Multimedia Tools and Applications (2022) 81:35195–35221
https://doi.org/10.1007/s11042-020-09172-2

A contemporary combined approach for query expansion

Dilip Kumar Sharma [1,2] • Rajendra Pamula [1] • D. S. Chauhan [2]

Received: 15 February 2020 / Revised: 6 May 2020 / Accepted: 4 June 2020 /
Published online: 3 July 2020

Abstract

The use of an automatic query expansion technique is to enhance the performance of the Information Retrieval System. Selecting the candidate terms for query expansion is an essential task to make query more precise to extract the most suitable documents. This paper provides a method to select the best terms for query enhancement. Firstly, the effect *of abbreviation resolution*, Lexical Variation, Synonyms, n-gram pseudo-relevance feedback, Co-occurrence method on baseline approaches of query expansion is analyzed.. In this work, we used the Okapi BM25 algorithm for ranking. We used Concept-based normalization to deal with concept terms. Here our results show the improvement in results than the baseline approach. A new combined technique that integrates lexical variation, synonyms, n-gram pseudo relevance feedback for query enhancement is proposed. For experimental purpose three English written datasets CACM, CISI, and TREC-3 is used. The obtained results show improvement in the performance of query expansion concerning mean average precision, F-measure, and precision-recall curve.

Keywords Query expansion · Pseudo-relevance feedback · Ontology on query expansion · Information retrieval system · Concept-based normalization

✉ Dilip Kumar Sharma
todilipsharma@gmail.com

Rajendra Pamula
rajendrapamula@gmail.com

D. S. Chauhan
pdschauhan@gmail.com

[1] Indian Institute of Technology(ISM), Dhanbad, India

[2] GLA University, Mathura, India

1 Introduction

Information Retrieval (IR) majorly describes methodologies and techniques that can distinguish the documents relevant to a query from the documents that are non-relevant, accurately in corpora. There are three major categories of IR architecture: representation of documents and queries, ranking function to match queries with documents, and query expansion (QE) [26, 28]. In literature, various methodologies and approaches [56, 59] have been developed and proposed for these categories to improve IR system performance [14, 27, 38, 40, 55]. However, still, there is a vast possibility for QE because of the queries given by users carry ambiguity and uncertainty [23]. Query expansion is a way to enhance the retrieval performance of the IR system. Consider an example when user submits a query car then the search engine needs to extract all the documents related to the car. Query expansion helps in this process by expanding car into cars, automobiles, vehicles. Although few researchers [23, 56, 59] achieve success to some extent, still, colossal scope exists.

Some of the researchers used ontology to capture the semantic meaning of the documents for improving the quality of user queries [34]. The authors identified suitable concepts that pronounce and illustrate the content of documents. The main challenge was to discard unrelated concepts and keeping the relevant concepts only.

Before going to investigate QE approaches, we need to understand the importance of QE and its role. *Why do we require query expansion approaches?* The answer is that users are not experts in giving their requirements in the form of queries. Studies show that most of the time, they give two or three words in queries [60]. This short length queries are ambiguous and uncertain by nature [18, 23]. For example, a user puts his/her request (query) like "What is Java." Through this query, a user might wish to get results for Java technology or be interested in different variants of Javacoffee or looking for Java island. The other reason for using QE approaches is low recall rate as most of the applications demand high précised information but not high recall necessarily [18]. QE is one of the highly used practices for handling the above-discussed issues. Query reformulation or expansion approaches handle the "vocabulary mismatch problem" [23, 54, 66].

There are various QE approaches, but "pseudo-relevance feedback*(PRF)*" [6] is the most popular and widely used approach [9]. In this approach, the user submits his/her request in the form of a text query to an IR system, which extracts relevant documents. Then, top extracted documents are selected and used to identify unique *terms*, called *term-pool*. Researchers used various term weighting schemes to find suitable terms in the literature so far [9, 10, 23, 25, 50, 57]. The *PRF*method has one limitation, i.e., query-drift. If *PRF* returns irrelevant documents, then non-suitable terms would be added in the query, which decreases the performance [23, 25].

In this paper, we used three datasets are us, i.e., CACM, Centre for Inventions and Scientific Information (CISI), and Text REtrieval Conference-3 (TREC-3), to execute all the experiments. CACM dataset consists of titles and abstracts from the CACM journal. It consists of 3204 documents. CISI is a public dataset used for the information retrieval process. Here each dataset contains Unique id, title, abstract, cross-language references to other documents. The trec-3 dataset consists of 5500 labeled questions in the training dataset and 500 questions in the testing dataset. We conduct all the experiments on randomly selected 50 queries from each dataset. The organization of the remainder of the paper is as follows. Section 2 outlines the research objectives of this paper. The related work done in the fields is described in Section 3. Section 4 presents our QE framework for IR. This framework contains various modules, such as document pre-processing and indexing. Section 5 consists of all outcomes and their discussion. This paper includes a conclusion in Section 6.

2 Research objectives

- Firstly the effect of adding synonyms, n-gram PRF method, Abbreviation resolution, Co-occurrence features on the performance of baseline approaches
- We proposed a new combined techniques based query expansion approach. The proposed Technique includes the inclusion of WordNet, Concept-based normalization, use of PRF and co-occurrence features. The integration of these methods includes the benefits of each approach which is resulted in enhanced performance.
- We compared with recently developed similar types of QE techniques on three benchmark datasets, such as CACM, CISI, and TREC-3.

3 Related work

According to the literature, most of the queries contain one, two, or three words only [33]. This problem of short queries increases ambiguity due to the more than one possibility of different interpretations, the same or similar query terms. This problem is called a vocabulary mismatch problem. The positioning of high recall and high precision in IR has been a cause for increasing the work in QE.

Cooper et al. [15] developed various graphical relations for different items to improve the performance of IR. Horng et al. [29] implemented a new term weighting scheme for QE using a genetic algorithm. Gong et al. [22] proposed a new query expanding approach using WordNet. The authors applied the concept of substitution and hypernym/hyponymy associations in their approach. This proposed approach obtained significantly improved results. Fattahi et al. [21] used Domain-related topical and non-topical terms to develop a QE model. Latiri et al. [41] in his proposal for a QE approach, used association rule mining techniques. Also, the term weighting scheme, presented by Bendersky et al. [5] used a genetic algorithm to determine the weighting factor for the query.

Recently, less work has been reported using data sources for QE by researchers Anand et al. [1]; Azad et al. [3, 4]; Bouchoucha et al. [7]; Hsu et al. [30]; Kotov et al. [37]; explored the usage of Wikipedia to enhance QE for retrieving relevant documents for an original query. Few Authors suggested the use of automatically built glossary in query expansion. To create the thesaurus, they used the relationships between concepts like antonym, synonym, part-of, or hyponym. Adding the terms to query using ontology may retrieve irrelevant documents, also Li et al. [42]; Macdonald et al. [43]. Mahler [44] proposed a holistic query expansion (HQE) method. Wang Zhixiao et al. [64] dispensed an approach for QE that relies on Global Analysis and Ontology (GAO) along with statistical and semantic approaches. Chang et al. [12] implemented a new fuzzy-based QE approach to improve the IR system. The results were satisfactory. Nowacka et al. [46] developed a novel method for QE using fuzzy logic to improve the process. Khennak et al. [35] developed the QE approach using accelerated particle swarm. Using the MEDLINE dataset on their approach and got exceptional results. Singh et al. [56] proposed a different approach. They used different lexical co-occurrence techniques that relied on the corpus. Also, choosing the best grouping of query terms from the term pool obtained using pseudo-feedback based QE. Gupta et al. [23] introduced a new QE approach based on co-occurrence and PRF, verifying their results using CISI, CACM, and TREC-3 datasets, which improves the implementation of IR.

