International Journal of Information Science and Management
Vol. 16, No. 2, 2018, 19-42

A Useful Framework for Identification and Analysis of Different Query Expansion Approaches based on the Candidate Expansion Terms Extraction Methods

Mohammad Reza Keyvanpour
Associate Prof., Department of Computer Engineering, Alzahra University, Tehran, Iran
Corresponding Author, Keyvanpour@alzahra.ac.ir

Zahra Karimi Zandian
MSc, Data Mining Lab, Department of Computer Engineering, Alzahra University, Tehran, Iran
z.karimizandian@yahoo.com

Zahra Abdolhosseini
MSc, Data Mining Lab, Department of Computer Engineering, Alzahra University, Tehran, Iran
Zabdolhossein@gmail.com

Abstract

Query expansion is a method for improving retrieval performance by supplementing an original query with additional terms. This process improves the quality of search engine results and helps users to find the required information. In the recent years, different methods have been proposed in this area. In addition to such a variety of different approaches in this area and necessity of the study of their characteristics, the lack of a comprehensive classification based on candidate expansion terms extraction methods and also suitable and complete criteria to evaluate them, make the precise study, comparison and evaluation of methods for query expansion and choosing appropriate method based on need difficult for researchers. Therefore, in this paper a new useful framework is presented. In the proposed framework, in addition to the identification of three basic approaches based on the candidate expansion terms extraction methods for query expansion and expressing their properties, appropriate criteria for qualitative evaluation of these methods will be described. Next, the proposed approaches will be evaluated qualitatively based on these criteria. Using the systematic and structured framework proposed in this paper leads a useful platform for researchers to be provided for the comparative study of existing methods in the field, investigating their features specially their drawbacks to improve them and choosing appropriate method based on their needs.

Keywords: Query, Query Expansion, Classification, Document Content, External Knowledge Resources.

Introduction

In the last years the growth of the web in both content and users and the vast improvements in search engine technology have mainly changed how to collect and share knowledge and information. Nowadays the search engine plays the important role for users to access the required information. Although search engine technologies have grown and achieved a lot of success, however, there are still some problems to solve. For example, we

can refer to low precision in returned results or high volume of results. In this field, among the introduced methods for solving such problems, query expansion is one of the most important ones. Query expansion (QE) is a technology studied in the field of computer science, particularly within the scope of natural language processing, information retrieval (IR) (Kanaan, Al Shalabi, Ghwanmeh, & Bani Ismail, 2007) and data mining. Information retrieval focuses on finding documents whose content matches with a user query among a large document collection (Rivas, Iglesias, & Borrajo, 2014). Data mining is a process that uses data analysis tools to uncover and find patterns and relationships among data that may lead to extract new information from the large database (Keyvanpour & Imani, 2013; Karimi Zandian & Keyvanpour, 2017). QE is the process of refining the user query by adding new terms or reweighting query terms. Actually QE helps users find their required information (B. M. Kim, J. Y. Kim, & J. Kim, 2001; Lee, Huang, & Hung, 2007; Li & Agrawal, 2000). Some typical reasons to use the query expansion are introduced as follows:

- Increasing the quality of search results particularly from the view point of precision, recall and relevance measures, which are proposed in this field (Andreou, 2005).
- Helping the users create a good query that provides what the user really needs.
- Disambiguating the problems due to natural language processing and using single word to express a concept (Wang, Du, & Zhang, 2010).

Mainly query expansion has three main different types: Manually query expansion, interactive query expansion and automatic query expansion (Segura, García Barriocanal, & Prieto, 2011). In the manually QE, the user plays the most important role. User expands the query through adding terms to or removing terms from his/her query without any help from information retrieval system. Users can formulate the query using the facilities in the search engine interface to change language, and select documents that must be returned and decide which keyword is shown as a search result. In interactive query expansion, information retrieval system proposes expansion terms through interaction with the user. Finally, in third type, system proposes the expansion terms without user assistance (Cui, 2009).

Independent of query expansion types, this process has two phases. The first phase includes searching process for relevance terms and adding them to main query. These terms would be candidates for expansion. The second phase includes integration of expanded terms for new query (Gaillard, Bouraoui, De Neef, & Boualem, 2010). Regarding the existing methods for QE, some methods immediately work on query and expand it after applying the query. Most of these methods use external resources for QE, but others use returned documents for query expansion. We can use both strategies for QE. Figure 1 shows combination of first and second strategies.

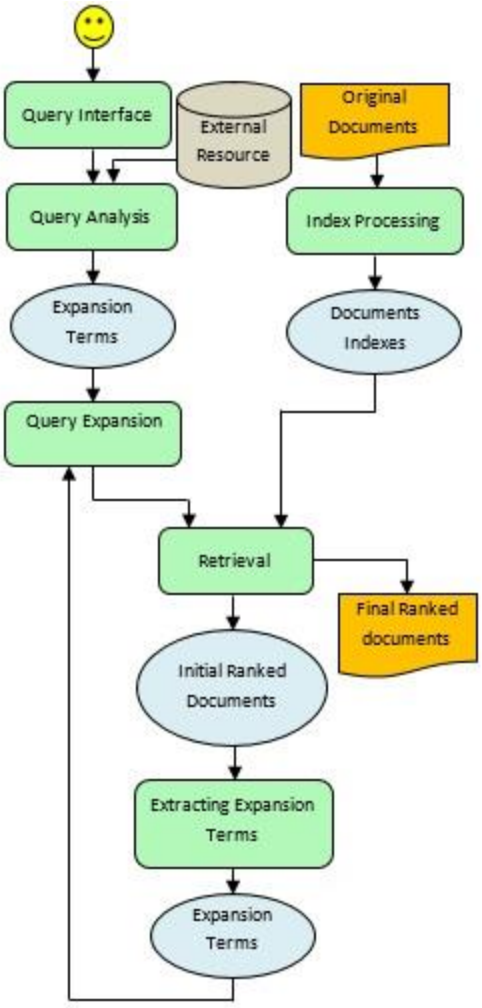

Figure 1. Architecture of query expansion

There are some challenges in the query expansion area, including:

- Problem of semantic understanding and dealing with language (Carpineto & Romano, 2012; Croft, Cronen Townsend, & Lavrenko, 2001)
- Needing speed in the search engine to reply to query (Bhogal, MacFarlane, & Smith, 2007; Carpineto & Romano, 2012)
- Difficulty of controlling the query expansion degree and the fact that the modified queries may contain lots of irrelevant terms, which can be seen as noise (Bhogal et al., 2007; Carpineto & Romano, 2012)
- "Query Drift" problem that is being acquired irrelevant and demanded results by expanding the query such as "out weighting" that is specifying more relevant result as a more irrelevant result and vice versa (Carpineto & Romano, 2012; Mitra, Singhal, & Buckley, 1998).

Query expansion is used in some applications, which employ information retrieval systems and search engines, like:

- Multi documents summarization (Zhao, Wu, & Huang, 2009)

- Image retrieval and search (Asbaghi, Keyvanpour, & Amiri, 2008; Hoque, Strong, Hoeber, & Gong, 2011; Keyvanpour & Tavoli, 2013)
- Question answering systems (Abouenour, Bouzouba, & Rosso, 2010; Jia, Sun, & Li, 2008; Liu, Fang, Hu, & Chen, 2008)
- Multimodal information retrieval (Díaz Galiano, Martín Valdivia, & Ureña López, 2009)
- Web information retrieval (Fattahi, Wilson, & Cole, 2008; Khozooii, Haratizadeh, & Keyvanpour, 2013)

Studying the methods in query expansion field shows that there are different methods and strategies for query expansion. Query expansion has been sought as a solution to the problem of imprecise query since the 1960s. Therefore, QE has a long history and various methods in this field have been proposed (Wollersheim, 2005). QE includes techniques such as finding synonyms, finding all the various morphological forms of terms by stemming each word in the search query, fixing spelling errors and automatically searching for the corrected form or suggesting it in the results and reweighting the terms in the original query (Farhoodi, Mahmoudi, Bidoki, Yari, & Azadnia, 2009; Pinto & Pérez Sanjulián, 2008).

