A Novel Information Retrieval Approach using Query Expansion and Spectral-based

Sara Alnofaie, Mohammed Dahab, Mahmoud Kamal
Computer Science
King Abdul-Aziz University
Jeddah, Saudi Arabia

***Abstract*—Most of the information retrieval (IR) models rank the documents by computing a score using only the lexicographical query terms or frequency information of the query terms in the document. These models have a limitation as they does not consider the terms proximity in the document or the term-mismatch or both of the two. The terms proximity information is an important factor that determines the relatedness of the document to the query. The ranking functions of the Spectral-Based Information Retrieval Model (SBIRM) consider the query terms frequency and proximity in the document by comparing the signals of the query terms in the spectral domain instead of the spatial domain using Discrete Wavelet Transform (DWT). The query expansion (QE) approaches are used to overcome the word-mismatch problem by adding terms to query, which have related meaning with the query. The QE approaches are divided to statistical approach Kullback-Leibler divergence (KLD) and semantic approach P-WNET that uses WordNet. These approaches enhance the performance. Based on the foregoing considerations, the objective of this research is to build an efficient QESBIRM that combines QE and proximity SBIRM by implementing the SBIRM using the DWT and KLD or P-WNET. The experiments conducted to test and evaluate the QESBIRM using Text Retrieval Conference (TREC) dataset. The result shows that the SBIRM with the KLD or P-WNET model outperform the SBIRM model in precision (P@), R-precision, Geometric Mean Average Precision (GMAP) and Mean Average Precision (MAP).**

Keywords*—*Information Retrieval; Discrete Wavelet Transform; Query Expansion; Term Signal; Spectral Based Retrieval Method

I. Introduction

Many ranking functions or similar functions such as Cosine and Okapi do not take into consideration the query terms proximity. proximity-base ranking functions based on the supposition, when the query terms closeness to each other, the document becomes more relevant to the query [1].The document that contains the query terms in one sentence or paragraph is more related than the document, which includes the query terms that far from each other. In a document, the closeness of the query terms is a significant factor as much as their frequency that must not ignore in the information retrieval (IR) model.

The Spectral-Based Information Retrieval Model (SBIRM) ranks the documents according to document scores that combine the frequency and proximity of the query terms [2]. It compares the terms of the query in the spectral domain instead of the spatial domain to take proximity in consideration without computing many comparisons. It creates a signal for a term, which maps the term frequency and position into the frequency domain and time domain respectively. To score the documents in SBIRM, compare the query terms spectrum that obtained by performing a mathematical transform such as Fourier Transform (FT) [3], Discrete Cosine Transform (DCT) [4] or Discrete Wavelet Transform (DWT) [5].

The conventional IR model lexicographic matches the query terms with the documents collection. In natural language, two terms can be lexicographically different although they are semantically similar. Therefore, directly matching the user query, which can include terms that are not present in documents leads to failure to retrieve the related documents that have other words with the same meaning. The query expansion (QE) approaches overcome vocabulary mismatch issues and enhance the performance of the retrieval by expanding the query with additional relevant terms without users' intervention. The query is expanded by subjoining either statistically related terms to the terms of the original query or semantically related terms chosen from some lexical database. Some statistical QE approaches in [6, 7, 8, 9] and semantic QE approaches in [10, 11, 12, 13, 14, 15, 16, 17] expand a query outperform IR model that ignores the proximity.

This research aims to design a QESBIRM that can retrieve the document relevant to the query terms using a proximity base IR model and QE techniques. This model combines two models: first, the SBIRM model using the DWT [5] that takes the proximity factor in its ranking function, and second, the statistical QE and semantic QE which overcomes vocabulary mismatch. With this merging, one can benefit from proximity ranking function and extend the query with more informative terms to enhance the performance of the IR model.

A thorough literature review will be presented along with a discussion of the proposed model in section two of this paper. The experiment is described in section three followed by results analysis. The conclusions and suggestions for future work will are outlined at the end of the paper.

II. Literature Review

A. Proximity-Base Information Retrieval Model

The proximity-base Model assumption is based on the fact that the document is extremely relevant to the query when the query terms occur near to each other. It uses spatial location information as a new factor to compute the document score in

information retrieval rather than touching the surface of the document by counting the query terms. The shortest substring retrieval model is one of the proximity-base Model proposed by Clarke in [18]. In this model, the document scores based on the shortest substring of text in the document that matches the query. This is done by creating a data structure called a Generalized Concordance List (GCL). These GCLs contain the query terms position in the document. This model does not consider term frequency in the documents when computing the document score although it is an important factor. It also takes long query time to create GCL and do not compute the score to the document that contains one term.

In the fuzzy proximity model [19], the document score is computed using the fuzzy proximity degree of the query terms appearance. The drawbacks of this model are that all the query terms have to occur in the document. If one query term does not occur or query terms are away from each other more than closeness parameter, the document score becomes zero. In addition, the model does not consider the frequency of the query term in the document.

Some research combines the proximity information to frequency scoring function [20, 21, 22]. The proximity IR model [20, 21] does not improve the performance significantly while the BM25P model [22] improves the performance but it is sensitive to the window size.

The Markov Random Field model considers Full Independence, Sequential Dependence, and Full Dependence between query terms but it is also sensitive to the window size [23].

In the proximity model, each query term positions is compared with the other query terms to calculate the document score. Subsequently, the comparisons number grows combinatorially if the query terms number grows [24, 25, 26]. This problem was overcome in SBIRM [2] by comparing the terms of the query in the domain of the spectral. In addition, the previous proximity models measure the proximity of the query terms only in specific region or window while SBIRM measures the proximity of the query terms in the whole document.

Briefly, the SBIRM steps are: first, the term signal is created. Then, the term signals transform into term spectra by using a spectral transform. After that, all terms spectral signal in is stored in each document. Next, the query terms signal is retrieved for every document. Finally, the document score is obtained by combing the spectra of the query terms. In the spectral domain, the query term frequency and position are represented by magnitude and phase values.

