A Novel Term-Term Similarity Score Based Information Foraging Assessment

Ilyes Khennak(✉), Habiba Drias, and Hadia Mosteghanemi

Laboratory for Research in Artificial Intelligence,
USTHB, Algiers, Algeria
{ikhennak,hdrias,hmosteghanemi}@usthb.dz

Abstract. The dramatic proliferation of information on the web and the tremendous growth in the number of files published and uploaded online each day have led to the appearance of new words in the Internet. Due to the difficulty of reaching the meanings of these new terms, which play a central role in retrieving the desired information, it becomes necessary to give more importance to the sites and topics where these new words appear, or rather, to give value to the words that occur frequently with them. For this aim, in this paper, we propose a novel term-term similarity score based on the co-occurrence and closeness of words for retrieval performance improvement. A novel efficiency/effectiveness measure based on the principle of optimal information forager is also proposed in order to assess the quality of the obtained results. Our experiments were performed using the OHSUMED test collection and show significant effectiveness enhancement over the state-of-the-art.

Keywords: Information retrieval · Information foraging theory · Query expansion · Term proximity · Term co-occurrence

1 Introduction

The experimentation and evaluation phase is the main phase to judge the success and failure of the implemented systems and achieved programs. The performance and quality of a retrieval system is measured on the basis of effectiveness and efficiency [2]. From the theoretical standpoint, the effectiveness is indicated by returning only what user needs and efficiency is indicated by returning the results to the user as quickly as possible [3,9]. From the practical standpoint the effectiveness, or relevance, is determined by measuring the precision, recall, etc., and efficiency is determined by measuring the search time [7]. The Reliance on these two measures varies from one community to another. The information retrieval community, for example, is focusing too much on the quality of the top ranked results while the artificial intelligence community, which started paying attention to information retrieval, ontologies and the Semantic Web [16], is focusing on the retrieval process cost. Accordingly, and in order to establish consensus between communities and adopt both effectiveness and efficiency measures, the Information Foraging Theory has been proposed to do so. The information foraging is a

R. Giaffreda et al. (Eds.): IoT360 2014, Part I, LNICST 150, pp. 29–41, 2015.
DOI: 10.1007/978-3-319-19656-5_5

theory that describes information retrieval behavior [13,14]. It is derived from a food foraging theory called optimal foraging theory that helps biologists understand the factors determining an animal's food preference and feeding strategies. The basis of foraging theory is a cost and benefit assessment of achieving a goal.

We introduce in the early part of this study the basic principles and concepts of information foraging theory which is employed later in the experimental section to evaluate the retrieval systems. The second part is devoted to presenting and discussing our suggested approach for improving retrieval performance. The major purpose of this proposed approach is to provide an effective similarity measurement for query expansion based on the co-occurrence and closeness of terms. This approach assigns importance, during the retrieval process, to words that frequently occur in the same context. For instance, the term *'COIOTE'* is often recurred in the same sites where the words *'Conference'*, *'Italy'*, and *'2014'* are clustered. the reliance on this principle was not a coincidence but was the result of studies carried out recently regarding the evolution and growth of the Web. All of these studies have shown an exponential growth of the Web and rapid increase in the number of new pages created. In his study [15], Ranganathan estimated that the amount of online data indexed by Google had increased from 5 exabytes in 2002 to 280 exabytes in 2009. According to [22], this amount is expected to be double in size every 18 months. Ntoulas et al. [12] read these statistics in terms of the number of new pages created and demonstrated that their number is increasing by 8 % a week. The work of Bharat and Broder [1] went further and estimated that the World Wide Web pages are growing at the rate of 7.5 pages every second. This revolution, which the Web is witnessing, has led to the appearance of two points:

- The first point is the entry of new words into the Web which is estimated, according to [21], at about one new word in every two hundred words. Studies by [8,20] have shown that this invasion is mainly due to: neologisms, first occurrences of rare personal names and place names, abbreviations, acronyms, emoticons, URLs and typographical errors.
- The second point is that the users employ these new words during the search. Chen et al. [5] indicated in their study that more than 17 % of query words are out of vocabulary (Non dictionary words), 45 % of them are E-speak (lol), 18 % are companies and products (Google), 16 % are proper names, 15 % are misspellings and foreign words (womens) [19].

