# Conceptual Weighing Query Expansion based on User Profiles

Amita Jain
Assistant Professor,
Computer Science &
Engineering, Ambedkar Institute
of Advanced Communication,
Technology and Research,
Geeta Colony, Delhi

Kanika Mittal
M.Tech (Information Security),
Ambedkar Institute of Advanced
Communication, Technology
and Research, Geeta Colony,
Delhi

Smita Sabharwal
M.Tech (Information Security),
Ambedkar Institute of Advanced
Communication, Technology
and Research, Geeta Colony,
Delhi

## ABSTRACT

Proper query terms significantly affect the performance of information retrieval systems. In this paper, a conceptual weighting method for query expansion is proposed with the help of user profile. Here, the users' initial queries and the retrieved documents based on the user's query (top n relevant documents) are analyzed and then the relevant terms from the documents retrieved are weighted. The terms having higher weight and the terms from the previous searches with a greater threshold weight will be selected and are used to derive the concepts in the concept network which are then connected to the phrases. Based on the matching of those phrases with that of the query phrases, additional query terms are selected and based on those additional query terms, the user's original query is expanded and the search is enhanced.

## Keywords

Natural Language Processing, Query Expansion, Term-weighting, concept networks, user profiles, information retrieval

## 1. INTRODUCTION

Natural Language Processing (NLP) is an area of research that describes how computers can be utilized to understand and manipulate natural language text or speech to do useful things. Huge amount of knowledge has been gathered by NLP researchers on how human beings understand and interpret the language, which in turn help in developing tools and techniques to make computer systems understand and manipulate natural languages to perform the desired tasks [26]. It is involved with the development of computational models of aspects of human language processing and the reasons for such development are: a) Language Processing tool development and b) Better understanding of human communication [27]. Some of the major fields of study of NLP include text summarization, information extraction, information retrieval, machine translation etc. As number of challenges is faced by user while searching for a relevant text document, so NLP plays an important role in the field of information retrieval for retrieving the information according to user's needs and requirements.

Information retrieval (IR) is meant for addressing the problems of storage, retrieval and evaluation of the content from those documents which are relevant to user's query. Unlikely database systems which work on highly structured data, IR systems work with unstructured natural text also.

Query expansion is the process of adding additional terms to the original query in order to improve retrieval performance [1], [29]. Generally the queries input by users contain terms that do not match the terms used to index the majority of the relevant documents (either controlled or full text indexing) and sometime the un-retrieved relevant documents are indexed by a different set of terms than those in the query or in most of the other relevant documents. In order to solve this problem and to improve the query performance, it is necessary to modify the user's query. So in order to modify the query, researchers have proposed query expansion to help the user to formulate what information is actually needed. Through query expansion, the effects of the word mismatch problem are reduced which is a result of different terms being used in reference to a single concept, both in the documents and in the user queries [2]. Query reformulation can be done to formulate the initial query through query expansion and term weighting. The query reformulation involves two basic steps: expanding the initial query with new terms and weighting the terms in the expanded query. These approaches are grouped in three categories [3]: (1) approaches based on feedback information from the user, (2) approaches based on global information from the document collection and, (3) Pseudo relevance feedback (PRF) approaches based on information derived from the set of initially retrieved documents.

The construction of user profiles plays an important role as different users have different perceptions about the information required. Due to abundant information available to the wide spectrum of users, most of the times, the users are not clear about what information they actually needed but has a rough idea about the documents he want to search. In this case, user profiles are efficient in a way that they store user preferences which gives the system an opportunity to provide more personalized search [24].

The main approach is to use a concept network to generate a conceptual query expansion for web information retrieval [4]. A graph consisting of concepts and phrases is defined as a concept network. Within the concept network, the degree to which the phrase has shown to be a good indicator of the concept weighted is represented by edges between the concepts and the phrases. The statistical analysis of concept-phrase co-occurrence in a concept hierarchy such as the Open Directory Project can be used to create a concept network. Query expansion is performed by selecting additional phrases from those that are connected to the same concepts as the user's query phrases.

Generally the user may not be clear about the exact usage of words in this query so as to obtain the necessary result. Therefore he might use some irrelevant words to represent his query which may lessen the chances of retrieving the required documents. In order to improve this problem, query re-weighing [9] plays an important role in finding out the relevant query terms and separating them from irrelevant ones to give more appropriate documents.

