Review of Probabilistic and Ontology-based Methods in Query Expansion

Vahideh Reshadat, Sara Khoshvatan, Mohammad-Reza Feizi-Derakhshi

***Abstract*—Currently World Wide Web has caused large amount of information become available for users as web pages due to distribution and low cost of content production. Information retrieval systems observe documents as words and if there is a relevant document which doesn't include query keywords, will not retrieved. One of the methods of improving information retrieval performance is to expand and refine of query which means adding words into user's query in order to improve the retrieval results. Finding these words can be done with or without user interference. In order to select words, probabilistic or ontology-based methods can be used. In this paper we will review different available methods for query expansion.**

I. Introduction

Nowadays World Wide Web faces with new challenges including abundant volume of information, heterogeneity, and non-structured information due to distribution and low cost of content production. As World Wide Web and multi-purpose use of web search engines expands, the number of users who use information retrieval systems is increased. In this situation, collection, organization, and useful sharing of information have significant importance. On the other hand, existing of so many text pages in World Wide Web caused results which search engines provided, contain a lot of irrelevant information. Consequently, finding the information which user needs has become difficult and complicated and proper retrieval of information becomes more important. Most significant reasons of retrieval of irrelevant documents and low precision of search engines are summarized as follow items:

- Most existing search engines still rely solely on the keywords contained in the queries to search and rank relevant documents.
- Most of the search engines compare the terms in lexical level rather than semantic level.
- Users formulate short query.
- The web is not a well-organized information source where innumerous "authors" created and are creating their websites independently. Therefore, the "vocabularies" of the authors vary greatly.
- Users usually tend not to use the same terms appearing in the documents as search terms.
- Natural language is ambiguous inherently. So natural language queries sometimes are ambiguous.

Most of the users who use information retrieval systems are ordinary users. It is generally accepted that such users typically are not likely to introduce complete and comprehensive queries. Even experienced users of IR systems when don't know the collection well cannot formulate a good query. It seems search engine should try to solve these problems themselves. So search engines use several ways to improve retrieval performance. One of the most popular and efficient mechanisms is query expansion. Query expansion means reconfiguration of query or adding terms in order to improve results. Search engines use different methods for query expansion including finding and adding synonym words, finding and adding all derivable words of a term, correcting words which have false spelling, reweighting of query terms, using users' feedback and ontology.

The remainder of this paper is organized as follows: a general definition of query expansion and its types is expressed in section II and a review of probabilistic and ontology methods is done in this section. In section III different methods is compared to each other and the paper is concluded with a summary of the discussed issues in section IV.

II. Query Expansion

Query expansion is the process of augmenting the user's query with additional terms in order to improve results. An important matter in query expansion is the ability of expansion words in distinguishing the documents. Suitable words are those that can retrieve more relevant documents and don't retrieve irrelevant results. Query expansion has some problems. The worst problem is query drifting, which is moving the query away from the user's intention. This happens frequently when the query is ambiguous. For example, the query "apple" might be about apple as fruit or apple computers. Outweighting is a specific kind of query drift. Outweighting is occurred when the augmentation terms are strongly related to the individual query terms but not to

Manuscript received July 30, 2011. This work was supported in part by the Department of Computer, Islamic Azad University, Shabestar branch.

Vahideh Reshadat is with the Department of Computer, Islamic Azad University, Shabestar branch, Shabestar, Iran (e-mail: v_reshadat@ yahoo.com).

Sara Khoshvatan is with the Department of Computer, Islamic Azad University, Shabestar branch, Shabestar, Iran (e-mail: sarakhoshvatan@ yahoo.com).

Mohammad-Reza Feizi-Derakhshi is with the Department of Computer, University of Tabriz, Tabriz, Iran (e-mail: mfeizi@tabrizu.ac.ir).

the whole query [4]. For tackle these problems many solution proposed. Reformulation and expansion the query divided in three sections: Manual query expansion in which query formulation done in some iteration. After reading some of the documents in the initial result set, the user might append additional terms to the query or remove terms from the original query. Semi-automatic query expansion which in a list of candidate terms is presented to the user who makes the final decision and Automatic query expansion that query is expanded automatically and is hided from the users [5].

