Optimizing an Arabic Query using Comprehensive Query Expansion Techniques

Mohammed Otair, Ph.D
Amman Arab University
Jordan - Amman

Ghassan Kanaan
Amman Arab University
Jordan - Amman

Raed Kanaan, Ph.D
Amman Arab University
Jordan - Amman

ABSTRACT
By utilizing a search engine for the inquired object, the user may get what he has looked for. However, in the average; the number of words the user comes up with for a query is two or three in general [23]. This mostly causes a number of problems. To overcome such problems, various query expansion techniques have been developed. However, none of them are asserted to present the optimal solution, especially in Arabic language because its complex morphological structure. Thus, the main objective of this paper is to optimize Arabic queries using comprehensive combination of these expansion techniques that can be used to enhance the process of query expansion and to retrieve the maximum number of the relevant documents for the Arabic user's query. The paper found that the developed system improved the recall and precision over couples of separated techniques. This method gets the benefits of both expansion approaches: interactive and automatic query; because the inquired object is automatically expanded and users are discretely engaged in query expansion.

General Terms
Information Retrieval, Query Expansion.

Keywords
Information Retrieval, Query Expansion, Thesaurus, Relevance Feedback, Interactive Query Expansion.

1. INTRODUCTION
One of the major causes of limiting going into search is the poor request of forming performance. In the early process of searching users usually form obscure, inexact or unfinished query. That is because they lack the initial knowledge about the needed information, or users do not present established expressions or well-known synonymous, therefore, their search results end up to be poor [15]. There are hundreds of words in the Arabic language and they may be derived or drawn out from one single original root [10]. Thus, without recovering all documents that contain all variants of the terms' query, and Arabic engines come up with poor results. This is in addition to that user query expressions are mostly short (on average number of words less than two [34]) which is vague and of the wrong recipe [21]. Several techniques of query expansion should be used to be able to surmount these problems. Query can be expanded by increasing a few main words of the query to many related main words (variants or synonymous), in the purpose of increasing the recall of the submitted query. The techniques of query expansion are quite significant to the retrieval system of Arabic language based on its specific and complex morphology. For the purpose of concept building, Arabic query words can be expanded by synonymous words which require an automatic classification of Arabic documents [33].

The query is expanded using some terms which have a much related meaning to those in the original query. The expansion terms process can be done using thesauri [27], and generally there are two types of thesauri: hand-crafted thesauri and corpus-based automatically constructed thesauri. Thesauri have repeatedly been integrated with most of information retrieval (IR) systems as a tool for the identification of variants expressions and linguistic objects that are similar from different perspective but superficially distinct. Thesauri have been proposed to expand query automatically of research for relatively some decades, and many methods have been invented [19]. Many techniques and algorithms for IR Systems such as building thesauri and query expansion have been developed and proposed in the literature [20].

Utilizing specific arena thesauri is been reported that it is coming up with good results for query expansion [11]. However, query expansion without term reweighing not use relevance feedback as it will be described later in this paper. The terms that can be added are looked up in thesauri, and it can be manually, semi-automatically, or automatically built. At the present there is very little number of specific arenas of thesauri, where the number of such arenas in the world is quite in large numbers. Unlike, setting up such type of manually thesauri requires a lot of human efforts from some domain experts and will be taking plenty of time. At the other side, query expansion techniques which use a general purpose thesaurus have mentioned by a number of researchers that these techniques failed to consolidate information retrieval performance [38, 40]. The query of the user is expanded via many ways such as global query expansion [45], local query expansion [44], relevance feedback [36] and user's logs [13] methods. The user's query can be expanded via the previous methods automatically without the intervention of the user based on some information such as, the frequency of index words or the classification of data and other information [26].

The search engine brings about more precise results to users and that depends on a revised word list, which can be brought by improving or expanding the initial query expression or expressions. Generally for query expansion, there are two approaches that are mostly used: (a) interactive query expansion (IQE) [29] (b) automatic query expansion (AQE) [31]. In automatic query expansion, the user query is automatically expanded based on to some more useful information such as document categorization, co-occurrence data and so forth [31]. In the opposite approach in respect of interactive query expansion, a user had to add some extra search expressions and then the elementary query is improved or expanded according to these chosen expressions. Relevance feedback is another alternative of IQE [35].]. In the relevance feedback technique, a user has to include the relevant documents in starting set of the documents being retrieved. Then a modified query is formed depends on the basic of these similar documents that were chosen by the user. This is in addition to that the thesaurus has taken significant concern for query formulation or expansion such as: synonym-based thesaurus [16] and the similarity thesaurus [28]. To set up synonym-based thesaurus, a group of synonym expressions or words should be chosen from an index words from special

dictionary. However, the expression or word of co-occurrence data is not treated within the thesaurus of synonym-based. In the similarity thesaurus [32], term similarity (which is a term-to-term relationship) is identified and then the thesaurus for a word is set from words with high resemblance to that word.

