithms. It appears that the ranked order of terms is more important than the actual scores of the terms. Therefore, we performed a *t*-test comparison between pair-wise term ranking algorithms using *rank_norm* term reweighting. The *p*-values are given in Table 3; only *p*-values lower than 0.01 and 0.05 are reported. As shown, the RSV and the LCA performed best with the *rank_norm* reweighting method,

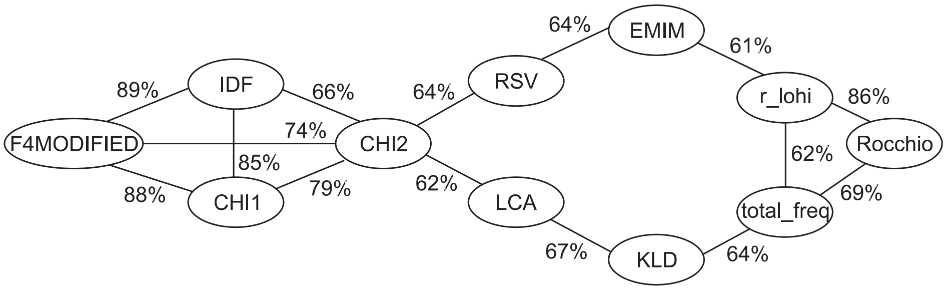

Figure 1. Percentage of average overlapping expansion terms for fifteen high-overlapping pairs of term ranking algorithms with the default parameter setting (R = 10, E = 25). IDF: inverse document frequency, EMIM: expected mutual information measure, LCA: local context analysis, KLD: Knullback-Leibler divergence, RSV: Robertson selection value.

Table 3. Results from the paired *t*-test between term ranking algorithms when the *rank_norm* term reweighting was applied (R = 10 and E = 25)

	CHI1	CHI2	EMIM	F4	total_freq	IDF	KLD	LCA	r_lohi	Rocchio
CHI2	-									
EMIM	-	-								
F4	-	-	-							
total_freq	-	-	-	-						
IDF	<0.05	-	-	<0.01	-					
KLD	-	-	-	-	-	-				
LCA	-	-	-	-	<0.05	-	-			
r_lohi	-	<0.05	<0.05	-	-	-	-	<0.01		
Rocchio	-	-	-	-	-	-	-	-	-	
RSV	-	-	-	-	<0.05	-	<0.05	-	<0.01	-

EMIM: expected mutual information measure, F4: F4MODIFIED, IDF: inverse document frequency, KLD: Knullback-Leibler divergence, LCA: local context analysis, RSV: Robertson selection value.

doi: 10.4258/hir.2011.17.2.120

significantly outperforming total_freq, r_lohi and KLD. This indicates that RSV and LCA performed better than the other algorithms at ranking the most useful terms near the top of the list using the default parameter settings. Although none of the term ranking methods resulted in a statistically significant improvement when applied in conjunction with the *rank_norm* term reweighting (Table 2), it may be valuable to note that most of the term ranking methods performed better for the rank_norm term reweighting than for the Rocchio term reweighting.

2. Performance at the Maximum Performance Parameters

The maximum performance values of the R and E parameters were fixed at R = 50 and E = 15 because these parameter values gave the best performance improvement for the OHSUMED test collection. Table 4 shows the comparative results for the parameters, and Table 5 gives the *t*-test comparisons between pair-wise term ranking algorithms for the results of the *rank_norm* term reweighting.

A comparison of Tables 2 and 4 indicates that the LCA term ranking shows the best performance, regardless of the term reweighting algorithm applied. The performance improvement, about 12%, was achieved by LCA when the expanded query was re-weighted by *rank_norm*. Furthermore, for *rank_norm* term reweighting, LCA significantly outperformed all of the other methods (Table 5). This suggests that LCA can select more useful terms when the pseudo-relevant documents provided are large enough to infer co-occurrence

Table 4. Comparison of different term ranking algorithms for different term reweighting methods (R = 50 and E = 15)

