I.J. Modern Education and Computer Science, **2016, 2, 49-53**
Published Online February 2016 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2016.02.06

Fuzzy Latent Semantic Query Expansion Model for Enhancing Information Retrieval

Olufade F.W Onifade
Department of Computer Science, University of Ibadan, Nigeria.
E-mail: fadowilly@yahoo.com

Ayodeji O.J. Ibitoye
Department of Computer Science and Information Technology, Bowen University, Nigeria.
E-mail: ibitoye_ayodeji@yahoo.com

Abstract—One natural and successful technique to have retrieved documents that is relevant to users' intention is by expanding the original query with other words that best capture the goal of users. However, no matter the means implored on the clustered document while expanding the user queries, only a concept driven document clustering technique has better capacity to spawn results with conceptual relevance to the users' goal. Therefore, this research extends the Concept Based Thesaurus Network document clustering techniques by using the Latent Semantic Analysis tool to identify the Best Fit Concept Based Document Cluster in the network. The Fuzzy Latent Semantic Query Expansion Model process achieved a better precision and recall rate values on experimentation and evaluations when compared with some existing information retrieval approaches.

Index Terms—Concepts, Concept Based Thesaurus Network, Latent Semantic Analysis, Information Retrieval System, Information Retrieval, Best Fit Concept Based Document Cluster.

I. Introduction

Since the internet has become a bank for all information, instead of going to the local libraries, nowadays people search the net for the information they need. As a result, there have been rapid growths in the volume of online text documents from network resource such as Internet, digital libraries and company-wide intranets [1]. This rapid development keeps posing great challenges to information retrieval experts when it comes to organization, management and retrieval of relevant document in respect to users query. This is because having data from different sources as made the world witnessed an exponential growth of digital information triggered by continuous development of information technology (IT) in the past years [2]. As part of the means to obtain a more relevant document from query posted by different users, Query refinement [3] became an essential information retrieval approach that interactively recommends new terms related to a particular user query in other to improve the quality of such query. Although, from the perspective of a user, the traditional way to meet her information need would be to perform an exhaustive review of contents from physical and digital documents [4]. However, with query refinement, whenever users interact with a retrieval system, the system provides term excerpts that are considered relevant to a particular user query as an ideal behavior. These retrieved documents though relevant to user query, may not thoroughly reflect the user's information need and sometimes the retrieved documents are also not relevant. It is not unusual that this existing methods engenders irrelevant documents since it cannot effectively interpret what users want and need. More so, in a situation wherein a concept that is fit for query refinement can be expressed by different word or a word have different meanings [5], it is possible that information that does not meet the users' needs are retrieved. These factors have made search to be ambiguous, boring, frustrating and unsatisfactory. In solving this problem, a new generation of information retrieval model [6] has been drawn from the world cognitive view that is based on the conjecture that the meaning of a text (word) depends on conceptual relationships to objects in the world rather than to linguistic or contextual relations found in texts or dictionaries. In this view, sets of words, names, noun-phrases, terms, etc. are mapped to encoded concepts. This process has been widely accepted has Concept-Based Information Retrieval [7]. Therefore, since relevance is subjective to different users, this research suggest that information retrieval systems must be able to retrieve documents that will meet users' search goal at different search times. Hence, to achieve this objective, retrieved documents must be structured from document clustering that is concept driven in order to retrieve documents that are specific to the need of a user. Thus, section two discussed on related works while section three gave an overview of the concept based thesaurus network (CBTN) and it role in the proposed research. However, in section four, an insight to what the research stands to achieve from the identified problems was illustrated while section five stated the methodology implored before experimentation and resulted obtained was showcased in section six. In section seven, we gave a summary of work done with detailed evaluation.

II. Related Works

Information retrieval (IR) is a broad area in Computer Science and one of its goal is on fulfilling the user need in finding information of their interest. Thus, with the increase in usage of Web search engines it is easy to develop a system [8] that extracts probabilistic correlations between query terms and documents term by collecting and using the user query logs. The process of determining the relevance of a retrieved document to a user hinders it performance since these correlations are used to find good expansion terms for new queries based on the assumption that the documents visited by users are relevant to the query. In semantic search [9] identified four approaches that can be applied. Contextual analysis, which emphasizes on how to disambiguate queries, reasoning, which can infer additional information from existing facts in the system. Natural language understanding, which aim to identify the entity in a sentence and ontology, which enrich the retrieval of specific domain related. With the new semantic search algorithms spread which account not only for keywords as present in the tradition keyword search system [10] but for semantic entities, relations, personalized information and many more. One fundamental problems associated with this techniques is query drift. Query drift is caused as a result of adding terms which have no association with the topic of relevance of the query. Ref. [11] introduced a subset of important terms Instead of using all the terms obtained through feedback for query refinement while [12] proves that a combination of semantic information from a Linked Data graph can lead to an improved ranking. Also [13] proposed a positional relevance model which assign more weight to the terms in the document which are nearer to the query terms while [14] describes a method to enrich search queries via a conjunctive extension based on the underlying semantic ontology. The method is able to retrieve documents provided only with a description instead of a search query. This leads to results without an overlap of keywords between query and document. PageRank algorithm was extended using Linked Data knowledge [15] to extract semantic knowledge from a Web document. However, this version of the algorithm is not able to scale on Web data

