# A Novel Term Selection based Automatic Query Expansion Approach using PRF and Semantic Filtering

**Yogesh Gupta, Ashish Saini**

***Abstract*: *Query expansion is the technique to make user's query precise by providing more relevant terms and term selection is one of the methods of it. This method removes irrelevant and redundant terms from the top ranked documents and enhances the efficiency of Information Retrieval System. There are several term selection methods and each method has its own strength and weakness. This paper introduces an approach which utilizes the strengths of each term selection method and overcomes the weaknesses. The proposed approach is based on pseudo relevance feedback and fuzzy logic. A novel semantic filter is also developed in this work to avoid query drifting problem. Three benchmark datasets CACM, CISI and TREC-3 are used to perform all the experiments and the results are compared with recent state of art in terms of MAP, precision-recall and F-measure. The results demonstrate the superiority of proposed approach over other compared approaches.***

***Index Terms*: *Automatic query expansion, fuzzy logic, pseudo relevance feedback, semantic filter, term selection method.***

## I. INTRODUCTION

Information Retrieval (IR) is a retrieval system, which retrieves relevant documents from a dataset for a user query. Generally, a user query is short and contains only one or two terms. These queries are ambiguous in nature as written in natural languages. Therefore, most of the documents are non-relevant in extracted documents. The term mismatch is the main cause of it. This problem occurs when users and indexers do not use the same words to represent the query. It is also known as vocabulary problem [1]. One of the most effective approach to deal with this problem is query expansion.

Query Expansion enhances the quality of user query and improves the performance of IR by adding few new relevant terms to user query. Few approaches related to query expansion are also reported in literature [2-4]. Term selection is one of them which determines a set of most suitable terms for QE.

A new automatic query expansion (AQE) approach is introduced in this paper. The introduced approach is based on various term selection methods, pseudo relevance feedback (PRF) [5] and fuzzy logic. In PRF based query expansion assumes that the top ranked documents, those are retrieved for the first time are relevant documents. Although PRF based AQE has been shown in improving IR performance [6-10] in literature, but it suffers from few problems. The main problem is that the consideration of all the unique terms of top ranked retrieved documents important for query expansion and some of them may not be relevant. There are several methods to select the suitable terms for PRF based query expansion such as Jaccard coefficient, Dice coefficient and others and no method is perfect. Therefore, a new term selection approach is proposed in this paper, which overcomes the weaknesses of individual term selection methods and utilizes their strengths. The proposed approach determines the most suitable terms after combining the weights of each unique term of top ranked documents using fuzzy rules. These unique term are the outcomes of each term selection method.

It is observed from literature survey that query term selection methods can be categorized into three categories as: co-occurrence information based [11], class based [12-13] and corpus based [14-15]. Each method determines the importance of expansion terms by calculating weighting scores and ranked them according to their individual score. It is not necessary that all important terms are presented in top ranked documents; these may be somewhere in the middle ranked documents. Therefore, many important terms, those are not in top-ranked terms, cannot be used by using individual term selection methods for AQE. Thus, it is natural to combine these term selection methods using fuzzy logic to improve performance. It is obvious that not all obtained query expansion terms are related to user query semantically and become essential to filter the noisy terms to avoid query drifting problem. Therefore, a semantic filtering approach [21] is used in this paper, which fulfils the required purpose. The proposed approach is compared with original query, Fuzzy and semantic (using WordNet ontology) based PRF approach given by Tomiye et al. [16] and Parapar et al. [17] AQE approach. Three benchmark datasets CACM, CISI and TREC-3 are taken to perform all the experiments.

This paper is categorized into following sections: section 2 gives the details of preliminaries and theoretical foundation of term selection methods. Section 3 discusses proposed term selection approach for AQE using semantic filtering. Section 4 presents the results and analysis. At the end, section 5 concludes the whole work.

**Revised Manuscript Received on December 22, 2018.**

**Yogesh Gupta**, Department of Computer Science and Engineering, Manipal University Jaipur, Jaipur, INDIA.

**Ashish Saini**, Department of Electrical Engineering, Dayalbagh Educational Institute, Agra, INDIA.

## II. THEORETICAL FOUNDATION AND PRELIMINARIES

This section discusses pseudo relevant feedback process and term selection using co-occurrence-based methods and corpus-based methods.

### A. *Document selection using PRF*

In pseudo relevant feedback process, the candidate terms are selected from initially retrieved top ranked documents to expand the user query. Suppose, top $n$ documents are retrieved from the corpus initially. This initial retrieval of documents is fully dependent on proper selection of similarity function. *Okapi-BM25* is used in this work for the same purpose.

After submitting query to an IR system, a set of relevant documents is obtained and then, top n documents are taken. These documents are used to identify unique terms to from a term pool. All the terms of this term pool can be ranked by anyone of the several term selection methods. These term selection methods are presented in following subsections.

### B. *Co-occurrence coefficient based Term selection method*

This method uses co-occurrence to find the relationship among query terms and candidate terms [11, 19]. The mathematical expressions of some co-occurrence coefficients related to such method are as follows:

$$Jaccard\_co(t_i,t_j)=\frac{d_{ij}}{d_i+d_j-d_{ij}} \quad (1)$$

$$Dice\_co(t_i,t_j)=\frac{2*d_{ij}}{d_i+d_j} \quad (2)$$

$$Cosine\_co(t_i,t_j)=\frac{d_{ij}}{\sqrt{d_i d_j}} \quad (3)$$

where $d_i$ and $d_j$ represent the number of documents containing $t_j$ and $t_j$ respectively, whereas $d_{ij}$ denotes the number of documents containing both $t_i$ and $t_j$ together.