Rank	Reweight				
	Rocchio	Ide	Probabilistic	Max_norm	Rank_norm
CHI1	0.2169	0.1741	0.1805	0.2139	0.2146
	(+0.28)	**(-19.51)**[b]	**(-16.55)**[b]	(-1.11)	(-0.79)
CHI2	0.2255	0.1848	0.1932	0.2179	0.2245
	(+4.25)[a]	**(-14.56)**[b]	(-10.68)	(+0.74)	(+3.79)
EMIM	0.2133	0.1871	0.2156	0.2115	0.2254
	(-1.39)	(-13.50)	(-0.32)	(-2.22)	(+4.21)
F4MODIFIED	0.2166	0.1718	0.1843	0.2137	0.2149
	(+0.14)	**(-20.57)**[b]	**(-14.79)**[b]	(-1.20)	(-0.65)
total_freq	0.2143	0.1802	0.2027	0.2054	0.2164
	(-0.92)	**(-16.69)**[b]	(-6.29)	(-5.04)	(+0.05)
IDF	0.2152	0.1694	0.1764	0.2127	0.212
	(-0.51)	**(-21.68)**[b]	**(-18.45)**[b]	(-1.66)	(-1.99)
KLD	0.2174	0.1838	0.2141	0.2014	0.2237
	(+0.51)	(-15.03)	(-1.02)	**(-6.89)**[a]	(+3.42)
LCA	0.2271	0.2054	0.23	0.2395	0.242
	(+4.99)[a]	(-5.04)	(+6.33)	**(+10.73)**[b]	**(+11.88)**[b]
r_lohi	0.212	0.1829	0.2032	0.2128	0.2082
	(-1.99)	**(-15.44)**[a]	(-6.06)	(-1.62)	(-3.74)
Rocchio	0.2143	0.1799	0.2037	0.2118	0.2136
	(-0.92)	**(-16.83)**[a]	(-5.83)	(-2.08)	(-1.25)
RSV	0.2172	0.1918	0.2185	0.2017	0.2273
	(+0.42)	(-11.33)	(+1.02)	**(-6.75)**[a]	(+5.09)

The mean average precision is presented for a combination of term ranking (rows) and term reweighting methods (columns), including, in parentheses, the percent (%) improvement from the unexpanded queries.

EMIM: expected mutual information measure, F4: F4MODIFIED, IDF: inverse document frequency, KLD: Knullback-Leibler divergence, LCA: local context analysis, RSV: Robertson selection value.

[a]($p < 0.05$) and [b]($p < 0.01$) are in bold.

Table 5. Results of the paired *t*-test between term ranking algorithms when *rank_norm* term reweighting was applied (R = 50 and E = 15)

	CHI1	CHI2	EMIM	F4	total_freq	IDF	KLD	LCA	r_lohi	Rocchio
CHI2	<0.01									
EMIM	-	-								
F4	-	<0.01	-							
total_freq	-	-	-	-						
IDF	-	<0.01	-	<0.05	-					
KLD	-	-	-	-	-	-				
LCA	<0.01	<0.01	<0.05	<0.01	<0.01	<0.01	<0.01			
r_lohi	-	<0.01	<0.05	-	-	-	<0.05	<0.01		
Rocchio	-	-	<0.05	-	-	-	-	<0.01	<0.05	
RSV	-	-	-	-	<0.05	-	-	<0.05	<0.01	<0.05

EMIM: expected mutual information measure, F4: F4MODIFIED, IDF: inverse document frequency, KLD: Knullback-Leibler divergence, LCA: local context analysis, RSV: Robertson selection value.

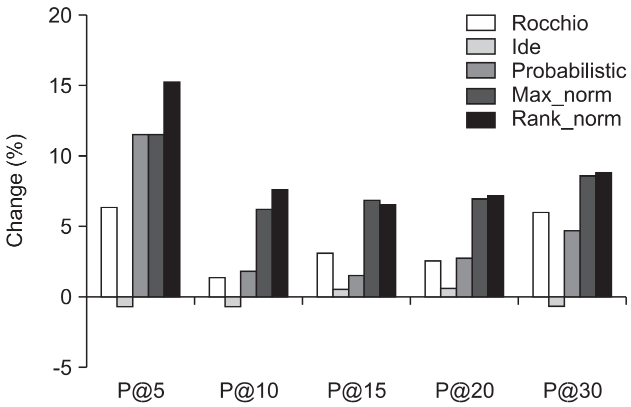

Figure 2. Comparison of different term reweighting algorithms for queries expanded with the local context analysis term ranking algorithm in terms of the percentage change of precision in the top 5, 10, 15, 20, and 30 retrieved documents from the original queries.

of terms. Throughout our experiments, LCA consistently performed the best at automatic query expansion from the set of OHSUMED retrieved documents.

With regard to the effect of term reweighting algorithms, the performance change of precision in top retrieved documents and the performance of MAP are shown in Figure 2. As shown in the figure, when comparing the traditional term reweighting methods, such as Rocchio, Ide, and probabilistic, the methods based on the results of term ranking algorithms, such as *max_norm* and *rank_norm*, showed better improvements with fewer retrieved documents.

