BM25 Beyond Query-Document Similarity

Billel Aklouche[1,2,4(✉)], Ibrahim Bounhas[1,4], and Yahya Slimani[1,3,4]

[1] LISI Laboratory of Computer Science for Industrial System, INSAT, Carthage University, Tunis, Tunisia
bounhas.ibrahim@gmail.com, yahya.slimani@gmail.com
[2] National School of Computer Science (ENSI), La Manouba University, Manouba, Tunisia
billel.aklouche@ensi-uma.tn
[3] Higher Institute of Multimedia Arts of Manouba (ISAMM), La Manouba University, Manouba, Tunisia
[4] JARIR: Joint group for Artificial Reasoning and Information Retrieval, Manouba, Tunisia
http://www.jarir.tn/

Abstract. The massive growth of information produced and shared online has made retrieving relevant documents a difficult task. Query Expansion (QE) based on term co-occurrence statistics has been widely applied in an attempt to improve retrieval effectiveness. However, selecting good expansion terms using co-occurrence graphs is challenging. In this paper, we present an adapted version of the BM25 model, which allows measuring the similarity between terms. First, a context window-based approach is applied over the entire corpus in order to construct the term co-occurrence graph. Afterward, using the proposed adapted version of BM25, candidate expansion terms are selected according to their similarity with the whole query. This measure stands out by its ability to evaluate the discriminative power of terms and select semantically related terms to the query. Experiments on two ad-hoc TREC collections (the standard Robust04 collection and the new TREC Washington Post collection) show that our proposal outperforms the baselines over three state-of-the-art IR models and leads to significant improvements in retrieval effectiveness.

Keywords: Query expansion · Co-occurrence graph · BM25 · Term discriminative power · Ad-hoc IR

1 Introduction

The main purpose of information retrieval (IR) systems is to provide a set of relevant documents according to a user's specified need. A number of ranking models have been proposed in the literature [4,22,32], all of which intend to retrieve the most relevant documents in response to a query. The difference in

N. R. Brisaboa and S. J. Puglisi (Eds.): SPIRE 2019, LNCS 11811, pp. 65–79, 2019.
https://doi.org/10.1007/978-3-030-32686-9_5

retrieval results from one model to another is manifested in the set of returned documents and in the order of their appearance. Among these models, Okapi BM25 [22] is a pre-eminent probabilistic model, which has proven its effectiveness as a state-of-the-art IR model and has been widely used, especially in TREC experiments. The BM25 model incorporates information about both terms and documents, which includes local terms frequencies, global terms frequencies and document length. Since it was introduced, several studies have been presented proposing extensions and improvements [5,13,15,21,23,26].

However, despite the improvements that can be made to ranking models, the user's query remains the key factor that controls the relevance of retrieval results. Indeed, it is often too short and insufficient to allow the selection of documents that meet the user needs. In most cases, the latter does not know exactly what he wants or how to express it. Therefore, the returned results are unlikely to be relevant. To overcome this problem, Query Expansion (QE) refers to techniques that reformulate the original query by adding new terms to those entered by the user to better express his need and improve retrieval performance.

A main challenge in QE is the selection of good expansion terms which do not hurt, but improve, retrieval performance. The strength of the BM25 model is that it allows capturing the behavior of terms not only in a document, but also in the entire collection. It assumes that a good document descriptor is a quite frequent term in this document, which is relatively infrequent in the entire document collection [14]. Based on these assumptions, we propose an approach to QE by adapting BM25 to work on term co-occurrence graphs. The main motivation is to model the discriminative power of terms using a measure analogous to the inverse document frequency (IDF) factor of TF-IDF [25]. We define a good expansion term as one that frequently co-occurs with the query terms and has a relatively rare co-occurrence with the rest of the vocabulary.

We evaluate our proposal using two ad-hoc TREC collections: the standard TREC Robust04 collection with 249 queries (TREC 2004 Robust Track) and the newest TREC Washington Post collection with 50 queries (TREC 2018 Common Core Track). Experimental results show that our proposal outperforms the baselines by significant margins in terms of MAP and precision.

The remainder of the paper is organized as follows. In the next section, we discuss some related work on QE. We describe the proposed adapted version of BM25 for QE in Sect. 3. The Experimental setup and the obtained results are presented in Sect. 4. Finally, Sect. 5 concludes the paper and provides insights for future work.

2 Related Work

For several years, great effort has been devoted to the development of new QE approaches [8]. Corpus-based QE approaches are among the most popular techniques that have been widely applied [29]. The corpus itself serves as a source for selecting expansion terms. Indeed, the broad range of corpus-based QE approaches can be divided into two main classes: local approaches and global approaches [28,29].