### C. *Term selection method based on term distribution in corpus*

This section describes two methods as follows:

#### Kullback Leibler Divergence (KLD)

The central idea of this method is relying on the distribution of terms in pseudo relevant document and in the entire dataset. *KDL* method can be formulated as (4) -(6).

$$KLD(t)=\sum_t\left\{P_r(t)\times\frac{P_r(t)}{P_c(t)}\right\} \quad (4)$$

$$P_r(t)=\frac{\sum_{d\in r}tf(t/d)}{\sum_{d\in r}\sum_{t\in d}tf(t/d)} \quad (5)$$

$$P_c(t)=\frac{\sum_{d\in c}tf(t/d)}{\sum_{d\in c}\sum_{t\in d}tf(t/d)} \quad (6)$$

The where, $P_r(t)$ and $P_c(t)$ are probabilities of appearing of term $t$ in top retrieved documents and in the entire corpus respectively.

#### Robertson Selection Value (RSV)

This method is proposed by Robertson [19]. According to this method, if the weight of candidate term is *wt* then the class (relevant document/non-relevant document) containing this term will add wt to the values of their matching function. RSV score for candidate expansion terms can be defined as (7).

$$RSV\_score=\sum_{t\in d}w(t,d)(P_{tr}-P_{tnr}) \quad (7)$$

where, $\mu_R$ and $\mu_N$ presents the mean of classes of *relevant* and *non-relevant* documents respectively.

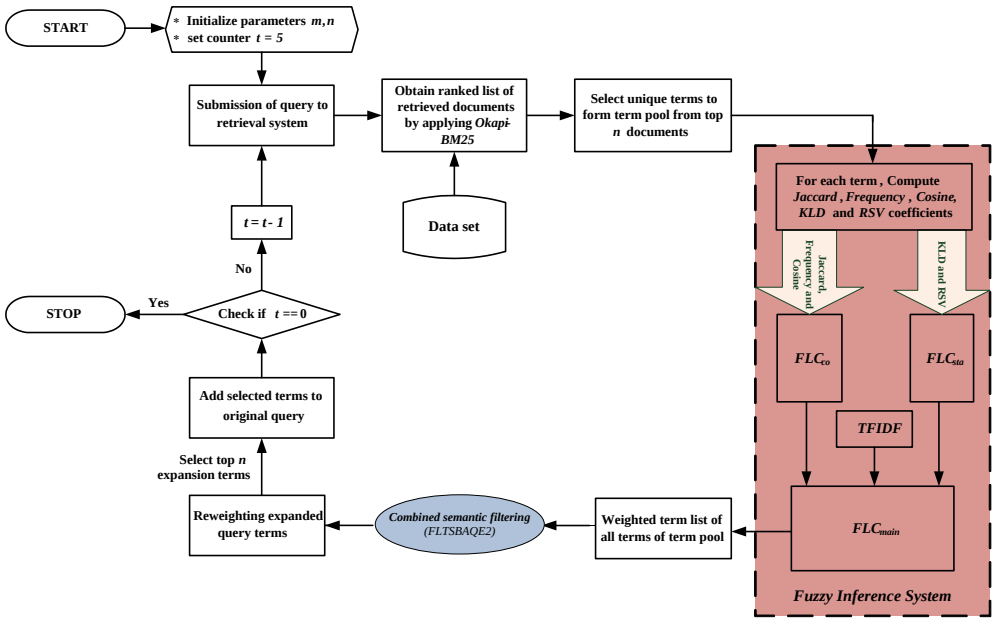


**Fig. 1:** An architecture of Proposed Term Selection based *AQE*

The above mentioned term selection methods for *PRF* based *AQE* have certain demerits as discussed above. A suitably designed fuzzy logic based term selection method for query expansion may be more effective in comparison to above mentioned co-occurrence based and statistics based term selection methods. Therefore, such type of term selection based *AQE* approaches are proposed in this work.

**Table I.** Details of entities and variable used in FIS.

| | Notation | Description |
|---|---|---|
| Variable | M | Number of top retrieved documents |
| | N | Number of top terms for query expansion |
| | T | Counter to check iteration |
| | Jaccard_Co | The Jaccard Coefficient Score value of a term |
| | Frequency_Co | The Frequency Coefficient Score value of a term |
| | Cosine_Co | The Cosine Coefficient Score value of a term |
| | TFIDF | TFIDF score value of a term |
| | KLD | Score value computed by KLD method of a term |
| | RSV | Score value computed by RSV method of a term |
| | Wco | Intermediate output representing term weight for FLCco |
| | Wsta | Intermediate output representing term weight for FLCsta |
| | wfinal | Output of FLCmain representing final term weight |

## III. PROPOSED TERM SELECTION BASED *AQE* APPROACHES

The proposed approach has *two* phases. In first phase, terms are combined obtained through various term selection methods at different levels using fuzzy logic. In second phase, a new *semantic filter* is used to remove noisy terms. *Okapi-BM25* ranking function is used to retrieve the relevant documents against a query. Fig. 1 represents the block diagram of proposed term selection based *AQE* approach and Table 1 describes the notations and description of different entities and variables used in Fig. 1. A collection of documents is retrieved from dataset after submitting query to IR system. Top ranked *n* numbers of documents are selected as *PRF* documents. All the unique terms selected from *PRF* documents are used to form a *candidate term pool*. Thereafter, fuzzy based approach which uses different term selection methods as discussed above to weight the terms of *candidate term pool*. These coefficients are used as input to Fuzzy Inference System (*FIS*), which determines final weight of each term of *candidate term pool*. The details of *FIS* used in this work are explained in *subsection A*.

A weighted list is formed containing all the terms and the final weights associated with these terms. A *semantic filter* [21] is also used to remove noisy terms from the list. The top *m* terms are selected to be added in original query for expansion. This procedure is repeated for *five* times to ensure best possible selection of suitable terms for query expansion.

