Issue 4, Vol.4 (June-July 2014)
ISSN 2249-6149

A Detailed Study on Query Expansion Techniques in Information Retrieval

Neha Soni[#1], Jaswinder Singh[#2]

#1 Student, M. Tech, Department of Computer Science and Engineering, G.J.U.S&T, Hisar, Haryana, India

#2 Assistant Professor, Department of Computer Science and Engineering, G.J.U.S&T, Hisar, Haryana, India

ABSTRACT

Query expansion is a scheme that helps searchers to formulate improved query statements to retrieve better search results. It attempts to expand user's query through the analysis of initially retrieved documents. The objective of this technique is to find additional keyword for the query that improves the quality of recovered information. Numerous techniques have been proposed and tried for improving the effectiveness of searching the World Wide Web for documents relevant to given topic of interest. In this paper various techniques of expanding the query are discussed.

Key words: Genetic Algorithm, Information Retrieval,Similarity Measure, Query Expansion.

Corresponding author: Neha Soni

1. INTRODUCTION

The www these days is the source of choice for information for everything. So the amount of information on the World Wide Web is growing fast and rapidly, as well as the number of new users unskilled in the art of web research. At the same time, the number of queries that search engines can handle has grown incredibly too. The Internet revolution has given rise to search engine, a tool whose task is to identify among the billions of existing websites those that are relevant to user's query. When a user submits his or her query the search engine analyses its repository of stored websites and returns the list of hyperlinks to those that contain information requested by the user's query. This list is in sorted order so that the most relevant websites come up first. Many mechanisms have been proposed to assess the degree of website's relevance, among them keyword frequency in the document, the average time spent on a given web page by other users, interlinking with other important websites and various combinations of all these three approaches[1]. The rate of handling queries must be hundred to thousand per second. Today these tasks are becoming increasingly difficult as the Web grows. The results that a user is interested in are often wash out by junk results. One of the main cause for this is that the number of documents in the indices has been increasing by many orders of magnitude, but the ability of user to look at documents has not. People are only willing to look at the first few of the results. In this situation, the retrieval of documents relevant to the user's need is of foremost importance.

2. INFORMATION RETRIEVAL (IR)

Information retrieval is the study of how to determine and retrieve from a collection of stored information, the sections that are responsive to particular information need. IR is concerned with text representation, storage, organization and retrieval of stored information that are similar in some sense to information requests received from users. The information need of the user must first be translated into a query which will be processed by the IR system. The set of keywords received from this translation will summarize the description of the user information needed. The key goal of IR system is to retrieve information which may be relevant to the user, given the user query. The integral component of any information retrieval system is ranking. It is common that web search queries have thousands or millions of results. Web users do not have the time and patience to go through all of them to find the ones they are interested in. Most web users do not look beyond the first page of results. Therefore, it is important that ranking function should output the desired results within the top few pages, otherwise the search engine is rendered useless.

2.1 Components of IR System

There are three basic components of IR system [2].

1) Query Subsystem: It is a system that allows users to formulate their queries and present the relevant documents retrieved by the system for user's query.

2) Matching Function: It compares both query and documents in database and gives a value which measures the similarity between query and document.

3) Document Database: It is the storage space where all the documents are stored.

The Architectural Diagram of Information Retrieval System is shown below:

