Journal of Theoretical and Applied Information Technology
10th November 2013. Vol. 57 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A SURVEY ON SEMANTIC QUERY EXPANSION

[1]R.JOTHILAKSHMI, [2]N.SHANTHI, [3]R.BABISARASWATHI

[1]Associate Professor, Department Of Information Technology, RMD Engineering College, TamilNadu.
[2]Professor and Head, Department Of Information Technology, KSR College of Technology, TamilNadu.
[3]Associate Professor, Department Of Computer Science &Engg., KSR College of Technology, TamilNadu.

E-mail: [1]rjothilakshmi@gmail.com

ABSTRACT

The ease of use of a large quantity of information source has spurred a great amount of attempt in the growth and enhancement of Information Retrieval techniques. Users' information needs are expressed in the form of natural language and keyword queries. The successful retrieval is very much dependent on the effective communication of the intended purpose. The relative ineffectiveness of information retrieval systems is mostly caused by the inaccuracy of a query formed by a few keywords. The well known method to overcome this limitation is the Query Expansion (QE), in which the user entered queries are augmented with new concepts with similar meaning. In this paper we discuss about various Query Expansion approaches with some of the case studies. Finally we discuss the issues related to Information Retrieval (IR) and further research directions towards the effective Information Retrieval.

Keywords: *Information Retrieval, Ontology, Query Expansion*

1. INTRODUCTION

Information Retrieval (IR) is the main task for the searcher in the field of World Wide Web since the content of the web is increasing day by day and dynamically. In order to retrieve the accurate (semantic) information the query which is given by the searcher is playing a very important role in IR. The current information systems are location finder rather than content finder. Even though now many systems are developed for content based retrieval still it is in research to increase the efficiency of the system in terms of precision and recall. Information space in online retrieval systems is relatively larger than conventional information retrieval systems and pooled with the uncertainty of the English language, a search query quite often results in a long list of results being returned, much of which are not always relevant to the user's information needs. The main difference in online retrieval systems and conventional information retrieval systems is that the former are usually web-based and as a result the document collection is more dynamic. To increase the number of relevant documents retrieved, queries need to be disambiguated by looking at their context. Query expansion techniques range from relevance feedback mechanisms to use of knowledge models such as ontologies to resolve ambiguities.

Normally Search engines expand the query by stemming the words, expanding the Acronyms and correcting the misspellings. Earlier query expansion techniques are based on statistics measures by counting the frequency of the term in the retrieved documents. To reformulate the query the important terms in the query are extracted and finding the relevant terms including synonym, polynum etc. To find the relevant terms to expand the query, this uses the computational linguistics, Information science etc. Currently the query expansion techniques are based on semantic resources, local analysis, global document analysis, Query log analysis, association rules, and neural networks.

Current search engines are based on keyword indexing systems and Boolean logic queries. Keyword list depicts the contents of documents or information objects but do not state about relations (either semantic or statistical) between keywords. The main problem with keyword-based information retrieval is that it performs string matching and thus it does not take into account the correlation between words or phrases. In lieu of this, much research efforts have emerged in the development of concept-based information retrieval. With concept based information retrieval, the search for information is based on the meaning of words or phrases rather than merely the presence of keywords within the index. Typically, a query that specifies the intent of a user's search goal is

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

provided to a retrieval system (e.g. web search engines, specialized search systems etc) with the hopes of obtaining the most relevant set of results for an information need. Given the large information space in online retrieval systems as well as the fact that a lot of valuable information is lost in translation when users try to represent their actual information need in the form of a query (also coined as the intention gap), a search query quite often results in a long list of results being returned and this requires a tedious evaluation of the huge amount of potential resources obtained.

By stemming a user-entered term, more documents are matched, as the alternate word forms for a user entered term are matched as well, increasing the total recall. This comes at the expense of reducing the precision. By expanding a search query to search for the synonyms of a user entered term, the recall is also increased at the expense of precision. This is due to the nature of the equation of how precision is calculated, in that a larger recall implicitly causes a decrease in precision, given that factors of recall are part of the denominator. It is also inferred that a larger recall negatively impacts overall search result quality, given that many users do not want more results to comb through, regardless of the precision.

The goal of query expansion in this regard is by increasing recall, precision can potentially increase (rather than decrease as mathematically equated), by including in the result set pages which are more relevant (of higher quality), or at least equally relevant. Pages which would not be included in the result set, which have the potential to be more relevant to the user's desired query, are included, and without query expansion would not have, regardless of relevance. At the same time, many of the current commercial search engines use word frequency (Tf-idf) to assist in ranking. By ranking the occurrences of both the user entered words and synonyms and alternate morphological forms, documents with a higher density (high frequency and close proximity) tend to migrate higher up in the search results, leading to a higher quality of the search results near the top of the results, despite the larger recall.