Limited and a few works have presented classification of query expansion techniques based on expansion terms extraction ways (Andreou, 2005; Wollersheim, 2005). According to Andreou (2005) query expansion methods were divided to probabilistic and ontological methods. According to Wollersheim (2005) query expansion approaches were divided to algorithmic QE and interactive QE. According to Bhogal et al. (2007) ontology-based query expansion was divided into two categories: Query expansion using corpus independent knowledge models and Query expansion using corpus dependent knowledge models. Carpineto and Romano (2012) have classified automatic query expansion into five categories: linguistic analysis, corpus- specific techniques, query- specific techniques, search log analysis and web data. According to Ooi, Ma, Qin and Liew (2015) query expansion methods were divided into three classes: query expansion using corpus dependent knowledge model, query expansion using relevance feedback, query expansion using language model.

Investigating the methods offered in this area indicates various approaches for query expansion. In addition to such a variety of different approaches in this area and necessity of the study of their characteristics, the lack of a comprehensive classification based on candidate expansion terms extraction methods and also suitable and complete criteria, make the precise study, comparison and evaluation of methods for query expansion and choosing appropriate method based on need difficult for researchers.

Accordingly, in this paper, in addition to query expansion methods classification and expressing their properties, the approaches are evaluated qualitatively and compared technically based on new and useful criteria.

To achieve this goal, first, various methods suggested in query expansion field are studies. Next, their characteristics whether their advantages or disadvantages are determined and extracted. Then, the main ideas of these identified methods are specified and extracted. The methods are classified based on the obtained ideas and expressed properties of each class individually. Next, appropriate criteria are presented to evaluate the approaches qualitatively. Finally, these proposed approaches are studied comparatively based on the proposed criteria.

The research steps of the proposed useful framework for the classification of various query expansion methods includes:

- Classification of query expansion methods based on the candidate expansion terms extraction methods;
- Expression of descriptions and properties of query expansion approaches introduced in the first step individually;
- Explanation of useful criteria to evaluate query expansion approaches qualitatively;
- Evaluation and comparison of query expansion approaches to each other based on proposed criteria qualitatively.

The rest of the paper is organized as follows. First section deals with the classification of query expansion methods. In the following section the proposed criteria for evaluating the proposed approaches are discussed. Query expansion methods in the framework of the proposed classification and criteria are evaluated qualitatively in next section. Finally, in last section conclusion is presented.

Classification of query expansion methods based on the candidate expansion terms extraction methods according to the proposed framework

As shown in Figure 2, based on the proposed classification in this paper, query expansion methods based on the candidate expansion terms extraction methods are divided into 3 categories: document content based methods, external knowledge resources based methods and hybrid methods.

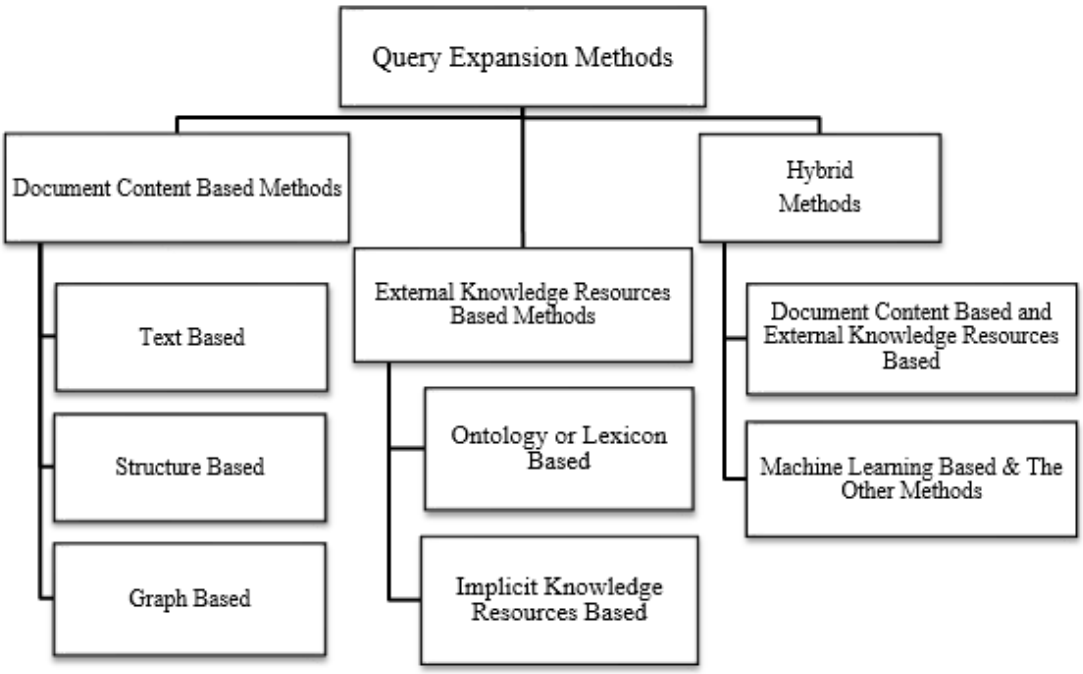

Figure 2. The proposed classification of query expansion methods

Document Content Based Methods

According to the proposed classification in this paper, in this category, there are methods that work on documents content.

In document content based methods, global analysis or local analysis based on selected document set is used to analyze and process (Chang & Ma, 2008). Due to extensive use of these concepts in text based methods, the kinds of analysis are explained in text based methods sub section.

Independent of the type of document content based methods, there are two approaches to process documents: statistical approach and natural language processing (NLP) approach. In statistical approach, statistical methods are used to extract candidate expansion terms. While, in NLP approach this process is done grammatically and linguistically. What is important is that the base of statistical methods is terms frequency. Terms frequency is calculated in documents. Terms with the highest frequency are introduced as candidates for expansion (Joho, Sanderson, & Beaulieu, 2004). Different methods to calculate terms frequency are included:

- Co-occurrence terms method: In this method the co-occurring terms with original query terms are selected as candidate terms for expansion. In fact, the idea is based on the association hypothesis (Imran & Sharan, 2010): "If an index term is good at discriminating relevant from non-relevant documents then any closely associated index term is likely to be good at this." This method supposes that co-occurrence pair of terms in documents often present similar topic. Therefore, co-occurrence terms with original query are good options for proposing as expansion terms. In other definition, we can say that this method is a probabilistic method and based on frequency of co-occurrence terms in training corpus (Farhoodi et al., 2009). Some works make similarity thesaurus based on co-occurrence terms. In similarity thesaurus the relation between terms and queries is calculated. In fact, this thesaurus is a matrix which contains similarity between terms. In this thesaurus, the term relation with query concept is the sum of its weighted relations with each terms in query. The query is expanded by n top terms with the highest weight, actually the terms that are the most similar to query will be selected (Tu, He, & Luo, 2009). In global analysis, co-occurrence terms are calculated based on all documents but in local analysis the expansion terms are extracted from retrieved top documents. Main question in this method is that how co-occurrence terms are extracted. There are standard measures for this work. For example, According to Imran and Sharan (2010) two coefficients namely jaccard and frequency were used.

$$jaccard\_co(t_i + t_j) = \frac{d_{ij}}{d_i + d_j - d_{ij}} \quad (1)$$

Where d_i and d_j are the number of documents in which terms t_i and t_j occur, respectively, and d_{ij} is the number of documents in which t_i and t_j co-occur.

$$freq\_co(t_i + t_j) = \sum_{d \epsilon D} (f_{d,t_i} \times f_{d,t_j}) \quad (2)$$

Where t_i and t_j are two terms of the query, $f_{d,ti}$ is frequency of term t_i in document d, $f_{d,tj}$ is frequency of term t_j in document d and D is number of top ranked documents used. We can apply these coefficients to measure the similarity between terms represented by the vectors.

