A Study of the Effect of Term Proximity on Query Expansion

Olga Vechtomova

Department of Management Sciences, University of Waterloo, Waterloo, Canada

Ying Wang

Toshiba of Canada Limited, Markham, Canada[1]

Correspondence to: Olga Vechtomova, 200 University Avenue West, Waterloo, Canada.
ovechtom@uwaterloo.ca

Abstract

Query expansion terms are often used to enhance original query formulations in document retrieval. Such terms are usually selected from the entire documents or from windows or passages surrounding query term occurrences. Arguably, the semantic relatedness between terms weakens with the increase in the distance separating them. In this paper we report a study that was conducted to systematically evaluate different distance functions for selecting query expansion terms. We propose a distance factor that can be effectively combined with the statistical term association measure of Mutual Information for selecting query expansion terms. Evaluation on TREC collection shows that the distance-weighted Mutual Information is more effective than Mutual Information alone in selecting terms for query expansion.

Keywords: information retrieval; query expansion; term proximity; word collocation; mutual information

[1] This work was done while Ying Wang was a Masters student in the Department of Management Sciences, University of Waterloo, Waterloo, Canada.

1. Introduction

Query expansion is a technique commonly used in Information Retrieval [1, 2] to improve the retrieval performance by reformulating the original query - either adding new terms or reweighting the original terms. Query expansion terms can be automatically extracted from the documents or taken from knowledge resources, such as thesauri. The advantages of the former techniques are that expansion terms are extracted from the collection, thereby there is no vocabulary mismatch problem, and secondly, no expensive human-constructed knowledge bases are required. Typically either top-ranked documents in the initially retrieved document set (blind or pseudo-relevance feedback) or documents judged relevant by the user in the retrieved set (relevance feedback) are used to extract query expansion terms. For short and incomplete queries, a substantial improvement can be achieved by using expanded queries [3, 4]. Query expansion terms ideally should have the following characteristics: (a) be semantically related to the original query terms; (b) be good at discriminating between relevant and non-relevant documents.

Different approaches to select query expansion terms following relevance or pseudo-relevance feedback have been proposed. For instance, in a probabilistic model of IR [4] query expansion terms are selected on the basis of how well they can discriminate between the relevant and non-relevant documents. There have been also attempts to select query expansion terms on the basis of the strength of their association with the original query terms. There exist several statistical association measures to identify significant word associations (also referred to as co-occurrences or collocations), and these measures are commonly used in corpus linguistics to estimate the degree of semantic closeness between words. Such measures include the Mutual Information (MI), t-score, Log-Likelihood and Chi-square test [5]. These measures have been used to identify significant word collocations in a variety of applications, such as, multiword unit extraction [6], speech recognition [7], information retrieval [8], lexicography and lexical analysis [9], word sense disambiguation [10], and analysis of aligned corpora [11]. A few studies have been conducted to evaluate the use of term association measures in query expansion to select words that are closely related to query terms [12, 13]. A common approach in these studies is to select query expansion terms from either the entire document or from sections of the document (passages or windows) surrounding the occurrences of query terms. One disadvantage of such approaches is that a decision needs to be made about the span of text surrounding the occurrences of query terms from which query expansion terms are extracted. Span lengths are usually determined empirically, and the optimal span length may vary depending on the collection. Also, within these spans, the association scores of query expansion terms are not affected by how far they are from the original query terms. A study by [14] demonstrates, however, that the strength of association between words decays exponentially with the increase in distance. It is, therefore, possible that more accurate estimates of the strength of association between two words can be achieved, if term association measures are combined with a decaying distance factor.

The term *collocation* has been used differently in the literature. For example, Manning and Schütze understand word collocations as grammatically bound elements occurring in a certain order and having limited compositionality [5]. On the other hand, Hoey [15] understands collocation as long-distance relationships between semantically related words. We recognise the existence of two types of collocation:

- Collocation due to lexical-grammatical or habitual restrictions. These restrictions limit the choice of words that can be used in the same grammatical structures. Collocations of this type occur within short spans, i.e. within the bounds of a syntactic structure, such as a noun phrase (e.g. "rancid butter", "white coffee", "mad cow disease", "stainless steel").
- Collocation due to a typical occurrence of a word in a certain thematic environment: two words hold a certain lexical-semantic relation, i.e. their meanings are close semantically, therefore they tend to occur in the same topics in texts. These type of collocations may span over longer distances in text than collocations of the previous type. Examples of some of the collocates of this type, identified using the Z-score statistic, are "nitrogen–pollution", "school – education" [12].

In our approach to the selection of query expansion terms, no explicit distinction is made between these two types of collocates, therefore collocates of both types can potentially be selected for query expansion using the proposed method. However, we assume that the closer a word occurs to the user's query term in text, the more likely it is related to it semantically, and propose a method that rewards words co-occurring closer to the instances of the original user's query terms.