V. Discussion

Ranking and reweighting of terms are the main processes of PRF. The basic features to be considered for PRF algorithms are the frequency heuristic and the probabilistic and statistical analysis of the terms. Because the general characteristics of medical documents are different from the characteristics of other collections, we have attempted to evaluate the retrieval performance of ranking and reweighting method pairs for the OHSUMED collection, which is a subset of the MEDLINE documents.

In our search for the best PRF algorithm for the medical area, we examined the core effects of term ranking algorithms used for selecting expansion terms. Table 2 shows two interesting results. The first result shown is the improved performance of the CH1, CH2, and F4MODIFIED term ranking algorithms for Rocchio term reweighting using the default parameter settings. These algorithms favor infrequent terms in the collection for expansion. In addition, it is likely that a few unique terms, expanded by a specific term ranking algorithm, may significantly improve the performance. However, the effect was not consistent with all of the other reweighting algorithms, such as Ide, probabilistic, *max_norm*, and *rank_norm*.

The second interesting result is the performance of LCA using the maximum performance parameter settings (R = 50, E = 15). LCA showed consistent improvement across all of the reweighting algorithms, except for the Ide (Table 4). The LCA performance result was statistically significant compared to all other ranking algorithms. This result suggests that LCA can select more useful terms when enough pseudo-relevant documents are provided to infer co-occurrence of terms in the whole collection. Moreover, the effect was magnified as the number of included documents increased.

On average, OHSUMED contains approximately 22 documents found to be definitely relevant for each of the 101 queries. Accordingly, a large number of non-relevant documents can be included as the number of pseudo-relevant documents increases. However, despite a large number of non-relevant documents, LCA showed a notable improvement in our experiments. Our results with LCA show that term co-occurrence played a more important role compared to other features, in a medical context. Thus, the co-occurrence feature should be seriously considered in the design of clinical query systems.

The limitations of our study are that we used the outdated OHSUMED test collection because of non-availability of modern test collections for evaluating retrieval algorithms against real clinical queries and that our findings might not be generalizable to real data collections because of our ad-hoc experiments of only one test collection. Although further experiments should be performed in the future, out findings are important in understanding the behaviors of various term ranking and reweighting algorithms on a subset of MEDLINE documents and clinical queries.

New comparative experiments on term ranking algorithms were performed in the context of a subset of MEDLINE documents. Among the various term ranking algorithms, LCA significantly outperformed all of the other methods in our experiments when the top 15 terms obtained from 50 retrieved documents were expanded with a modified weight by a rank_norm reweighting method. With medical documents, LCA, which uses co-occurrence with all query terms, significantly outperformed various term ranking methods based on both frequency and distribution analyses. To maximize the performance improvement even further, the weight of the terms in the expanded query should be appropriately adjusted. Furthermore, the results of the experiments demonstrate that the term rank-based reweighting method contributed to a remarkable improvement in mean average precision.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgements

This work was supported in part by the Advanced Biomedical Research Center (ABRC) funded by KOSEF and in part by the National Research Foundation of Korea (NRF) grant, which was funded by the Korean government (MEST) (No. 2009-0075089).

References

1. Hersh WR. Information retrieval: a health and biomedical perspective. 2nd ed. New York: Springer; 2003.
2. Nankivell C, Wallis P, Mynott G. Networked information and clinical decision making: the experience of Birmingham Heartlands and Solihull National Health Service Trust (Teaching). Med Educ 2001; 35: 167-172.
3. Haux R, Grothe W, Runkel M, Schackert HK, Windeler HJ, Winter A, Wirtz R, Herfarth C, Kunze S. Knowledge retrieval as one type of knowledge-based decision support in medicine: results of an evaluation study. Int J Biomed Comput 1996; 41: 69-85.
4. Bernstam EV, Herskovic JR, Aphinyanaphongs Y, Aliferis CF, Sriram MG, Hersh WR. Using citation data to improve retrieval from MEDLINE. J Am Med Inform Assoc 2006; 13: 96-105.
5. Vechtomova O, Wang Y. A study of the effect of term proximity on query expansion. J Inf Sci 2006; 32: 324-333.
6. Srinivasan P. Exploring query expansion strategies for MEDLINE. Technical Report 1528. Ithaca, NY: Cornell University; 1995.
7. Srinivasan P. Retrieval feedback in MEDLINE. J Am Med Inform Assoc 1996; 3: 157-167.
8. Hersh W, Buckely C, Leone TJ, Hickam D. 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. Dublin, IE: Springer-Verlag Inc.; 1994. p192-201.
9. Yu H, Kim T, Oh J, Ko I, Kim S, Han WS. Enabling multi-level relevance feedback on PubMed by integrating rank learning into DBMS. BMC Bioinformatics 2010; 11 Suppl 2: S6.
10. States DJ, Ade AS, Wright ZC, Bookvich AV, Athey BD. MiSearch adaptive PubMed search tool. Bioinformatics 2009; 25: 974-976.
11. Yoo S, Choi J. On the query reformulation technique for effective MEDLINE document retrieval. J Biomed Inform 2010; 43: 686-693.

