American Journal of Computer Science and Information Engineering
2017; 4(3): 24-30
http://www.aascit.org/journal/ajcsie
ISSN: 2381-1110 (Print); ISSN: 2381-1129 (Online)

American Association for Science and Technology

Keywords
Search Engine,
Pseudo Relevance Feedback,
Fuzzy,
Recall,
Precision

Received: April 16, 2017
Accepted: April 27, 2017
Published: July 5, 2017

Automatic Query Expansion for Information Retrieval: A Survey and Problem Definition

Olalere A. Abass[1], Olusegun Folorunso[2], Babafemi O. Samuel[1]

[1]Department of Computer Science, Tai Solarin College of Education, Ijebu Ode, Nigeria
[2]Department of Computer Science, Federal University of Agriculture, Abeokuta, Nigeria

Email address

olaabas@gmail.com (O. A. Abass), folorusoo@funaab.edu.ng (O. Folorunso), princefm@live.com (B. O. Samuel)

Citation

Olalere A. Abass, Olusegun Folorunso, Babafemi O. Samuel. Automatic Query Expansion for Information Retrieval: A Survey and Problem Definition. *American Journal of Computer Science and Information Engineering*. Vol. 4, No. 3, 2017, pp. 24-30.

Abstract

An ideal information retrieval system is expected to retrieve only the relevant documents while irrelevant ones are ignored towards ensuring throughput of the retrieval system and reduce the time user spend on the search engines as well as serving a motivation for continue the search. The process of IR consists of locating relevant documents on the basis of user query, such as keywords. One of the most fundamental research questions in information retrieval is how to operationally define the notion of relevance so that we can score a document with respect to a query appropriately. The most critical language issue for retrieval effectiveness is the term mismatch problem because both the indexers and the users do often not use the same words. This scenario is called *vocabulary problem*. Consequently, IRS users spend much time and resources to obtain their information need after querying the system. One solution to this problem is known as query expansion via pseudo relevance feedback which is intelligent technique for boosting the overall performance in IR. This paper reviews the intelligent method of query expansion and fashion out future work on the implementation of intelligent information retrieval for the purpose of removing "noise" (irrelevant documents) from the lists of retrieved documents.

1. Introduction

In the current era of advancement in Information and Communication Technology (ICT), Information Retrieval (IR), a subfield of computer science, has emerged as an important research area that is concerned with the searching and retrieving of knowledge-based information from database [1] and also deals with the representation, storage, and access of information [2]. Hence, IR focuses on the organization and retrieval of information from large database collections [3].

The field of IR has developed along with the field of databases. In the traditional IR model, it is assumed that there exist a large number of documents and data contained in such unstructured documents are without any associated schema. The process of IR consists of locating relevant documents on the basis of user query, such as keywords. The World Wide Web (WWW) provides a convenient way to interact with information sources across the Internet. IR has played a critical role in making the web a productive and useful tool, especially for researchers [4] for the purpose of efficiently retrieving relevant documents. Examples of information retrieval system (IRS) are online library catalogs and online document-management systems like storing newspaper articles. Data

of these types are organised as a collection of *documents*. In the context of the web, usually each hypertext terminal markup language (HTML) page is considered to be a document. However, a persistent problem facing the web is the explosion of stored information with little guidance to help the user to locate what is highly interesting in timely manner.

As at today, the WWW technology has grown extremely large in terms of unimaginable usage as information repository for the purpose of knowledge reference. In the web-based IR, the information explosion has created diverse challenges. IR along with other areas like database, web mining techniques, natural language processing (NLP), machine learning etc. can be used to solve the above challenges. Ineffectiveness of IRS is often caused by the query inaccuracy. Retrieving information from the internet using an IRS need precise keywords to achieve the best result because the system requires the exact keywords to return a high quality result lists as thousands of irrelevant documents are returned if the selected keywords are too general. This has become a problem for users when they are not sure about the nature of the content they need or the difficulties of describing the nature of the context of the information needs in just a few keywords [5].