Out of these two points which the web is witnessing and due to the difficulty, or better, the impossibility to use the meanings of these words, we proposed a method based on finding the locations and topics where these words appear, and then trying to use the terms which neighbor and occur with the latter in the search process. We will use the best-known instantiation of the Probabilistic Relevance Framework system: Okapi BM25, and the Blind Relevance Feedback: Robertson/Sparck Jones' term-ranking function as the baseline for comparison, and evaluate our approach using OHSUMED test collection. The main contributions of our work in this paper are the following:

– The adoption of an external correlation measure in order to evaluate the co-occurrence of words with respect to the query features.
– The determination of an internal correlation measure in order to assess the proximity and closeness of words relative to the terms of the query.

In the next section, we will introduce the information foraging theory. The BM25 model and the Blind Feedback approach are presented in Sect. 3. In Sect. 4 we will explain our proposed approach and finally we will describe our experiments and results.

2 Information Foraging Theory

The information foraging is a theory proposed by [14]. It is becoming a popular theory for characterizing and understanding web browsing behavior [6]. The theory is based on the behavior of an animal deciding what to eat, where it can be found, the best way to obtain it and how much energy the meal will provide.

For example, imagine a predator that faces the recurrent problem of deciding what to eat. Energy flows into the environment and comes to be stored in different forms. Different types of habitat and prey will yield different amounts of net energy. By analogy, imagine an academic researcher that faces the recurrent problems of finding relevant information. Information flows into the environment to be represented in different types of external media. The different information sources will have different profitabilities in terms of the amount of valuable information.

The basis of foraging theory is a *cost* and *benefit* assessment of achieving a goal where cost is the amount of resources consumed when performing a chosen activity and the benefit is what is gained from engaging in that activity.

Conceptually, the *optimal forager* finds the best solution to the problem of maximizing the rate of benefit returned by effort expended given the energetic profitabilities of different habitats and prey, and the costs of finding and pursuing them. By analogy, the *optimal information forager* finds the best solution to the problem of maximizing the rate of valuable information gained per unit cost.

Reference [14] expressed the rate of valuable information gained per unit cost, by the following formula:

$$R = \frac{G}{T} \tag{1}$$

where:

R, is the rate of gain of valuable information per unit cost,
G, is the ratio of the total amount of valuable information gained,
T, is the total amount of time spent.

In order to adopt both effectiveness and efficiency measures during the experimentation phase, we suggest considering the parameters G and T in representing the effectiveness and efficiency measures, respectively.

Accordingly, we propose to evaluate and compare the quality of the results obtained by our suggested approach relying on the principle of optimal information forager and using the basis of the rate R, where G and T represent respectively the total number of relevant documents returned by the retrieval system, and the total amount of search time.

3 Probabilistic Relevance Framework

The probabilistic Relevance framework is a formal framework for document retrieval which led to the development of one of the most successful text-retrieval algorithms, Okapi BM25. The classic version of Okapi BM25 term-weighting function, in which the weight w_i^{BM25} is attributed to a given term t_i in a document d, is obtained using the following formula:

$$w_i^{BM25} = \frac{tf}{k_1((1-b)+b\dfrac{dl}{avdl})+tf} w_i^{RSJ} \tag{2}$$

where:
tf, is the frequency of the term t_i in a document d;
k_1, is a constant;
b, is a constant;
dl, is the document length;
$avdl$, is the average of document length;
w_i^{RSJ}, is the well-know Robertson/Sparck Jones weight [17]:

$$w_i^{RSJ} = \log \frac{(r_i+0.5)(N-R-n_i+r_i+0.5)}{(n_i-r_i+0.5)(R-r_i+0.5)} \tag{3}$$

where:
N, is the number of documents in the whole collection;
n_i, is the number of documents in the collection containing t_i;
R, is the number of documents judged relevant;
r_i, is the number of judged relevant documents containing t_i.