Research on automatic query expansion (or modification) was already under way before 1960 when initial requests were enlarged on the grounds of statistical evidence .The idea was to obtain additional relevant documents through expanded queries based on the co-occurrence of terms. Currently the idea has to incorporate new expansion techniques to retrieve documents by expanding the query semantically to improve precision and recall. The significant issue for retrieval effectiveness is the term mismatch problem. The indexers and the users do often not use the same words [20]. This is known as the vocabulary problem, compounded by synonymy (same word with different meanings, such as ’bank‘) and polysemy (different words with the same or similar meanings, such as ’car ‘and ’automobile‘). Synonymy, together with word inflections (such as with plural forms, “television” versus “televisions”), may result in a failure to retrieve relevant documents, with a decrease in recall (i.e., the ability of the system to retrieve all relevant documents). Polysemy may cause retrieval of erroneous or irrelevant documents, thus implying a decrease in precision(i.e., the ability of the system to retrieve only relevant documents).To overcome the vocabulary problem, various methods have been proposed, such as interactive query refinement, relevance feedback, word sense disambiguation, and search results clustering. The most likely and successful technique is to expand the original query with additional words that best capture the actual user goal, or to generate useful query to retrieve relevant documents.

To deal with the vocabulary problem, numerous approach have been proposed including interactive query refinement, relevance feedback, word sense disambiguation, and search results clustering. One of the mainly usual and doing well technique is to expand the original query with other words that best capture the tangible user intention, or that simply produce a more useful query to retrieve relevant documents.

The rest of the paper is organized as, chapter 2 describes QE, chapter 3 describes approaches to QE, and chapter 4 describes issues related to Information Retrieval and chapter 5 as conclusion.

2. QUERY EXPANSION

Query expansion (QE) is the process of reformulating a seed query to improve retrieval performance in information retrieval. The types of query expansion can be classified as Manually, Automatically and interactively. Without detailed knowledge of the collection make-up and of the retrieval environment, most users find it difficult to formulate queries which are well designed for retrieval purposes. The users might to spend huge amount of time to reformulate the queries to accomplish the effective retrieval. To deal this

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

difficulty, first query reformulation should be treated as an initial attempt to retrieve the relevant information. Following that, the documents initially retrieved could be examined for relevance and new improved query formulation could be constructed to retrieve additional useful documents. The query reformulation consists of two basic steps: 1) expanding the original query with new terms and 2) reweighting the terms in the expanded query. The query expansion technique may also uses the standard thesauri and/or domain specific thesauri for adding new terms. The overall design of the semantic query expansion is shown in figure:1, which is placed after the reference section of the paper.

2.1 Problems Commonly Arise In Web Queries

2.1.1 Short term and long term query

The short terms queries lead to more ambiguities than the long terms query. The ambiguity comes from the terms in the initial query or how the terms are related in knowledge repository (ontology).

2.1.2 Intentional and extensional query

From the query terms we have to identify whether the user is interested additional knowledge or the specific knowledge of the terms in the query.

2.1.3 Disambiguation

The Documents that deal with the same domain can use different terms for describing the same concepts.

2.1.4 Identifying the relevant results

It may be difficult to detect whether a given result of the query is valid. A result is valid if it belongs to the expected result. The problem is that the expected result is in the mind of the user. A result is also valid if it belongs to the intersection of the intended domains.

2.2 Problem in Query Expansion

The added (after expansion) terms, may cause query drift - the change of the focus of a search topic caused by improper expansion [27] – thus hurting precision. The main cause be when an expansion term is correlated with a single term of the original query rather than with the entire query it may easily match unrelated concepts. Sometimes, QE achieves better precision in the sense that it has the effect of moving the results toward the most popular or represented meaning of the query in the collection at hand and away from other meanings; e.g., when the features used for QE are extracted from web pages [30], or when the general concept terms in a query are substituted by a set of specific concept terms present in the corpus that co-occur with the query concept [28]. QE is also useful for improving precision when it is required that several aspects (or dimensions) of a query must be present at once in a relevant document. This is another facet of query disambiguation, in which query expansion can enhance those aspects that are underrepresented in the original user query ([29], [31]).