There are two main approaches to query expansion covered in the literature. The dominant one is that of probabilistic query expansion. Probabilistic query expansion use statistical measures such as co-occurrence measures in documents to select expansion terms. Selecting terms that are most related to query terms. Ontological methods suggest an alternative approach which uses semantic relations drawn from the ontology to select terms [4][7]. The main goal in query expansion is expanding the query in a way which system removes ambiguities from the query, gets a correct understanding from user intention and retrieves relevant document.

A. Probabilistic Methods

Most probabilistic methods can be categorised as global or local. Global techniques extract their co-occurrence statistics from the whole document collection. When users submit a query, the system selects the most related terms according to the pre-calculations and expands the query. Local techniques extract their statistics from the n-top documents returned by an initial query. Local techniques are based on the hypothesis that the n-top documents are relevant to the query. They must be fast because they delay the response of the system. All calculations for local methods are done on-line [4][7][13].

One of the earliest global analysis techniques is Term Clustering which groups document terms into clusters based on their co-occurrences. Term clustering is based on the association hypothesis. Namely that terms related in some corpus tend to co-occur in the documents of that corpus. The most representative work on term clustering was conducted by Spark Jones in the late 60ʹs and early 70ʹs. Expansion terms were selected from the clusters which contained the query terms [4][34]. Well-constructed term clusters can improve retrieval performance. A serious problem with term clustering is that it cannot handle ambiguous terms. If a query term has several meanings, term clustering will add terms related to different meanings of the term and make the query even more ambiguous. In this case, it will lower retrieval effectiveness [34].

Latent Semantic Analysis is a numerical approach which uses terms co-occurrence. LSA is dimensional reduction technique which can distinguish most important dimension and removes low important dimension and improves Vector Space Model. In LSA terms vector map to the high level concept in the low dimension. LSA hope to solve the problem of Vector Space Model specially existence of synonyms and polysynonym terms. Despite the potential claimed by its advocates, retrieval results using LSA so far have not shown to be conclusively better than those of standard vector space retrieval systems. It is expensive and if a query term is ambiguous, terms related to different meanings of the term will have similar reduced representations. This is equivalent to adding unrelated terms to the query [21][34].

Abdelali [2] proposed a novel QE approach which use LSA mechanism. This approach uses a total sense vector that represents the entire query semantics. Then use the new total query vector to find the closest matching set of words/documents vectors in the corpus. The obtained list will be used to expand the query. Experiments results show a significant enhancement of recall and precision because it uses words and documents to expand the query which are low but relevant [2].

Frei and Qiu in [24] used Similarity Thesaurus to expand query. A Similarity Thesaurus is a matrix which is built considering the term to term relationship rather than simple co-occurrence data. Each component of this term-term matrix shows probability of co-occurrence of *i-th* and *j-th* terms. It is built automatically. Terms for expansion are selected based on similarity to the whole query rather than their similarity to individual term. When Similarity Thesaurus constructed the first r terms of the collection or terms crossing a similarity threshold can be used for query expansion. Experiments show that this kind of query expansion results in a notable improvement in the retrieval effectiveness when measured using both recall-precision and usefulness. But construction of such a Similarity Thesaurus for a large database is computationally expensive [13][24].

Jing and Croft [34] proposed technique which is called PhraseFinder to construct collection dependent association thesauri automatically using large full-text document collections. PhraseFinder considers co-occurrences between phrases and terms as associations. For association generation, sentence by sentence, PhraseFinder reads texts, recognizes phrases and terms and generates association data. Each association is a triple:

<termeid, phraseid, association_frequency>

The association frequency is equal to term frequency times phrase frequency. The range which used to generate associations were natural paragraph and 3-10 sentences per paragraph is sufficient for full texts. Each concept is represented as a set of tuples which show the words and phrases related to the concept. Given a query Q, The highest ranked concepts are used for query expansion. It seems that the larger the collection, the better the association thesaurus performs. The approach to the construction is realistic and general [16][34].

