A Comparative analysis of automatic query expansion approaches based on fuzzy logic for document retrieval

Jyotsana Bhardwaj[1], Mrs. Shalini[2]

[1]*Computer Engineering Department, Jaipur Institute of Technology, Jaipur, Raj, India*
[2]*Asst. Prof. of CSE, Jaipur Institute of Technology, Jaipur, Raj, India*

ABSTRACT

This paper presents a comparative analysis of recently developed query expansion approaches using fuzzy logic to retrieve relevant documents from large datasets for a given user query. Two query expansion approaches are compared and analyzed in different manner for two benchmark datasets: CISI and CACM. These approaches are based on fuzzy logic and term selection methods. On the basis performance evaluating parameters such as precision, recall, MAP and precision-recall graph, it is found that the approach proposed in [13] improves document retrieval in comparison to the approach proposed in [32].

Keywords: Fuzzy logic, query expansion, term weighting, term selection, precision, recall.

I. INTRODUCTION

Document retrieval system retrieves the most relevant documents from a large data corpus according to user's queries. Document retrieval consists of various components like searching, querying, document indexing and ranking. It is used in many areas such as information filtering, digital libraries, media search, recommender system, news retrieval, blog search, search engines (mobile search, enterprise search, web search and federal search) and many others.

The main issue with document retrieval effectiveness is "term mismatch problem". This problem states that it is not necessarily the same words are used by users and indexers who (who performs indexing) for searching the documents. This problem is also known as the "*vocabulary problem"* [1]. Synonymy and polysemy make this problem more complex. Synonymy means that the same words have different meanings such as "apple". This word has two meanings: an organization and a fruit. Polysemy means that different words have the same meaning such as "television" and "tv". Synonymy decreases recall by not retrieving all for any word "television" and "tv". Similarly, Polysemy decreases precision by retrieving more non- relevant documents in comparison to relevant documents.

One of the most successful and widely used techniques is query expansion to resolve term mismatch problem. Query expansion assists users in framing new queries from original queries. It has been observed in literature

that the average length of query is 2.30 words [2], which were already reported by Lau and Horvitz [3]. Thought, a slight increment in query length has been noticed for long queries (of five or more words) [3], but still many queries most of the queries consist of one, two or three words at most. Such type of situation makes vocabulary problem more critical. This shows that there is a huge scope for optimized query expansion approaches.

An initial query given by user is always incomplete and inadequate to represent user's need. Therefore, query expansion technique is used, which helps to select most suitable terms to be added with original query and in this way, the document retrieval performance can be enhanced [3]. To overcome above mentioned problems to a certain extent, researchers have proposed may query expansion approaches to formulate better queries [4-8]. Conceptually, the performance must be improved after applying query expansion approaches, but this is not the case always. Adding a new term to query creates a risk of query drift (the original query is changed topically) and diverts the searching into another direction. Therefore, there is need of extensive research to explore query expansion approaches with respect to their efficiency in improving retrieval effectiveness.

In last few decades, researchers analyzed various aspects of automatic query expansion (AQE) and have done research in several domains. The first work was reported by Van Rijsbergen in 1979 [4]. The proposed work was based on relevance feedback. Yang and Korfaghe [9] used real coded Genetic algorithm (GA) with random mutation and two-point crossover operators for improving the performance of query expansion. Sanchez et al. proposed GA based query expansion approach using user relevance feedback. GA was used to determine weights of all possible expanded terms for Boolean queries [10]. They tested their approach on patent dataset consisting 479 documents. Robertson and Willet [11] used evolutionary algorithm to identify the upper bound of relevance feedback for automatic query expansion technique in vector space model based document retrieval systems. They compared their results with Robertson et al.'s retrospective relevance weighting technique [12]. The results were satisfactory.

In recent years, Pseudo Relevance Feedback (PRF) based AQE is used widely and improved query expansion performance and retrieval processes. PRF is a type of local query expansion technique. However, there are a lot of limitations in PRF based QE in term of accuracy and computational complexity. To overcome these limitations, some other techniques were used with PRF i.e. semantic filtering such as WordNet etc [13]. However, it is also reported in literature that WordNet alone does not improve query expansion to large extent. Therefore, different variants were also introduced in recent years [13]. The use of concept and context of queries and documents is another way to enhance PRF based query expansion. Later on, some researchers also used soft computing technique to improve the performance of query expansion.