Local approaches use the top-ranked documents, retrieved in response to the initial query, in order to select expansion terms, mostly using pseudo-relevance feedback (PRF), where the top k ranked documents in the initial retrieval results are assumed to be relevant. For example, authors in [28] presented a PRF technique called LCA (Local Context Analysis) in which candidate expansion terms are selected on the basis of their co-occurrence relationship with query terms within pseudo-relevant documents. They showed the effectiveness of the proposed PRF technique using different languages. In [12], authors presented a concept-based PRF technique. They built a directed query relations graph to extract concepts that are related to the query. The query relations were mined using association rules. Authors in [27] discussed the contribution of linear methods for PRF. They used an inter-term similarity matrix to get expansion terms. In [31], authors presented a matrix factorization technique using pseudo-relevant documents. They considered PRF as a recommendation task for selecting useful expansion terms. They demonstrated the effectiveness of this technique on two retrieval models: the language model and the vector space model.

Unlike local QE, global approaches allow selecting expansion terms without regard to the initial retrieval results. In this case, expansion terms are selected by analyzing the entire corpus in order to discover term associations and co-occurrence relationships [29]. For example, authors in [33] proposed a technique to expand short queries for microblog retrieval. They explored the use of Wikipedia, DBpedia and association rules mining for selecting semantically related terms to the queries. In [6], authors addressed QE by using co-occurrence relationships and inferential relationships between terms. They proposed to integrate QE into language modeling and demonstrated the feasibility of this integration.

The use of term co-occurrence statistics is one of the earliest QE approaches, in which terms that are statistically related to the query are considered as potential expansion candidates. However, a basic issue in this approach is the selection of discriminative terms using co-occurrence statistics. Usually, the selected terms tend to occur frequently in the entire collection and thus are unlikely to be discriminative. This limitation is mainly due to the way in which the similarity between terms is measured [19].

Several measures have been used to evaluate the similarity between pairs of terms. We may cite Cosine similarity, Jaccard index, Dice coefficient and Mutual Information [8]. Recently, QE based on word embedding [1,3,30] leads to an interesting improvement on retrieval effectiveness by exploring word relationships from embedding vectors. In these methods, term co-occurrence statistics are employed to learn word vector representations using word embedding algorithms such as word2vec [18] and Glove [20]. Indeed, terms co-occurrence within the same context window is used to produce word vectors [30]. We use the same approach, i.e. a context window-based approach applied over the entire corpus, in order to build our term co-occurrence graphs.

3 An Adaptation of BM25 for Query Expansion Based on Term Co-occurrence Graphs

In this section, we describe our QE approach and we present the proposed adaptation of BM25 for term co-occurrence graphs. Figure 1 depicts the general architecture of our QE system. We select semantically related terms to the query following two steps. First, a term co-occurrence graph is constructed over the entire corpus using a context window-based approach. This approach has been used in multiple IR and Natural Language Processing (NLP) tasks such as word embedding [18,20]. Indeed, the co-occurrence of terms within a specified context window is used to capture semantic relations between terms. For instance, given the sentence "The SPIRE conference covers research on string processing and information retrieval." and taking "conference" as the target term with a window-size equal to 2, its context terms will be "The", "SPIRE", "covers" and "research". Second, using an adapted version of the BM25 model to measure the similarity of terms in co-occurrence graphs, candidate expansion terms are scored according to their similarity with the query as a whole.

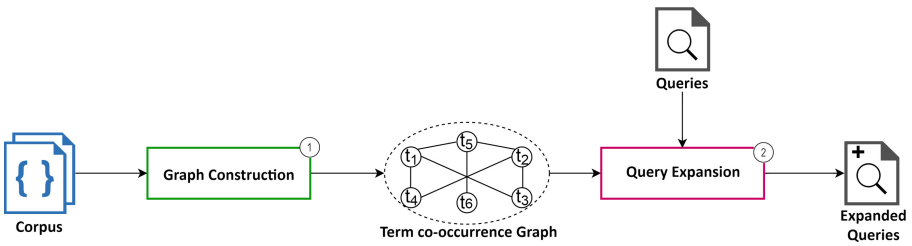

Fig. 1. General architecture of the proposed QE approach.