### A. *Fuzzy Inference System used in proposed approach*

The *FIS* used in this work is composed of *three* fuzzy logic controllers (FLC) as shown in Fig. 1. $FLC_{co}$ deals with *three co-occurrence coefficients* and gives an output for each term i.e. term weight $w_{co}$ based on fuzzy rules. Similarly, $FLC_{sta}$ gives term weight $w_{sta}$ for each term as output after fuzzyfying *two statistical distribution term selection measures* (*KLD* and *RSV*). The weights $w_{co}$ and $w_{sta}$ along with *TFIDF* score of each term are used as inputs for $FLC_{main}$, which gives final weight $w_{final}$ for each term. Therefore, *FIS* gives a weight signifying cumulative effect of *co-occurrence coefficients*, *statistical distribution term selection measures* and *TFIDF* measure for better term selection. Then, *semantic filter* is used to filter less-suitable terms from obtained list. At the last, top *m* terms are selected to expand the query as per their *similarity*.

**Table II.** Domain knowledge used to frame the fuzzy rules.

| FLC | Domain Knowledge | Number of fuzzy rules framed | Total number of fuzzy rules |
|---|---|---|---|
| $FLC_{co}$ | If a term has low *Jaccard_Co* value, low *Frequency_Co* value and low *Cosine_Co* value then $w_{co}$ will be low.<br>If a term has high *Jaccard_Co* value, high *Frequency_Co* value and high *Cosine_Co* value then $w_{co}$ will be high. | 27 | 63 |
| $FLC_{sta}$ | If a term has low *KLD* score and low *RSV* score, then $w_{sta}$ will be low.<br>If a term has high *KLD* score and high *RSV* score then $w_{sta}$ will be high. | 9 | |
| $FLC_{main}$ | If a term has low value of $w_{co}$, low value of $w_{sta}$ and *TFIDF* is also low then the value of $w_{final}$ is likely to be low.<br>If a term has high value of $w_{co}$, high value of $w_{sta}$ and *TFIDF* is also high then the value of $w_{final}$ will be high. | 27 | |

**Table III.** Algorithm used for proposed approach

Step 1: Initialize the parameters (*n*, *m* and set counter *t = 5*).
Step 2: User submits a query to IR system.
Step 3: An ordered list of documents from dataset is extracted by using *Okapi-BM25* ranking functions.
Step 4: Select top *n* documents from ranked list.
Step 5: Term pool of all unique terms is created.
Step 6: Apply following term selection methods to determine suitability scores to all terms of term pool.
*Step 6.1*: Co-occurrence-based method
*Step 6.1.1*: Compute suitability score of terms using *Cosine* coefficient
*Step 6.1.2*: Compute suitability score of terms using *Jaccard* coefficient
*Step 6.1.3*: Compute suitability score of terms using *Frequency* coefficient
*Step 6.2*: Statistical distribution term selection method
*Step 6.2.1*: Compute suitability score of terms using *KLD* coefficient
*Step 6.2.2*: Compute suitability score of terms using *RSV* coefficient
Step 7: Combine the weights of all terms of term pool obtained from statistical and co-occurrence based methods of Step 6 using $FLC_{co}$ and $FLC_{sta}$ respectively at first level *FIS*.
Step 8: Combine $FLC_{co}$, $FLC_{sta}$ and *TFIDF* to determine final weights as suitability score of each term of term pool.
Step 9: Select top *m* terms from ranked list of terms.
Step 10: Add these *m* terms to original query and used to retrieve the documents from dataset.
Step 11: If counter *t ==0*, then stop the process. Otherwise perform *t = t-1* and move to Step 2.

A fuzzy rule base is formed for all the three fuzzy logic controllers of *FIS* to control output variables ($w_{co}$, $w_{final}$ and $w_{sta}$). The domain knowledge used in this paper to frame fuzzy rules is tabulated in Table 2. All the fuzzy rules carry equal weight. The *AND* operator is used to obtain a fuzzy set representing antecedent in canonical form of that particular fuzzy rule in this *FIS*. *Implication* is performed by applying fuzzy operator, which gives an output fuzzy set for consequent part of each fuzzy rule. In this work, *centroid method* [20] is used as defuzzification method. The algorithm used for proposed term selection based *AQE* is given in Table 3 as follows.

## IV. EXPERIMENTAL ANALYSIS AND DISCUSSION

The We have used *three* benchmark and widely used datasets TREC-3, CISI and CACM to perform all the experiments. We have selected total fifty queries randomly from these datasets to test the performance of proposed *AQE* approach and to compare with other similar type of approaches. We have analysed the results in following three ways:

1. Query specific analysis
2. Overall performance analysis
3. Statistical analysis

In this paper, we have compare the results of proposed *AQE* approach with original query; *Fuzzy and semantic* (using *WordNet* ontology) *based PRF approaches* [16, 17].