Local analysis can be traced at least back to Attar and Fraenkel [1] which used a similar global approach to term clustering to select expansion terms, but, needless to say, since it was a local method the clusters were created from terms of the n-top results of an initial query. Expansion terms were selected from the clusters which contained the query terms. If a large fraction of the documents are relevant and

clusters well-constructed then retrieval performance will improve else retrieval performance will degrade [4].

One of the famous local techniques is relevance feedback. The idea of relevance feedback (RF) is to involve the user in the retrieval process so as to improve the final result set. First the user issues a query, the system returns an initial set of retrieval results, the user marks some returned documents as relevant or non-relevant and then the system computes a better representation of the information need based on the user feedback. In reconstruction the query it used Rocchio [26] algorithm which introduced in and popularized by Salton‟s SMART system around 1970. Rocchio proposed the following formula, which is used to improve an initial query q:

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

Where q0 is the original query vector, Dr and Dnr are the set of known relevant and non-relevant documents respectively, and α, β, and γ are weights attached to each term. These control the balance between trusting the judged document set versus the query. The relevance feedback methods depend heavily on the kind and quality of the user relevance information. If the user distinguish the relevant document well the precision and recall improve. The major problem of relevance feedback is that it is difficult to get relevance information provided by users. Users are often reluctant to provide explicit feedback, or in general do not wish to prolong the search interaction [7][21][22][34].

To overcome the difficulty due to the lack of sufficient relevance judgments, Pseudo-Relevance Feedback (also known as Blind Feedback) is commonly used provides a method for automatic local analysis. It automates the manual part of relevance feedback. Local feedback mines relevance feedback by assuming the top-ranked documents to be relevant. Expansion terms are extracted from the top-ranked documents to formulate a new query for a second cycle retrieval. Experiments show that Pseudo Relevance Feedback act better than global technique. This method has obvious drawback: if a large fraction of the documents assumed relevant is actually non-relevant, then the words added to the query (drawn mostly from these documents) are likely to be unrelated to the topic and the quality of the documents retrieved using the expanded query is likely to be poor. Thus the effects of Pseudo Feedback strongly depend on the quality of the initial retrieval [21][22].

In recent years, many improvements have been obtained on the basis of local feedback, one them is proposed by Mitra [22] which involves improving precision in top ranks documents. So the initially retrieved documents are examined for additional, strong indications of relevance. To use K documents in the feedback process, retrieve a large number T of documents using the original query, for each retrieved documents, compute a new similarity score. Mitar used Boolean constraints, Proximity constraints and Fuzzy Boolean operators in computing new score. Rerank the documents according new score and select the top K documents in the new ranking and use them in the Rocchio relevance feedback process to expand the query. Then expanded query is used to retrieve the final list of documents returned to the user. Experiments results show that refining the set of documents used in query expansion often prevents the query drift caused by blind expansion and yields substantial improvements in retrieval effectiveness, both in terms of average precision and precision in the top twenty documents [22].

Croft [34] proposed a new technique, called Local Context Analysis, which combines the advantages of a PhraseFinder and local feedback. In general, LCA is one of the most successful and well established query expansion methods. The main goal of Croft was to solve problems of existening expansion methods specially PhraseFinder and Local Feedback. Croft showed that combining method has better performance. More specifically, expansion terms are selected not based on their frequencies in the top-ranked documents but rather on their co-occurrences with query terms. When user submits the query, system retrieves a set of documents as response and n-top documents are selected from the set. Nouns and noun phrases are extracted from the passages. For each nouns and noun phrase the degree of co-occurrence with query terms is computed. Then r-top concepts which co-occurrence with more query terms and have higher co-occurrence degree are used as expansion terms. Experiment results showed more than 20% improvement in average precision [34].