The Okapi BM25 model calculates the score of a document D given a query Q as follows [22]:

$$BM25(Q, D) = \sum_{t \in Q} IDF(t) \times \frac{(k_1 + 1) \times tf(t, D)}{k_1 \times (1 - b + b \times \frac{dl}{avgdl}) + tf(t, D)} \quad (1)$$

where:

- $IDF(t)$ is the Inverse Term Frequency of t and it is computed as follows:

$$IDF(t) = log \frac{N - df(t) + 0.5}{df(t) + 0.5} \quad (2)$$

- N is the number of documents in the collection.
- $df(t)$ is the number of documents containing term t.

- $tf(t, D)$ is the term frequency of t, i.e., the number of occurrences of term t in the document D.
- dl and $avgdl$ denote the length of document D and the average document length in the collection, respectively.
- k_1 and b are free hyper-parameters.

We propose to contribute in developing one-to-many association measures which are computed on a symmetric co-occurrence graph. This measure is inspired from Okapi BM25 [22]. A symmetric co-occurrence graph is an undirected graph $G = (V, E)$, where V is a set of nodes and E is a set of weighted edges. We also define the symbols cited hereafter as follows:

- n_i: the node number i in the graph.
- $e(n_i, n_c) = e(n_c, n_i)$: the weight of the edge linking n_i and n_c.
- $co\_degree(n_i)$: the number of nodes in G having n_i as destination.
- $sum\_e(n_c)$: the sum of the weights of the edges having n_c as destination.
- $avgsum\_e$: the average of the previous parameter (i.e. $sum\_e(n_c)$) over all the possible destination nodes in G.
- N is the number of all possible destination nodes in G.

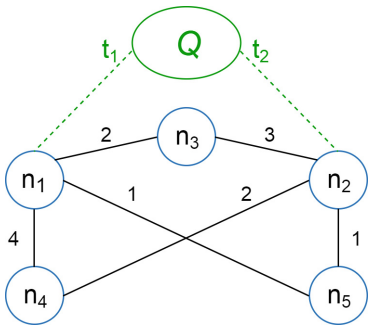

Fig. 2. Example of a query projected on the graph. n_1, n_2 are the query terms and n_3, n_4, n_5 are candidate expansion terms.

We formalize similarity calculus in graphs as follows. Let $C = \{n_1, \ldots, n_m\}$ be the projection of query $Q = \{t_1, \ldots, t_m\}$ on the co-occurrence graph G and n_c be a candidate node in G. An example of a query Q projected on a co-occurrence subgraph is illustrated in Fig. 2. We propose to compute the relevance of a node n_c given C in a co-occurrence graph. C is considered as a query, n_c as a document and the relevance is assessed using an adapted version of BM25. That is, we follow other research which used IR models to compute similarities between queries and terms [9,10]. We define the relevance of a node n_c given another node n_i as a product of a local weight $L_{i,c}$ and a global weight G_i:

$$Sim(n_i, n_c) = L_{i,c} \times G_i \tag{3}$$

Local Weights. The main hypothesis behind local weights is that terms which co-occur frequently are likely to be similar. Local weights apply to the following constraints:

- $L_{i,c} = 0$ if $e(n_i, n_c) = 0$.
- $L_{i,c}$ increases with $e(n_i, n_c)$.
- $L_{i,c}$ approaches a maximum value of 1.

In classical document indexing with TF-IDF [25], local weights are normalized by document length, which is equivalent in our case to the sum of the weights of the edges linking to n_c (cf. Formula 4). A term which appears n times in a short document is more significant compared to the case in which it appears the same number of times in a longer one. In our case, n_i is more relevant to n_c when the latter has a lower degree.

$$L_{i,c} = \frac{e(n_i, n_c)}{sum\_e(n_c)} \tag{4}$$

In BM25, local weights are computed as parameterized frequencies based on a 2-Poison model:

$$L_{i,c} = \frac{e(n_i, n_c)}{k_1 + e(n_i, n_c)} \tag{5}$$

$L_{i,c}$ is nonlinear to k_1. This is justified by Robertson et al. [24] by the fact that *"the information gained on observing a term the first time is greater than the information gained on subsequently seeing the same term"*. Robertson and Walker [22] considered two hypotheses, namely Verbosity and Scope. The first hypothesis allows to handle synonyms. Let consider two synonyms represented by the nodes n_c^1 and n_c^2 (i.e. both terms are similar to an input node n_i). Let also suppose that n_c^1 has a greater co-occurrence value with n_i (i.e. $e(n_i, n_c^1) > e(n_i, n_c^2)$). We say that one of the two nodes is more verbose (i.e. more likely to be used) than the other. According to this hypothesis, we should normalize the co-occurrence values to obtain close values of similarity for both nodes. However, applying to the scope hypothesis, we should not normalize as we would prefer to return n_c^1. We suppose in this case that the more frequent term is more likely to represent the sense which is shared by both nodes. However, n_c^2 is less frequent and thus unable to add much information to the original query. Both hypotheses are complementary. In real co-occurrence graphs, both scenarios are present, each of them constitutes a partial explanation. To insure that both hypotheses are respected, document lengths are normalized. In our case, we compute the average of the weighted degrees of nodes as follows:

$$avgsum\_e = \frac{\sum_i sum\_e(n_i)}{N} \tag{6}$$

Then weighted degrees are normalized as follows:

$$norm\_sum\_e(n_i) = \frac{sum\_e(n_i)}{avgsum\_e} \tag{7}$$

To allow adjust and tune this score, it is reformulated as follows:

$$norm\_sum\_e(n_i) = 1 - b + b \times \frac{sum\_e(n_i)}{avgsum\_e} \tag{8}$$

The constant b determines the scaling by the degree of the target node (document length in query-document matching). $b = 1$ means fully scaling the term weight, while $b = 0$ disables normalization. The quantity obtained by Eq. 8 is used to normalize the local weight computed by Eq. 5. Thus we have:

$$L_{i,c} = \frac{e(n_i, n_c) \times (k_1 + 1)}{k_1 \times (1 - b + b \times \frac{sum\_e(n_c)}{avgsum\_e}) + e(n_i, n_c)} \tag{9}$$

Global Weights. Global weights are defined according to the Probabilistic Model of Robertson and Spärck-Jones [16]. Given a node n_i, we would like to know if a node n_c is relevant (i.e. similar to n_i) based on the probabilistic IR framework [7]. The contingency table (Table 1) defines the main parameters used to estimate the probability of relevance of n_c. This table is defined in a scenario of relevance feedback where a user selects the terms which are relevant to a given query.

Table 1. Contingency table of main parameters

	n_c is relevant	n_c is not relevant	Total
n_c co-occurs with n_i	s	$co\_degree(n_i)$-s	$co\_degree(n_i)$
n_c does not co-occur with n_i	S-s	(N- $co\_degree(n_i)$)-(S-s)	N-$co\_degree(n_i)$
Total	S	N-S	N

In this table, s is the number of terms which are relevant to the query which co-occur with n_i . The relevance of n_c may be estimated as follows:

$$INF(n_i) = log \frac{\frac{s}{S-s}}{\frac{co\_degree(n_i)-s}{(N-co\_degree(n_i))-(S-s)}} \tag{10}$$

In this model, it is fairly standard to add 0.5 to the quantities which may be null (i.e. the cells of the second and the third column of the contingency table) [16]. A second variant is thus defined as follows:

$$INF(n_i) = log \frac{\frac{s+0.5}{S-s+0.5}}{\frac{co\_degree(n_i)-s+0.5}{(N-co\_degree(n_i))-(S-s+0.5)}} \tag{11}$$

Using 0.5 is a kind of smoothing which is justified by the limits of maximum likelihood estimate (or MLE) which penalizes rare events. Smoothing allows handling events which has never been seen nor observed [16]. In the absence of

relevance feedback, we have s=S=0. If we adopt Formula 10 (without smoothing), we obtain:

$$INF(n_i) = log\frac{N - co\_degree(n_i)}{co\_degree(n_i)} \tag{12}$$

With smoothing (Formula 11), we get:

$$INF(n_i) = log\frac{N - co\_degree(n_i) + 0.5}{co\_degree(n_i) + 0.5} \tag{13}$$

In all the cases, $INF(n_i)$ reflects how much a term is distributed over the others. It just checks if it is common or rare across all the other terms. That is, terms which tend to co-occur with many terms (e.g. stop words) will get null or low values. However, it provides an absolute evaluation of the discriminative power of a term which does not depend on the original query.

In Eq. 3, we replace global and local weights computed respectively by Formulas 13 and 9. Besides, we compute the sum of the similarity of n_c and all the terms of C. Thus, using an adapted version of BM25, noted here $BM25_{cog}$ (BM25 for co-occurrence graphs), we calculate the score of each candidate node n_c as follows:

$$BM25_{cog}(C, n_c) = \sum_{n_i \in C} INF(n_i) \times \frac{(k_1 + 1) \times e(n_i, n_c)}{k_1 \times (1 - b + b \times \frac{sum\_e(n_c)}{avgsum\_e}) + e(n_i, n_c)} \tag{14}$$

The constant k_1 determines how relevance changes when the number of co-occurrence $e(n_i, n_c)$ increases. A null value of k_1 means disabling term weight (using only $INF(n_i)$). If k_1 is large, the term weight component would increase nearly linearly with $e(n_i, n_c)$. Using the default value of this parameter means that after three or four co-occurrences, additional co-occurrences will have a little impact [24].