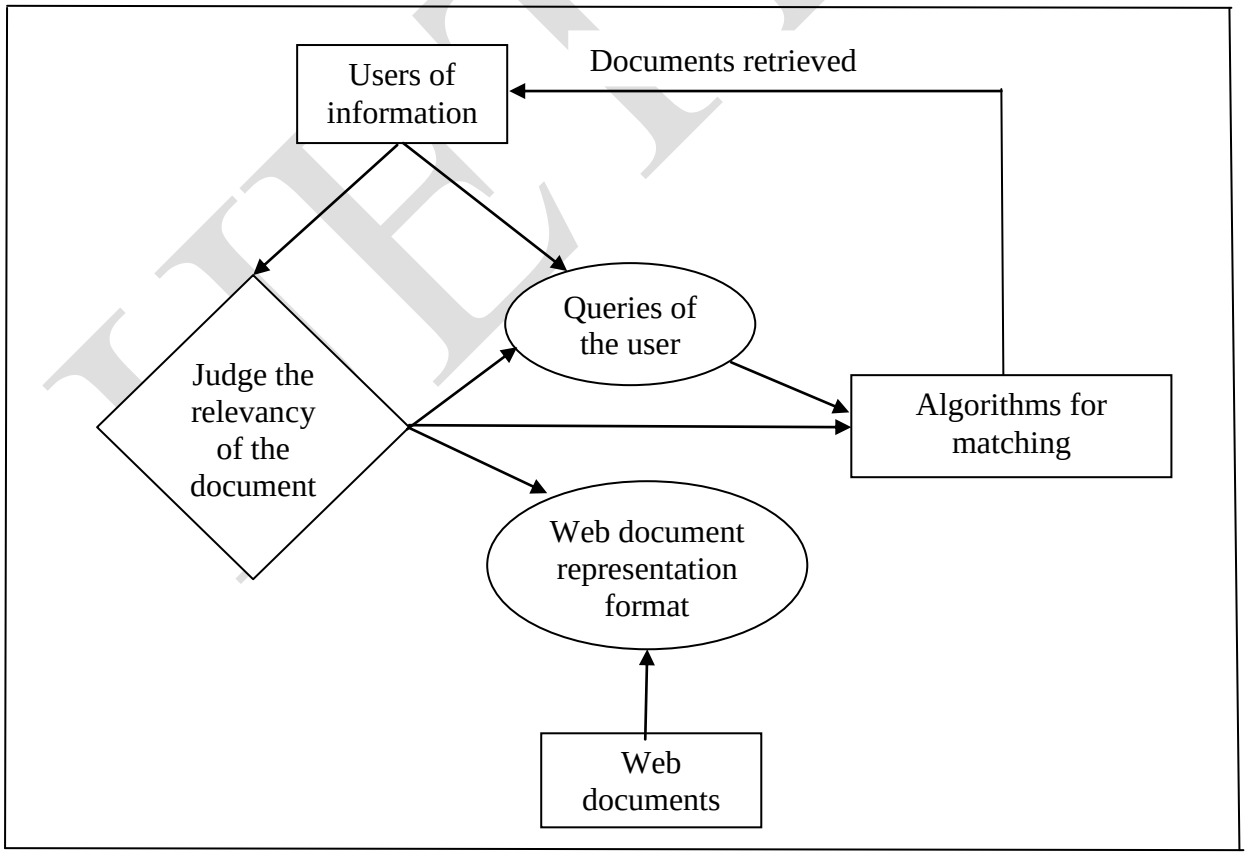

Fig 1: Basic Architecture of Information Retrieval System

2.2 Classic Models of IR System [3]

1. Boolean Model: This model is most common exact-match model and is based on Boolean logic and classical set theory in that both the documents to be searched and the user's query are conceived as sets of terms. Retrieval of document is based on whether or not the documents contain the query terms. It uses Boolean queries and Boolean operators like OR, AND, NOT.
2. Vector Space Model: It Represents queries and documents in multi-dimensional space. Here, documents as well as queries are represented as vectors of weights. Each weight denotes the importance of the corresponding term respectively in the document or in the query.
3. Probabilistic Model: It is based on the Probability Ranking Principle, which states that an IR system is supposed to rank the documents based on their probability of relevance to the query, given all the clue available. The principle of this model takes into account that there is uncertainty in the representation of the queries and documents.

2.3 Quality Evaluation of IR System

There are several criteria to measure this aspect, with precision and recall being the most used. Precision is a standard IR performance measure. It is defined as the number of relevant documents retrieved divided by the total number of documents retrieved. Recall is also a standard IR performance measure. It is defined as a number of relevant documents retrieved divided by the total number of relevant documents in the collection.

2.4 Process of IR System [4]

For each collection, each query is compared with all the web documents, using any similarity measure. This gives a list describing the similarities of each query with all documents of the collection. Then this list is ranked in decreasing order of similarity degree. After making a training data consists of the top N (predefined document cutoff) document of the list with a corresponding query, the keywords (terms) from the training data and the terms which are used to form a query vector are retrieved automatically. Lastly adapt the query vector using the genetic approach.

3. SIMILARITY MEASURES

It is function use to measure the degree of similarity between query and documents. It measures how much the query and document is similar to each other. This measure gives a value which decides the degree of similarity. First, query and document are converted into vector form in order to find the similarity between them and if the query and document vector do not have any term in common then similarity score is very document are converted into vector form in order to find the similarity between them and if the query and document vector do not have any term in common then similarity score is very low. Some of the popular measures are cosine, jaccard, dice and various other measures are defined [5-13].