Htun et al. [31] performed two user studies based on different collaborative search interfaces and information access scenarios. They used several IR, Collaborative IR, and Collaborative Information Seeking evaluation metrics to compare the interfaces. Authors concluded that various features of a query like time spent on a query, popularity of a query, and effectiveness of a query might be useful to get information about retrieval performance. Singh et al. [58] used rank fusion and semantic notion methods in query expansion. Raza et al. [48] presented a survey on using Statistical approaches for query expansion. Fang et al. [20] suggested a semantic sequential dependence model (SSDM). To fetch biomedical documents, they used the combination of semantic content and conventional SDM. The authors used BioASQ is used as a dataset here, which depicted satisfactory results. Sharma et al. [52] provides a comparative study of recently established query expansion methods using fuzzy logic to obtain relevant documents from broad datasets for a specific user query. Dahab et al. [16] used spectral analysis information retrieval to consider term proximity. Here, the Authors used the distribution approach (KLD) and WordNet to select the candidate features distribution. Zingla et al. [70] used association rules for term retrieval. In this, the authors focused on combining external sources with association rule mining to enhance query representation. To deal with query drift in the future, one can add more significance to the user query terms. Huang et al. [32] proposed code changes approach for query expansion. The result shows the enhancement in precision by up to 52% to 62%. This method gives an effective code search.

Azad et al. [2] presented a survey on QE techniques implemented between 1960 and 2017. They compared these techniques in terms of various features like data sources used weighting and ranking methodologies, user participation, and applications. Bounhas et al. [8] presented a morpho-semantic knowledge graph from classical Arabic vocalized corpus. They mixed Ghwanmeh stemmer and MADAMIRA to retrieve a multi-level lexicon from an Arabic vocalized corpus. Perez et al. [47] proposed an automatic QE approach based on collaborative feature location models. Azad et al. [3] developed a new QE approach using Wikipedia and WordNet. To address the relationship problem, the query terms, the authors suggested a novel method using Wikipedia –WordNet as data sources. The authors used In-link and tf-idf approach to give weight to expansion terms. They obtained satisfactory results. Sharma et al. [53] used cuckoo search and particle swarm optimization to present a new QE technique. The authors used a cuckoo search and fuzzy logic-based technique for query expansion. The authors used the first Okapi-BM25 ranking function to collect documents, then created a candidate term pool. Then Cuckoo search is applied to get the best candidate expansion terms to append in the query. Datasets used were CISI, CACM, and TREC-3. This technique was tested on three benchmark datasets extracting satisfactory results.

Chandra et al. [11] demonstrate the use of query enhancement in the Cross-Lingual IR(CLIR) system. Here, the user can input their query in any known language and get the output in the desired language. They used Hindi-English CLIR to convert the Hindi search into English documents to enhance MAP results. However, in the future, it is required to increase the quality results and decrease the associated time spend in fetching relevant documents.Kumar et al. [39] propose the use of user profile information like user ratings and user tags to fetch relevant information. Here PRF method is used to grasp social information. There is a possibility to use all user features to extract relevant items in the future. Vocabulary mismatch problem occurs in case of microblog texts due to their shorter length, so to address this challenge, authors wang et al. [65] used a k-Nearest Neighbor (kNN) algorithm to create termsfrom local word embeddings to increase the initial query. They used official TREC Twitter corpora as a dataset, and the result indicates that this method accelerates the retrieval performance. Azad et al. [4] try to predict the relationship between terms and determining the relationship of terms to the user query. The authors suggested a query expansion procedure that uses term frequency and inverse document frequency algorithm, use of k-nearest neighbor (kNN) based cosine similarity, and the calculation of correlation score. They used the FIRE dataset to examine the proposed method. This method gives satisfactory retrieval performance in comparison to the existing approach. Dalton et al. [17] suggested word-based and entity-based methods to deal with complex topics, which result in a 20% improvement in MAP value. Nasir et al. [45] discussed the novel Query expansion approach that uses both corpus-based techniques and relevance feedback. In this semantic similarity is determined between the user query and the retrieved candidate features using different knowledge which helps in query reformulation. Torjmen-Khemakhem et al. [63] discussed a novel QE process for medical text depending on retro-semantic mapping. In this text database and UMLS concepts are used for medical image retrieval, then the most appropriate concept is selected. Gupta et al. [24] suggested a novel swarm-based approach which used KHM and fuzzy logic for making quality data clustering. Two standard datasets CASM and CISI, were used to evaluate the clustering results. This approach combines with document clustering in future work. Wu et al. [67], the authors combine pattern mining, topic modeling, and relevance ranking techniques for multi-document summarization. The authors used DUC 2006 and DUC 2007 datasets for the experimental purpose, which gives better performance than existing methods.

Esposito et al. [19] propose a hybrid approach using lexical resources and word embedding for query expansion to design a Question-Answering system for the Italian language and considering the cultural heritage domain. There is a need to consider other languages and different domain's knowledge to generalize the result as a future scope. This method is a standard statistical approach. Chaudhary et al. [13] proposed a framework, Image Retrieval using Knowledge Embedding (ImReKE), for embedding knowledge into images and queries using a knowledge base, allowing retrieval approaches to understanding the context of queries and images in a better way. The improvement achieved is maximum when knowledge embedding created using the proposed knowledge base in comparison to existing KBs, ConceptNet, and ImageNet. Khennak et al. [36] suggested two correlation measures for query reorganization. The one is called external Correlation, which relies on features co-occurrence and another, is an internal correlation using term proximity. The authors used MEDLINE as a dataset, and the experiments showed the robustness of this method compared to existing work.

In the approaches mentioned above, suitable terms identified and used for QE without considering the effect of various necessary conditions to obtain the optimized query using knowledge sources and other IRS parameters. This paper explores the effect of various processes on query expansion. Then according to the effect of these processes, we are combining and presenting a framework for Automatic Query Expansion (AQE).

4 Proposed query expansion framework

In this section, we explain various components or elements of the presented AQE framework, as shown in Fig. 1. The modules of this framework are discussed in detail in the following subsections. This model composes of five components: Document preprocessing, Query preprocessing and expansion, Indexing, Querying, and Document retrieval. This proposed method combines several techniques to enhance the performance of the IR system. Contextual information is also used here. WordNet to applied to find the association between query terms and document terms. Our proposed method improved the pseudo relevance feedback method by adding a Linear document context combination, which is the specialty of our method. This method helps removal of non-relevant documents that are retrieved during the pseudo relevant feedback process.

4.1 Document preprocessing

All three datasets CACM, CISI, and Trec-3 are preprocessed using these steps:

1. **Segmentation of sentences:** Each document contains a few paragraphs, and these paragraphs are composed of many sentences. We have segmented all the sentences within paragraphs using the Lucene framework written in Java. This step identifies the boundaries of each sentence among words. Generally, punctuation marks in written language are used for sentence boundaries, which differentiate sentences from each other. Here, dictionary-based segmentation is used, which is more flexible and appropriate for query expansion. It uses a lexicon tool to decide the boundary of words [68].
2. **Replacement of characters:** After the segmentation of sentences, we did character replacements in sentences. Textual characters replace all character images. We also replace all roman numbers with Arabic numbers. The replacement of characters is essential and helps to capture the right semantic terms in query and document matching.

   In this work, we have also replaced the hyphens in sentences such as T.V. is replaced with Television, and Hewlett-Packard is replaced with Hewlett Packard. Similarly, we have made many other changes in the character replacement process. This step brings all the words in a single format.