Lastly, vocabulary mismatch is also one of the reasons for the ineffectiveness of the IRS. It is the fundamental problem for the IR [6] and it is a common phenomenon that exists in natural language where the same concept or item has different meaning. To deal with the vocabulary problem, several approaches have been proposed including interactive query refinement, query expansion, relevance feedback, word sense disambiguation, search results clustering and re-ranking. One of the most successful techniques is to expand the original query with other words that best capture the actual user's intent or simply produce a more useful query that is more likely to retrieve relevant documents [7]. These techniques tackle the problem of ineffectiveness in documents retrieval by modifying the query to improve the quality of the query since many believe that the inaccurate query is the major cause of the problem that exists in IR [5]. Hence, this paper reviews the intelligent method of query expansion and fashion out future work on the implementation of intelligent information retrieval using datasets of two test collections (Forum for Information Retrieval Evaluation, FIRE, and ClueWeb) for the purpose of removing "noise" (irrelevant documents) from the lists of retrieved documents. The research will only be limited to FIRE and ClueWeb despite availability of other test collections.

The remainder of this paper is structured as follows. We describe the architecture of IRS and challenges in web-based IR in Section 2. The concepts and challenges of intelligent IR are the focus of Section 3 while Section 4 provides an overview of AQE and PRF. Finally, we conclude the paper with the direction of future research work on implementation of intelligent information retrieval for the purpose of removing "noisy" (irrelevant) documents from the lists of retrieved documents.

2. Information Retrieval System

a. Architecture of IRS

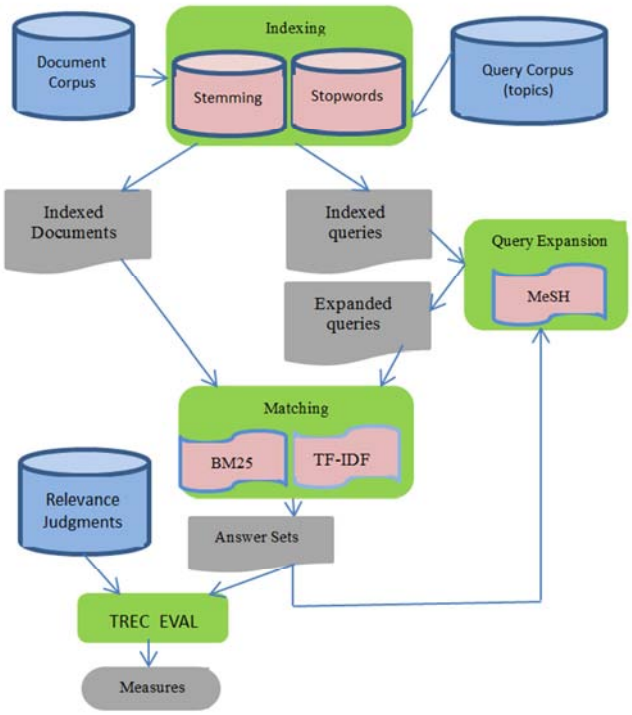

Figure 1. *The Information Retrieval Process [9].*

The basic architecture and the processes of IRS are shown in Figure 1. There are three basic processes an IRS has to support: (i) the representation of the content of the documents, (ii) the representation of the user's information need, and (iii) the comparison of the two representations. Representing the documents is usually called the *indexing* process which takes place off-line and end-user of the IRS is not directly involved. The indexing process results in a representation of the document. Users have a premeditated need for information before searching, the process of representing their information need is often referred to as the *query formulation* process and the resulting representation is the query [8]

Comparing the two representations is known as the *matching* process. Retrieval of documents is the result of this process.

b. Information Retrieval and Search Engines

A search engine is a resource that provides the ability to search information on the Internet. [10] opine that search engines provide three main facilities: (i) gathering of a set of webpages that form the documents which a searcher can retrieve information, (ii) representation of pages in these documents in a way to capture their content, and (iii) allowing searchers to issue queries, then use IR algorithms to find the most relevant pages from these

webpages/documents.

A search engine can gather new pages for its documents in two ways. First, individuals/companies who create webpages may directly contact the search engine to submit their new pages. Second, search engines employ the so called web `spiders/crawlers' which traverse known webpages (link to link) to search new materials. Differences among spiders determine the database of documents that a given search engine accesses as well as the timeliness of its contents [10]. The design of search engine is meant for searching for information on the WWW and the results generated are presented to the user in a list of results commonly called 'hits'.