Cosine

$$Cos(Q, D_i) = \frac{\sum_{j=1}^{t} w_{q,j}\, d_{i,j}}{\sqrt{\sum_{j=1}^{t}(w_{q,j})^2 \sum_{j=1}^{t}(d_{i,j})^2}}$$

Jaccard

$$Jaccard(Q, D_i) = \frac{\sum_{j=1}^{t} w_{qj}\, d_{ij}}{\sum_{j=1}^{t}(d_{ij})^2 + \sum_{j=1}^{t}(w_{qj})^2 - \sum_{j=1}^{t} w_{qj}\, d_{ij}}$$

Dice

$$DS = \frac{2\sum_{i=1}^{n} A_i \cdot B_i}{\sum_{i=1}^{n}(A_i)^2 + \sum_{i=1}^{n}(B_i)^2}$$

Issue 4, Vol.4 (June-July 2014)
ISSN 2249-6149

Czekanowski

$$S_{cze} = \frac{2\sum_{i=1}^{d} \min(P_i, Q_i)}{\sum_{i=1}^{d}(P_i + Q_i)}$$

4. EVOLUTIONARY APPROACH-GENETIC ALGORITHM (G.A)

The application of GA to IR holds interesting promises in Information Retrieval [8, 14]. GA is a heuristic search algorithm based on natural selection and genetic ideas. GA has ability to find global solution in many difficult problems. It uses the idea of survival of fittest individuals within a given population. A population of strings or we can say solutions to a specified problem are created and maintained by the GA. GA then iteratively create new populations from the old by ranking the strings and then choose the fittest for interbreeding to create new strings that are hopefully will be closer to the optimum solution for the problem. GA operations can be used to generate new and better generations.

4.1 GA Components

1. Chromosome representation for the feasible solutions to the optimization problem.
2. Initial population of feasible solutions.
3. A fitness function use to evaluate each solution.
4. Genetic operators that generate new population from existing population.
5. Control parameters such as probability of genetic operators (crossover, mutation), number of generations and population size.

4.2 GA Approach to IR [2, 3, 10]

The diagram of intelligent search approach for the use of GA in IR is shown below:

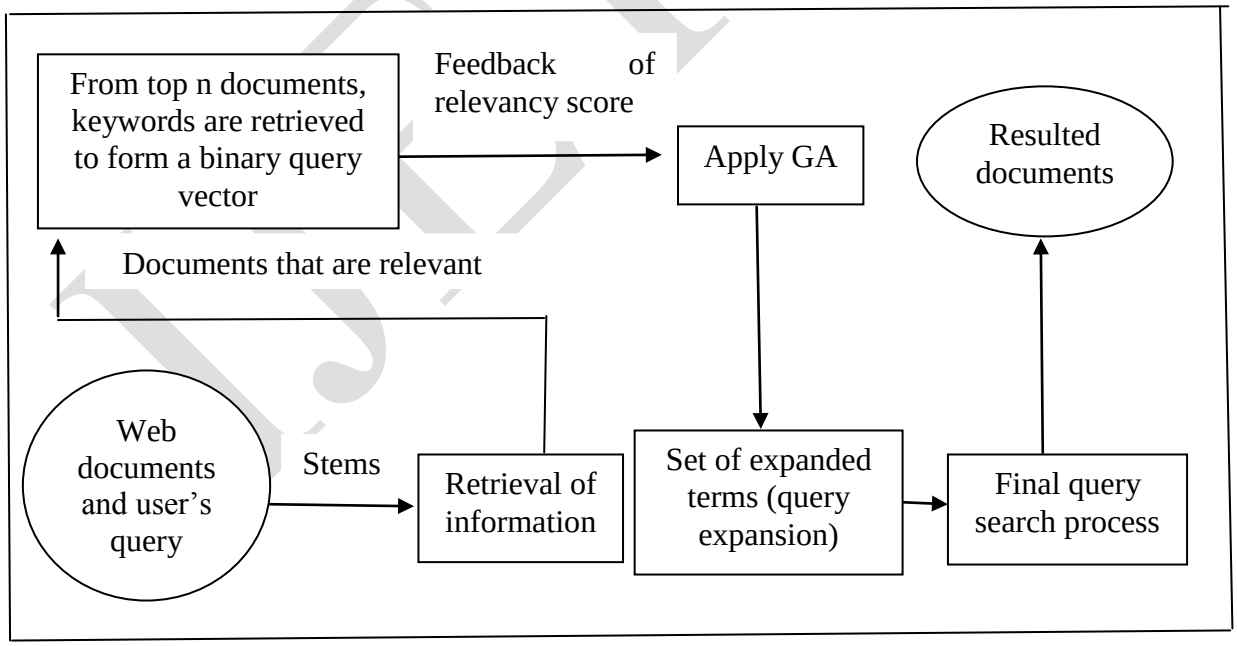

Fig 2: Diagram of intelligent search approach

The terms in Documents and queries are pre-processed by creating a list of keywords from each document. Keywords retrieved from IR is received as initial population chromosome by GA approach. The number of keywords of documents retrieved from a user query defines the length of chromosome. Then Documents and queries are indexed using Normalized Vector Space Model (VSM) and weights are assigned to the terms in each document. The weights of

the keywords are formed as query vector and document vector. Then Documents are matched with queries using fitness function.