12. Aronson AR, Rindflesch TC, Browne AC. Exploiting a large thesaurus for information retrieval. In: Proceedings of Recherche d'Information Assistee par Ordinateur (RIAO). New York: ACM; 1994. p197-216.
13. Yang Y, Chute CG. An application of least squares fit mapping to text information retrieval. In: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Pittsburgh, PA: ACM; 1993. p281-290.
14. Yang Y, Chute CG. Words or concepts: the features of indexing units and their optimal use in information retrieval. Proc Annu Symp Comput Appl Med Care 1993: 685-689.
15. Yang Y. Expert network: effective and efficient learning from human decisions in text categorization and retrieval. In: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Dublin, IE: Springer-Verlag Inc.; 1994. p13-22.
16. Hersh W, Price S, Donohoe L. Assessing thesaurus-based query expansion using the UMLS metathesaurus. In: Proceeding of the AMIA Symposium. Los Angeles, CA: Hanley & Belfus; 2000. p344-348.
17. Chu WW, Liu Z, Mao W, Zou Q. A knowledge-based approach for retrieving scenario-specific medical text documents. Control Eng Pract 2005; 13: 1105-1121.
18. Efthimiadis EN. Query expansion. Ann Rev Inf Sci Tech 1996; 31: 121-187.
19. Harman D. Relevance feedback revisited. In: Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Copenhagen, DK: ACM; 1992. p1-10.
20. Efthimiadis EN, Brion PV. UCLA-Okapi at TREC-2: query expansion experiments. In: Proceedings of 2nd Text Retrieval Conference. Gaithersburg, MD: National Institute of Standards and Technology; 1994. p200-215.
21. Carpineto C, de Mori R, Romano G, Bigo B. An information theoretic approach to automatic query expansion. ACM Trans Inf Syst 2001; 19: 1-27.
22. Fan W, Luo M, Wang Li, Xi W, Fox EA. Tuning before feedback: combining ranking discovery and blind feedback for robust retrieval. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Sheffield, UK: 2004. New York, NY: ACM; 2004. p138-145.
23. Xu J, Croft WB. Improving the effectiveness of information retrieval with local context analysis. ACM Trans Inf Syst 2000; 18: 79-112.
24. Buckley C. Implementation of the SMART information retrieval system. Technical Report TR85-686. Ithaca, NY: Cornell University; 1985.
25. Lovins JB. Development of a stemming algorithm. Mech Transl Comput Linguist 1968; 11: 22-31.
26. Robertson SE, Walker S. Okapi/Keenbow at TREC-8. In: Proceedings of the 8th Text Retrieval Conference (TREC-8). Gaithersburg, MD: National Institute of Standards and Technology; 2000. p151-162.
27. Baeza-Yates R, Ribeiro-Neto B. Modern information retrieval. New York: ACM Press; 1999.
28. Robertson SE. On relevance weight estimation and query expansion. J Doc 1986; 42: 182-188.
29. Robertson SE. On term selection for query expansion. J Doc 1990; 46: 359-364.
30. Rocchio JJ. Relevance feedback in information retrieval. In: Salton G, ed. The SMART retrieval system: Englewood Cliffs, NJ: Prentice-Hall; 1971: p313-323.
31. Ide E. New experiments in relevance feedback. In: Salton G, ed. The SMART retrieval system: Englewood Cliffs, NJ: Prentice-Hall; 1971: p337-354.
32. Carpineto C, Romano G, Giannini V. Improving retrieval feedback with multiple term-ranking function combination. ACM Trans Inf Syst 2002; 20: 259-290.
33. Hull D. Using statistical testing in the evaluation of retrieval experiments. In: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Pittsburgh, PA: ACM; 1993. p329-338.

doi: 10.4258/hir.2011.17.2.120