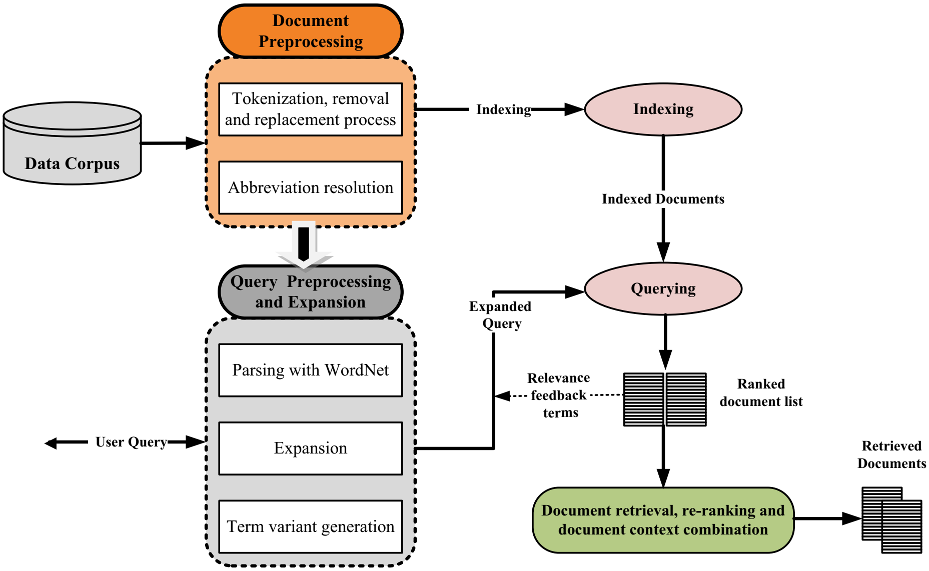

Fig. 1 Framework of proposed automatic query expansion

3. **Removal of text and special characters:** We remove all HTML tags, headings, and abbreviations to reduce the matching complexity between a query and a document. We have removed concise sentences as they produce some noise in determining precision and recall. Concise sentences indicate sentences consist of one or two words in it. We also removed the special characters like hyphens, asterisks, slashes to smoothen the matching process.
4. **Abbreviation Resolution:** After the removal of special characters. In this method, we expand the specific abbreviation, which helps in retrieving relevant results.e.g.inc. Means incorporate.
5. **Tokenization:**It transforms the text of a document into a list of words or tokens. It usually requires some preprocessing on text documents like translating all words in the lower case, eliminating punctuation characters etc. Individual tokens are identified upon initial preprocessing. If the text of the document has been tokenized, it is crucial to determine which words to be used. It means it has to be determined which descriptors are useful in the joint position of defining the content of the document and separating the document in the set from the other document.
6. **Stop Word Removal:**Stop word removal is the process of removal of the least essential terms from documents and queries. Stop words consist of prepositions, conjunctions, and articles. The reason for this process is that these terms are not functional and are not useful for identifying high rank documents with respect to a user query. Even though these stop words have a grammatical function and are essential for the comprehension of sentences, they are of slight use in discriminating some documents from other documents as part of IR.
7. **Stemming:**Stemmingis a technique, which removes suffixes from terms in order to reduce them to a common stem form called root form. Stemming typically removes gerunds (ing*)*, plurals, and past tenses.

4.2 Document indexing and query expansion

All the documents are indexed after preprocessing the whole corpora, aWe used the Lucene framework to create an index for all documents. We did indexing on each document of the corpus. Further, queries are also preprocessed as documents are done then queries are parsed to find out the phrases for similar concepts. Instead of using any explicit method, we have used knowledge sources to deduce fitting terms that can be used for QE in various cases. These sources of knowledge are much related to the data corpora used in this paper and can be categorized as ontological assets and automatically produced knowledge sources. In this work, our primary focus to determine the best word association for query expansion, and it is query-specific. These word associations are as follows:

- Lexical analysis: In this work, we find out variants of each query term. These variants may be singular, plural, orthographic, spelling variation, and morphological.
- Semantic analysis: When we perform lexical analysis, the main problem is with the synonyms. Typically, synonyms are lexically distinct but meaning wise; these are similar, which is known as semantic similarity.
- Ontological analysis: We used Ontological resources such as WordNet API. We used these resources to find out relationships between query terms and expansion terms. The

ontological analysis provides specialized or generalized associations between terms. It shows a has-a relationship or is-a-part-of relationship. Therefore Wordnet parsing is applied to find all the possible synset of candidate expansion terms.

- Co-occurrence relationships analysis: We determined co-occurrence analysis from each corpus Gupta et al. [23]. Two or more terms appear in corpus many times; then, these terms used for co-occurrence analysis. For example, the expansion comes with a query many times; then, QE can be considered as co-occurred terms. Two types of co-occurrence relationships are being used in this work: *PRF* and statistically regular n-grams taken out from corpus.

4.3 Query normalization based on concepts

In this work, we are using the Okapi-BM25 model Robertson et al. [49, 50]; Gupta et al. [28] for ranking the documents. This ranking model is a widely used method and rank the documents for queries based on similarity among them. However, this method has few issues, also like Okapi-BM25is not able to differentiate query terms and concept terms. The other issue is corpus related problem, when we use it for TREC-3, The documents and queries in TREC-3 dataset usually contain more than one concept terms, and Okapi-BM25 fails to discriminate that. Here, the length of the documents also plays an important role. If document length is less and has one conception only, a higher rank given to this document as compared to another document, whose length is more and contains more than one concept as this document has other concepts too.

Zhou et al. [69] presented a solution to overcome the issues mentioned above using a conceptual IR model. Stokes et al. [61] and Stokes et al. [62] also proposed a new and more effective solution, i.e., concept-based query normalization, to resolve these issues. We have used this method to overcome these issues. To solve the former issue, we have categorized terms into two categories: concept terms t_c), and non-concept terms (t_n). To divide query terms into these categories, we define query terms as non-concept terms that are not in any knowledge source. In contrast, query terms called concept terms; those have entries in knowledge sources. Therefore, the similarity between a document d and a given query q with both concept and non-concept terms can be determined as follows [62]:

$$sim_{score}(q,d) = nsim_{score}(q,d) + csim_{score}(q,d) \tag{1}$$

Where $nsim_{score}$ *(q, d)* represents the similarity score for the non-concept query, $csim_{score}$ *(q, d)* represents the similarity score for concept query. These similarity scores can be defined mathematically as:

$$nsim_{score} = \sum\nolimits_{t\in Q_n} sim_t(q,d) \tag{2}$$

$$nsim_{score} = \sum\nolimits_{t\in Q_n} r_{p,t}.w_t.r_{q,t} \tag{3}$$

where Q_n is the collection of all non-concept terms in query q, and

$$r_{p,t} = \frac{(k_1+1).f_{p,t}}{k_1\cdot\left[(1-b)+b\cdot\frac{W_p}{avgW_p}\right]+f_{p,t}} \tag{4}$$

$$W_t = \log \frac{N - f_t + 0.5}{f_t + 0.5} \tag{5}$$

$$r_{q,t} = \frac{(k_3 + 1).f_{q,t}}{k_3 + f_{q,t}} \tag{6}$$

Where k_1 and b are constant, and the values changed to *1.2* and *0.75*, respectively. k_3 is set to *7* in this work after empirical study. W_p indicates document length, $avg(W_p)$ indicates the mean document length in whole data collection. In this context, the total number of documents in the collection are represented by N, while f_t shows the number of documents having the term t. The similarity score for concept terms can be defined as:

$$csim(q, d) = \sum\nolimits_{C \in Q_c} sim_c(q, d) \tag{7}$$

$$csim(q, d) = \sum\nolimits_{C \in Q_c} Norm(sim_{tc1}(q, d), \ldots\ldots, sim_{tcN}(q, d)) \tag{8}$$

Where Q_c is the collection of all concept terms in query q, and C represents a concept in Q_c. t_{ci} is a concept term belonging to C.

4.4 Pseudo relevant N-gram feedback

After gathering the concepts, We explore the Pseudo Relevant N-gram feedback method. The steps for it are listed below:

1. First, we retrieve relevant documents using Okapi and select only the top 100 most relevant.
2. Afterward, we find all unique terms, unigrams, bigrams, and trigrams from these top 100 relevant documents. We formed tokens formed using the Lucene framework.
3. After collecting all N-grams, we compute the suitability score of each term using the method proposed by Gupta et al. [23]. Gupta et al. [23] presented a new method for term reweighting that relies on four IR factors, while they used this method to compute the suitability of terms for query expansion. Further, they arranged all terms in decreasing order of suitability score, and at last, we select the top 10 terms for query expansion.

4.5 Document context combination

Although, the selection of terms of QE using the above-discussed method improves the quality of a query and also shows the significant improvement in MAP too. However, some drop can be observed in MAP for some of the queries in both the datasets. Gupta et al. [23] have been witnessed this negative impact on document retrieval. The reason for it to select non-relevant documents using PRF. To resolve this problem, we have used a linear document context combination method in this paper, which is described below:

1. We divide all selected top-ranked 100 relevant documents according to different concept levels as the number of concepts these documents have.

2. Then we re-compute the score of each document as Eq. (9) also, re-rank the documents in each group [62].

$$S_i = P_i + \alpha * \frac{D_i}{D_{max}} * P_{max} \tag{9}$$

Where D_{max} and P_{max} are the maximum similarity scores among all top 100 documents. We set Alpha to 0.5 after empirical study.