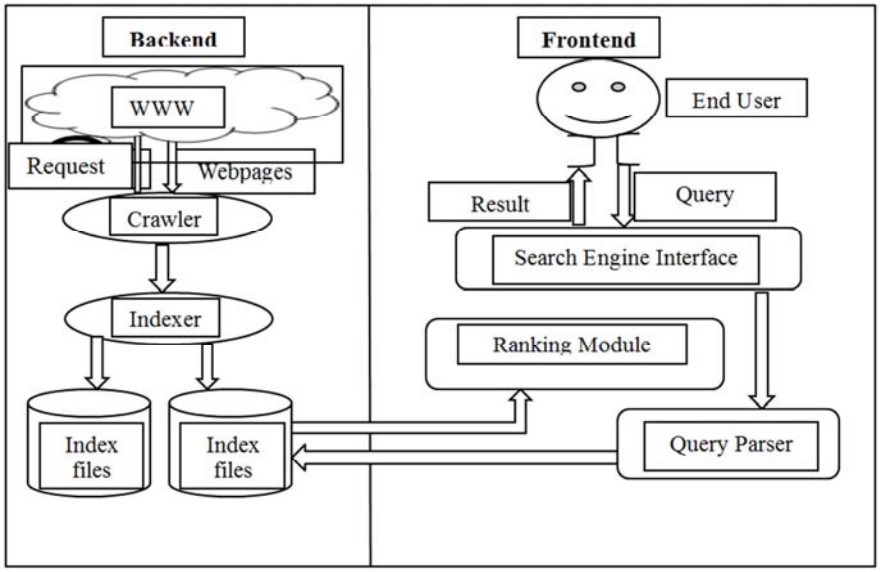

Figure 2. *Web Search Engine Architecture [11].*

In the frontend, we have (i) search engine interface, (ii) query processor and (iii) ranking module. Crawler finds the web via recursively visiting webpages through links between webpages and then downloads the pages. The indexer extracts the keywords from the downloaded webpages and analyse them. Then indexer builds index files having a table of keywords and their corresponding webpages. In the frontend where the user enters a query in a search engine interface, query processor analyses the query and breaks it into keywords.

The query keywords is then matched with index files keywords and then it returns a list of webpages. Finally, ranking mechanism is done by ranking all the returned webpages according to chosen ranking algorithm [11]. This mechanism is handled by the *ranking module* which plays an essential role in web search engine because for ranking search results. For a user's query, it determines the order of the pages in the result. Generally, the order of the webpages depends on *popularity* (or *PageRank*) of the pages. Webpages having high popularity comes at the top of returned results i.e. results should be arranged in descending order of PageRank. PageRank calculation is very critical part of search engine. Hence, search engine uses page ranking algorithms.

c. Challenges in Web-based Information Retrieval

Any IRS attempts to rank documents optimally given a query so that relevant documents would be ranked above non-relevant ones. In order to achieve this goal, the system must be able to score documents so that a relevant document would ideally have a higher score than a non-relevant one [12]. Hence, IRS aims at retrieving the relevant documents according to user's need. Concretely, a search engine computes a similarity between the user's query and the indexed documents; the documents that contain the query terms are retrieved and ordered according to their decreasing similarity with the query [13]. In practice, this problem is usually mentioned as a ranking problem, which aims to be solved according to the degree of relevance (similarity) between each document and the user's query [14]. However, the quality of queries submitted to IRS directly affects the quality of search results generated by these systems [15]. For this reason, the issue of how to improve search queries has been of great interest in IR research. One of the most fundamental research questions in IR is how to operationally define the notion of relevance so that we can score a document with respect to a query appropriately. The most critical language issue for retrieval effectiveness is the term mismatch problem because both the indexers and the users do often not use the same words. This scenario is called *vocabulary problem*. This is compounded by *synonymy* (same word with different meanings, such as "java") and *polysemy* (different words with the same or similar meanings, such as "tv" and "television"). 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 [7]. One solution to this problem is known as *query expansion* (QE) via *pseudo relevance feedback* (PRF).

QE via PRF is an intelligent technique for boosting the overall performance in IR. The technique assumes that top-ranked documents in the first-pass retrieval are relevant and then used as feedback documents in order to refine the representation of the original queries by adding potentially related terms or adjusting the weights of query terms [16].