Through this paper, a technique is derived to improve the performance of query expansion using concept networks combined with weighting technique for finding out the new phrases which can be matched with the user's initial query phrases. The relevant query phrases are obtained by weighting technique and selecting the term with higher weight which in turn can be matched with the phrases connected to the concepts and finally the union of the original query terms and those selected phrases from concept network will provide the additional query terms for the expanded query.

The rest of the paper is organized as follows: In Section 2, the related work is mentioned, in section 3, we will discuss about the basic methodology used, in section 4, the proposed query expansion technique is explained. In section 5, we conclude our research.

## 2. RELATED WORK

Huge amount of work has been done in the field of natural language processing and information retrieval in order to gain a better understanding of human communication which in turn leads to user satisfaction. Siddiqui and Tiwari [27] explained the importance of NLP for information retrieval and explained various aspects of NLP like language modeling, syntactic analysis, semantic analysis etc. and said that NLP is associated with computational aspects of human language processing.

There are many works done in the field of automatic query reformulation that improves initial queries through query expansion and term weighting without user intervention [30]. The methods like automatic query reformulation do not rely on users to make relevance judgments [17]. They are often based on concept based retrieval [13], language analysis [12], term co-occurrences, PRF [3].

Bodner and Song [12] said that when there is a need for deep understanding of queries and documents, it always require huge computational cost for language analysis approaches As it is nearly difficult to evaluate, there are many things to variant such as pseudo relevance feedback (PRF). In PRF, only those documents which are on the top of the list are considered as relevant. This procedure has been found to be highly effective in some settings, most likely those in which the original query statement are long and precise [3].

Concept based retrieval is another related work in which query words are treated as concepts but not as literal strings of letters, and then they can retrieve relevant documents even if they do not contain the specific words used in the query. Concept-based retrieval often tested the effects of thesaurus-based query expansion on Boolean retrieval performance. In [22], Chen et al. proposed extended fuzzy concept networks for dealing with user's query. In [21], Chen et.al, proposed an interval valued fuzzy concept networks for information retrieval, in which the interval values are used to represent the degrees of association between concepts. Chen et al. in [20] proposed a fuzzy valued concept network in which fuzzy numbers are used to determine the degree of association between the concepts.

In [25], Jung et al. proposed a terms weighting scheme which considers "absence terms" along with "occurrence terms", in finding the degrees of similarity among document descriptor vectors, in which the "absence terms" means terms which are not present in a specific document and they are provided negatively weighted rather than assigning zero weighted .In Klink's study [14], he proposed an automatic reformulation method for improving the original query. Kim et al. in [8] proposed a query term expansion and reweighting method which considers the term co-occurrence within the feedbacked documents. According to Daoud and Boughanem [18], based on the queries given by the user, a user profile is created for the same search session which allows re-ranking the search results of the query. Kang and Cheol [10] said that the classification information generated from the upper ranked documents can be used to generate a cluster which can be then selected by the user corresponding to the requirement. In [24], Koutrika and Loannidis, proposed query re-writing algorithm for disambiguation of query based on user profiles so as to personalize the searches.

## 3. METHODOLGY USED

The concept based query expansion model proposed in this paper; generate relevant query terms by weighing the terms from the top n retrieved documents. By using a classifier, the top ranked (n) documents are obtained from the first retrieval [10]. On the basis of the documents retrieved, the query terms are chosen and the expansion of the initial query is done based on the combination of higher weighing terms (using re-weighting technique) and the terms having weighted threshold> Wp (using updates user profiles related to browsing history) in order to obtain the concepts and the respective phrases. Those phrases are finally joined with that the original user's query terms thereby expanding the query. This leads to an improvement in the search results and the relevant document retrieval for a particular user's query. The results are then returned to the user to achieve maximum user satisfaction higher retrieval effectiveness. The user profile is then updated and used for future reference to perform faster retrieval. The weighing of terms based on concept network for query expansion model is shown in Fig..1.

### 3.1 User Profiles

The information presented to the user must be in accordance with the user's needs. But in practice, user is not clear about the exact information he wants to search. A query may not represent a unique information need, resulting in generation of many irrelevant answers For example, the user searching information on a specific fruit may enter the query as "Orange", and then he might be encountered with results specifying information about orange business services which is not relevant for the user information need. Thereby, results ranking is important which comprises of re-ordering the results returned by the underlying search engine based on user preferences. As different users have different perceptions, variable needs so based on the user's interest, personal information is collected and analyzed and the result is stored in user profile.