2.2 Limitations and Assumptions

2.2.1 Information need and query type

An important aspect for consideration while dealing with information retrieval is the information need expressed by users. Usually the users express their need thro query which consists of one or more keywords. We have to classify whether the user is *searching or querying* for information of their interest .Search goals can be generally classified as exploratory, information lookup (fact-finding) and comprehensive [25]. In exploratory search, the users are interested in a general idea of a topic where high recall or precision is not required but fairly having an only some documents as a reference is adequate. In fact-finding search, the precision of the results returned is very important as compared to comprehensive search in which high recall is of greater importance. An information need, whether exploratory, comprehensive or fact-finding, is articulated in the form of informational, navigational and transactional queries. Though, given that these are natural language queries, they may differ in language of their structure and length. Users may devise natural language queries which could be grammatically correct or incorrect and may vary in terms of its length. The term mismatch problem occurs since users may not aware of the terms used in the retrieval system (terms used by the user may not match with terms used in the document or index)

2.2.2 Query structure and languages

Informational, navigational or transactional queries submitted by users are structured in a variety of ways. It may be contain

Journal of Theoretical and Applied Information Technology
10th November 2013. Vol. 57 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

some bag of words (keywords) or complete sentence which represent the information of their need. Apart from the structure, several studies have been conducted to examine the language of the query, [jansen et all] conducted a study over a Web log with 1.5M queries, and found that 2.1% of the queries contained Boolean operators,7.6% contained other query syntax, primarily double-quotation marks for phrases.[White *et al]* examined interaction logs of nearly 600,000 users, and found that 1.1% of the queries contained one or more operators,8.7% of the users used an operator at any time. The keyword based queries are broadly classified as single word queries, context queries (phrase, proximity), Boolean queries and Natural Language queries. The short term query is having more of word disambiguation problem than the long term queries. However, long queries may not always have done well with direct query expansion as such queries may consist of multiple important concepts. A verbose query may not always contain explicit information in the description itself to indicate which of these concepts is more important. Secondly, such queries may contain two or more equally essential key concepts. We believe it is thus vital to identify the key concepts expressed in a query to ensure the main search intent is emphasized.

3. APPROACHES OF QUERY EXPANSION

There are variety of approaches for improving the initial query formulation through query expansion and term reweighting. These approaches are grouped into three main categories: 1) Based on user feedback 2) based on local analysis (from initially retrieved documents) 3) Global analysis (global information such as thesaurus and Ontologies). The detailed taxonomy of query expansion approaches are shown in figure 2 which is placed after the reference section of the paper. In this section we provide the detailed characteristics of the approaches along with case studies.

3.1 Relevance Feedback

Relevance feedback takes the outcome that are primarily returned from a given query and uses information provided by the user about whether or not those results are relevant to perform a new query. The content of the assessed documents is used to alter the weights of terms in the original query and/or to add words to the query. Relevance feedback effectively reinforces the system's original decision, by making the expanded query more analogous to the retrieved relevant documents.

The specific data source from which the expansion features are generated using relevance feedback may be more reliable than the sources generally used by query expansion, but the user must assess the relevance of the documents. The expected effect is that the new query will be moved towards the relevant documents and away from the non-relevant ones.

The advantages of relevance feedback approaches are 1)The user has to provide the relevance judgment on initially retrieved documents 2) breaks down the search process into small steps 3) it provides a controlled process to highlight some terms(which are relevant) and de- highlight other terms(which are non-relevant)

In pseudo relevance feedback where the top ranked ***n*** documents are assumed to be relevant. Terms from these documents are selected and used for expanding the query. In either of the approaches the term selection method is a key factor in the performance of expanded queries. It needs to consider how to weight the new terms, whether to exclude the original query terms, whether to include all of the new terms or just some of them and if so how many new terms to include.

3.1.1 Case studies

[2] It is based on Interests retention models in which the User interests can be described as a set of concepts that users are familiar with. Interests retentions may be related to user recent interests and possible queries from users. Based on the hierarchical granular structure and inspired by the basic level for human problem-solving, they defined a *starting point, SP* that consists of a user identity (e.g. a user name, a URI, etc.) and a set of nodes that serve as the background for the user (e.g. user interests, friends of the user, or other user familiar or related information)

[16]In this paper, they proposed a novel relevance feedback method called Translation Enhancement (TE), which uses the extracted translation relationships from relevant documents to revise the translation probabilities of query terms and to identify extra available translation alternatives so that the translated queries are more tuned to the current search.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3.2 Linguistics Analysis

In order to decipher the exact intent of a query, queries may be scrutinized based on their linguistic characteristics. However, a full study of natural language query characteristics in terms of morphological, syntactical and semantic properties can be a challenging task as there is much depth involved. Thus, through a survey of computational linguistics studies, three major linguistic properties of queries (i.e. morphological, syntactical and semantics) that can be extracted for the purpose of query expansion are identified.

3.2.1 Morphological analysis

Morphology is the identification, analysis and description of the structure of a word in a given document such as root words, affixes, parts of speech etc. The lexeme and lemma of each word in the documents are identified. Individual words are analyzed into their components and non-word tokens such as punctuation are separated from the words. This process will usually assign syntactic categories to all the words in the sentence.