A new co-occurrence based query expansion techniques was proposed for improving document retrieval and tested on CACM, CISI datasets [14]. This approach successfully enhanced the performance of the system. Two different query expansion approaches using local collocation and global collocation were proposed [15]. These approaches were based on long span collocates. A new semantic similarity based query expansion approach using clusters was proposed to overcome the limitation of ambiguous and short queries [16]. This approach constructed various clusters of documents those are retrieved by the original query, and each cluster ranked on the basis of content similarity with the query. At last, this approach was suggesting terms from these ranked clusters to disambiguate the query.

A new query expansion technique using WordNet lexical chains was proposed by Gong et al. [17]. The proposed work was based on synonym and hypernym/hyponymy relations in WordNet. They used lexical chains as expansion rules. This approach improved query performance dramatically. Bendersky et al. [18] presented a new term reweighting method for query expansion to enhance the performance of document retrieval. They used GA to reweight a user's query vector in their approach. The proposed approach was based on the user's relevance feedback. Cooper et al. [19] proposed a GUI for users with graphical relations between different items by lexical neighborhoods for prompted query refinement. A novel term weighting based query expansion approach was proposed by Horng et al. [20]. They used GA to adapt the query term weights in order to get the closest query vector to the optimal one. Chen et al. [21] framed association rules to find out the degrees of similarity among terms and constructed a tree structure of these terms to select for query expansion. Chang et al. [22] proposed a new query expansion approach using fuzzy rules. The results were satisfactorily. Chang et al. [23] proposed a novel query expansion approach using weighting and re-weighting methods to enhance the performance of document retrieval**.** Chang et al. [24] framed fuzzy rules for user relevance based query expansion approach for document retrieval. Carlos et al. [25] proposed an approach to learn terms which actually helped to bridge the terminology gap existing between initial query and the relevant documents**.** Tayal et al. [26] used fuzzy logic to give weights to each query term using fuzzy triangular membership function. Gupta et al. [33] used fuzzy logic for constructing ranking function to enhance the performance of document retrieval process. Sharma et al. [34] presented the concepts of deep web and its analysis for web searching.

Rivas et al. [27] applied developed query expansion technique in biomedical document retrieval system. They combined text preprocessing with query expansion approach to improve the performance of document retrieval. They used one of the part of MEDLINE dataset, called Cystic Fibrosis for all the experiments. Li et al. [28] analyzed various query weighting approaches on L2R dataset for two transfer ranking frameworks: AdaRank and LambdaMART.

Singh et al. [29-30] combined various term selection based query expansion approaches to improve document retrieval performance. They also used Word2vec for selecting query expansion terms semantically. They obtained satisfactorily results. Singh et al. [31] proposed PRF and corpus-based term co-occurrence approach to find suitable terms for query expansion. They tested their approach on two datasets: FIRE and TREC-3. Singh et al. [32] presented a novel query expansion approach based on fuzzy logic. Authors obtained better recall and precision for the proposed AQE approach. Gupta et al. [13] proposed a novel automatic query expansion approach based on term weighting to extract relevant documents from datasets. The proposed approaches were compared with existing approaches and found improvement in document retrieval process.

This paper presents a comparative analysis of recently proposed fuzzy logic based query expansion approaches [13, 32]. Both the query expansion approaches are compared in terms of adopted approaches, number of membership functions, framed fuzzy rules and results. The rest of the paper is presented as follows. The theoretical background of both the query expansion approaches is discussed in section 2. In section 3, the results and analysis of performance of both the approaches are discussed. Finally, in section IV, the conclusion of the paper is presented.

II THEROTICAL FOUNDATION OF FUZZY LOGIC BASED QUERY EXPANSION APPROACHES

This section describes the theoretical background of fuzzy logic based query expansion in terms of approaches adopted in these approaches, membership functions and fuzzy rule base used in both the approaches.

2.1 Adopted approaches

In [32], a fuzzy logic based query expansion approach is proposed. This approach considers top-retrieved document as the most relevant documents to find suitable expanded terms. This approach includes various terms selection methods for query expansion approaches. These term selection methods determine the importance of all unique terms in terms of relevance score. All unique are selected from top-retrieved documents. The proposed method combines these weights of each term by using fuzzy logic to determine the weights of possible expanded query terms. Then, a new query vector is created by combining additional query term weights and original query term weights.