The RSJ weight can be used with or without relevance information. In the absence of relevance information (the more usual scenario), the weight is reduced to a form of classical idf:

$$w_i^{IDF} = \log \frac{N-n_i+0.5}{n_i+0.5} \tag{4}$$

The final BM25 term-weighting function is therefore given by:

$$w_i^{BM25} = \frac{tf}{k_1((1-b)+b\dfrac{dl}{avdl})+tf} \log \frac{N-n_i+0.5}{n_i+0.5} \tag{5}$$

Concerning the internal parameters, a considerable number of experiments have been done, and suggest that in general values such as $1.2 < k_1 < 2$ and

$0.5 < b < 0.8$ are reasonably good in many cases. Robertson and Zaragoza [18] have indicated that published versions of Okapi BM25 are based on specific values assigned to k_1 and b: $k_1 = 2, b = 0.5$ As part of the indexing process, an inverted file is created containing the weight w_i^{BM25} of each term t_i in each document d.

The similarity score between the document d and a query q is then computed as follows:

$$Score_{BM25}(d, q) = \sum_{t_i \in q} w_i^{BM25} \tag{6}$$

During the interrogation process, the relevant documents are selected and ranked using this similarity score.

3.1 Blind Relevance Feedback for Query Expansion

One of the most successful techniques to improve the retrieval effectiveness of document ranking is to expand the original query with additional terms that best capture the actual user intent. Many approaches have been proposed to generate and extract these additional terms. The Blind Relevance Feedback (or the Pseudo-Relevance Feedback) is one of the suggested approaches. It uses the pseudo-relevant documents, i.e. the first documents retrieved in response to the initial query, to select the most important terms to be used as expansion features.

In its simplest version, the approach starts by performing an initial search on the original query using the BM25 term-weighting and the previous document-scoring function (formula 6), suppose the best ranked documents to be relevant, assign a score to each term in the top retrieved documents using a term-scoring function, and then sort them on the basis of their scores. One of the best-known functions for term-scoring is the *Robertson/Sparck Jones* term-ranking function, defined by formula 3. The original query is then expanded by adding the top ranked terms, and re-interrogated by using the BM25 similarity score (formula 6), in order to get more relevant results.

In addition to the BM25 Model, we will use the Robertson/Sparck Jones term-scoring function for Relevance Feedback as a baseline to compare the results of our proposed approach.

4 The Closeness and Co-occurrence of Terms for Effectiveness Improvement

The main goals of our proposed method is to return only the relevant documents. For that purpose, we have introduced the concept of co-occurrence and closeness, during the search process. This concept is based, at first, on finding for each query term the locations where it appears and then selecting, from these locations, the terms which frequently neighbor and co-occur with that query term. To put it simply, we recover for each query term the documents where it appears, and then assess the relevance of the terms contained in these documents to the query term on the basis of:

1. The co-occurrence, which gives value to words that appear in the largest possible number of those documents.
2. The proximity and closeness, which gives value to words in which the distance separating them and the query term within a document, with respect to the number of words, is small.

These words are then ranked on the basis of their relevance to the whole query and the top ranked ones are added to that query in order to repeat the search process.

we started our work by reducing the search space through giving importance to documents which contain at least two words of the initial query. This means that the terms, which will be added to the original query, will depend only on this set of documents. The following formula allows us to select the documents that contain at least two words of the query, i.e. to pick out any document d whose $Score_{Bigram}$ to a query q is greater than zero:

$$Score_{Bigram}(q,d) = \sum_{(t_i,t_j)\in q}^{i\neq j} (w_i^{BM25} + w_j^{BM25}) \tag{7}$$

