QEA: A New Systematic and Comprehensive Classification of Query Expansion Approaches

Fatemeh Serpush[a*], Mohammadreza Keyvanpour[b]

[a] *Faculty of Computer and Information Technology Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran*

[b] *Department of Computer Engineering Alzahra University, Tehran, Iran*

Abstract

A major problem in information retrieval is the difficulty to define the information needs of user and on the other hand, when user offers your query there is a vast amount of information to retrieval. Different methods , therefore, have been suggested for query expansion which concerned with reconfiguring of query by increasing efficiency and improving the criterion accuracy in the information retrieval system. Accordingly in this paper, in addition to propose a new coherent categorization for approaches, we proceed to detailed identify them, and proper functional criteria to evaluate each of these approaches are suggested.

Keywords: information retrieval; query expansion; knowledge models; relevant feedback

1. Introduction

There is a vast amount of information to retrieval, but information which is user interest, should be retrieved. One way to increase efficiency is query expansion. Query expansion is a technique used to boost performance of document retrieval engine by expanding and reconfiguring the user's query. In "Fig. 1", are visible documents retrieve based on query expansion. The query expansion is performed in order to provide the possibility of making the user query unambiguous. The users can choose to refine their search terms. This step will add semantically related and additional contextual information to the query. Accordingly, a recent query expansion attracted the attention of many researchers in this field and several approaches have been proposed. But some of these approaches despite its high influence on retrieval are faced to challenges which include semantic constraints [1] and longer query [2] and ambiguity [3]. In this paper, in addition to proposing a new coherent categorization for these approaches, we proceed to identify and introduce these approaches and their challenges, advantages and proper functional criteria to analyze and evaluate them are suggested which can lead to a more precise understanding of them and the accurate and systematic use of them based on need. The rest of the paper is organized as

* Corresponding author. Email: f.serpush@yahoo.com

follows: section 2, In addition to providing a review of related works, introduces several proposed definitions for query expansion.

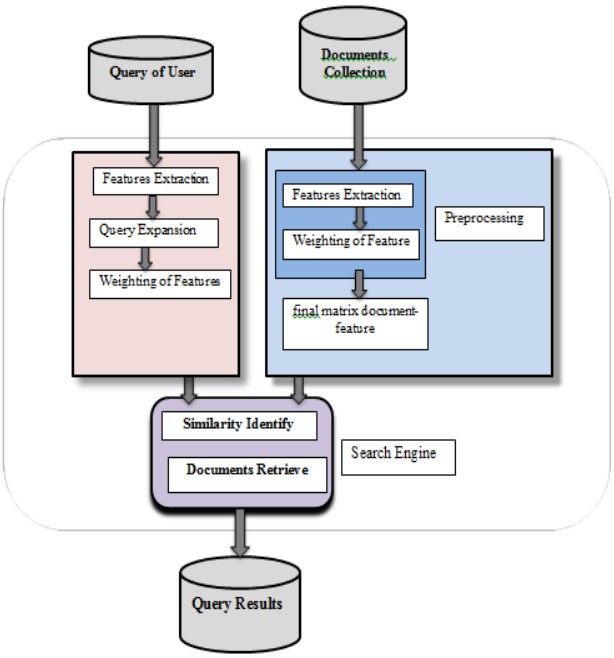

Fig. 1. process of documents retrieve based on query expansion

Section 3 Then in addition to providing a coherent classification on query expansion approaches in terms of performance, to introduce, identification and detailed analysis of each approach are discussed. Section 4, offers five criteria to study and analyze different QEA which is presented in section 3 and based on the proposed criteria, mentioned approaches are evaluated. Section 5 presents the results of this new classification for QEA.

2. Related Work

Retrieval means using the available evidence to predict the degree to which a document is relevant or useful for a given user need but been retrieve some unrelated documents. A supporter theory for query expansion is to pay attention to issues of synonymies, multiply and multiplicity of meanings, this process may provide better results for increasing recall and precision [4] [10]. Contradiction and incompatibility between user's query and documents have a highly influence on the performance of today's search engines. A number of researches have focused on query expansion and tried to solve the ambiguity problem of short queries [7], [5], [8]. Query expansion is a process that tries to offer more relevant documents to the user which does not necessarily contain the same words in the original query [6] which divided a work on automatic query expansion into two classes: global analysis and local analysis [7]. Some researchers were identified two general strategies for query expansion: ontology-based and semantic-based, and in the [9] divided two categories based on the structure of knowledge and based on search results or in the [4] divided two categories based on collection and search results but some methods have not been studied and is generally Expanding a query with synonyms or hyponyms is one example of an ontology-based query expansion. Semantic-based methods focus on the collection of documents [23]. Actions of query expansion can be based on various ideas. The point is how to determine the correlation between each pair of keywords. Query expansion has many applications, Including retrieval system based on learning , geographical information retrieval, personal queries, knowledge organization, the responsiveness systems to questions, the related systems to medical field, patent retrieval, image search, distributed information retrieval, cross-language retrieval and disambiguates, the organization of electronic information and also expert searches [9]. In this paper, query expansion approaches were reviewed, classified and evaluated.

3. Proposed Classification and analysis of query Expansion Approach (QEA)

To match the user's information need and documents, several techniques have been proposed for query expansion. The general rule is using meanings

of the words and similar or related phrases, to be appeared in the original query.

User of retrieval systems that use word matching as basis for retrieval are faced with the challenge of phrasing their queries in the vocabularies of the documents they wish to retrieval. Hence, the issue of query expansion is proposed which efficiently improves of web information retrieval [4], [13], [21]. Since there are different approaches for query expansion in information retrieval, providing a general categorization which examines each approach according to their key features, seems necessary. In this section, as shown in "Fig. 2" query expansion approaches is based on knowledge structure or search results that we discuss these approaches. Then in the "Table I", these methods analysis.

A. QEA Based on Search Results

In this way, initially the user submits the query, search engine retrieves a number of documents according to initial query then documents will be analyzed and related words with entered query will be added to the query and retrieved and reload again [20]. So in methods based on the search results, analysis is performed on limited volume of documents for expansion of user query. Detailed study of the technical content of this method shows that there are three general approaches for query expansion based on query results that these approaches are introduced separately [4].