This adapted version of the BM25 model stands out by the following aspects. First, it allows both one-to-one and one-to-many associations. On the other hand, the INF factor allows to evaluate the discriminative power of terms. That is, terms that co-occur with many other terms are penalized. Moreover, it has two hyper-parameters, which may be tuned to enhance results. We used $BM25_{cog}$ in a PRF scenario in [2]. The obtained results showed significant improvements over the state-of-the-art baselines.

4 Experiments

In this section, we first present our test collections and describe the experimental setup. Then we discuss the experimental results.

4.1 Experimental Setup

We used two TREC collections in our experiments. The first is the standard Robust04 collection which is available in TREC disks 4 and 5[1]. It consists of

[1] https://trec.nist.gov/data/cd45/.

news articles from different sources. This collection was used in TREC 2004 Robust Track. The second is the newest TREC Washington Post collection[2] provided by TREC 2018 Common Core Track, which consists of news articles and blog posts published by Washington Post from January 2012 through August 2017. Statistics of these collections are presented in Table 2.

Table 2. TREC collections statistics.

Collection	Document set	#docs	Size	#query	#qrels
Robust04	TREC Disks 4 & 5 minus Congressional Record	528k	1.9 GB	249	17,412
WAPOST	TREC Washington Post Corpus	608k	6.9 GB	50	3,948

All experiments were conducted using the Terrier 4.2 IR platform[3]. For both collections, preprocessing involved stopword removal using the Terrier's standard stopword list and stemming using the Porter stemmer. We only considered the title of the TREC topics as queries (i.e., short queries).

We use Mean Average Precision (MAP), precision at top 5 documents (P@5) and precision at top 10 documents (P@10) as evaluation measures. MAP serves as the objective evaluation measure for parameter tuning. Statistically significant differences in terms of retrieval performance are computed using the two-tailed paired t-test at a 95% confidence level.

4.2 Parameter Tuning

In order to construct the term co-occurrence graph, we need to choose the value of the window size parameter. We explored different values of this parameter to see the effect they have on effectiveness. Window size values of 2–10, plus a window size equal to sentence length, were tested. We found that a window size of 7 terms gives the best results on the Robust04 collection, whereas best results on the Washington Post collection are obtained using a dynamic window size equal to sentence length. This is consistent with previous research [11,17], stating that the best choice of context window size is collection-dependent. We should note that we used a symmetric window size, i.e., a window size of n means n terms to the left and n terms to the right of the target term. The optimal parameter value for each collection was used to construct the term co-occurrence graph.

The model hyper-parameters were tuned using 5-fold cross-validation over the queries of each collection, where topics were randomly split into 5 folds. The hyper-parameters were tuned on 4-of-5 folds and tested on the final fold. This process is carried out 5 times, each time using one fold. The results presented are the mean of the 5 runs. We varied the value of b from 0.1 to 0.9 and the value of k_1 from 0.1 to 3.0 in increments of 0.1. The number of expansion terms was empirically set to 10.

[2] https://trec.nist.gov/data/wapost/.
[3] http://terrier.org/.

4.3 Results

In this subsection, we evaluate the effectiveness of the proposed approach. We consider three state-of-the-art IR models as baselines, namely: Okapi BM25 model [22], Language Model with Jelinek-Mercer smoothing [32] and Divergence from Randomness (DFR) PL2 model [4]. Besides, we consider the classical PRF approach and embedding-based QE approach by using word2vec[4] (W2V) to train word vectors over the target corpus. The obtained results are reported in Tables 3 and 4. Superscripts 1/2/3 indicate that the improvements over the unexpanded baselines, PRF and W2V, respectively, are statistically significant (t-test with $p\_value < 0.05$).

Table 3. Retrieval results on the Robust04 collection.