The fitness function is a performance measure, used to evaluate how well each solution is. Given a chromosome, the fitness function must return a numerical value that represent the chromosome. This value will be used in the selection process. The fitness function used in the proposed IR system is based on similarity measure. GA module then applies the standard GA functions (selection, crossover). Then select highest n keywords frequency from the top N(cutoff) documents of the list with a corresponding query. These keywords are retrieved and are used to form a query vector. Output list containing the first results is generated. Adapt the query vector using the GA approach, to get an optimal or near optimal query vector. The Queries are then expanded by adding terms that are most relevant to the original query, to be used in the search system.

5. RELATED WORK

Eman Al Mashagba, Feras Al Mashagba and Mohammad Othman Nassar [13] different similarity measures in VSM, different GA approaches and compared them to find the best one that could be used when data is Arabic language. Priya I. Borkar and Leena H. Patil [2], presented a model of HGAPSO for web information retrieval. It expanded the keywords to produce the new keywords that were related to the user search. Noor Ali Ameen Albayaty and Nushwan Yousif Baithoon [3], proposed several contributions towards the query improvement and query expansion and compared the result with original query. It was concluded that QE methods increased the precision and recall rate. Abdelmgeid A. Aly [4], presented an adaptive method using GA to modify user's queries, based on relevance judgments. The algorithm showed the effects of applying GA to improve the effectiveness of queries in IR systems. Jose R. Perez-Agiiera [15], proposed a new GA used to change the set of terms that compose a user query without the supervision of user, by complementing an expansion process based on the use of a morphological thesaurus. Suhail S. J Owais, Pavel Kromer, and Vaclav Snasel [16], investigated the use of GA in IR in the area of optimizing a Boolean query. Abdelmgeid Amin Aly [17], presented a query expansion method to reformulate the query. Experiments on test collections showed that the improvement increases with the size of collection and with the number of additional search terms that expand the original query. Andre L. Vizine and Leandro N. de Castro & Ricardo R. Gudwin [18], described a system for automatically generating group profiles for web documents by selecting a suitable library of keywords, and a search agent that generates and optimizes, via a genetic algorithm (GA), search queries for the Google search engine. M. Shamim Khan and Sebastian Khor [19] described a scheme that attempts to automatically expand the user query through the analysis of initially retrieved documents. Xiaomin Zhang, Sandra Zilles and Robert C. Holte [20] proposed two new query suggestion systems using query search. Li Ming [21], proposed a new query expansion method based on Bayesian belief network and relevance feedback. Aysh Alhroob, Hayel Khafajeh and Nisreen Innab [22], conducted the comparison between two query expansion techniques (global and local query) to determine the query effectiveness. Gaihua Fu, Christopher B. Jones and Alia I. Abdelmoty[23], introduced an ontology-based spatial query expansion method. It was specially designed to resolve a query that involved a fuzzy spatial relationship. Hang Cui, Ji-Rong Wen, Jian-Yun Nie, Wei-Ying Ma[24] and Zhu Kunpeng, Wang Xiaolong, Liu Yuanchao[25] took the advantage of the query logs available in various websites and used them as a means for query expansion. Laurence A. F. Park and Kotagiri Ramamohanarao[26], presented a new method of automatic query expansion using a collection dependent thesaurus built with probabilistic

latent semantic analysis. Hazra Imran and Aditi Sharan[27], addressed the basic issues related to the query expansion along with some important strategies that have been used for automatic query expansion. Claudio Carpineto and Giovanni Romano[28] presented a unified view of a large number of recent approaches to Automatic Query Expansion that leveraged various data sources and employed very different principles and techniques. Many questions related to AQE were addressed. Jose R. Perez-Aguera and Lourdes Araujo[29] presented a study of two different approaches, co-occurrence and probabilistic distributional analysis, for query expansion. Claudio Carpineto, Renato de Mori, Giovanni Romano and Brigitte Bigi[30] presented a computationally simple and theoretically justified method for assigning scores to candidate expansion terms. Such scores were used to select and weight expansion terms within Rocchio's framework for query reweighting. Bhawani Selvaretnam , Mohammed Belkhatir[31] identified the factors that influence the performance of query expansion methods. Mohammed Otair, Ghassan Kanaan and Raed Kanaan[32] optimized the Arabic queries using comprehensive combination of the expansion techniques. Ashish Kishor Bindal and Sudip Sanyal[33], presented a stochastic based approach for optimizing the query vector without user involvement. The document search space using particle swarm optimization was explored and the search space of possible relevant and non-relevant documents for adaption of query vector was exploited. Yogesh Kakde[34], presented the important work done on Query Expansion (QE) between the period 1970 to 2012. Bodo Billerbeck and Justin Zobel[35], explored the alternative methods for reducing query evaluation costs and proposed a new method based on keeping a brief summary of each document in memory. Ali Asghar Shiri, Crawford Revie and Gobinda Chowdhury[36], provided a review of the literature related to the application of domain-specific thesauri in the search and retrieval process. The review consists of two main sections covering, firstly studied on thesaurus-aided search term selection and secondly those dealing with query expansion using thesauri.

