International Journal of Civil Engineering and Technology (IJCIET)
Volume 8, Issue 8, August 2017, pp. 910–915, Article ID: IJCIET_08_08_095
Available online at http://http://www.iaeme.com/ijciet/issues.asp?JType=IJCIET&VType=8&IType=8
ISSN Print: 0976-6308 and ISSN Online: 0976-6316

© **IAEME** Publication **Scopus** Indexed

ENHANCED QUERY EXPANSION FOR WEB INFORMATION RETRIEVAL

Dr. Ramesh Shahabadkar, Y. Vijaya Bhaskar Reddy
Department of Computer Science & Engineering
Vardhaman College of Engineering, Hyderabad, India

BSS. Murali Krishna, T. Durga Devi
Department of Computer Science & Engineering
MLR Institute of Technology, Hyderabad, India

ABSTRACT

Due to the increase in demand of web documents, finding the required document is an important issue for the user. The user is unable to access the relevant documents due to inappropriate query and improper knowledge. In order to enhance the searching efficiency of relevant documents, the original user query is needed to be reformulated. In this paper, a novel Enhanced Query Expansion based Classifier (EQC) technique is proposed for web document retrieval. It uses feedback based documents for query expansion, reformation and optimization. First, the relevant documents for the given query are obtained by means of NTCIR-6 algorithm. From that, the topmost k relevant documents are selected to represent the feature terms. Then the documents are classified using Naïve Bayes classifier which provides good feedback documents. Then the unique terms are extracted and they are ranked using co-occurrence based approach. Rocchio Algorithm was implemented to reweight the unique terms and to expand the query. By using Binary Group Search Optimizer (BGSO) Algorithm, optimum query is selected for document retrieval. The original query is reformulated and feedback is given to the dataset for searching the relevant document. The efficiency of expanded query is measured in terms of Precision (P), Recall (R), F-measure and Mean Average Precision (MAP). The precision, recall and F-measure values are increased with 1%, 3% and 3.5% respectively. The performance measures show the improvement in relevant document retrieval scheme with reduced computational complexity.

Key words: Retrieval Feedback, Query Expansion, optimization, pseudo relevant documents, Naïve Bayes classifier, Co-occurrence approach.

Cite this Article: Enhanced Query Expansion For Web Information Retrieval, Dr. Ramesh Shahabadkar, Y. Vijaya Bhaskar Reddy, BSS. Murali Krishna and T. Durga Devi, International Journal of Civil Engineering and Technology, 8(8), 2017, pp. 910–915.
http://www.iaeme.com/IJCIET/issues.asp?JType=IJCIET&VType=8&IType=8

1. INTRODUCTION

The web has turned out to be fundamental means for many individual's regular activities and owing to its broad search engines, has turned into an essential tool for retrieving and searching data. The process of storing and accessing large amount of data has become a challenging issue due to the continuous growth of online data. To enhance data, users search experience, major Websites provide new options [2]. Web information retrieval system is used to give the useful information based on the user requirements [3].

Information Retrieval (IR) is a division of computer science which manages storage, maintenance and searching data within huge volume of data. The data contains audio, video and all kind of text documents [4].The most critical use of web search is Information retrieval (IR) which is referred as to finding a list of files which are related to the user query [5]. Keyword is a user query in many information retrieval systems to retrieve data. In this keyword based query model, keywords are extracted from the documents and different methods applied for keywords to assign weight [6].

Google and Yahoo are the famous search engine to retrieve data from internet. Traditionally, users enter the keyword to this search engine and the search engine provides all the web pages which are matched to the keyword string [7]. To determine document-query matching two theoretical models are introduced which are vector space and probabilistic models [8]. These models provide all the relevant data to the user from the data collection. User's information needs are satisfied by the search engine [9-10].

One of the difficulties in keyword search is that the client utilizes distinctive words in the inquiry than the descriptors utilized for indexing. Another test is that clients regularly give a short, ambiguous or badly shaped question. Keeping in mind the end goal to discover important outcomes, the question must be extended with relevant, related words, for example, synonyms [11].