In [32], firstly, authors used Okapi-BM25 ranking function to retrieve the relevant documents from the dataset against original query. Then top retrieved documents were selected as PRF documents and all unique terms were identified from these documents to form a candidate term pool. Further, three types of term weighting methods such as class based, statistics based and co-occurrence based methods are used to give the weights to all terms of the term pool. Fuzzy logic is used to combine these methods at two levels: at first level, three different fuzzy logic controllers are developed and at second level, another fuzzy logic controller is developed. After that, a semantic filter is used to remove noisy and redundant terms; those are obtained for query expansion. Then after, all the terms are ranked in decreasing order of relevance score and top ranked terms are selected for query expansion.

In [13], authors proposed two approaches to extract relevant documents from large datasets. First approach was a new query expansion approach, which was based on term weighting scheme and second approach was a new combined semantic filter. In first approach, they used Particle Swarm Optimization (PSO) to determine the optimal weights of information retrieval evidences for all terms. Further, fuzzy logic was used to make PSO dynamic by controlling its parameters such as acceleration coefficients and inertia during the optimization process. In second approach, noisy terms were removed using proposed combined semantic filtering method. Then after, Rocchio method [16] was used to reweight the terms. The proposed approach improved the performance of document retrieval process effectively.

2.2 Membership functions used in approaches

In [32], triangular membership function is used for Fuzzification process. The ranges of membership for all variables are represented by three linguistic terms as high, medium and low. In [13], authors also used three membership functions for input and output variables such as High, Medium and Low. In both the approaches, Triangular type of membership is used to express the membership functions in this approach.

2.3 Fuzzy rule base

Fuzzy rules are framed in both the approaches on the basis of domain knowledge. In [32], total 21 fuzzy rules are framed and domain knowledge is tabulated in Table 1.

Table 1. Domain Knowledge for framing fuzzy rules in [32].

S.No.	Domain Knowledge Base
1	If ("Wstatistical is High") and ("Wclass is High") and ("Wco-occurrence is High") then "Wcombine is High".
2	If ("Wstatistical is Medium") and ("Wclass is Low") and ("Wco-occurrence is High") then "Wcombine is Medium".
3	If ("TFIDF is Low") and ("Wcombine is Low") then "Wfinal is Low".
4	If ("TFIDF is High") and ("Wcombine is Medium") then "Wfinal is Medium".

In [13], total 27 fuzzy rules are framed and the following domain knowledge is used to create these rules:

Table 2. Domain Knowledge for framing fuzzy rules in [13].

S.No.	Domain Knowledge Base	Examples
1	If "*Normalized gbest* is *low", "UN* is *low"* and "acceleration coefficients (c_1 and c_2) are also *low"* then "variation in c_1 (Δc_1)" and "variation in c_2 (Δc_2)*"* are likely to be "*medium"*.	"If *Normalized gbest* is *Low* and *UN* is *Low* and c_1, c_2 are *Low* then Δc_1 and Δc_2 are *Medium"*
2	If "*Normalized gbest* is *low"* and "*UN* is *low"* and "acceleration coefficients (c_1 and c_2) are *High"* then "variation in c_1 (Δc_1) and variation in c_2 (Δc_2)*"* are likely to be "*high"*.	"If *Normalized gbest* is *Low* and *UN* is *Low* and c_1, c_2 are *High* then Δc_1 and Δc_2 are *High"*
3	If "*Normalized gbest* is *low"* and "*UN* is *medium"* and "inertia (ω) is *medium"* then "variation in inertia ($\Delta\omega$)" is likely to be "*low"*.	"If *Normalized gbest* is *Low* and *UN* is *Medium* and ω is *Medium* then $\Delta\omega$ is *Low"*
4	If "*Normalized gbest* is *high"* and "*UN* is *low"* and "inertia (ω) is *low"* then "variation in inertia ($\Delta\omega$) is likely to be *high"*.	"If *Normalized gbest* is *High* and *UN* is *Low* and ω is *Low* then $\Delta\omega$ is *High"*

III EXPERIMENTAL RESULTS AND ANALYSIS

To analyze both the approaches, CACM and CISI datasets are used as benchmark datasets. Random fifty queries are selected from each dataset. The analysis is presented in two ways: analysis for overall effectiveness and query wise analysis.