3.2.2 Syntactic analysis

Linear sequences of words are transformed into structures that show how the words relate to each other. Syntactic analysis must exploit the results of morphological analysis to build a structural description of the sentence. The goal of this process, called parsing, is to convert the flat list of words that forms the sentence into a structure that defines the units that are represented by that flat list. The important thing here is that a flat sentence has been converted into a hierarchical structure and that the structure corresponds to meaning units when semantic analysis is performed.

3.2.3 Semantic analysis

The structures created by the syntactic analyzer are assigned meanings. It must map individual words into appropriate objects in the knowledge base or database. It must create the correct structures to correspond to the way the meanings of the individual words combine with each other. Lexical ambiguity can be addressed by algorithmic methods that automatically associate the appropriate meaning with a word in context, a task referred to as word sense disambiguation.

3.2.4 Case studies

[5]They used **MDL(**Minimum description Logic) which minimizes grammar length and data compression length. The words are separated by removing the affixes, and Part of Speech tags are identified by an unsupervised learner of morphology to bootstrap the acquisition of lexical categories.

[6] In their work the morphologically related word forms are identified by stemming and lemmatization. The fundamental idea is to detect terms for query expansion, within a collection of texts, which have morphological relations with the terms of the query.

[7] They presented a novel syntactically-based query reformulation (SQR) technique, which is based on shallow syntactic evidence induced from various language samples. Then the performance of the system is evaluated by combining with pseudo relevance technique.

[12] They have proposed a simple and original technique, relying on an analogy-based learning process, able to automatically detect morphological variants within documents and use them to expand query terms.

3.3 Corpus Dependent

In this approach the documents are analyzed to select the suitable concepts which are similar to the terms in the query. The term selection and term clustering are the two methods which applied in corpus dependent approach.

3.3.1 Case studies

[3]The BIC-aiNet algorithm as the biclustering task can be viewed as a multimodal combinatorial optimization problem, in this work we have considered an immune-inspired algorithm to accomplish this task, denoted BIC-aiNet (Artificial Immune Network for Biclustering),

[13]They proposed a novel concept-based query expansion technique, which allows disambiguating queries submitted to search engines. The concepts are extracted by analyzing and locating cycles in a special type of query relations graph. This is a directed graph built from query relations mined using association rules. The concepts related to the current query are then shown to the user who selects the one concept that he interprets is most related to his query. This concept is used to expand the original query and the expanded query is processed instead.

[14] They proposed a new technique based on conceptual semantic theories, in contrast to the structuralist semantic theories upon which other techniques are based. The source of the query expansion information is the concept network

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

knowledge base. Query terms are matched to those contained in the concept network, from which concepts are deduced and additional query terms are selected

[17] They designed a mechanism that can automatically select corpus words that are semantically related to the query, in the expansion process. The new expansion process is guided by the semantics relations between the original query and the expanding words, in the context of the utilized corpus. The major advantage of our approach is its well thought out mathematical approach in selecting the query expanding words/documents from the corpus. Only corpus words with the largest sense vector similarity (within some ‘‘cosine’’ threshold) to the sense vector of the ‘‘entire’’ initial query will be selected for the expansion process.

[19]First, they propose a novel query expansion mechanism on fields by combining field evidence available in corpora. Second, they propose an adaptive query expansion mechanism that selects an appropriate collection resource, either the local collection, or a high-quality external resource, for query expansion on a per-query basis. Combining field evidence can refine the statistics for query term reweighting, and improve query expansion. Moreover, we have devised an adaptive query expansion mechanism that automatically selects the appropriate collection resource for query expansion. The proposed adaptive mechanism applies query performance prediction, selective query expansion, and collection enrichment techniques.

3.4 External Knowledge Sources Based Expansion

The external knowledge based sources like Ontology, Word Net, Query logs, Wikipedia etc, can be used to expand the query by solving the semantic ambiguity.

3.4.1 Word sense disambiguation

Word sense disambiguation (WSD) is natural language processing, which governs the process of identifying which sense of a word (i.e. meaning) is used in a sentence, when the word has multiple meaning. A disambiguation process requires two strict things: a dictionary to specify the senses which are to be disambiguated and a corpus of language data to be disambiguated

3.4.2 Thesaurus

A thesaurus is a reference work that lists words grouped together according to similarity of meaning (containing synonyms and sometimes antonyms), in contrast to a dictionary, which contains definitions and pronunciations. In information technology, a thesaurus represents a database or list of semantically orthogonal topical search keys. Terms are the basic semantic units for conveying concepts. They are usually single-word nouns, since nouns are the most concrete part of speech. "Term relationships" are links between terms. *Hierarchical* relationships are used to indicate terms which are narrower and broader in scope. The *equivalency* relationship is used primarily to connect synonyms and near-synonyms. *Associative* relationships are used to connect two related terms whose relationship is neither hierarchical nor equivalent