3. Intelligent Information Retrieval (IIR)

a. IIR Concepts

The concept of IIR was first mooted in the late 1970s but had lost currency within the IR community by the early 1990s. With the popularity of the concept of 'intelligent agents', it implies that the idea of IIR is again in general vogue. IIR is a machine (or program) doing something *for* the user or taking over some functions that previously had to be performed by human beings either user or intermediary [17]. According to [18], IIR is increasingly used in the literature to refer to techniques aimed at the definition of system that offers a flexible access to the huge availability of documents in digital form. Personalized indexing, relevance feedback, text categorization, text mining, cross-lingual information retrieval, question-answering tools, flexible user interfaces are examples of such techniques.

b. Overcoming IR Challenges using Intelligent Information Retrieval

[17] highlighted two limitations of the traditional IRS as follows: (i) it is difficult to understand the needs of the user in order to search for the specified sets of keywords, and (ii) traditional websites or machines do not have the ability to learn and were not capable enough to assume the search criteria; and hence were not able to find the results based on the inputs entered by the users in the past.

The solution to traditional IR challenges has reveals the need for IIR to handle access over the internet, distributed, collaborative and context-sensitive retrieval and these challenges in IIR relate to query representation from user specification, clustering and indexing, classification, question answering issues, meta-search, distributed information retrieval, matching, ranking the result by relevance, language modelling, performance measure and user feedback in terms of recall and precision [1]. IR on the Internet is particularly challenging for the non-expert user seeking technical information with specialized terminology. The user can be assisted during the required search tasks with intelligent agent technology delivered through a decision making support system. Hence, to make IRS more effective, a big deal of research in IR is aimed at trying to add some kind of intelligence to IRSs. A component of an intelligent behaviour is flexibility, intended as the capability of learning a context and adapting to it. This is to adapt to the users' needs: the *certainty* is a way to allow a more natural expression of user's needs, the *precision* is the capability of eliciting from the user her/his actual information preferences [18].

c. The Concepts of Query Expansion and Relevance Feedback

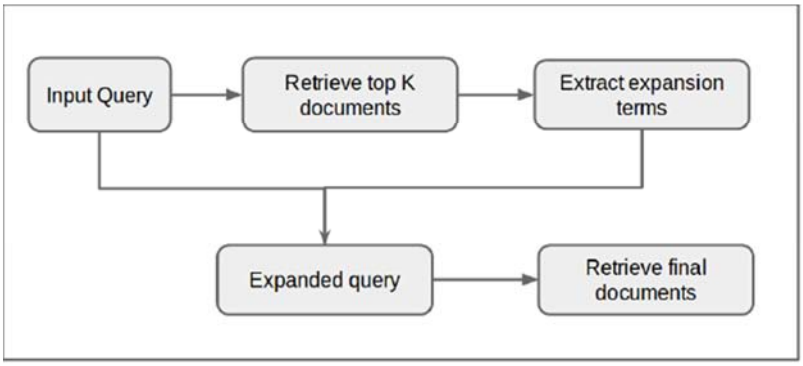

Figure 3. *Architecture of Pseudo Relevance Feedback based system [21].*

In QE, adding additional terms into query can either be automatic, manual or user-assisted. Early approaches to QE were manual in nature as search engine users were asked to extract expansion terms from top-ranked documents and reformulate their own queries. This makes manual QE method not popular since it required user intervention and equally some knowledge of the underlying retrieval system [19]. That is, manual QE depends on user input to decide which terms will be added to the original query while AQE is a technique that relies on the terms weighing. Terms with the highest weight will be added to the original query. A proper weight is needed in order to receive a useful result [5]. As earlier stated, retrieving relevant documents that can fulfil user's need is one of the major challenges in the IRS. One of the most feasible and successful techniques to handle this problem is PRF-based QE, where some top documents retrieved in the first iteration are used to expand the original user query. Considering the above problem, there is a need for automatic PRF-based QE techniques that can automatically reformulate the original user's query. Some years back, it has been observed that the volume of data available online has dramatically increased while the number of query terms searched remained very less [20]. While there has been a slight increase in the number of long queries (five or more words), the most prevalent queries are still those of one, two, and three words. In this situation, the need and the

scope for AQE have increased.

Consequently, IRSs work very well if the user is able to convey his information need in form of query but such query provided by the user is often unstructured and incomplete. An incomplete query hinders a search engine from satisfying the user's information need. In practice, we need some representation which can correctly and completely express the user's information need [21]. If we retrieve just based on occurrence of query terms, we might mark document as irrelevant while such document might well be relevant to the user's information need. Thus such documents cannot be retrieved if query is not modified. Hence, it is intuitive that query needs to be refined and expanded.