Query expansion (QE) is used to enhance retrieval performance in information retrieval operations. It uses additional terms with the original queries. A few strategies are utilized for getting terms for query expansion, including thesaurus based techniques, relevance feedback-based techniques and co-occurrence-based techniques [12]. QE is utilized as a part of different applications, for example, multimedia data (Audio, video) retrieval, medicinal, health and social. Inquiry development has potential in complex event recognition [13].

Query expansion can extensively be characterized into three classifications. To begin with classification misuses gathering based or worldwide investigation, which utilizes setting worldwide of terms in an accumulation to discover comparative terms with query terms [14-15]. Second class incorporates query based or nearby investigation in which the setting of terms is diminished to littler subsets of data, which is given from relevance feedback or pseudo-relevance feedback and collaboration information like client profile and passed inquiries. Last classification is Knowledge-based approach, which comprises of investigation of the learning in outside information sources [16].

Queries submitted to a Web search engine are often ambiguous. Query expansion approaches which aim to overcome the ambiguity of natural language and furthermore the trouble in utilizing a single term. It is for the most part conducted by supplementing original queried terms by morphological variations or semantically related terms [17]. Query expansion requires a term determination stage where the system presents the query expansion terms to the users in a reasonable order. The order should preferably be one in which the terms that are most likely to be useful are close to the top of the list [18]. Tuning IR systems to optimize these assessment measures may create unsuitable outcomes when redundancy and ambiguity are considered [19]. In information retrieval community, query expansion involves evaluating a user's input and expanding the search query to match extra documents [20].

2. RELATED WORK

Francesco Colace *et al.* proposed a novel query expansion method to enhance accuracy of text retrieval systems. It makes use of a minimal relevance feedback to expand the initial query with a structured representation composed of weighted pairs of words. Such a structure was gained from the relevance feedback through a technique for pairs of words selection based on the Probabilistic Topic Model. This method when compared with other baseline query expansion plans and techniques proved very effective.

Jianqiang L *et al.* introduced a novel semantic-based approach to attain the diversity-aware retrieval of EMRs, i.e., both the relevance and novelty are considered for EMR ranking. Firstly, mine all the potential semantics from a client query and expand them to display the different query perspectives with the help of medical domain Ontology. After that a novel diversification strategy is used that considers not only the aspect importance but also the similarity. A real-world pilot study, which utilizes the proposed medical search approach to enhance the second use of the EMRs, is reported.

Arantxa Otegi *et al.* investigated the utilization of knowledge-based semantic related methods to locate the vocabulary mismatch between the query and reports. This approach is based on IR and Passage Retrieval methods used for questions and answers. The analysis shows that our models and PRF are complementary; i.e. PRF is better for easy queries, and our models are stronger for difficult queries. Our models generalize better to other collections, being more robust to parameter adjustments. In addition, we show that our method has a positive impact to end-to-end question answering systems for Basque. It can be also readily applied to other knowledge bases. Applying our search method on Wikipedia gave better results than using other knowledge structures (medical ontology and linked data repositories) on the same Wikipedia.

Luca Soldaini *et al.* investigated the utility of bridging the gap between a novice's and an expert's vocabularies, our approach adds the most appropriate expert expression to queries submitted by users, a task we call query clarification. We evaluated the impact of query clarification. Using three different synonym mappings and conducting two task-based retrieval studies, users were asked to answer medically-related questions using interleaved results from a major search engine. Our results show that the proposed system was preferred by users and helped them answer medical concerns correctly. The correct percentage got increased by 7 % when compared to the use of query that was used without any query clarification. Finally, we introduced a supervised classifier to select the most appropriate synonym mapping for each query, which further increased the fraction of correctness to 12 %.

Jagendra Singh *et al.* presented a new method for QE, based on fuzzy logic. This method considered the top-retrieved document as the relevant feedback document for mining additional QE terms. Different QE term selection method, calculates the importance of all unique terms in the top-retrieved documents collected for mining additional expansion terms. These methods give different relevance scores for each term. The proposed method combines different weights of each term by using fuzzy rules to infer the weights of the additional query terms. Then, the weights of the additional query terms and the weights of the original query terms are used to form the new query vector, and we use this new query vector to retrieve documents. All the experiments are performed on TREC and FIRE benchmark datasets. The proposed QE method increases the precision rates and the recall rates of information retrieval systems for dealing with document retrieval. It gets a significant higher average recall rate, average precision rate and F measure on both datasets.

