International Journal of Pure and Applied Mathematics
Volume 120 No. 6 2018, 3617-3630
ISSN: 1314-3395 (on-line version)
url: http://www.acadpubl.eu/hub/
Special Issue

http://www.acadpubl.eu/hub/

OPTIMIZED QUERY EXPANSION BASED CLASSIFIER FOR WEB INFORMATION RETRIEVAL

Dr. Anandam Velagandula[1], Priyadarshini Chatterjee[1],
Ch. Mamatha[1], K.Rajesh[2]
[1]Department of Computer Science and Engineering
and Department of Information Technology,
Vardhaman College of Engineering, Shamshabad
jinipriya@gmail.com
[2]Department of Computer Science and Engineering,
MLR Institute of Technology

June 13, 2018

Abstract

Due to the increase in availability of web documents, retrieving 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 increase the probing performance of relevant documents, original user query is needed to be reformulated. A unique optimized Query Extended Classifier (OQC) 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 Okapi-BM25 algorithm. From that, the topmost k relevant documents are selected to represent the feature terms. Then the documents are classified using k-Nearest Neighbor (k-NN) classifier which provides good feedback documents. Then the unique terms are extracted and they

are ranked using simultaneous occurrence based approach. Another algorithm implemented to reweight the unique terms and to expand the query. By using Universal Group Search Optimizer (UGSO) 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 can be counted in terms of Recall, Precision, Mean Average Precision (MAP),F-measure. These values are increased as 1%, 3% and 3.5% respectively. The performance measures show the improvement in relevant document retrieval scheme with reduced computational complexity.

***Keywords*:**Retrieval Feedback, Query Expansion, optimization, pseudo relevant documents, KNN classifier, Co-occurrence approach.

1 Introduction

The web has turned out to be fundamental means for many individuals regular activities and because of its broad searching engines, it has turned into an essential way for retrieving, 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 Fetching is a division that manages maintenance, storage and finding 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. Users information needs are satisfied by the search engine [9-10].

There is a difficulty on keyword based search; the client utilizes distinctive nouns in the inquiry than the caption utilized as tokens. Next way is; clients regularly give a small, ambiguous and 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 words for extended query, as thesaurus based techniques, importance feedback-based techniques and concurrency-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 words with other, words [14-15]. Another class incorporates questions and nearby investigation that contains setting of words is diminished to smaller sets of data that is from importance feedback or false-importance feedback and collaboration data like client information and completed inquiries. Final classification is learning-based approach that contains investigation of the learning in outside information sources [16].

Queries submitted to a Web search engine are often ambiguous. Extended queries can remove the uncertainty of simple language. Most of the part is derived after semantically and morphologically varying the original terms [17]. The terms that considered are generally at the top of the list [18]. The information fetching system can be tuned to maximize the suitable outcomes after considering redundancy and uncertainty

[19, 20].

2 Related work

Francesco Colace et al. proposed a unique query extension technique to enhance correctness of text fetching systems. It uses least importance feedback to extend the original query with a formatted structure that consists some pairs of words. This kind of structure was gained from the importance feedback using a technique that uses pairs of words . This method when compared with other base query extension plans and techniques proved very effective. Jianqiang L et al. introduced a unique syntactic-based approach to attain the equality of EMRs, i.e., both the importance and uniqueness are considered for EMR ranking. Firstly, 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 unique bifurcation strategy is used which will consider not only the matter of relevance but also the equableness.

Arantxa Otegi et al. investigated the utilization of learning-related syntactic related methods to identify the glossary mismatch between the reports and query. This approach is based on data retrieval and Path Retrieval techniques related to questions and answers. The study shows that our system and PRF are parallel; i.e. PRF gives better results for easy queries, and our proposed system is better for complex queries. The proposed system gives better results related to collective queries. It is robust to attribute adjustment .We can also show that our technique has a better results on question answering system. It can be also readily applied to other knowledge bases.

Luca Soldaini et al. investigated the benefit of joining the gap between an experts vocabularies and a novices, our proposal gives the most suitable accomplished expression to queries introduced by the person, a task we by which we can initiate query clarification where we can estimate query clarification impact. By making use 3 different synonym plotting and observing two task based retrieval survey, people were interrogate to answer medically-pictured questions using inter leaved conclusion from a

major search engine. Our results represent proposed system which has been referred by users and helps them in answering medical related 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 propose an idea of a supervised classifier to prefer the most relevant synonym plotting for each query, which in future can increase the fraction of correctness to 12%.