Like LCA Abdelmgeid [3] proposed a method which combine global and local techniques. More precisely he applied two query expansion methods in sequence to reformulate the query. One method is Similarity Thesaurus based expansion and the other is Local Feedback method [3].

Cui [7] proposed a new method for query expansion based on query logs. The central idea of his method is that if a set of documents are often selected for the same queries, then the terms in these documents are strongly related to the terms of the queries. Thus some probabilistic correlations between query terms and document terms can be established based on the query logs. These probabilistic correlations can be used for selecting high-quality expansion terms from documents for new queries. A series of experiments showed that the log-based method can achieve substantial performance improvements. Computing the term correlations offline and considering user interests are the advantages of this method. The number of relevant expansion terms suggested this method is rather than LCA. When the system developed newly, log file is empty or has not enough information and it is impossible to use this method [7].

Jiang [15] propose a probabilistic query expansion based on a new User Interest Model which is constructed and updated automatically. At the time when people first use this system, people need to fill in a form which log people‟s initial interests, then in the process of using this system, system can update these interests automatically by querying logs. Because of the user may be interested in many sorts of interests in coordinately in a period of time, user interest can be figured with key words and its weight. User Interest Tree

is a tree structure. It can show the information about the topic and key words of user interest. We use UIT to select the expansion terms. The advantage of this method is clustering the terms automatically according to their meaning. In contrast the weight of key word should change every time and this is time consuming [15].

Query expansion technology and Formal Concept Analysis (FCA) are two effective methods for improving the query precision in the information retrieval field. So Zhang [35] proposed a query expansion mechanism based on user interest topics. Zhang first use TREC as formal contexts, and build concept lattices based on FCA. The expansion source can be produced by using concept lattice. When expansion source produced, it searches for nodes in Open Direct Project which are matching with expansion source. The Open Directory Project (ODP) is the largest, most comprehensive human-edited directory of the Web. It is constructed and maintained by a vast, global community of volunteer editors. It describes the hierarchy relations among user interest topics by popular nouns. The nodes which retrieve as results show topics and sub topics which are related with user query and the key words of them added to user query. This method can improve the precision of the query in the case of basically maintaining the recall [35].

B. Ontology Methods

Ontology is a conceptual model which describes objects in a domain and relationships among them [30]. According to Gruber [11] an ontology is a "shared specification of a conceptualization" and Borst [6] defined it as "a formal, explicit specification of a shared conceptualization [11]. The basic building blocks of ontologies are concepts and relationships. Concepts appear as nodes in the ontology graph and relationships usually connect two or more concepts. *is-a* and *part-of* are the main relations in ontology [4]. WordNet is one of the ontologies which is not limited to a specific domain .WordNet is a manually-constructed lexical system developed by George Miller and his colleagues at the Cognitive Science Laboratory at Princeton University .The basic object in WordNet is a set of strict synonyms called a synset. By definition, each synset in which a word appears is a different sense of that word. There are four main divisions in WordNet, one each for nouns, verbs, adjectives, and adverbs. Within a division, synsets are organized by the lexical relations defined on them. [32].

Using ontologies for query expansion can be dated at least up to Voorhees [32]. In her paper, Elen Voorhees outlines a method for using ontologies for query expansion, Most approaches use large lexical ontologies (usually WordNet or Cyc) because they are not domain specific and because their relations are not sparse [4]. A word may have a number of different meanings depending on its context. Query may contain vague words, the systems will retrieve all sets of documents while the users usually want only one set. This drawback leads to poor precision. To solve this problem, sense disambiguation can be employed [14]. The method most commonly used is that first, the query terms is disambiguated so that they map to a unique ontology concept. Then terms related in the ontology to the disambiguated concepts are added to the query [4].