Park et al. [3] used the FT in SBIRM model. This model called Fourier Domain Scoring (FDS). Unfortunately, the FDS has a large index storage space [27]. To overcome this problem, the SBIRM use the DCT to perform document ranking [4] .The SBRM high precision still achieved by this model. The frequency information is extracted from the signal as a whole using the FT and DCT transforms.

Many data mining problems use the Wavelets transform as efficient and effective solutions [28] because it has properties [29] such as multiresolution decomposition structure. Therefore, Park and others used the DWT in SBIRM [5]. The DWT in document ranking is able to concentrate at different resolutions on the signal portions [5].The signal is break into wavelets of different scales and positions, so that it can analyze the patterns of the terms in the document at various resolutions (whole, halves, quarters, or eighths).

Using the signal concept as representation model with DWT led to improvement in the performance of text mining tasks like document clustering [30], document classification [31, 32, 33] and recommender system on Twitter [34].

B. Automatic Query Expansion Approaches

In respect of information retrieval application, there is a long history for the QE. The experimental and scientific reached by this application reached to maturity especially in laboratory settings like Text Retrieval Conference (TREC). The QE is a process of broadening the query terms using words that share statistical relationships or meaning with query terms. Usually, the queries consist of two or three terms, which are sometimes not enough to understand the expectations of the end user and fail to express topic of search. Various approaches used to expand the query over IR model that ignore the proximity information. Some of this approaches use an external resource or use target corpus or both.

The target corpus approaches is classified to local and global. The global approaches analyze the whole corpus to explore terms that co-occurred. When the terms co-occurs frequently with query term, they are consider as related terms. One of the global approaches constructs automatically in the indexing stage and named as co-occurrence thesaurus. On the other hand, the local approaches use the top relevant documents of the initial search results. The global approaches are less effective than local because they relies on the collection frequency features but are irrelevant with the terms of the query [10].

The latent semantic indexing (LSI) is classified as global approach [35]. It computes the singular value decomposition of the term-document matrix to replace the document features with smaller new features set. This new generated features are then used to expand the query. The Rocchio's is one of the sample Local approaches [36]. It expands the query with the top relevant documents terms that re-weights by sum weights of that term in all top relevant documents. Rivas and other are well known to enhance the performance of the IR using Rocchio's with the biomedical dataset [37]. The limitation of this approach is the term weight that reflects the significance of that term to the entire collection instead of its usefulness to the user query. Local approaches based on distribution analysis, which distinguishes between useful expansion terms and bad expansion terms by comparing the appearance in relevant documents with the query with that in all documents. In other words, the score of the appropriate expansion term becomes high when its frequency is high in relevant documents compared with the collection. One of this statistical comparative analysis approach uses a chi-square variant to select the pertinent terms [8]. On the other hand, The Robertson Selection Value approach Uses Swets theory [7]. Carpineto et al. [6] proposed an effective approach that depends on the terms probability distributions in the related

documents and in the corpus. In average, the Kullback-Leibler divergence (KLD) performance outperforms the previous expansion approaches based on distribution analysis when applied to selecting and weighting expansion terms [6]. Amati [9], calculates the divergence between the distributions of the terms using Bose-Einstein statistics (Bo1) and the KLD. In a different study [39], the KLD gave a good performance compared with the Bo1.

The Local context analysis (LCA) [38] is a local approach base on co-occurrence analysis. It computes term co-occurrence degree with whole query terms using co-occurrence information of the top-ranked documents. Pal and Mandar [39] proposed newLCA that tries to improve LCA [39]. The Relevance Models (RM1) is another co-occurrence approach [40]. The LCA, newLCA and RM1 sometimes do not perform at the expected level.

The external resource approaches use esources such as Dictionaries, Thesaurus, WordNet, Ontology and other semantic resources [10]. Many of the works have concentrated on the use of WordNet to improve the IR performance. Many studies extended the query using all synonyms contained in a synset which contains query terms [11, 41, 42].The rest of other approaches set all synonyms of the synset, which contain query terms as CET. They then use the word sense disambiguate (WSD) approach to determine the right sense synsets. Finally, they consider the synonyms of the right sense synsets as expansion terms. Giannis and others use the most common sense WSD approach [12]. Recently, Meili et al. [14] used the synonym of the synsets that has the same parts-of-speech with query term to extend each query terms. Fang [15] used the Jaccard coefficient to expand the query using synonyms of the synset that contain query terms and have high overlap between its glosses and the query terms glosses. Tyar and Then [16] proposed considering the glosses of the Synonyms, Hypernyms and Hyponyms synsets in the Gloss overlap WSD that using Jaccard coefficient. The drawback of these approaches are usually sensitive to WSD and the expansion terms independently of the content of the corpus and query [10].

The target corpus and external resource approaches first, use corpus as a source of candidate expansion terms (CET). Then, compute semantic similarity score of this CET with query terms using WordNet. Finally, it add the terms, which have a high score to the query. The semantic similarity measure in [17] using edge base counting approaches while in [13], the gloss overlap approach is used. The drawback of edge base counting approach is that it measures semantic similarity between two terms only if they have the same part of speech.

C. Query Expansion and Proximity-Based Information Retrieval Model

Park [3] expands the query using the Rocchio approach over the FDS model. The precision of the expanded query over the FDS model is less than the FDS without expansion [2]. Audeh [43] studied the effect of the QE on the proximity base IR. He uses LSI and WordNet synonyms to extend the query over fuzzy proximity model. The experiment showed an inadequate performance of the QE approaches over the fuzzy proximity model. The WordNet synonyms low performance can be improved by taking only the right sense instead of all query terms synonyms while the LSI can be improved by considering enough number of pseudo-documents. The fuzzy proximity model is high selectivity model. For some queries; it got less than five documents. Unlike these papers, the current study use better proximity base IR model and good performance query expansion approach. Over the BM25P He et al. [22] expanded the query using KLD QE approach that sometimes leads to a degraded performance. The performance of the MRF model improves by expanding the query using RM1 approach [44]. Unlike these papers, the current work uses better proximity base IR model and good performance query expansion approach.