6. QUERY EXPANSION

From the review of literature it can be said that the query submitted by the user can be expanded by various methods for efficient web document retrieval. The explosive growth of www is making it difficult for a user to locate information that is relevant to his/her interest. Though existing search engines works well to a certain extent but they still face problems like word mismatch which arises because the majority of IR systems compare query and document terms on lexical level rather than on semantic level. The average length of queries by the user is less than two or three words. Short queries and the incompatibility between the terms in user queries and web documents strongly affect the relevant document retrieval. Query expansion is a technique to increase the effectiveness of the information retrieval. It is the process of supplementing additional terms or phrases to the original query to improve the retrieval performance. The main problem of query expansion is the selection of expansion terms based on which user's original query is expanded.

Query Reformulation approaches is divided into Query Expansion and Query Reweighting approaches. Both of which are further categorized into techniques based on Thesaurus, techniques based on Relevance Feedback and technique based on information extracted from collection of documents.

Types of Query Expansion Techniques: Divided into

1. Global Analysis Technique
2. Local Analysis
3. Context Analysis Method

Global Analysis: In this technique the list of candidate terms is generated from the whole collection. The earliest global analysis technique is term clustering, which groups document terms into clusters based on their co-occurrences. User's Queries are expanded by the terms that are in the same cluster. Other global techniques include Latent Semantic Indexing, similarity thesauri, and Phrase Finder. It requires statistics such as statistics of co-occurrences of pairs of terms, which results in a similarity matrix among those terms. To expand the query, terms which are the most similar to the query terms are identified and added. Local analysis uses only some initially retrieved documents for query expansion. A well-known local analysis technique is relevance feedback which alters a query based on user's relevance judgments of the retrieved documents. The expansion terms are extracted from relevant documents. Local context analysis combines the local and global analysis approach. In this concepts are selected based on co- occurrence with query terms, chosen from the top ranked documents and the best passages are used instead of whole documents [22][27].

Query expansion techniques can broadly be classified in two categories: those based on the search results and those based on some form of knowledge structure.The former group of techniques depends on the search process and uses relevance feedback in an earlier iteration of search as the resource to identify the query expansion terms and the latter group of techniques is independent of the search process and additional query terms are derived by traversing a semantic network built up according to knowledge structure [23].