This paper presents a new methodology for building an automatic thesaurus and using several techniques of query expansion for Arabic document retrieval. The proposed thesaurus construction approach and the proposed query expansion approach can improve the performance of IR systems for dealing with document retrieval. The proposed method emphasizes on query expansion techniques applied to Arabic text. The developed query expansion techniques usually lead to a reduction in the number of retrieved documents, and to an increment in the recall and the precision at the same time.

2. ARABIC LANGUAGE

23 countries speak an Arabic language as their official language, and it is one of the official languages of the United Nations. It is estimated that with nearly 422 million native speakers. Now, Arabic language is the highest number of being the mother language to people who speak it after Chinese [http://encarta.msn.com/encyclopedia 7615706474/]. The Arabic alphabet consists of 28 letters (16 of them have one dot, two or three dots); it is distinguished by its use of connected letters in writing. Arabic characters change shape based on their position within words. This extends the Arabic alphabet to ninety different character representations [30]. An Arabic letter might have four different shapes: isolated, initial, medial, and final. In Arabic language, quite big numbers of words are morphologically deduced from a dedicated list of original roots. These roots are the form of armless verbs; they can be: pentaliteral, quadrilateral or trilateral. Most of the Arabic roots are made up from the three consonants [47]. Arabic words are grouped into: nouns, verbs, and particles. Some nouns and all verbs are deduced from a root. Arabic sentences are either nominal or verbal. Nominal sentences begin with a subject followed by a noun, an adjective, a prepositional phrase, or an adverb. Verbal sentences contain a verb before a nominative noun (the subject), and may contain complements.

The 28 letters of Arabic language and its main characteristic is that the words are can be resolved down-to shared roots and constructed from shared roots. Exclusions to this trend or rule are shared particles and nouns [16]. The Arabic language is mainly inflectional language that has 85% of words deduced from trilateral roots. Nouns and verbs are deduced from a wide range set of roots which contains about 10,000 of original roots [4]. Three genders: neuter, feminine, and masculine; and three numbers: plural, dual, and singular are found in Arabic language [16]. Arabic Language has particular morphology characteristics that make it very complex to build natural language processing methods for retrieval systems in this language [16]. The Arabic language is distinct from other language in respect of number of substantial aspects, one of them that Arabic sentences and words are written and read from right to left. This language does not have vowels because it is principally a consonant language in its written forms. The noun and the verb are the major components of speech in Arabic languages, which are organized of trilateral original roots [16].

Most of Arabic word consists of threefold or fourfold roots, which are updated according to a number of rules to formulate a set of words which have very closed or similar meanings. This is done by combining suffixes, prefixes and infixes to the roots. Presently, the Arabic language has common forms which are: the slang language, the classical Arabic, and the modern standard Arabic [1]. Further information about the Arabic language Characteristics can be found at [7].

3. THESAURUS

A thesaurus (plural is thesauri) which is the main significant tool in information retrieval systems in respects of searching and indexing process. Thesaurus can be used with a concern for application of words with a view to enhance or expand the query (i.e. query expansion) [25]. Actually, there are many different definitions of thesauri; an example of a moderate definition is given by Schütze [37]: "simply a mapping from words to other closely related words". Thesauri can help the searcher in the searching process to find some related terms to the initial query. There are two types of constructed thesaurus: the manual (linguistic) and the automatic (statistical). A manual thesaurus is often too broad or it can be narrow. Building and maintaining thesauruses manually could be expensive and time consuming task and it may be affected by the experience of the persons who build them. Because there are many limitations that encounter the building process of the manual thesauri, the computer tools' use to construct them is needed, or its help in building the thesauri as the thesaurus can be constructed automatically. It had added the variations with the user query by itself and there is no interaction of the user query and besides enhancing the retrieval performance of the IR system.

In the other hand, building an automatic thesaurus is more original diverse methods that were taken for constructing this thesaurus such as similarity thesauri [31]. The available methods for automatic query expansion could be considered as global or local. In the local query expansion methods, in order to expand or improve an initial query, these methods adopt and utilize a few numbers from the top-ranked of the retrieved documents [8]. But if a limited number of the selected documents from these top ranked are relevant to the submitted initial query of the user needs, then the performance of the retrieving process will go down to a great deal and this will result a poor recall. Lately, suggested query expansion methods are depends on the technique of the user's relevance feedback [9]. The methods that adopt this technique depends on the analyzing the retrieved documents that are relevant to the user's query, in order to expand the original query to improve the performance of the retrieval process. On the opposite way around the global query expansion methods need some computations and statistics such as find the co-occurrence of data in the collection which needs to have large amounts of computer resources to be able to compute. For instance, the term – clustering method is one of the global methods [24] which create a cluster of words by categorizing them based on their co-occurrence. After that, these clusters can be used in the process of query expansion.