III. Overview of Concept Based Thesaurus Network

Unlike other document clustering techniques that has been used for information retrieval, the Concept Based Thesaurus Network (CBTN) is a document clustering technique that has the capacity to engender documents that contains a higher degree of similarity, lower level of disorderliness, and low entropy between concepts and documents in the documents clustered from user query at different search times [16] [17]. The conceptual network implored in this process caters predominantly for all documents in the document cluster that has the presence of the user query and it associated synonyms [18]. Queries are restructured based on concepts using Fuzzy Concept Network (FCN) relationships [19] in conjunction with other parameters like term frequency, degree of relationship, and degree of effect between terms. With CBTN, various concepts contain quite a number of documents while these documents belong to various concepts with some level of associated degrees [20]. Moreover, concepts are also linked together with some level of fuzzy positive, negative, special or general degrees using the fuzzy concept network standard rules. CBTN which forms the basis for our research has been proven to be an important framework in Concept Based Information Retrieval Systems (CBIRS). This is because it clusters all documents with the presence of the user query, but yet provides a significant degree and level of association between concepts and concepts, concepts and documents, and documents and documents respectively for better and relevant document retrieval that is aimed at satisfying users' intention.

IV. Research Motivation

The CBTN document clustering techniques has only provided a platform for retrievable documents to be concept driven. It has not decrease the volume of search that was generated by the user query to form the clustered documents; neither has it decreased the search time to get the potential result. More so, the precision and recall rates of clustered documents seem to have little or no difference with existing keywords and automatic thesaurus construction based document clustering technique. Hence, we are encouraged to increase the precision and recall rates, reduce the search volume by extracting from the CBTN the cluster that is more relevant to the users' query. The ability to achieve this foresight will no doubt reduce query search time while also increasing the relevance ratio of retrieved documents based on users' query.

V. Proposed Fuzzy Latent Semantic Query Expansion Model

Here, aside the initial process of generating the CBTN through users query, this research uses the Latent Semantic Analysis tool on the CBTN, the document represented in the corpus and the terms that are presented in the users query for the second time. Fig.1 gives a proper illustration on how Latent semantic analysis (LSA) is used to manipulate the trio elements in order to identify the concept cluster that is more similar to the user query. The extracted unique cluster from the CBTN is tagged the Best Fit Concept Based Document Cluster (BFCBDC).

From fig 1, LSA is used to find the level of association that coexists between the trios of users 'query, CBTN and the document representation. The LSA accepts the user query at once while there exists a bi-directional flow of process between the LSA, CBTN and the document representation until the last concept and document is considered. The action of LSA is aided through the construction of the following matrixes.

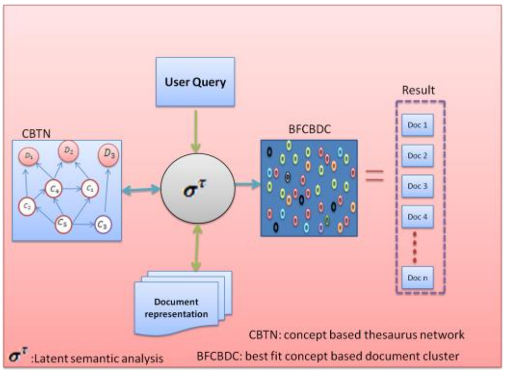

Fig.1. Diagram Showing the Best Fit Document Cluster

1. **Relevance matrix**; this is a query (term) by document matrix TD, where each column describes a document, each row a query (term).

$$R_m = \begin{array}{c} \\ t_1 \\ t_2 \\ \vdots \\ t_n \\ \\ \end{array} \left\{ \begin{array}{ccc} d_1 & d_2 \dots\dots.. & d_n \\ V_{11} & V_{12} \dots\dots. & V_{1n} \\ V_{21} & V_{22} \dots\dots. & V_{2n} \\ \vdots & \vdots & \vdots \\ V_{n1} & V_{n2\dots\dots\dots\dots} & V_{nn} \end{array} \right\}$$

It is also a fuzzy matrix where the element V_{ij} represents the relevance degree between term t_i and document d_j , $V_{ij} \epsilon [0,1]$ and n is the number of documents.