Additionally, users tend to input short queries even when the information need is complex. Irrelevant documents are retrieved as answers because of the ambiguity of the natural language (words with multiple senses). If we know that some of retrieved documents are relevant to the query, terms from those documents can be added to the query in order retrieve more relevant documents. This is called *relevance feedback*. Often, it is not possible to ask the user to judge the relevance of the retrieved documents. In this case PRF methods can be used where it is assumed that the first few retrieved documents are relevant and use the most important terms from them to expand the query [22].

The effectiveness of IRS is usually evaluated taking into account both *recall* and *precision*. Using a combined recall/precision measure, the overwhelming majority of recent experimental studies agree that AQE results in better retrieval effectiveness, with improvements of the order of 10% and larger (e.g., [23], [24], [25], [26]). Such findings are important to support the claim that AQE is an effective technique, but this may be not sufficient for the cases when we are primarily interested in precision. However, several recent studies have pointed out that AQE does not necessarily hurt precision [7].

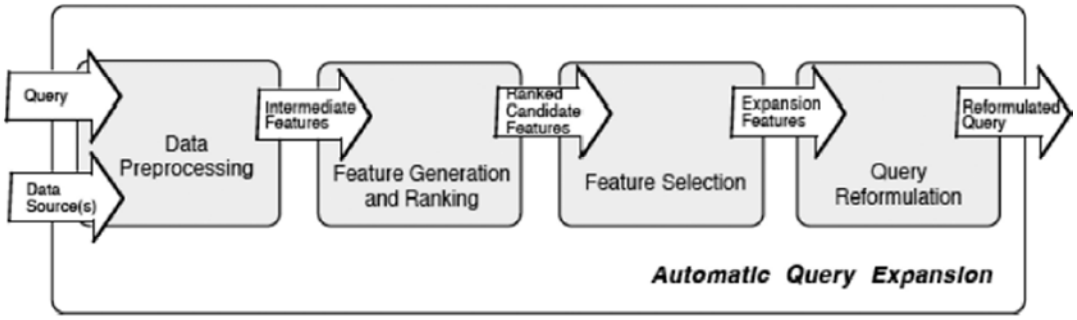

Figure 4. *Main Steps of Automatic Query Expansion [7].*

Sometimes, AQE achieves better precision in the sense that it has the effect of moving the results toward the most popular or representative meaning of the query in the collection at hand and away from other meanings; e.g., when the features used for AQE are extracted from Webpages [24], 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]. AQE 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.

d. Automatic Query Expansion Techniques

An age-long issue in the field of IR is the word mismatch between query and documents. Hence, the alternative strategies for solving the vocabulary problem are the use of different AQE techniques. According to [7], some of the other related techniques of QE are (i) interactive query expansion/refinement, (ii) relevance feedback, (iii) search results clustering, (iv) thesaurus, and (v) semantic network. The modification of the search process is to improve the effectiveness of an IRS by incorporating information obtained from prior relevance judgments. The basic idea is to do an initial query, get feedback from the user (or automatically) as to what documents are relevant and then add term from known relevant document(s) to the query.

In *automatic* relevance feedback (also called *pseudo* or *blind* relevance-feedback), only the few top-ranked retrieved documents are treated as relevant, without any user involvement; if (in addition), the few bottom-ranked retrieved documents are treated as non-relevant and they do participate in the feedback, we have *full automatic relevance-feedback* [29]. In other words, PRF is one of the three approaches in relevance feedback where the candidate terms for expanding the user's query are selected from an initially retrieved set of documents. Therefore, the initial step is to retrieve the top n documents from the corpus. The retrieved documents are considered to be relevant.

4. Fuzzy Principles for Improving Information Retrieval

Fuzzy concepts affect most phases of IR process. They are deployed during document indexing, query formulation and search request evaluation. In general, document is interpreted as a fuzzy set of document descriptors and queries as a composite of soft search constraints to be applied on documents. Document-query evaluation process is based on fuzzy ranking of the documents in documentary collection according to the level of their conformity to the soft search criteria specified via user queries. The document-query matching has to deal with the uncertainty arising from the nature of fuzzy decision making and from the fact that user information needs can be recognized, interpreted, understood

only partially and the document content is described only in a rough, imperfect way [30].