3.4.3 Ontology

Ontology formally represents knowledge as a set of concepts within a domain, and the relationships between pairs of concepts. It can be used to model a domain and support reasoning about entities. The main components are Individuals, classes, attributes, relations, rules, axioms and events. A domain ontology (or domain-specific ontology) models a specific domain, which represents part of the world. Particular meanings of terms applied to that domain are provided by domain ontology. An upper ontology (or foundation ontology) is a model of the common objects that are generally applicable across a wide range of domain ontologies. It employs a core glossary that contains the terms and associated object descriptions as they are used in various relevant domain sets

3.4.4 Case studies

[8]They presented a method of automatic query expansion using a collection dependent thesaurus built with probabilistic latent semantic analysis. They showed how to build the thesaurus using probabilistic latent semantic term relationships in their work.

[10] In this paper, we describe our query expansion approach sub-mitted for the Semantic Enrichment task in Cultural Heritage in CLEF (CHiC) 2012. They make use of an external knowledge base such as Wikipedia and DBpedia. It consists of two major steps, concept candidate’s generation from knowledge bases and the selection of K-best related concepts. For selecting the K-best concepts, they ranked them according to their semantic relatedness with the query. They used Wikipedia-based Explicit Semantic Analysis to calculate the semantic relatedness scores.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[11] The integration of semantic closeness measures into the matching function yields ranked results for multi-concept queries, where the ranking is based on conceptual distance in the thesaurus. This faceted query expansion over thesaurus relationships has potential to reduce the vocabulary problem associated with highly specific queries and indexing descriptors. They discussed the potential of query expansion techniques using the semantic relationships in a faceted thesaurus. An extract of the National Museum of Science and Industry's collections database, indexed with the Getty Art and Architecture Thesaurus (AAT), was the dataset for the research.

[22] They used WordNet Lexical Chains and WordNet semantic similarity to assign terms in the same query into different groups with respect to their semantic similarities. For each group, they expand the highest terms in the WordNet hierarchies with *Hypernym* and *Synonym*, the lowest terms with Hyponym and Synonym, and all other terms with only Synonym.To determine expansion dimensions for each word in the same group, find their relative positions in the WordNet hierarchies.

[23] In this approach the query expansion uses the lexical databases and user feedback to improve the Boolean search query.

[32]It uses context ontologies for identifying domain semantics to specify and expand(refine) web queries. The context of the web query is established by comparing context ontologies against search terms given by the user.

[33] presents query exapnsion method based on linguistic and semantic knowledge.The methodology uses semantic knowledge(ResearchCyc) and linguistic knowledge(WordNet) to expand the query terms and refine the query for user intensioanl one.

3.5 Query log Based Expansion

In this approach the documents are analyzed based on the previous query log which was given by the user.

3.5.1 Case studies

[20] They proposed a new method of obtaining expansion terms, based on selecting terms from past user queries that are associated with documents in the collection. First they submit a query to the search system to get the top *R* answer documents. From this initial retrieval run, it is possible to identify a set of candidate expansion terms; these may be based on the top *R* documents, or surrogates corresponding to these documents. Then the top *E* expansion terms are selected, using Robertson and Walker's term selection value formula. Finally, selected terms are appended to the original query, which is then run against the target text collection.

[21]They presented a novel method for automatic query expansion based on the query logs. The main thrust of this method is to establish probabilistic correlations between query terms and document terms through mining the query logs, which is an effective way to bridge the gap between the query space and the document space. Then for the newly coming queries, high-quality expansion terms can be selected from the document space on the basis of these probabilistic correlations

3.6 Association Rule Mining Based Expansion

In this approach the terms to expand are identified by using the Association Rule Mining method. The most frequent terms are identified by applying the association rules.

3.6.1 Case studies

[1] They proposed a new Minimal Generic Basis, called *MGB*, for only retaining irredundant association rules. The design of this basis relies on the Formal Concept Analysis (FCA) mathematical settings.

[34]In their work, the contextual properties of important terms discovered by association rules, and ontology entries are added to the query by disambiguating word senses. Applying a hybrid querying expansion algorithm that combines association rules and ontologies to derive queries targeted to match and retrieve additional documents similar to the positive examples.

3.7 Other Types of Expansion

The query expansion approach utilizes the Conditional Random Field and Hidden Markov Model to find the hidden concept for expanding the terms in the query.

3.7.1 Case studies

[15] They proposed a robust query expansion technique based on the Markov random field model for information retrieval. The technique, called latent concept expansion, provides a mechanism for modeling term dependencies during expansion. The latent concept expansion technique captures two semi orthogonal types of dependence as (i) syntactic dependence which covers phrases, term proximity, and term co-occurrence .These methods capture the fact that queries implicitly or explicitly

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

impose a certain set of positional dependencies and semantic dependency.