5 Experimental analysis

As we have discussed that CACM and CISI documents are in plain text format, whereas TREC-3 documents are XML documents. CACM consists of 3204 English text documents and 64 queries. CISI has 1460 English written text documents on IRand contains 100+ queries. TREC-3 has more than eight lacs of documents. In this research work, MAP, F-measure, and precision-recall plots are the terms used to compute the performance of the proposed method. The average precision is the mean of the precision scores after each relevant document is retrieved. MAP is average precision across multiple queries. The mathematical representations for evaluating parameters are as follows:

$$\text{Precision} = \frac{\text{number of retrieved relevant documents}}{\text{number of retrieved documents}} \tag{10}$$

$$\text{Recall} = \frac{\text{number of retrieved relevant documents}}{\text{number of relevant documents}} \tag{11}$$

$$\text{F}-\text{measure} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision} + \text{Recall}} \tag{12}$$

5.1 Comparing baseline approaches

All the experiments have been divided into two types of evaluation analysis. The first analysis is done on comparing baseline performance without adopting the QE approach, and comparison is also made with the Okapi ranking algorithm along with other baseline approaches for all three datasets. In this analysis, we have observed the effect of various techniques used in the proposed QE approach on results. We did the second analysis on the combined effect of all different techniques on results. In this type of analysis, we have compared the results with recently developed QE approaches of Gupta et al. [25], Khennak et al. [35], and Sharma et al. [53].

The main objective of performing this experiment is to demonstrate the improvements in results using the baseline model, which can be used for the proposed QE approach. Table 1 shows the comparison of MAP values obtained by Okapi-BM25 and our concept-based normalization baseline model for all three datasets.

Table 1 Comparison of MAP values obtained by the baseline approach and Okapi-BM25

Dataset	Okapi-BM25	Baseline model
CACM	0.1873	0.2617
CISI	0.1586	0.2261
TREC-3	0.1957	0.2708

This table gives the overall performance analysis of the models. However, it is also interesting to observe query-wise analysis for both the datasets. In this work, 50 queries were selected from datasets CACM and CISI, while a selection of 50 queries in number (151–200) was taken from TREC - 3 datasets for query wise analysis. We compute F-measure for all queries, and analysis is done for the top 50 documents. Figures 2, 3 and 4. illustrates results. These figures clearly illustrate that the Baseline model gives higher F-measure values than Okapi-BM25 for all three datasets.

Figure 2 depicts that our baseline model gets better F-measure values for all selected 50 queries as compared to Okapi-BM25 in the case of CACM datasets. A similar type of results can be shown in Figs. 3 and 4 for CISI and TREC-3 Datasets.

5.2 Analyze the effect of lexical variation on query expansion

In Section 4, we have discussed various types of terms in QE approaches. These are lexical variation, synonymy, ontologically related words, and co-occurring terms. These types make much impact on finding terms for query expansion. In this subsection, we are considering term equivalence through lexical variation as a relation type. As we discussed in the above section about splits of words at breakpoints, for instance, a hyphen. In this work, we use WordNet as a lexical database. Table 2 shows the comparison of *MAP* values obtained by lexicon variation based QE with the baseline model. We can see that there is a slight improvement in results after using lexicon variation based query expansion.

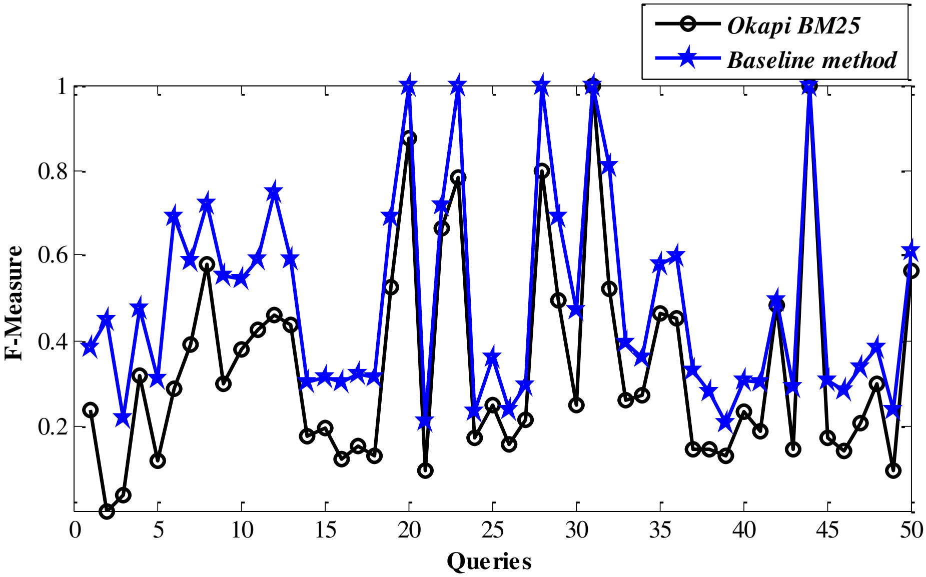

Fig. 2 Comparison of F-measure values obtained by the Baseline approach with Okapi-BM25 for CACM

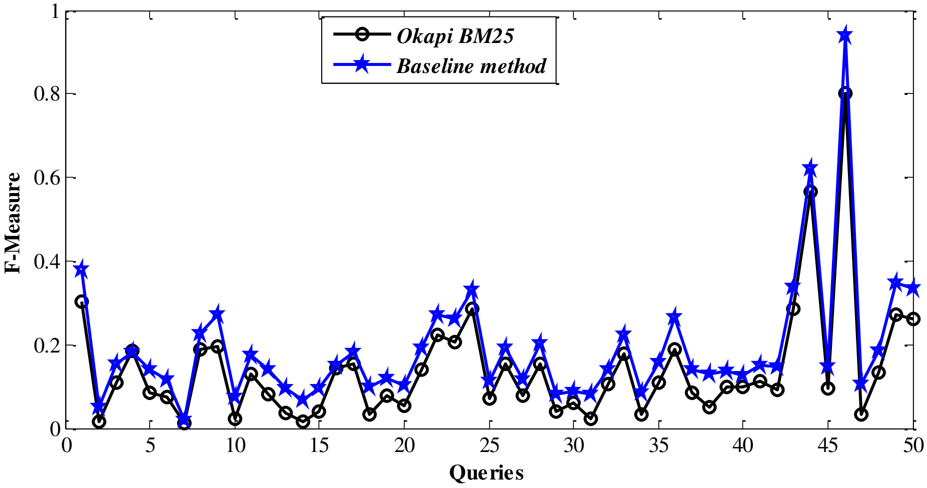

Fig. 3 Comparison of F-measure values obtained by the Baseline approach with Okapi-BM25 for CISI

The query wise analysis is also done to check the effectiveness of lexicon variation based query expansion. Figures 5, 6 and 7 shows the results for CACM, CISI, and TREC-3 datasets.

Figure 5 presents the comparison of the Baseline approach with the Lexical variation based approach for the CACM dataset. It is clear from this figure that Lexical variation has a favorable impact, as seen on the prior performance of the QE method. Similarly, the comparison is shown in Figs. 6 and 7 for CISI and TREC-3 datasets, respectively.