Information can be captured in two ways: explicitly, by asking user for feedback such as preferences or rating; and implicitly, by analyzing behavior of user such as the time spent reading an online document. User's profiles are represented as weighted concept hierarchies in which the concepts having more classified items received higher weights. We construct each user profile based on the following two methods [16]: (a) Pure browsing history, in which we assume that the preferences of each user consist of the following two aspects: (1) persistent (or long term) preferences, (2) ephemeral (or short term) preferences. In persistent preferences, incremental user profile development is done periodically and it is stored for use in later sessions and in ephemeral preferences, the only the current session information used to construct each user profile is gathered, and it is immediately exploited for executing some adaptive process for personalizing the current interaction (b) Modified collaborative filtering in which predictive algorithms can be used to predict a term weight in each user profile. In other words, each user profile is computed based on term weights in a Web page the user browsed [16]. The user profile can also be constructed by combining graphs based on query profiles in same search sessions [18]. The construction of user profile is important in a way that it presents right information to the right user at right time i.e. Personalization [15] is achieved.

## 3.2 Selection of top n relevant documents

In order to select the top ranked documents from the documents retrieved by the search engine, document classifier (TAXON) [6, 10] is used. It basically classifies the relevant documents retrieved from the initial user natural language query and creates categories. TAXON identifies the relevance of the vector about documents and query subjects, and then classifies the documents. It also performs document classification, using a thesaurus tools, based on concepts by acquiring closely related meanings with the subjects of document. The relevant query terms that occur frequently are selected from the ranked documents retrieved, which are then weighed.

## 3.3 Weighing of query terms

In general, more preference is given to the query terms representing the user search intention and vice versa, the terms with higher frequency can reflect the user's search intentions more accurately. Based on this principle [9], depending on the queries submitted by the user to the search engine, we derive users' search intentions and weight the term in the Query. There have several researches and methodologies for weighting the query terms. Here we use the method in [7].

$$w_{iq} = \left(0.5 + 0.5 * \frac{\mathrm{tfiq}}{\max(\mathrm{tfiq})}\right) * \mathrm{IDF}_i$$

Where, tfik denotes the occurrence frequency of term ti in document dk and tfiq denotes the occurrence frequency of term ti in the user’s query Q. IDF is the inverse document frequency [3].Based on the weight, the higher weighing terms are selected. The main purpose of weighing of query terms is to improve the efficiency of the search and relevancy of documents [8].

## 3.4 The Concept Network

Concept networks have been widely used in the field of information retrieval. In [21], Chen et.al, proposed an interval valued fuzzy concept networks for information retrieval, where interval values are used to represent the degrees of association between concepts. Chen et al., in [19],[28] proposed an information retrieval method based on extended fuzzy concept networks, where one of the four fuzzy relationships can be used to determine the relation between the concepts, i.e., fuzzy positive association relationship, fuzzy negative association relationship, fuzzy generalization relationship and fuzzy specialization relationship [22]. Also they proposed methods for dealing with user’s query based on extended fuzzy concept networks. In [20], they proposed a fuzzy valued concept networks for information retrieval, where the degrees of association between concepts are represented by arbitrary shapes of fuzzy numbers. In [23], Chang et al. presented their work to allow multiple fuzzy relationships between each pair of concepts in fuzzy concept networks, where each relationship has its own strength.

A concept network [5] is a graph including nodes and directed links, where each node represents a concept or a document. A basic concept network CN = {C, P, E} consists of a set of concept nodes C, a set of phrase nodes P and a set of edges E, say, = {ci, pj, wij}, where ci ϵ C, pj ϵ P, and ci and pj are related with a weight wij. The weight wij of an edge in the concept network represents the degree to which the phrase represents the intension of the concept. However, a concept network can be constructed provided a set of concepts and documents assigned to these concepts. During this process, a bag-of-words approach is used to count the occurrences of noun phrases within each document. The noun phrase frequency is used to calculate the weight values wij.

For example, consider the set of documents Di = {di1. . . din} which are a subset of the extension of the concept ci ϵ C. For each document dik, the set of phrases used in this document is Pik = {p1, ik. . . pm ,ik}. We define a function f (dik, pj) as the occurrence count of phrase pj in document dik. The value for the edge weight between concept ci and phrase pj is given by [11]:

$$w_{ij} = \frac{\sum_{k=1}^{n} \frac{f(d_{tk}\ ,p_j)}{\sum_{l=1}^{m} f(d_{tk}\ ,p_l,t_k)}}{n}$$