One of the first tasks in this area is done in [31]. She used the extended vector space model of information retrieval that was introduced by Fox. She describes an automatic indexing procedure that uses the *is-a* relations contained within WordNet and the set of nouns contained in a text to select a sense for each polysemous noun in the text. The particular disambiguation technique used in this work is based on the idea that a set of words occurring together in context will determine appropriate senses for one another despite each individual word being multiply ambiguous. Retrieval experiments comparing the effectiveness of these sense-based vectors vs. stem-based vectors show the stem-based vectors to be superior overall, although the sense-based vectors do improve the performance of some queries. The overall degradation is due in large part to the difficulty of disambiguating senses in short query statements [31].

Another expansion process involves semantic/lexical relations within WordNet is proposed by Voorhees. She tries four expansion strategies: expansion by synonyms only, expansion by synonyms plus all hyponyms, expansion by synonyms plus the parent hypernyms plus all hyponyms, expansion by synonyms plus the relevant synset. The result indicates that all of the expansion strategies just improve the retrieval performance a little, the semantic/lexical relations don"t make significant advantage. The process of query expansion that is based on WordNet benefits short query statement more than long statement [32].

Smeaton in [28] tried to expand the queries of the TREC-4 collection with various strategies of weighting expansion terms, along with manual and automatic word sense disambiguation techniques. Unfortunately all strategies degraded the retrieval performance [28].

Instead of matching terms in queries and documents, Richardson in [25] used WordNet to compute the semantic distance between concepts or words and then used this term distance to compute the similarity between a query and a document. Although proposed two methods to compute semantic distances, neither of them increased the retrieval performance [25].

Navigli [23] proposed a disambiguated method which is done by Creation of semantic networks for each sense of word and intersection and scoring. After pruning stop words, for each word in the initial query WordNet synonym sets are extracted. Each possible combination of synonym sets for initial query is considered as a configuration. In each configuration semantic network for each sense is created and intersected and a score is assigned and finally configuration with high score is selected as words sense. Five sense-based expansion methods is used to choose expansible words. The following expansion methods are explored: synset, hyperonym, Gloss synset, Gloss words expansion, common nodes expansion. Although size of the experiment was limit but the results indicate that all of five expansion strategies produce an improvement [23].

Another method is proposed by Liu [18] that generally in which noun phrases in queries are identified and then Word

sense disambiguation is done by use of adjacent words in the query. Whenever the sense of a query term is determined, its synonyms, hyponyms, words from its definition and its compound words and also terms from n-top ranked documents are considered for possible additions to the query. Results show that this approach yields significant improvements in for secor ptypical short queries [18].

Song [29] presents systems that in the initial stage of retrieval process, a set of user-provided instances are used and system retrieves a sample of documents. On the retrieved document set, important terms and phrases are selected by applying IR and natural language processing techniques. The Apriori algorithm is used to mine association rules from retrieved documents. Association rules are applied to query expansion and after disambiguating, WordNet is used as ontology to find relevant entries semantically and syntactically for query expansion. In comparison with cosine similarity, SLIPPER and Okapi BM25, this method neco eeh successful [29].

WordNet may bring many noises for the expansion due to its collection independent characteristic. Furthermore, it may not catch current state of words and their relationships so Gong has employed a Term Semantic Network (TSN) which is created with respect to word co-occurrence in the collection. Since TSN is directly extracted from the collection, it can overcome these shortages. He used TSN both as a filter and a supplement for WordNet. Indeed words with highest confidence and support with words which are not described in WordNet but can be used to expand the original query. To eliminate some noise words, those with lowest confidence and support are removed. For expanding query hypernyms, hyponyms and synonyms relations are combined. Experiments shows that WordNet with TSN filtering yields a significant increase in the number of correct documents retrieved and combiningrpo WordNet-TSN-Filtering with TSN expansions is much improved than any other alone. It is assumed that queries are single-word queries and is employed in Web image search system [9]. To overcome this disadvantage Gong [10] proposed a method for multi-term query expansions. WordNet assign terms in the same query into different groups with respect to their semantic similarities. For each group the highest terms in the WordNet hierarchies are expanded b hypernym and synonym, the lowest terms b hyponym and synonym, and all other terms b only synonym. Furthermore, with use of collection related TSNdthe low-frequency and unusual words in the expansions are removed. Group terms with the similarity threshold as 0.05 in the same query improve the query performance dramatically [10].