Term proximity has been explored extensively in document ranking studies [16, 17, 18, 19, 20], where several distance factors were proposed. Two common intuitions underlie these approaches: (1) the closer the terms are in a document, the more likely they are topically related, and (2) the closer the query terms are in a document, the more likely the document is relevant to the query. The reasoning behind these intuitions is that a document may contain different topics or talk about different aspects of the same topic, therefore, if terms occur close to each other they are more likely used in related senses and discuss the same subject than if they are located in different parts of the document. For example, Clarke et al. [16] proposed a technique of scoring documents based on term proximity and density. They introduced the notion of *cover*, which is the shortest span of text containing instances of all query terms. Document score is calculated based on two assumptions: (1) the shorter the cover, the more likely the corresponding document is relevant, and (2) the more covers are in a document, the more likely the document is relevant. The evaluation on TREC data set proved the effectiveness of the method compared to some standard approaches to QE. A similar technique was proposed by Hawking and Thistlewaite [17, 18], which also demonstrated promising results on TREC data set. Approaches to document ranking based on term proximity are particularly suitable for distributed information retrieval systems, as they do not rely on collection-wide statistics such as inverse document frequency and average document length.

A related approach that utilises term proximity implicitly in query expansion is the use of context-independent or query-biased summaries for the selection of query expansion terms. For example Lam-Adesina and Jones [21] have evaluated query-biased and context-independent summarisation techniques for QE following pseudo-relevance feedback and achieved improvements over standard QE methods using complete documents on TREC data set.

Gao et al. [22] proposed a combined measure of Mutual Information with an exponential decaying distance factor, which they evaluated for query translation task in CLIR (Cross-language IR). Their approach will be discussed in more detail in section 2.2. The goal of our study is to systematically evaluate a number of decaying distance factors in combination with the term association measure of Mutual Information, and their effectiveness in query expansion. We propose a new decaying distance function that shows improved performance over the distance factor proposed by Gao et al. [22].

The following two hypotheses were formulated and investigated in this study:

Hypothesis 1:

Query expansion terms ranked by a combination of Mutual Information with a decaying distance factor lead to a significant performance improvement over the original query terms.

Hypothesis 2:

Query expansion terms ranked by a combination of Mutual Information with a decaying distance factor lead to a significant performance improvement over the query expansion terms ranked by Mutual Information alone.

The rest of the paper is organised as follows: in section 2 we present the experiments conducted in this study, section 3 contains the analysis of results, and section 4 concludes the paper and outlines future work.

2. Experiments

2.1. Collections and Evaluation

Okapi was used as the testbed IR system in this study [3]. Experiments were conducted on TREC document collections FT (*Financial Times*) and LA (Los Angeles *Times*)[2], and query topics 301 – 450 from the ad hoc tracks of TREC-6, TREC-7 and TREC-8 [23]. Query terms were taken from the "Title" section of the topics, as they most closely resemble the queries users formulate in real search scenarios. The total size of the document collection is 342,054 documents with the average document size of 333 words. The retrieval performance results were averaged over 150 topics. Retrospective evaluation using relevance feedback was conducted, whereby the

original queries were expanded with the terms extracted from the top 10 ranked relevant documents, and expanded queries were applied to the same set of documents (the whole collection) from which the documents used for query expansion were derived. The experimental runs were also replicated using frozen-rank evaluation technique to test the effectiveness of the QE methods in retrieving unseen relevant documents.

To retrieve a ranked document set in response to query terms from the "Title" section of TREC topics, Okapi BM25 function was used. Tuning constant k_1 (controlling the effect of within-document term frequency) was set to 1.2 and b (controlling document length normalisation) was set to 0.75 [4].

2.2. Query expansion term selection

The assumption behind this work is that semantically related words are usually located in proximity, and the distance between two words could indicate the strength of their association. The goal of our experiments is to explore the role of term proximity in selecting query expansion terms.

Gao et al. [22] introduced a decaying co-occurrence model used for cross-language query translation. The association measure they proposed was used in selecting a translation word for a query term, that was most strongly associated with other query terms' translations. To calculate the strength of association of a query term y with a set of other query terms T, they introduced the notion of Cohesion (Equation 1) and a pairwise term similarity score $SIM(x, y)$, which is a combination of Mutual Information and a distance factor (Equation 2). The distance factor exponentially decreased as the distance between terms increased (Equation 3).

$$Cohesion(y,T) = \log(\sum_{x \in T} SIM(x,y)) \quad (1)$$

$$SIM(x,y) = MI(x,y) * df(x,y) \quad (2)$$

$$df(x,y) = e^{-\alpha*(D(x,y)-1)} \quad (3)$$

$D(x, y)$ is the average distance between words x and y in all the documents in the corpus; α is the decaying rate for the exponential function. The decaying rate α=0.8 demonstrated the best performance [22].

[2] TREC Research Collection Volumes 4 and 5.