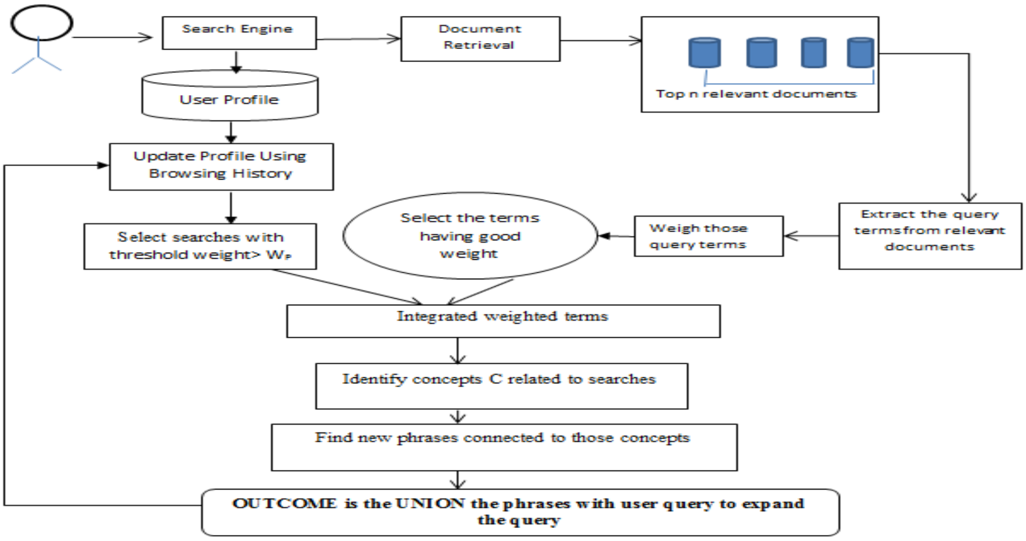
Search Engine
Document Retrieval
Top n relevant documents
User Profile
Update Profile Using Browsing History
Extract the query terms from relevant documents
Weigh those query terms
Select the terms having good weight
Select searches with threshold weight> Wₚ
Integrated weighted terms
Identify concepts C related to searches
Find new phrases connected to those concepts
OUTCOME is the UNION the phrases with user query to expand the query

**Fig. 1: Concept based query expansion model**

## 4. PROPOSED QUERY EXPANSION

The conceptual weighing query expansion method based on the user profile is given as follows:-

Consider a concept network CN = {C, P and E}, and a query Q = {$q_1$ . . . $q_n$} consisting of query phrases $q_i$, and the weighted thresholds W = {$W_{P,}$ We, $W_D$}. **The weighted thresholds are evaluated based on the number of top 'n' documents retrieved as a result of query expansion. Determination of threshold also depends on the frequently used query terms from various profiles of the users running the query**. The process of query expansion is as follows:

1. Concept network consists of a phrase set P, consisting of phrases.
2. Consider the searches made by the user (using user profile) having weight greater than the threshold weight of the searches i.e. WP. Select the terms from those searches.
3. If there are terms in the searches with threshold < WP then, don't consider those terms for further use. Else obtain the set of concepts C for the given concept network related to the searches, connected to the phrases P having weight greater than the threshold weight we and matching with that of query phrases.
4. Find the set of phrases P' which are connected to the concepts C. We use a weight threshold parameter WD to derive all phrases connected to the concepts in C" with a weight greater than WD which are then chosen as the phrases P".
5. The original query phrase and the new set of phrases are then combined to obtain the query expansion: QE = Q U P".
6. This can be further used to update user profile using browsing history which can be referred for future search to fasten the search and improve the retrieval efficiency.

The pseudo code for the above proposed method is given in Fig. 2:

```
Step1. Consider a function F_Search_Engine with a parameter P_Data {P_Data-> input dataset, P_concept-> subset of concept network initially passed as NULL.}
This will return a set of documents P_list_of_docs.
Step2. The above returned docs will be passed to a function F_Classifier which will classify the documents and will return the top n relevant_documents
Step3. The relevant_documents will be then passed to another function F_Extracted_Terms which will return the query terms having higher weights.
Step4. Using the user profile identify the previous searches with threshold greater than WP {WP ->weight of previously searched keywords}.
Step5. Combine the above query terms (from Step 3) and searches (from Step 4) to get a set of integrated weighted terms.
Step6. Identify the concepts based on the integrated weighted terms
Step7. Derive the phrases connected to those concepts.
Step8. The above derived phrases will be passed to a function F_Union. {F_Union->union of terms extracted in step 3 and phrases derived in step 6}
Step9. Update the user profile using browsing history as a base for new search in future reference {Refer Section 3.1}.
```