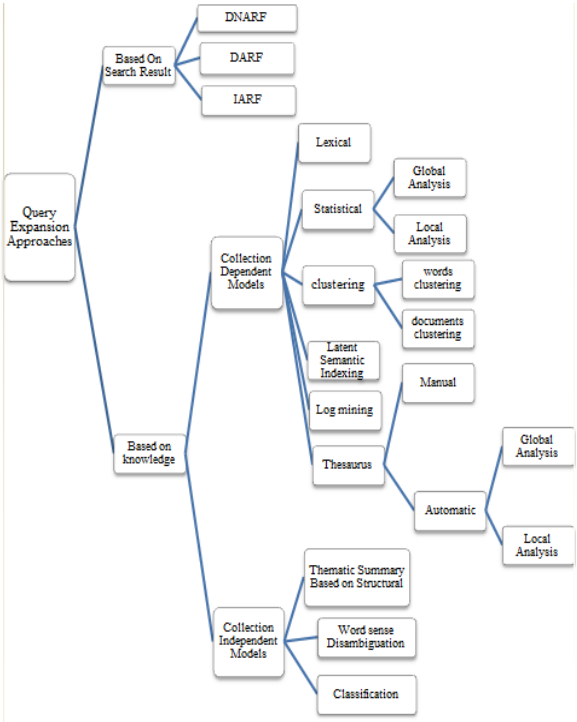

Fig. 2. proposed categorization for query expansion methods

1) Direct non-automated relevant feedback (DNARF)

This method is to involve the user in the retrieval process so as to improve the final result set. In particular, the user gives feedback on the relevance of documents in an initial set of results. According to "Fig.3", in this approach, initially user enter your query, search engine based on it retrieves pages and then the user from retrieved pages chooses a number, the search engine based on user feedback, again retrieves pages and provides to the user [29].

DNARF idea is to improve the final result set, the user is involved in the retrieval process, thus the user gives feedback on the documents relevance with his/her need in an initial collection of results [4], [14], [34].

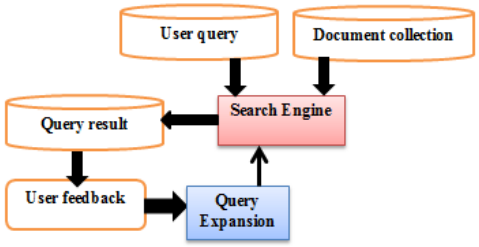

Fig. 3. query expansion based on user feedback

Challenges that this approach faces include spelling mistakes, lack of cross-language information retrieval and mismatched dictionary of searcher against corpus dictionary [2]. In addition to these challenges the use of DNARF leads to a long query and these results in non-efficient of information retrieval systems, so it can be acknowledged that the DNARF, by itself, cannot be retrieved successfully. Rocchio algorithms is a classic algorithm for implementing of this approach, which uses the method of documents displaying , so each document and query are equivalents to a vector that "(1)" represents this algorithm [4], [14], [34].

$$\vec{q}_m = \alpha\vec{q}_0 + \beta\frac{1}{|D_r|}\sum_{\vec{d}_j\in D_r}\vec{d}_j - \gamma\frac{1}{|D_{nr}|}\sum_{\vec{d}_j\in D_{nr}}\vec{d}_j. \quad (1)$$

Dr, collection of relevant documents and retrieved; Dnr, collection of non-relevant documents and retrieved; $\vec{q}_0$ is initial query and $\vec{q}_m$ is query has been developed and α , β , γ are coefficients.

2) Direct automatic relevant feedback (DARF)

One of the problems of DNARF was that users have not willing to provide feedback, in order to DARF become alternative DNARF, in this way acts that primarily on the user's initial query does retrieval operation but the results are not displayed to the user, instead of those documents, K the first document assumes that the relevant document, and after obtaining the center of gravity, the Rocchio algorithm which is explained in(1), and runs on and shows retrieval secondly to the user[2], [34]. In this approach, if the loop is repeated only once, good results can be achieved but with repeated n times, not only does not improve retrieval, instead become more irrelevant with needs of user, because the noise gradually increases [30] and irrelevant documents added to the cycle and Leading to deviate query. Sometimes, once repeat, there will be a bad result [4], [2], [34] that the process is described in "Fig. 4".

3) Indirect automatic relevant feedback approach (IARF)

This approach automatically acts and its difference with two previous approaches is that in the DNARF, the users knowingly choose a set of documents as relevant documents and in the DARF, the user does not have any involvement in retrieval and the system internally would simulation procedures DNARF, in this approach, users unknowingly give feedback that is, the user history of previous searches, the system identifies the user's preference [1], [12]. In this approach, a user query, by adding the words that will change and a new query is searched that in "Fig. 5" it can be well observed.

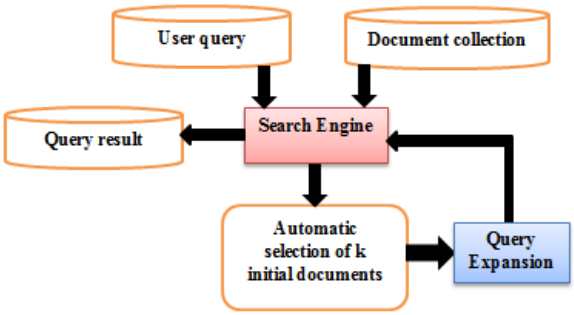

Fig. 4. query expansion based on retrieved k initial document

B. QEA Based on Knowledge

Usually, the query submitted by a novice user is short and ambiguous, and thus many irrelevant results are retrieved. Accordingly, a knowledge base system (KBS) is very suitable for solving this problem. After the expertise of query formulation is transformed into the knowledge base, the inference engine of the KBS can infer other appropriate keywords to expand the original query and get more relevant results.

Furthermore, the knowledge can be reused by others and easily adapted to different scenarios [7]. In "Fig. 6", information retrieval and query expansion based on knowledge structure are shown. Detailed study of the technical content of these methods represents two general models based on the knowledge structure to expand the query: corpus dependent knowledge models and corpus independent knowledge models. At the rest of the section, these two models and their approaches are presented separately.

1) Collection Dependent Models

Users generally require tools to help them sift through large collections of information and retrieve only those items of interest. Query expansion using collection dependent knowledge (CDK) is more suitable for the collection of statistical documents. In web collection, knowledge models will have to be updated and rebuilt constantly, because the web has more dynamic property. This model can be divided into six categories: lexical methods, statistical methods, latent semantic indexing (LSI), clustering, log mining, thesaurus [9], [13], [29].