Retrieval model	Method	MAP	P@5	P@10
BM25	Baseline	0.2363	0.4691	0.4100
	PRF	0.2537	0.4378	0.3835
	W2V	0.2396	0.4627	0.4133
	$BM25_{cog}$	$\mathbf{0.2589^{1,3}}$	$\mathbf{0.4723^{2}}$	$\mathbf{0.4289^{1,2,3}}$
LM	Baseline	0.2155	0.3952	0.3651
	PRF	0.2437	0.4008	0.3747
	W2V	0.2259	0.4201	0.3743
	$BM25_{cog}$	$\mathbf{0.2469^{1,3}}$	$\mathbf{0.4394^{1,2,3}}$	$\mathbf{0.3940^{1}}$
PL2	Baseline	0.2239	0.4578	0.4032
	PRF	0.2287	0.4185	0.3763
	W2V	0.2303	**0.4683**	0.4092
	$BM25_{cog}$	$\mathbf{0.2464^{1,2,3}}$	0.4667^{2}	$\mathbf{0.4221^{1,2,3}}$

According to these tables, the proposed QE approach outperforms the state-of-the-art baselines in terms of MAP, P@5 and P@10 in all cases. The MAP improvements are always statistically significant in both collections. As for precision, we can see that improvements are also statistically significant in most cases. This shows that our QE approach, which generates semantically related terms to the query as a whole, leads to improvement in retrieval performance of the state-of-the-art models.

Comparing our QE and the classical PRF approach, we observe in Table 4 that the latter outperforms our approach in terms of MAP in the Washington Post collection. This shows in this case the advantage of local analysis over global analysis. Whereas, we see in Table 3 that our approach outperforms PRF in terms of MAP in the Robust04 collection. In terms of precision, we can remark that our approach outperforms PRF by significant margins in both collections. Besides,

[4] We used the CBOW implementation of word2vec and we set the vectors dimension to 300.

Table 4. Retrieval results on the Washington post collection.

Retrieval model	Method	MAP	P@5	P@10
BM25	Baseline	0.2385	0.4920	0.4300
	PRF	**0.2865**	0.4640	0.4160
	W2V	0.2436	0.4400	0.4080
	$BM25_{cog}$	$0.2687^{1,3}$	$\mathbf{0.5120^{3}}$	$\mathbf{0.4400^{3}}$
LM	Baseline	0.2065	0.3760	0.3560
	PRF	**0.2612**	0.4000	0.3780
	W2V	0.2119	0.3600	0.3560
	$BM25_{cog}$	$0.2505^{1,3}$	$\mathbf{0.4720^{1,3}}$	$\mathbf{0.4080^{1,3}}$
PL2	Baseline	0.2274	0.4880	0.4100
	PRF	**0.2754**	0.4400	0.4080
	W2V	0.2330	0.4800	0.4080
	$BM25_{cog}$	$0.2599^{1,3}$	$\mathbf{0.5120^{2}}$	$\mathbf{0.4640^{1,2,3}}$

by comparing PRF and the unexpanded baselines, it can be observed that PRF hurts the precision in the majority of cases. This shows that our proposal is able to generate better expansion terms and can filter out non-discriminative ones, which co-occur with too many terms, thus improving the precision at top-ranked documents.

By comparing our results to those obtained using word embedding-based QE, we can remark that our proposal yields better results on both collections with significant margins in the majority of cases. This confirms the effectiveness of the proposed approach for QE.

In this set of experiments, retrieval models were used with their suggested default parameters. These default settings are unlikely to be optimal for different collections and query lengths. Therefore, we next investigate the impact of parameters tuning on retrieval performance in both collections. To this end, the three models were extensively tuned using 5-fold cross-validation over the queries of each collection. Optimal parameter settings are listed in Table 5. We tuned parameters b and λ for the BM25 model and the LM model, respectively, from 0.10 to 0.90 in increments of 0.01. For the PL2 model, parameter c was tuned from 1.0 to 20.0 in increments of 0.1. We tuned the k_1 parameter of the BM25 model but it had little impact on retrieval effectiveness. We therefore used the default value in Terrier (k_1=1.2). Table 6 presents the MAP results achieved by the proposed QE approach and the baselines for each of the three IR models. Superscript 1 indicates that the improvements over the baselines are statistically significant (t-test with $p\_value < 0.05$). We can see that, in both collections, our QE outperforms the unexpanded baselines with statistically significant improvements in all cases. These results confirm the effectiveness of our proposal regardless of the ranking model. Furthermore, it is worth noting that our best result on the Washington Post collection is equal to the result of the best official automatic run in TREC 2018 Common Core Track.

Table 5. Optimal parameter settings.