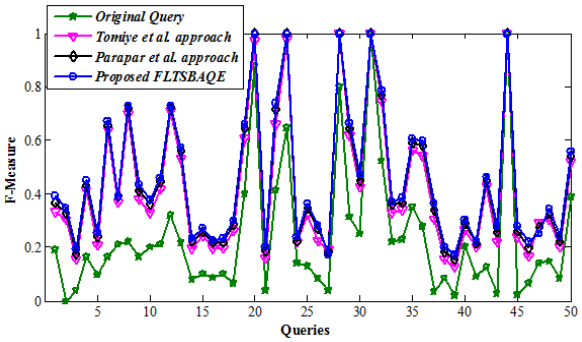


**Fig. 2:** Comparison of F-measure at top ten cut-off against fifty queries for CACM

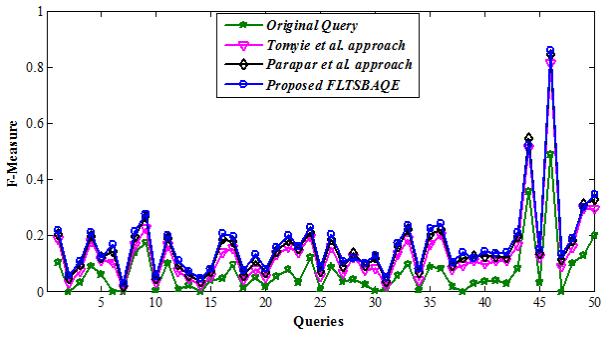


**Fig. 3:** Comparison of F-measure at top ten cut-off against fifty queries for CISI

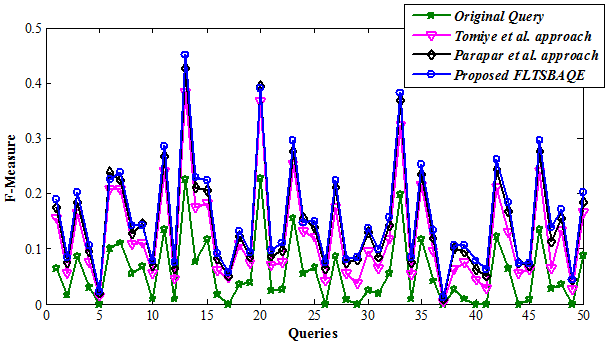


**Fig. 4:** Comparison of F-measure at top ten cut-off against fifty queries for TREC-3

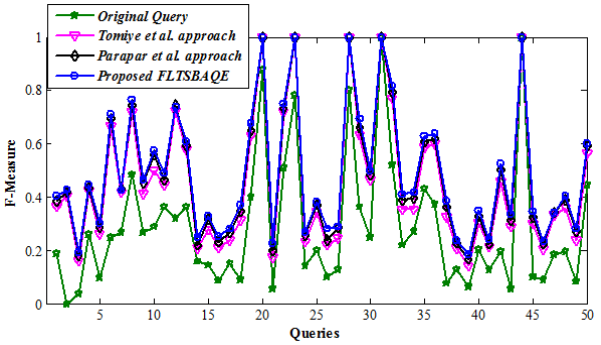


**Fig. 5:** Comparison of F-measure at top twenty cut-off against fifty queries for CACM

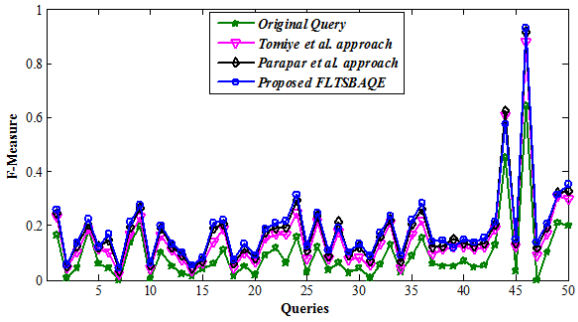


**Fig. 6:** Comparison of F-measure at top twenty cut-off against fifty queries for CISI

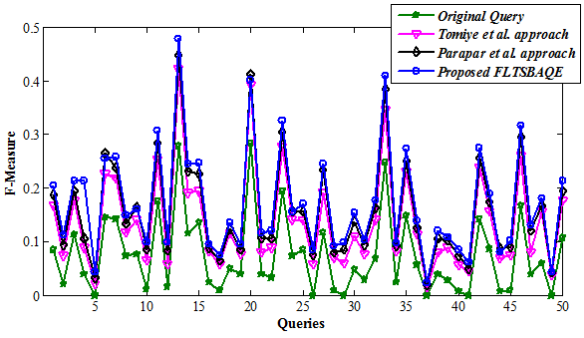


**Fig. 7:** Comparison of F-measure at top twenty cut-off against fifty queries for TREC-3

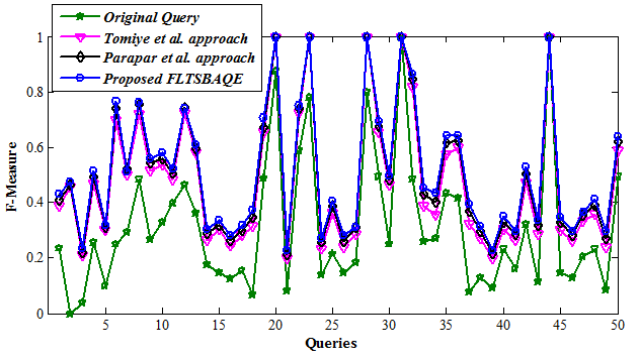


**Fig. 8:** Comparison of F-measure at top thirty cut-off against fifty queries for CACM

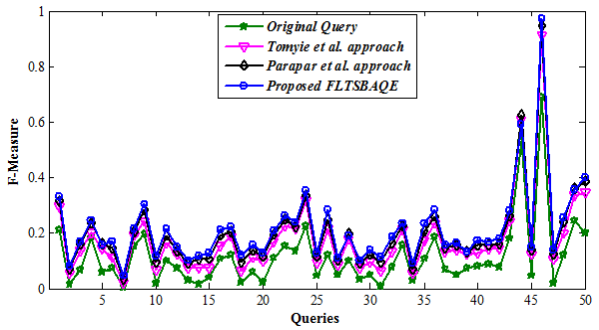


**Fig. 9:** Comparison of F-measure at top thirty cut-off against fifty queries for CISI

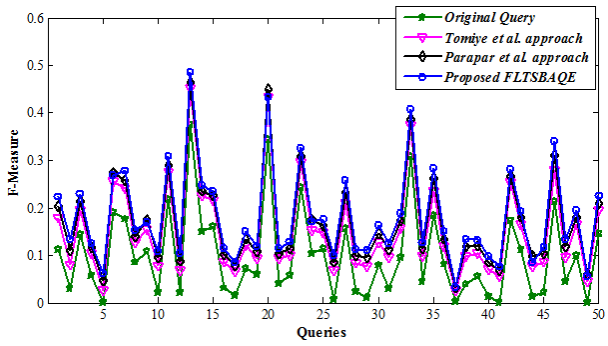


**Fig. 10:** Comparison of F-measure at top thirty cut-off against fifty queries for TREC-3

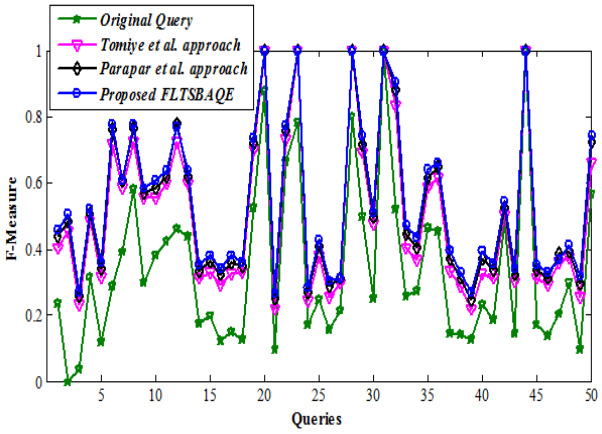


**Fig. 11:** Comparison of F-measure at top fifty cut-off against fifty queries for CACM

The proposed term selection based AQE approach is based on PRF. It is much required to select the best values of n and m. In this paper, the values for n and m are set empirically as 10 and 6 for CACM, CISI respectively and 30 and 10 for TREC-3 respectively [21]. In proposed approach, first top n documents are taken from initial run and then all unique terms are listed. Further, the weight is computed for each term using proposed approach. In this work, different weights are given

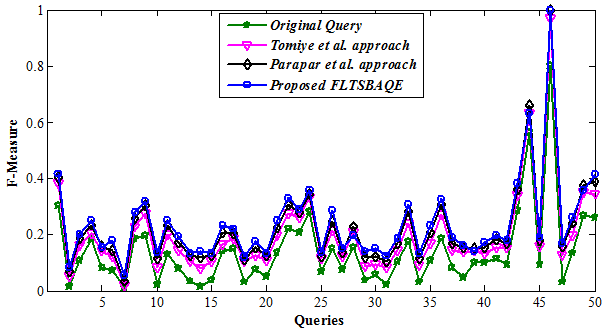


**Fig. 12:** Comparison of F-measure at top fifty cut-off against fifty queries for CISI

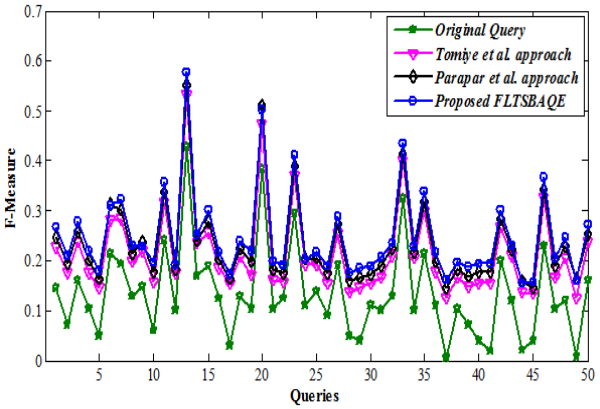


**Fig. 13:** Comparison of F-measure at top fifty cut-off against fifty queries for TREC-3

to original query terms and supposed to be added terms for expansion.