Also there is another method in this topic namely local context analysis (LCA) that is one of the most successful methods (Sheng & Jiang, 2003). Clearly, this method selects expansion terms based on terms co-occurrence with original query terms. Simplified form of LCA formula for weighting terms is shown in Equation (3):

$$f(t,Q) = \frac{\prod_{q_i \epsilon\, Q}(\sum_{d \epsilon S} tf(t,d) * tf(q_i,d))}{N} * idf(t) \quad (3)$$

Where t is a document term, Q is the query, q_i is a query term, d is a document in n-top-ranked document set, tf(t,d) is the frequency of term t in document d, tf(q_i,d) is the frequency of q_i in document d, idf(t) is the inversed document frequency of term t and N is the number of top-ranked documents (N is a normalization factor and doesn't influence on ranking). Above formula is compared with tf*idf formula that is shown in Equation (4):

$$tf * idf = \frac{\sum_{d \epsilon S} tf(t,d)}{N} * idf(t) \quad (4)$$

This comparison shows that LCA method weights the term frequency with regards to effect of the query terms frequency. Then the high frequency terms that are presented in documents with high frequency of query terms are weighted by higher scores. Also, here it is tried to emphasize on terms that are simultaneously co-occurrence with all query terms.

- Clustering method: In this method clustering can be done on all documents or retrieved documents based on the original query. In first case it is supposed that all of documents are divided into sets of clusters (Xu & Hu, 2010) based on documents textual features. Notice that documents clustering in this approach are independent of query. Actually one of the most successful initial global analysis methods was terms clustering which is based on the association hypothesis (Andreou, 2005). Basically in this approach, the documents which are similar in majority of terms are clustered together. Discriminative terms from each cluster are used for query expansion. In second case clustering is dependent on query. Figure 3 shows the process of cluster based query expansion.

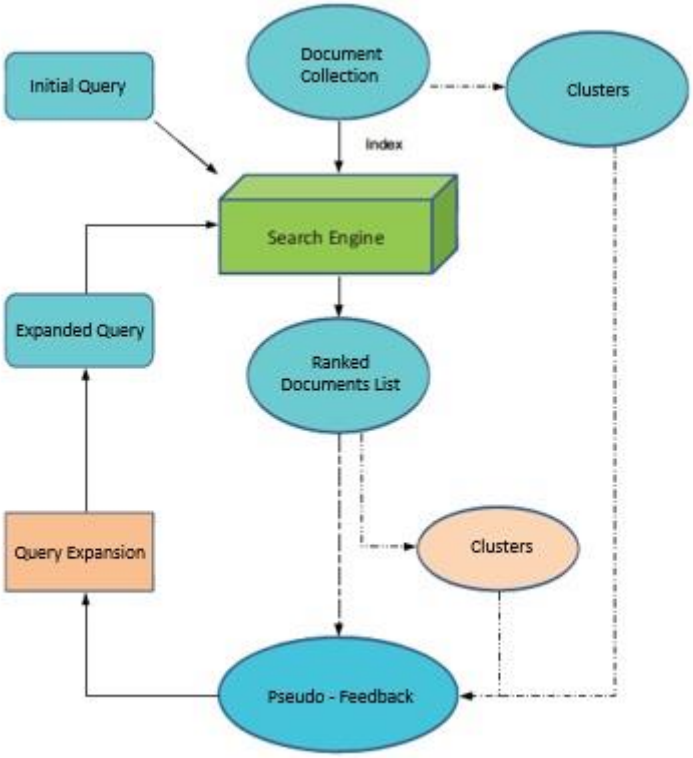

Figure 3. Cluster-Based Query Expansion (Xu & Hu, 2010)

- Relevance feedback method: The main idea of this method is to correct the original query by previous retrieved documents. The first user applies the query; next set of ranked documents are shown. User selects relevant documents from the retrieved documents. In this method expansion terms are selected from these documents that user confirms. Actually this method uses user interaction for specifying low numbers of relevance documents and then more similar documents are retrieved (Jia et al., 2008; Wollersheim, 2005). In query expansion based on relevance feedback, terms weights are updated based on relevance information. For example, terms weight in relevant documents can be increased and M top terms can be selected as candidate expansion terms. A lot of methods are suggested for terms weighting using relevance feedback. Robertson and Spark Jones function (Robertson & Jones, 1976) is explained as a method using relevance feedback. In this model w, weight of term t is shown in Equation (5):

$$w = \log \frac{p(1-\overline{p})}{\overline{p}(1-p)} \quad (5)$$

Where p is the probability of t occurring in relevant documents, and $\overline{p}$ is the probability of t occurring in non-relevant documents. p and $\overline{p}$ are calculated based on Equations (6) and (7),

$$p = \frac{r}{R} \quad (6)$$

$$\overline{p} = \frac{n-r}{N-R} \quad (7)$$

Where N is the total number of documents in a collection, n is the number of documents containing t, R is the number of relevant documents for a topic, and r is the

number of relevant documents in which t occurs. The term weight can be calculated for all terms in relevant documents. Then, the terms can be ranked in decreasing order of weight, and the top ranked terms can be used for expanding an existing query (Robertson & Jones, 1976). Other definition in this method is pseudo relevance feedback that is similar to relevance feedback with this difference that pseudo relevance feedback does not use the user interaction and supposes retrieved N top documents are relevant. Okapi BM25 (Matsumoto, Kurohashi, Nyoki, Shinho, & Nagao, 1991; Pant & Srinivasan, 2005) is the most widely-used pseudo relevance feedback algorithm, which uses the result of the first retrieval as a resource to extract more query terms for the second retrieval. According to Lin, Li, Hsu and Wu (2010) this algorithm was used as the basic pseudo relevance feedback. Used formula in this algorithm has been shown as follows. The similarity between query Q and document D_n is calculated by the Equations (8), (9) and (10):

$$Sim(Q, D_n) = \sum_{T \epsilon Q} \mathrm{w}^{\mathrm{l}} \frac{(k_I + 1)\mathrm{tf}(k_3 + 1)\mathrm{qtf}}{(\mathrm{k} + \mathrm{tf})(k_3 + \mathrm{qtf})} \quad (8)$$

Where

$$\mathrm{w}^{\mathrm{l}} = \log \frac{(\mathrm{r}+0.5)/(\mathrm{R}-\mathrm{r}+0.5)}{(\mathrm{n}-\mathrm{r}+0.5)/(\mathrm{N}-\mathrm{n}-\mathrm{R}+\mathrm{r}+0.5)} \quad (9)$$

$$k = k_I(1 - \mathrm{b}) + \mathrm{b}\frac{\mathrm{dl}}{\mathrm{avdl}} \quad (10)$$

Where N is number of documents in the collection, n is the number of items containing a specific term, R is the number of items known to be relevant to a specific topic, r is the number of these containing the term, tf is the frequency of occurrences of the term within a specific document, qtf is the frequency of occurrences of the term within a specific query, dl is the document length, avdl is the average of document length and k_i , b are the constants used in various best matching (BM) functions.

According to proposed classification, these methods are divided into three categories: text based methods, graph based methods, structure based methods.