III. Architecture of Proposed Model

The QESBIRM used in this work retrieves more relevant documents to the query by using good performance proximity model (SBIRM) and expands it with semantic relevant terms that overcome the mismatch problem. This is achieved by finding semantic similar term using on average the best distribution approach (KLD), the best target corpus and external resource approaches (P-WNET), and finally combining these approaches.

The proposed model is composed of two stages: text preprocessing and indexing stage; and processing query stage. The text preprocessing and indexing stage consist the following steps:

Text preprocessing, create term signals, apply weighting scheme on the term signal, apply wavelet transform on the signal and create an inverted index. It is all done in offline mode. The processing query stage steps are as follow: first, preprocessing the query. Second, apply weighting scheme on the query term. Third, retrieve query terms transformed signal. After that, compute the documents Scores. Then, the retrieved top ranking documents are sent to automatic query expansion model to extract the related terms as expansion features. Finally, the new query is sent to the spectral-based retrieval model to retrieve the final rank documents. The model architecture is shown in Fig. 1. In the following paragraphs more details on each model steps is provided.

D. Text Preprocessing

The text preprocessing is an essential part of any text mining application. At this stage, a combination of four common text-preprocessing methods were used: tokenization, case folding, stop word removal and stemming [2, 31]. First, the tokenization step, which is the task of converting a raw text file into a stream of individual tokens (words) by using spaces and line breaks and removing all punctuation marks, brackets, number and symbols [31].

Next, the case folding step which involves converting the case of every letter in the tokens to a common case. Usually, the lower case is the common case [2]. The following Stop Word Removal step ignores many terms that are not useful such as and, a, and the, in the English language because they are very common. If they are used in a query, nearly all of the documents in the set would return because every document would contain these words. If they are included in the index, the term weight would be very low.

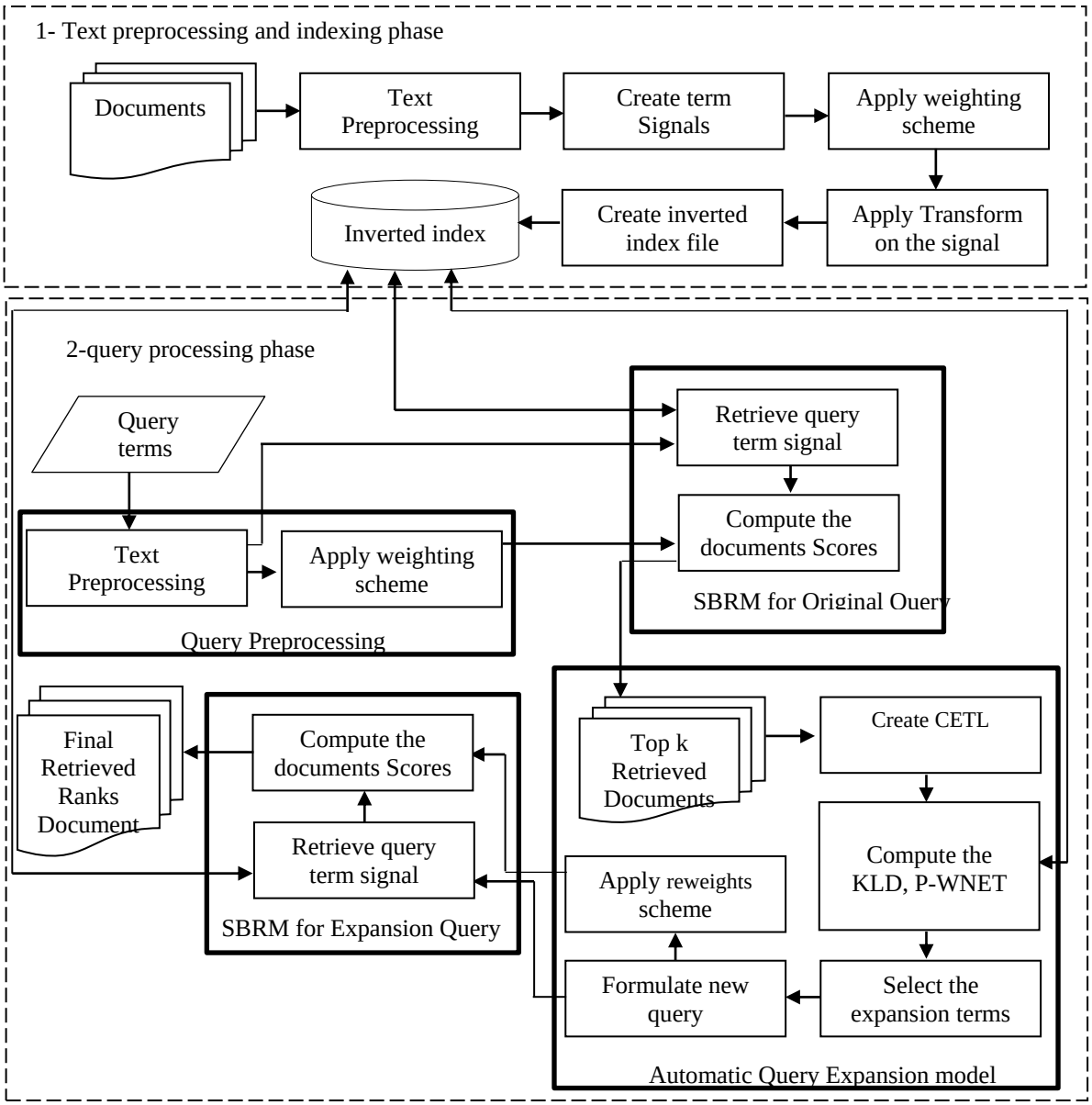

Fig. 1. General Architecture of a proposed model

Therefore, these terms are ignored. By doing this the terms number contained in the document set lexicon is reduced. Therefore, the amount of processing done by the indexer also reduces [2]. Finally, the Stemming step converts each term to its stem by removing all of a term's prefixes and suffixes [2]. The information retrieval model applies a stemming process in text preprocessing because it makes the tasks less dependent on particular forms of words. It also reduces the size of the vocabulary, which might otherwise have to contain all possible word forms [31]. In general, porter2 is the best overall stemming algorithm [45].