In [19] author presents a novel query expansion algorithm, namely PSS-QE (Phrase Semantic Similarity based Query Expansion). Phrases instead of wordsoeœ used to expand. This method considers query as the expression, extracts key phrases from original results, and calculates the semantic similarity between the query phrase and each phrase are extractedob using the semantic similarity algorithm based on WordNet, and then expands the query with the most similar phrases to search again. Experiments results show that the proposed algorithm can provide more precision.

In [20] another method is proposed which establish the domain knowledge database by adopting ontology as the describing tool, then parse the query string into terms, and then construct semantic diagraph with each term as the first vertex based on the domain knowledge and calculate the semantic distance between the first vertex and each vertex in the semantic diagraph, then according to the threshold, select the expanded terms of each semantic diagraph, and at last combine all the term gotten from the semantic diagraph with logic operator and then obtain the result of query expansion.

In [12] a hybrid method is presented which employs a combination of ontology-based collaborative filtering and neural networks to improve query expansion. Information about expanding the user query sc achieved from similar users, similarities with other users eœ reviewed and the most relevant web documents and their corresponding terms from these similar users‟ queries eœ acquired. External ontological sources including users‟ personal information is employed, some similar users eœ found so initial knowledge about all users and their interests can be provided from these external ontologies. After finding the similar users, and further constructs the training data of relevant documents retrieved by similar users, at last predict document relevancy and discover the most relevant web documents and their corresponding terms. The method can improve the precision and only requires users to provide less query information at the beginning.

Wang in [33] proposed an expansion model of semantic query based on ontology which is composed of four parts: user interface, query request handle module, retrieval module and results handle module. With the four parts work cooperatively, the model can better understand user‟s willingness and improve the recall and precision ratios of information retrieval. query request handle module can optimize the initial query, and make the query keywords more accuracy. It includes two parts: concept extraction module and query expansion module. Query expansion module is based on the keywords from concept extraction module to perform semantic expansion. The module expands the query keywords by using of the semantic relations and reasoning mechanism in ontology. It employs query expansion handle algorithm to extract the synonym relation, father relation, son relation and brother relation in ontology and generate pre-expansion words. It computes the relativity between expansion words and query words and puts the high relativity words as expansion words. Experimental results show that this method can improve the precision and recall ratios of web information retrieval and retrieval precision is highly improved when adding 6 expansion words.

In [27] a semantic based query expansion method is represented which has five steps: (i) preprocessing, (ii) clustering the query words, (iii) semantic network construction, (iv) performing spreading activation algorithm, and (v) filtering the candidate words. Spreading activation was proposed as a method to search. The search process is initiated by activation a set of source nodes and then their

activation iteratively spread out to other nodes linked to the source nodes until some termination specification is met. The constructed semantic network in the previous phase is as input for spreading activation algorithm. The weights of nods that are in original submitted query are assigned to „1" and weights of other nodes is assigned to „0". This approach can deal with vague words and is naturally robust to noise words and avoids the outweighting problem in query expansion.

III. Comparison

Stability of automatic query expansion is less than stability of interactive query expansion results. In this way, there is possibility of query deviation. However, in both manual and interactive approaches, this possibility is less due to the presence of user. But automatic approaches are better that manual approaches because they do not need user and his/her interfering in the process. Moreover users who use information retrieval systems do not have great tendency to participate in the process of information retrieval. Probabilistic methods consider lexical relation between terms and need a lot of computational operations so they are expensive. If the query was ambiguous they do nothing. In contrast ontologies consider semantic relation of terms and have preprocessing stage where the query is disambiguated. The key problem is how to use the knowledge of ontologies in query expansion. Global techniques extract their co-occurrence statistics from the whole document collection. The global analysis techniques are relatively robust but corpus wide statistical analysis consumes a considerable amount of computing resources. Moreover, since it only focuses on the document side and does not take into account the query side, global analysis cannot address the term mismatch problem well. Global methods need semantic similarity and terms disambiguation. In contrast local techniques extract their statistics from the top-n documents returned by an initial query. It only focuses on original query. Resource storing and reducing user interfere are the main advantages of these approaches. These methods are not robust, if a large fraction of the documents assumed relevant, it is actually non-relevant, then the words added to the query (drawn mostly from these documents) are likely to be unrelated to the topic and the quality of the documents retrieved using the expanded query is likely to be poor.