In the fuzzy-enabled IR frameworks, soft search criteria could be specified using linguistic variables. User search queries can contain elements declaring level of partial importance of the search statement elements. Linguistic variables such as "probably" or "it is possible that", can be used to declare the partial preference about the truth of the stated information. The interpretation of linguistic variables is then among the key phases of query evaluation process. Term relevance is considered as a gradual (vague) concept. The decision process performed by the query evaluation mechanism computes the degree of satisfaction of the query by the representation of each document. This degree, called Retrieval Status Value (RSV), is considered as an estimate of the relevance of the document with respect to the query. RSV = 1 corresponds to maximum relevance and RSV = 0 denotes no relevance. The values within the range (0; 1) correspond to particular level of document relevance between the two extremes 0 and 1 [31].

Possibility theory together with the concept of linguistic variable defined within fuzzy set theory provides a unifying formal framework to formalize the processing of imperfect information. Inaccurate information is inevitably present in IRS and textual databases applications. The automatically created document representation based on a selection of index terms is invariably incomplete and far worse than document representations created manually by human experts who utilize their subjective theme knowledge when performing the indexing task. Automated text indexing deals with imprecision since the terms are not all fully significant to characterise the document content and their statistical distribution does not reflect their relevance to the information included in the document necessarily. Their significance depends also on the context in which they appear and on the unique personality of the inquirer. During query formulation, users might have only a vague idea of the information they are looking for and therefore face difficulties when formulating their information needs by the means of query language of particular IRS. A flexible IRS should be designed to provide detailed and rich representation of documents, sensibly interpret and evaluate soft queries and hence offer efficient IR service in the conditions of vagueness and imprecision [30]. Hence, the use of fuzzy logic controller in the implementation of AQE via PRF is imperative in search for relevant documents.

5. Conclusion and Future Work

The quest for retrieving not all but only relevant documents in the list of retrieved documents remains vital to both the users and researchers of IRS. Although solutions to vocabulary problem still remains totally unsolved, but AQE which is an intelligent method of improving the performance of IRS via PRF has been proposed to overcome the problem. The advance of AQE techniques has been confirmed by a number of experimental tests on classical benchmark. There exist reports of evaluation studies indicating not only in *recall* but also in *precision* with remarkable improvements in average retrieval effectiveness.

Our future research will attempt to study the performance of IIR for retrieving relevant documents using fuzzy logic method on two test collections: Forum for Information retrieval Evaluation (FIRE) and ClueWeb datasets

References

[1] Sharma Manish & Patel Rahul (2013). A Survey on Information Retrieval Models, Techniques and Applications. *International Journal of Emerging Technology and Advanced Engineering*, volume 3, Issue 11, pp. 542-545.

[2] Mohameth-François Sy., Sylvie Ranwez., Jacky Montmain., Armelle Regnault., Michel Crampes & Vincent Ranwez Pezzoli (2012). User-centered and ontology based information Retrieval system for life sciences, *BMC Bioinformatics*, 1471-2105.

[3] Akram Roshdi & Akram Roohparvar (2015). Review: Information Retrieval Techniques and Applications. *International Journal of Computer Networks and Communications Security* vol. 3, No. 9 pp. 373-377.

[4] Silberschatz, A., Henry F. Korth. & S. Sudarshan (2011). Database System Concepts. 6th Edition. Published by McGraw-Hill. Pg 915-943.

[5] Jessie Ooi, Xiuqin Ma, Hongwu Qin & Siau Chuin Liew (2015). A Survey of Query Expansion, Query Suggestion and Query Refinement Techniques. Proceedings of 4th International Conference on Software Engineering and Computer Systems (ICSECS), Kuantan, Pahang, Malaysia. pp. 112-117.

[6] Xu, J. & Croft, W. B. (1996). Query expansion using local and global document analysis. In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, 4–11.

[7] Carpineto, C. & Romano, G. (2012). A survey of automatic query expansion in information retrieval. *ACM Computing Surveys*, Vol. 44, No. 1, Article 1 pp.1-50.

[8] Hiemstra Djoerd (2009). Information Retrieval Models. Published in Goker, A., and Davies, J. Information Retrieval: Searching in the 21st Century. John Wiley and Sons, Ltd., ISBN-13: 978-0470027622, Pg 1-2.

[9] Rivas, A. R., Iglesias, E. L. & Borrajo, L. (2014). Study of Query Expansion Techniques and Their Application in the Biomedical Information Retrieval, *The Scientific World Journal*. Volume 2014, Article ID 132158 10 pages http://dx.doi.org/10.1155/2014/132158

[10] Fan W, Gordon M. D, & Pathak P (2000) Personalization of search engine services for effective retrieval and knowledge management. Proceedings of International Conference on Information Systems (ICIS), Brisbane, Australia.