### A. *Query wise retrieval effectiveness*

In this paper, we have computed F-Measure to analyse the performance of each query of proposed approach and compared the results with other approaches. We have determined F-measure value at four cut-offs. These cut-offs are top *ten*, *twenty*, *thirty* and *fifty* retrieved documents for TREC-3, CISI and CACM datasets. We have obtained better F-measure values using proposed approach in comparison to other approaches for all the datasets as shown in Figs 2-13.

Fig. 2 shows the comparison of approaches for top *ten* cut-off for *CACM*. It is clear from this figure that proposed *FLTSBAQE* gets better *F-measure* for *forty* queries out of *fifty* queries in comparison to other query expansion approaches. Fig. 3 shows the comparison for *CISI* at top *ten* retrieved documents cut-off. This figure shows that better *F-measure* are obtained by proposed *FLTSBAQE* approach over Tomiye et al. approach, Parapar et al. approach and *original query* for *forty-four* queries. *F-measure* values of *FLTSBAQE* approach are equal to Parapar et al. approach for *one* query only. Fig. 4 demonstrates the results for *TREC-3* at top *ten* cut-off.

This figure depicts that proposed approach outperforms other query expansion approaches for *forty-three* queries.

Figs. 5-7 show the results for top *twenty* cut-off for all *three* datasets. These figures reveal that the proposed approach achieve better *F-measure* for *forty-one* queries in comparison with other query expansion approaches for *CACM*. Similarly, proposed approach gets better results as compared to others for CISI and TREC-3.

Likewise, Figs. 8-13 also illustrate that proposed *FLTSBAQE* approach outperforms other query expansion approaches for top *thirty* and top *fifty* retrieved documents in case of all *three* datasets. It is also observed that *FLTSBAQE* lags for only *three* queries, *five* queries and *six* queries from Parapar et al. approach at cut-off *fifty* retrieved documents for *CACM, CISI* and *TREC-3* respectively.

The improvement in query wise *precision* for *FLTSBAQE* with respect to Parapar et al. approach is also analysed. In Figs. 14-16, the length of each bar depicts the range of *precision* variations of *FLTSBAQE* over Parapar et al. approach for top *fifty* retrieved documents.

Fig. 14 shows that the performance of proposed *FLTSBAQE* degrades than Parapar et al. approach for *three* queries only in case of *CACM* dataset.

