Journal of Big Data

RESEARCH Open Access

A hybrid semantic query expansion approach for Arabic information retrieval

Hiba ALMarwi[1*], Mossa Ghurab[1] and Ibrahim Al-Baltah[2]

*Correspondence:
Hebh.Almarwi@gmail.com
[1] Computer Science Department, Sanaa University, Sanaa, Yemen
Full list of author information is available at the end of the article

Abstract

In fact, most of information retrieval systems retrieve documents based on keywords matching, which are certainly fail at retrieving documents that have similar meaning with syntactical different keywords (form). One of the well-known approaches to overcome this limitation is query expansion (QE). There are several approaches in query expansion field such as statistical approach. This approach depends on term frequency to generate expansion features; nevertheless it does not consider meaning or term dependency. In addition, there are other approaches such as semantic approach which depends on a knowledge base that has a limited number of terms and relations. In this paper, researchers propose a hybrid approach for query expansion which utilizes both statistical and semantic approach. To select the optimal terms for query expansion, researchers propose an effective weighting method based on particle swarm optimization (PSO). A system prototype was implemented as a proof-of-concept, and its accuracy was evaluated. The experimental was carried out based on real dataset. The experimental results confirm that the proposed approach enhances the accuracy of query expansion.

Keywords: Query expansion, Word embeddings, Particle swarm optimization, Information retrieval, WordNet, Term frequency

Introduction

Information retrieval (IR) is an active research field that aims at extraction of the most relevant documents from large datasets. User query plays an important role in this process. A numerous efforts have been done to retrieve the relevant documents which are written in English language. Nevertheless, Arabic language has not received the deserved effort due to some inherent difficulties with the language itself. In fact, Arabic language is one of the richest human languages in its terms, varieties of sentence constructions, and diversity of meaning [1]. The sentence in Arabic language is made up of interconnected terms based on grammatical relation [2–4]. User query in most cases is too short which may neither be sufficient nor effective enough to express what the user needs [2]. Vocabulary mismatch is one of the most critical issues in IR where the user and indexer use different terms [5, 6]. Consequently, IR systems could not retrieve the documents which match the user needs. A well-known and effective strategy to resolve this issue is to perform query expansion (QE).

Query expansion is a technique that expands the initial query by adding more terms which are semantically similar to the original user query. As a result, several approaches have been introduced to process user queries. Traditional query expansion methods rely on statistical models such TF/IDF, and BM25 [7, 8]. The statistical methods depend on a term-based document retrieval which generates queries that capture the user's interests from a collection of documents. Although these methods are effective, they are not able to provide accurate information to the user query. Since those methods consider terms as atomic units of information, disregarding syntactic and semantic similarities between terms. An alternative to the statistical method is the semantic method, which attempts to find the candidate terms based on the representative meaning of the query in its context [9]. The semantic methods rely on external semantic resources such as WordNet and domain ontology. However, a complex language such as Arabic language suffers from the lack of semantic resources. Therefore, a hybrid method is needed to enhance the performance of query expansion especially for Arabic language.

Therefore, this paper proposes a hybrid semantic query expansion approach for Arabic information retrieval which calculates the weight of each term based on three information retrieval evidences namely word embedding, WordNet, and term frequency. To remove noise from the generated terms, a particle swarm optimization (PSO) is used as a semantic filtering to avoid query drift problems. The rest of the paper is organized as follows: "Background and preliminaries" section presents a brief background and preliminaries. "Related work" section reviews the related query expansion studies. The proposed approach is presented in "Framework for proposed approach" section. "Experiments and evaluation" section presents the experimental results and discussion. Finally, "Conclusion and future work" section concludes the study and outlines the future work.

Background and preliminaries

There are several approaches in query expansion. Query expansion methods which are related to our approaches are presented in the following subsections.

Term frequency

In document retrieval theory, document and query are represented as a vector in vector space. Each term in the vector has a weight which represents the importance of the term in the document as a whole. Several researchers have suggested different weighting functions. Luhn [10] studied term distribution to assign a weight to a term according to its frequency. A few years later, researchers enhanced term frequency performance by computing number of terms related to the document length. Ounis [11] studied the effect of the document length in the collection. Although this method is accepted due to its simplicity and efficiency, yet it ignores the order and semantic relations between terms. In addition, it suffers from data sparsity. As a result, this limitation makes its usage undesirable to measure words similarity.

WordNet

Most query expansion methods utilize the knowledge resource such as WordNet. WordNet is a global lexical database which organizes the terms holding identical meanings into sets called synset [12]. These synsets are connected to each other through pre-defined lexical

relations. Arabic WordNet has been constructed by the adaption of the Euro WordNet construction [13]. The Arabic wordNet contains 11,269 concepts [12], comparing with English wordNet which contains 155,287 concepts [14]. Arabic WordNet is commonly used for query expansion where appropriate senses are linked to the original query to provide the desired conceptual information. Voorhees [15] mentioned that this approach makes a little difference in retrieval effectiveness when the initial query is not well molded. On the other hand, the well molded query will improve retrieval effectiveness significantly. Furthermore, using a lexical resource alone for query expansion can cause a topic drift. This is because, the inappropriate changes in the query will cause query to match semantically other similar terms. Unfortunately, using WordNet in a query expansion process generates some noisy terms. Gong, Cheang, and Hou [16] used term semantic network to filter out the noisy terms.

Word embeddings

To obtain effective semantic term representations, term representation may implicitly be learnt by using latent dirichlet allocation [17] and latent semantic analysis [18]. These methods still consider corpora as "bag of words". Hence, they are not effective in capturing the semantic behind the text. Recently, neural network language models [19] have been used to model languages with promising results. Word embeddings are a set of language modeling such as word2vec [20] and Glove [21]. Word embeddings map each word to a vector of a real number. The vector values are learning in a way that resembles a neural network. Consequently, the technique is regularly lumped into the field of deep learning. The main idea behind word embeddings is to find dense, low-dimensions and real-valued vectors for each term within its context. The generated embeddings represent the syntax and semantic relations between terms. In embedding spaces, the words that have the similar meanings should have the similar representations. In addition, the embedding spaces show straight structure that generated word embeddings can be deciphered as relations [20]. This allows vector-oriented reasoning based on the offsets between words.

Particles swarm optimization

Particle swarm optimization is a population stochastic nonlinear optimization technique. It is inspired from the social behavior of birds. It looks for an optimal solution in search space [22]. Each solution in a search space is called a particle. All particles are initialized with velocity, position, and fitness value which are calculated by using an objective function. The algorithm is guided by personal experience (pbest), overall experience (gbest), and the present movement of the particles to decide their next positions in the search space. Further, the experiences are accelerated by two factors known as **c1** and **c2**, and two random numbers are generated between [0, 1]. In each iteration, the pbest and gbest values are calculated. After finding the two best values, particle updates its velocity and positions.