Jagendra Singh et al. proposed a new approach for QE, substantiated on fuzzy logic. This approach contemplates the top-retrieved document as significant feedback report for mining supplementary QE provision. Peculiar QE term alternate method, calculates the importance of entire exclusive provisions in the top retrieved reports gathered for mining supplementary extension provisions. These approaches provide disparate important scores for each and every term. The method which is proposed combines disparate weights of each provision based on fuzzy rules to interpret the weights of increased query provisions. Then, we make use of all the weights of both additional query the initial query provisions form an advanced query vector which can be further used retrieve relevant reports. All the research can be implemented on TREC and FIRE benchmark data sets. The approach which is a proposal of QE scheme raises the precision degree which reminiscence degree of information rescue systems for negotiating document and report retrieval. This gets a proficient greater average recollection rate and an average precision degree and F measure which conclude on data sets.

3 Proposed Optimized Query Expansion based Classifier (OQC) Approach

When retrieving the web documents, the original query submitted by the user is not sufficient to fetch the significiant 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 OQC approach is proposed.

In a Vector Space Model, it contains N number or range of documents and M number or range of terms. The term and a collection of document is represented as
$t_i(1 \leq i \leq M), d_j(1 \leq j \leq N)$.
For M dimensional vector, the document is represented as,

$$d_j = (w_{1j}, w_{2j}, ..., w_{Mj})\tau \tag{1}$$

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

$$q_k = (w_{1k}, w_{2k}, ..., w_{Mk})\tau \tag{2}$$

Where, L is the length of the query, $1 \leq k \leq L$ and w_{ik} is the i th term's weight of query q_k .

The overall block diagram of our proposal approach is shown in Fig. 1. The dataset contains collection of documents and they are retrieved with Okapi-BM25. 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 k-NN 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 Rocchio approach. Number of possible expansion terms is produced and they are optimized with UGSO. Finally the best terms are selected for user query to retrieve relevant documents.

Implementation: In this paper, the relevant documents are selected and similarity score is calculated using Okapi BM-25. Top 10 documents are selected and top 20 terms are extracted for classification. K-NN algorithm is used for classification. The terms from good feedback document are extracted and top 5 unique terms are selected using Rocchio weighing approach. An optimization algorithm UGSO is used to elect the prime expanded query from number of possible queries. The proposed technique is compared against Okapi-BM25 model, Classifier Based Query Expansion (CBQE) [26]. The comparison performance the

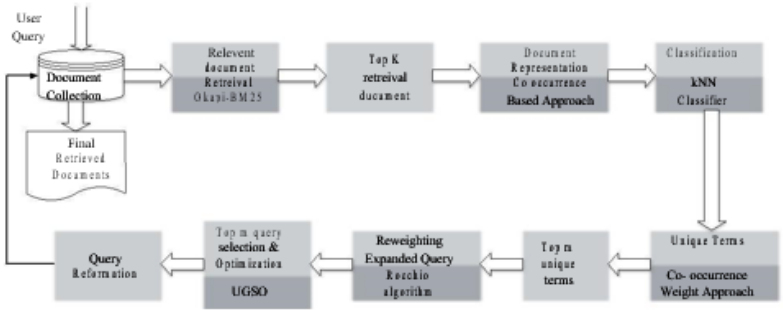

Figure 1: The diagram of Optimized Query Expansion based Classifier (OQC) model

Table 1: Comparison of Precision, Recall, F-measure and values for Okapi-BM25, CBQE, Proposed OQC

Techniques	Precision	Recall	F-measure
Okapi-BM25	0.2317	0.1172	0.1556
CBQE	0.2637	0.1427	0.1754
Proposed OQC	0.3319	0.4211	0.5158

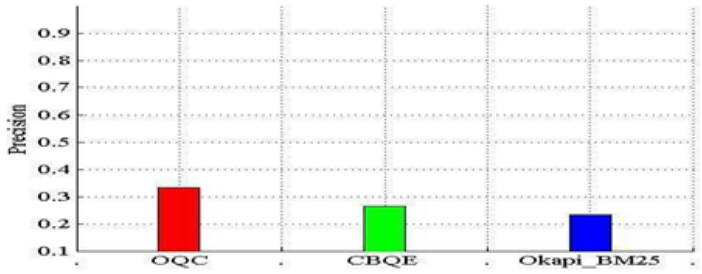

Figure 2: Comparison of precision values with current techniques

approaches intermits precision, recall and F-measure is shown in table II.