Fig. 15 show that proposed *FLTSBAQE* lags from Parapar et al. approach for *five* queries in case of *CISI* dataset but performance is increased for rest of the queries. Similarly, Fig. 16 shows that *FLTSBAQE* enhances the performance for *forty-four* queries.

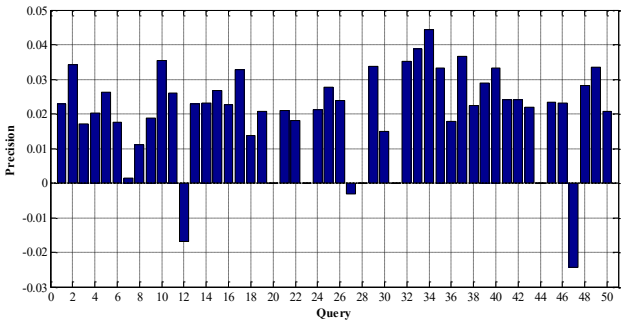


**Fig. 14:** Query-wise precision variation of *FLTSBAQE* against Parapar et al. approach for CACM

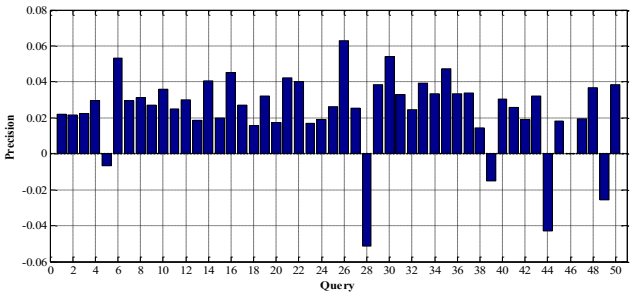


**Fig. 15:** Query-wise precision variation of *FLTSBAQE* against Parapar et al. approach for CISI

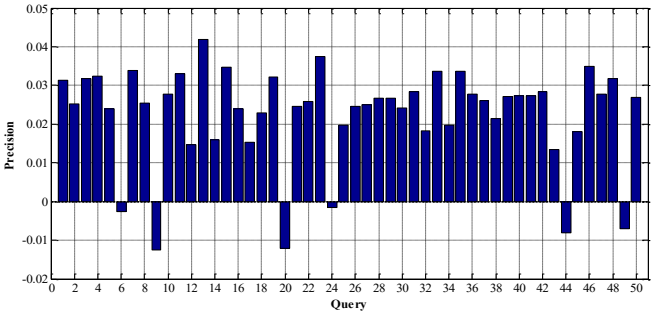


**Fig. 16:** Query-wise precision variation of *FLTSBAQE* against Parapar et al. approach for TREC-3

**Table IV:** Query wise comparison for CACM

| Q. No. | User Query | | Tomiye et al. approach | | Parapar et al. approach | | *FLTSBAQE* | |
|---|---|---|---|---|---|---|---|---|
| | ***R*** | ***P*** | ***R*** | ***P*** | ***R*** | ***P*** | ***R*** | ***P*** |
| **14** | 0.4545 | 0.1762 | 0.7727 | 0.2582 | 0.7954 | 0.2802 | **0.7954** | **0.3051** |
| **26** | 0.3667 | 0.2136 | 0.7000 | 0.3274 | 0.7667 | 0.3409 | **0.8000** | **0.3613** |
| **63** | 0.2500 | 0.2780 | 0.5000 | 0.3882 | 0.5000 | 0.4012 | **0.6250** | **0.4289** |

**Table V:** Query wise comparison for CISI

| Q. No. | User Query | | Tomiye et al. approach | | Parapar et al. approach | | *FLTSBAQE* | |
|---|---|---|---|---|---|---|---|---|
| | ***R*** | ***P*** | ***R*** | ***P*** | ***R*** | ***P*** | ***R*** | ***P*** |
| **2** | 0.0384 | 0.1000 | 0.2307 | 0.1573 | 0.2692 | 0.1792 | **0.3076** | **0.2112** |
| **12** | 0.3076 | 0.1201 | 0.6154 | 0.1506 | 0.6154 | 0.1824 | **0.6154** | **0.2043** |
| **34** | 0.3421 | 0.1583 | 0.5263 | 0.2163 | 0.5526 | 0.2547 | **0.5789** | **0.2982** |

**Table VI:** Query wise comparison for TREC-3

| Q. No. | User Query | | Tomiye et al. approach | | Parapar et al. approach | | *FLTSBAQE* | |
|---|---|---|---|---|---|---|---|---|
| | ***R*** | ***P*** | ***R*** | ***P*** | ***P*** | ***R*** | ***P*** | ***R*** |
| **159** | 0.3329 | 0.1936 | 0.4164 | 0.2609 | 0.4776 | 0.3029 | **0.4776** | **0.3108** |
| **177** | 0.1167 | 0.0794 | 0.3226 | 0.1607 | 0.3434 | 0.1878 | **0.3613** | **0.2282** |
| **198** | 0.2413 | 0.1427 | 0.4649 | 0.293 | 0.4811 | 0.3195 | **0.5036** | **0.3423** |

It is quite interesting to see the performance of few queries in terms of precision and recall. Therefore, three queries are randomly selected from all three datasets to perform experiments as shown Table 4-6. Precision and Recall are represented as P and R respectively in these tables. It is clear from these tables that proposed FLTSBAQE outperforms other query expansion approaches.

B. *Overall retrieval effectiveness*

We have computed *MAP* and *P@rank* to test the overall performance of proposed automatic query expansion approach. The comparative results are tabulated in Table 7-10. Table 7 presents the comparison of *MAP*. It is evident from this table that proposed approach gets better *MAP* than *original query*, Tomiye et al. approach and Parapar et al. approach for all *three* datasets.