[18] They described a Markov chain framework that combines multiple sources of knowledge on term. A Markov chain model allows chaining of multiple inference steps with different link types to perform "semantic smoothing" on language models.

4. ISSUES IN INFORMATION RETRIEVAL (IR)

4.1 Data Cleansing

Data cleansing is the process of removing the noise or incorrect information from the document. It is the very first and significant task in IR, because if the document contains inconsistent or incorrect values, and the data is collected form heterogeneous formats then the performance of the system degrades.

4.2 Indexing and Querying

Indexing is the process of classifying or summarizing the document by index terms to increase the search process by IR system. The main issues related to indexing are the selection of the index terms and representation of the index by choosing the appropriate structure. Usually queriest gives the query consists of set of keywords, which may differ from the terms used in the document collection. To solve this difficulty, QE plays an vital role to increase the performance of the system. At rest QE depends of further factors like lack of domain ontology and similarity measures.

4.3 Document Ranking

Ranking is the central part of many IR problems such as document retrieval, collaborative filtering etc. Usually the IR systems rank the retrieved documents which are relevant to the information need. In future we expect the system with good feature selection techniques to rank the documents.

4.4 Document Classification and Clustering

The process of analyzing the set of documents and classify them in to some predefined classes. The documents with same properties are tending to closer in same cluster. In future we need the distributed system with efficient techniques to cluster the documents since the nature of dynamic content evolves in the digital world.

4.5 Document Visualization

The process of delivering the information in the document to the user is a critical one. The documents are represented after analyzing the metadata and type of text present in the document.

5. CONCLUSION

In this study we have analyzed the various QE approaches. The Word Sense Disambiguation techniques and Ontology (Upper or Domain) plays a vital role to expand the user queries to increase the precision and recall of the system. But, people still find it difficult with current Information Retrieval (IR) system, to retrieve information relevant to their information needs. Thus, in the dynamic world of the Web, large digital libraries, Question Answering system and of some repositories, people will have a question of, which technique will allow a faster and high accurate quality? With increasing demand for access, the quick response is becoming more and more critical factor. Also the quality of the IR system is greatly affected by the user interaction with the system. Based on the above issues in future the researcher has to develop an IR system with fast indexing and quick query response also the study of user behavior is important since it affect the design of IR system. The IR system should be designed with deployment of new strategies to convene the requirements of searchers in future.

REFERENCES

[1].Chiraz Latiri · Hatem Haddad · Tarek Hamrouni, "Towards an effective automatic query expansion process using an association rule mining approach",*J Intell Inf Syst*,

[2]. Yi Zeng · Ning Zhong · Yan Wang et all, "User-centric query refinement and processing using granularity-based strategies", *Knowl Inf Syst*,2011,vol- 27,pp:419–450

[3]. Pablo A. D. de Castro Æ Fabrı´cio O. de Franc¸a, "Query expansion using an immune-inspired biclustering algorithm" , Nat Comput,2010,vol- 9,pp:579–602

[4]. Bhawani Selvaretnam ·Mohammed Belkhatir, " Natural language technology and query expansion :issues, state-of-the-art and perspectives", *J Intell Inf Syst* ,2011.

[5]. John Goldsmith, Yu Hu and Irina Matveeva, "Using Morphology and Syntax Together in Unsupervised Learning, *Proceedings of the Second Workshop on Psychocomputational*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Models of Human Language Acquisition, pp: 20–27,

[6]. Moreau, F., Claveau, V., & Sébillot, P.” Automatic Morphological Query Expansion Using Analogy-Based Machine Learning”. In *Paper presented at the 29th european conference on IRresearch* ,2007,pp. 222–233

7. C. Lioma , I. Ounis,” A syntactically-based query reformulation technique for information retrieval”, *Information Processing and Management*,2008,vol-44,pp: 143–162

8. Laurence A. F. Park and Kotagiri Ramamohanarao, “Query expansion using a collection dependent probabilistic latent semantic thesaurus” *Advances in Knowledge Discovery and Data Mining, Lecture Notes in Computer Science, Volume* 4426, 2007, pp 224-235

9. Bai, j., song, d., bruza, p. D., nie, j. Y. And cao, g.,:Query expansion using term relationships in language models for information retrieval”. *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*”, 31 October – 5 November 2005, pp. 688-695.

10. Nitish Aggarwal and Paul Buitelaar,” Query Expansion using Wikipedia and DBpedia” *Workshop at CLEF: Conference and Labs of the Evaluation Forum Information Access*,2012

11. Douglas Tudhope, Ceri Binding, Dorothee Blocks, Daniel Cunliffe,“ Query expansion via conceptual distance in thesaurus indexed collections”,*journal of Documentation,*2006.