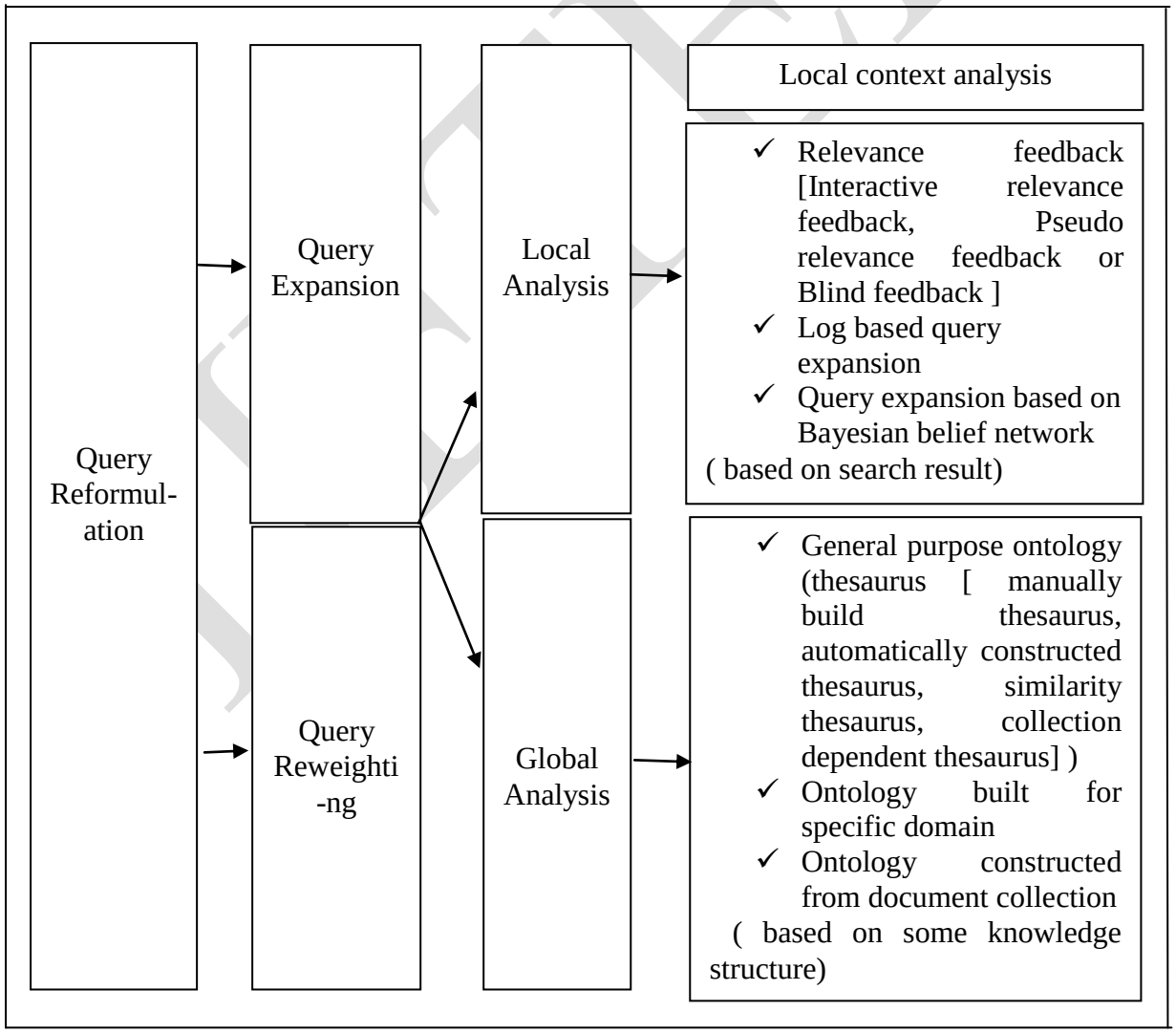

Fig 3: Query Expansion Techniques

Knowledge structures can either be a general-purpose ontology (or thesaurus) or an ontology built for a specific domain, or an ontology constructed from document collection based on the term clustering. The common use of ontologies in Semantic Web is to enrich the current Web resources with some well-defined meaning to enhance the search capabilities of existing web searching systems. This ontology based technique is distinguished from conventional ones in that a query is expanded by derivation of its geographical query footprint. Topic ontologies are means of classifying web pages based on their content. The main problem of query expansion is the selection of the expansion terms based on which user's original query is expanded.

Thesaurus helps to solve this problem[26]. Query expansion that is done using a thesaurus, adds synonyms, broader terms, and other appropriate words. Thesaurus are built manually or automatically. Building the thesaurus manually requires the study of words meanings and the relation between these words according to the meaning like synonyms and antonyms. The automatic methods to build the thesauri are characterized by high precision in the determination of relations between words and the possibility of using the same method for more than one language and a great number of corpuses can be used to build automatic thesaurus. The similarity thesaurus calculates the relevance between terms and queries and is constructed by interchanging the role of documents and terms in retrieval model. In similarity thesaurus the relevance of a term to the concept of the query is the sum of the weighted relevance of the term to each term in the query. User's queries are expanded by adding top n relevant terms that are most similar to the concept of the query rather than selecting terms that are similar to the query terms. A collection dependent thesaurus is automatically built and use the term frequencies found with a document collection. A collection dependent thesaurus is a square table that contain all the words which are found within the collection and their relationship to each other. Relevance feedback is local query expansion because it uses only a subset of document set to calculate the set of expansion terms. Thesaurus expansion is global query expansion because it uses the whole document set when calculating the set of expansion terms. In this method, the initial search is conducted through the system and uses the user's query. The system retrieves a set of ranked documents and then the user determines which of these documents is relevant to the query. The system reforms the query based on user judgment means the determination of the relevant and irrelevant documents. The system then repeats the retrieval process by using the modified query. The process is repeated until the user gets the suitable documents for him. Probabilistic retrieval is based on estimating the probability of relevance of web document to user for the given user's query. The relevance feedback from a few documents is used to establish the probability of relevance for other documents in the collection. Interactive relevance feedback extends the involvement of the retrieval system in IR process to rely on user feedback. Pseudo relevance feedback or blind feedback is a non-interactive version of relevance feedback method. To remove user interaction and hence speed up the query process, the retrieval system does not question the user about the relevance of the top matching documents to query, but instead assumes that the documents that match the query are relevant. The terms are then extracted from this set of documents and used to build the expanded query. In log-based query expansion, relevance judgments are obtained from the query logs which record user interactions with search system. The log-based query expansion [24] overcomes the difficulties of relevance feedback and pseudo-relevance feedback. Sufficient information of user relevance judgments can be educe by analysing query logs while eliminating the step of collecting feedbacks from users for ad-hoc queries. User logs help establishing probabilistic correlations between terms in both user queries and documents and with these term-term correlations, accurate expansion terms can be selected from the documents. Beside full text query, documents can be retrieved by the attributes of the document in enterprise document