1.1)The lexical queryapproachbased onCDK

Lexical network combines weight and annotations on typed relations between terms and concepts. Some inference mechanisms are applied to the network to improve its quality and coverage[43] and it is another important source for deriving context. It contains domain specific vocabularies and relationships among them which have been automatically extracted from the collection.

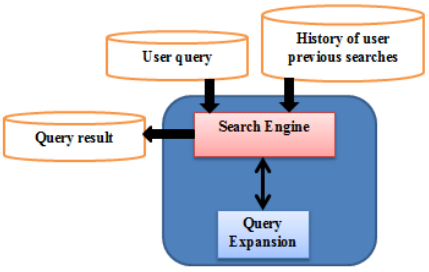

Fig. 5. query expansion based on history

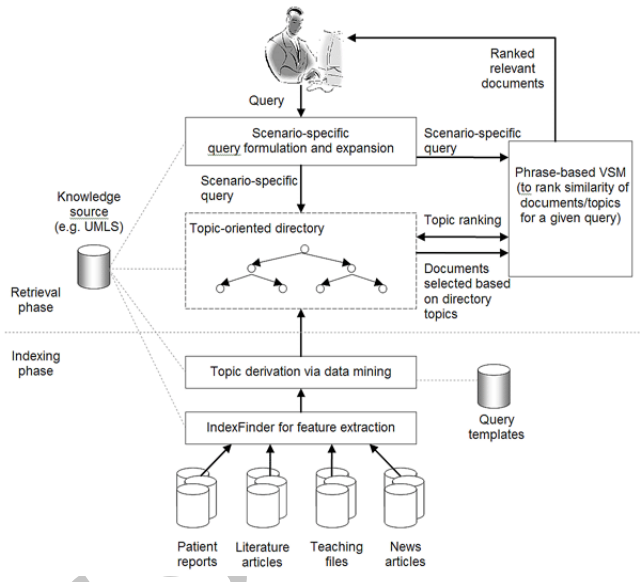

Fig. 6. Retrieval System Based on Knowledge [30]

The general tendency is to disambiguate terms during the search process and not store the disambiguated terms prior to the search [13]. Lexical network is to analyze radiological report in order to extract terms and semantic relations between them [43].

Approaches leverage global language properties, such as synonyms and other linguistic word relationships (e.g. hyponyms). These approaches are typically based on dictionaries or other similar knowledge representation sources such as Word Net. Lexical query expansion approaches can be effective in improving recall but word sense ambiguity can frequently lead to topic drift, where the semantics of the query changes as additional terms are added [29], [22].The type of lexical network where are working with is a graph with lexical items or concepts as nodes connected through arcs interpreted as relations between items. Those relations are semantically typed and represent (typical) lexical or ontological relationship possible between terms (hypernym, synonym, antonym, part of, cause, consequence, typical location, telic role, semantic role and so on)[43], [44], [45].

In the lexical query expansionmethod, first, the query is analyzed in order to extract the lexical affinities(LA) it contains. Let A a LA of the query Q. We can define a quantity of information for A and a similarity between a document and a query using only LAs as done in (2) for lemmas. The similarity using single lemmas (S_{lem}) and the one using LAs (S_{LA}) are combined:

$$S(D,Q) = \beta \cdot S_{lem}(D,Q) + (1-\beta).S_{LA}(D,Q) \qquad (2)$$

where β is a coefficient ($\beta \leq 1$). We have used the empirical value $\beta = 0.7$.

In fact, when the query is expanded with synonyms or stems, an associated word is considered as the lemma itself (equivalence) for the construction of the LAs[25]. Recently word sense was frozen into the lexicon; Therefore researchers discovered that full lexical knowledge comes from the texts themselves proposed the use of a generative lexicon to disambiguate word sense [13], [43]. The expansion procedure used in this work relies heavily on the information recorded in Word Net. Word Net's basic object is a set of strict synonyms. "Fig. 7" shows a piece of Word Net. The figure contains is-A relation between organisms.

1.2)The statistical approach based on CDK

Previous studies show that the expansion of a query with synonyms or hyponyms results in little impact on information retrieval performance [1]. Statistical approaches are data-driven and attempt to discover significant word relationships based on term co-occurrence analysis and feature selection [29]. Related methods include term clustering [17] and Latent Semantic Indexing. In this approach, the words suggested to the user are based on user profile [46], personal information repository (PIR) [11], and user log mining [12]. Statistical methods for query expansion are focused on the document and can be divided into two general categories: local and global analysis [1].

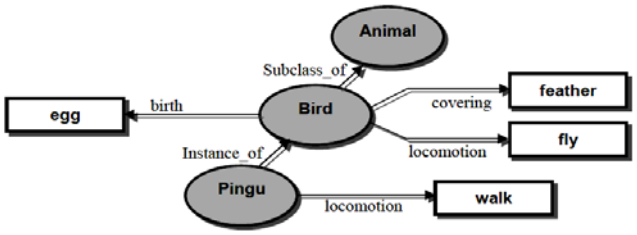

Fig. 7. an example of Relations in WordNet [24]

1.2.1) The statistical approach based on local analysis

This approach assumes that it returns a certain number of classy documents which is related to the initial query because it usually requires less human intervention. However, these methods are not strong because it is impossible for almost all the search engines or explore methods to restore only the relevant documents [14]. Local analysis extracts highly relevant terms from relevant documents retrieved by the original query or from the data mining results. In fact the initial query returns the classy documents, but the words are selected based on the co-occurrence of query terms."Fig.8" depicts the local analysis that uses latent semantic indexing or association rule algorithm to extract the top co-occurrence terms from each original term in top N retrieved documents. Most local analysis methods use the notion of Rocchio's ideal query as a start point. Rocchio's query expansion is a method for detecting the ideal query. The ideal query is the one that has maximal similarity with relevant documents and minimal

similarity with the irrelevant ones. Assuming a vector space retrieval model this query Q is given by the following formula:

$$Q = \frac{1}{|D_r|}\sum_{d_r \in D_r} d_r - \frac{1}{|D_i|}\sum_{d_i \in D_i} d_i \qquad (3)$$

D_r is the set of relevant documents and D_i the set of irrelevant documents. In other words Rocchio's query expansion finds the average term frequency in relevant documents, the average term frequency in irrelevant documents, subtracts the latter from the former and thus calculate a per term weight. In this way terms that appear with high frequencies in relevant documents and low frequencies in irrelevant documents will get higher weight.