Query	Original			Expanded		
Retrieval model	BM25	LM	PL2	BM25	LM	PL2
Parameter	b	λ	c	b	λ	c
Robust04	0.322	0.616	8.480	0.404	0.684	7.500
WAPOST	0.418	0.470	5.380	0.545	0.478	4.460

Table 6. Comparison of MAP results between the expanded queries and the baselines with optimal parameter settings for the three retrieval models.

Collection	Retrieval model	Baseline	$BM25_{cog}$
Robust04	BM25	0.2498	0.2643^{1} (+5.80%)
	LM	0.2285	0.2593^{1} (+13.48%)
	PL2	0.2529	$\mathbf{0.2697^{1}}$ (+6.64%)
WAPOST	BM25	0.2506	0.2725^{1} (+8.74%)
	LM	0.2180	0.2642^{1} (+21.19%)
	PL2	0.2481	$\mathbf{0.2761^{1}}$ (+11.29%)

5 Conclusion

In this paper, we proposed an adaptation of the state-of-the-art probabilistic model BM25 to measure the similarity between terms in a co-occurrence graph for QE. The proposed measure allows to evaluate the discriminative power of terms and to obtain semantically related terms to the whole query. Besides, it takes advantage of the BM25's hyper-parameters that can be adjusted to improve retrieval results.

Experiments on the TREC Robust04 and Washington Post collections show significant improvements over the baselines in terms of MAP and precision for three state-of-the-art IR models.

As part of our future work, we plan to investigate the use of external resources (e.g. Wikipedia) to build the term co-occurrence graph. In addition, investigating the use of asymmetric context windows to construct the co-occurrence graph is also an interesting research direction. Another direction for extending this work is to study the use of the new similarity measure for other IR tasks, such as Query Reweighting and Word Sense Disambiguation (WSD).

References

1. Aklouche, B., Bounhas, I., Slimani, Y.: Query expansion based on NLP and word embeddings. In: Proceedings of the The Twenty-Seventh Text Retrieval Conference (TREC 2018), Gaithersburg, Maryland, USA (14–16 November 2018)

2. Aklouche, B., Bounhas, I., Slimani, Y.: Pseudo-relevance feedback based on locally-built co-occurrence graphs. In: Welzer, T., Eder, J., Podgorelec, V., Kamisalic Latific, A. (eds.) Advances in Databases and Information Systems, vol. 11695, pp. 105–119. (2019). https://doi.org/10.1007/978-3-030-28730-6_7
3. ALMasri, M., Berrut, C., Chevallet, J.-P.: A comparison of deep learning based query expansion with pseudo-relevance feedback and mutual information. In: Ferro, N., Crestani, F., Moens, M.-F., Mothe, J., Silvestri, F., Di Nunzio, G.M., Hauff, C., Silvello, G. (eds.) ECIR 2016. LNCS, vol. 9626, pp. 709–715. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-30671-1_57
4. Amati, G.: Probability models for information retrieval based on divergence from randomness. Ph.D. thesis, University of Glasgow, UK (2003)
5. Ariannezhad, M., Montazeralghaem, A., Zamani, H., Shakery, A.: Improving retrieval performance for verbose queries via axiomatic analysis of term discrimination heuristic. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, pp. 1201–1204. ACM, 7–11 August 2017
6. Bai, J., Song, D., Bruza, P., Nie, J.Y., Cao, G.: Query expansion using term relationships in language models for information retrieval. In: Proceedings of the 14th ACM International Conference on Information and Knowledge Management, Bremen, Germany, pp. 688–695. ACM, 31 October–5 November 2005
7. Bounhas, I., Elayeb, B., Evrard, F., Slimani, Y.: ArabOnto: experimenting a new distributional approach for building arabic ontological resources. Int. J. Metadata, Semant. Ontol. **6**(2), 81–95 (2011). https://doi.org/10.1504/IJMSO.2011.046578
8. Carpineto, C., Romano, G.: A survey of automatic query expansion in information retrieval. ACM Comput. Surv. (CSUR) **44**(1), 11–150 (2012). https://doi.org/10.1145/2071389.2071390
9. Elayeb, B., Bounhas, I., Khiroun, O.B., Evrard, F., Saoud, N.B.B.: A comparative study between possibilistic and probabilistic approaches for monolingual word sense disambiguation. Knowl. Inf. Syst. **44**(1), 91–126 (2015). https://doi.org/10.1007/s10115-014-0753-z
10. Elayeb, B., Bounhas, I., Khiroun, O.B., Saoud, N.B.B.: Combining semantic query disambiguation and expansion to improve intelligent information retrieval. In: Duval, B., van den Herik, J., Loiseau, S., Filipe, J. (eds.) ICAART 2014. LNCS (LNAI), vol. 8946, pp. 280–295. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-25210-0_17
11. Fagan, J.: Automatic phrase indexing for document retrieval. In: Proceedings of the 10th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New Orleans, Louisiana, USA, pp. 91–101. ACM (3–5 June 1987)
12. Fonseca, B.M., Golgher, P., Pôssas, B., Ribeiro-Neto, B., Ziviani, N.: Concept-based interactive query expansion. In: Proceedings of the 14th ACM International Conference on Information and Knowledge Management, Bremen, Germany, pp. 696–703. ACM (31 October – 05 November 2005)
13. He, B., Huang, J.X., Zhou, X.: Modeling term proximity for probabilistic information retrieval models. Inf. Sci. **181**(14), 3017–3031 (2011). https://doi.org/10.1016/j.ins.2011.03.007
14. Jones, K.S., Walker, S., Robertson, S.E.: A probabilistic model of information retrieval: development and comparative experiments: Part 2. Inf. Process. Manag. **36**(6), 809840 (2000). https://doi.org/10.1016/S0306-4573(00)00016-9