Based on [3], the main goals of a thesaurus are ultimately to:

(a) Provide a standard vocabulary for indexing and searching tasks;
(b) Support users with determination terms for suitable query formulation;
(c) Supply categorized hierarchies that let the expanding and narrowing of the original query request based on the user information needs.

3.1 Automatic Thesaurus Generation

Thesaurus automatically is considered as another option to solve the problem of the cost of creating a manual thesaurus;

it is created by analyzing the documents in the corpus semantically. In order to build such types of automatic thesaurus, two main methodologies are found: shallow grammatical and word co-concurrence. The first methodology is accomplished by utilizing an analysis of shallow grammatical on the text to benefit from the grammatically relationships between the words [12]. While in the words co-occurring, a document may be having alike or related meaning and count statistics of words to find the most alike words [12]. The easiest way to compute and then build a co-occurrence thesaurus is done based on term-term similarities. As a summary, to generate the thesaurus automatically the collection should be analyzed by using one of the following approaches [12]:

1. Word co-occurrences (words co-occurring many times are more likely to be classified to the same query result)
2. a shallow grammatical analyzes to find the dependencies between words

Despite that thesauri are mainly implemented in commercial and experimental information retrieval systems. However, experiments have demonstrated harsh effects on effectiveness of the retrieval, and there will be lack of prolific automatic methodologies for constructing thesaurus [46]. The following matters should be taken independently for improving the effectiveness of the retrieval [46]:

- *Construction*: There are two types of thesauri: manual and automatic. One of the goals of this paper is on how to construct a thesaurus automatically.
- *Access*: For a given query, the thesaurus must be used to access in some way to enhance or expand the query.
- *Evaluation*: After a thesaurus is constructed, it is important to know the goodness of it. A manual thesaurus is evaluated in terms of classification coverage, the soundness, and thesaurus item selection. At the other hand, the evaluation of automatic thesauri is commonly done by query expansion to know whether the retrieval performance is enhanced or not.

3.2 Categories of Thesaurus

3.2.1. Global Vs Local Thesaurus

The classes of global thesaurus are constructed as a whole depend on word co-occurrence and their relationship and reliance in the corpus. Both queries and documents are indexed in these classes. Whereas, the local thesaurus [34] is constructed actively for each processed query and uses the retrieved documents responding to a given query to amend only this query. Despite that the global analysis methodologies are relatively strong; yet, the statistical analysis that be accomplished along the collection needs a big amount of computing resources. In addition, because they focus on the documents' part without taking the query's part and come up with a partial solution to the word mismatch trouble [22].

3.2.2. Manual Vs Automatic thesaurus

The methodologies for constructing thesaurus in general may be classified into four classes: General purpose Hand crafted manually built thesaurus, Co-occurrence based automatically built thesaurus, Similarity based automatically built thesaurus, Head Modifier based automatically built thesaurus [22]. Building this kind of thesaurus manually is very dense work and there is lack in a specific domain, because it is mainly involves the general words. The synonymous relationships between words are represented in the Handcrafted thesaurus [42]. Voorhees [41] expands a query by utilizing a wordent, a manually builds grid of lexical dependencies and discover that expansion aids mainly when the queries are very short, hence a constructing of automatic thesaurus are not useful.

4. QUERY EXPANSION

Query expansion (QE) is a technique of adding a new list of terms to the initial user query in the context of an IR to enhance the performance of the system [2]. In this case, some of the retrieved documents to the user query may does not contain the words as they are formulated by the user, but words either variously formulated or having a very similar meaning. The new list of added terms created by a QE technique would let the system to consider these relevant documents.

4.1 Sources of Search Terms

Arabic search engines mainly return many irrelevant documents especially when an initial user's queries are not accurate enough. Based on study by Spink [39], there are five origins to select or append terms in the query expansion or reformulation, they are: term relevance feedback, user interaction intermediary, the question statement, thesaurus, and the human intermediary [22]. However, the three main sources to select the related words varying in their level of specificity [18]:

A) *Query specific*: Identifying new terms by locating words in a subset of documents retrieved by a given query. The user can locate these words manually or by using automatic local analysis based on the top *K* of the retrieved documents. The major problem of this technique is the time consuming after the initial user's query is submitted, which can be a significant problem for interactive systems [18].

B) *Corpus specific*: Uses the whole documents within a corpus/collection to define the relationships between words to formulate a global thesaurus. This technique is also called automatic global analysis [34]. In such type of analysis, statistics are required overall the collection, some types of these statistics is the terms co-occurrences (i.e. co-occurrences between every couple of terms will be found and computed). As a result of this computing process is matrix of similarity will be created which forms global association thesauri [14]. Via the terms that are the mostly similar to query terms, the query will be expanded and improved.