retrieval. An approach to expand the full text query with attribute query improves the retrieval performance. Firstly the full text search is conducted. The relevance of retrieveddocuments judged by the user are the data source to construct the attribute query. The possibility of match between the keywords and attributes is derived from the Bayesian belief network [21] and the match with maximum possibility is used as the attribute query to expand the full text query. Later the documents are searched again with the expand query.
Weighting method assumes that all relevant documents are known before a query is submitted, a situation which is not realistic but suggestive of a method of relevance feedback after some knowledge is gained about relevance. Probabilistic weighting schemes provide a useful method for relevance feedback, especially in the field of term reweighting. The local feedback method is similar to the traditional relevance feedback method, which alter queries by using the result of initial retrieval, except that the traditional method uses the judgment set for calculating re-weighting while the local feedback method assumes that terms in the top ranked n documents are relevant to the user's request. Then expand the Queries by adding the weight of terms in relevant documents and reducing the weight of terms in the last m documents of initial retrieval. Instead of single word terms being extracted, it is possible to extract the phrases. If two phrases appear separately at different parts of same document, then they are considered as two different phrases and thus may co-exist.

CONCLUSION

The difficulty faced by internet user while searching the documents related to topic about which he or she is having only partial knowledge is well known. Users don't have ability to build a query string containing appropriate keywords. From the review of literature, it is concluded that there exists several methods to expand a query that can then improve the outcome of search significantly. From the study, it can also be concluded that GA allows to improve the original query.

REFERENCES

[1] W. S. Alhalabi, M. Kubat and M. Tapia, Search engine ranking efficiency evaluation tool, ACM SIGCSE Bulletin, vol. 39 (2), pp. 97–101, 2007.

[2] P. I. Borkar and A. P. L. H. Patil, A model of hybrid genetic algorithm-particle swarm optimization (hgapso) based query optimization for web information retrieval, IJRET, vol. 2(1),pp. 59 – 64, 2013.

[3] N. A. A. Albayaty and N. Y. Baithoon, File Search with Query Expansion in a Network System (s), Information and Knowledge Management, vol.3, pp. 23– 30, 2013.

[4] A. Abdelmgeid, Applying genetic algorithm in query improvement problem, International Journal Information Technologies and Knowledge, Vol. 1, 2007.

[5] A. Huang, Similarity measures for text document clustering, in Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand, pp. 49–56, 2008.

[6] S.S. Choi, S.H. Cha and C.C. Tappert, A Survey of Binary Similarity and Distance Measures,Journal of Systemics, Cybernetics & Informatics, vol. 8(1), 2010.

[7] D. Ellis, J. Furner-Hines, and P. Willett, Measuring the degree of similarity between objects in text retrieval systems, Perspectives in Information Management, vol. 3(2) , pp.

128–149, 1993.

[8] B. Klabbankoh and Q. Pinngern, Applied genetic algorithms in information retrieval, Faculty of Information Technology, King Mongkuts Institute of Techology Ladkrabang, 2000.

[9] S. Renjith and C. Anjali, Fitness Function in Genetic Algorithm based Information Filtering-A Survey, 2013.

[10] J. Usharani, Dr K Iyakutti, A Genetic Algorithm based on Cosine Similarity for Relevant Document Retrieval, International Journal of Engineering Research & Technology, Vol. 2(2), Feb-2013.

[11] Manoj Chahal, Jaswinder Singh,Effective Information Retrieval Using Similarity Function: Horng and Yeh Coefficient, International Journal of Advanced Research in Computer Science and Software Engineering. Vol. 3(8), Aug 2013.

[12] S.H. Cha, Comprehensive survey on distance/similarity measures between probability density functions, vol. 1(4), 2007.