As we previously mentioned, we will find, in the first step, the terms which often appear together with the query terms. Finding these words is done by assigning more importance to words that occur in the largest number of documents where each term of the query appears. We interpret this importance via the measurement of the external distance of each term t_i of the R_c' vocabulary to each term $t_{j_{(q)}}$ of the query q (R_c, is the set of documents returned by using the formula (7)). This distance, which does not take in consideration the content of documents, computes the rate of appearance of t_i with $t_{j_{(q)}}$ in the collection of documents R_c. In the case where t_i appears in all the documents in which $t_{j_{(q)}}$ occurs, the value of the external distance will be 1.0; and in the case where t_i does not appear in any of the documents in which $t_{j_{(q)}}$ occurs, the value of the external distance will be 0.0. Based on this interpretation, the external distance $ExtDist$ of t_i to $t_{j_{(q)}}$ is calculated as follows:

$$ExtDist(t_i, t_{j_{(q)}}) = \frac{\sum_{d_k\in R_c} x_{(i,k)} * x_{(j,k)}}{\sum_{d_k\in R_c} x_{(j,k)}} \tag{8}$$

where:

$x_{(i,k)} = \begin{cases} 1 \text{ if } t_i \in d_k, \\ 0 \text{ else.} \end{cases}$

d_k, is a document that belongs to R_c.

The total external distance between a given term t_i and the query q is estimated as follows:

$$ExtDist(t_i, q) = \sum_{t_{j_{(q)}}\in q} ExtDist(t_i, t_{j_{(q)}}) \tag{9}$$

Our dependence on this distance came as a result of the remarkable outcomes achieved in [10,11]. The distance was used during the indexing process to compute the external distance between each pair of terms of the dictionary which appear in at least one document. After that and during the search process, the original query was expanded by adding, for each term t of the initial query, the term whose external distance to t is the highest.

In the second step, we will find the terms which are often neighbors to the query terms. Therefore, we attribute more importance to terms having a short correlation with the query keywords. We interpret this importance via the measurement of the internal correlation between each term t_i of V_R and each term $t_{j_{(q)}}$ of the query q. This correlation, which takes into consideration the content of documents, computes the correlation between t_i and $t_{j_{(q)}}$ within a given document d in terms of the number of words separating them. The more t_i is close to $t_{j_{(q)}}$, the greater is its internal correlation. For this purpose, we used the well-known Gaussian kernel function to measure the internal correlation $IntDist$ between t_i and $t_{j_{(q)}}$ within a given document d:

$$IntDist_{(d)}(t_i, t_{j_{(q)}}) = exp\left[\frac{-(i-j)^2}{2\sigma^2}\right] \tag{10}$$

where:
i (resp. j), is the position of the term t_i (resp. $t_{j_{(q)}}$) in d;
σ, is a parameter to be tuned.

The terms t_i and $t_{j_{(q)}}$ may appear more than once in a document d. Therefore, the internal distance between the term pair $(t_i, t_{j_{(q)}})$ is estimated by summing all possible $IntDist_{(d)}$ between t_i and $t_{j_{(q)}}$. Thus, the preceding formula becomes:

$$IntDist_{(d)}(t_i, t_{j_{(q)}}) = \sum_{occ(t_i, t_{j_{(q)}})} exp\left[\frac{-(i-j)^2}{2\sigma^2}\right] \tag{11}$$

where:
$occ(t_i, t_{j_{(q)}})$, is the number of appearance of the term pair $(t_i, t_{j_{(q)}})$ in the document d.