12. Fabienne Moreau, Vincent Claveau, and Pascale Sebillot, “Automatic morphological query expansion using analogy-based machine learning”, *journal of documentation*, 2006,vol-62,pp:509-533.

13. Bruno M. Fonseca,, Paulo Golgher, “ConceptBased Interactive Query Expansion”, *CIKM’05*, October 31–November 5, 2005.

14. Orland Hoeber, Xue-Dong Yang, and Yiyu Yao, “Conceptual Query Expansion” *Advances in Web IntelligenceLecture Notes in Computer Science*,Volume 3528, 2005, pp 190-196

15. Donald Metzler, W. Bruce Croft, “Latent Concept Expansion Using Markov Random Fields” *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* ,2007, pp: 311-318

16. Daqing He, Dan Wu, “Enhancing query translation with relevance feedback in translingual information retrieval”, *Information Processing and Management* ,2010,vol-47,pp: 1–17

17. Ahmed Abdelali , Jim Cowie , Hamdy S. Soliman, “Improving query precision using semantic expansion”, *Information Processing and Management*, 2007,vol-43,pp: 705–716

18. Kevyn Collins-Thompson Jamie Callan, Kevyn Collins-Thompson Jamie, **“**Query expansion using random walk models “,*CIKM '05 Proceedings of the 14th ACM international conference on Information and knowledge management* ,pp: 704-711

19. Ben He , Iadh Ounis, “Combining fields for query expansion and adaptive query expansion”, *Information Processing and Management* ,2007,vol-43,pp: 1294–1307

20. Bodo Billerbeck Falk Scholer Hugh E. Williams Justin Zobel, “Query Expansion using Associated Queries”, *CIKM’03,* November 3–8, 2003.

21. Hang Cui, Ji-Rong Wen, “Probabilistic Query Expansion using Query Logs”, *WWW 2002*, May 7-11, 2002..

22. Zhiguo Gong, Chan Wa Cheang, and Leong Hou U, “Multi-term Web Query Expansion Using WordNet”, *LNCS 4080*,2006,pp. 379 – 388

23. Erich Schweighofer, Anton Geist, “Legal Query Expansion using Ontologies and Relevance Feedback”, *Proceedings of LOAIT 07*, 2007,pp.149-159.

24. http://en.wikipedia.org/

25. Ricardo Baeza-Yates, “Modern Information Retrieval” Pearson publication.

26. http://google.com,Google search engine

27.] Mitra, M., Singhal, A., and Buckley, C,”. Improving automatic query expansion”, *In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* Pp: 206–214.

28. Chu, W. W., Liu, Z., and Mao, W,” Textual Document Indexing and Retrieval via Knowledge Sources and Data Mining”, *Communication of the Institute of Information and Computing Machinery(CIICM) 5*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

29. Crabtree, D., Andreae, P., and Gao, X.," Exploiting underrepresented query aspects for automatic query expansion", *In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*,pp: 191–200.

[30]Cui, H., Wen, J.-R., Nie, J.-Y., and Ma, W.-Y. 2003," Query expansion by mining user logs", *IEEE Transactions on Knowledge and Data Engineering* ,vol-15, issue-4, pp:829–839

[31]Jelinek, F., 1969," A fast sequential decoding algorithm using a stack", *IBM journal or research and development* 13, pp:675–685

[32] Roger H.L. Chiang, Cecil Eng Huang Chua, Veda C. Storey, "A smart web query method for semantic retrieval of web data", *Data&Knowledge Engineering*, 2001, vol-38, pp:63-84.