2. **Document Pointer matrix:** this is a document by concept matrix DC, where each row describes a document, each column a concept; representing each concept as a vector in the Cartesian coordinate system for the standard vector space. Here we have a set of document D = $(d_1, d_1 \dots\dots\dots\dots\dots\dots\dots.d_m)$ and a set of concept C = $(c_1, c_1 \dots\dots\dots\dots\dots\dots\dots.c_n)$

$$T = \begin{array}{c} \\ d_1 \\ d_2 \\ \vdots \\ d_m \\ \\ \end{array} \left\{ \begin{array}{ccc} c_1 & c_2 \dots\dots.. & c_n \\ t_{11} & t_{12} \dots\dots. & t_{1n} \\ t_{21} & t_{22} \dots\dots. & t_{2n} \\ \vdots & \vdots & \vdots \\ t_{m1} & t_{m2\dots\dots\dots\dots} & t_{mn} \end{array} \right\}$$

where m is the number of documents n is the number of concepts and t_{ij} is the relative relevance degree between document and concepts., $t_{ij} \epsilon [0,1]$, $1 \le i \le m$, $1 \le j \le m.$

These matrixes allow LSA to find more meaningful query (term) by document and document by document similarity test between terms, concepts and documents. It also helps to get a more differentiated view on the data set and its underlying relations while the obtained output from this research is the BFCDC. The Best Fit Concept Document Clustered as shown in fig 1 above is the best concept cluster from the CBTN that is more relevant to the user query. The process of identifying the BFCDC is the real essence of this research. Therfore, the algorithm below is presented to illustrate the procedures of LSA on the trio of user query, CBTN and document representation has described in fig 1.

1. Input: the user query-terms set Q = $\{q_1, q_2, \ldots, q_n\}$, the constructed concept based thesaurus network and the respective document representation from the database.
2. Let K be the set of concepts from the concept based thesaurus network; k = $(c_1, c_2, \dots\dots\dots\dots\dots..c_n)$
3. Let P be the relevant degree vector of set K, where P = $(dr_1, dr_2, \dots\dots\dots\dots\dots..dr_n)$
4. Let D be the set of documents that contains the user query term and other concepts from the CBTN
5. Decompose the co-occurrence matrix M between set of query Q and set of document D to find their level of association using $DTD_K = \ T_K S_K^{\frac{1}{2}}$
6. Decompose the co-occurrence matrix D between set of document D and set of concept to find their level of association using $DCD_K \ = \ D_K S_K^{\frac{1}{2}}$
7. Choose query term q_i that has the largest Inverse Dependent Frequency (IDF) value form the query term set Q (assume that f_q denotes the IDF value of term q_i) and let that term q_i be the center of the query expansion.
8. Find the cosine values of every query term q_i with respect to the set of concept c_i and set of documents D respectively using $CD_K = D_K S_K$
9. Recompose the matrixes to get the relative relevance degree of interception between user query, concept and document using $M = DC^t * T_k * S_k^+ * D_k$ Where S_k^+ is the inverse of S_k
10. Find document clusters from training document clusters D for which their cluster center contains users query q_i , wherein the degree of effect of concepts from it set of concept based thesaurus network is larger than the other terms in the cluster center.
11. Output: Ranked best fit concept based retrieved documents

Algorithm. 1: Algorithms for BFCDC

It is from the drived Best Fit Concept Document Cluster that we retrieve document that are relevant and closer to the users intention.

VI. Experiments and Results

To quantify how good the retrieval performance of an Information Retrieval (IR) system is, two basic measures of retrieval performance are used in the IR community. Recall is the fraction of relevant documents that are retrieved by the system out of the entire collection, while Precision is the fraction of documents retrieved by the system that are relevant. Thus, the research experiment unfolds the precision and recall rate of Fuzzy Latent Semantic Query Expansion Model (FLSQEM) while the obtained result is compared with the traditional keyword based search information retrieval technique and automatic query expansion (AQE) retrieval technique. In the next section we show the experiment conducted and graphical representation of the obtained result. The tested queries are illustrated in table 1. These queries are generated to retrieve relevant document from a set of clustered documents in the field of computer science.