In our study, Cohesion score and Similarity score were formulated in the same way as those in [22]. *D(x, y)* was calculated as the average distance between query terms *x* and *y* in the relevant set, i.e. the top-ranked documents retrieved in response to the user's query and judged relevant by the user. As our goal is to explore the use of distance in ranking query expansion terms, we investigated other distance functions.

Mutual Information

The (pointwise) mutual information (MI), which has its origins in the information theory [24], is a measure for discovering interesting word collocations [9]. The mutual information score between a pair of words compares the probability that the two words occur as a joint event with the probability that they occur individually and that their co-occurrences are simply a result of chance [11]. The mutual information score grows with the increase in frequency of word co-occurrence. If two words co-occur mainly due to chance their mutual information score will be close to zero. If they occur predominantly individually, then mutual information will be a negative number.

The standard formula for calculating Mutual Information is:

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)} \qquad (4)$$

where P(x,y) is the probability that words x and y occur together;

P(x) and P(y) are the probabilities that x and y occur individually.

Mutual Information (Equation 4) is usually applied where term *x* immediately follows term *y* in text, e.g. as used in [9]. The probability that two words occur as a joint event $P(x,y)$ is estimated as $f(x,y)/N$, where joint frequency $f(x,y)$ denotes the number of times that *y* appears immediately after *x*. As we deal with collocates which can occur on either sides of the query term within a distance of more than one word, the original MI formula is not suitable. Therefore, we used the MI statistic as proposed in [12], which provides for unordered co-occurrence within a distance of more than one word (Equation 5). For more details and the justification of the formula see [12].

$$MI(x,y) = \log_2 \frac{\frac{f_r(x,y)}{R * V_x(D)}}{\frac{f_r(x)}{R} * \frac{f_c(y)}{N}} \qquad (5)$$

Where: $f_r(x, y)$ - the joint frequency of pair (x, y) in the known relevant documents;

$f_c(y)$ – frequency of y in the corpus;

$f_r(x)$ – frequency of x in the known relevant documents;

$V_x(D)$ – average length of the known relevant documents;

N – corpus size;

R – number of words in the known relevant documents.

2.3. Exponential Distance Factor

In our query expansion experiments, the following parameters were used. These parameters demonstrated best performance in the preliminary set of experiments [25, 26]:

- Top 10 relevant documents were used in relevance feedback. These documents were taken from the ranked document set retrieved using the user's original query, and their relevance was determined using TREC relevance judgements.
- The original user's query was expanded by top 20 ranked collocates.
- Collocates were extracted from the entire document.

Two baselines were set in our experiments: No_Expan is the run with only the original queries, and is used as the baseline to test Hypothesis 1; MI is the run using Mutual Information (Equation 5) alone as the Similarity function in Equation 2 (i.e. without a distance factor) to rank query expansion terms. It is used as the baseline run to test Hypothesis 2. The run No_Expan has the Average Precision (AveP) of 0.2426, and MI has the AveP of 0.3872.

First, the effect of using the exponential distance factor proposed by Gao et al. on ranking query expansion terms was tested (run MID_Exp in Table 1). The average precision of MID_Exp (0.3030) is slightly higher than No_Expan, and significantly lower than the average precision of the MI run (0.3872). After the results were analysed further, it was found that the average distance between a pair of terms was usually longer than 20 words and it resulted in some extremely small values of the distance factor. The exponential distance is likely to be compatible with the method that restricts collocates and query terms within smaller windows, for example, sentences. If the distance between a pair is always within the span of a sentence, the exponential distance factor would be more reasonable to use.

The results of MID_Exp indicate that the exponential distance factor as proposed in [22] does not work well for ranking query expansion terms. Therefore, alternative distance factors are proposed and evaluated in this work.

2.4. Logarithm Distance Factors

The exponential distance factor could be replaced by a logarithm-based distance factor, such as the following:

$$df(x, y) = \log_2(1 + 1/D(x, y)) \quad (6)$$

In Equation 6, *df(x, y)* is always within the range (0, 1] as *D(x, y)* is always greater than or equal to 1. Similarly, another distance factor (Equation 7) was proposed.

$$df(x, y) = \log_2(2 + 1/D(x, y)) \quad (7)$$

The *df(x, y)* in Equation 7 is always within the range (1, $\log_2 3$]. It emphasises the importance of *MI(x, y)* in the Similarity score (see Formula 2) and de-emphasises the distance factor *df(x, y)*.