Text Based Methods

In this category, texts are used for QE. Text mining technique is usually used to expand the query. Large volume of text documents which are available in the web, lead to discovering hidden knowledge from document corpuses. Text Mining is the discovery of new, implicit, and previously unknown information by automatically extracting information from text documents (Imani, Keyvanpour, & Azmi, 2013; Keyvanpour & Imani, 2013). Considering to be more important kind of analysis in text based methods, according to proposed framework in this paper, text based methods are divided into global analysis based and local analysis based (Figure 4).

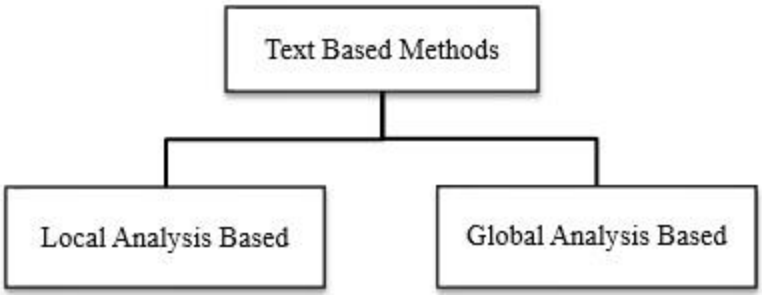

Figure 4. Classification of text based methods

✓ In the local analysis, documents returned based on initial query are processed. The whole of calculations are done online in this case and local analysis case uses subset of documents to build relations between terms and avoid analyzing the whole document space (Chang & Ma, 2008).

✓ In the global analysis, the whole of corpus is investigated and the records of predefined keywords which are relative to documents are stored. These keywords define main scope of documents (Xu & Croft, 2017). Usually in this case the calculation of terms correlation will happen only once at the time of system creation.

Structure Based Methods

There are some methods that use structural information of documents. Namely according to Yin, Tu, Xu and Zhang (2009) semantic skeleton was built for query and documents contents. The semantic skeleton includes keywords and their position in query sentences. Yin et al. (2009) presented a method to utilize the match of the semantic skeleton to search the documents. Using this method can avoid the difficulties of complex comprehension of natural language, effectively increasing the recall and speed of retrieval. In some works, regardless of other contents of pages only links existing between corpus pages namely Wikipedia corpus are used. Main idea of using links is that the links existing in pages refer to pages with similar contexts. In fact, in these works structural information of pages is used. According to Lin and Wu (2008) Wikipedia was used and QE was performed via link analysis. Wikipedia is an online free encyclopedia which can be edited by anyone online. Every entry in Wikipedia has links to related entries in Wikipedia or related web pages in other websites. In this paper, researchers believe that the anchor text of such links must contain related keywords. Therefore, they treat with these anchor texts as candidates of query expansion. Also, according to Leelapatra and Netisopakul (2008) new technique was proposed to improve query expansion using a link analysis technique called Hypertext Induce Topic Selection algorithm (HITS). In this work, there were two modules: link analysis module and query expansion module. The output of link analysis module was sent to query expansion module. According to Qian, Qian, Wei, Wang and Zhou (2002) to enable the structure expansion, first, a structure thesaurus was built based on the analysis of the XML corpus. With receiving a query, the structure thesaurus was examined and for each tag in the original query, one or more tags were retrieved from the same group.

Graph Based Methods

In this sub section, we refer to some works that use graph for QE. The construction of concept graph is one of artificial intelligence methods for presenting the hidden knowledge in documents. The nodes of this graph are concepts and its links are relations between concepts. One of the graph applications is finding the other correlative concepts for user query. This graph can be built by two main methods: the methods based on NLP and the methods based on statistical. In the first method concept graph is built by language processing and analyzing the content of texts. The resources used for these methods can be dependent on or independent of corpus. According to Farhoodi et al. (2009) concept graph was built by Wikipedia corpus. But there is another method for constructing concept graph. According to Amiri, AleAhmad, Rahgozar and Oroumchian (2008) statistical calculation was used to extract concepts and build graph. According to Zhao et al.(2009) a graph based method as a query expansion method was used for multi document summarization. This work uses both the relations between two sentences and relations between sentences and words to select appropriate discriminative words from document set and performs query expansion by them.

Properties of Document Content Based Methods

According to what is expressed in the last sub sections, we can summarize properties of document content based methods in Table 1.

Table 1
Document Content Based Methods and their properties

		Advantages	Disadvantages	Description
Text Based Methods	Global Analysis Based	1- Because of corpus global analysis, there is not the problem of non-reliability about relative or irrelative to initial retrieved results 2- Because performing QE is not dependent on initial retrieved results in this method and is done once, QE is done faster than local analysis method.	1- Very static 2- Needing documents containing keywords 3- This approach consumes a big part of calculation resources 4- Difficulty in reliability on completeness and correctness of corpus documents.	Frequently suitable terms for query expansion are found from statistical calculations and terms frequency in document set. These methods can use statistical preprocessing for query expansion.
	Local Analysis Based	1- These methods are query oriented 2- These methods avoid global analysis of corpus	1- If initial retrieved results be irrelative, performance is low because expansion terms are selected from these documents. 2- Needing calculation of term correlation for every query in run time. 3- Difficulty in reliability from completeness and correctness on corpus documents.	In this approach suitable terms are extracted by initial retrieved results processing. Frequently, statistical methods are used to extract candidate expansion terms. In this approach all calculations are performed after applying the query by user and results retrieval. Therefore, this approach is dependent on query.

	Advantages	Disadvantages	Description
Structure based methods	1- Increasing the recall 2- This method can avoid the difficulties of complex comprehension of natural language 3- Frequently this approach is independent of language. 4- Increasing the speed of retrieval.	1- Some structure based methods can be applied to only special corpuses such as Wikipedia	There are some methods that use structural information of documents instead of consideration of text content and in the many cases link analysis will be performed for query expansion. In the other word, connections and positions of words are considered.
Graph Based Methods	1- Query expansion may be performed semantically 2- This approach can receive suitable information with lower noise than traditional methods. 3- Concept graph can be used for natural language as semantic based language	1- Language dependent if concept graph is constructed by linguistic processing. 2- When corpus global analysis is needed, the process of graph construction is time consuming and requires high computational resource. 3- Valuation of concept graphs is a big issue. Because in concept graphs the number of concepts and their relations are quite high and this makes it impossible to evaluate the accuracy of all the relations directly.	In this approach the graph is constructed based on document analysis. Analysis method can be NLP or statistical method. Generally, the nodes are concepts and links show similarity between concepts. Similarity between terms is calculated by constructed graph or network and query expansion is performed based on that.

External Knowledge Resource-Based Methods

In this category, there are methods that use knowledge resources for query expansion. In these resources the relations between items have been defined explicitly or implicitly. The resources used in these methods can be independent of or dependent on corpus (Abouenour, Bouzoubaa, & Rosso, 2009). Based on proposed classification, external knowledge resource-based methods are divided into two categories: ontology or lexicon-based methods and implicit knowledge resources based methods.