C) *Language specific*: Using a general dictionary to find thesauri that is independent from any collection. Because of the vagueness, this type of thesaurus is difficult to implement because it includes several meanings for most words [18].

4.2 Query Expansion Techniques

Researchers such in [5] categorized Query expansion techniques into relevance feedback, and automatic query expansion.

A) *Relevance feedback*: it is a mechanism of revising a search process by using knowledge obtained from a preparatory search in a best or optimal search [6]. One approach of this mechanism is the manual query expansion, which demands the user's intervention. Relevance feedback is mainly executed using variants of the Rocchio algorithm [35]. To eliminate or reduce the need for manual user feedback, some systems simply automatically assume that the top N retrieved documents are relevant. This is motivated by the assumption that the top results are more relevant than a random subset and any important co-occurrence patterns found within this set are

more likely to be relevant to the query. This approach is called pseudo-relevance feedback or blind feedback [43].

B) *Automatic query expansion*: To improve the users' queries, they would not always need their intervention. Automatic query expansion automatically extracts of assumed to be salutary terms from the documents and improving user queries by adding these terms. This kind of query expansion technique somehow is different from an alternative technique called an interactive query expansion (IQE). IQE enables the users to interfere and identify suitable terms from the list of terms extracted automatically from the documents by systems' espousing IQE [36].

From another perspective (which is Thesaurus approaches), many researchers mainly classify existing query expansion approaches into two main categories:

A. *Global analysis*: it involves a constructing of thesaurus in order to help and aid the users to reformulate and enhance their initial queries. Moreover, after the relationships via all documents in the collection are analyzed and the co-occurrences statistics of the terms are found, then a thesaurus could be constructed automatically [14].

B. *Local analysis*: The initial subset of the retrieved documents is used to extract the terms and discover the related terms in order to expand the submitted query.

5. THE DEVELOPED COMPREHENSIVE SYSTEM

In general, users retrieve documents through IR systems according to theory submitted queries. However, the query terms that submitted by the users generally do not provide good enough information to retrieve the most relevant documents. For this reason the query expansion method has been devised [24], it has been the main method for enhancing the performance of IR systems.

This paper implements a special Arabic thesaurus to expand the user's query which improves the performance of IR system. When the user submits his/her query terms, the system chooses the term which has the highest IDF value among the query terms as "the center of the query expansion terms" and chooses the terms having higher degrees of relationship with respect to the center of the query expansion terms as query expansion terms. Then, the system calculates the degree of relationship and the weight (which measure the relatively importance) of each expansion term. Finally, it expands the original query terms by expansion terms.

As mentioned before in this paper, many users find it difficult to formulate queries that are well designed for retrieval purposes even the Thesaurus has been implemented. Yet, most users often need to reformulate their queries to obtain the results of their interest. Thus, the first query formulation should be treated as an initial attempt to retrieve relevant information. Documents initially retrieved could be analyzed for relevance and used to improve initial query. The process of query modification is commonly referred as they are known as relevance feedback methods:

This paper is focused on adopting a term-term similarity technique to construct an automatic Arabic thesaurus which can be implemented in a specific application to enhance the expansion process and to get more relevance documents for the user's query. Moreover, the developed system tries to implement the most of query expansion techniques via six phases as comprehensive methodology to try to optimize the query expansion technique and then to optimize the performance of the IR, as follows: Traditional IR processes, Pseudo Relevance Feedback, Query-Specific Clustering, Statistical Thesauri, Relevance Feedback, and finally Interactive Query Expansion. All phases enhanced the performance of the IR system and each phase takes the benefits from the previous phase and added a significant enhancement of the system. However, the experiments concluded that the Query-Specific Clustering enhanced the performance factors (precision and recall) more than a conventional information retrieval system and the other enhancement techniques.

Phase 1: Traditional IR processes

After the user Builds Database using the developed system, it creates the following database schema:

Stems(TermID, Term, ni, idf)

Variations(Variation, Document, TermID)

Then, the following steps will be executed:

1. For each document in the collection, apply the following:
   a. Apply stemming on words using light stemmer.
   b. Remove the stop words from the documents. The system uses Arabic stop word list contains 1459 words.
   c. Insert all the stems into Stems-table.
2. For every stem which has a variation(s), insert these terms with their stemmed variations and some information (such as: their positions and in which document(s) they found) into Variations-table.
3. 60 standard queries have been submitted to the system.
4. The system removes the stop words from each query.
5. The system returns the stems of the terms in the submitted query.
6. The system will match that stems with the table (Stems-table).
7. When stem is found, then the program will return the documents that include its variations.

Phase 2: Pseudo Relevance Feedback

1. Submit a query in Arabic (from the standard 60 queries).
2. An elementary group of documents will be retrieved.
3. The top 20 ordered of the returned documents will be marked as non relevant or relevant.
4. Then, the system will apply a relevance feedback as will describe in phase 5.
5. Based on the actions in step 3 and 4, a modified group of documents will be retrieved.