**Table VII:** Comparison of MAP of proposed FLTSBAQE with Original query, Tomiye et al. and Parapar et al. approach

| Corpus | User Query | Tomiye et al. approach | Parapar et al. approach | *FLTSBAQE* |
|---|---|---|---|---|
| ***CACM*** | 0.1873 | 0.2629 | 0.2752 | **0.2818** |
| ***CISI*** | 0.1586 | 0.2345 | 0.2429 | **0.2516** |
| ***TREC-3*** | 0.1957 | 0.2884 | 0.2847 | **0.3079** |

**Table VIII:** Comparison of P@rank of proposed FLTSBAQE with Original query, Tomiye et al. and Parapar et al. approach for CACM

| | User Query | Tomiye et al. approach | Parapar et al. approach | *FLTSBAQE* |
|---|---|---|---|---|
| **P@5** | 0.4042 | 0.6982 | 0.7443 | **0.7931** |
| **P@10** | 0.3585 | 0.6545 | 0.6970 | **0.7384** |
| **P@20** | 0.2917 | 0.5887 | 0.6244 | **0.6627** |
| **P@30** | 0.2476 | 0.5390 | 0.5739 | **0.6119** |
| **P@50** | 0.1610 | 0.4526 | 0.4851 | **0.5173** |

**Table IX:** Comparison of P@rank of proposed FLTSBAQE with Original query, Tomiye et al. and Parapar et al. approach for CISI

| | User Query | Tomiye et al. approach | Parapar et al. approach | *FLTSBAQE* |
|---|---|---|---|---|
| **P@5** | 0.3648 | 0.5995 | 0.6295 | **0.6683** |
| **P@10** | 0.3202 | 0.5549 | 0.5931 | **0.6241** |
| **P@20** | 0.2536 | 0.4766 | 0.5029 | **0.5289** |
| **P@30** | 0.1983 | 0.4297 | 0.4542 | **0.4782** |
| **P@50** | 0.1128 | 0.3456 | 0.3698 | **0.3938** |

**Table X:** Comparison of P@rank of proposed FLTSBAQE with Original query, Tomiye et al. and Parapar et al. approach for TREC-3

| | User Query | Tomiye et al. approach | Parapar et al. approach | *FLTSBAQE* |
|---|---|---|---|---|
| **P@5** | 0.5087 | 0.7043 | 0.7158 | **0.7302** |
| **P@10** | 0.4783 | 0.6584 | 0.6735 | **0.6938** |
| **P@20** | 0.4431 | 0.6079 | 0.6230 | **0.6437** |
| **P@30** | 0.4073 | 0.5645 | 0.5815 | **0.6059** |
| **P@100** | 0.2905 | 0.4633 | 0.4745 | **0.4908** |

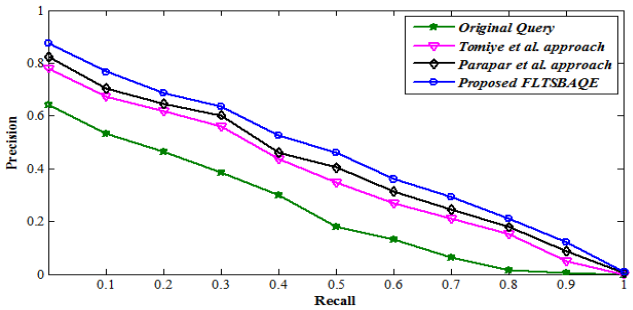


**Fig. 17:** Precision-Recall curves of all approaches for CACM

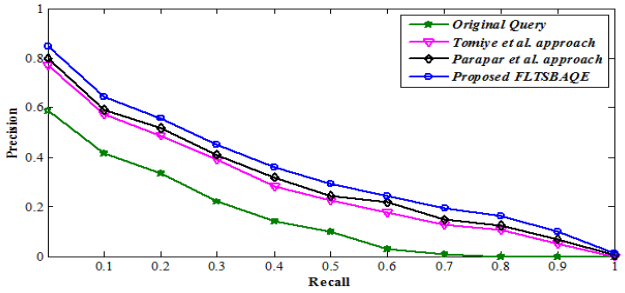


**Fig. 18:** Precision-Recall curves of all approaches for CISI

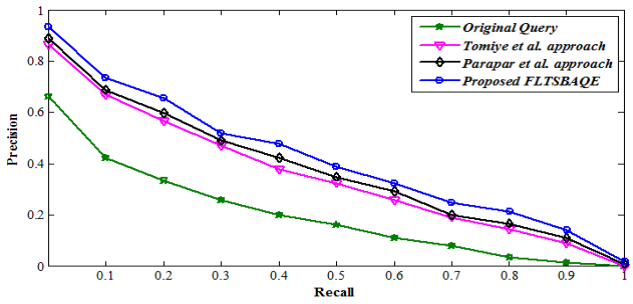


**Fig. 19:** Precision-Recall curves of all approaches for TREC-3

Tables 8-10 tabulates the comparison of proposed *FLTSBAQE* with *original query*, Tomiye et al. approach and Parapar et al.'s approach in terms of *P@rank*. It can be observed from these tables that *proposed approach* outperform all other approaches.

*Precision-Recall* curves are drawn to check the performance of proposed approach for all *three* datasets. Figs 17-19 clearly illustrate that *proposed FLTSBAQE* gets higher *precision* values at all levels of *recall* for *CACM*, *CISI* and *TREC-3*.

### C. *Statistical analysis*

Statistical analysis is performed in terms of *paired t-test* on all datasets. This analysis determines the statistically difference between approaches. The results are presented in Table 11. This table shows the statistically significant improvement in results obtained by *FLTSBAQE* over other approaches at $p = 0.05$.