1.2.2) The statistical approach based on global analysis

Global analyses pays attention to all documents for the extraction of co-occurrence from related words and include words categorization, latent indexing and similar full thesaurus. One of the major challenges of global analysis methods is thatit needs semantic similarity and disambiguating of these words [1], [13], [9].

1.3)Documents clustering based on CDK

Text clustering is one of the key issues in the area of data mining and information retrieval [39] and was introduced as a means for improving intermittent search performance. Due to known human limitations, it is very difficult for people (even expert) to discover useful information by reading large quantities of unstructured text. This difficulty has inspired the creation of a more specific technique for unsupervised document organization, automatic topic extraction and fast information retrieval or filtering to aid human beings' information discovery.

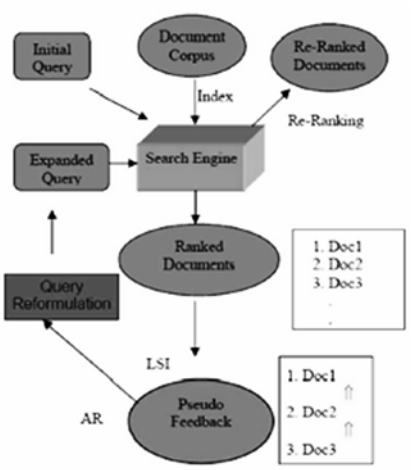

Fig. 8. Local analysis process based on the statistical approach [1], [9]

From one perspective, clustering divided into two categories, documents clustering and clustering of words. Document clustering has been studied in the field of information retrieval for several decades. The purpose of document clustering is to group similar documents into clusters on the basis of their contents. When the query is posed the search engine, considering to which cluster it belongs, suggests other phrases to the user that in "Fig.9" is shown. Similarity of two documents can be obtained using Euclidean distance or cosine similarity which is shown respectively in "(4)" and "(5)".

$$L_1(x, y) = \sum_{i=1}^{m} (x_i - y_i)^2. \qquad (4)$$

$$L_2(x, y) = \sum_{i=1}^{m} |(x_i - y_i). \qquad (5)$$

x, y are vectors of the documents.

Purity is a simple and transparent evaluation measure and compute purity, each cluster is assigned to the class which is most frequent in the cluster, and then the accuracy of this assignment is measured by counting the number of correctly assigned documents and dividing by the number of total documents. As shown in "(6)".

$$Purity(\varnothing, C) = \frac{1}{N}\sum_{k}(\max_{j} | w_k \cap c_j |). \qquad (6)$$

Where $\emptyset = \{w_1, w_2, \ldots, w_k\}$ is the set of clusters and $C=\{c_1, c_2, \ldots, c_{kj}\}$ is the set of classes. And w_k is the set of documents and c_j is the set of documents.

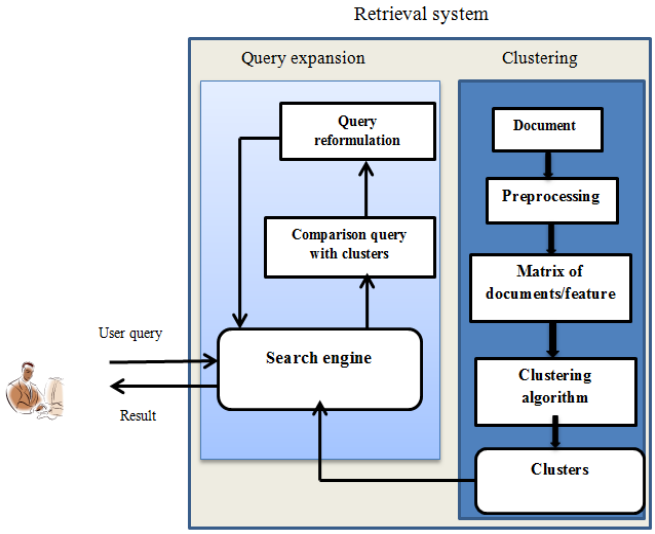

Fig. 9. query expansion and information retrieval system (IRS) based on clustering[9], [4].

Among the challenges that this approach is facing: belonging a term only to one cluster, recognizing the differences between represented document clusters and term clusters at the same time, usefulness of this technique was found to be marginal due to poor clusters resulting from small document collections or insufficient in vocabulary between relevant and irrelevant documents [1], [4], [9], [13], [18].

1.4) Latentsemanticindexing based on CDK

Latent semantic models such as latent semantic analysis are able to map a query to its relevant documents at the semantic level where lexical matching often fails. These models address the problem of language discrepancy between web documents and search queries by grouping different terms that occur in a similar context into the same semantic cluster [47].

Vector space model is one of several methods that determine the similarity between the two documents. This model was developed for information retrieval, and especially is used in biomedical document retrieval. The main advantage of vector space model is ranking efficient and accurate related document according to their similarity to the user query [2]. But main problem of the vector space model is vocabulary mismatch, i.e. it faces with dissimilar words unrelatedly. The second problem is the retrieval of large documents that takes a long time to process and calculate. One of the limitations of the vector space retrieval is that if the query words do not appear in the document then it will not be returned as relevant. [38],to resolve this problem, one of the parameters of this model as latent semantic indexing is used which is , a way to view, review vocabulary and semantics for the retrieval of information and have been widely used in recent years. Latent semantic indexing is a vector space approach for modeling documents in query expansion scope and claims that this technique extracts embedded meanings from a set of documents [11], [18] and discovers conceptual solidarity in the words and uses them in the initial query [20]. The purpose of latent semantic indexing is to overcome the problems of lexical matching by using statistically derived conceptual indices instead of individual words for retrieval. This approach starts with a matrix of terms by documents. This matrix is then evaluated by the singular-value decomposition (SVD) to gain the latent semantic. The SVD decomposes a term-document matrix into three separate matrices, by

which documents and terms are projected into the same dimensional space. For example, the SVD X of $t \times d$ matrix of terms and documents is decomposed into the product of three other matrices and is shown in "(7)":

$$X = T_0 S_0 D_0' . \quad (7)$$

Where T_0 ($t \times m$ matrix) and D_0' ($m \times d$ matrix) have orthogonal, unit-length columns ($T_0'T_0=I$, $D_0'D_0=I$), and S0 ($m \times m$ matrix) is diagonal. T_0 and D_0 are the matrices of left and right singular vectors, and S_0 is the diagonal matrix of singular values [1], [9], [47].