Phase 3: Local Query Expansion (Query-specific clustering)

For each query do the following steps:

1. Define matrix F as follows:
   • A column for each document in D (m columns).
   • A row for each word in K (l rows).
   • $f_{i,j}$ = the frequency of word k_i in document d_j.
2. Define matrix C as the multiplication of F by its transposition F^t:

$$C = F \cdot F^t$$

$$c_{l,\,j} = \sum_{k=1}^{m} f_{l,\,k} \cdot f_{j,\,k}$$

- $c_{i,j}$ quantifies the frequency of co-occurrence of terms K_i and K_j in the answer documents.
- Normalization: Replace $c_{i,j}$ with

3. Define the p closest neighbors of a word k_i as follows:

• Consider row i of C.

- Pick the p columns of C with the highest values in this row.
- The corresponding terms are the p closest neighbors of k_i.
- Hence, the neighbors of k_i are the terms with the highest co-occurrence with k_i in the answer set.

Phase 4: Global Query Expansion or Global Clustering (statistical thesauri)

1. Construct the document-term matrix.
   - This familiar matrix is populated with weights $w_{i,j}$ each denoting the weight of term k_i in document d_j. The formula of $w_{i,j}$ is a full weighting schema as in step 3 Phase 5.
2. Convert the term-document matrix to a term-term similarity matrix, using a term similarity measure.
   - The construction in this process is similar to the matrix for query specific clustering.
   - It begins with a term-document matrix with weights (rather than frequencies).
   - The outcome is a matrix that expresses term-term similarity:

$$Si,\ j = \sum_{k=1}^{t} Wi,k \cdot Wj,k$$

	k_1			k_j			k_t
k_1	$s_{1,1}$	…	…	$s_{1,j}$	…	…	$s_{1,t}$
	.			.			.
	.			.			.
k_i	$s_{i,1}$	…	…	$s_{i,j}$	…	…	$s_{i,t}$
	.			.			.
	.			.			.
k_t	$s_{t,1}$	…	…	$s_{n,j}$	…	…	$s_{t,t}$

3. Choose a *threshold* that determines if two terms are similar enough to be in the same cluster. The value 10 is chosen as a threshold value. This data is stored in a new binary term-term similarity matrix.
4. Assign the terms to clusters (Using the Cliques clustering method)
5. Adding the expansion terms to each query depending on the synonyms provided by the constructed thesaurus.

Phase 5: Relevance Feedback

1. Find the frequency for each term in the documents collection.
2. Calculate the *idf* for each stem in all documents collection based on the following formula. Then, save these *idf* values into the Stems-table.
   - ***(idf$_i$) = log$_{10}$ (N/ n$_i$), where***
     Where: N= number of documents in the collection and n_i = number of document that contain the term *i*. (All these values stored in the database, calculated once)
3. When TF and IDF are joint into TF*IDF, the result is high weights for words that occur with medium frequency in every document and with low frequency in the collection. Thus, calculate the full weight schemes for each of stem using the following formula:
   - ***wij = tf$_{ij}$ * idf$_i$ = tf$_{ij}$ * log$_{10}$ (N/ n$_i$)***
     where, Term Frequency (tf): Term Frequency is the number of times a term *i* appears in document *j* (tf_{ij})

Then, calculate the similarities between the query and the document in the documents collection. This correlation is calculated by the cosine of the angle between the two vectors (query vector and document vector) according to the formula:

$$\vec{q} = (w_{1,q}, w_{2,q}, ..., w_{t,q}), \quad \vec{d}_j = (w_{1,j}, w_{2,j}, ..., w_{t,j})$$

$$sim(d_j, q) = \frac{\vec{d}_j \bullet \vec{q}}{|\vec{d}_j| \times |\vec{q}|} = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}}$$

4. After these calculations of the similarities, then the top ranked documents will be retrieved to the user as the relevant documents.
5. User will examine the most relevant retrieved documents to his search; he will consider only the top **N** documents (**N** is determined by the user to decide which documents are the most relevant).
6. But, if the user does not find all what he is searching for, then by click a check box beside each retrieved relevant document to his search (The non-relevant documents remain unchecked).
7. Then the system applies standard rochio to modify initial query and reweighing the query terms.

- Rocchio [35] builds the new query Q' from the original query Q by:

Where $$Qm' = \alpha Q + \beta \frac{1}{n_1} \sum_{i=1}^{n_1} Dr_i - \delta \frac{1}{n_2} \sum_{i=1}^{n_2} Dn_i$$

Qm' : the new Query(applied for each term in the old query).
Q: the old query.
α : constant = 1, β : constant =0.7, δ : constant= 0.4
Dr_i: Relevant Set from the retrieved Set. (User documents selection).
Dn_i: Non- Relevant Set from the retrieved Set. (User documents not select them).