IV. Conclusion

Query expansion approaches are divided into two main types including probabilistic and ontology. Most probabilistic methods can be categorized as global or local. Global techniques extract their co-occurrence statistics from the whole document collection while local techniques extract their statistics from the top-n documents returned by an initial query. Ontology-based methods use semantic relations between terms to select appropriate expansion terms. Words ambiguity in user query is one of the information retrieval problems. Without knowing exact meaning of terms, retrieval systems return a lot of irrelevant documents which contain different meanings of terms and this leads to precision reduction. To overcome this problem various methods using ontologies have proposed. WordNet is one of the general ontologies which researchers presented approaches based on it for query expansion. These techniques were various and obtained results have different impacts in information retrieval performance. The method most commonly used is that first the query terms is disambiguated so that they map to a unique ontology concept. Then terms related in the ontology to the disambiguated concepts are added to the query.

References

[1] R. Attar and A. Fraenkel, "Local feedback in full-text retrieval systems," *ACM*, vol.24, pp. 397-417, July 1977.

[2] A. Abdelali, J. Cowie, and H. Soleiman, "Improving query precision using semantic expansion," Information Processing and Management, pp. 705-716, 2007.

[3] A. Abdelmgeid, "Using a query expansion technique to improve document retrieval," Information Technologies and Knowledge: An International Journal, vol. 2, 2008.

[4] A. Andreou, "Ontology and query expansion," Master of Science thesis, School of Informatics University of Edinburgh, 2005.

[5] B. Billerbeck, "Efficient query expansion," A Thesis Submitted for the Degree of Doctor of Philosophy, School of Computer Science and Information Technology, Portfolio of Science, Engineering and Technology, RMIT University, Melbourne, Victoria, Australia, 2005.

[6] W. N. Borst, "Construction of engineering ontologies for knowledge sharing and reuse," Ph.D Thesis Series, Dutch Graduate School for Information and Knowledge Systems, September 1997.

[7] H. Cui, J. Wen, J. Nie, and W. Ma, "Probablistic query expansion using user logs," Proceeding of the 11th Word Wide Web Conference, Honolulu, Hawaii, USA. ACM 1-58113-449-5/02/0005, pp. 325-332, 2002.

[8] E. A. Fox, "Extending the Boolean and vector space models of information retrieval with P-norm queries and multiple concept types," Ph.D Thesis, Cornell University, 1983.

[9] Z. Gong, C. Wa Cheang, and U. Hou, "Web query expansion by WordNet," In Processing of the 16th International Conference on Database and Expert System Applications, Copenhagen, Demark, pp. 166-175, 2005.

[10] Z. Gong, C. Wa Cheang, and L. Hou U, "Multi-term Web query expansion using WordNet," pp. 379-388, DEXA 2006.

[11] T. R. Gruber, "A translation approach to portable ontology specifications," Knowledge Acquisition, pp. 199–220, 1993.

[12] L. Han and G. Chen, "HQE: A hybrid method for query expansion," Expert Systems with Applications, pp. 7985–7991, 2009.

[13] I. Hazra and Sh. Aditi, "Thesaurus and query expansion," International Journal of Computer Science & Information Technology (IJCSIT), Vol. 1, No. 2, November 2009.

[14] M. Indrawan and S. Loke, "The impact of ontology on the performance of information retrieval: A case of WordNet," International Journal of Information Technology and Web Engineering, Vol. 3, Issue 1, 2008.