3.1 Overall effectiveness

The overall effectiveness of both the approaches is compared in terms of *MAP, P@rank* and precision-recall graph. Table 3 tabulates the comparison of MAP values of both query expansion approaches. It is clear from this table that query expansion approach proposed in [13] gives better MAP values in comparison to the approach which is proposed in [32] for both datasets. Table 4-5 shows the results for both the approaches in terms of P@rank. These tables clearly show that query expansion approach proposed in [13] gives better results in comparison to the approach proposed in [32]. Precision-recall graphs are also plotted for CACM and CISI as shown in Figs. 1-2. These figures depict that approach proposed in [13] is better than the approach proposed in [32]. The approach proposed in [13] gets better precision values at all recall points in comparison to approach proposed in [32].

Table 3. Comparison of *MAP* for both query expansion approaches.

Dataset	Query expansion approach proposed in [32]	Query expansion approach proposed in [13]
CACM	0.2801	0.2842
CISI	0.2512	0.2553

Table 4. Comparison of *P@rank* for both query expansion approaches for *CACM* dataset.

	Query expansion approach proposed in [32]	Query expansion approach proposed in [13]
P@5	0.7847	0.8012
P@10	0.7351	0.7487
P@15	0.6895	0.7043
P@20	0.6519	0.6728
P@30	0.6051	0.6245
P@50	0.5104	0.5230

Table 5. Comparison of *P@rank* for both query expansion approaches for *CISI* dataset.

	Query expansion approach proposed in [32]	Query expansion approach proposed in [13]
P@5	0.6642	0.6756
P@10	0.6121	0.6308
P@15	0.5609	0.5771

P@20	0.5228	0.5398
P@30	0.4703	0.4814
P@50	0.3883	0.4087

3.2 Query wise analysis

To analyze the query wise performance of both the approaches, four queries are selected randomly and the values of recall and precision are computed. These results are tabulated in Table 6 and Table 7 for CACM and CISI respectively. It is clear from these tables that the approach proposed in [13] is gets more precision and recall values in comparison to the approach proposed in [32].

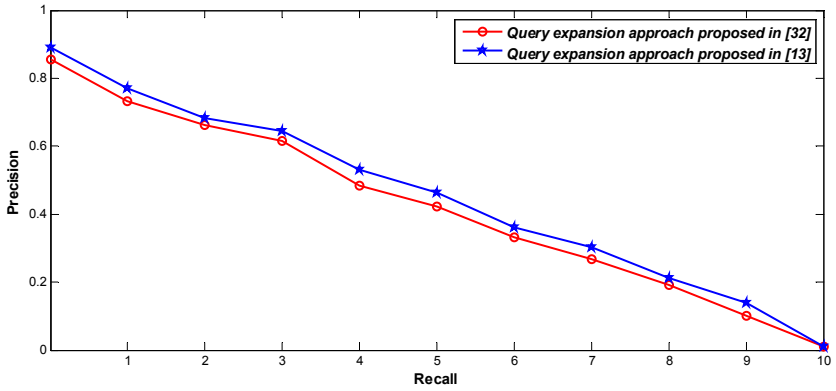

Fig. 1. Comparison of precision-recall graphs of approaches proposed in [13,32] for CACM.

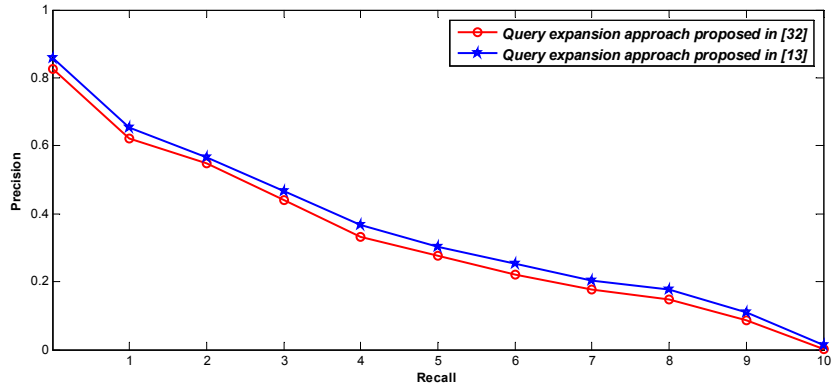

Fig. 2. Comparison of precision-recall graphs of approaches proposed in [13,32] for CISI.