The average internal correlation between t_i and $t_{j_{(q)}}$ in the whole R is then determined as follows:

$$IntDist(t_i, t_{j_{(q)}}) = \frac{\sum_{d_k \in R} IntDist_{(d_k)}(t_i, t_j)}{C(t_{j_{(q)}})} \tag{12}$$

The following formula calculates the total internal correlation between a given term t_i and the query q :

$$IntDist(t_i, q) = \sum_{t_{j_{(q)}} \in q} IntDist(t_i, t_{j_{(q)}}) \tag{13}$$

Finally, in order to compute the total correlation ($Dist$), the values of $ExtDist$ and $IntDist$ were normalized between 0 and 1. The overall correlation between t_i and q is obtained using the following formula:

$$Dist(t_i, q) = \lambda ExtDist(t_i, q) + (1 - \lambda) IntDist(t_i, q) \tag{14}$$

where:
λ, is a parameter to adjust the balance between the external and internal correlations ($\lambda \in [0, 1]$).
Using formula (14), we evaluate the relevance of each term $t \in V_R$ with respect to the query q. Then we rank the terms on the basis of their relevance and add the top ranked ones to the original query q. Based on the BM25 similarity score, presented in Sect. 3, we retrieve the relevant documents, as follows:

$$Score_{BM25}(d, q') = \sum_{t_i \in q'} w_i^{BM25} * \beta \tag{15}$$

where:
q', is the expanded query;
$\beta = \begin{cases} 1 \text{ if } t_i \in \text{ q,} \\ Dist(t_i, q) \text{ else.} \end{cases}$

5 Experiments

In order to evaluate the effectiveness of the proposed approach, we carried out a set of experiments. First, we describe the dataset, the software, and the effectiveness measures used. Then, we present the experimental results.

5.1 Dataset

Extensive experiments were performed on OHSUMED test collection. The collection consists of 348 566 references from MEDLINE, the on-line medical information database, consisting of titles and/or abstracts from 270 medical journals over a five-year period (1987-1991). In addition, the OHSUMED collection contains a set of queries, and relevance judgments (a list of which documents are relevant to each query).

In order that the results be more accurate and credible, we divided the OHSUMED collection into 6 sub-collections. Each sub-collection has been defined by a set of documents, queries, and a list of relevance documents. Table 1 summarizes the characteristics of each sub-collection in terms of the number of documents it contains, the size of the sub-collection, and the number of terms in the vocabulary (dictionary).

Regarding the queries, the OHSUMED collection includes 106 queries. Each query is accompanied by a set of relevance judgments chosen from the whole collection of documents. Partitioning the collection of documents into sub-collections leads inevitably to a decrease in the number of relevant documents

Table 1. Characteristics of the sub-collections used for evaluating the proposed approach.

Size of the collection:	**(#documents)**	**50000**	**100000**	**150000**	**200000**	**250000**	**300000**
	(Mb)	26.39	52.36	80.72	107.58	135.05	164.31
Number of terms in the dictionary		81937	120825	156009	184514	211504	237889

for each query. In other words, if we have n documents relevant to a given query q with respect to the entire collection, then surely we will have m documents relevant to the same query with respect to one of the sub-collections, where the value of n is certainly greater or equal to the value of m and, the probability of non-existence of any relevant document for a given query could be possible. In this case, in which the value of m is equal to 0, we have removed, for each sub-collection c, every query does not include any relevant document in c. Table 2 shows the number of queries (*Nb Queries*) for each sub-collection, the average query length in terms of number of words (*Avr Query Len*), the average number of relevant documents (*Avr Rel Doc*).

Table 2. Some statistics on the OHSUMED sub-collections queries.

#documents	**50000**	**100000**	**150000**	**200000**	**250000**	**300000**
Nb Queries	82	91	95	97	99	101
Avr Rel Doc	4.23	7	10.94	13.78	15.5	19.24
Avr Query Len	6.79	6.12	5.68	5.74	5.62	5.51

5.2 Software, Effectiveness Measures

The BM25 model, the Relevance Feedback technique presented in Sect. 3, and the proposed approach have been implemented in Python. All the experiments have been performed on a Sony-Vaio workstation having an Intel i3-2330M/2.20GHz processor, 4GB RAM and running Ubuntu GNU/Linux 12.04. The precision and the Mean Average Precision (MAP) have been used as measures to evaluate the effectiveness of the systems and to compare the different approaches. As indicated is Sect. 2, the principle of optimal information forager has been employed to assess the performance of the search methods.