Two runs were conducted: MID_lgd, using Equation 6 as the distance factor, and MID_lgd2, using Equation 7. Performance measures of these two runs are presented in Table 1. The Average Precision, and Precision at 10, 15, 20, 30 and 100 documents are significantly higher in MID_lgd2 compared to MID_lgd. However, they are not significantly different from those of the MI run. One problem in the above distance factors could be that the average distance between query terms in the relevant set, *D(x, y)*, is calculated in the same way for frequent and infrequent word pairs. If two words co-occur infrequently, for example, only once in the local set, the average distance is still used to calculate their association score. However, Mutual Information scores for infrequent pairs are not reliable: a pair of low-frequency words will get a higher MI score than a pair of high-frequency words, with all other parameters being equal [5]. Manning and Schütze reported two possible solutions that have been proposed: (1) to use a cutoff and to only consider collocations with a frequency of at least 3; (2) to multiply the MI score of a pair of words by their joint frequency *f(x,y)*. [5] They note that the first solution does not really remove the problem, but simply reduces its effect. In order to emphasise the frequency of word pairs in the local document set, we introduce the joint frequency into the distance formula, (Equation 8).

$$df(x, y) = \log_2(2 + f_r(x, y)/D(x, y)) \quad (8)$$

Where $f_r(x, y)$ is the joint frequency in the relevant document set.

The distance factor in Equation 8 (run MID_lgd3 in Table 1) puts more emphasis on frequent words, thereby reducing the problem of overweighting low-frequency words. The performance of MID_lgd3 is somewhat higher

than that of MID_lgd2. To test the effect of the constant in the logarithm function, Equation 9 (run MID_lgd4 in Table 1) was further proposed, which sets the constant as 3. The performance of MID_lgd4 is however worse than that of MID_lgd3.

$$df(x,y) = \log_2(3 + f_r(x,y)/D(x,y)) \quad (9)$$

Next, we evaluated the distance function without the logarithm format (Equation 10).

$$df(x,y) = f_r(x,y)/D(x,y) \quad (10)$$

The run MID_d5 was conducted, and its performance results are presented in Table 1. After continuous improvement of the distance factor, the best performance was obtained by using Equation 10. The results suggest that the distance factor *df(x, y)* performs better without the logarithm format of *D(x, y)* and $f_r(x, y)$. However, it is not clear what effect term proximity alone has on the ranking of query expansion terms. Therefore, we evaluated separately the effect of the joint frequency (Equation 11, Run MID_d6) and the effect of the distance as a linear function (Equation 12, Run MID_d7) on the performance.

$$df(x,y) = f_r(x,y) \quad (11)$$

$$df(x,y) = 1/D(x,y) \quad (12)$$

As seen from Table 1, MID_d6 performs worse than MID_d5, suggesting that term proximity information has some positive effect on query expansion term selection. On the other hand, linear reduction of the MI score with the increase in the average distance between two terms (Equation 12, Run MID_d7) leads to worse performance than combination of the joint frequency and the inverse average distance. This result suggests that it is necessary to reward high-frequency collocations to compensate for the MI bias towards low-frequency collocations. The 11-point precision-recall graphs of runs using different distance factors are presented in Figure 1. For comparison we also show in Table 1 the results obtained by using Robertson selection value (RSV) [27], which is a well-known QE term selection method demonstrating consistently high performance in TREC experiments. MID_d5 method does not perform better than RSV in these experimental settings. Our goal was to investigate the effect of term distance on the word association measure of Mutual Information. The evaluation demonstrates that for the QE task a combination of MI with the frequency and distance factors selects overall better terms than MI alone. A combination of MI with the frequency and distance factors may also be useful for other tasks in which MI is currently used, such as construction of lexical resources and word sense disambiguation.

It is noteworthy to mention that the evaluation was conducted on a collection of rather short documents from the newswire corpora, and the distance factor may make a stronger contribution to the quality of selected terms in collections of longer documents.

Table 1. Performance of the baseline and experimental query expansion runs (retrospective evaluation)

Run	Distance factor formula	AveP	P@5	P@10	P@15	P@20	P@30	P@100	R-Prec
No_Expan		0.2426	0.4453	0.3887	0.3476	0.3157	0.2740	0.1584	0.2795
MI		0.3872	0.8293	0.6400	0.5333	0.4650	0.3818	0.1938	0.3898
MID_Exp	$e^{-\alpha*(D(x,y)-1)}$	0.3030	0.6987	0.5387	0.4409	0.3810	0.3051	0.1647	0.3213
MID_lgd	$\log_2(1+1/D(x,y))$	0.3451	0.7907	0.6060	0.4871	0.4187	0.3371	0.1736	0.3533
MID_lgd2	$\log_2(2+1/D(x,y))$	0.3877	0.8307	0.6447	0.5360	0.4667	0.3809	0.1928	0.3910
MID_lgd3	$\log_2(2+f_r(x,y)/D(x,y))$	0.4220	0.8360	0.6727	0.5724	0.5013	0.4209	0.2138	0.4243
MID_lgd4	$\log_2(3+f_r(x,y)/D(x,y))$	0.4071	0.8333	0.6573	0.5493	0.4787	0.3996	0.2047	0.4085
MID_d5	$f_r(x,y)/D(x,y)$	0.4341	0.8187	0.6800	0.5884	0.5220	0.4342	0.2287	0.4247
MID_d6	$f_r(x,y)$	0.4053	0.7827	0.6407	0.5524	0.4960	0.4180	0.2251	0.4103
MID_d7	$1/D(x,y)$	0.3436	0.7947	0.5973	0.4822	0.4173	0.3367	0.1725	0.3521
RSV		0.4531	0.8480	0.7027	0.6062	0.5373	0.4478	0.2313	0.4501