Related work

In order to overcome word mismatch problem in information retrieval, many popular solutions have been proposed by the researchers. Most early studies in Arabic language in the field of IR have focused on morphological analysis of the documents. From another point of view, many efforts have focused on developing Arabic stemmer such

as [23–25], which depends on a set of rules and uses lookup table for roots. Al-Serhan and Ayesh [26] tackled this drawback by utilizing neural network to extract Arabic root. Although it significantly increases the IR performance, most stemming techniques introduce a large amount of noise in documents. Elayeb and Boun has [27] explained the limitations of morphological analysis in Arabic IR. Traditionally, document and query represent as a vector in a vector space, and each item in the vector has a weight which reflects its importance. Different weighting functions have been suggested. Luhn [10] assigned weight to the term based on its frequencies. Ung and Park [3] proposed a term weighting function which considers the occurrence and the absence of terms. In spite of the fact that has been gotten from these methods which is sensibly great, it does not consider the semantic similar terms. Bai [28] selected expansion terms by computing correlations between pairs of terms using the association rule [5].

One of the most well-known approaches to overcome the limitations of keyword-based method is using thesaurus and domain ontology which attempt to rephrase the query based on its context [29, 30]. Yokoyama and Klyuev [31] used Japanese WordNet for query expansion. Alzahrani and Salim [32] used fuzzy concept to assign the value from 1 to 0 to reflect the degrees of similarities between Arabic documents based on ontology. Chauhan, Zhai and Zhou [33, 34] exploited ontology of sport domain to develop semantic IR system. Khan [35] developed semantic web search based on ontology. Although these approaches are effective, most complex language like Arabic has scarce of semantic resources like lexicons and ontologies. Traditional information retrieval models treat queries as a set of unrelated terms, disregarding the semantic relationships interweaving them. To enhance the performance of information retrieval, semantic methods utilize document co-occurrence statistics [18], probabilistic latent semantic analysis [36] to represent terms. Although these models have already achieved good performance, they are very costly and time consuming. To learn a viable representation of term based on its context, distributed word representation which is also known as a word embedding has been introduced in information retrieval field [37, 38]. Diaz, Mitra, and Roy [39, 40] have used contextually associated words which have been generated from word embeddings to extend user query. Liu [41] used fuzzy rules to reweigh the expansion words which have been generated from word embeddings.

As it can be seen from the reviewed studies, some limitations were found. Of these limitations, some studies were focused on statistical method which depends on the exact matching to generate the expansion terms. This is in turns neglected any potential semantic matching. On the other hand, some studies attempted to tackle the aforementioned limitation by using semantic methods, which utilizes the knowledge base during the process of expansion terms generation. Yet, this method it suffers from the limited number of terms and relations that are included in the knowledge base. Therefore, this study proposes a hybrid approach which utilizes statistical and semantic method in order to overcome the mentioned limitations and to produces better results.

Framework for proposed approach

The architecture of the proposed approach is illustrated in Fig. 1 and Table 1 respectively. An overall architecture of proposed approach is presented in Fig. 1a, b, give detailed insight of proposed approach. First the query is submitted by user to retrieve

the desired results. This query represents the input of the proposed approach. Then, the initial user query is handled. The meaningful concepts are extracted and processed using Khoja algorithm [42]. In order to get a rich set of associated terms, the initial user query is expanded. This can be done by combining candidate terms from various kinds of information sources aforementioned including WordNet, word embeddings and term frequency. The query is refined in three different stages as shown in Fig. 1b. First, the synonyms for each t_i in a user query are obtained using Arabic WordNet. They are combined with the seed query terms for further expansion in second stage. In second stage, Word2Vec is used to extract more semantic similar terms for each t_i from the previous stage by computing its cosine distance from the original word

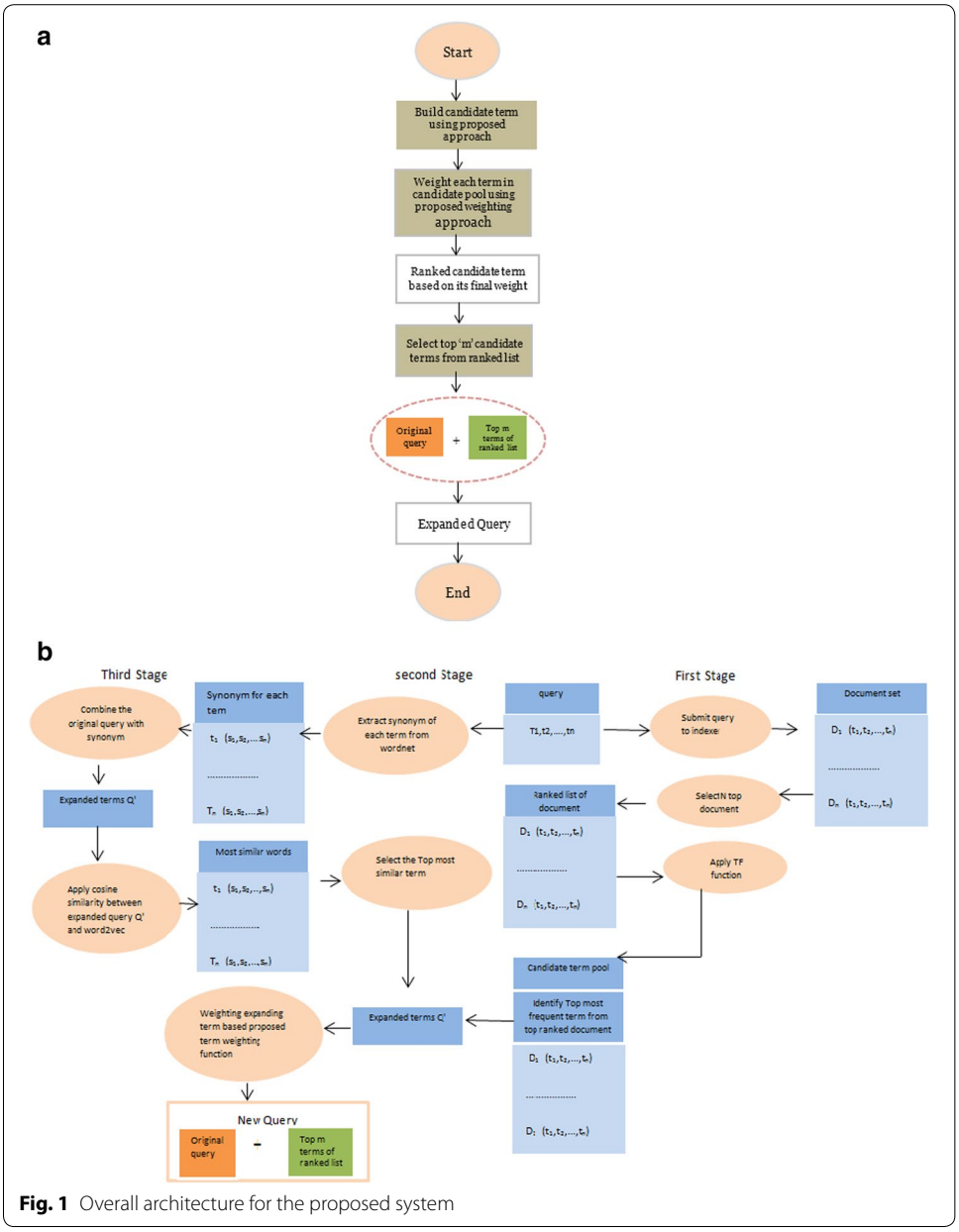

Fig. 1 Overall architecture for the proposed system

Table 1 Pseudo code of proposed system

Step 1	For each term t_i in query q Construct set of synonyms q_c based on wordnet
Step 2	Create extended query set q' by unifying original query q with q_c.
Step 3	For each term $t_{i'}$ *in* q'
	Extract the most similar relevant sense of the term within query Context based on word2vec(c)
Step 4	Select the most frequent m terms from PRD
Step 5	Unify **m** with **c** for generating final candidate term that produce the Sense of query context
Step 6	For each final candidate term tf from step 5
	Compute average weight using Eqs. 6, 7, 8
Step 7	Select optimal average weight for each term in final candidate term using PSO algorithm
Step 8	Unify the top optimal term from step 7 with original query q

in the vector space. Word2Vec (skip Gram) is chosen for our word embedding process because it has proven to be useful in capturing intensive representations of word based on its context [2]. In order to find further nominee expansion terms, most frequent terms are calculated in third stage. Frequent terms are calculated using rapid miner tool on a collection of documents that are retrieved at top ranks in response to the initial query.