E. Term signals

Rather than mapping a document to a vector that contains the count of each word, the SBIRM maps each document into a collection of term signals.

The term signal, introduced by [3], it is a vector representation that displays the spread of the term throughout the document. It shows the term occurrences number in specific partitions or bins within the document.

To create the signal of the term first, divide the document into an equal number of bins. Then, represent the term t signal in document d using (1):

$$\tilde{f}_{d,t} = [\tilde{f}_{d,t,0}\ \tilde{f}_{d,t,1} \ldots \tilde{f}_{d,t,B-1}] \quad (1)$$

where $\tilde{f}_{d,t,b}$ is the number of times term t occurred in bin b in document d.

For example, document d is divided into eight bins (B=8) and they contain two terms "computer" and "data".

Fig. 2. shows how the term signal creates for the terms "computer" and "data". As shown in Fig. 2, "computer" two times occurs inbin_3, one time in bin_5 , and two times in bin_7 ; "data" occurs one time in bin_0 , one time in bin_2, three times in bin_5. The term signals for "computer" and "data " are shown in (2).

$$\tilde{f}_{d,\text{computer}} = [0,0,0,2,0,1,0,2] \quad \tilde{f}_{d,\text{data}} = [1,0,1,0,0,3,0,0] \tag{2}$$

F. Weighting Scheme

In the index stage, once the term signal created for each term in the corpus, the weighting scheme should apply to minimize the impact of highly common terms or high frequency terms in documents [4]. The BD-ACI-BCA weighting scheme was chosen as document weighting scheme in this experiments, which is shown to be one of the best methods [46]. In term signals, to apply this weighting scheme, the need to modify it to weigh the term signal instead of weighing the term in the document like Vector Space Model. In this work, it is applied to each signal component considering each bin as separate document [4].

$$\omega_{d,t,b} = \frac{1+\log f_{d,t,b}}{\omega_d} \tag{3}$$

Where $\omega_{d,t,b}$ and $f_{d,t,b}$ is the weight of term t and occurrence number of term t in bin b in document d respectively.

$$\omega_d = (1-s) + s.\frac{\omega'_d}{av_{d\epsilon D}\omega'_d} \tag{4}$$

Where s is the slope parameter (0.7), ω'_d is the document vector norm and $av_{d\epsilon D}\omega'_d$ is the average of the documents vector norm in the collection.

$$\omega_{d,t} = 1 + log\, f_{d,t} \tag{5}$$

Where $f_{d,t}$ is the term t occurrence number in document d In query stage, the following BD-ACI-BCA scheme using to weighting the query term [5]:

$$\omega_{q,t} = (1 + \log (f_{q,t})) \log (1 + f_m / f_t) \tag{6}$$

Where $\omega_{q,t}$ and $f_{q,t}$ are the weight and the frequency of the term t in query q, respectively, f_t is the documents number, which term t occurrence in, f_m is the large value of f_t for all t.

G. Signal Transform

Different signal levels resolution provide by DWT. The DWT is a sequence of high-pass and low-pass filters. The HWT can be described by high-pass filter (wavelet coefficients) is $\left[1/\sqrt{2} \quad -1/\sqrt{2}\right]$. While the low-pass filter (scaling function) is $\left[1/\sqrt{2} \quad 1/\sqrt{2}\right]$ as appears in Fig. 3[5, 47].

For example, let $\tilde{f}_{d,t}$= [3, 0, 0, 1, 1, 0, 0, 0] is the term t signal in document d when perform HWT. The Signal Transform will be W.$\tilde{f}_{d,t}$=[$\frac{5}{\sqrt{8}}, \frac{3}{\sqrt{8}}, \frac{2}{\sqrt{4}}, \frac{1}{\sqrt{4}}, \frac{3}{\sqrt{2}}, \frac{-1}{\sqrt{2}}, \frac{1}{\sqrt{2}}$, 0].

The terms positions at many resolutions appear in the transformed signal. Each transformed signal component provides term occurrences information in the specific location. In the first component, (5/√8) show that there are five term appears. The term occurrence three times in the signal first half more than the second half as in the second component. There is two more term appearance in the first quarter than the second quarter. As the fourth component appears, there is one more term appearance in the third quarter than in the fourth quarter. The signal eighths compare in the next four components.

H. Inverted index

An inverted index can created to store the word vectors. In this model, the words in each document were represented as:

$$<b_1, f_1 > < b_2, f_2 > \ldots < b_y, f_y > \tag{7}$$

Where y is the non-zero bins component, b_a is the bin number and f_a is the spectral value of bin b_a [3].