15. Lv, Y., Zhai, C.: Lower-bounding term frequency normalization. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, Glasgow, Scotland, UK, pp. 7–16. ACM, 24–28 October 2011
16. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008)
17. Metzler, D., Croft, W.B.: A markov random field model for term dependencies. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil, pp. 472–479. ACM (15–19 August 2005)
18. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, Proceedings of the 26th International Conference on Neural Information Processing Systems, Lake Tahoe, Nevada, United States, pp. 3111–3119. 5–8 December 2013
19. Peat, H.J., Willett, P.: The limitations of term co-occurrence data for query expansion in document retrieval systems. J. Am. Soc. Inf. Sci. **42**(5), 378–383 (1991)
20. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, pp. 1532–1543. ACL 25–29 October 2014
21. Rasolofo, Y., Savoy, J.: Term proximity scoring for keyword-based retrieval systems. In: Sebastiani, F. (ed.) ECIR 2003. LNCS, vol. 2633, pp. 207–218. Springer, Heidelberg (2003). https://doi.org/10.1007/3-540-36618-0_15
22. Robertson, S.E., Walker, S.: Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In: Croft, B.W., van Rijsbergen, C.J. (eds.) SIGIR 1994, pp. 232–241. Springer, London (1994)
23. Robertson, S.E., Zaragoza, H.: The probabilistic relevance framework: Bm25 and beyond. Found. Trends Inf. Retrieval **3**(4), 333–389 (2009). https://doi.org/10.1561/1500000019
24. Robertson, S., Zaragoza, H., Taylor, M.: Simple bm25 extension to multiple weighted fields. In: Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management, Washington, D.C., USA, pp. 42–49. ACM, 08–13 November 2004
25. Salton, G., McGill, M.: Introduction to Modern Information Retrieval. McGraw-Hill Book Company, USA (1984)
26. Song, R., Taylor, M.J., Wen, J.-R., Hon, H.-W., Yu, Y.: Viewing term proximity from a different perspective. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 346–357. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-78646-7_32
27. Valcarce, D., Parapar, J., Barreiro, A.: Lime: Linear methods for pseudo-relevance feedback. In: Proceedings of the 33rd Annual ACM Symposium on Applied Computing, Pau, France, pp. 678–687. ACM, 09–13 April 2018
28. Xu, J., Croft, W.B.: Improving the effectiveness of information retrieval with local context analysis. ACM Trans. Inf. Syst. (TOIS) **18**(1), 79–112 (2000). https://doi.org/10.1145/333135.333138
29. Xu, J., Croft, W.B.: 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, pp. 4–11. ACM, 18–22 August 1996

30. Zamani, H., Croft, W.B.: Relevance-based word embedding. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, pp. 505–514. ACM, 7–11 August 2017
31. Zamani, H., Dadashkarimi, J., Shakery, A., Croft, W.B.: Pseudo-relevance feedback based on matrix factorization. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, Indianapolis, Indiana, USA, pp. 1483–1492. ACM, 24–28 October 2016
32. Zhai, C., Lafferty, J.: A study of smoothing methods for language models applied to ad hoc information retrieval. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New Orleans, Louisiana, USA, pp. 334–342. ACM, 9–13 September 2001
33. Zingla, M.A., Chiraz, L., Slimani, Y.: Short query expansion for microblog retrieval. In: Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 20th International Conference KES-2016, York, UK, pp. 225–234. Elsevier, 5–7 September 2016