The generated expanded terms from the three stages are called as candidate terms and build a candidate term pool. The values of three IR evidence namely word embeddings, WordNet and term frequency are computed for each term in candidate pool. Each evidence has its weight which represents the importance of candidate terms. PSO based term weighting approach is applied to find optimal weights for all the three IR evidences and to determine the final weight of each candidate term as it is shown in proposed PSO term-weighting approach section. After computing the weights of original query terms and candidate terms, all the terms are arranged in descending order of their final weights. And the top **K** terms are selected for query expansion. Finally, the selected expanded terms are added to original query.

Proposed AQE approach

Researchers proposed approach aims to retrieve more relevant documents. It is providing a convenient way of finding terms that are semantically related to any given query. In this section, researchers describe how extended query term set is obtained based on three IR evidence namely word embeddings, WordNet and term frequency. The researchers construct Qc, the set of synonyms for each t_i in a user query, giving a query Q consisting of **m** terms $\{t_i, \ldots, \ldots, t_m\}$ as Eq. 1

$$q_c = \{\{t_1, syn(t_1)\}, \{t_2, syn(t_2)\}, \ldots, \ldots, \{t_n, syn(t_n)\}\}. \tag{1}$$

First the synonyms for each t_i in a user query are obtained using Arabic WordNet, where $syn(t_i)$ are the synonyms of term t_i in user query. Next, researchers define an extended query term set (EQTS) q'as Eq. 2

$$q' = q \cup q_c. \tag{2}$$

q' is sent to second stage for further expansion. In this stage, Word2Vec is used to extract more semantically similar terms for each t_i from the previous stage by computing

its cosine distance from the original word in the vector space. Word2Vec (skip Gram) is chosen for our word embedding process because it has shown an efficient learning in generating high-quality word embeddings in large-scale unstructured text data [2]. Researchers define the set C of candidate expansion terms as Eq. 3

$$c = \bigcup_{t \in q'} NT(t). \tag{3}$$

where NT(t) is the set of **K** terms that are the nearest to t_i in the embedding space. Next, researchers define an extended query term set (EQTS) q' as Eq. 4

$$q' = q' \cup c. \tag{4}$$

In order to find further expansion terms, we select the most frequent **m** terms from a set of pseudo-relevant documents (PRD)—which are retrieved at top ranks in response to the initial query. The size of PRD and the number of selected terms **M** may be varied as a parameter. All the expanded terms are added to the original query which constitute a set of obvious candidates from which terms may be chosen and utilized to expand Q. In fact, some of the obtained candidate terms may not be related to the meaning of query as a whole. Therefore, term-weighting functions used to select most suitable terms by assigning weights to each term in candidate pool. A term weighting function is mathematically represented by Eq. 5 and discussed in below section. It is based on three evidence namely word embeddings, WordNet, and term frequency. Each IR evidence has its own weight which is multiplied to corresponding IR evidence value.

Proposed PSO term-weighting approach

In fact, some of obtained candidate terms are proximate neighbours of individual query terms, it is preferable to consider terms that are close to the meaning of query as a whole. The proposed PSO term-weighting function aims to select most suitable terms. It assigned weights to each term in candidate pool. It chose and included extra terms to a query. The proposed term weighting function is based on three evidence namely word embeddings, WordNet, and term frequency. The values of this evidence are computed for each term in candidate pool which is mathematically represented in Eq. 5

$$sim(q, t) = \sum_{t \in q'} w2v.wn.tf. \tag{5}$$

where sim(Q,t) is the similarity value between t_i in candidate pool and all the terms in Q. The first element of proposed term weighting function is w2v. The mathematical expression given as Eq. 6 is used to compute the mean cosine similarity between t_i in candidate pool and all the terms in Q in embedding space

$$sim(vec(t, q)) = \frac{1}{q} \sum_{t_i \in q} t.q_i. \tag{6}$$

The second element of proposed term weighting function is wn. This element indicates the mean cosine similarity between t_i in candidate pool and all terms in Q which is mathematically expressed as Eq. 7

$$sim(t,q) = \frac{1}{q}\sum_{t_i \in q} t.q_i. \tag{7}$$

The third element is tf which is one of the weighting IR evidence used in many term-weighting function. This element indicates the number of occurrences of a term in the collection. To restrict the search domain of candidate term, researchers consider only the number of times the candidate terms appear within PRD

$$tf(t) = \bigcup_{t \in PRD} count(t). \tag{8}$$

tf(*t*) is the number of times the candidate term appears in pseudo-relevant documents (PRD). Each evidence has its weight which represents the importance of candidate terms where the values of weight are between 0 and, 1. The sum of all the weights is ensured to be 1. The ideal values of the different evidences weights were founded out using PSO. The initial values of weights are taken as positions of particles. Each weigh is multiplied by its value. Then it is summed up to calculate the final weight of the term *w_Score*. That is mathematically represented in Eq. 9

$$w\_Score = w1 * w2v + w2 * wn + w3.tf. \tag{9}$$

The initial values of **c1**, **c2**, population size, *w*, velocities and a maximum number of iteration were set and *w_Score* used as the fitness value. In each iteration the fitness value of each particle is compared with other particles to get best value (best). To obtain global best value (gbest) the current population best fitness is compared with the previous population best fitness. At the end of each iteration, positions and velocities of each particle are updated. The maximum iteration is checked. The maximum weight of each term was maximized so the most appropriate candidate terms for query expansion could be identified. All terms were arranged in descending order according to their final weights. The top **M** terms were selected for query expansion. Finally, the chosen expanded terms were included to the original query.

Experiments and evaluation

In the following section, we present a set of experiments to evaluate the performance of our proposed approach. The results show that our proposed approach have achieved the best performance compared with all the other approaches.

Experimental environment

This experiment was run on a Dell desktop computer with a 64 bit i5-3470 processor CPU running at 3.20 MHz, with an 8 Mb RAM, running Microsoft windows 7 professional with service pack 1.