1.5)Log mining

In this method that is called query log mining, system in accordance with the previous users' successful queries which for them a log is made, when a new user makes a search, query will search in the log, if there is successful previous similar query, it'll suggest to the user to use it instead her/his query[29], [27].A large amount of user interaction information is available to search engine users. This information is stored in query logs and can be used to further improve user satisfaction [26]. In this approach performance for both short and long queries is similar, it acknowledges that this approach is an effective way to reduce the difference between short and long queries and also query expansion is influenced by the mismatches between the words of the query and the documents that this approach has overcome this problem. It means that it obtains the relationships between words of query and documents."Fig. 10" extracted query of a large set of loges. Each log includes a query and a set of documents that the user has clicked. If a set of documents for the same queries is selected, then the terms in this document are strongly related to queries words. Thus, the possible relationship between query terms and document terms based on user log is established [4], [31].

1.6)Query expansion based on thesaurus

Thesaurus be can defined from structured system of terms, concepts and metadata that are mainly related to a specialized field, with relationships between them. Thesaurus is a common approach in the area of indexing and retrieval of documents to improve search results [32] and in fact is a set of terms and relations between them, which is widely used in the expansion query by adding synonyms words for query [42] before starting the retrieval process [33], [34]. Thesaurus affiliated with the statistical relationships between words of two similar thesaurus, or relationships between words considering the meaning (synonyms, antonyms) in the thesaurus associated. Query expansion based on thesaurus can be divided into two main categories: manually and automatically. These two approaches are described in the following [4], [32], [33], [35], [42].

1.6.1) Query expansion based on manual thesaurus

Manual thesaurus is lexicon that includes set of synonymous for concepts and by an expert group created and maintained [42]. UMLS is a sample of this thesaurus that was used in the biomedical scope. The problem is that the amount of information is limited Thesaurus Manual and the other hand holding it is costly then came the automatic thesaurus [2], [4].

1.6.2) Query expansion based on automatic thesaurus

Automated methods for thesaurus construction apply high accuracy in determining the relationship between words, such methods possibly is used and lot of literature, can be used for automatic thesaurus construction [35]. Automatically [36] using co-occurrence words, or words related to in terms of grammar can be built [26]. But the problem faced by automatic thesaurus, it is difficult to ambiguity of words in the initial queries, and so the use of phrases, helps to reduce ambiguities of query terms, and lets query expansion technique to extract more relevant terms. One major problem is that there are evaluate

the results of [35]. This approach can be divided into two categories: global analysis and Local analysis [4].

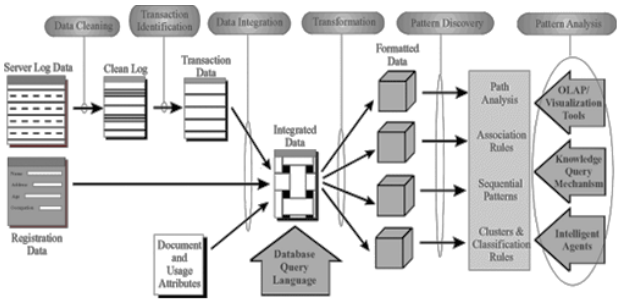

Fig. 10.query expansion based on log mining [31]

1.6.2.1) Global analysis

Global analysis is usually performed on large volumes of documents. It can search all documents needed for the search engine [36], analyses and find its related words (synonyms, similar, related, co-occurrence ...), not necessarily with the same meaning [2]. This method is time consuming and static, i.e. before the user can submit his /her query, he /she should analysis the whole collection, and words related to the identification should be known and the thesaurus should be built on it [37].

1.6.2.2) Local analysis

Local analysis uses only documents of top rank for query expansion [36]. In this case, on a limited set of documents, the thesaurus is constructed. This limited collection of documents is optioned from direct automatic relevant feedback and related words in this document are extracted and then a thesaurus is made that is useful to the query. Then using the Thesaurus It expands the query words and does the original search [2]. Therefore the local analysis is a dynamic state, and needs to compute term correlations, for each query in the runtime [37].

2) Collection Independent Knowledge

This model can be in the form of thesaurus or ontology [42], which is divided into three general categories that we'll describe them. Ontologies provide a structured way of describing knowledge. Ontology is a "shared specification of a conceptualization". The problem with corpus dependent knowledge (CIK) methods is that they are content driven. This can only work if there are sufficient relevant documents to work with and also that these documents contain a reasonable set of terms that represent the subject area for the query. Corpus independent knowledge models do not suffer from this drawback. Philosophically speaking ontology is the "metaphysical study of the nature of being and existence" (WordNet). Practically speaking, ontologies can be seen as special kinds of graphs describing the entities that exist in a domain, their properties and the relations between them. The basic building blocks of ontologies are concepts and relationships.

2.1)Structural thematic summary based on CIK

This method involves thematic analysis of texts, term disambiguation and analyses cohesion relations. Each term in the text is linked to a thematic node. With a conventional thesaurus, humans use their domain, common sense and grammatical knowledge to index documents. One of the methods provided in this topic is query expansion based on geographic features [15]. Topical terms(TTs) represent the subject content of documents. TTs are typically the terms which web searchers use to find relevant sources of informat ion. Terms such as 'globalization' , 'child abuse', 'Skin care', and 'Cosmetic plastic surgery' are examples of topical terms. Lists of subject headings (e.g., Library of Congress Subject Headings) and thesauri cover topical terms[27], and also MeSH is the National Library of Medicine's controlled vocabulary thesaurus. It consists of sets of terms naming descriptors in a hierarchical structure that permits searching at various levels of specificity and it consists a list of words in order to analyze the thematic of America medical information – biological. "Fig.11" shows hierarchical structure of "Diseases" term that these terms can use thematically for expansion "Diseases".

2.2)Word sense disambiguation based on CIK

Word sense disambiguation (WSD) algorithms uses semantic similarity measures is outlined as follows. WSD refers to the process of selecting the correct sense of a word from a set of possible senses or in terms of ontologies to map a term to the correct unique concept. One category of. In their approach they treat the ontology as a graph (network) and use Pagerank to disambiguate senses from that network. The Pagerank algorithm was originally designed to perform link analysis in web pages and the most important pages. The basic idea behind Pagerank is that, if there is link from page A to page B then the author of A is implicitly conferring some impor-tance to page B. More specifically it confers some of its own importance to page B, thus if A is important then B will be also become important but if A is not so important then B would only slightly benefit from the link from A. Thus importance is defined recursively and the algorithm runs in several iterations until convergence. Initially all pages have the same importance but after each iteration importance is concentrated in specific pages. An alternative way to view Pagerank is that it roughly expresses the probability of a random web serfer staying in a specific page[12]. Indexing, text classification, query formulation, multi-lingual information retrieval and concept mapping are done in this method. word sense ambiguity can frequently lead to topic drift, where the semantics of the query changes as additional terms are added[29].