Ontology or Lexicon-Based Methods

These approaches are typically based on dictionaries or other similar knowledge representation sources such as WordNet (Voorhees, 1994). An ontology is a formal naming and definition of the types, properties, and interrelationships of the entities that exist for a particular domain of discourse. The purpose of an ontology is to provide a context for the vocabulary it contains (Bhogal et al., 2007). Therefore, one type of dictionaries that consists of concepts and their relations is called ontology. The strategies based on tree structures and ontology can expand users query by the derivable semantic information from ontology. Generally, we can perform query expansion by external resources. So in many works this approach has been used. According to Barathi and Valli (2010) the similarity was calculated according to concepts semantic similarity and their relations. Concepts similarity is measured

by distance between concepts. The distance between different concepts is calculated based on their position in the concepts hierarchy. The position of a concept in a hierarchy is defined based on Equation (11), where k is a predefined factor and larger than l and L(n) is the depth of the node n in hierarchy.

$$Milestone(n) = \frac{\frac{1}{2}}{k^{L(n)}} \tag{11}$$

For the root of a hierarchy, L(root) is zero. For any two concepts c_1 and c_2 in the hierarchy having closest common parent (ccp), the distance d_c between these two concepts and their ccp is calculated by Equation (12) and the distance d_c between c_1 and ccp is calculated by Equation (13).

$$d_c(c_1, c_2) = d_c(c_1, ccp) + d_c(c_2, ccp) \tag{12}$$

$$d_c(c_1, ccp) = milestone(ccp) - milestone(c_1) \tag{13}$$

Thus, the similarity Sim_c between the two concepts c_1 and c_2 is calculated by Equation (14).

$$sim_c(c_1, c_2) = 1 - d_c(c_1, c_2) \tag{14}$$

Also, the similarity Sim_r between any two relations r_1 and r_2 is given by Equation (15).

$$sim_r(r_1, r_2) = 1 - d_r(r_1, r_2) \tag{15}$$

The distance between two relations is also calculated based on their positions in the relation hierarchy. In some works, domain ontology is used for query expansion like Segura, Vidal and Prieto (2010) and Zhang, Du, Li and Jia (2009) and the other are used general ontology for query expansion namely WordNet like Shabanzadeh, Nematbakhsh and Nematbakhsh (2010). WordNet is one of the most important lexical resources in information retrieval. In the works which use domain ontology, query terms domain is constrained to only special domain.

Implicit Knowledge Resources Based Methods

There are methods that use the resources with implicit or indirect relations between items for query expansion. Unlike ontology or lexicon based methods which contain direct relations between items, this category needs to analyze items and extract relations. For example, according to Ngok and Gong (2009) log information was used for QE, to analyze log items and extract relations between new applied query and last queries. So, we can introduce log as implicit knowledge resource. Notice in this method accessibility to log is considerable. Thus this method cannot be immediately used after building system. Log analysis method uses user click behavior to select expansion terms. In the other hand logs are analyzed to mining of relations between a query and observed items by user. Also we can introduce other methods based on log analysis such as: query expansion by similar queries like Bowman, Linden, Ortega and Spiegel (2014),Bowman and Spiegel (2011), Leblang, Ortega and Saunders

(2011), Moshfeghi, Velinov and Triantafillou (2016) and Zaïane & Strilets (2002), query expansion by queries which have returned similar URL like Baeza Yates, Hurtado, & Mendoza (2004), query expansion by the last query in query session like Carmel, Lewin Eytan, Libov, Maarek and Raviv (2017) and Chen, Cai, Chen and de Rijke (2017).

Properties of External Knowledge Resource Based Methods

According to what is expressed in the last sub sections, properties of external knowledge resource based methods have been summarized in Table 2.

Hybrid Methods

Nowadays many works use hybrid methods. These methods combine a number of methods for query expansion. In most works the main purpose of combining methods is access to semantic query expansion. Each method has many advantages and disadvantages. Applying these methods is useful because of increasing precision and maximizing the advantages and cover or decrease the disadvantages. We divide the existence methods to four categories based on the classification proposed in this paper. Figure 5 shows classification of these methods.

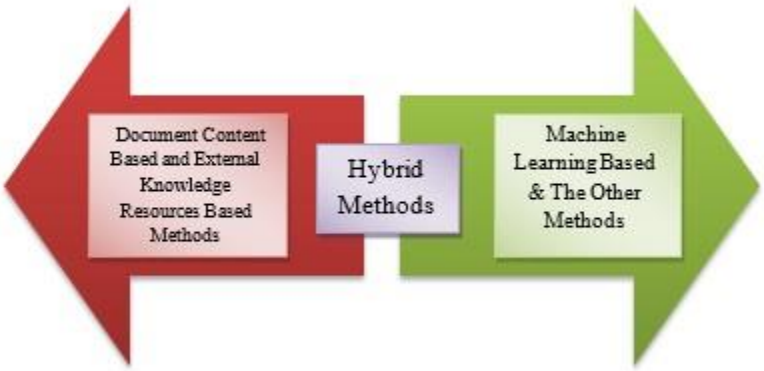

Figure 5. The Classification of Hybrid Methods

Document Content Based and External Knowledge Resources Based Methods

There are a lot of methods in this category. According to Chang and Ma (2008) a method has been proposed for finding the expansion terms by users' log analysis and topic clustering. According to Shabanzadeh et al. (2010) a method has been proposed by clustering semantically and using WordNet. In this method the candidate expansion terms are words that are relative to the whole of words in the cluster. According to Liu, Li, Zhang and Xiong (2008) a method has been proposed to calculate semantic similarity phrase. Also in this work initial retrieved results and WordNet are used and expansion units are phrases.

Table 2
External Knowledge Resource Based Methods and their properties

		Advantage	Disadvantage	Description
External Knowledge Resource Based Methods	**Ontology Based**	Semantically query expansion Speed of query expansion	Relations between concepts are complicated than defined relations in ontology because ontologies are static. Since words are categorized based on their type in ontology (e.g. noun, adjective, verb), two words with difference type cannot have a relationship. General ontologies don't have named entities. Ontology may be not matched with documents corpus. In some cases, ontology is domain dependent General ontology may contain ambiguity terms. Construction and maintenance costs are high. When query concepts are not related to ontology concepts, query expansion is not possible. Access to these semantic resources especially for some languages such as Arabic and Persian is difficult.	In this approach semantic relations between terms are considered in different formats such as hierarchical in knowledge structures. Actually semantic relations between initial query terms and the other terms in knowledge structure are used to query expansion.
	Implicit Knowledge Resources Based	Given that log data is created based on user options, user interests are payed attention more. Chosen words by this approach to expand are more appropriate and reliable.	If there are not any related queries in user log, query expansion process is failed.	This approach uses click behavior of users obtained from log system to choose expansion words. In the other word, logs are analyzed to mine relationships between a query and seen documents by users.

Machine Learning Based and The Other Methods

Unlike the pervious introduced methods, there are not many works in this approach and generally using machine learning techniques for query expansion has not been propagated yet. According to Al Shaor, Hmeidi and Najadat (2008) using genetic algorithm for query expansion was proposed. Also according to Krömer, Snášel, Platoš and Abraham (2010) genetic programming was used to extract suitable expansion terms. According to Han and Chen (2009) WordNet and RBF neural network were utilized for query expansion. In this work RBF neural network has been used to extract the most relative web documents and terms corresponding to them. According to Chen, Lin and Chang (2006) user judgment and neural network were used for query reweighting. This work consisted of two parts. In first part, the query vector formed by the weights of query terms was constructed and the degree of

similarity between each document vector and the query vector was calculated. The system retrieved and ranked the h-top documents having higher degrees of similarity with respect to the user's query, and the user marked each of the h'-top documents retrieved as a relevant document or an irrelevant document, where the values of h and h' were determined by the user and h>= h' >= 1. Then, the relevant documents among the h'-top documents retrieved were used to form the "cluster center vector". In second part the weights of query terms were adjusted based on a back propagation neural network.

The evaluation criteria of query expansion approaches based on the proposed framework

Performance evaluation of query expansion methods is often difficult, because different methods with a variety of approaches achieve this goal and applying different criteria to evaluate these methods is not possible. To evaluate the approaches outlined in the previous section, the following criteria have been considered and rankings are applied on four different levels: low, medium, high, very high except the Independence from Language that is Yes or No.