Building corpora (index) and query designing

Due to the lack of the available Arabic corpora, Arabic corpus was collected from different Arabic news websites using Vietspider program. Approximately 72 h were needed to collect 8 GB of Arabic Web Pages from different known news websites such as Al-Alam, BBC, CNN, and Al-Jazeera. The collected HTML pages were passed through a series of

pre-processes stages including removing non-essential HTML tags from HTML files. Arabic stop words lists such as conjunctions, prepositions, and articles do not have any effect in text mining process [43]. Furthermore, it increases the dimensionality of the text. Therefore, stop words were removed to reduce text dimensionality. Although some Arabic stop words lists are available in different studies such as [4], none of them has shown efficiency in Arabic information retrieval. Therefore, our own Arabic stop words list was used. Non-stop words were stemmed using Khoja algorithm [42]. For effective results, the process of removing stop words was combined with stemming process [5]. These preprocessing steps reduced corpora size to 1 GB. Statistical information about the dataset is provided in Table 2. The processed files were used for training and indexing purposes. For indexing creation purpose, the processed web pages were indexed using lucene. The processed files were then dumped as raw text for the purpose of training the neural network of Word2Vec framework. The parameters of Word2Vec were set as follows: word vector dimensionality 300; negative samples 25; and window size five words. These are as part of the parameter setting described in below section. On another hand, 40 query documents (QDocs) were designed manually by an expert of Arabic language to verify the correctness of our approach. Due to the paper restriction, Table 3 presents only eight queries which were selected randomly as an exemplary sample and shows the expansion terms that obtained by w2v, WordNet and the selected expansion term from PSO.

Table 3 illustrates the selected random queries. It presents expansion terms that obtained by WordNet, W2V, tf and the proposed approach. Table 4 presents the selected expansion term and their fitness value.

Parameter setting

The proposed query expansion approach has two unique parameters. **N**, that is the number of ranked documents selected as most relevant document for query expansion (size of PRD). **M** that is number of the candidate terms selected for query expansion. To find out the best performance of proposed approach, a set of experiments are performed to select suitable values of **N** and **M**. The results of these experiments are presented in following subsections.

Number of top ranked documents (N)

It is important issue to select proper number of PRD documents for query expansion. Therefore, set of experiments are performed to check size impact of PRD on IR performance. Table 5 presents the results for size of PRD varying from 5 to 20.

As it can be seen clearly from Table 5, the best performance of proposed approach in terms of a mean average precision MAP cannot be achieved effectively by a low or high number of documents. However, the highest precision results were achieved when **N** parameter is set to 10.

Table 2 Statistical about dataset

Size	Number of documents	Number of sentence	Number of words
1 GB	6464	9561	48,305

Table 3 Selected terms for the randomly chosen query

Query No	Query	Candidates obtained by WordNet	Candidates obtained by Word2Vec	Top 5 frequent terms in PRD	Selected terms obtained by PSO
1	ما أثر هجوم قوات حفتر على وسط ليبيا؟	محاولة، اعتداء، وسط، بيئة، محيط، منتصف.	عسكري، طرابلس، اعتداء، جراء، جنود، ميليشيات.	بنغازي، طرابلس، جيش، ميليشيات، حفتر.	ميليشيات، طرابلس، اد، عسكري، جنود، بنغازي.
	what is the influence of Hafter's attack in the middle of Libya?	Attempt, Assault, Middle, Environment, Surroundings. Center.	Martial, Tripoli, As a result, Led , Soldiers. Militias,	Benghazi, Tripoli, Arm, Militias, Hafter.	Militias, Tripoli , Led, Martial , Soldiers, Benghazi.
2	ما الحلول التي يمكن ان تقوم بها منظمة الصحة العالمية للمواجهة انتشار مرض الكوليرا في المها وتعز؟	NAN	ازدياد، ارتفاع، مجابهة، مصلوب، قطعة، الصلو.	الصلو، خيس، حزز، حصد، مجابهة	خيس، حزز، حصد، مصلوب. الصلو، مجابهة.
	What are the solutions can the World Health Organization (WHO handle to illuminate the spreading of cholera in Taiz and Makha?		Increase, Height, Confrontation, Musalloub, Qataba , Al. Salou.	Al. Salou, Khais, Score, Harvesting, Confrontation.	Khais, Score, Harvesting, Musalloub, Al. Salou, Confrontation.
3	كم عدد الضحايا الذين ماتوا غرقاً قبال السواحل الليبية بداية ٢٠١٧؟	زورقاً، حتف، شواطئ، بحر.	جرح، نحب، حتف، قتلوا، استهلال، شواطئ، بحر.	حتف، طرابلس، بدء، قتلوا، بلغ.	بلغ، حتف، جرح، شواطئ، بحر، استهلال.
	How many victims have been died sinking in the Libyan coast at the beginning of 2017?	Boat, Its end, beaches Sea, Start.	Wound, Death, its end, were killed, Starting, Beaches Sea.	Its end, Tripoli, Beginning, were killed, Arrive.	Arrive, Its end, Wound, beaches Sea, Starting.
4	ماذا تنص اتفاقية السلطة السورية من تنفيذ اتفاقية الفوعة كفريا الزبداني؟	ميثاق، نظام سياسي، حكومة، اداء، إجراء، تطبيق.	حكومة، نظام سياسي، سوريّة، النظام، دمشق، حلب.	سوريّة، حلب، دمشق، دستور، تطبيق.	النظام، نظام سياسي، حكومة، تطبيق. دمشق، حلب.
	what does the Syrian government agreements state from implementation of the agreement of the Foe Kafriya Zabadani?	Covenant, Political system, Government performance, Procedure, Implementation.	Government, Political system, Syria, System, Damascus, Aleppo.	Syria, Aleppo, Damascus, Constitution, Implementation.	System, Political system, Government, Implementation, Damascus, Aleppo.
5	لماذا رفض ملك المغرب لقاء رئيس الحكومة التونسية؟	عاهل، ملكية، مقابلة، عميد، مرشد، إمبراطور.	جلالة، دستور إمبراطور، عاهل، آل، عهد الملك.	آل، عاهل، رئيس، حكومة، مقابلة.	جلالة، آل، إمبراطور، عهد الملك، دستور، عاهل.
	Why did the King of Morocco refuse meet the President of Tunisian government?	King, Royal, Meeting, Dean, Director, Emperor.	Majesty, constitutional emperor, king, kings of Saudi Arabia, Reign of the king.	Kings of Saudi Arabia, King, President, Government, Meeting.	Majesty, king, Emperor, Reign of the king, Majesty, King.
6	ما التطورات المتسارعة التي يقودها حفتر لمواجهة الهجمات المحتملة في وغرب ليبيا؟	نقيض، عكس، ضد، شرق، الشرق، مجابهة.	المكس، القذافي، العقيد، الميليشيات، جهة الشرق.	اعتداء، الميليشيات، هجوم، طرابلس، القذافي.	نقيض، المكس، القذافي، العقيد، الميليشيات، جهة الشرق.
	what are the rapid developments that Haftar conducts to face the possible attacks in the south and west of Libya?	Contradictory, Opposite, Against, Confrontation, The east.	The opposite, Gaddafi, Colonel, The militias, East side.	Assault, The militias, Attack, Tripoli, Gaddafi.	Contradictory, Opposite, Gaddafi, Colonel, The militias, east side.
7	هل رافق مجزرة الهاشي كفريا والفوعة اعلام يحترم المهنة الإعلامية؟	لازم، صاحب، عربي حر، ليبرالي، متحرر.	زامن، اصطحب، مرافقة، مجزرة، مذبحة.	مجزرة، تغطية، مذبحة، صاحب، ضحايا.	زامن، لازم، صاحب، مذبحة، ليبرالي، متحرر.
	Did the free media which respect the media career accompany the Massacre of Kfariaien and Alfoaien people?	Escort, Accompanying , Arabic, Free, Liberal, Unrestricted.	Synchronize, Attach, Accompany Massacre, Butchery.	Massacre, Coverage, Butchery, Accompanying, Victims	synchronize, Escort, Accompanying, Massacre, Unrestricted.
8	ما لمطالب المحددة للرئيس حكومة الوفاق الوطني في ليبيا للإعادة الأمن والاستقرار في جنوب ليبيا؟	حكم، ولاية، نظام، نظام سياسي، عسكري.	حاسم، دبلوماسي، ليبرالي، حزبي، المطلوبة، الائتلاف	حاسم، طرابلس، عسكري، حفتر، معارضة.	طرابلس، عسكري، الائتلاف ، المطلوبة، معارضة، حاسم.
	what are the requirements of the President of National Accord government for the restoration of security and peace in southern Libya?	Rule, State, System, Political system.	Decisive, Diplomatic, Liberal, Party, Required, Opposition.	Decisive, Tripoli, Military, Hafter ,Opposition.	Tripoli, Military, Coalition, Required, Opposition, Decisive.