5.3 Results

Before proceeding to compare the quality of the suggested approach with the BM25 and the Pseudo-Relevance Feedback methods, we fixed the parameter σ of the internal correlation (formula (10)). For this aim, we considered the internal correlation as the total correlation ($Dist$), i.e. $\lambda = 0$, and systematically tested a set of fixed σ values from 1 to 40 in increments of 5. Table 3 presents the precision

Table 3. The best performance of the proposed approach for different σ.

PSD	*10*		*20*		*50*	
σ	*P@10*	*MAP*	*P@10*	*MAP*	*P@10*	*MAP*
1	0.1060	0.2110	0.1048	0.2208	0.1073	0.2193
5	0.1109	**0.2265**	0.1121	0.2252	0.1146	**0.2241**
10	0.1109	0.2253	0.1121	**0.2255**	0.1146	0.2231
15	0.1109	0.2231	0.1121	0.2245	0.1146	0.2228
20	0.1109	0.2230	0.1121	0.2245	0.1146	0.2235
25	0.1109	0.2230	0.1121	0.2245	0.1146	0.2233
30	0.1109	0.2230	0.1121	0.2245	0.1146	0.2233
35	0.1109	0.2230	0.1121	0.2245	0.1146	0.2233
40	0.1109	0.2230	0.1121	0.2245	0.1146	0.2233

values after retrieving 10 documents (P@10) and the Mean Average Precision (MAP) reached by the proposed approach, while using the sub-collection of 50000 documents. The number of pseudo-relevant documents (denoted by PSD) was tuned at 10, 20 and 50.

From Table 3, we can conclude that the appropriate values of σ, which bring the best performance, are 5 and 10.

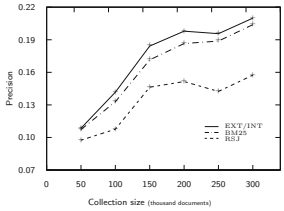
For all the following experiments, the parameters σ and λ were set to 5 and 0.5, respectively. Moreover, the number of expansion terms added to the initial query for the proposed system and the Pseudo-Relevance Feedback approaches was set to 10, which is a typical choice [4].

In the first stage of testing, we evaluated and compared the results of the suggested approach (EXT/INT), which use both the external and internal correlations, with those of BM25 and RSJ (Robertson/Sparck Jones algorithm for Relevance Feedback); where we computed the precision values after retrieving 10 documents (P@10). Figure 1 shows the precision values for the EXT/INT, the BM25 and the RSJ techniques.

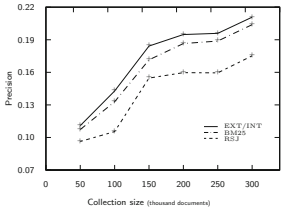
In the second stage of testing, we computed the Mean Average Precision (MAP) score to evaluate the retrieval performance of the EXT/INT, the BM25, and the Relevance Feedback method.

From Fig. 1a and b, we note an obvious superiority of the suggested approach EXT/INT compared with the BM25, and this superiority was more significant in comparison to the RSJ technique. Despite the superiority shown in Fig. 1c, the result was not similar to that observed in Fig. 1a and b, however, the precision values of the proposed approach were the best in all the sub-collections.

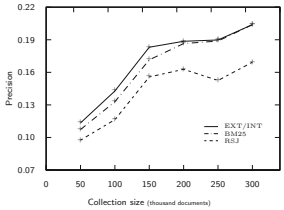
Through Fig. 1 we can conclude that the proposed method EXT/INT, compared with the rest of the search techniques, succeeded to improve the ranking of the relevant documents and made them in the first place. The precision values of the suggested system, after retrieving 10 documents, show a clear and significant superiority in front of each of BM25 and RSJ techniques, and this confirms the effectiveness of the EXT/INT approach.



(a) Precision after retrieving 10 documents (P@10), (PSD=10).