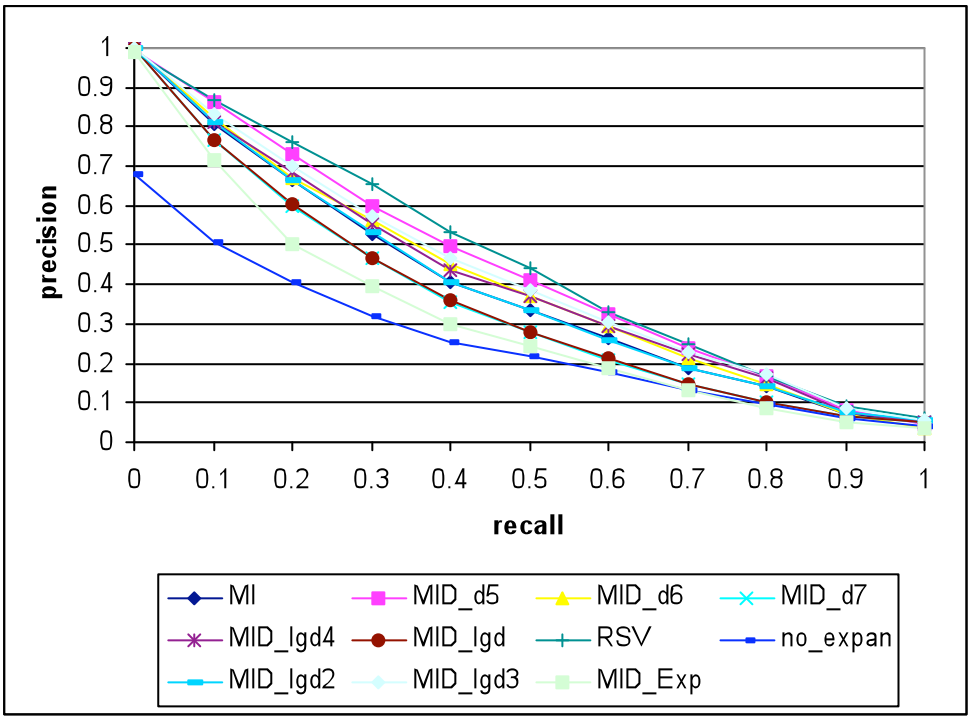

Fig. 1. Precision at 11-point recall levels of the baseline and experimental query expansion runs (retrospective evaluation).

The above runs were also replicated in a more realistic search scenario: instead of using for relevance feedback the top 10 retrieved relevant documents that can occur anywhere in the retrieved set, we used only those relevant documents which occur in the top 10 retrieved documents. This approach simulates a more realistic search scenario whereby a user looks at the top retrieved documents and judges their relevance. We also used a frozen-rank evaluation technique to evaluate the effectiveness of the above QE techniques in retrieving the unseen relevant documents. The technique consists in freezing the ranks of the top *n* (here 10) documents retrieved in the baseline run and evaluated for relevance feedback, and re-ranking the rest of the documents in the collection. The advantage of the rank freezing technique over the retrospective evaluation technique is that it evaluates how well a retrieval method ranks previously unseen relevant documents. The results of the frozen-rank evaluation (Table 2) follow the trend that was observed in the retrospective search evaluation: combination of the MI statistic with distance and frequency factors (MID_d5) performs better than MI alone, while both methods perform better than the original queries. Specifically, AveP of MID_d5 is 4.7% higher than that of MI, and the difference is statistically significant.

Table 2. Performance of the baseline and experimental query expansion runs (frozen rank evaluation)

Run	Distance factor formula	AveP	P@5	P@10	P@15	P@20	P@30	P@100	R-Prec
No_Expan		0.2426	0.4453	0.3887	0.3476	0.3157	0.2740	0.1584	0.2795
MI		0.2606	--	--	0.3698	0.3440	0.3042	0.1735	0.2983
MID_Exp	$e^{-\alpha*(D(x,y)-1)}$	0.2185	--	--	0.3484	0.3140	0.2604	0.1437	0.2617
MID_lgd	$\log_2(1+1/D(x,y))$	0.2369	--	--	0.3622	0.3300	0.2816	0.1557	0.2779
MID_lgd2	$\log_2(2+1/D(x,y))$	0.2597	--	--	0.3698	0.3457	0.3064	0.1726	0.2981
MID_lgd3	$\log_2(2+f_r(x,y)/D(x,y)$	0.2669	--	--	0.3747	0.3513	0.3096	0.1771	0.3024
MID_lgd4	$\log_2(3+f_r(x,y)/D(x,y)$	0.2682	--	--	0.3760	0.3510	0.3096	0.1785	0.3053
MID_d5	$f_r(x,y)/D(x,y)$	0.2729	--	--	0.3942	0.3643	0.3207	0.1826	0.3054
MID_d6	$f_r(x,y)$	0.2706	--	--	0.3867	0.3590	0.3189	0.1803	0.3051
MID_d7	$1/D(x,y)$	0.2358	--	--	0.3613	0.3297	0.2809	0.1553	0.2759
RSV		0.2888	--	--	0.3938	0.3707	0.3309	0.1947	0.3253