I. The Document Score

```
For each weighted term signal x in each document d
    Repeat
        n ← number of elements in x
        Compute half of x as n/2
        Initialize i to 0
        Initialize j to 0
        Set temp to empty list
        Set result to empty list
        While j < n–3
            temp[i] = (x[j] + x[j+1]) / √2
            temp[i+half] = (x[j] – x[j+1]) / √2
            i ← i + 1
            j ← j + 2
        temp[i] = (x[n–2] + x[n–1]) / √2
        temp[i+half] = (x[n–2] – x[n–1]) / √2
        result ← second have of temp
        x ← first have of temp
    Until number of elements in x = 1
```

Fig. 2. Haar wavelet transform

The SBRM [2], [5] compute the document score by using the phase and magnitude information of the query terms transform signal. The phase describes the proximity information while the magnitude value of the component describes the term frequency. To compute the score of the document, let the transformed signal of the query term t in the document d where a number of components B is $\tilde{\zeta}_{d,t} = \left[\zeta_{d,t,0}\, \zeta_{d,t,1}\, \ldots\, \zeta_{d,t,B-1}\right]$. First, for every spectral component, the magnitude and phase, defined by equation (8) and (9) are respectively calculated.

$$H_{d,t,b} = \left|\zeta_{d,t,b}\right| \tag{8}$$

and the phase which defined as

$$\phi_{d,t,b} = \frac{\zeta_{d,t,b}}{H_{d,t,b}} \quad (9)$$

Then, for each component the zero phase precision is calculated using equation (10)

$$\bar{\Phi}_{d,b} = \left| \frac{\sum_{t \in q, H_{d,t,b} \neq 0} \phi_{d,t,b}}{\#q} \right| \quad (10)$$

where q is the query terms and #(q) is the number of the query tokens. The components phases that have zero magnitudes ignores in the zero phase precision ($\bar{\Phi}_{d,b}$) *because* these phase values mean nothing. After that, the score is computed using equation (11):

$$s_{d,b} = \bar{\Phi}_{d,b} \sum_{t \in Q} H_{d,t,b} \quad (11)$$

Finally, the components scores are combined to obtain the document score:

$$S_d = \|\tilde{s}_d\|_p \quad (12)$$

where $\tilde{s}_d = [s_{d,0}\ s_{d,1}\ \ldots\ s_{d,B-1}]$ and $\|\tilde{s}_d\|_p$ is the l^p norm compute by:

$$\|\tilde{s}_d\|_p = \sum_{b=0}^{B-1} |s_{d,b}|^p \quad (13)$$

J. Kullback-Leibler divergence Query Expansion Approach

Carpineto et al. [6] proposed interesting query expansion approaches based on term distribution analysis. They used the KLD concept [48]. The distributions variance between the terms in the top relevant documents collection that is obtained from the first pass retrieval using the query and entire document collection is the base of the scoring function. The query expands with high probability terms in the top related document compared with low probability in the whole set. The KLD score of term in the CET are computed using the equation:

$$KLD(t) = P_R(t)\ log\ \frac{P_R(t)}{P_C(t)} \quad (14)$$

Where $P_R(t)$ is the term t probability in the top relevant documents R, and $P_C(t)$ is the term t probability in the corpus C, given by the following equations:

$$P_R(t) = \frac{\sum_{d \in R} f_{t,d}}{\sum_{d \in R} \sum_{v \in d} f_{v,d}} \quad (15)$$

$$P_C(t) = \frac{\sum_{d \in C} f_{t,d}}{\sum_{d \in C} \sum_{v \in d} f_{v,d}} \quad (16)$$

Where $f_{t,d}$ is the term t frequency in document d.

K. P-WNET Query Expansion approach

The scoring function of the P-WNET approach considers three parameter [13]. First, the semantic similarity between t and q_i using WordNet gloss overlap. Second, the t's rareness in the corpus. Finally, the similarity score of the top relevant document that contains t.

$$Rel_{t,q_i} = \frac{C_{t,q_i}}{C_t + C_{q_i} - C_{t,q_i}} \quad (17)$$

Where C_{t,q_i} is the number of common term between t and q_i definitions and C_t is the number of terms in t definitions.

$$idf_t = max(0.0001, log_{10} \frac{N - N_t + 0.5}{N_t + 0.5}) \quad (18)$$

$$s(t,q_i) = Rel_{t,q_i} * idf_t * \sum_{d \in R} \left(\frac{sim(d,q)}{\max_{d' \in R} sim(d',q)} \right) \quad (19)$$

$$S(t) = \sum_{q_i \in q} \frac{S(t,q_i)}{1 + S(t,q_i)} \quad (20)$$

L. Reweight scheme

After adding the expansion terms to the authentic query term, the new query must be reweighed. One of the best reweighing schemes is the scheme that is derived from KLD or P-WNET. The weight of the new query is computed using the following equation [13]:

$$\omega_{new}(t) = \alpha \frac{\omega_{orig}(t)}{\max_{v \in Q} \omega_{orig}(v)} + \beta \frac{score(t)}{\max_{v \in R} score(v)} \quad (21)$$

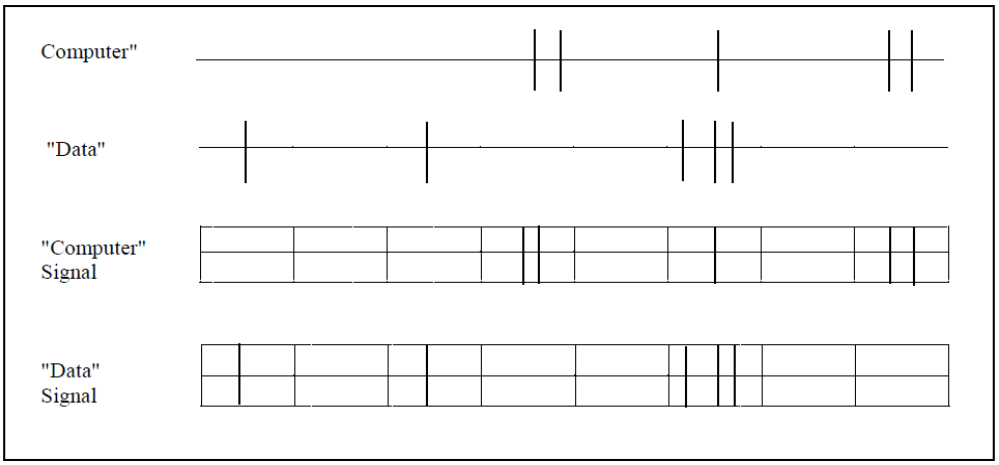

Fig. 3. The example of create the term signals

For every d document in the data set
For each term t in document d
– create term signal $\tilde{f}_{d,t}$
– weight the signal $\tilde{\omega}_{d,t}$
– Transform the signal ($\tilde{\zeta}_{d,t}$) =Spectral Transform ($\tilde{\omega}_{d,t}$)
– Stored the Transformed signals in an inverted index.