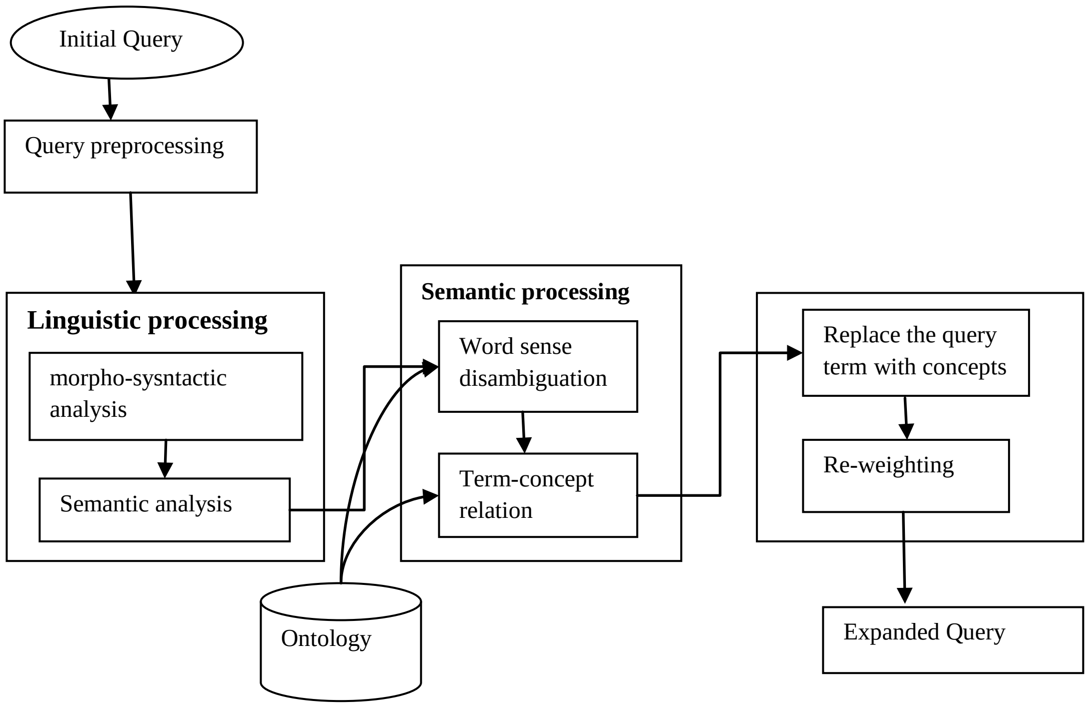
[33]Jorddi Consea, Veda C.Storey, Vijayan Sugumaran, "Improving Query processing through semantic knowledge", *Data&Knowledge Engineering*, 2008, vol-66, pp:18-34.

[34]. Min Song,II-Yeol Song et all, "Integration of association rules and ontologies for semantic query expansion", *Data&Knowledge Engineering*, 2007, vol-63, pp:63-75.

[35] J. Bhogal b, A. Macfarlane P. Smith, "A review of ontology based query expansion", *Information Processing and Management*,2007, vol- 43 pp:866–886

[36] Claudio Carpineto & Giovanni Romano, "A Survey of Automatic Query Expansion in Information Retrieval", *ACM Computing Surveys*, Volume 44, Issue 1, January 2012.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195



.

Figure 1: *Overall System Design for Semantic Query Expansion*

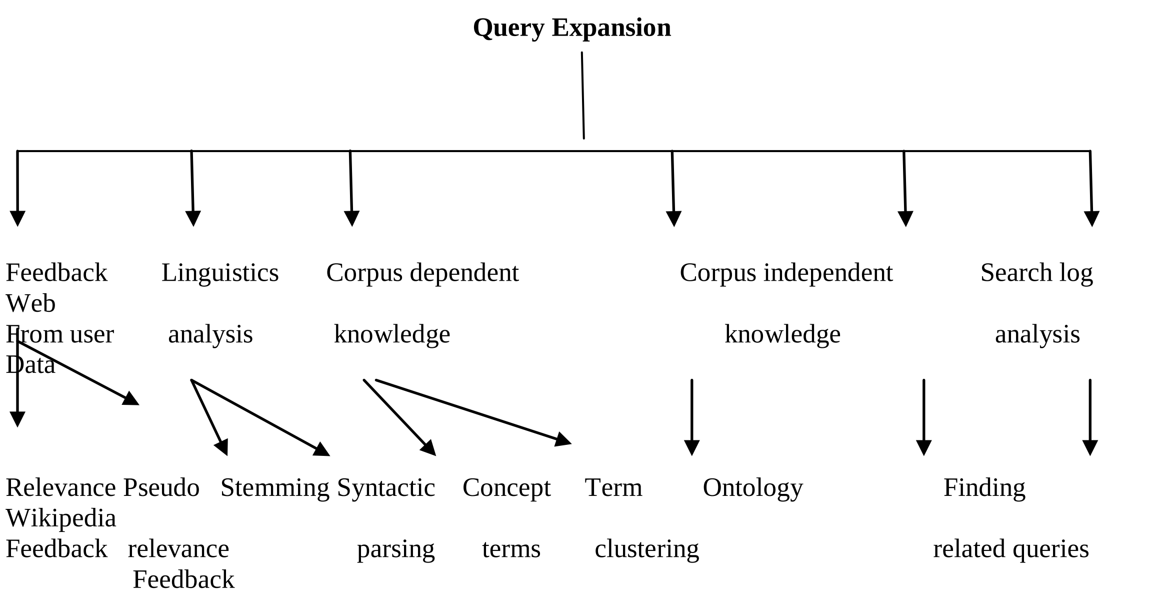


Figure 2: *Approaches of Query Expansion*