3. Results and Analysis

The purpose of the experiments conducted in this study was to systematically evaluate the effect of different distance factors on ranking collocates of the original query terms for query expansion following relevance feedback. The two hypotheses introduced in the beginning of the paper are discussed, and the results of the retrospective evaluation are analysed in this section.

Hypothesis 1

To determine whether the query expansion ranking method with the highest performance (MID_d5) performs better than No_Expan run, t-test was conducted. The results are presented in Table 3.

Table 3. T-test results of runs No_Expan and MID_d5.

	AveP	**P@5**	**P@10**	**P@15**	**P@20**	**P@30**	**P@100**
P value	<< .001	<< .001	<< .001	<< .001	<< .001	<< .001	<< .001
Significant?	Y	Y	Y	Y	Y	Y	Y

The above significance analysis supports Hypothesis 1. The differences between the average precision and precision at all cutoff points of runs in the frozen-rank evaluation are also significant. The use of query expansion terms selected using distance-weighted MI performs significantly better than the original query terms. MID_d5 outperforms original queries (run No_Expan) in 129 topics, improving the mean Average Precision by 129.2%. However, it deteriorates the Average Precision in 20 topics by 32.8%. One topic has the same average precision.

Hypothesis 2:

MID_d5 significantly outperforms the use of Mutual Information alone for ranking collocates in all but P@5 (precision at 5 documents) measures, as can be seen from Table 4. The differences between the average precision and precision at all cutoff points (except P@100) of runs in the frozen-rank evaluation are also significant.

Table 4. T-test results of runs MI and MID_d5.

	AveP	**P@5**	**P@10**	**P@15**	**P@20**	**P@30**	**P@100**
P value	<< .001	.526	.011	<< .001	<< .001	<< .001	<< .001
Significant?	Y	N	Y	Y	Y	Y	Y

MID_d5 has higher performance than MI in 104 topics, improving the mean Average Precision of these topics by 34%, and worse performance in 43 topics, degrading the mean AveP by 27.1%. Three topics have the same average precision. These results show that taking into consideration distance between collocates does help to select better query expansion terms, thereby improving the overall retrieval performance. Hypothesis 2 is, therefore, supported.

Table 5. Frequency of the top 20 query expansion terms selected by MID_d5 and MI in the known relevant document set, grouped by distance.

Distance	MID_d5			MI		
	Frequency	%	Cumulative %	Frequency	%	Cumulative %
1	6174	0.6	0.6	488	0.3	0.3
2	4881	0.5	1.1	555	0.3	0.6
3	4793	0.5	1.5	669	0.4	1.1
4	4534	0.4	2	659	0.4	1.5
5	4534	0.4	2.4	587	0.4	1.8
5-10	21954	2.1	4.6	3001	1.9	3.7
11-20	41400	4	8.6	5960	3.7	7.4
21-30	39089	3.8	12.4	5763	3.6	10.9
31-50	73502	7.2	19.6	71554	44.1	55.1
51-100	158438	15.5	35.1	10934	6.7	61.8
101-150	129466	12.6	47.7	24316	15	76.8
151-200	105171	10.3	58	20581	12.7	89.5
201+	429756	42	100	17010	10.5	100

To gain a better understanding of the effect the distance factor has on the type of selected query expansion terms, further analysis was conducted. We compared the number of instances of query expansion terms selected using MID_d5 and MI grouped into categories by their distance from the original query terms. Term frequencies were counted in the 10 known relevant documents per topic used for query expansion in the retrospective evaluation experiment described earlier in the paper. As can be seen from Table 5, 4.6% (MID_d5) and 3.7% (MI) term occurrences are within the distance of 10 words from the original query terms. Within the distance of 30 words there are 12.4% of occurrences of MID_d5-selected terms, and 10.9% of MI-selected terms. Interestingly, 55.1% of occurrences of the MI-selected terms are within the span of 50 words from the original query terms, however, only 19.6% of occurrences of MID_d5-selected terms are within this span. On average, however, terms selected by MID_d5 occur slightly closer to the original query terms (231 words) than terms selected by MI (238 words).