(b) Precision after retrieving 10 documents (P@10), (PSD=20).



(c) Precision after retrieving 10 documents (P@10), (PSD=50).

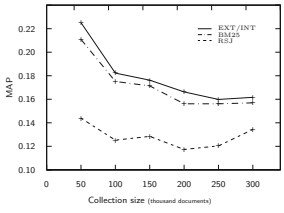
Fig. 1. Effectiveness comparison of the EXT/INT approach to the state-of-the-art.

Figure 2 shows a clear advantage of the EXT/INT approach compared to the RSJ in all the sub-collections. It also shows a slight superiority over the BM25 results.

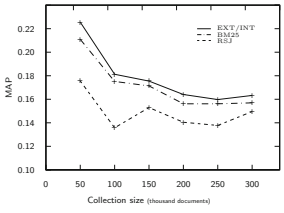
As previously explained in Sect. 2, we propose to use the principle of optimal information forager in order to adopt both effectiveness and efficiency measures in evaluating the quality of the obtained results. For this purpose, we calculate for each query the rate R, illustrated in formula 1, where the parameter G was taken as the number of relevant documents retrieved and the parameter T as the total amount of search time. The different rates, each of which is linked to a query, are then summed and divided by the total number of queries. As a result, we obtain an average rate $R(Q)$ defined as follows:

$$R(Q) = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{G_{q_i}}{T_{q_i}} \tag{16}$$

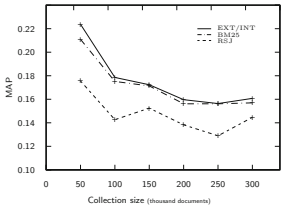
where:
$|Q|$, is the total number of queries,
G_{q_i}, is the number of relevant documents retrieved for query q_i,
T_{q_i}, is the total time spent in processing q_i.



(a) Mean Average Precision (MAP), (PSD=10).



(b) Mean Average Precision (MAP), (PSD=20).



(c) Mean Average Precision (MAP), (PSD=50).

Fig. 2. Mean Average Precision (MAP) results of the EXT/INT approach, the BM25 and the RSJ methods.

Table 4. R(Q)-score achieved by EXT/INT, BM25 and RSJ.

#documents	50000	100000	150000	200000	250000	300000
EXT/INT	14.9049	10.7110	10.1994	9.9643	9.0079	9.2276
BM25	16.0251	11.7612	11.2681	10.9097	10.0364	10.2747
RSJ	11.1365	7.3431	7.0448	6.6212	5.7615	5.9524

It can be seen from Table 4 that the BM25 overcame the EXT/INT approach in terms of $R(Q)$ values. This superiority is mainly due to the short time taken by BM25 during the search process as it used only the original query words. However, it is clear that the proposed method produces the best results over the RSJ in all cases.

6 Conclusion

In this paper, we proposed a novel term-term similarity score based on the co-occurrence and closeness of words for retrieval performance improvement. We have introduced in this work, the concept of the External/Internal similarity of terms.

We thoroughly tested our approach using the OHSUMED test collection. The experimental results show that the proposed approach EXT/INT achieved a significant improvement in effectiveness.

Although the main purpose of relying on the principle of optimal information forager, and in particular the $R(Q)$ Score, in assessing the quality of retrieval systems was not to get better results compared to BM25 and RSJ methods, but rather to introduce a new measure in order to compare the performance of retrieval systems, taking into account both effectiveness and efficiency measures.

Even though our methods perform quite well, there are some remaining issues that need to be investigated further. One limitation of this work is the use of a single test collection. The other one is that the semantic aspect of terms was not exploited in order to improve the search effectiveness.