5.3 Analyze the effect of synonyms on query expansion

This section discusses the effects of synonyms on query expansion. In this work, we convert terms into synonyms for QE. The experiment analysis is done as above: overall and query wise. Table 3 compares Mean Average Precision (MAP) values obtained by approaches. This

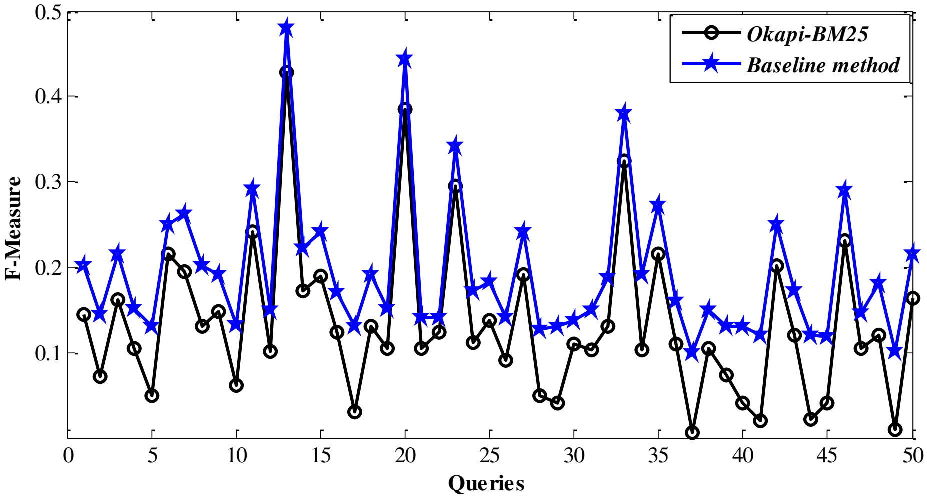

Fig. 4 Comparison of F-measure values obtained by the Baseline approach with Okapi-BM25 for TREC-3

Table 2 Comparison of MAP values obtained by the baseline approach and Lexical variation based QE approach

Dataset	Baseline model	Lexical variation based Query Expansion
CACM	0.2617	0.2733
CISI	0.2261	0.2448
TREC-3	0.2708	0.3012

table shows that MAP values have minor improvements, which cannot be considered as significant.

The comparison of F-measure values is also presented in Figs. 8, 9 and 10 for query wise analysis. Figure 8 shows the comparison for the CACM dataset. This figure depicts the usage of synonyms in the QE approach improves the performance slightly. A similar analysis is made for CISI and TREC-3 in Figs. 9 and 10, respectively. Synonyms enhanced the performance in CISI and TREC-3 datasets also. Finally relevant documents are collected using the synonyms feature.

5.4 Analyze the effect of abbreviation on query expansion

The abbreviation is one type of synonymy and used very frequently in domain-specific documents such as science. As we have discussed earlier in the above sections that there are three types of using abbreviations. First, creating a set of long-form and short-form pairs (Schwartz et al. [51]), second, using a resource, and last, converting all abbreviations to their long forms in the corpus.

We have used the last one in this work: converting all abbreviations into their extended forms. Table 4 shows the comparison of MAP values achieved by various approaches. The query wise analysis is also shown in Figs. 11, 12 and 13.

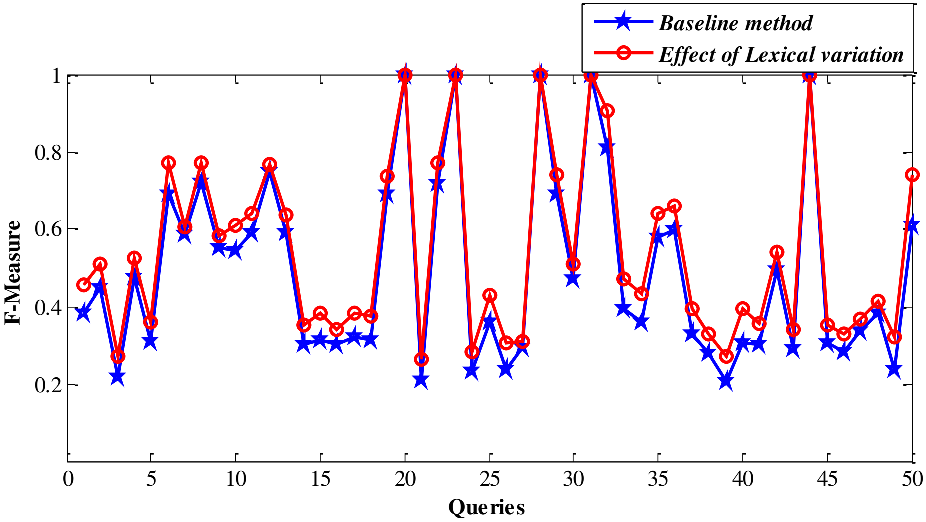

Fig. 5 Comparison of Baseline approach with Lexical variation based QE in terms of F-Measure for CACM

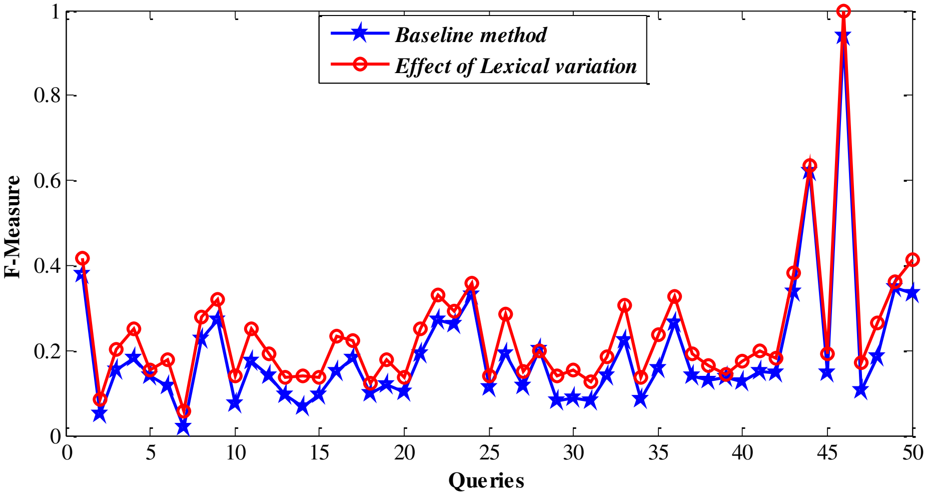

Fig. 6 Comparison of Baseline approach with Lexical variation based QE in terms of F-Measure for CISI

Figure 11 demonstrates the results for CACM datasets, and it is clear from this figure that using an abbreviation for QE decreases F-Measure values for almost all 50 queries. Similarly, the performance of QE is dropped in the case of the other two datasets CISI and TREC-3 after using abbreviations, as shown in Figs. 12 and 13, respectively.

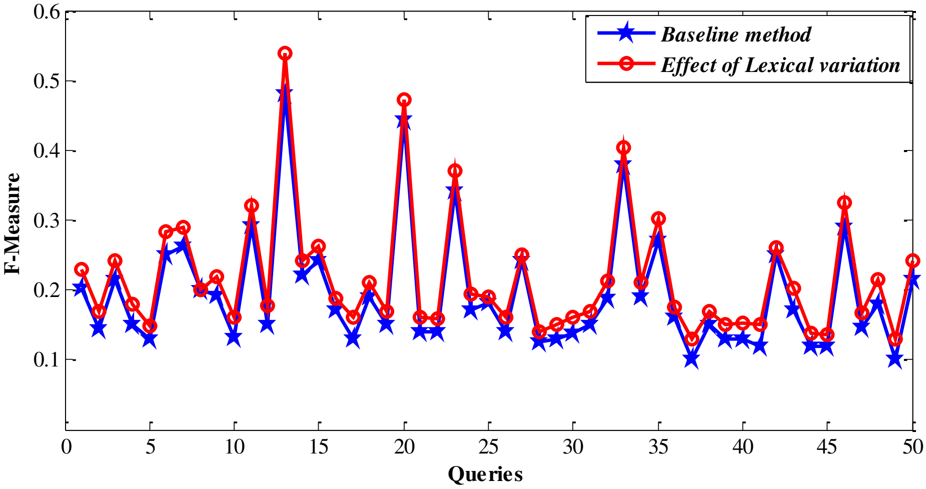

Fig. 7 Comparison of Baseline approach with Lexical variation based QE in terms of F-Measure for TREC-3

Table 3 Comparison of MAP values obtained by the baseline approach and Synonym based QE approach

Dataset	Baseline model	Synonyms based Query Expansion
CACM	0.2617	0.2649
CISI	0.2261	0.2304
TREC-3	0.2708	0.2796