Number of candidate terms selected for query expansion (M)

A number of candidate expansion terms were generated by the proposed approach. Selection the most relevant candidate terms **M** from a whole candidates set is an important issue. The number of candidate terms parameter, that will be added into the submitted queries, was tuned through performing several experiments as shown in Table 6. This is to ensure the accuracy of the obtained results.

Table 4 Expansion terms and fitness values

Q #	Term	Fitness Value	Term	Fitness Value	Term	Fitness Value	Term	Fitness Value	Term	Fitness Value	Term	Fitness Value
1	ميليشيات	0.79	طرابلس	0.62	ادى	0.87	عسكري	0.89	جنود	0.65	بنغازي	0.60
2	خيس	0.96	حرز	0.86	حصد	0.68	مصلوب	0.57	الصلو	0.32	مجابهة	0.29
3	بلغ	0.93	حتف	0.92	جرح	0.90	تواطئ	0.77	بحر	0.63	استهلال	0.39
4	النظام	0.79	نظام سياسي	0.67	حكومة	0.61	تطبيق	0.52	دمشق	0.38	حلب	0. 37
5	جلالة	0.93	آل	0.65	إمبراطور	0.64	عهد الملك	0.64	دستور	0.46	عاهل	0.12
6	نقيض	0.95	العكس	0.57	القذافي	0.43	العقيد	0.41	الميليشيات	0.158	جهة الشرق	0.113
7	زامن	0.84	لازم	0.82	صاحب	0.64	مذبحة	0.64	ليبرالي	0.44	متحرر	0.41
8	طرابلس	0.87	عسكري	0.84	الائتلاف	0.77	المطلوبة	0.69	معارضة	0.68	حاسم	0.65

Table 5 Performance versus size of pseudo relevance documents

	Size of pseudo relevance documents			
	5	**10**	**15**	**20**
MAP	0.2179	*0.514*	0.3475	0.1875

Table 6 Performance versus a number of expansions term

	Number of expansions terms				
	2	**4**	**6**	**8**	**10**
MAP	0.44179	0.5175	*0.534*	0.3375	0.2937

As it can be seen clearly from Table 6, the best performance of proposed approach in terms of a mean average precision MAP cannot be achieved effectively by a low or high number of candidate terms. However, the highest precision results were achieved when **M** parameter is set to 6. It is clear from the results of experiments that applying composition of parameters (size of pseudo relevance documents and number of candidate terms) indeed effects on IR performance in terms of MAP positively.

Experimental results

To evaluate the performance of the proposed approach, two analyses is done. First, researchers use standard evaluation metrics: precision, recall and F-measure. Second, to make the result more reliable, statistical analysis is done. During the evaluation, the designed queries were used. The query relevant documents were determined by the Arabic domain specialist who provided the test queries.

Overall performance

F-measure for top ten retrieved documents are computed for the proposed approach and compared with original query, TF, WordNet and W2V approaches respectively as shown in Fig. 2. It demonstrates the higher F-measure values. It was obtained by the proposed approach in compassion with other approaches. It is clear from the experimental results shown in Fig. 2, that the proposed approach is performing better than the

original query for all forty queries. The proposed approach also obtains higher values of F-measure for 38 and 36 queries in comparison to W2V and WordNet approaches respectively. However, F-measure values are equal for two and four queries in comparison to W2V and WordNet approaches respectively. Furthermore, higher F-measure values are achieved by proposed approach over TF approach for 39 queries.

Four experiments were also conducted to evaluate the accuracy of proposed approach. The accuracy of proposed approach was calculated and compared against the accuracy of four different approaches including queries without expansion, w2v-based, TF-based and the WordNet based approaches. The accuracy calculation was carried out using Eq. 10. The following subsection presents the evaluation results.

$$Accuracy = \frac{relevant\ documents \cap retrieved\ documents}{retrieved\ documents} \tag{10}$$

The comparison accuracy between the proposed approach and without expansion

The comparison results between the proposed approach and queries without expansion are represented in Fig. 3. As it is shown by this figure; the average accuracy percentage for queries without expansion is 11% which is relatively less than the accuracy of the proposed approach. However, the highest accuracy obtained by without expansion approach is for query 4, which is still lower than the accuracy of the same query using proposed approach. It is clear from the experimental results shown in Fig. 3 that, proposed approach is more accurate than the without expansion approach for all the queries.

The comparison accuracy between the proposed approach and WordNet based approach

The comparison results between the proposed approach and WordNet are represented in Fig. 4. As it is shown by this figure; the average accuracy percentages for proposed approach and the WordNet based are 53% and 19%, respectively. Accordingly, the WordNet Based approach obtains very low accuracy for more than 32 input queries, while

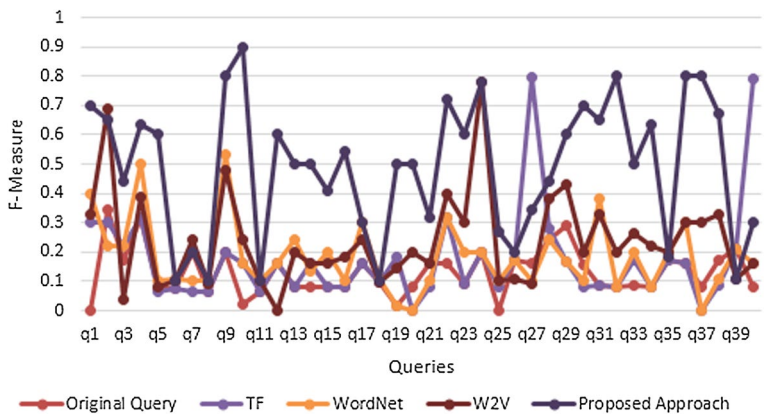

Fig. 2 F-measure values of top ten retrieved documents with respect to 40 queries