References

1. Bharat, K., Broder, A.: A technique for measuring the relative size and overlap of public web search engines. Comput. Netw. ISDN Syst. **30**(1), 379–388 (1998)
2. Cambazoglu, B.B., Aykanat, C.: Performance of query processing implementations in ranking-based text retrieval systems using inverted indices. Inf. Process. Manage. **42**(4), 875–898 (2006)
3. Cambazoglu, B.B., Baeza-Yates, R.: Scalability Challenges in Web Search Engines. In: Melucci, M., Baeza-Yates, R. (eds.) Advanced Topics in Information Retrieval. The Information Retrieval Series, vol. 33, pp. 27–50. Springer, Heidelberg (2011)
4. Carpineto, C., Romano, G.: A survey of automatic query expansion in information retrieval. ACM Comput. Surv. **44**(1), 1–50 (2012)

5. Chen, Q., Li, M., Zhou, M.: Improving query spelling correction using web search results. In: EMNLP-CoNLL 2007: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 181–189. ACL, Stroudsburg (2007)
6. Dix, A., Howes, A., Payne, S.: Post-web cognition: evolving knowledge strategies for global information environments. Int. J. Web Eng. Technol. **1**(1), 112–126 (2003)
7. Dominich, S.: The Modern Algebra of Information Retrieval. Springer, Heidelberg (2008)
8. Eisenstein, J., OConnor, B., Smith, N.A., Xing, E.P.: Mapping the geographical diffusion of new words. In: NIPS 2012: Workshop on Social Network and Social Media Analysis: Methods, Models and Applications (2012)
9. Frøkjær, E., Hertzum, M., Hornbæk, K.: Measuring usability: are effectiveness, efficiency, and satisfaction really correlated? In: CHI 2000: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 345–352. ACM, New York (2000)
10. Khennak, I.: Classification non supervisée floue des termes basée sur la proximité pour les systèmes de recherche d'information. In: CORIA 2013: Proceedings of the 10th French Information Retrieval Conference, pp. 341–346. Unine, Neuchâtel (2013)
11. Khennak, I., Drias, H.: Term proximity and data mining techniques for information retrieval systems. In: Rocha, Á., Correia, A.M., Wilson, T., Stroetmann, K.A. (eds.) Advances in Information Systems and Technologies. AISC, vol. 206, pp. 477–486. Springer, Heidelberg (2013)
12. Ntoulas, A., Cho, J., C. Olston.: What's new on the web?: the evolution of the web from a search engine perspective. In: WWW 2004: Proceedings of the 13th International Conference on World Wide Web, pp. 1–12. ACM, New York (2004)
13. Pirolli, P.: Information Foraging Theory: Adaptive Interaction with Information. Oxford University Press, Oxford (2007)
14. Pirolli, P., Card, S.: Information foraging. Psychol. Rev. **106**(4), 643–675 (1999)
15. Ranganathan, P.: From microprocessors to nanostores: rethinking data centric systems. IEEE Comput. **44**(1), 39–48 (2011)
16. Ramos, C., Augusto, J.C., Shapiro, D.: Ambient intelligence the next step for artificial intelligence. IEEE Intell. Syst. **23**(2), 15–18 (2008)
17. Robertson, S.E., Jones, K.S.: Relevance weighting of search terms. J. Am. Soc. Inform. Sci. **27**(3), 129–146 (1976)
18. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: BM25 and beyond. Found. Trends Inf. Retrieval **3**(4), 333–389 (2009)
19. Subramaniam, L.V., Roy, S., Faruquie, T.A., Negi, S.: A survey of types of text noise and techniques to handle noisy text. In: AND 2009: Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data, pp. 115–122. ACM, New York (2009)
20. Sun, H.M.: A study of the features of internet english from the linguistic perspective. Studies in Literature and Language **1**(7), 9–103 (2010)
21. Williams, H.E., Zobel, J.: Searchable words on the web. Int. J. Digit. Libr. **5**(2), 99–105 (2005)
22. Zhu, Y., Zhong, N., Xiong, Y.: Data explosion, data nature and dataology. In: Zhong, N., Li, K., Lu, S., Chen, L. (eds.) BI 2009. LNCS, vol. 5819, pp. 147–158. Springer, Heidelberg (2009)