2.3)Classification documents based on CIK

Classification means building a model that can classify a group of objects, so that it is able to predict the classification or missed features from the objects that will encounter them in the future (objects which their classes may not be known). Hence we can say that the classification consists of two steps:

Supervised learning from the training dataset to build a model and data classification based on the model [41]. Often, a class is a more general subject area like China or coffee. Such more general classes are usually referred to as topics, and the classification task is then called text classification, text categorization, topic classification or topic spotting. An example for China appears in "Fig.12". The notion of classification is very general and has many applications within and beyond information retrieval. For instance, in computer vision, a classifier may be used to divide images into classes such as landscape, portrait, and neither. In text classification, we are given a description d ∈X of a document, where X is the document space; and a fixed set of c lasses C = $\{c_1, c_2, \ldots, c_J\}$. Classes are also called categories or labels. Typically, the document space X is some type of high-dimensional space, and the classes are human defined for the needs of an application, as in the examples China and documents that talk about multicore computer chip s above. We are given at raining set D of labeled documents (d,c) ,where (d,c) ∈X ×C [9], [19].

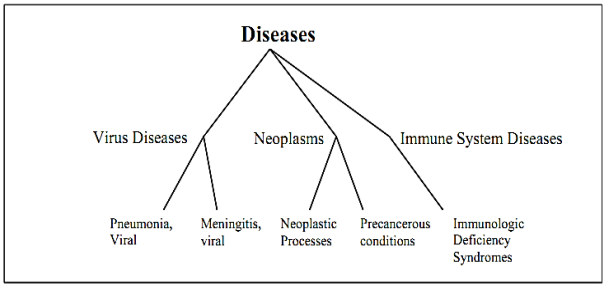

Fig. 11. terms related to "Diseases" term thematically [28]

4. Proposed Criteria And Evaluation Of Knowledge Structure-Based Methods

In this section, we will introduce the five functional criteria to study and evaluation of the various methods for query QEA: user intervention, kind of performance, implementation, precision and recall [4], [9]. Each of these criteria from different perspectives is considered in query expansion methods that results in "Table II" is shown.

A. User intervention

Based on this criterion, different methods of query expansion are performed in three forms: automatic, manual and interactive. In the manual query expansion, the users, according to their skills, decide which words can be used in the original query but in interactive query expansion the system according to the phrases in the collection offers the phrases to the users [24]. In the automatic query expansion unlike the two previous

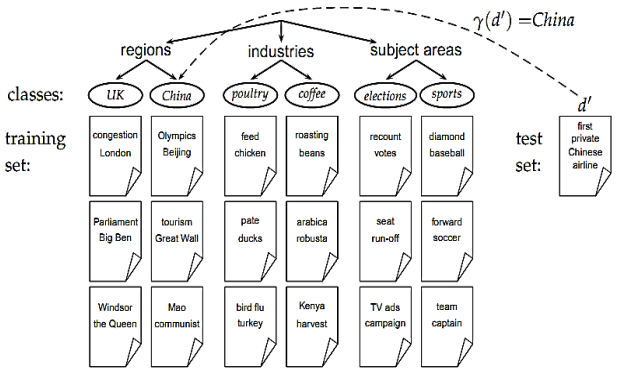

Fig. 12. text classification in information retrieval [19]

B. Kind of performance

According to this criterion, the methods presented for knowledge-based query expansion are either dynamic or static. In the static mode, before the user presents query, the entire pre-defined set, based on the relevant approach, should be analyzed once. But considering that information on the web is constantly changing, this collection should be updated. In the dynamic mode, according to changes on the Web, at any time, any approach returns different results.

C. *Implementations*

Implementation, as the third criterion proposed in this paper addresses practicality and feasibility of various approaches for knowledge-based query expansion.

D. Precision

In fact, this criterion is ratio of the number of relevant documents retrieved to all retrieved documents, which are in the "(8)" is shown:

$$P = \frac{tp}{tp + fp}. \quad (8)$$

tp: the number of documents, correctly, retrieved (related)

fp: the number of documents, mistakenly, retrieved (unrelated)

E. Recall

The fifth proposed criterion for evaluation is recall that is also called the retrieval ability. This criterion is the ratio of the number of related retrieved documents to all related documents, is shown in "(9)":

$$R = \frac{tp}{tp + fn}. \quad (9)$$

fn: the number of documents, mistakenly, not retrieved.

Table 1
Analysis of query expansion Methods

Challenges	Advantages	Main Idea	approach		
• Lack of application, where there are spelling mistakes • Lack of application in cross-language information retrieval • Lack of application, where is mismatch between the searcher vocabulary and set dictionary • prolongation of query • The cost of user time	• Recall increase • Using user manner	To improve the final result set, the user is involved in retrieval process	**DNARF**		**based on search results**
• Query stray • Creating noise	• Lack of user involvement • High performance if once the process is repeated	Automating the DNARFA	**DARF**		
• Query stray	• Using user tastes unconsciously	User unconsciously give feedback , namely identification of user tastes of on the history of user previous searches	**IARF**		
• The need for high volume Explore the queries	• Retrieve of relevant documents • Ambiguity reduce • Performance equally for short and long queries • Overcome the mismatch problem between query words and documents	Search engine suggest to the user words that will complete your query	**Log mining**		**Collection Dependent Models**
• Lack of use ambiguous terms • topic drift	• disambiguate terms • improving recall	contains domain-specific vocabularies and relationships	**Lexical**		
• impossible for almost all the search engines • query expansion based on N intial document	• extracts highly relevant terms • Lack of human intervention • extract the top co-occurrence terms	data-driven and discover significant word relationships based on original query	**Local statistical**	**statistical**	
• need to semantic similarity • needs to disambiguating of terms	• words categorization • latent indexing	pays attention to all documents for the extraction of co-occurrence	**Global statistical**		
• Choose the number of clusters, correctly • Show style • High computational complexity in dealing with large numbers of documents and features • Highly dependent on the definition of similarity measures between data • Lack resolve the simultaneously all the needs	• Groupings by Similarity • Unsupervised Learning	Classification of documents based on relevance (the synonymy, co-occurrence, …)	**Clustering**		