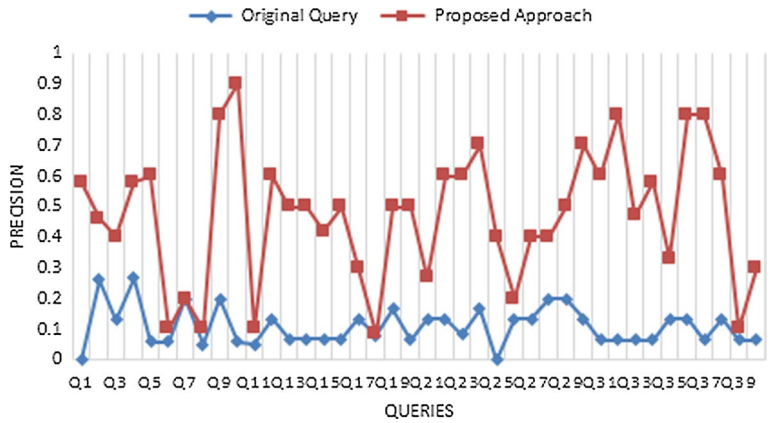

Fig. 3 Accuracy of proposed approach versus queries without expansion

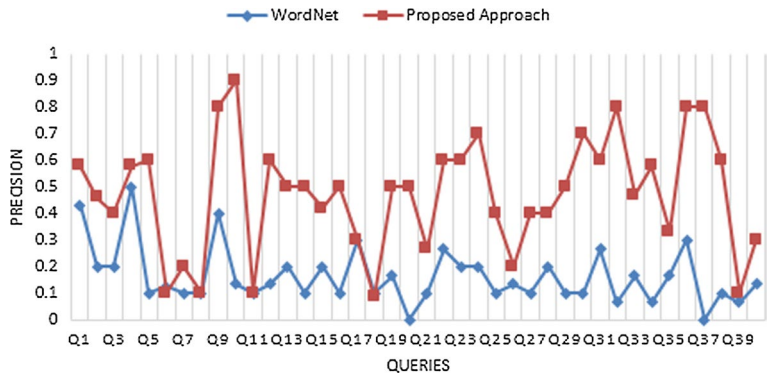

Fig. 4 Accuracy of proposed approach versus wordNet based approach

the accuracy of proposed approach is higher than the WordNet-based for more than 13 input queries. It is worthy to point out that, WordNet based approach is more accurate than without expiation approach, yet less accurate than the proposed approach.

The comparison accuracy between the proposed approach and TF based approach

The comparison results between the proposed approach and TF based approach are represented in Fig. 5. As it is shown by this figure; the average accuracy percentage for TF based approach is 14% which is relatively less than the accuracy of the proposed approach. It is clear from the experimental results shown in Fig. 5 that, proposed approach is more accurate than the TF based approach for all the queries expected for query number 15.

The comparison results between the proposed approach and W2v based approach

The comparison results between proposed approach and the w2v-based approach are shown in Fig. 6. As depicted in this figure, the average accuracy percentage for proposed approach is 53%. This means that proposed approach obtains very high accuracy

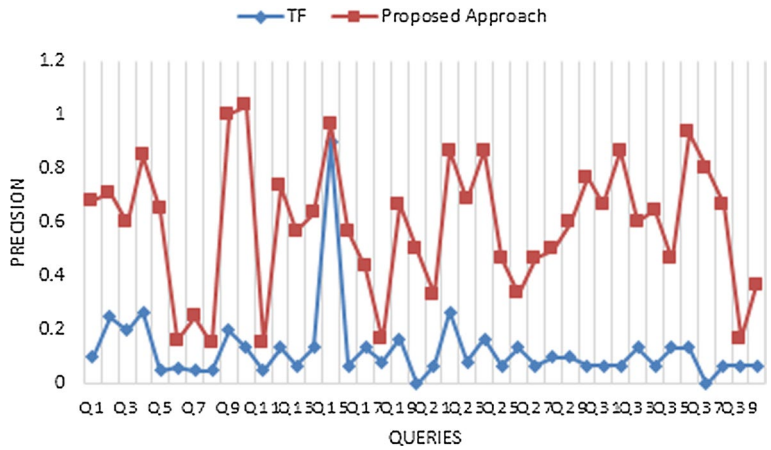

Fig. 5 Accuracy of proposed approach versus TF based approach

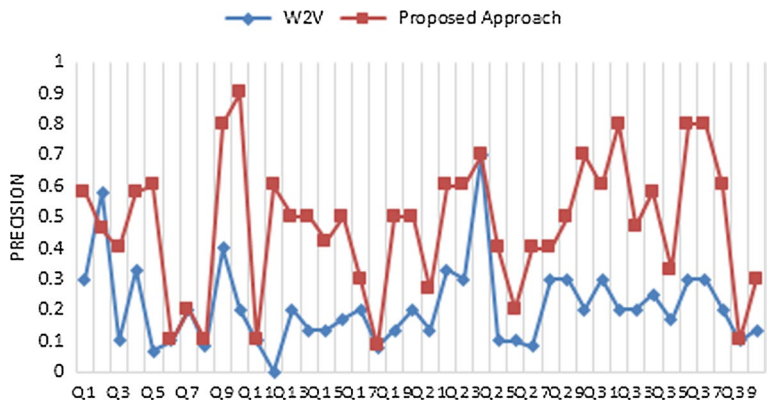

Fig. 6 Accuracy of proposed approach versus w2v-based approach

for more than 21 input queries. While the average accuracy percentage for w2v-based approach is 27.37%. W2v based approach provided higher accuracy values for only 11 input queries. The w2v-based approach was compared to the WordNet-based approach in terms of accuracy. It is worthy to point out that, the average accuracy percentage of WordNet-based approach was only 19%. This means that w2v-base approach obtains high accuracy for more than 3 input queries over wordNet based approach. As shown in Fig. 6; for all tested queries, the accuracy values of the proposed approach are higher than the accuracy values of the wordNet-based and w2v-based approaches.

Precision and recall values also computed and compared for the above selected queries using above mentioned approaches as shown in Table 7.

To check the overall performance of the proposed approach, Recall and Precision values are computed and compared with Original query, TF, WordNet, W2V based approaches as shown in Table 8.

As it can be seen clearly from Table 8, the proposed approach outperforms all other query expansion approaches. Figure 7 shows the comparison of Recall–Precision for all approaches.

it is clear from the above results the proposed approach outperform other query expansion approaches due to, proper selection of expansion terms from candidate pool.

Statistical analysis

To make the result more reliable, statistical paired t-test analysis is also computed. Table 9 shows the improvement of proposed approach against other approach is statistically significant at $\propto = 0.05$. The proposed approach statistically outperform other approaches as p-values are 0.0257, 0.0258, 0.0295 and 0.0330 for Original query, tf-based, WordNet-based and W2v-based approaches respectively (Fig. 8).

The results from different analysis demonstrate that proposed approach have achieved the best performance compared with all the other approaches.