– The relevant document vectors are added to the initial query vector.
– The non-relevant document vectors are subtracted.

After Modifying the query terms and have new weights for Qm' , The system rebuilds the vector space model and re-execute the above steps for vector space model. So, the similarities of most relevant documents increased.

Phase 6: Interactive Query Expansion

1. After applied the previous 5 phases, then the number of documents will be reduced (by exclude/subtract the documents that do not include anyone of stems or their variations).
2. The system will display all the possible variations of the query stems to the user (As a list of checkboxes). Then he can to select/deselect the words from the list that he believes they are relevant to his information needs. This step enables the user to make an interactive query expansion.
3. Then, the system logically will create a unique key that will be used later as a stem for the selected words.
4. The system excludes/subtract the deselected words from Stems-table that the used excludes. So the variations of a stem remained in the selection variations field.
5. The system will put the new stem and its variations in a Stems-table (the variations were selected). The user may be involved in submitting the system's feedback to achieve quite precise thesaurus with very good precision ratio. Then the user will be able to intervene in selecting the similar and related terms to broaden the query in order to improve it. Consequently, the degree of the relevancy between the initial word the new word will be increased and lessen the relevancy the word he uses and those that are not needed.

6. EXPERIMENTS

The performance will be measured based on precision and recall, after building and comparing the results by using a collection of 242 Arabic abstracts and by building a collection of 60 Arabic queries as a standard for these documents. All these abstracts involve computer science and information system. Experts found the relevant document for each query manually. The system implemented in C#.NET language. Table 1 shows the details of the used Arabic collection.

Table1. Collection statistics

Number of documents	242
Number of queries used	60
Average words per query	2.5
Average number of relevant documents per query	43.0

Note: Relevant documents for these were found manually by Experts users.

In order to evaluate any system, the standard measurements must be applied to find the efficiency of its performance, and the relevancy of the search outcome, which is basically, measured using the Precision-Recall metric. Recall is the fraction of the relevant documents which has been retrieved, i.e.

$$\text{Recall} = \frac{\text{Number of Retrieval and Relevant Documents}}{\text{Number of Total Retrieval Documents}}$$

Precision is the fraction of the retrieved document, i.e.

$$\text{Precision} = \frac{\text{Number of Retrieval and Relevant Documents}}{\text{Number of Total Retrieval Documents}}$$

The comparison of all experiments will be based totally on average recall, precision. Results of the experiments demonstrated that the performance of the system (the recall and precision ratios) is enhanced when the suggested accumulative stages were implemented incrementally. In other words, when the third phase is applied, then it applies the steps of the first and the second phases; after that it applies the steps of phase3… and so on.

Table 1 shows a comparative for all experimental results of all phases, which shows that best results come from applying the six accumulative phases.

Table 1. Average precision for all phases at 11-recall levels

	Average that resulted after apply each phase					
Recall Level	Phase1	Phase2	Phase3	Phase4	Phase5	Phase6
0	0.44	0.55	0.74	0.82	0.9	1
0.1	0.4	0.46	0.71	0.78	0.88	1
0.2	0.36	0.44	0.64	0.72	0.86	0.9
0.3	0.33	0.4	0.6	0.68	0.81	0.88
0.4	0.3	0.36	0.57	0.64	0.76	0.82
0.5	0.24	0.32	0.5	0.6	0.7	0.77
0.6	0.21	0.24	0.44	0.58	0.66	0.72
0.7	0.12	0.18	0.39	0.46	0.6	0.7
0.8	0.09	0.1	0.25	0.38	0.57	0.68
0.9	0.04	0.08	0.21	0.31	0.5	0.66
1	0.03	0.05	0.18	0.27	0.47	0.61

Figure 1 shows 11-point interpolated precision using the proposed six phases. A local method gives better results than global methods. However, as a general common sense each phase added a value in terms of enhancing the average of the precision and recall ratios. It can be noticeable that the phase 6 which implements all the previous phases from 1 to 5 and after that it applies its steps as Interactive Query Expansion technique. It gives the most optimal precision/ recall results.

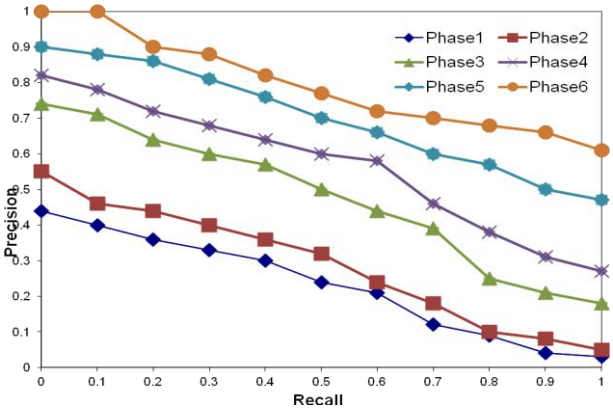

Fig 1: Average Precision and Recall of 60 queries

Another behaviour that can be seen in the figure above is, by applying phase three (local query expansion), an improvement of the average of precision increases more than the first two phases. However, the best average of precision resulted from applying all phases from phase 1 to 6. In other words, this figure is another representation of presented results in table 1.