Fig. 4. The text preprocessing and indexing stage steps

Where w_orig (t) is the weight of the term t in the original query that normalized using the maximum query terms weight. The Score(t) is the KLD score or P-WNET score of the term t that also normalized using the maximum the terms score in the top document. The steps that are used in the Text preprocessing and indexing stage and the query processing stage appear in Fig. 4 and 5 respectively.

IV. Experiments

In this work, python language is used to program the proposed information retrieval model. The TREC dataset was used for the Linguistic Data Consortium (LDC) in Philadelphia, USA. The documents set is the Associated Press disk 2 and Wall Street Journal disk 2 (AP2WSJ2) which consist of 154,443 documents. The query set is the queries number from 51 to 200 from TREC 1, 2, and 3. Queries, which also called 'topics' in TREC, have special SGML mark-up tags such as narr, desc and title.

Only the queries title field used contain in average 2.3 word length. Relevance judgments are also part of the TREC collection. In fact, the relevance judgment is marked each document in the documents set as either irrelevant or relevant with every query.

To examine the performance of the QESBIRM, two experiments were conducted using the data set. The first experiment was for SBIRM model, and the second experiment was for QESBIRM that used KLD, P-WNET expansion approach.

The retrieval performance of the QE approach is affected by two parameters. One of the parameter is the top ranked documents number that known as pseudo-relevance set. The second is the informative expansion terms number that, add to the query. The parameters set to D=10 and T=20, 40, 60 respectively, which perform a good improvement base on the studies in [6, 10, 13].

V. Results

As see in Fig.6 and 7 and Table 1, the SBIRM using KLD or P-WNET improve the performance of the SBIRM. Main reasons behind this are the mismatch issues. It is concluded that the hybrid approach used in this work, i.e. SBIRM using KLD and P-WNET produces high performance in retrieve more relevant document by considering the proximity and expand the query.

VI. Conclusions

This research studies the impact of extending the query by adding statistical and semantic related terms to the original query terms on proximity base IR model. This is done by combining the SBIRM model with KLD or P-WNET. The QESBIRM using KLD, P-WNET were tested and evaluated. The experiment results show that the QESBIRM using KLD and P-WNET approach outperformed the SBIRM in precision, GMAP and MAP metric.

Recommended future work is to investigate the impact of the other QE approaches and combine KLD and P-WordNet in the SBIRM proximity base IR. It is also of interest to evaluate the model developed with samples written in other languages like Arabic. Another possible research direction is to discover the performance of other proximity base retrieval models with extending the query using QE approaches. Finally, the semantic features can use in text mining models such as text classification and clustering that consider the proximity.

TABLE I. The Performance Results of the SBIRM and QESBIRM

Approach	*P@10*	*P@15*	*P@20*	*Map*	*G-MAP*	*R-precion*
SBIRM	0.439	0.421	0.406	0.232	0.111	0.270
SBIRM with KLD (D=10,T=60)	0.465	0.447	0.421	0.244	0.113	0.271
SBIRM with KLD (D=10,T=40)	0.469	0.448	0.430	0.249	0.115	0.277
SBIRM with KLD (D=10,T=20)	0.467	0.438	0.422	0.252	0.114	0.281
SBIRM with P-WNET (D=10,T=60)	0.459	0.436	0.422	0.251	0.119	0.284
SBIRM with P-WNET (D=10,T=40)	0.472	0.444	0.422	0.245	0.117	0.277
SBIRM with P-WNET (D=10,T=20)	0.467	0.44	0.423	0.249	0.119	0.281

1) For each query term t ∈ Q
   - Retrieve inverted list I_t containing Transforming term signals $\{\tilde{\zeta}_{0,t}, \tilde{\zeta}_{1,t}, ..., \tilde{\zeta}_{d,t}\}$
2) Compute the score for each d document in set using Transform signal ($\zeta_{d,t}$).
   a) For each magnitudes of the spectral component $\zeta_{d,t,b} \in \tilde{\zeta}_{d,t}$
      i. Calculate the magnitudes of the signal component using (8)
      ii. Calculate the unit phase of the signal component using (9)
      iii. In the spectra of the word signal , For each b component
         A. Calculate the Zero phase precision using (10)
         B. Compute the score of the component as (11)
            $s_{d,b} = {}^{\Phi}{}_{d,b} \sum_{t \in Q} w_{q,t} H_{d,t,b}$
   b) Combine component score to obtain document score using (12)
3) Sort the document base on the document score.
4) Select top D retrieved documents.
5) CET list that contains all unique terms of top k retrieved documents.
6) Compute the KLD or P-WNET score for each term in CET equation using (14, 20) respectively.
7) Select T top score terms expansion terms from CET.
8) Add expansion terms to the original query to formulate the new query.
9) Re-weight the new query.

Repeat step 1, 2 and 3 with the new query.