- Calculation Cost: The amount of documents' volume that is required to analyze and the number of required analysis and calculation to achieve to best solution.
- Semantic Level: The amount of semantic relationship between the words selected for expansion and the words in the query (Farhoodi et al., 2009).
- QE Speed: Speed of retrieve of best information to answer query.
- Precision: The number of relevant documents for each query is denoted as Relevant, the number of documents retrieved for the query is denoted as Retrieved, and the number of relevant documents correctly retrieved is denoted as Relevant ∩ Retrieved. The precision is defined in Equation (16) (Gao, Liu, Wang, Gu, & Yong, 2015).

$$Precision = \frac{\text{Relevant} \cap \text{Retrieved}}{\text{Retrieved}} \quad (16)$$

- Independence from Language: Is the query expansion approach independent of document language or not?

Evaluation of query expansion approaches using the proposed criteria

We introduced the measures for evaluation of query expansion approaches qualitatively in previous section. The results of comparing different mentioned techniques are shown in Table 3. Notice that we initial the values of each measure based on the majority role between current methods of each approach.

In text based methods, calculation cost of methods based on global analysis is more than methods based on local analysis, generally; because as mentioned before, in global analysis we have to investigate whole of text but in local analysis, some parts of the text are used to create relationships between words not whole of the text (Barathi & Valli, 2010). Due to process of parts of text retrieved based on initial query in local analysis based methods, semantic level and precision are usually better than global analysis based methods. It is clear that investigation based on a particular query causes better semantic level and precision (Xu

& Croft, 2000, 2017). In contrast, because in methods based on global analysis, the global techniques examine word occurrences and relationships in the carpus as a whole and use this formulation to expand any particular query (Xu & Croft, 2000, 2017), speed of query expansion in methods based on global analysis is higher than methods based on local analysis.

Table 3
The Evaluation of Query Expansion Approaches

			Calculation Cost	Semantic Level	QE Speed	Precision	Independence from Language
Document Content Based Methods	Text Based	Global Analysis Based	Statistical Methods: High NLP Methods: Very high	Statistical Methods: Low NLP Methods: Medium	High	Statistical Methods: Low NLP Methods: Medium	Statistical Methods: Yes NLP Methods: No
		Local Analysis Based	Statistical Methods: Low NLP Methods: High	Statistical Methods: Medium NLP Methods: High	Statistical Methods: Medium NLP Methods: Low	Statistical Methods: Medium NLP Methods: High	Statistical Methods: Yes NLP Methods: No
	Structure Based		Low	High	High	High	Yes
	Graph Based		Global Analysis: High Local Analysis: Medium	High	Global Analysis: High Local Analysis: Low	High	Statistical Methods: Yes NLP Methods: No
External Knowledge Resources Based Methods	Ontology Based		Low	High	High	Domain dependence: High Domain independence: Medium	No
	Implicit Knowledge Resources Based		High	High	Low	Domain dependence: High Domain independence: Medium	No
Hybrid Methods			Medium	Very high	Medium	Very high	No

According to the proposed classification, we can divide methods based on global analysis and local analysis into statistical and NLP methods and examine them. By investigating the existing methods in this area, it seems that although statistical methods have less calculation cost but semantic level and precision of query expansion are less than NLP methods (Strzalkowski & Vauthey, 1992). Because statistical methods use probabilistic correlations between query terms (Cui, Wen, Nie, & Ma, 2002) and used techniques are independent of document language but candidate expansion terms extraction in NLP methods is done

grammatically and linguistic (Grefenstette, 1997; Strzalkowski & Vauthey, 1992), so these methods are dependent on document language.

In the structure based approach, due to using structural information of documents query expansion based on this approach is independent of language. As a result, speed and precision of query expansion are improved (Yin et al., 2009), so semantic level criterion is high too. Calculation cost in these methods is low because of independent of the language and text (Abouenour et al., 2010; Dalton, Dietz, & Allan, 2014; Leelapatra & Netisopakul, 2008).

As mention before, the graph based methods create graph of concepts to show hidden knowledge of documents. Given that this approach uses relationships between concepts existed in document and query directly (Farhoodi et al., 2009), semantic level and precision of graph based methods are high (Amiri et al., 2008; Zhao et al., 2009). To express calculation cost and speed of this approach, it is important if the method is based on global analysis or local analysis. According to what was stated, in the global analysis based methods, the calculation cost and query expansion speed are higher than local analysis based methods (Xu & Croft, 2000, 2017). If to create and analyze of graph, statistical methods are used, this approach is independent of language of document but using NLP methods lead to be dependent on language.

In the external knowledge resources based methods, calculation cost of implicit knowledge resources based methods is high. In contrast, calculation cost of ontology based methods is low. Because in ontology based methods, the resources that are independent of carpus and created before starting query expansion (Bhogal et al., 2007) are used; so one time these resources are supplied and after that for each query expansion are used. While in implicit knowledge resources based methods, the resources that are dependent on carpus and click behavior of users obtained from log system to choose expansion words are used (Abouenour et al., 2009) and for each query expansion this process is repeated. As a result of what is expressed, QE speed in ontology based methods is higher than another type of methods in external knowledge resources based methods category. Both types of external knowledge resources based methods are dependent on language and resources related to each document and each language are used. Because in ontology based approach, ontology is the dictionary of concepts and their relation and implicit knowledge resources based methods use the resources with implicit or indirect relations between items for query expansion, semantic level of both of them are high. If for obtaining resources, domain of the document is paid attention and then the resources are provided precision of query expansion in both of approaches is high but if acquiring resources is independent of domain of document precision of query expansion in both of approaches is medium because in this case conditions and situation of the specific domain are not paid attention.

As mentioned before, in hybrid methods whether document content based and external knowledge resources based methods or machine learning based and the other methods, it is tried to improve advantages of methods and reduce their disadvantages to expand query and obtained results of both of them are the same so one row exists in Table 3 called hybrid methods. Precision and semantic level of hybrid methods usually are very high because using more than one technique and using their advantages lead to improve precision of query expansion and semantic level. But in contrast cost calculation and speed of them are medium

because running the hybrid methods of query expansion is longer than running one of them and it is necessary to spend more time to do process of query expansion. Finally, in hybrid methods used for query expansion, combining methods causes analysis to do simpler and amount of necessary calculation become lower. These methods are dependent on language (Chang & Ma, 2008; Chen et al., 2006; Han & Chen, 2009; Hmeidi, Najadat, & Al Sha'or, 2008; Krömer et al., 2010; Liu et al., 2008).

Conclusion

Query expansion process improves the quality of search engine results and helps users find the required information. In the recent years, different methods have been proposed in this area. Lack of a comprehensive classification in terms of candidate expansion terms extraction methods in this regard, expressing their features of the approaches and specified evaluation criteria are the main challenges facing researchers. Therefore, in this paper, a useful framework for identification, classification and analysis of query expansion approaches was presented. The proposed model provides a suitable platform for the comparative study of existing methods, investigating their features specially their drawbacks to improve them and choosing appropriate method based on their needs.

According to the proposed framework, query expansion methods could be divided based on terms of candidate expansion terms extraction methods into three categories: document content based methods, external knowledge resources based methods and hybrid methods. Assessment of the proposed techniques in each approach showed, with regard to advantages and drawbacks of each approach, that to achieve a comprehensive method with more advantages, the hybrid methods are more suitable. This combination usually eliminates failures in any of the methods and allows them to benefit from the advantages of others.

References

Abouenour, L., Bouzouba, K., & Rosso, P. (2010). An evaluated semantic query expansion and structure-based approach for enhancing Arabic question/answering. *International Journal on Information and Communication Technologies, 3*(3), 37-51.

Abouenour, L., Bouzoubaa, K., & Rosso, P. (2009). *Structure-based evaluation of an Arabic semantic Query Expansion using the JIRS Passage Retrieval system.* Paper presented at the Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages, Athens, Greece.

Al Shaor, A., Hmeidi, S., & Najadat, H. (2008). *Application of genetic algorithm in automatic query expansion.* Paper presented at the International Arab Conference on Information Technology, Sfax, Tunisia.