7. CONCLUSION

Query expansion techniques are mostly used in increasing the recall ratio. However, a precision may be significantly decreased by a query expansion technique, because it could contain vague terms. Thus, multiple types of thesauri were proposed for query expansion in IR and combining these thesauruses with pseudo-relevance feedback method. The fundamental concept underlying the suggested methods is that every kind of thesaurus has its own distinct features and integrating them which brings a precious resource for query expansion. To avoid some troubles of expansion terms, the techniques of designing a weighting term should be used.

This paper presented a new comprehensive approach for automatic thesaurus construction and query expansion techniques for document retrieval. The proposed thesaurus construction method and the proposed implementation query expansion techniques can improve the performance of IR systems for dealing with document retrieval.

8. REFERENCES

[1] Abdelali, A., "Localization in Modern Standard Arabic," Journal of the American Society for Information Science and technology Volume 55, Number 1, 2004.

[2] Abouenour L., Karim B., and Paolo R., "An evaluated semantic query expansion and structure-based approach for enhancing Arabic question/answering," International Journal on Information and Communication Technologies, Vol. 3, No. 3, 2010.

[3] Alekcandov V.V , Kuleshov S.V , Shannaq B. , "Phenomenon of identification‖ information-measuring and operating systems Journal, http://www.radiotec.ru/catalog.php?cat=jr.(2010), 2010.

[4] Al-Fedaghi and Al-Anzi, "A new algorithm to generate Arabic root-pattern forms," In Proceedings of the 11th

National Computer Conference, King Fahd University of Petroleum & Minerals, pages 04–07, 1989.

[5] Al-Kabi M., Kanaan G., Al-Shalabi R., Noor Z., Zaher S. and Maen H., "Stem-Based Query Expansion for Arabic Corpus," ABHATH AL-YARMOUK: "Basic Sci. & Eng." Vol. 18, No. 2, pp. 227- 246, 2009.

[6] Billerbeck B. and Zobel J., "Questioning Query Expansion: An Examination of Behavior and Parameters," In Schewe & Williams, H.E. (Ed.), Proceedings of the Australasian Database Conference, 15, Melbourne, Australia: RMIT University 69- 76, 2004.

[7] Boumedyen S., Kuleshov S., "Super Arabic morphological analyzer (SAMA1)," Information-Measuring and Operating Systems Journal, 2009.

[8] Buckley, C., Salton, G., Alan, J. and Singhal, A., "Automatic query expansion using SMART," Proceedings of the 3rd Text Retrieval Conference (TREC-3), pp.69-80, 1995.

[9] Chang Y. , Chen S. and Liau C., "A new query expansion method based on fuzzy rules," Proceedings of the 2003 Joint Conference on AI, Fuzzy System, and Grey System, 2003.

[10] Chekayri A., "La Structure Des Racines en Arabe," Ph.D. dissertation, University Paris VIII, 1999.

[11] Chen H. et al, "Automatic Thesaurus Generation for an Electronic Community System," Journal of American Society for Information Science, 46(3), 175–193, 1995.

[12] Christopher D. , Prabhakar R. and Schütze H., "Introduction to Information Retrieval," Cambridge University Press. 2008.

[13] Cui H., et al., "Query Expansion by Mining User Logs," IEEE Transaction on Knowledge and Data Engineering 15:829-839, 2003.

[14] Cui H., Wen J., Nie J. and Ma W., "Query Expansion for Short Queries by Mining User Logs," IEEE Trans. Knowl. Data Eng., 15(4), 829- 839, 2002.

[15] Daniel M., Olivier N., "Building Virtual Communities for Information Retrieval," In proceedings CRIWG 2003, 371-379, 2003.

[16] Farag A., and Andreas N., "Corpora based Approach for Arabic/English Word Translation Disambiguation," Speech and Language Technology, Volume 11, 2009.

[17] Fellbaum C., "An Electronic Lexical Database," MIT Press, Cambridge, MA, 1998.

[18] Gauch S., Wang J. and Rachakonda S., "A Corpus Analysis Approach for Automatic Query Expansion and its Extension to Multiple Databases," ACM Transactions on Information Systems (TOIS), 17(3), 250- 269, 1999.

[19] Grootjen F., Th.P. van der Weide, "Conceptual query expansion," Data & Knowledge Engineering, Volume 56, Issue 2, Pages 174–193, 2006.