Fig 2 shows the precision value of Okapi-BM25, CBQE and Proposed OQC techniques. For the existing approaches, the values are in the range of 0.2317 and 0.2637. This value is increased up to 0.3319 using the proposed technique. The recall also increased to 0.4211 which is shown in Fig3. When compared with proposed approach, the conventional techniques give a recall value of 0.1172 and 0.1427. The increase in precision and recall value shows the improvement of performance in document retrieval.

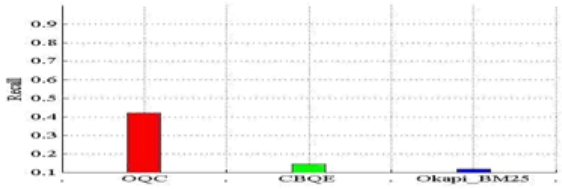

Figure 3: Recall comparison values with alive techniques

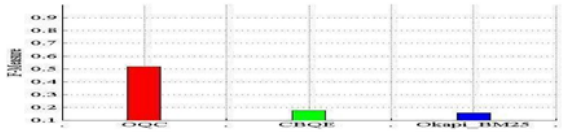

Figure 4: F-measure comparison with current techniques

Fig.4 shows the comparison result of F-measure that is above 0.5 for OQC. But in case of Okapi-BM25 and CBQE, the resultant values are below 0.2. In Fig. 5, the graph is plotted between precision and recall values in the range of 0.6 and 1. The precision range is between 0.3 and 0.35 for the recall value 0.6. Hence the retrieved number of document is nearer to the amount of significant

documents. It enhances the performance of document retrieval in web information retrieval.

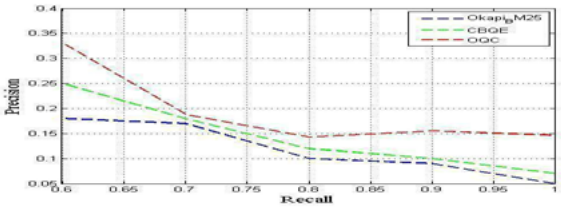

Figure 5: Precision vs. Recall

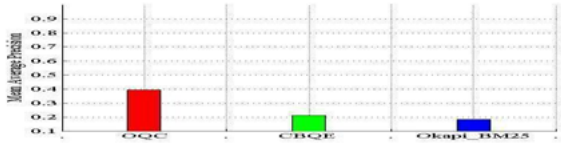

Figure 6: Comparison of Mean Average Precision

The precision value calculated is analyzed for number of queries. The average precision result is compared with the average precision of Okapi-BM25 and CBQE. The obtained evaluation 0.5158 is higher for proposed OQC. In modern search systems, it is common to return the ranked documents and it can be accomplished efficiently. Hence these results show the improvement of document retrieval system than others.Table 2 and figures 2 to 6 proves that the approach which proposed OQC application out performs the Okapi-BM25 and CBQE in all compact on TREC dataset.

4 Conclusion

In this paper, OQC approach was proposed for retrieving the web information through expanded query. The features from relevant

feedback documents were used for expanding the query. Okapi-BM25 was used in the initial phase of retrieval to get the relevant set of documents. The testing set documents were classified with k- NN classifier. The unique terms were extracted from the pair of relevant documents and 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 Universal Group Search Optimizer algorithm and they were added with the original query. Again the final retrieved documents were weighted with Okapi-BM25 measures. The precision, recall, F measure, average precision score were calculated to show the enhanced performance of proposed OQC.

References

[1] Leturia, Igor, Antton Gurrutxaga, Nerea Areta, Inaki 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] Snasel, Vaclav, Ajith Abraham, Suhail Owais, Jan Platos, and Pavel Kromer, "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, Samus 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.