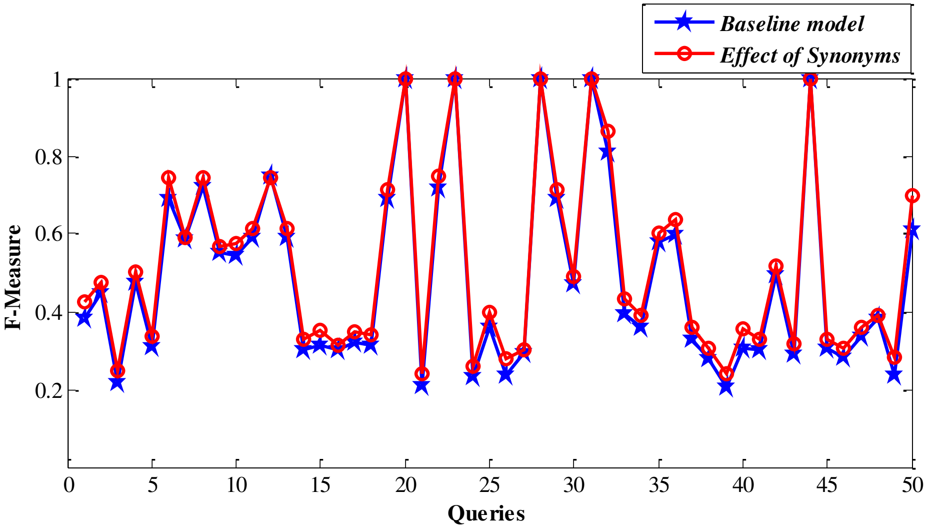

Fig. 8 Comparison of Baseline approach with synonym based QE in terms of F-Measure for CACM

From the results mentioned above, it can be analyzed that ambiguity becomes an issue in using abbreviation query expansion. For example, some type of abbreviation may have more than one long-form or full form. Therefore, most of the queries are suffered from query-drift problem, and the QE approach could not perform better.

5.5 Analyze the effect of related and co-occurring terms on query expansion

Till now, ontological resources were used for analyzing query expansion. In this section, we are focusing on the use of the co-occurring terms are for query expansion. In a log-likelihood association metric, we take the document length as the window size for a co-occurrence pair.

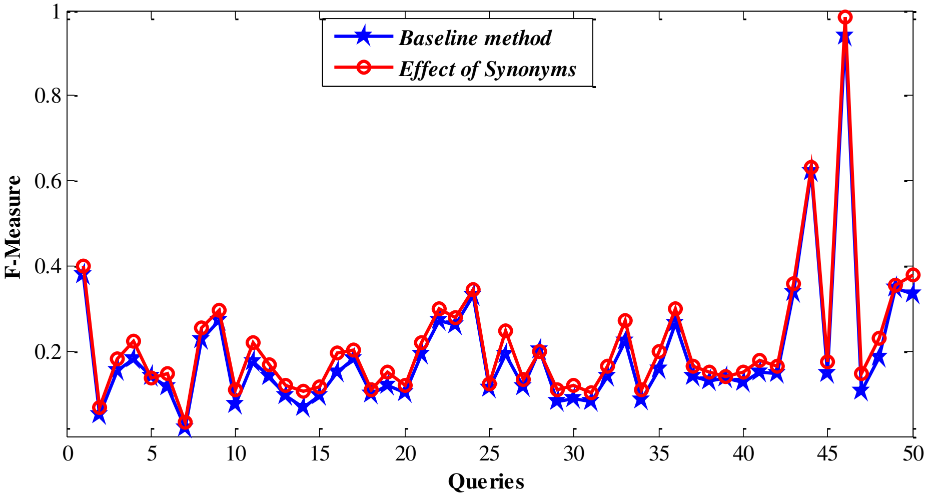

Fig. 9 Comparison of Baseline approach with synonym based QE in terms of F-Measure for CISI

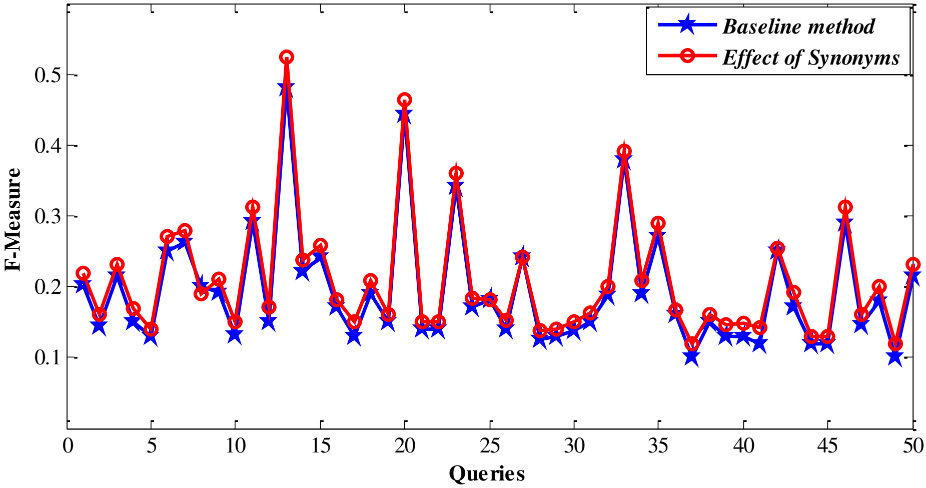

Fig. 10 Comparison of Baseline approach with synonym based QE in terms of F-Measure for TREC-3

We find the terms; that are frequently appearing together in documents. At last, we add these co-occurring terms to the original query.

Table 5 presents the results for MAP values obtained from our expansion experiments using co-occurring terms. We have also compared the results with the Baseline approach. This table clearly shows a little improvement in the performance of QE as compared to the Baseline approach. For query wise analysis, we compute F-measure values for each query using various approaches, as demonstrated in Figs. 14, 15 and 16 for all three used datasets.

Figure 14 clearly shows that using co-occurrence for QE does not make much impact on the improvement of performance for CACM as the values are almost the same for most of the queries. The similar types of results and analysis exist for CISI and TREC-3 in Figs. 15, and 16 respectively.

5.6 Analyze the effect of PRF relevance N-gram on query expansion

This section shows the effect of using the PRF relevance N-gram method for query expansion. The proposed approach is based on pseudo relevant N-grams feedback, as discussed in above Section. In this work, we have taken unigrams, bigrams, and trigrams as co-occurring terms from datasets. Further, we remove all stop words from obtained all N-grams. We have also removed less frequent N-grams; those are occurring less than two. Then, we have determined the top ten associated phrases for the given user query. Table 6 tabulates the results for MAP values obtained from our expansion experiments using N-gram PRF based QE and Baseline approach. This table clearly shows that PRF using unigram and bigram improve MAP values significantly in comparison to others.

Table 4 Comparison of MAP values obtained by the baseline approach and abbreviation based QE approach

Dataset	Proposed Concept-based normalization (Baseline) model	Abbreviation based Query Expansion
CACM	0.2617	0.2593
CISI	0.2261	0.2239
TREC-3	0.2708	0.2659

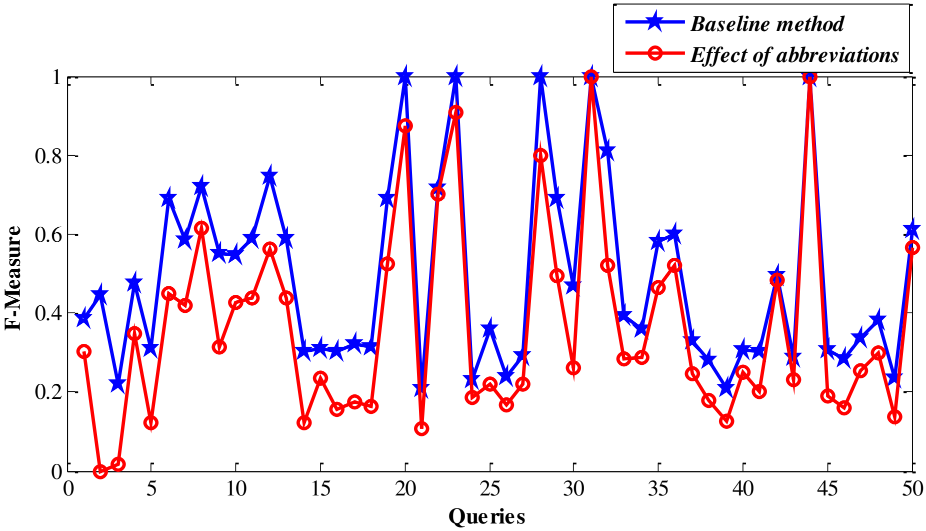

Fig. 11 Comparison of Baseline approach with Abbreviation based QE in terms of F-Measure for CACM

For query wise analysis, we compute F-measure values for each query using various approaches, as demonstrated in Figs. 17, 18 and 19 for all three datasets. Figure 17 shows some improvement in performance using N-gram PRF based QE for the CACM dataset. Figure 18 also shows some improvement in performance for CISI, and a similar type of enhancement can be observed for TREC-3 in Fig. 19.