As can be seen from Table 5, terms selected by MID_d5 have substantially more occurrences in the relevant document set used for query expansion than terms selected by MI. This is due to the inclusion of the joint frequency $f_r(x, y)$ into the distance factor. Comparison of the number of postings (documents) in the entire collection containing the query expansion terms selected by MI and MID_d5 shows even more radical differences between the type of terms selected by the two measures (Table 6). Only 2.5% of terms selected by MID_d5 have 1 posting, compared to 30% of terms selected by MI. This means that 30% of MI-selected query expansion terms only occur in the documents which the user has already seen, and are not useful for retrieving other relevant documents.

Table 6. Mean, Median and Standard deviation of the numbers of postings containing query terms selected by MID_d5 and MI.

	MID_d5	MI
Mean	12684.25	43.884
Median	2366.5	4
Stdev	24661.14	233.2018

Table 7. Top 20 query expansion terms for the topic 303 "Hubble Telescope Achievements" selected using MID_d5 and MI (terms present in both lists are highlighted)

MID_d5	**MI**
universe	supertelescope
astronomer	photometric
VLT	Waelkens
cosmology	VLT
galaxy	Paranal
space	Christoffel
Earth	Silla
squirrel	ESO
observatory	Hofstadt
Saturn	Glaswerke
light	DM450m
surpass	interstellar
star	concave
mirror	Atacama
ESO	astrophysicist
helium	**astronomer**
Edwin	**galaxy**
distant	**observatory**
observe	innermost
quasar	Cerro

Table 7 shows an example of the top 20 query expansion terms selected by MID_d5 and MI measures for the TREC topic 303 "Hubble Telescope Achievements". As evident from the table, MI-selected terms contain more rare terms, such as proper names, than the terms selected by MID_d5.

4. Conclusions and Future Work

In this study, motivated by earlier studies on term proximity ranking [16, 17, 20], experiments on query expansion [1, 3, 4, 12, 28], and word association models [14, 22, 29], we systematically investigated the use of term proximity information for the ranking and selection of query term collocates as query expansion terms. Through continuous improvement of the distance factor in the collocate-weighting formula we developed a query

expansion term selection method, which shows significant performance gains over the use of original user queries. Combining the collocation distance, collocation frequency and Mutual Information helps select better query expansion terms than the use of MI alone. The difference between the two methods is statistically significant in all but one measure (Precision at 5 documents). The evaluation conducted in this study suggests that distance-weighted MI is an effective query expansion technique. The proposed distance-weighted MI function might also be useful for other applications, such as word sense disambiguation and lexicography, which rely on word association measures. However, more task-based evaluations need to be conducted.

Other possible avenues for improvement of the query expansion term selection that could be explored are:

Use of the Part-Of-Speech (POS) information: Jing and Croft [30] studied the effect of different POS on the selection of word co-occurrences, and found that noun phrases achieved the best performance. Xu and Croft [28] only used noun phrases for query expansion. Evert and Kermes [31] also suggested applying linguistic filters before extracting collocations. A method using noun phrases only in the query expansion may improve the retrieval performance.

Use of the number of distinct relevant documents in query expansion term selection. Collocates located in the proximity of query terms are promising expansion terms as shown in this paper. However, words co-occurring with query terms in several relevant documents are likely to be more useful than words co-occurring with query terms in only one relevant document. Therefore, taking into account the number of relevant documents that a word pair co-occurs in may improve the selection of query terms.

References

[1] M. Beaulieu and S. Jones, Interactive searching and interface issues in the Okapi best match probabilistic retrieval system. *Interacting With Computers* 10 (1998) 237-248.

[2] J. J. Rocchio, Relevance feedback in information retrieval. In: G. Salton (ed.), *The SMART Retrieval System – experiments in automatic document processing.* (Englewood Cliffs, New Jersey, Prentice-Hall, 1971) 312-323.

[3] S.E. Robertson, S. Walker and M. Beaulieu, Okapi at TREC-7: automatic, ad hoc, filtering, VLC and interactive track. In: E.M. Voorhees and D.K. Harman (eds.), *Proceedings of the 7th Text REtrieval Conference (TREC-7),* (Gaithersburg, MD, NIST, 1999) 253-264.

[4] K. Spärck Jones, S. Walker and S.E. Robertson, A probabilistic model of information retrieval: development and comparative experiments. Parts 1 and 2. *Information Processing and Management* 36 (2000) 779-808, 809-840.

[5] C.D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing* (The MIT Press, Cambridge, Massachusetts, 1999).

[6] T. Dunning, Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1) (1993) 61–74.

[7] F. Jelinek, Self-organised language modelling for speech recognition. In: A. Waibel and K. Lee (eds.) *Readings in Speech Recognition.* (San Mateo, California, Morgan Kaufmann Publishers, 1990).

[8] C. J. Van Rijsbergen, A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation* 33(2) (1977) 106-119.