3. PROPOSED ENHANCED QUERY EXPANSION BASED CLASSIFIER (EQC) APPROACH

When retrieving the web documents, the original query submitted by the user is not sufficient to retrieve the relevant documents. This may be due to the lack of user knowledge. The dataset contains collection of documents and they are responding to user queries to provide the relevant documents. The queries are needed to be translated for enhancing the searching efficiency which is a major issue with document retrieval. For efficient document retrieval, a novel EQC approach is proposed.

In a Vector Space Model, it contains P number of documents and L number of terms. The term and a collection of document is represented as $t_i\,(1 \le i \le L), h_j\,(1 \le j \le P)$. For L dimensional vector, the document is represented as,

$$e_j = (x_{1j}, x_{2j}, ..., x_{Lj})^S \tag{1}$$

Where, x_{ij} represent weight of the term t_i in document h_j and S represent transpose of all weighing terms. The query for the document is denoted as

$$r_k = (x_{1k}, x_{2k}, ..., x_{Lk})S \tag{2}$$

Where, N is the length of the query, $1 \le k \le N$ and x_{ik} is the i^{th} term's weight of query q_k.

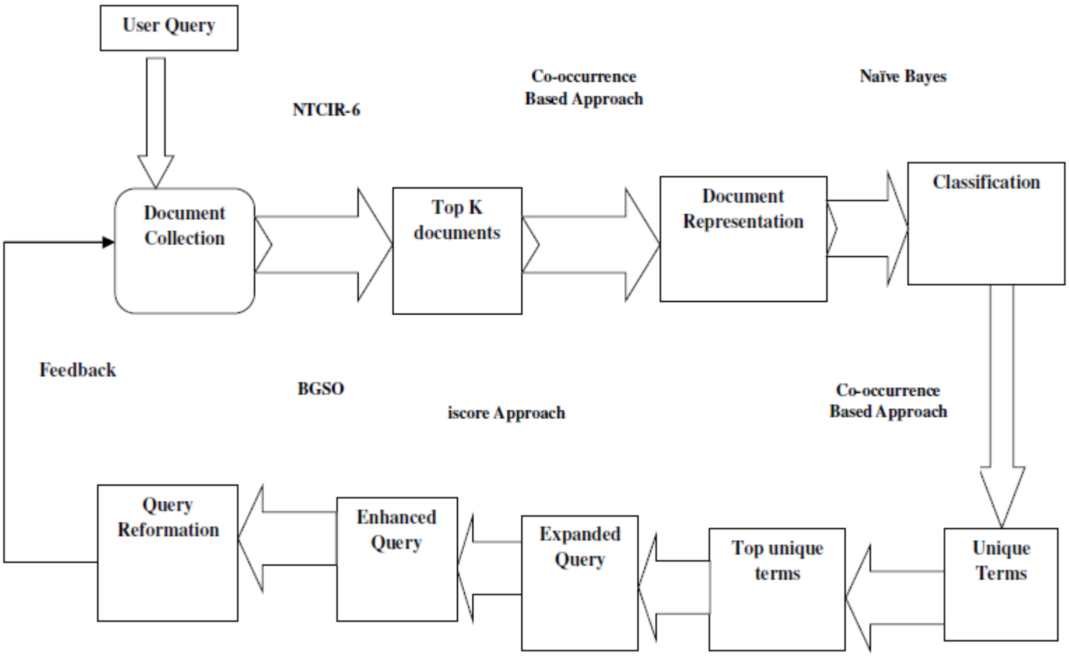

Figure 1 The diagram of Enhanced Query Expansion based Classifier (EQC) model

The overall block diagram of proposed method is shown in Fig. 1. The dataset contains collection of documents and they are retrieved with NTCIR-6. Based on the rank, top k relevant documents are retrieved for a given user query. Based on the co- occurrence features, the document is represented and classified with Naïve Bayes classifier algorithm. All unique terms are extracted from the classified document set and weighted to get top m unique terms. These terms are added with the original query and re weighted using iSCORE approach. Number of possible expansion terms is produced and they are optimized with BGSO. Finally the best terms are selected for user query to retrieve relevant documents.