**Table XI.** Paired t-test results

| | Dataset | ***FLTSBAQE*** | | |
|---|---|---|---|---|
| | | *h-value* | *p-value* | *CI* |
| ***User Query*** | *CACM* | 1 | 0.0000 | [-0.2542, -0.1515] |
| | *CISI* | 1 | 0.0000 | [-0.2326, -0.1365] |
| | *TREC-3* | 1 | 0.0000 | [-0.2764, -0.1559] |
| ***Tomiye et al. approach*** | *CACM* | 1 | 0.0000 | [-0.0952, -0.0591] |
| | *CISI* | 1 | 0.0000 | [-0.0741, -0.0493] |
| | *TREC-3* | 1 | 0.0000 | [-0.0772, -0.0501] |
| ***Parapar et al. approach*** | *CACM* | 1 | 0.0000 | [-0.0549, -0.0317] |
| | *CISI* | 1 | 0.0000 | [-0.0466, -0.0287] |
| | *TREC-3* | 1 | 0.0000 | [-0.0489, -0.0308] |

## V. CONCLUSION

A new automatic query expansion approach based on term selection (*FLTSBAQE*) using *semantic filtering* is introduced in this paper. Firstly, multiple terms selection methods are combined to improve the performance of automatic query expansion. Secondly, the weights of the terms are determined using fuzzy logic. The performance of proposed *FLTSBAQE* is analysed and tested on *CACM*, *CISI* and *TREC-3*. The results were compared and analyzed in terms of MAP, F-measure, recall and precision. The paired t-test is also used to authenticate the results. The results are obtained from the proposed approach are motivating as compared to *original query*; Tomiye et al.'s fuzzy ontology based automatic query expansion approach and Parapar et al. approach.

The present research work is an effort to improve query expansion approach using fuzzy logic and semantic filtering. The work can be extended in many directions in future as new linguistics variables can be used to frame new fuzzy rules. The robustness of proposed term selection method can be tested on other TREC datasets.

## REFERENCES


1. G. W. Furnas, T. Landauer, L. Gomez, and S. Dumais, "The vocabulary problem in human-system communication," Communications of the ACM, vol. 30, no. 11, pp. 964-971, 1987.
2. H. Chen, J. Yu, K. Furuse, and N. Ohbo, "Support IR query refinement by partial keyword set," Proceedings of the second international conference on web information systems engineering, Singapore, pp. 245–253, 2001.
3. B. Kim, J. Kim, and J. Kim, "Query term expansion and reweighting using term co-occurrence similarity and fuzzy inference," Proceedings of the joint ninth IFSA world congress and 20th NAFIPS international conference, Vancouver, Canada, pp. 715–720, 2001.
4. Y. Chang, S. Chen and C. Liau, "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.
5. B. Yates, and R. Neto, "Modern Information Retrieval," Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA, 1999.
6. Y. Bade, R. Bhat, and P. Borate, "Optimization techniques for improving the performance of information retrieval system," International Journal of research on advanced technology, vol. 2, no. 2, pp. 263–267, 2014.
7. K. C. Thompson, "Reducing the risk of query expansion via robust constrained optimization," Proceeding of the 18th ACM conference on Information and knowledge management, NY, USA, pp. 837–846, 2009.
8. K. Raman, R. Udupa, P. Bhattacharyya, and A. Bhole, "On improving pseudo-relevance feedback using pseudo-irrelevant documents," Proceedings of ECIR, pp. 573–576, 2010.
9. R. White, and G. Marchionini, "Examining the effectiveness of real-time query expansion," Information Processing and Management, vol. 43, no. 3, pp. 685–704, 2007.
10. Z. Ye, J. Huang, and H. Lin, "Finding a good query-related topic for boosting pseudo-relevance feedback," Journal of the association of Information Science and Technology, vol. 62, no. 4, pp. 748–760, 2011.
11. J. Singh, and A. Sharan, "Context window based co-occurrence approach for improving feedback based query expansion in information retrieval," International Journal of Information Retrieval, vol. 5, no. 4, pp. 31–45, 2015.
12. Y. Li, S. Chung, and J. Holt, "Text document clustering based on frequent word meaning sequences," Data and Knowledge Engineering, vol. 64, no. 1, pp. 381–404, 2008.
13. J. Aguera, and L. Araujo, "Comparing and combining methods for automatic query expansion," Advances in natural language processing and applications, research in computing science, vol. 33, pp. 177–188, 2008.
14. M. Valdivia, M. Galiano, A. Raez, and L. Lopez, "Using information gain to improve multi-modal information retrieval systems," International Journal on Process Management, vol. 44, no. 3, pp. 1146–1158, 2008.
15. C. Carpineto, and G. Romano, "A survey of automatic query expansion in information retrieval," ACM Computer Survey, vol. 44, no. 1, pp. 1–50, 2012.
16. A. Tomiye, A. Samuel, A. Ijesunor, and I. Udo, "A fuzzy-ontology based information retrieval system for relevance feedback," International Journal of Computer Science, vol. 18, no. 1, pp. 382-389, 2011.
17. J. Parapar, M. Quindimil, and A. Barreiro, "Score Distributions for Pseudo Relevance Feedback," Information Sciences, vol. 273, pp. 171-181, 2014.
18. S. Robertson, "On term selection for query expansion," Journal of documentation, vol. 46, no. 4, pp. 359-364, 1990.
19. J. Swets, "Information retrieval systems," Journal of Science, vol. 141, pp. 245-250, 1963.
20. C. Lee, "Fuzzy logic in control systems: Fuzzy logic controller, Parts I and II," IEEE Transaction on System, Man and Cybernetics, vol. 20, pp. 404–435, 1990.
21. Y. Gupta, and A. Saini, "A novel Fuzzy-PSO term weighting automatic query expansion approach using semantic filtering," Knowledge Based System, vol. 136, pp. 97-120, 2017.