• Weaker of vector space of the runtime performance • Computationally expensive	• Reduce the problem of ranking for less significant dimensions • Extraction of connotation • Efficient	Data display, exploring the use of vocabulary and semantics science for information retrieval	**latent semantic indexing**	
• limited the amount of information • Maintenance is costly	• High quality	Dictionaries that humans makes and it does maintenance	**thesaurus**	
• Precision reduce sometimes • Time consuming and static • Evaluation of results difficult	• Offline • The creation of once • Used in the cross-language retrieval • use	Search Engine with analyzing on the large volumes of documents makes Thesaurus	**Automatic thesaurus by global analyzing**	
• Online	• Dynamically • Used in the cross-language retrieval • High precision	In this method, the search engine makes thesaurus on a limited set of documents.	**Automatic thesaurus by local analyzing**	
• Need to information update • High cost • Lack of public use • Need to expert	• Link term to a thematic node • Use in the Special domain • Use of grammatical knowledge	involves thematic analysis of texts, term disambiguation and analyses cohesion relations	**Thematic summary based on structural**	**Collection Independent Models**
• to topic drift	• text classification • query formulation • multi-lingual • concept mapping • disambiguation	process of selecting the correct sense of a word from a set of possible senses	**Word sense disambiguation**	
• Need to Supervised learning • Need to training dataset	• able to predict the classification • able to predict missed features from the objects	building a model that can classify a group of objects	**Classification**	

5. Conclusion

Concept of the query expansion as one of the most important issues in the field of information retrieval has been introduced. Then, different methods and definitions provided by researchers to query expansion were expressed. Afterwards, based on these definitions and methods, a coherent classification and comprehensive for these approaches was identified and introduced then we explained idea and advantages and disadvantages all approach. Next, to create an appropriate context for studying, evaluating and analyzing each of the approaches, five proper function criteria were proposed. These criteria include user intervention, the kind of performance, implementation, precision and Recall.

Table 2

Current Status, Knowledge-Based Query Expansion Approaches, from the Perspective of the Proposed Criteria

Proposed Criteria								
Recall	**Precision**	**Kind of performance**	**Implementation**	**User Intervention**	**Methods**			
High	Low	Dynamic	Practical (user dissatisfaction)	Interactive	DNARF			**Based on search results**
Low	Low	Dynamic	Practical	Automatic	DARF			
High	Low	Dynamic	Practical	Automatic	IARF			
High	Medium	Dynamic	Practical	Manual	Thematic Summary Based on Structural			**Collection independent models**
					Word Sense Disambiguation			
					Classification			
High	High	static	Practical (but costly)	Automatic	Latent Semantic Indexing			**Collection Dependent Models**
Medium	Low	static	Practical	Automatic	Clustering			
Medium	High	static	Practical (need to disambiguation)	Automatic	Global Analysis		Statistical	
High	High	Dynamic	Practical (but not accuracy)	Automatic	Local Analysis			
High	Medium	Practical (but frozen)	static	Automatic	Lexical			
High	High	Dynamic	Practical	Interactive	Log Mining			
High	High	Simi-Dynamic	Practical (Difficult maintain)	Manual	Manual		Thesaurus	
High	Medium	Dynamic	Practical	Automatic	Global Analysis	Automatic		
Medium	Medium	static	Practical	Automatic	Local Analysis			

Based on these criteria, collection dependent knowledge approaches and collection independent knowledge approaches and search result were evaluated and analyzed and their position were clarified. As a result, providing the possibility of consistent and correct use of the knowledge-based query expansion approaches based on needs, using proposed criteria, as an important achievement of this study is noteworthy. But it should be noticed that the proposed criteria pay less attention to relations between approaches, the operations complexity and the techniques used for every approach, therefore the continuation of this research is possible through definition and application of the criteria which consider mentioned cases.

References

[1] X.Xu, "Cluster-based query expansion using language modeling for biomedical literature retrieval," PhD diss., Drexel University, 2011.

[2] S. Bozzon, A. Brambilla, M. Della Valle, E. Fraternali, S. Quarteroni, "An Introduction to Information Retrieval," In Web Information Retrieval, Springer Berlin Heidelberg, pp. 3-11 , 2013.

[3] H. Cui, J. R. Wen, J. Y. Nie, W. Y. Ma, "Query expansion by mining user logs," Knowledge and Data Engineering, IEEE Transactions on, 15(4), pp.829-839, 2003.

[4] F. serpush, MR. keyvanpour, "A Proposed Framework for Query Expansion Approaches for Web information retrieval". 4th international conference on information technology. ISC. pp:46-56, 2014.

[5] X. Xu, Z. Weizhong, Z. Xiaodan, H. Xiaohua, and S. Il-Yeol, "A comparison of local analysis, global analysis and ontology-based query expansion strategies for bio-medical literature search," In Systems, Man and Cybernetics, SMC'06. IEEE International Conference on, vol. 4, pp. 3441-3446. IEEE, 2006.

[6] M. Farhoodi., M. Mahmoudi., A. M. ZareBidoki., A.Yari, M. Azadnia, "Query expansion using Persian ontology derived from Wikipedia," World Applied Sciences Journal 7, no. 4 , pp. 410-417, 2009.

[7] W. C. Shih, S.S. Tseng, "A Knowledge-based Approach to Retrieving Teaching Materials for Context-aware Learning," Educational Technology & Society, 12(1), 82-106 , 2009.

[8] R .Fattahi., C. S. Wilson, F.Cole, "An alternative approach to natural language query expansion in search engines," Text analysis of non-topical terms in Web documents. Information Processing & Management, 44(4), pp.1503-1516 ,2008.

[9] F.serpush, MR. keyvanpour, "Categorization and Assessment of Approaches of QEBkS for Information Retrieval". The 4rd joint conference of AI & Robotics and 6th robocup Iranopen international Symposium., 2014.

[10] R.Campos, G.Dias, A.M.Jorge, A.Jatowt, "Survey of temporal information retrieval and related applications", *ACM Computing Surveys (CSUR)*, *47*(2), 15, 2014.

[11] P. A. Chirita, C. S. Firan, W. Nejdl, "Personalized query expansion for the web," In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval ,pp. 7-14, 2007.

[12] Andreou, Agissilaos. "Ontologies and query expansion." Univ. of Edinburgh, 2005.