Amiri, H., AleAhmad, A., Rahgozar, M., & Oroumchian, F. (2008). Keyword suggestion using conceptual graph construction from Wikipedia rich documents. Paper presented at the International Conference on Information and Knowledge Engineering, Universal Conference Management Systems and Support, California, USA.

Andreou, A. (2005). *Ontologies and query expansion.* Master of Science diss., University of Edinburgh.

Asbaghi, S., Keyvanpour, M., & Amiri, A. (2008). *Learning-based approach for semantic image retrieval by using a dynamic semantic network.* Paper presented at the 19th International Workshop on Database and Expert Systems Application (DEXA'08), Turin, Italy.

Baeza Yates, R., Hurtado, C., & Mendoza, M. (2004). *Query recommendation using query logs in search engines.* Paper presented at the International Conference on Extending Database Technology, Berlin, Heidelberg.

Barathi, M., & Valli, S. (2010). Ontology based query expansion using word sense disambiguation. *International Journal of Computer Science and Information Security, IJCSIS*, 7(2), 22-27.

Bhogal, J., MacFarlane, A., & Smith, P. (2007). A review of ontology based query expansion. *Information processing & management, 43*(4), 866-886. http://doi.org/10.1016/j.ipm.2006.09.003.

Bowman, D., Linden, G., Ortega, R. E., & Spiegel, J. R. (2014). Identifying items relevant to a current query based on items accessed in connection with similar queries: Google Patents.

Bowman, D., & Spiegel, J. R. (2011). Identifying the items most relevant to a current query based on items selected in connection with similar queries: Google Patents.

Carmel, D., Lewin Eytan, L., Libov, A., Maarek, Y., & Raviv, A. (2017). *The demographics of mail search and their application to query suggestion.* Paper presented at the Proceedings of the 26th International Conference on World Wide Web, Perth, Australia.

Carpineto, C., & Romano, G. (2012). A survey of automatic query expansion in information retrieval. *ACM Computing Surveys (CSUR), 44*(1), 1-50, http://doi.org/10.1145/2071389.2071390.

Chang, P., & Ma, H. (2008). *The Theme-Mine in Query Expansion.* Paper presented at the IEEE Symposium on Advanced Management of Information for Globalized Enterprises (AMIGE 2008), Tianjin, China.

Chen, S. M., Lin, H. C., & Chang, Y.C. (2006). A new method for query reweighting for document retrieval based on neural networks. *International journal of information and management sciences,* 17(4), 95-110.

Chen, W., Cai, F., Chen, H., & de Rijke, M. (2017). *Personalized query suggestion diversification.* Paper presented at the Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan.

Croft, W. B., Cronen Townsend, S., & Lavrenko, V. (2001). *Relevance Feedback and Personalization: A Language Modeling Perspective.* Paper presented at the DELOS Workshop: Personalisation and Recommender Systems in Digital Libraries.

Cui, H., Wen, J. R., Nie, J. Y., & Ma, W. Y. (2002). *Probabilistic query expansion using query logs.* Paper presented at the Proceedings of the 11th international conference on World Wide Web, Honolulu, Hawaii, USA.

Cui, J. (2009). *Query Expansion Research and Application in Search Engine Based on Concepts Lattice.* Master of Science diss., School of Computing, Sweden.

Dalton, J., Dietz, L., & Allan, J. (2014). *Entity query feature expansion using knowledge base links.* Paper presented at the Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, Gold Coast Queensland, Australia.

Díaz Galiano, M. C., Martín Valdivia, M. T., & Ureña López, L. (2009). Query expansion with a medical ontology to improve a multimodal information retrieval system. *Computers in biology and medicine,* 39(4), 396-403. http://doi.org/10.1016/j.compbiomed.2009.01.012.

Farhoodi, M., Mahmoudi, M., Bidoki, A. Z., Yari, A., & Azadnia, M. (2009). Query expansion using persian ontology derived from Wikipedia. *World Applied Sciences Journal,* 7(4), 410-417.

Fattahi, R., Wilson, C. S., & Cole, F. (2008). 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. http://doi.org/10.1016/j.ipn.2007.09.009.

Gaillard, B., Bouraoui, J. L., De Neef, E. G., & Boualem, M. (2010). *Query expansion for cross language information retrieval improvement.* Paper presented at the Fourth International Conference on Research Challenges in Information Science (RCIS), Nice, France.

Gao, G., Liu, Y. S., Wang, M., Gu, M., & Yong, J. H. (2015). A query expansion method for retrieving online BIM resources based on Industry Foundation Classes. *Automation in construction,* 56, 14-25. http://doi.org/10.1016/j.autcon.2015.04.006.

Grefenstette, G. (1997). Short query linguistic expansion techniques: Palliating one-word queries by providing intermediate structure to text. *Information Extraction A Multidisciplinary Approach to an Emerging Information Technology* (pp. 97-114). Berlin: Springer.

Han, L., & Chen, G. (2009). HQE: A hybrid method for query expansion. *Expert Systems with Applications,* 36(4), 7985-7991. http://doi.org/10.1016/j.eswa.2008.10.060.

Hmeidi, I. I., Najadat, H. M., & Al Sha'or, A. J. (2008). *Application of genetic algorithm in automatic query expansion.* Paper presented at the Proceedings of the International Arab Conference on Information Technology (ACIT).

Hoque, E., Strong, G., Hoeber, O., & Gong, M. (2011). *Conceptual query expansion and visual search results exploration for Web image retrieval.* Paper presented at the Advances in Intelligent Web Mastering–3, Berlin, Heidelberg.

Imani, M. B., Keyvanpour, M. R., & Azmi, R. (2013). A novel embedded feature selection method: a comparative study in the application of text categorization. *Applied Artificial Intelligence,* 27(5), 408-427. http://doi.org/10.1080/08839514.2013.774211.

Imran, H., & Sharan, A. (2010). Selecting effective expansion terms for better information retrieval. *International Journal of Computer Science and Applications,* 7(2), 52-64.

Jia, K., Sun, Y., & Li, Z. (2008). *Answer extraction based on query expansion in Chinese question answering system.* Paper presented at the International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE'08), Beijing, China.

Joho, H., Sanderson, M., & Beaulieu, M. (2004). *A study of user interaction with a concept-based interactive query expansion support tool.* Paper presented at the *Advances in Information Retrieval*, Berlin, Heidelberg.

Kanaan, G., Al Shalabi, R., Ghwanmeh, S., & Bani Ismail, B. (2007). *A comparison between interactive and automatic query expansion applied on Arabic language.* Paper presented at the 4th International Conference on Innovations in Information Technology (IIT'07), Dubai, Dubai.

Karimi Zandian, Z., & Keyvanpour, M. (2017). Systematic identification and analysis of different fraud detection approaches based on the strategy ahead. *International Journal of Knowledge-based and Intelligent Engineering Systems,* 21(2), 123-134. http://doi.org/10.3233/KES-170357.

Keyvanpour, M. R., & Imani, M. B. (2013). Semi-supervised text categorization: Exploiting unlabeled data using ensemble learning algorithms. *Intelligent Data Analysis,* 17(3), 367-385. http://doi.org/10.3233/ida-130584.

Keyvanpour, M., & Tavoli, R. (2013). Document image retrieval: Algorithms, analysis and promising directions. *International Journal of Software Engineering and Its Applications,* 7(1), 93-106.

Khozooii, N. S., Haratizadeh, S., & Keyvanpour, M. R. (2013). An Analytical Framework for Web Information Filtering Techniques. *International Journal of Hybrid Information Technology*, 6(6), 345-358. http://doi.org/10.14257/ijhit.2013.6.6.31.