Fig. 5. The query processing stage steps

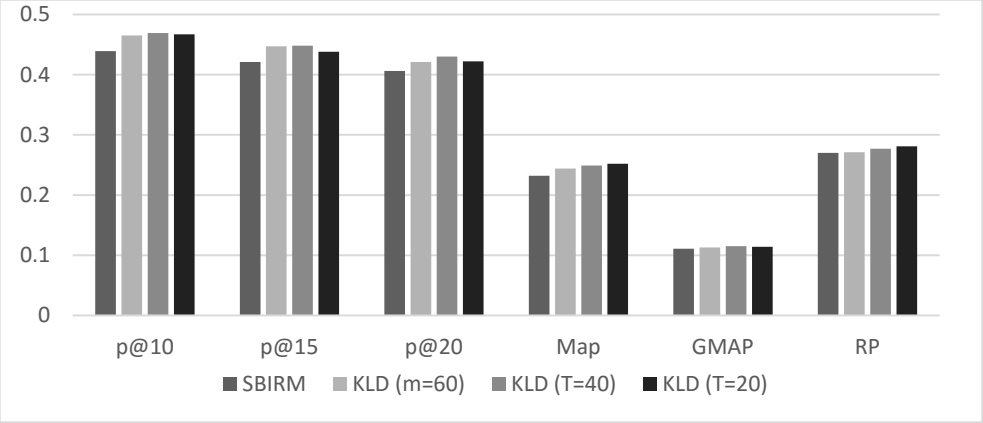

Fig. 6. Comparison of the SBIRM and KLD over the SBIRM

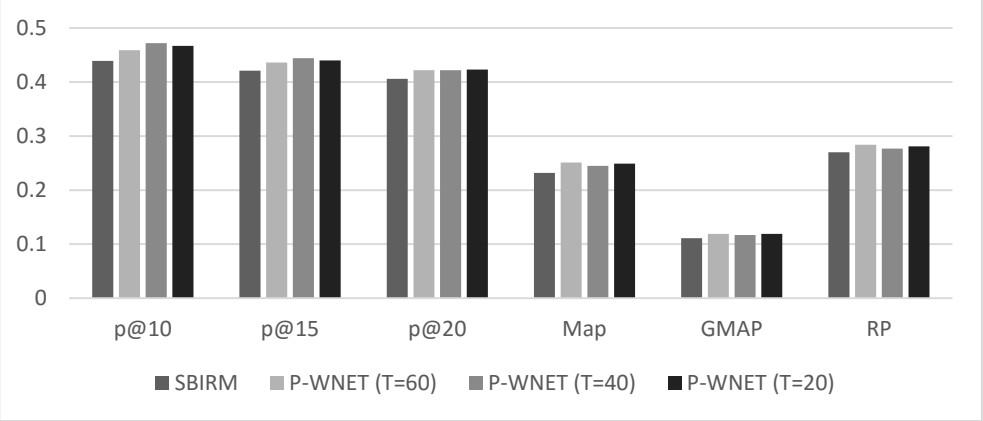

Fig. 7. Comparison of the SBIRM and P-WNET over the SBIRM

REFERENCES

[1] D. Hawking and P. Thistlewaite, Relevance weighting using distance between term occurrences, Australian National University, 1996.

[2] P. Laurence, "Spectral Based Information Retrieval," PhD thesis, Electrical and Electronic Engineering Department, Melbourne University, Melbourne, Australia, 2003.

[3] P. Laurence, K. Ramamohanarao, and M. Palaniswami, "Fourier domain scoring: A novel document ranking method," IEEE Transactions on Knowledge and Data Engineering, vol. 16, no.5, pp. 529-539, 2004.

[4] P. Laurence, K. Ramamohanarao, and M. Palaniswami, "A novel document ranking method using the discrete cosine transform," IEEE transactions on pattern analysis and machine intelligence, vol. 27, no.1, pp. 130-135, 2005.

[5] P. Laurence, K. Ramamohanarao, and M. Palaniswami, "A novel document retrieval method using the discrete wavelet transform," ACM Transactions on Information Systems, vol. 23, no.3, pp. 267-298, 2005.

[6] C. Carpineto, R. De Mori, G. Romano, and B. Bigi, "An information-theoretic approach to automatic query expansion," ACM Transactions on Information Systems, vol. 19, no.1, pp. 1-27, 2001.

[7] E. Robertson, "On term selection for query expansion," Journal of documentation, vol. 46, no. 4, pp. 359-364, 1990.

[8] T. Doszkocs, "AID, an associative interactive dictionary for online searching," Online Review, vol. 2, no. 2, pp. 163-173, 1978.

[9] G. Amati, C. Joost, and V. Rijsbergen, "Probabilistic models for information retrieval based on divergence from randomness," 2003.

[10] C. Carpineto and G. Romano, "A survey of automatic query expansion in information retrieval," ACM Computing Surveys, vol. 44, no. 1, pp. 1-50, 2012.

[11] A. Barman, J. Sarmah, and S. Sarma, "WordNet Based Information Retrieval System for Assamese," in Proceedings of IEEE 15th International Conference on Computer Modelling and Simulation, pp. 480-484, 2013.

[12] G. Varelas, E. Voutsakis, P. Raftopoulou, E. Petrakis, and E. Milios, "Semantic similarity methods in wordNet and their application to information retrieval on the web," in Proceedings of the 7th annual ACM international workshop on Web information and data management, pp. 10-16, 2005.

[13] D. Pal, M. Mitra, and K. Datta, "Improving query expansion using WordNet," Journal of the Association for Information Science and Technology, vol. 65, no. 12, pp. 2469-2478, 2014.

[14] M. Lu, X. Sun, S. Wang, D. Lo, and Y. Duan, "Query expansion via wordnet for effective code search," in Proceedings of IEEE 22nd International Conference on Software Analysis, Evolution, and Reengineering, pp. 545-549. 2015.

[15] H. Fang, "A Re-examination of Query Expansion Using Lexical Resources," in Association for Computational Linguistics, pp. 139-147, 2008.