Table 6. *Recall* and *Precision* values in case of *CACM*.

Query No.	Query expansion approach proposed in [32]		Query expansion approach proposed in [13]	
	Recall	Precision	Recall	Precision
14	0.7954	0.3065	0.7954	0.3105
26	0.7667	0.3591	0.8000	0.3676

36	0.5172	0.3418	0.5172	0.3569
63	0.5000	0.4504	0.6250	0.4520

Table 7. *Recall* and *Precision* values in case of *CISI*.

Query No.	Query expansion approach proposed in [32]		Query expansion approach proposed in [13]	
	Recall	Precision	Recall	Precision
2	0.3207	0.2178	0.3461	0.2225
12	0.6154	0.1949	0.6423	0.2091
28	0.4333	0.3394	0.4500	0.3472
34	0.5789	0.2918	0.6052	0.3078

IV CONCLUSION

A detailed comparison of recently developed fuzzy logic based query expansion approaches is presented in this paper. The performance of these query expansion approaches is compared in two ways: overall effectiveness of the approaches and query wise analysis of approaches. These approaches are also compared in terms of methodologies followed to develop in both approaches, membership functions and fuzzy rules framed to compute relevance score. CACM and CISI are used as benchmark dataset for performing all experiments. The comparison clearly shows that the approach proposed in [13] is better than the approach proposed in [32]. The query expansion approach proposed in [13] is superior in dealing with uncertainty, vagueness and impreciseness of queries and documents written in natural language in comparison to the approach proposed in [32]. As fuzzy logic is used in approach proposed in [32] for deciding the weights of various term selection methods whereas in [13], fuzzy logic and PSO are used to decide the weights of evidences.

REFERENCES

Journal Papers:

[1] Furnas, G., Landauer, T., Gomez, L., Dumais, S.: The vocabulary problem in human-system communication. ACM. (1997) 30, 11, 964–971.

[2] www.hitwise.com/us/press-center/press-releases/2009/google-searches-oct-09/

[3] Lovins, J.: Development of a stemming algorithm. Mechanical Translation and Computational Linguistics. (1968) 11 (1-2), 22-31.

[4] Rijsbergen, C.: Information Retrieval, second edition, Butterworth, USA. 1979.

[5] Sakai, T., Robertson, S.: Flexible pseudo relevance feedback using optimization tables. Louisiana. (2001) 396-397.

[6] Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing and Management. (1988) 24(5), 513–523.

[7] 7. Witten, I., Moffat, A., Bell, T.: Managing Gigabytes: Compressing and Indexing Documents and Images. Morgan Kaufmann. (1999).

[8] Molto, M., Svenonious, E.: Automatic Recognition of title page names. Information Processing and Management. (1991) 27 (1), 83-95.

[9] Yang, J., Korfhage, R.: Query modifications using genetic algorithms in vector space models. International Journal of Expert Systems. (1994) 7 (2) 165-191.

[10] Sanchez, E., Miyano, H., Brachet, J.: Optimization of fuzzy queries with genetic algorithms. In proceedings of Applications to a data base of patents in biomedical engineering, VI IFSA Congress, Sao-Paulo, Brazil, (1995) 293-296.

[11] Robertson A., Willet, P.: An upperbound to the performance for ranked-output searching: optimal weighting of query terms using a genetic algorithm. Journal of Documentation. (1996) 52 (4) 405-420.

[12] Robertson, S., Jones, S.: Relevance weighting of search terms. Journal of the American Society for Information Science. (1976) 27, 129-145.

[13] Gupta, Y., Saini, A.: A novel Fuzzy-PSO term weighting automatic query expansion approach using semantic filtering. Knowledge Based System. (2017) 136, 97-120.

[14] Xu, croft.: Query Expansion using Local and Global Document Analysis. ACM SIGIR conference on research and development in information retrieval. (1996).