Pseudo relevance N-gram feedback improves the performance as the top-ranked documents contain two or more than two query concepts. These remove any ambiguity from each other mutually. Hence it improves the likelihood of feedback terms for query expansion.

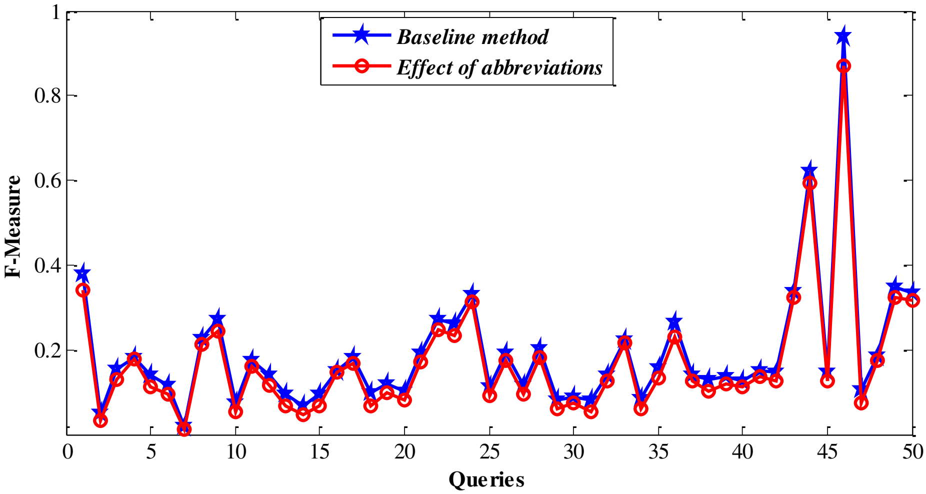

Fig. 12 Comparison of Baseline approach with Abbreviation based QE in terms of F-Measure for CISI

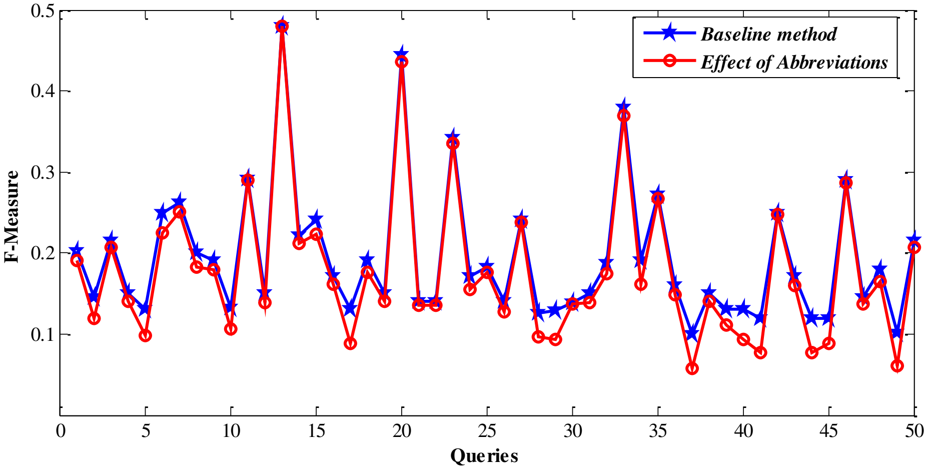

Fig. 13 Comparison of proposed Baseline approach with Abbreviation based QE for TREC-3

5.7 Analyze the combined effect of all techniques on query expansion

This section analyzes the combined effect of all techniques discussed above on query expansion. We include Lexical variation, Synonyms, Co-occurrences, and PRF relevance N-

Table 5 Comparison of MAP values obtained by the baseline approach and Co-occurrence based QE approach

Dataset	Baseline model	Co-occurrence based Query Expansion
CACM	0.2617	0.2631
CISI	0.2261	0.2278
TREC-3	0.2708	0.2753

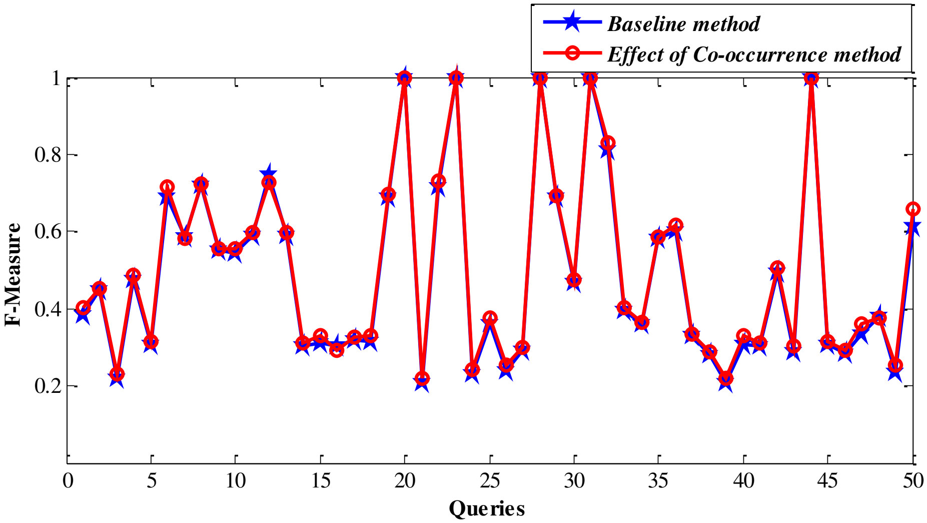

Fig. 14 Comparison of Baseline approach with co-occurrence method in terms of F-Measure for CACM

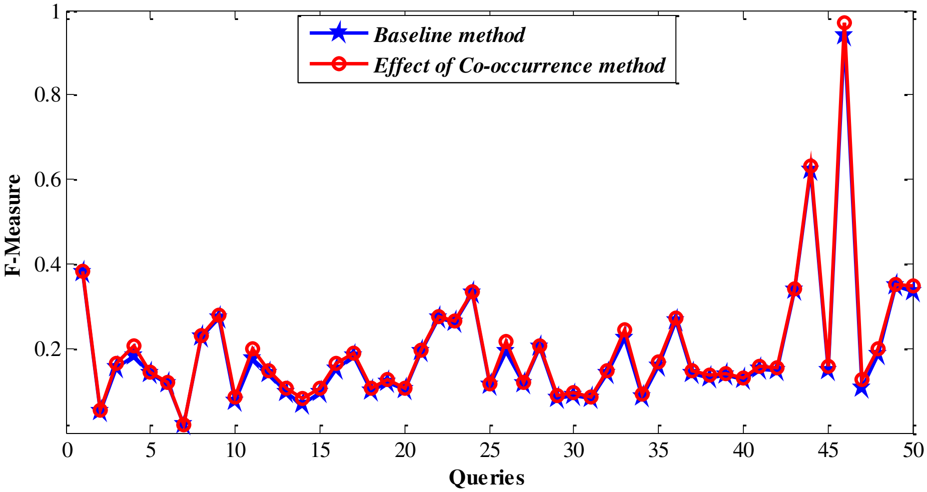

Fig. 15 Comparison of Baseline approach with co-occurrence method in terms of F-Measure for CISI