[13] E. Al Mashagba, F. Al Mashagba, and M.O Nassar, Query Optimization Using Genetic Algorithms in the Vector Space Model, International Journal of Computer Science Issues (IJCSI), vol. 8(5), 2011.

[14] M. A. Kausar, M. Nasar and S. K. Singh, A Detailed Study on Information Retrieval using Genetic Algorithm, Journal of Industrial and Intelligent Information, vol. 1(3), 2013.

[15] L. Araujo, H. Zaragoza, J. R. Pérez-Agüera, and J. Pérez-Iglesias, Structure of morphologically expanded queries: A genetic algorithm approach, Data & Knowledge Engineering, vol. 69(3), pp. 279–289, 2010.

[16] S. S. Owais, P. Krömer, and V. Snaˇsel, Query optimization by Genetic Algorithms, in DATESO, vol. 129, pp. 125–137, 2005.

[17] A. Abdelmgeid Amin, Using a Query Expansion Technique to Improve Document Retrieval, 2008.

[18] A. L. Vizine, L. N. de Castro, and R. R. Gudwin, An evolutionary algorithm to optimize web document retrieval, in Proceedings of the IEEE international conference on integration of knowledge intensive multi-agent systems, pp. 273– 278, 2005.

[19] M. Shamim Khan and S. Khor, Enhanced web document retrieval using automatic query expansion, Journal of the American Society for Information Science and Technology, vol. 55(1), pp. 29–40, 2004.

[20] X. Zhang, S. Zilles, and R. C. Holte, Improved query suggestion by query search, in KI 2012: Advances in Artificial Intelligence, Springer, pp. 205–216, 2012.

[21] L. Ming, An approach to query expansion based on Bayesian belief network and

relevance feedback,International Journal of Digital Content Technology & its Applications, vol. 6(4), 2012.

[22] Alhroob, Aysh, Hayel Khafajeh, and Nisreen Innab, Evaulation of different queryexpansion techniques for arabic text retrieval system, American Journal of Applied Sciences vol. 10(9), 2013.

[23] G. Fu, C. B. Jones, and A. I. Abdelmoty, Ontology-based spatial query expansion in information retrieval, in on the move to meaningful internet systems 2005: Coop IS, DOA, and ODBASE, Springer, pp. 1466–1482, 2005.

[24] H. Cui, J.R. Wen, J.Y. Nie, and W.Y. Ma, Query expansion for short queries by mining user logs, IEEE Trans. Knowl. Data Eng, vol. 15(4), pp. 829–839, 2002.

[25] Z. Kunpeng, W. Xiaolong, and L. Yuanchao, A new query expansion method based on query logs mining,International Journal on Asian Language Processing, vol. 19,pp. 1–12, 2009.

[26] L. A. Park and K. Ramamohanarao, Query expansion using a collection dependent probabilistic latent semantic thesaurus, in Advances in Knowledge Discovery and Data Mining, Springer, pp. 224–235, 2007.

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

[28] C. Carpineto and G. Romano, A survey of automatic query expansion in information retrieval,ACM ComputingSurveys (CSUR),vol. 44(1), 2012.

[29] J. R. Pérez-Agüera and L. Araujo, Comparing and combining methods for automatic query expansion, 2008.

[30] C. Carpineto, R. De Mori, G. Romano, and B. Bigi, An information-theoretic approach to automatic query expansion, ACM Transactions on Information Systems(TOIS), vol. 19(1), pp. 1–27, 2001.

[31] B. Selvaretnam and M. Belkhatir, Natural language technology and query expansion: issues, state-of-the-art and perspectives,Journal of Intelligent Information Systems, vol. 38(3), pp. 709–740, 2012.

[32] Otair, Mohammed, Ghassan Kanaan, and Raed Kanaan, Optimizing an Arabic Query using Comprehensive Query Expansion Techniques, International Journal of Computer Applications pp. 42-49, 2013.

[33] A. K. Bindal and S. Sanyal, Query Optimization in Context of Pseudo Relevant Documents, in 3rd ItalianInformation Retrieval (IIR) workshop, 2012.

[34] Y. Kakde, A Survey of Query Expansion until June 2012,Indian Institute of Technology, Bombay, 2012.

[35] B. Billerbeck and J. Zobel, Techniques for efficient query expansion, in String

Processing and InformationRetrieval, pp. 30–42, 2004.

[36] A. A. Shiri, C. W. Revie, and G. Chowdhury, Thesaurus-assisted search term selection and query expansion: a review of user-centred studies,Knowledgeorganization, vol. 29(1), pp. 1–19, 2002.