Table 1. Sample Queries for System Testing

S/N	Queries
Q_1	Explain data structure
Q_2	Network security and types
Q_3	What is Software management
Q_4	How do I declare Variables in programming?
Q_5	Difference between Artificial intelligence and neural networks
Q_6	Network protocols
Q_7	Possible programming operators with definition
Q_8	Human computer interaction
Q_9	What is transaction management
Q_{10}	Why is database security important

A. Precision rate

Precision measures how many of the documents retrieved in a search process are actually relevant, that is, how much of the result set is relevant in the retrieved documents based on user query. For example, a 75 percent precision rate means that 75 percent of the documents retrieved are relevant, while 25 percent of those documents have been misidentified as relevant**.** In calculating the precision rate for the user queries, we use the (1) as illustrated below:

$$Precision\ rate\ =\ \frac{R_r}{R_d} \qquad (1)$$

Where R_r denotes the number of relevant retrieved document, R_d denotes the number of retrieved document. So, fig. 2 is used to showcase the precision rate of traditional keywords based information retrieval technique, automatic query expansion (AQE) information retrieval technique and the proposed FLSQEM in performance comparison.

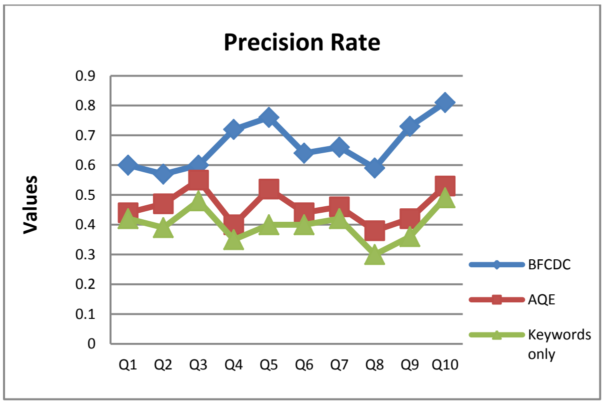

Fig.2. Comparative Precision Rate

In fig.2, the precision rate of our proposed methodology hovers above the threshold value of 0.5. The value 0.5 was chosen because the degree of relevance of any retrieved documents to user query must be between 0 and 1. The obtained result as indicated in fig. 2 shows that the proposed FLSQEM performs better than the existing keyword and automatic query expansion retrieval technique respectively.

B. Recall rate

Recall measures how many of the relevant documents in a collection have actually been retrieved or assessed as relevant, that is, how much of the target set has been found based on users query. For example, a 50 percent recall rate means that 50 percent of all relevant documents in a collection have been found, and 50 percent have been missed. In finding the recall rate from table 1, we apply (2)

$$\text{Recall rate}\ =\ \frac{\mathrm{R_r}}{\mathrm{R_e}} \qquad (2)$$

Where R_r denotes the number of relevant retrieved document, and R_e connotes the number of relevant documents in the collection. The fig. 3 shows the recall rate of 10 different user queries as presented in table 1 Thereafter, we analyze the result obtained from the values on the graph.

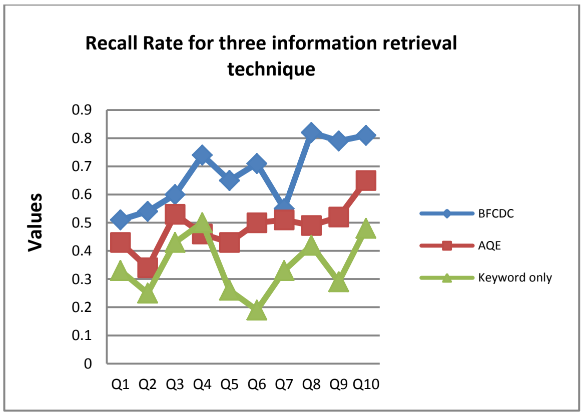

Fig.3. Recall Rate in Comparison

In fig. 3, we saw that the recall value for the proposed FLSQEM is above the threshold value of 0.5. The value of 0.5 was set as benchmark for the recall rate since the relevance degree of any retrieved document to users query must be between 0 and 1. Based on the queries in placed in table 1, the recall rate recorded a 100% result when compared with the keyword search and automatic query expansion information retrieval technique.

VII. Conclusion

In this paper, an improved method of document clustering for automatic query expansion tagged concept based thesaurus network was extended in other to obtain from it an optimal cluster that is more relevant to the users query. Most of the methods implored in information retrieval only select expanded terms based on their frequency in the relevant document set and the meaning of words is not directly brought into account. Meanings, which are subjective was the major concern for the formation of concept based thesaurus network and the motive to minimize search time, reduce search volume and extract from the network a more meaningful document is why we introduce the use of latent semantic indexing on the concept based thesaurus network. At the end, we have been able to obtain to a satisfactory state of result that portrays users' intention from experimented queries. However, the efficiency of this work can be improved by adding more issues that will generate a more optimal result. Due to different existing page rank technique, direct incorporation of vital page rank technique as part of algorithm that will extract the Best Concept Cluster from the network may be required. Rather than the Latent Semantic Analysis approach which was implored, the methodology can be modified to enhance a better and efficient result. Our goal must however remain the need to display hierarchically the Best Fit Concept Based Document Cluster that will reflect or indicate the users' intention to a large extent.