Discussions

Table 7 lists the accuracy-Recall values of the four approaches for eight randomly selected input queries. By exploring these observations, there are two fundamental key discoveries: the cases that the proposed approach outperformed the other approaches and vice versa. First, for some queries such as queries Q#2, Q#7, Q#6, Q#3 and Q#1, the proposed approach results outperformed the other approaches. For example, in case of query no. 7 the term مذبحه is synonym to the third query term and term متحرر appears with query term اعلام in many documents. Therefore, these terms can be added as new term for query expansion. Similarly, for query no. 2 term الصلو and term مصلوب comes with the تعز in many documents (refer to Table 3). Consequently, these terms are added in original query using proposed approach and it improves the accuracy.

In case of query no. 3 the term مطلع and term خطوة أولى generates as synonym for term بدأيه using wordNet method. Besides, the terms بحر شواطئ,بلغ,زورقا,قُتلوا,حتف,نحب,جرح generates as expanded terms using w2v-base approach. The computed accuracy for expanded query using wordNet is 0.4 and for modified query using w2v-base method is 0.5. The terms مطلع, أول خطوة , نحب,حتف is not selected as expanded terms using proposed approach, hence it improves the accuracy from 0.4 and 0.5 to 0.7. The reason behind this improvement is that, PSO plays an important role in selecting the suitable terms for query expansion and make query more specific. Therefore, most of the retrieved documents are relevant, and hence the results of the proposed approach were better than the other approaches. Table 8 presents the comparison of MAP of proposed approach with other query expansion approaches.

Second, for some queries including Q#4, Q#5, and Q#8, the WordNet, TF and w2v-based approaches fetched better results than the proposed approach. For example, in query Q#4 the terms النظام, نظام, سياسي,حكومة is added as new term for query expansion using proposed approach. In such cases, the proposed approach fails to remove inappropriate terms from the candidate pool which cause lags our approach behinds other approaches.

Conclusion and future work

In this paper, a hybrid query expansion approach for Arabic information retrieval was proposed. This approach combines statistical and semantic method to utilize the advantages and strengths of each method. Thus, the term mismatch limitation of the statistical

Table 7 Recall and Precision values for the above selected queries using different query expansion approaches

Query no.	Original query		TF		WordNet		W2V		Proposed approach	
	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision
1	0.145	0.088	0.194	0.083	0.244	0.21	0.36	0.3	0.6	0.6
2	0.25	0.166	0.255	0.166	0.260	0.251	0.56	0.475	0.875	0.76
3	0.124	0. 076	0.136	0.064	0.295	0.166	0.29	0.2	0.8	0.8
4	0.253	0.177	0.257	0.1397	0.29	0.1331	0.68	0.68	0.43	0.41
5	0.150	0.075	0.17	0.11	0.81	0.75	0.78	0.78	0.69	0.65
6	0.133	0.10	0.43	0.33	0.58	0.43	0.66	0.59	0.80	0.74
7	0.135	0.094	0.100	0.038	0.168	0.166	0.278	0.22	0.78	0.68
8	0.170	0.09	0.18	0.16	0.2412	0.21	0.58	0.55	0.43	0.41

Table 8 Recall and Precision values using different query expansion approaches

Original query		TF		WordNet		W2V		Proposed approach	
Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision
0.17375	0.1194	0.1985	0.1488	0.2163	0.1974	0.2737	0.2706	0.53062	0.4728

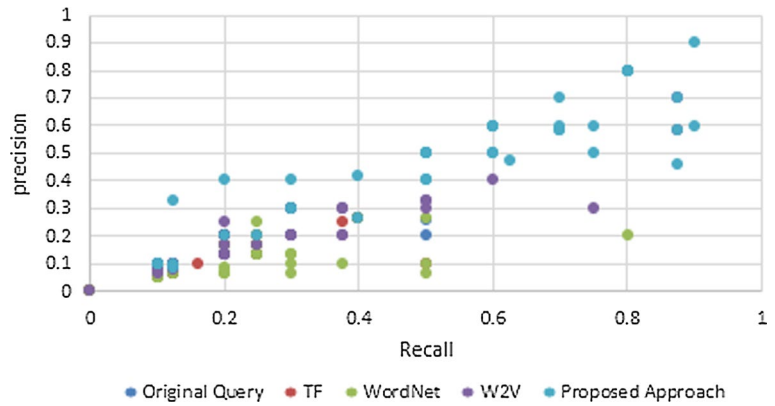

Fig. 7 Recall–Precision graph for different approaches

Table 9 Paired t-test results

Approach	h-value	p-value
Original query	1	0.0257
tf-based approach	1	0.0258
WordNet-based approach	1	0.0295
W2v-based approach	1	0.0330

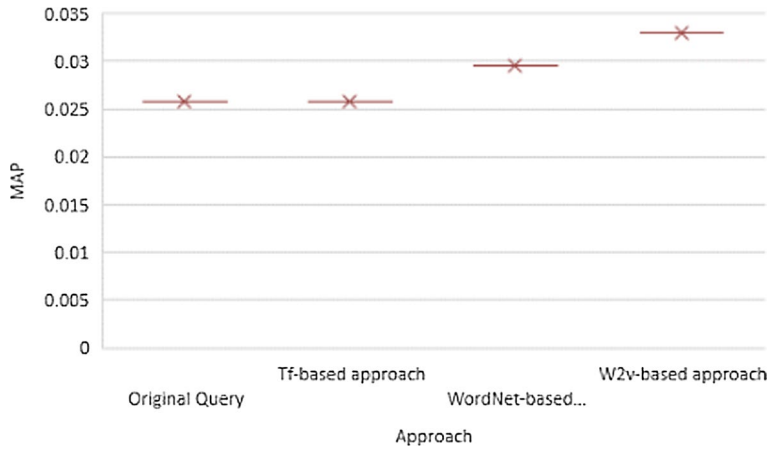

Fig. 8 Paired t-test results

method is tackled by allowing semantically similar words to contribute to the scoring function. The proposed approach generates expansion terms closely related to the meaning of a query as a whole. To find the degree of the importance of the expanded terms to the query as a whole, the weights of each candidate term were computed based on three evidence namely word embedding, WordNet, and term frequency. Moreover, the particles swarm optimization was used to select the most suitable terms for query expansion. The experimental results that were carried out on real dataset confirmed that the proposed approach increased the value of accuracy in terms of information retrieval and demonstrated the effectiveness of the proposed approach.

An interesting point for future work would be to extend the evaluation to include more domains to ensure the applicability of the proposed approach in different domains.

Abbreviations

QE: Query expansion; PSO: Particle swarm optimization; IR: Information retrieval; W2v: Word to vector; TF/IDF: Term frequency/inverse-document frequency; TF: Term frequency; pbest: Personal best value; gbest: Global best value.

Acknowledgements

This paper and the research behind it would not have been possible without the exceptional support of my supervisor, Dr. Mossa Ghurab. His enthusiasm, knowledge and exacting attention to detail have been an inspiration and kept my work on track. I have been extremely lucky to have a supervisor who cared so much about my work, and who responded to my questions and queries so promptly. I would also like to thank the Dr. Ibrahim Al-Baltah for his valuable and constructive suggestions during the editing of this research work.

Authors' contributions

HA took on the main role performed the literature, designed, performed experiments, analyzed data and wrote the paper. MG supervised the research and co-wrote the paper. IA reviewed the manuscript language and helped in edit manuscript. All authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

Data will not be shared.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

Not applicable.

Author details

[1] Computer Science Department, Sanaa University, Sanaa, Yemen. [2] Information Technology Department, Sanaa University, Sanaa, Yemen.