4. CONCLUSION

In this paper, EQC approach was proposed for retrieving the web information through expanded query. The features from relevant feedback documents were used for expanding the query. NTCIR-6 was used in the initial phase of retrieval to get the relevant set of documents. The testing set documents were classified with Naïve Bayes classifier. The unique terms were extracted from the set of relevant documents and the terms were weighted using Rocchio algorithm. The extracted terms where combined to create a number of possibilities for query expansion. The optimal terms were selected with Binary Group Search Optimizer algorithm and they were added with the original query. Again the final retrieved documents were weighted with NTCIR-6 measures. The precision, recall, F measure, average precision score were calculated to show the enhanced performance of proposed EQC.

REFERENCES

[1] Leturia, Igor, Antton Gurrutxaga, Nerea Areta, Iñaki Alegria, and Aitzol Ezeiza, “Morphological query expansion and language-filtering words for improving Basque web retrieval”, Springer, Language resources and evaluation, volume. 47, no. 2, pp. 425-448, 2013.

[2] Durao, Frederico, Karunakar Bayyapu, Guandong Xu, Peter Dolog, and Ricardo Lage, “Expanding user’s query with tag-neighbors for effective medical information retrieval”, Springer, Multimedia tools and applications, volume. 71, no. 2, pp. 905-929, 2014.

[3] Tao, Xiaohui, Yuefeng Li, and Ning Zhong, “A personalized ontology model for web information gathering”, IEEE transactions on knowledge and data engineering, volume. 23, no. 4, pp. 496-511, 2011.

[4] Snášel, Václav, Ajith Abraham, Suhail Owais, Jan Platoš, and Pavel Krömer, “Optimizing information retrieval using evolutionary algorithms and fuzzy inference system”, Springer, In Foundations of Computational Intelligence, Volume 4, pp. 299-324, 2009.

[5] Liu, Bing, “Information retrieval and Web search.” Springer, In Web Data Mining, pp. 211-268, 2011.

[6] Lee, Ming-Che, Kun Hua Tsai, and Tzone I. Wang, “A practical ontology query expansion algorithm for semantic-aware learning objects retrieval”, Elsevier, Computers & Education, volume. 50, no. 4, pp.1240-1257, 2008.

[7] Tamine-Lechani, Lynda, Mohand Boughanem, and Mariam Daoud., “Evaluation of contextual information retrieval effectiveness: overview of issues and research”, Springer, Knowledge and Information Systems, volume. 24, no. 1, pp. 1-34, 2010.

[8] Ghorab, M. Rami, Dong Zhou, Alexander O’Connor, and Vincent Wade, “Personalised information retrieval: survey and classification”, Springer, User Modeling and User-Adapted Interaction, volume. 23, no. 4, pp. 381-443, 2013.

[9] Zhou, Dong, Séamus Lawless, and Vincent Wade, “Improving search via personalized query expansion using social media”, Springer, Information retrieval, volume. 15, no. 3-4, pp. 218-242, 2012.

[10] Malizia, Alessio, Kai A. Olsen, Tommaso Turchi, and Pierluigi Crescenzi, “An ant-colony based approach for real-time implicit collaborative information seeking”, Elsevier, Information Processing & Management, volume. 53, no. 3, pp. 608-623, 2017.

[11] Jalali, Vahid, and Mohammad Reza Matash Borujerdi, “Information retrieval with concept-based pseudo-relevance feedback in MEDLINE”, Springer, Knowledge and information systems, volume. 29, no. 1, pp. 237-248, 2011.

[12] Gao, Ge, Yu-Shen Liu, Meng Wang, Ming Gu, and Jun-Hai Yong, “A query expansion method for retrieving online BIM resources based on Industry Foundation Classes”, Elsevier, Automation in Construction, volume. 56, pp. 14-25, 2015.