REFERENCES

[1] D Christopher, Manning, P.R., Schtze, H.: An Introduction to Information Retrieval. Cambridge University, 2009.

[2] P. H. Cleverley and S. Burnett, "Retrieving haystacks: a data driven information needs model for faceted search," Journal of Information Science, vol. 41, no. 1, pp. 97–113, 2015.

[3] C. D Manning, P. Raghavan, and H Schutze,. Introduction to Information Retrieval. Cambridge University Press, 2008.

[4] M. C. de Andrade and A. A. Baptista, "Researchers' information needs in the bibliographic database: A literature review," Information Services and Use, vol. 34, no. 3, pp. 241–248, 2014.

[5] Hele-Mai and Tanel Mauri, A survey of concept based information retrieval tool, 2008.

[6] B He, and I Ounis. Studying Query Expansion Effectiveness. Proceedings of ECIR'09 European Conference in Information Retrieval, 2009.

[7] H Cui, J Wen,.R., J.-Y Nie, and Ma, W.-Y. Query expansion by mining user logs. Knowledge and Data Engineering, IEEE Transactions on, 15(4):829–839, 2003.

[8] G. Sudeepthi, G. Anuradha, P. M. Surendra, and P. Babu, "A Survey on Semantic Web Search Engine," vol. 9, no. 2, pp. 241– 245, 2012.

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

[10] G Cao, J.-Y Nie, J Gao, and Robertson, S. Selecting good expansion terms for pseudo-relevance feedback. In SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pages 243{250, New York, NY, USA. ACM. 2008.

[11] Xin He and Mark Baker. xhrank: Ranking entities on the semantic web. In ISWC Posters & Demos'10.

[12] Y Lv, and C Zhai. Positional relevance model for pseudo-relevance feedback. In Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval, SIGIR '10, pages 579{586, New York, NY, USA. ACM. 2010.

[13] Axel Ngonga. Generating conjunctive queries for keyword search on rdf data. In In Sixth ACM WSDM (Web Search and Data Mining) Conference, 2013.

[14] Julia Stoyanovich, Srikanta J. Bedathur, Klaus Berberich, and Gerhard Weikum. Entityauthority: Semantically enriched graph-based authority propagation. In WebDB, 2007.

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

[16] K Sugimoto, H Nishizaki, and Y Sekiguchi, "Effect of document expansion using web documents for spoken documents retrieval," in Proceedings of the 2nd Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, pp. 526–529, 2010.

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

[18] M Beigbeder, and A. Mercier. An information retrieval model using the fuzzy proximity degree of term occurences. Proceedings of SAC '05. New York, USA: ACM Press. 2005.

[19] F Clarizia, L Greco, P Napoletano: A new technique for identification of relevant web pages in informational queries results. In: Proceedings of the 12th International Conference on Enterprise Information Systems: Databases and Information Systems Integration. 8-12 June 2010.

[20] T Akiba. and K Honda., "Effects of query expansion for spoken documnet passage retrieval," in Proceedings of International Conference on Speech Communication and Technology, pp. 2137–2140, 2011.

Authors' Profiles

Olufade F.W Onifade obtained a PhD in computer science from Nancy 2 University, Nancy, France in 2009. He is currently a Senior Lecturer at the Computer Science department, University of Ibadan, Ibadan, Nigeria. He has published over 70 papers in both local and International referred journals and conferences and has held several fellowships including ETTMIT and the CV Raman Fellowship for African Researchers in India. His research interests include Fuzzy Learning, Information Retrieval, Biometrics and Pattern Matching. Dr. Onifade is a member of IEEE, IAENG and CPN.

Ayodeji O.J Ibitoye lectures at the Department of Computer Science and Information Technology, Bowen University, Iwo, Nigeria. He obtained his B.Sc. and M.sc Computer Science from the prestigious University of Ilorin, Ilorin Kwara State and University of Ibadan, Ibadan, Oyo state in 2009 and 2014 respectively. He is a young innovative and resourceful researcher with great analytic and programming skills. His Research Interest is in Big Data Analytics, Information Retrieval, Biometrics and Fuzzy learning. He has several peered reviewed journals and conferences in his field of expertise.