Received: 19 June 2019 Accepted: 22 May 2020
Published online: 29 June 2020

References

1. Atwan J, Mohd M, Rashaideh H, Kanaan G. Semantically enhanced pseudo relevance feedback for arabic information retrieval. J Inf Sci. 2016;42(2):246–60.
2. Sadowski C, Stolee KT, Elbaum S. How developers search for code: a case study. In: Proceedings of the 2015 10th joint meeting on foundations of software engineering. 2015. p. 191–201.
3. Jung Y, Park H, Du D-Z. An effective term-weighting scheme for information retrieval. Computer Science Technical Report TR008. Department of Computer Science, University of Minnesota, Minneapolis, Minnesota. 2000. p. 1–15.
4. Lau T, Horvitz E. Patterns of search: analyzing and modeling web query refinement. In: UM99 user modeling. Springer; 1999. p. 119–28.
5. Carpineto C, Romano G. A survey of automatic query expansion in information retrieval. ACM Comput Surv. 2012;44(1):1–50.
6. Furnas GW, Landauer TK, Gomez LM, Dumais ST. The vocabulary problem in human–system communication. Commun ACM. 1987;30(11):964–71.
7. 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. 2004. p. 42–9.

8. Beil F, Ester M, Xu X. Frequent term-based text clustering. In: Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining. 2002. p. 436–42.
9. Shaalan K, Al-Sheikh S, Oroumchian F. Query expansion based-on similarity of terms for improving Arabic information retrieval. In: International conference on intelligent information processing. Springer; 2012. p. 167–76.
10. Luhn HP. The automatic creation of literature abstracts. IBM J Res Dev. 1958;2(2):159–65.
11. He B, Ounis I. Term frequency normalisation tuning for bm25 and dfr models. In: European conference on information retrieval. Springer; 2005. p. 200–14.
12. ElKateb S, Black W, Rodríguez H, Alkhalifa M, Vossen P, Pease A, Fellbaum C. Building a wordnet for arabic. In: LREC. 2006. p. 29–34.
13. Gonzalo J. Sense proximity versus sense relations. GWC. 2004;2004:5.
14. Fellbaum C. A semantic network of english verbs. WordNet Electron Lex Database. 1998;3:153–78.
15. Voorhees EM. Query expansion using lexical-semantic relations. In: SIGIR'94. Springer; 1994. p. 61–9.
16. Gong Z, Cheang CW, Hou UL. Web query expansion by wordnet. In: International conference on database and expert systems applications. Springer; 2005. p. 166–75.
17. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res. 2003;3(Jan):993–1022.
18. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. J Am Soc Inf Sci. 1990;41(6):391–407.
19. Sergienko R, Gasanova T, Semenkin E, Minker W. Collectives of term weighting methods for natural language call routing. In: Informatics in control, automation and robotics. Springer; 2016. p. 99–110.
20. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. 2013. p. 3111–9.
21. Pennington J, Socher R, Manning CD. Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014. p. 1532–43.
22. Clerc M, Kennedy J. The particle swarm-explosion, stability, and convergence in a multidimensional complex space. IEEE Trans Evol Comput. 2002;6(1):58–73.
23. Kadri Y, Nie J-Y. Effective stemming for arabic information retrieval. In: Proceedings of the challenge of arabic for NLP/MT conference, Londres, Royaume-Uni. 2006. p. 68–74.
24. Boudchiche M, Mazroui A, Bebah MOAO, Lakhouaja A, Boudlal A. Alkhalil morpho sys 2: a robust arabic morpho-syntactic analyzer. J King Saud Univ Comput Inf Sci. 2017;29(2):141–6.
25. Pasha A, Al-Badrashiny M, Diab MT, El Kholy A, Eskander R, Habash N, Pooleery M, Rambow O, Roth R. Madamira: a fast, comprehensive tool for morphological analysis and disambiguation of arabic. LREC. 2014;14:1094–101.
26. Al-Serhan H, Ayesh A. A triliteral word roots extraction using neural network for arabic. In: 2006 International conference on computer engineering and systems. IEEE; 2006. p. 436–40.
27. Elayeb B, Bounhas I. Arabic cross-language information retrieval: a review. ACM Trans Asian Low-Resour Lang Inf Process. 2016;15(3):1–44.
28. Bai J, Nie J-Y, Cao G, Bouchard H. Using query contexts in information retrieval. In: Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval. 2007. p. 15–22.
29. Shen X, Xu Y, Yu J, Zhang K. Intelligent search engine based on formal concept analysis. In: 2007 IEEE international conference on granular computing (GRC 2007). IEEE; 2007. p. 669
30. Froud H, Lachkar A, Ouatik SA. Stemming versus light stemming for measuring the similarity between arabic words with latent semantic analysis model. In: 2012 colloquium in information science and technology. IEEE; 2012. p. 69–73.
31. Yokoyama A, Klyuev V. Search engine query expansion using Japanese wordnet. In: 2010 3rd international conference on human-centric computing. IEEE; 2010. p. 1–5.
32. Alzahrani SM, Salim N. On the use of fuzzy information retrieval for gauging similarity of Arabic documents. In: 2009 second international conference on the applications of digital information and web technologies. IEEE; 2009. p. 539–44.
33. Chauhan R, Goudar R, Sharma R, Chauhan A. Domain ontology based semantic search for efficient information retrieval through automatic query expansion. In: 2013 international conference on intelligent systems and signal processing (ISSP). IEEE; 2013. p. 397–402.
34. Zhai J, Zhou K. Semantic retrieval for sports information based on ontology and sparql. In: 2010 international conference of information science and management engineering, vol. 1. IEEE; 2010. p. 395–8.
35. Khan HU, Saqlain SM, Shoaib M, Sher M. Ontology based semantic search in holy quran. Int J Fut Comput Commun. 2013;2(6):570.
36. Hong L. A tutorial on probabilistic latent semantic analysis. 2012. arXiv preprint arXiv:1212.3900.
37. Zhou G, He T, Zhao J, Hu P. Learning continuous word embedding with metadata for question retrieval in community question answering. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (Vol. 1: Long Papers). 2015. p. 250–9.
38. Zhang M, Liu Y, Luan H, Sun M, Izuha T, Hao J. Building earth mover's distance on bilingual word embeddings for machine translation. In: Thirtieth AAAI conference on artificial intelligence. 2016.
39. Diaz F, Mitra B, Craswell N. Query expansion with locally-trained word embeddings. 2016. arXiv preprint arXiv :1605.07891.
40. Roy D, Paul D, Mitra M, Garain U. Using word embeddings for automatic query expansion. 2016. arXiv preprint arXiv :1606.07608.
41. Liu Q, Huang H, Lut J, Gao Y, Zhang G. Enhanced word embedding similarity measures using fuzzy rules for query expansion. In: 2017 IEEE international conference on fuzzy systems (FUZZ-IEEE). IEEE; 2017. p. 1–6.
42. Khoja S. Stemming arabic text. Lancaster: Computing Department, Lancaster University; 1999.
43. Jansen BJ, Booth DL, Spink A. Determining the informational, navigational, and transactional intent of web queries. Inf Process Manage. 2008;44(3):1251–66.

Publisher's Note