**Fig. 2: Pseudo code for Proposed Query Expansion Method**

Let us illustrate this method by the use of a simple example shown through Fig.. 3

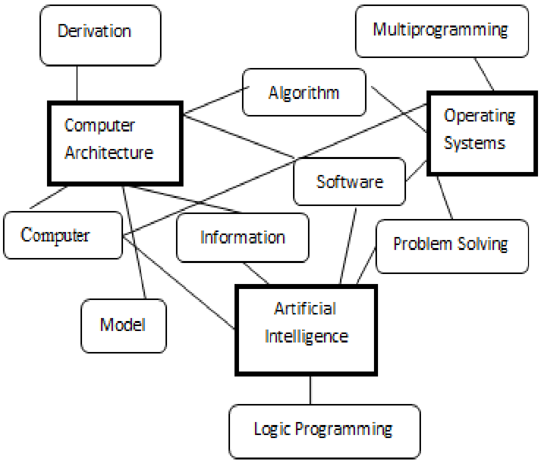

**Fig. 3: An example of concept network**

In this figure, the concepts are represented by the solid line boxes i.e. Computer Architecture, Operating Systems, Artificial Intelligence and the phrases are represented by the rounded rectangle boxes. In the interest of clarity, distance is used to represent the edge weights, rather than displaying the weight values; phrases with very low weights. Suppose we are given a user query Q = {How INFORMATION is derived for PROBLEM solving through use of SOFTWARE}. So here, let us say the query terms after removing stop words and other synonyms in Q are {"information", "problems", "software",} and considering weight thresholds We = 0.05 and WD = 0.1 (with reference to section 4), Hence we follow certain steps.

1. The function F_Extracted_terms will extract the relevant query terms from the documents i.e. Q= {Information, problems, software}. These terms are selected by weighing the terms and hence choosing only the terms having larger weights by weighing technique [9].
2. The other additional terms, if any, from the previous searches having weighted threshold> $W_P$ are identified by user profiles and are then integrated with the terms extracted in step 1.
3. Now the concept network consists of concepts C= {Computer architecture, Operating systems, artificial intelligence} and respective phrases connected to each concept.
4. From concepts C, derive those concepts C' having $w_{e,}$ greater than other concepts. So the 'Computer Architecture' is connected to maximum phrases so chosen as candidate concept C'.
5. Now derive the new phrases connected to the concept chosen i.e. {model, computer, information and derivation}.
6. Match these phrases with the query terms already derived in step 2. And now join the terms to expand the query. Hence the result {information, problems, software, model, computer, derivation}.

Hence, through this example we have observed that the user's query is expanded by adding some relevant terms thereby refining the search results. The proposed approach is more appropriate for query expansion than the other conceptual approaches for expansion used; it considers combination of terms from user's previous searches and terms having good weight (using re-weighting technique) for deriving concepts and the respective phrases for expansion. Thus giving a query more specific than the original query and improving the quality of expansion.

## 5. CONCLUSION

In this paper, we proposed a technique of query expansion for expanding the query terms in order to perform faster and relevant retrieval of documents based on the user's query. This helped in increasing user's satisfaction due to more appropriate search results. We have made use of concept network for deriving the concepts related to searches, thus obtaining the phrases connected to those concepts and also query weighting technique is used for extracting terms having good weight from relevant documents. The user profile is constructed to analyze the previous searches made by the user and thus the profiles are updated based on browsing history. This improved the quality of search and time of retrieval.

## 6. REFERENCES

[1] Efthimis N. Efthimiadis. Query expansion, Annual Review of Information Systems and Technology (ARIST), 31, 1996.

[2] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais, The vocabulary problem in human-system communication. Communications of the ACM, 30(11), 1987

[3] R. Baeza-Yates, B. Ribeiro-Neto, Modern information retrieval, Addison Wesley, 2011.

[4] Y. Qiu and H. P. Frei. Concept based query expansion, In Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, 1993.

[5] Martin Kracker, A Fuzzy Concept Network Model and its applications ,IEEE International Conference on fuzzy systems , 1992

[6] H.K Kang, B.M. Ryu and I.K Whang, An effective Document Classification System Based on Concept Probability Vector, Lecture Notes in Computer Science 3309, pp.457-462, 2005.

[7] G.Salton, C.Buckley, Term-weighting approaches in automatic text retrieval, Information processing & Management, Volume 24, No 5, pp. 513-523, 1988.