[13] R. C. Bodner, F. Song "Knowledge-based approaches to query expansion in information retrieval," Springer Berlin Heidelberg,, pp. 146-158, 1996.

[14] J. Bhogal, A. MacFarlane, P. Smith, "A review of ontology based query expansion," Information processing & management, 43(4), pp.866-886, 2007.

[15] W. Zhu, "Text clustering and active learning using a LSI subspace signature model and query expansion," Doctoral dissertation, Drexel University, 2009.

[16] N.Cardoso, M. J. Silva, "Query expansion through geographical feature types," In Proceedings of the 4th ACM workshop on Geographical information retrieval ,pp. 55-60, 2007.

[17] M. Sanderson, "Retrieving with good sense," Information retrieval, 2(1), pp.49-69, 2000.

[18] Q. Zhao, S. S. Bhowmick, "Association rule mining: A survey," Nanyang Technological University, Singapore, 2003.

[19] C. D. Manning, P. Raghavan, H. Schütze, "Introduction to information retrieval," (Vol. 1). Cambridge: Cambridge University Press, 2008.

[20] E. Hoque, G.Strong, O. Hoeber, M. Gong, "Conceptual query expansion and visual search results exploration for Web image retrieval," In Advances in Intelligent Web Mastering–3 , Springer Berlin Heidelberg, pp. 73-82, 2011.

[21] A. Salamanca, E. Le´o , "An Integrated Architecture for Personalized Query Expansion in Web Search," In Proceedings 6th AAAI Workshop on Intelligent Techniques for Web Personalization & Recommender Systems, Chicago ,pp. 20-28 ,2008.

[22] B. D. Brewer, O. Hurst-Hiller, "Incremental query refinement," U.S. Patent No. 7,890,526. Washington, DC: U.S. Patent and Trademark Office, 2011.

[23] E. M. Voorhees, "Query expansion using lexical-semantic relations," In SIGIR'94 , London, pp. 61-69, 1994.

[24] D. Stenmark, "Query expansion using an intranet-based semantic net," Proceedings of IRIS-26, 2003.

[25] A. Lourdes, J. Perez-Iglesias, "Training a classifier for the selection of good query expansion terms with a genetic algorithm," In: Evolutionary Computation (CEC), IEEE Congress on, p. 1-8, 2010.

[26] C. de Loupy, P. Bellot, M. El-Beze, P. F. Marteau, "Query expansion and classification of retrieved documents," In *TREC* pp. 382-389, 1998.

[27] R. Fattahi, C. S.Wilson, F. Cole, "An alternative approach to natural language query expansion in search engines: Text analysis of non-topical terms in Web documents," *Information Processing & Management*, *44*(4), 1503-1516, 2008.

[28] S. Blott, F. Camous, C. Gurrin, G. J. Jones, "On the use of clustering and the MeSH controlled vocabulary to improve MEDLINE abstract search," 2005.

[29] A.P.Natsev, A. Haubold, J.Tešić, L.Xie, R.Yan, "Semantic concept-based query expansion and re-ranking for multimedia retrieval," In Proceedings of the 15th international conference on Multimedia, pp. 991-1000, 2007.

[30] Liu, Zhenyu. A knowledge-based approach to scenario-specific medical free-text retrieval. University of California at Los Angeles, 2005.

[31] R.Kosala, & H.Blockeel, "Web mining research: A survey" ACM Sigkdd Explorations Newsletter, 2(1), 1-15., 2000.

[32] Meusel, R., Niepert, M., Eckert, K., Stuckenschmidt, H. "Thesaurus extension using web search engines," In The Role of Digital Libraries in a Time of Global Change, 2010, pp. 198-207, Springer Berlin Heidelberg.

[33] S.Gauch, & J.B.Smith, "An expert system for automatic query reformulation," Journal of the American Society for Information Science, 44(3), 124-136, 1993.

[34] K.Park, H.Jee, T.Lee, S.Jung, H.Lim, "Automatic extraction of user's search intention from web search logs," Multimedia tools and applications, 61(1), 145-162, 2012.

[35] A.Alhroob, H.Khafajeh, N.Innab, "Evaluation Of Different Query Expansion Techniques For Arabic Text Retrieval System" American Journal of Applied Sciences, 10(9), 2013.

[36] J.Xu, & W.B.Croft, "Query expansion using local and global document analysis," In Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval ,, pp. 4-11, 1996.

[37] N.Yousef, I.Al-Bidewi, M.Fayoumi, "Evaluation of Different Query Expansion Techniques and using Different Similarity Measures in Arabic Documents," European Journal of Scientific Research, ISSN 1450-216X Vol.43 No.1, pp.156-166, 2010.

[38] Garron, A., & Kontostathis, A. "Latent Semantic Indexing with selective Query Expansion," In TREC, 2011.

[39] L.Liu, J.Kang, J.Yu, Z.Wang, "A comparative study on unsupervised feature selection methods for text clustering," In Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE'05. Proceedings of 2005 IEEE International Conference on, pp. 597-601, IEEE., 2005.

[40] C.Carpineto, & G.Romano, "Towards more effective techniques for automatic query expansion," In Research and Advanced Technology for Digital Libraries, pp. 126-141, Springer Berlin Heidelberg., 1999.

[41] Q. Zhao, S. S. Bhowmick, "Association Rule Mining: A Survey", CAIS, Nanyang Technological University, Singapore, Technical Report, 2003.

[42] H.Imran, A.Sharan, "Thesaurus and query expansion," International journal of computer science & information Technology (IJCSIT), 1(2), pp.89-97, 2009.

[43] L.Ramadier, et al. "Spreading Relation Annotations in a Lexical Semantic Network Applied to Radiology", *Computational Linguistics and Intelligent Text Processing*. Springer Berlin Heidelberg, 40-51, 2014.

[44] M.Zarrouk, M.Lafourcade, A. Joubert, "About inferences in a crowdsourced lexical-semantic network", *EACL*, 174, 2014.

[45] F.Ingrosso, A.Polguère, "How Terms Meet in Small-World Lexical Networks: The Case of Chemistry Terminology", In Terminology and Artificial Intelligence, pp. 167-171, 2015.

[46] I.Ruthven, "Re-examining the potential effectiveness of interactive query expansion," In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pp. 213-220, 2003.

[47] Y.Shen, X.He, J.Gao, L.Deng, G.Mesnil, "A latent semantic model with convolutional-pooling structure for information retrieval", InProceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, pp. 101-110, 2014.