[15] Z. Jiang and Z. Yu, "A new technology of query expansion based on User Interest Model," IEEE 2010.

[16] Y. Jing and W. Bruce Croft, "An association thesaurus for information retrieval," In Proceedings of RIAO-94, 4th International Conference, pp. 146-160, New York, 1994.

[17] S. Jones and D. M. Jackson, "the use of automatically obtained keyword classifications for information retrieval," 1970.

[18] S. Liu, F. Liu, C. Yu, and W. Meng, "An effective approach to document retrieval via utilizing WordNet and recognizing phrases," In

Proceedings of the 27th Annual International ACM SIGIR Conference, Sheffield, pp. 266-272, July 2004.

[19] Y. Liu, C. Li, P. Zhang, and Z. Xiong, "A query expansion algorithm based on phrases semantic similarity," International Symposiums on Information Processing, 2008.

[20] L. Ma, L. Chen, Y. Gao, and Y. Yang, "Ontology based query expansion in Vertical search engine," 6th International Conference on Fuzzy Systems and Knowledge Discovery, 2009.

[21] CH. D. Manning, P. Raghavan, and H. Schutze, "An introduction to information retrieval,"œCambridge University Press, England, 2009.

[22] M. Mitra, A. Singhal, and C. Buckley, "Improving automatic query expansion," In Proceedings of the 2lst Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 98, Melbourne, Australia, Aug 1998.

[23] R. Navigli and P. Velardi, "An analysis of ontology-based query expansion strategies," Work Shop on Adaptive Text Extraction and Mining, In 14th European Conference on Machine Learning ECML, pp. 22–26, September 2003.

[24] Y. Qiu and H. P. Frei, "Concept based query expansion," In Proceeding of The ACM-SIGIR Intl, Conference on Research and Development in Information Retrieval, pp. 160-169, 1993.

[25] R. Richardson and A. F. Smeaton, "Using WordNet in a knowledge-based approache to information retrieval," Technical Report CA-0395, School of Computer Applications, Dubline City University, 1995.

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

[27] M. Shabanzadeh, M. A. Nematbakhsh, and N. Nematbakhsh, "A semantic based query expansion to search," International Conference on Intelligent Control and Information Processing, Dalian China, pp. 13-15, August 2010.

[28] A. F. Smeatone and C. Berrut, "Running TREC-4 experiments: A chronological report of query expansion experiments carried out as a part of TREC-4," Technical Report CA-2095, School of Computer Science, Dubline City University, 1995.

[29] M. Song, I. Y. Song, X. Hu, and R. B. Allen, "Integration of association rules and ontologies for semantic query expansion," Data Knowledge and Engineering, pp. 63-75, 2007.

[30] R. Studer and S. Staab, "Handbook on ontologies," International Handbooks on Information Systems, Springer Berlin Heidelberg, 2004.

[31] E. M. Voorhees, "Using WordNet to disambiguate word senses for text retrieval," In R. Korfhage, Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval Pittsburgh, New York ACM Press, pp. 171-180, July 1993.

[32] E. M. Voorhees, "Query expansion using lexical-semantic relations," Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland, New York Springer-Verlag, pp. 61-69, July 1994.

[33] H. Wang, J. Qin, and H. Shao, "Expansion model of semantic query based on ontology," School of Information Science and Engineering, Shenyang University of Technology, Shenyang, China, 2009.

[34] J. Xu and B. Croft, "Improving the effectiveness of information retrieval with Local Context Analysis,"o Computer Science Department, University of Massachusetts, ACM Transactions on Information Systems (TOIS), Vol. 18, Issue 1, Jan 2000.

[35] B. Zhang, Y. Du, H. Li, and Y. Wang, "Query expansion based on topics," Punlished in: Proceeding FSKD '08 Proceedings of the 15th International Conference on Fuzzy Systems and Knowledge Discovery, Vol. 02, 2008.