[13] Kuo, Yin-Hsi, Kuan-Ting Chen, Chien-Hsing Chiang, and Winston H. Hsu, “Query expansion for hash-based image object retrieval”, ACM, pp. 65-74, 2009.

[14] Melucci, Massimo, “A basis for information retrieval in context”, ACM, Transactions on Information Systems (TOIS), volume. 26, no. 3, pp.14, 2008.

[15] Song, Wei, and Soon Cheol Park, “Latent semantic analysis for vector space expansion and fuzzy logic-based genetic clustering”, Springer, Knowledge and Information Systems, volume. 22, no. 3, pp. 347-369, 2010.

[16] Leong, Chee Wee, Samer Hassan, Miguel Enrique Ruiz, and Rada Mihalcea, “Improving query expansion for image retrieval via saliency and picturability”, Springer, pp. 137-142, 2011.

[17] de Boer, Maaike, Klamer Schutte, and Wessel Kraaij, “Knowledge based query expansion in complex multimedia event detection”, Multimedia Tools and Applications, volume. 75, no. 15, pp. 9025-9043, 2016.

[18] Durao, Frederico, Karunakar Bayyapu, Guandong Xu, Peter Dolog, and Ricardo Lage, “Expanding user’s query with tag-neighbors for effective medical information retrieval”, Springer, Multimedia tools and applications, volume. 71, no. 2, pp. 905-929, 2014.

[19] Lee, Ming-Che, Kun Hua Tsai, and Tzone I. Wang, “A practical ontology query expansion algorithm for semantic-aware learning objects retrieval”, Elsevier, Computers & Education, volume. 50, no. 4, pp. 1240-1257, 2008.

[20] Lu, Zhiyong, Won Kim, and W. John Wilbur, “Evaluation of query expansion using MeSH in PubMed”, Springer, Information retrieval, volume. 12, no. 1, pp. 69-80, 2009.

[21] Colace, Francesco, Massimo De Santo, Luca Greco, and Paolo Napoletano, “Weighted word pairs for query expansion”, Elsevier, Information Processing & Management, volume. 51, no. 1, pp.179-193, 2015.

[22] Li, Jianqiang, Chunchen Liu, Bo Liu, Rui Mao, Yongcai Wang, Shi Chen, Ji-Jiang Yang, Hui Pan, and Qing Wang, “Diversity-aware retrieval of medical records”, Elsevier, Computers in Industry, volume. 69, pp. 81-91, 2015.

[23] Otegi, Arantxa, Xabier Arregi, Olatz Ansa, and Eneko Agirre, “Using knowledge-based relatedness for information retrieval”, Springer, Knowledge and Information Systems, volume. 44, no. 3, pp. 689-718, 2015.

[24] Soldaini, Luca, Andrew Yates, Elad Yom-Tov, Ophir Frieder, and Nazli Goharian, “Enhancing web search in the medical domain via query clarification”, Springer, Information Retrieval Journal, volume. 19, no. 1-2, pp. 149-173, 2016.

[25] Singh, Jagendra, and Aditi, “A new fuzzy logic-based query expansion model for efficient information retrieval using relevance feedback approach”, Springer, Neural Computing and Applications, pp. 1-24, 2016.

[26] Rahul Shankarrao Khokale and Mohammad Atique, Web Information Retrieval Using Automatic Multi-Document Summarization, Volume 5, Issue 3, March (2014), pp. 107-114, International Journal of Computer Engineering and Technology.

[27] Anbazhagu, U. V., Deepalakshmi, P. and Praveen, J. S. Defeating SQL Injection Using Query String Attack Prevention Technique. International Journal of Computer Engineering and Technology, 6(10), 2015, pp. 42-41.

[28] M. Asokan and Dr. P. Arul. Load Testing for J-query Based Mobile Websites Using Borland Silk Performer™. International Journal of Computer Engineering and Technology, 6(9), 2015, pp. 12-20.

[29] Jagendra Singh, and Aditi Sharan, "A novel model of selecting high quality pseudo-relevance feedback documents using classification approach for query expansion", In Computational Intelligence: Theories, Applications and Future Directions (WCI), IEEE Workshop, ISBN: 978-1-4673-8215-1, pp. 1-6, 2015.