[9] K. Church, W. Gale, P. Hanks and D. Hindle, Using statistics in lexical analysis. In: U. Zernik (ed.), *Lexical Acquisition: Using On-line Resources to Build a Lexicon*, (Englewood Cliffs, NJ, Lawrence Elbraum Associates, 1991) 115-164.

[10] D. Yarowsky, Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. In: *Proceedings of COLING-92,* (Nantes, 1992) 454-460.

[11] T. McEnery and A. Wilson, *Corpus Linguistics* (Edinburgh, 1996).

[12] O. Vechtomova, S.E. Robertson and S. Jones, Query expansion with long-span collocates. *Information Retrieval* 6(2) (2003) 251-273.

[13] K. Ishikawa, K. Satoh and A. Okumura, Query term expansion based on paragraphs of the relevant documents. In: E.M. Voorhees and D.K. Harman (eds.), *Proceedings of the 6th Text Retrieval Conference (TREC-6),* (Gaithersburg, MD, USA, 1997) 577-584.

[14] D. Beeferman, A. Berger and J. Lafferty, A Model of Lexical Attraction and Repulsion. In: *Proceedings of the 8th conference on European chapter of the Association for Computational Linguistics*, (Madrid, Spain, 1997) 373-380.

[15] M. Hoey, *Patterns of Lexis in Text* (Oxford University Press, 1991).

[16] C.L.A. Clarke, G.V. Cormack and E.A. Tudhope, Relevance Ranking for One to Three Term Queries. *Information Processing and Management* 36(2) (2000) 291-311.

[17] D. Hawking and P. Thistlewaite, Proximity operators - so near and yet so far. In: D.K. Harman (ed.), *Proceedings of the 4th Text Retrieval Conference (TREC-4),* (Gaithersburg, MD, USA, 1995) 131-143.

[18] D. Hawking and P. Thistlewaite, Relevance Weighting Using Distance between Term Occurrences, (Unpublished manuscript, Joint Computer Science Technical Report Series, TR-CS-96-08, The Australian National University, 1996).

[19] B. Pôssas, N. Ziviani and W. Meira, Enhancing the Set-Based Model Using Proximity Information. In: *Proceedings of the 9th International Symposium on String Processing and Information Retrieval (SPIRE),* (Lisbon, Portugal, 2002, Springer-Verlag, Lecture Notes in Computer Science #2476) 104-116.

[20] Y. Rasolofo and J. Savoy, Term Proximity Scoring for Keyword-based Retrieval Systems. In: *Proceedings of the 25th European Conference on Information Retrieval Research*, (Pisa, Italy, 2003, Springer-Verlag, Lecture Notes in Computer Science #2633) 207-218.

[21] A. M. Lam-Adesina and G. J. F. Jones Applying summarization techniques for term selection in relevance feedback. In: *Proceedings of the 24th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, (New Orleans, Louisiana, United States, 2001) 1-9.

[22] J. Gao, J. Nie, H. He, W. Chen, and M. Zhou, Resolving query translation ambiguity using a decaying co-occurrence model and syntactic dependence relations. In: *Proceedings of the 25th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, (Tampere, Finland, 2002) 183-190.

[23] E.M. Voorhees and D.K. Harman (Eds.) *Proceedings of the Eighth Text Retrieval Conference (TREC),* (NIST, Gaithersburg, MD, USA, 2000)

[24] R. Fano, *Transmission of information* (Cambridge, Mass., MIT Press, 1961).

[25] Y. Wang, The Use of Term Proximity in Collocate-Ranking for Query Expansion. (Unpublished manuscript, Masters Thesis, Department of Management Sciences, University of Waterloo, Canada, 2004).

[26] Y. Wang and O. Vechtomova, Exploring the Use of Term Proximity in Collocate-Ranking for Query Expansion. In: *Proceedings of the 17th Joint ACH/ALLC (Association for Computers and the Humanities/Association for Literary and Linguistic Computing) Conference,* (Victoria, BC, Canada, 2005) 257-259.

[27] S.E.Robertson, On term selection for query expansion, *Journal of Documentation*, 46 (1990) 359-364.

[28] J. Xu and W. B. Croft, Query Expansion Using Local and Global Document Analysis. In: *Proceedings of the 19th annual international ACM SIGIR conference on Research and Development in Information Retrieval,* (Zurich, Switzerland, 1996) 4-11.

[29] T.R. Niesler and P.C. Woodland, Modeling word-pair relations in a category-based language model. In: *Proceedings of the International Conference on Acoustics, Speech, and Signal*, (Munich, Germany, 1997) 795-798.

[30] Y. Jing and W.B. Croft, An Association Thesaurus for Information Retrieval. (Unpublished manuscript, Technical Report, UM-CS-1994-017, University of Massachusetts, Amherst, MA, USA, 1994).

[31] S. Evert and H. Kermes, The influence of linguistic pre-processing on candidate data. In: *Proceedings of the Workshop on Computational Approaches to Collocations*, (Vienna, Austria, 2002).