[20] Hammo B, Sleit A., El-Haj M., "Effectiveness of Query Expansion in searching the Holy Quran," Proceeding of the Second International Conference on Arabic Language Processing, 1-10, 2007.

[21] Hang C., Ji-Rong W., Jian-Yun N., Wei-Ying M., "Probabilistic Query Expansion Using Query Logs," 2002.

[22] Hazra I. and Aditi S., "Thesaurus and Query Expansion," International Journal of Computer science & Information Technology (IJCSIT), Vol 1, No 2, 2009.

[23] Jansen B. and Spink A., "Methodological approach in discovering user search patterns through web log analysis," Bulletin of the American Society for Information Science and Technology, Vol. 27, pp. 15-17, , 2000.

[24] Jones K., "Automatic Keyword Classification for Information Retrieval," Butterworths, London, UK, 1971.

[25] Kanaan G., and Wedyan M., "Constructing an Automatic Thesaurus to Enhance Arabic Information Retrieval System," The 2nd Jordanian International Conference on Computer Science and Engineering, JICCSE 2006, 89-97, 2006.

[26] Khafajeh H., Yousef N., Kanaan G., "Automatic Query Expansion for Arabic Text Retrieval Based on Association and Similarity Thesaurus," In: Proceedings of EMCIS, 2010.

[27] Kristensen J., "Expanding End-Users Query Statements for Free Text Searching with a Search-Aid Thesaurus," Information Processing and Management, 29(6), 733–744, 1993.

[28] Li W., and Agrawal D., "Supporting web query expansion efficiently using multigranularity indexing and query processing," Data and Knowledge Engineering, Vol. 35, pp. 239-257, 2000.

[29] M. Magennis and C. J. van Rijsbergen, "The potential and actual effectiveness of interactive query expansion," in Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in information Retrieval, pp. 324-332, 1997.

[30] Nwesri A., "Effective Retrieval Techniques for Arabic Text," Doctor of Philosophy thesis, RMIT University, 2008.

[31] Qiu Y. and Frei H., "Concept based query expansion," Proceedings of the 16th Annual International ACM Conference on Research and Development in Information Retrieval, NY, pp.160-169, 1993.

[32] Qiu Y. and Frei H., "Improving the retrieval effectiveness by a similarity thesaurus," Technical Report No. 225, Dept. Computer Science, Swiss Federal Institute of Technology (ETH), 1995.

[33] Rachidi T. et al, "Arabic user search Query correction and expansion," In Proc. of COPSTIC'03, Rabat, 2003.

[34] Ricardo Baeza-Yates, "Modern Information retrieval," Addison Wesley, 1999.

[35] Rocchio J., "Relevance feedback in information retrieval," In The SMART Retrieval System, G. Salton Ed., Prentice-Hall, Englewood Cliffs, NJ, 313–323, 1971.

[36] Ruthven I., and Lalmas M., "A Survey on the Use of Relevance Feedback for Information Access Systems," Knowledge Engineering Review, 18, 95– 145, 2003.

[37] Schütze H. and Pedersen J., "A cooccurrence-based thesaurus and two applications to Information Retrieval," Information Processing and Management 33(3): 307-318, 1997.

[38] Smeaton, A. and Berrut, C., "Thresholding Postings Lists,Query Expansion by Word- Word Distances and POS Tagging of Spanish Text," In Proceedings of The Fourth Text Retrieval Conference, 1996.

[39] Spink A. and Saracevic T., "Interaction in information retrieval : selection and effectiveness of search terms," Journal of the American Society for Information Science 48 ,No 8,p.741-761,1997.

[40] Stairmand M., "Textual Context Analysis for Information Retrieval," In Proceedings of the 20th ACM-SIGIR Conference, pp. 140–147, 1997.

[41] Voorhees E., "Query expansion using lexical-semantic relations," In Proceedings of the 17th ACM-SIGIR Conference, pp. 61-69, 1994.

[42] Wang Y., Vandendorpe J., and Evens M., "Relational thesauri in information retrieval," Journal of the American Society for Information Science, 36(1):15-27, 1985.

[43] Xu J. and Croft W., "Query expansion using local and global document analysis," In 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 4–11, 1996.

[44] Xu. J and Croft W., "Improving the effectiveness of information retrieval with local context analysis," ACM Trans. Inf. Syst. 18(1):79-112, 2000.

[45] Yuen- H. et al, "Global and local term expansion for text retrieval," Proceedings of NTCIR-4, Tokyo, 2003.

[46] Yufeng J., Croft W., "An Association Thesaurus for Information Retrieval," In RIAO 94 Conference Proceedings, 1994.

[47] Z. Moukdad Hidar, "Stemming and root-based approaches to the retrieval of Arabic documents on the Web," Webology, 3(1), Article 22, 2006