[16] S. Tyar and M. Than, "Sense-based Information Retrieval System by using Jaccard Coefficient Based WSD Algorithm,"in Proceedings of 2015 International Conference on Future Computational Technologies, pp. 197-203, 2015.

[17] J. Singh and A. Sharan, "Co-occurrence and Semantic Similarity Based Hybrid Approach for Improving Automatic Query Expansion in Information Retrieval," in Proceedings of International Conference on Distributed Computing and Internet Technology, pp. 415-418, 2015.

[18] C. Clarke and G. Cormack, "Shortest-substring retrieval and ranking," ACM Transactions on Information Systems (TOIS), vol. 18, no. 1, pp. 44-78, 2000.

[19] M. Beigbeder and A. Mercier, "An information retrieval model using the fuzzy proximity degree of term occurences," in Proceedings of the 2005 ACM symposium on Applied computing, pp. 1018-1022, 2005.

[20] J. Peng, C. Macdonald, B. He, V. Plachouras, and I. Ounis, "Incorporating term dependency in the DFR framework," in Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 843-844, 2007.

[21] S. Buttcher C. L. A. Clarke, and B. Lushman, "Term proximity scoring for ad-hoc retrieval on very large text collections," in Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 621-622, 2006.

[22] B. He, J. X. Huang, and X. Zhou, "Modeling term proximity for probabilistic information retrieval models," Information Sciences, vol. 181, no. 14, pp. 3017-3031,2011.

[23] D. Metzler and W. Croft,. "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. ACM, 2005.

[24] A. El Mahdaouy, E. r. Gaussier, and S. d. O. El Alaoui, "Exploring term proximity statistic for Arabic information retrieval," in 2014 Third IEEE International Colloquium in Information Science and Technolog, pp. 272-277, 2014.

[25] Y. Lv , and C. Zhai. "Positional language models for information retrieval,"in Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. ACM, 2009.

[26] J. Zhao, J. Huang and B. He. "CRTER: using cross terms to enhance probabilistic information retrieval", In Proceedings of the international ACM SIGIR conference on research and development in information retrieval , pp. 155-164, 2011.

[27] K. Ramamohanarao and L. A. F. Park, "Spectral-based document retrieval," in Advances in Computer Science-ASIAN 2004. Higher-Level Decision Making: Springer, pp. 407-417, 2004.

[28] T. Li, Q. Li, S. Zhu, and M. Ogihara, "A survey on wavelet applications in data mining," ACM SIGKDD Explorations Newsletter, vol. 4, no. 2, pp. 49-68, 2002.

[29] J. Walker, A primer on wavelets and their scientific applications,CRC press, 2008.

[30] h. Almofareji, "Web Document Clustering Using Discrete Wavelet Transforms," M.S. thesis, Computer Science Department, King Abdulaziz University, Jeddah, Saudia Arabia, 2015.

[31] A. Diwali, M. Kamel, and M. Dahab, "Arabic Text-Based Chat Topic Classification Using Discrete Wavelet Transform," International Journal of Computer Science Issues, vol. 12, no, 2, p. 86, 2015.

[32] S. Thaicharoen, T. Altman, and K. J. Cios, "Structure-based document model with discrete wavelet transforms and its application to document classification," in Proceedings of the 7th Australasian Data Mining Conference-Volume 87, pp. 209-217, 2008.

[33] G. Xexéo, J. Souza, P. Castro, and W. Pinheiro, "Using wavelets to classify documents," in Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01, pp. 272-278, 2008.

[34] G. Arru, D. Feltoni Gurini, F. Gasparetti, A. Micarelli, and G. Sansonetti, "Signal-based user recommendation on twitter," in Proceedings of the 22nd International Conference on World Wide Web Steering Committee/ACM, pp. 941-944, 2013.

[35] J. Rocchio, "Relevance feedback in information retrieval," In The SMART Retrieval System- Experiments in Automatic Document Processing, pp. 313-323, 1971.

[36] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, "Indexing by latent semantic analysis," Journal of the American society for information science, vol. 41, no. 6, pp. 391, 1990.

[37] A. g. Rivas, E. Iglesias, and L. Borrajo, "Study of query expansion techniques and their application in the biomedical information retrieval," The Scientific World Journal, 2014.

[38] J. Xu and W. B. Croft, "Improving the effectiveness of information retrieval with local context analysis," ACM Transactions on Information Systems , vol. 18, no. 1, pp. 79-112, 2000.

[39] D. Pal, M. Mitra, and K. Datta, "Query expansion using term distribution and term association,",arXiv preprint arXiv:1303.0667, 2013.

[40] V. Lavrenko and W. Croft, "Relevance based language models," in Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 120-127, 2001.

[41] R. Nawab, M. Stevenson, and P. Clough, "Retrieving candidate plagiarised documents using query expansion," in Proceedings of

European Conference on Information Retrieval, Springer Berlin Heidelberg, pp. 207-218, 2012.

[42] R. Selvi and E. Raj, "An Approach to Improve Precision and Recall for Ad-hoc Information Retrieval Using SBIR Algorithm," in Proceedings of IEEE 2014 World Congress in Computing and Communication Technologies, pp. 137-141, 2014.

[43] B. Audeh, "Experiments on two Query Expansion Approaches for a Proximity-based Information Retrieval Model," in Rencontre des Jeunes Chercheurs en Recherche d'Information 2012, pp. 407-412, 2012.

[44] M. Lease, "An improved markov random field model for supporting verbose queries." in Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. ACM, pp. 476-483, 2009.

[45] M. Porter, "An algorithm for suffix stripping," Program, vol. 14, pp. 130-137, 1980.

[46] J. Zobel and A. Moffat, "Exploring the similarity space," in ACM SIGIR Forum, vol. 32. no. 1, pp. 18-34,1998.

[47] R. Hamming and L. Trefethen, "Haar wavelets," 1999.

[48] T. Cover and J. Thomas, "Elements of information," TheoryWiley, New York, 1991.