Kim, B. M., Kim, J. Y., & Kim, J. (2001). *Query term expansion and reweighting using term co-occurrence similarity and fuzzy inference.* Paper presented at the Joint 9th IFSA World Congress and 20th NAFIPS International Conference, Vancouver, BC, Canada.

Krömer, P., Snášel, V., Platoš, J., & Abraham, A. (2010). *Evolutionary improvement of search queries and its parameters.* Paper presented at the 10th International Conference on Hybrid Intelligent Systems (HIS), Atlanta, GA, USA.

Leblang, J., Ortega, R., & Saunders, C. (2011). Identifying the items most relevant to a current query based on user activity with respect to the results of similar queries: Google Patents.

Lee, H. M., Huang, C. C., & Hung, W. T. (2007). Mining navigation behaviors for term suggestion of search engines. *Journal of information science and engineering*, 23(2), 387-401.

Leelapatra, W., & Netisopakul, P. (2008). *Improving query expansion using link analysis.* Paper presented at the 5th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON 2008), Krabi, Thailand.

Li, W.S., & Agrawal, D. (2000). Supporting web query expansion efficiently using multi-granularity indexing and query processing. *Data & Knowledge Engineering*, 35(3), 239-257. http://doi.org/10.1016/S0169-023X(00)00024-0.

Lin, M. C., Li, M. X., Hsu, C. C., & Wu, S. H. (2010). *Query Expansion from Wikipedia and Topic Web Crawler on CLIR.* Paper presented at the Proceedings of NTCIR-8 Workshop Meeting, Tokyo, Japan.

Lin, T. C., & Wu, S. H. (2008). *Query expansion via wikipedia link.* Paper presented at the International Conference on Information Technology and Industrial Application.

Liu, M., Fang, F., Hu, Q., & Chen, J. (2008). *Question Analysis and Query Expansion in CS-CS IR4QA.* Paper presented at the Proceedings of NTCIR-7 Workshop Meeting, Tokyo, Japan.

Liu, Y., Li, C., Zhang, P., & Xiong, Z. (2008). *A query expansion algorithm based on phrases semantic similarity*. Paper presented at the International Symposiums on Information Processing (ISIP), Moscow, Russia.

Matsumoto, Y., Kurohashi, S., Nyoki, Y., Shinho, H., & Nagao, M. (1991). User's guide for the juman system, a user-extensible morphological analyzer for japanese. *Nagao Laboratory, Kyoto University*.

Mitra, M., Singhal, A., & Buckley, C. (1998). *Improving automatic query expansion.* Paper presented at the Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, Melbourne, Australia.

Moshfeghi, Y., Velinov, K., & Triantafillou, P. (2016). *Improving Search Results with Prior Similar Queries.* Paper presented at the Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, Indianapolis, Indiana, USA.

Ngok, P., & Gong, Z. (2009). *Log mining to support web query expansions.* Paper presented at the International Conference on Information and Automation (ICIA'09), Zhuhai, Macau, China.

Ooi, J., Ma, X., Qin, H., & Liew, S. C. (2015). *A survey of query expansion, query suggestion and query refinement techniques.* Paper presented at the 4th International Conference on Software Engineering and Computer Systems (ICSECS), Kuantan, Malaysia.

Pant, G., & Srinivasan, P. (2005). Learning to crawl: Comparing classification schemes. *ACM Transactions on Information Systems (TOIS),* 23(4), 430-462. http://doi.org/10.1145/1095872.1095875.

Pinto, F. J., & Pérez Sanjulián, C. F. (2008). Automatic query expansion and word sense disambiguation with long and short queries using WordNet under vector model. *Actas de los Talleres de las Jornadas de Ingeniería del Software y Bases de Datos, 2*(2), 17-23.

Qian, W. n., Qian, H. l., Wei, L., Wang, Y., & Zhou, A. y. (2002). *Structure-Based Query Expansion for XML Search Engine.* Paper presented at the Proceedings of the 12th International Conference on New Information Technology.

Rivas, A. R., Iglesias, E. L., & Borrajo, L. (2014). Study of query expansion techniques and their application in the biomedical information retrieval. *The Scientific World Journal,* 2014. http://doi.org/10.1155/2014/132158.

Robertson, S. E., & Jones, K. S. (1976). Relevance weighting of search terms. *Journal of the Association for Information Science and Technology,* 27(3), 129-146. http://doi.org/10.1002/asi.4630270302.

Segura, N. A., García Barriocanal, E., & Prieto, M. (2011). An empirical analysis of ontology-based query expansion for learning resource searches using MERLOT and the Gene ontology. *Knowledge-Based Systems,* 24(1), 119-133. http://doi.org/10.1016/j.knosys.2010.07.012.

Segura, N. A., Vidal, C. C., & Prieto, M. M. (2010). *Query expansion based on domain ontology for learning Objects search.* Paper presented at the 3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT), Chengdu, China.

Shabanzadeh, M., Nematbakhsh, M. A., & Nematbakhsh, N. (2010). *A Semantic based query expansion to search.* Paper presented at the International Conference on Intelligent Control and Information Processing (ICICIP), Dalian, China.

Sheng, X., & Jiang, M. (2003). *An information retrieval system based on automatic query expansion and hopfield network.* Paper presented at the Proceedings of the International Conference on Neural Networks and Signal Processing, Nanjing, China.

Strzalkowski, T., & Vauthey, B. (1992). *Information retrieval using robust natural language processing.* Paper presented at the Proceedings of the 30th annual meeting on Association for Computational Linguistics, Newark, Delaware.

Tu, X., He, T., & Luo, J. (2009). *Term relevance estimation for Chinese query expansion.* Paper presented at the International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE 2009), Dalian, China.

Voorhees, E. M. (1994). *Query expansion using lexical-semantic relations.* Paper presented at the Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, Dublin, Ireland.

Wang, C., Du, Y., & Zhang, P. (2010). Optimization of Query Expansion Source in Formal Concept Analysis. *Journal of Convergence Information Technology,* 5(7), 133-140. http://doi.org/10.4156/jcit.vol5.issue7.17.

Wollersheim, D. (2005). *Dynamic query expansion for information retrieval of imprecise medical queries.* PhD diss., La Trobe University, Bundoora, Victoria.

Xu, J., & Croft, W. B. (2000). Improving the effectiveness of information retrieval with local context analysis. *ACM Transactions on Information Systems (TOIS),* 18(1), 79-112. http://doi.org/10.1145/333135.333138.

Xu, J., & Croft, W. B. (2017). *Quary expansion using local and global document analysis.* Paper presented at the SIGIR'96 Proceedings of the 19th annual international ACM,

SIGIR conference on Research and development in information retrieval, Zurich, Switzerland.

Xu, X., & Hu, X. (2010). *Cluster-based query expansion using language modeling in the biomedical domain.* Paper presented at the IEEE International Conference on the Bioinformatics and Biomedicine Workshops (BIBMW), Hong Kong, China.

Yin, W., Tu, P., Xu, F., & Zhang, H. (2009). *The Query Expansion Method Based on Semantic Skeleton.* Paper presented at the International Workshop on Intelligent Systems and Applications (ISA 2009), Wuhan, China.

Zaïane, O. R., & Strilets, A. (2002). *Finding similar queries to satisfy searches based on query traces.* Paper presented at the International Conference on Object-Oriented Information Systems, Berlin, Heidelberg.

Zhang, B., Du, Y., Li, H., & Jia, L. (2009). *The Method of Query Expansion Based on Domain Ontology.* Paper presented at the Pacific-Asia Conference on Circuits, Communications and Systems (PACCS'09), Chengdu, China.

Zhao, L., Wu, L., & Huang, X. (2009). Using query expansion in graph-based approach for query-focused multi-document summarization. *Information processing & management,* 45(1), 35-41. http://doi.org/10.1016/j.ipm.2008.07.001.