[15] Olga V.: Query expansion with long-span collocates Information Retrieval. American Society for Information Science and Technology. (2009) 60 (2).

[16] Barathi, M., Valli, S.: Query Disambiguation Using Clustering and Concept Based Semantic Web Search for efficient Information Retrieval. Life Science Journal. (2013)10 (2).

[17] Gong, Z., Cheang, C., Hou, L.: Multi-term Web Query Expansion Using WordNet. Database and Expert Systems Applications. Lecture Notes in Computer Science. (2006) 4080, 379-388.

[18] Bendersky, M., Metzler, D., Bruce, W.: Effective Query Expansion with Multiple Information Sources. Fifth ACM International Conference on Web Search and Data Mining. ACM, USA. (2012).

[19] Cooper J., Byrd, R.: BIWAN—a visual interface for prompted query refinement. Proceedings of the 31st Hawaii international conference on system sciences, Hawaii. (1998) 2, 277–285.

[20] Horng, J., Yeh, C.: Applying genetic algorithms to query optimization in document retrieval. Information Processing and Management. (2000) 36, 737–759.

[21] Chen, H., Yu, J., Furuse, K., Ohbo, N.: Support IR query refinement by partial keyword set. Proceedings of the second international conference on web information systems engineering, Singapore. (2001) 11, 245–253.

[22] Chang, Y., Chen, S., Liau, C.: A new query expansion method based on fuzzy rules. Proceedings of the seventh joint conference on AI, Fuzzy system, and Grey system, Taipei, Taiwan, Republic of China. (2003).

[23] Chang, Y., Chen, C.: A New Query Reweighting Method for Document Retrieval Based on Genetic Algorithms. IEEE Transactions On Evolutionary Computation. (2006) 10 (5), 617-622.

[24] Chang, Y., Chen, S., Liau, C.: A new query expansion method for document retrieval based on the inference of fuzzy rules. Journal of Chinese Institute of Engineers. (2007) 30 (3), 511-515.

[25] Carlos, M., Maguitman, A.: A semi-supervised incremental algorithm to automatically formulate topical queries. Information Sciences. (2009) 179, 1881–1892.

[26] Tayal, D., Sabharwal, S., Jain, A., Mittal, K.: Intelligent query expansion for the queries including numerical terms. National Conference on Communication Technologies and its impact on Next Generation Computing. (2012)

[27] Rivas, A., Iglesias, E., Borrajo, L.: Study of Query Expansion Techniques and Their Application in the Biomedical Information Retrieval. The Scientific World Journal. (2014)

[28] Pengfei Li, Mark Sanderson, Mark Carman, Falk Scholer, On the Effectiveness of Query Weighting for Adapting Rank Learners to New Unlabelled Collections, CIKM'16 , October 24-28, 2016, Indianapolis, IN, USA.

[29] Jagendra Singh, Aditi Sharna, Relevance Feedback-based Query Expansion Model using Ranks Combining and Word2Vec Approach, Journal of IETE Journal of Research, Volume 62, 2016 - Issue 5, Pages 591-604, 2016.

[30] Jagendra Singh and Aditi Sharan, Relevance Feedback Based Query Expansion Model Using Borda Count and Semantic Similarity Approach, Computational Intelligence and NeuroscienceVolume 2015, Article ID 568197, 1-13 pages, 2015.

[31] Jagendra Singh; Aditi Sharan; Mayank Saini, Term co-occurrence and context window-based combined approach for query expansion with the semantic notion of terms, Int. J. of Web Science » 2017 Vol.3, No.1, vol.3, No.1, pp.32 – 57, 2017.

[32] Jagendra Singh; Aditi Sharan, A new fuzzy logic-based query expansion model for efficient information retrieval using relevance feedback approach, Journal Neural Computing and Applications archive Volume 28 Issue 9, September 2017 Pages 2557-2580

[33] Gupta, Y., Saini, A., Saxena, A.: A new fuzzy logic based ranking function for efficient Information Retrieval system. Expert System with Application. (2015) 42, 1223-1234**.**

[34] Sharma, D., Sharma, A.: Search Engine: A Backbone for Information Extraction in ICT Scenario. International Journal of Information Communication Technologies and Human Development. (2011) 3(2), 38-51.