[11] Rawat Poonam., Shri Prakash Dwivedi & Haridwari Lal Mandoria. (2014). An Adaptive Approach in Web Search Algorithm. *International Journal of Information & Computation Technology*. Vol. 4, No. 15, pp. 1575-1581.

[12] Zhai, C. X. (2008). Statistical language models for information retrieval a critical review. Found. *Trends Inf. Retr.*, 2:137-213, March.

[13] Ermakova Liana & Mothe Josiane (2015). Query Expansion by Local Context Analysis. Insitut de Recherche en Informatique de Toulouse, Perm State National Research University.

[14] Maitah Wafa., Mamoun. Al-Rababaa & Ghasan. Kannan (2013). Improving the Effectiveness of Information Retrieval System using Adaptive Genetic Algorithm. *International Journal of Computer Science & Information Technology,* Vol 5, No 5, pp. 91-105.

[15] Croft, W. B. & Thompson, R. H. (1987). I3R: A new approach to the design of document retrieval systems. *Journal of the American Society for Information Science*, 38(6): 389-404.

[16] Ye Zheng & Huang Jimmy Xiangji (2014). A Simple Term Frequency Transformation Model for Effective Pseudo Relevance Feedback. ACM SIGIR '14, July 6-11, pp. 323-332.

[17] Ahuja Sudhir & Goyal Rinkaj (2012). Information Retrieval in Intelligent Systems: Current Scenario & Issues. *International Journal of Computer Engineering Science* (IJCES), Volume 2 Issue 5, pp. 1-8.

[18] Pasi Gabriella (2003). Intelligent Information Retrieval: Some Research Trends. *Advances in Soft Computing*, pp. 159-171, DOI: 10.1007/978-1-4471-3744-3_16

[19] Zhou Dong, Mark Truran, Jianxun Liu & Sanrongf Zhang (2013). Collaborative Pseudo-Relevance Feedback. *Expert Systems with Applications*, 40, 6805-6812.

[20] Singh J. & Sharan A. (2015) Context window based co-occurrence approach for improving feedback based query expansion in information retrieval. *Int J Inform Retr Res* 5(4):31–45.

[21] Kankaria Ashish (2015). Query Expansion. Indian Institute of Technology Bombay, Mumbai.

[22] Inkpen Diana (2010). Information Retrieval on the Internet. A Lecture Note, University of Ottawa, Canada, K1N 6N5, pp. 1-30.

[23] Mitra, M., Singhal, A. & Buckley, C. (1998). Improving automatic query expansion. In *Proceedings of the 21st Annual International ACM SIGIR* Conference on Research and Development in Information Retrieval. ACM Press, 206–214.

[24] Carpineto, C., Romano, G. & Giannini, V. (2002). Improving retrieval feedback with multiple term-ranking function combination. *ACM Trans. Info. Syst. 20*, 3, 259–290.

[25] Liu, S., Liu, F., Yu, C., & Meng, W. (2004). An effective approach to document retrieval via utilizing wordnet and recognizing phrases. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 266–272.

[26] Lee, K. S., Croft, W. B. & Allan, J. (2008). A cluster-based resampling method for pseudo-relevance feedback. In Proceedings of the 31th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, 235–242.

[27] Cui, H., Wen, J. R., Nie J. Y. & Ma, W. Y. (2003). Query expansion by mining user logs. *IEEE Trans. Knowl. Data Engin. 15*, 4, 829–839.

[28] Chu, W. W., Liu, Z. & Mao, W. (2002). Textual document indexing and retrieval via knowledge sources and data mining. *Comm. Institute of Info. Comput. Machinery 5*, 2.

[29] Xu Yunjie (2005). Information Retrieval with a Hybrid Automatic Query Expansion and Data Fusion Procedure. *Information Retrieval*, 8, 41–65, 2005

[30] Alhenshiri, Anwar A. (2013). Web Information Retrieval and Search Engines Techniques. *Al-Satil Journal*, PP: 55-92.

[31] Bordogna Gloria & Pasi Gabriella (2001). Modelling vagueness in information retrieval. Lectures on Information Retrieval, volume 1980, pp. 207-241.