[8] Kim, B.M., Kim, J. Y. and Kim, J., Query term expansion and reweighting using term co-occurrence similarity and fuzzy inference, Proceedings of the Joint 9th IFSA World Congress and 20th NAFIPS International Conference, Vancouver, Canada, Vol. 2, pp. 715-720, 2001.

[9] C.Wang, Yajun DU, P.Zhang, B.Han, A Term-Reweighting Method for Query Expansion, Journal of Computational Information Systems 6:11, pp. 3779-3785, 2010.

[10] John W.Kang, Hyun-Kyu.Kang, A Term Cluster Query Expansion Model based on Classification Information in Natural Language Information Retrieval, International Conference on Artificial Intelligence and Computational Intelligence, 2010.

[11] O. Hoeber, X.D. Yang, Y. Yao, Conceptual Query Expansion, Advances in web intelligence, Springer, LNAI 3528, pp. 190-196, 2005.

[12] R. Bodner, F .Song, Knowledge-based approaches to query expansion in information retrieval. In McCalla, Advances in Artificial Intelligence, pp. 146-158, 1996.

[13] Qiu, Y. and Frei, H.P, Concept based query expansion, In Proceedings of the 16th annual international ACM SIGIR conference on Research and Development in Information Retrieval, ACM Press, pp. 160-170, 1993.

[14] S. Klink, Query reformulation with collaborative concept-based expansion, In Proceedings of the First International Workshop on Web Document Analysis WDA, 2001.

[15] M. Speretta, S. Gauch, Personalized search based on user search hierarchies, Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05), 2005.

[16] K. Sugiyama, K. Hatano, M. Yoshikawa, Adaptive Web Search Based on User Profile Constructed without Any Effort from Users, ACM 1-58113-844-X/04/0005, 2004.

[17] Youjin Chang, I. Choi, J. Choi, M. Kim, V.V. Raghavan , Conceptual Retrieval based on Feature Clustering of Documents, 2002.

[18] M. Daoud, L.T. Lechani, M. Boughanem, A Session Based Personalized Search Using an Ontological User Profile, ACM 978-1-60558-166-8/09/03, 2009.

[19] S.M. Chen, Y.J. Horng, Fuzzy query processing for document retrieval based on extended fuzzy concept networks, IEEE Trans. Systems Man Cybernet.—Part B: Cybernet. 29 (1) pp.126–135, 1999.

[20] S.M. Chen, Y.J. Horng, C.H. Lee, Document retrieval using fuzzy valued concept networks, IEEE Trans. Systems Man and Cybernet.—Part B: Cybernet. 31 (1) pp.111–118, 2001.

[21] S.M. Chen, W.H. Hsiao, Y.J. Horng, A knowledge-based method for fuzzy query processing for document retrieval, Cybernet. Systems Internat. J. 28 (1) pp. 99–119, 1997.

[22] M. Kracker, A fuzzy concept network model and its applications, Proc. First IEEE Internat. Conf. on Fuzzy Systems, San Diego, USA, pp. 761–768, 1992.

[23] S.M. Chen, Y.J. Horng, C.H. Lee, Fuzzy information retrieval based on multi-relationship fuzzy concept networks, Elsevier, 2002.

[24] Koutrika, Loannidis, A Unified User Profile Framework for Query Disambiguation and Personalization, Proceedings of the Workshop PIA, 2005.

[25] Y. Jung, H. Park and D. Du, An Effective Term Weighting Scheme for Information Retrieval, Computer Science Technical Report TR008, Department of Computer Science, University of Minnesota, Minneapolis, minnesota, pp. 1-15, 2000.

[26] GG Chowdhury, Natural Language Processing, Annual Review of Information Science and Technology, 2003.

[27] Tanveer Siddiqui and U. S. Tiwari "Natural Language Processing and Information Retrieval" Oxford University press, 2008.

[28] S.J,Chen, H.C.Chu, A New Method for Fuzzy Query Processing of Document Retrieval based on Extended Fuzzy Concept Networks, International Conference on Electronics and Information Engineering, 2010.

[29] M.Dragoni, Celia da Costa Pereira, Andrea G.B. Tettamanzi, A Conceptual Representation of Documents and Queries for Information Retrieval System using Light Ontologies, Expert Systems with Applications 39 (2012) pp.10376–10388, Elsevier, 2012.

[30] H. Imran, A. Sharan, Thesaurus and Query Expansion, International Journal of Computer science & Information Technology (IJCSIT), vol. 1, No 2, November 2009.