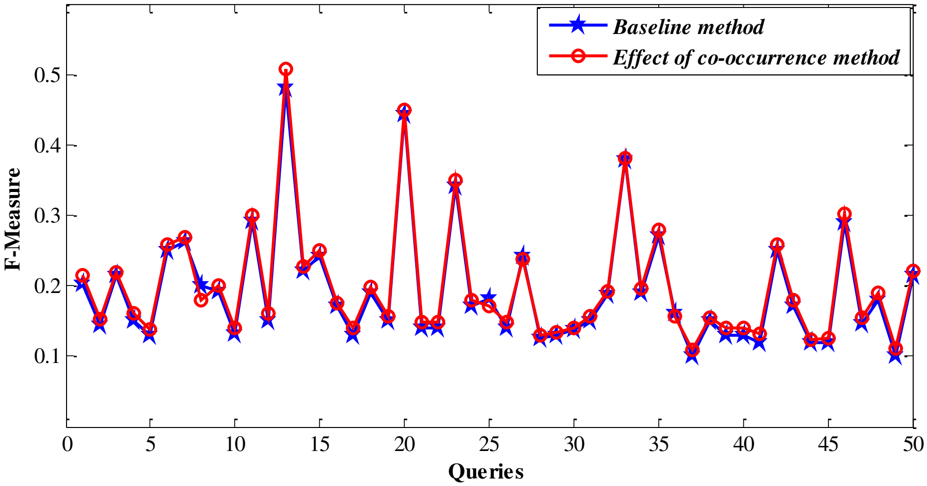

Fig. 16 Comparison of Baseline approach with co-occurrence method in terms of F-Measure for TREC-3

gram techniques to implement the QE approach. The abbreviation is not taken as it degrades the performance. Table 7 tabulates the results for MAP values obtained from our combined techniques based QE approach. Their results are now compared with the Baseline method and recently developed QE approaches of Gupta et al. [25], Khennak et al. [35], and Sharma et al.

Table 6 Comparison of MAP values obtained by the baseline approach and PRF relevance N-gram based QE approach

Dataset	Baseline model	N-gram PRF based Query Expansion
CACM	0.2617	0.2684
CISI	0.2261	0.2396
TREC-3	0.2708	0.2881

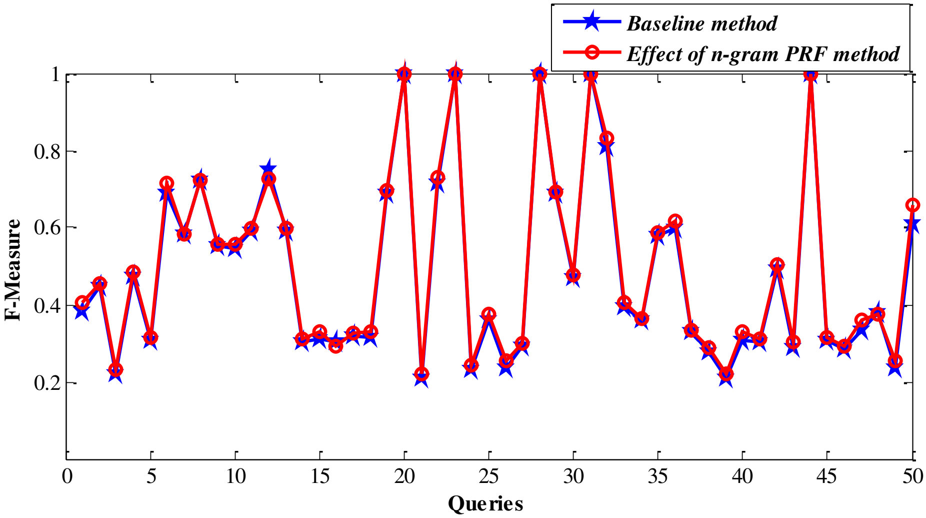

Fig. 17 Comparison of Baseline approach with N-gram PRF based QE in terms of F-Measure for CACM

[53], Table 7 shows that our combined techniques based QE approach gets higher values of MAP as compared to the rest of the approaches for CACM and CISI datasets. In the case of TREC-3, the performance of our proposed approach turned to be better than the Baseline method approaches, Khennak et al. [35], and Sharma et al. [53], but it lacks behind from Gupta et al. approach [25]. Our proposed method reduces the MAP value by 1.71%.

The proposed combined techniques based QE causes improvement in results as compared to other approaches because this approach takes the advantages of each method. The abbreviation method decreases the performance; therefore, it is not included in the proposed combined techniques based approach.

Query wise analysis and comparison is also performed to check the performance of all QE approaches. Figures 20, 21 and 22 present the comparison of F-measure values for all three

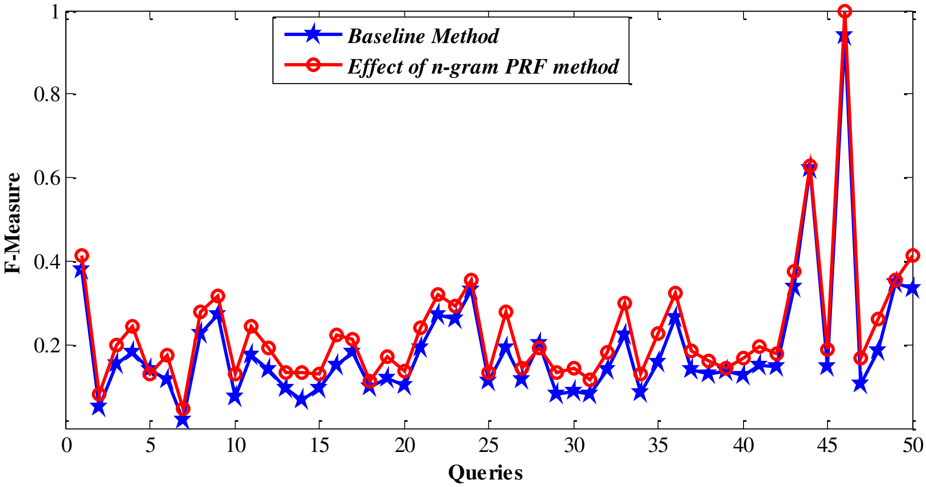

Fig. 18 Comparison of Baseline approach with N-gram PRF based QE in terms of F-Measure for CISI

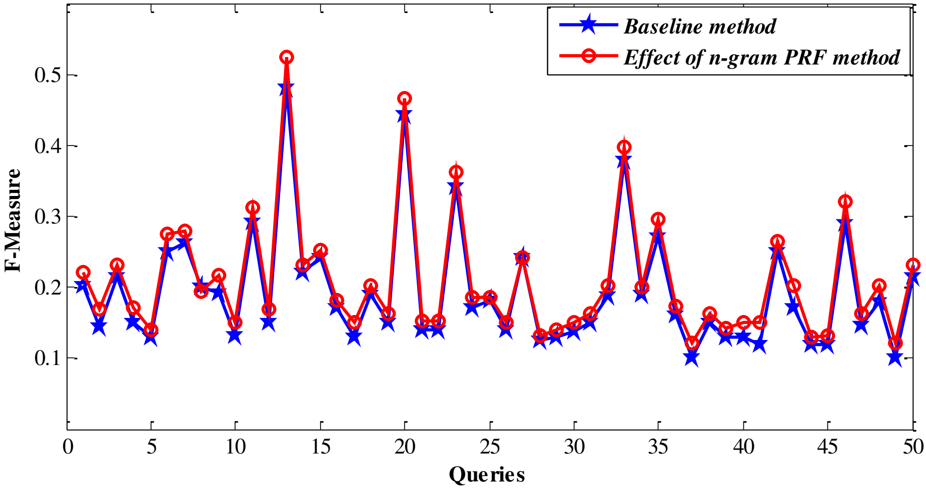

Fig. 19 Comparison of Baseline approach with N-gram PRF based QE in terms of F-Measure for TREC-3

datasets. Figure 20 presents the results for the CACM dataset. It shows that our baseline method is inferior to Khennak et al. [35] approach. However, combined techniques based QE outperforms Khennak et al. [35] approach as well as Sharma et al. approach for 44 queries. While for the rest six queries, their results are the same. This figure also depicts the superiority of our combined technique based QE over the Gupta et al. [25] approach. Figure 21 presents the comparison of performance using the CISI dataset. This figure shows that combined techniques based on QE get higher values of F-measure in comparison to approaches like Khennak et al. [35], Sharma et al. [53], and Gupta et al. [25] Similarly, Fig. 22 shows the outcomes for the TREC-3 dataset. This figure depicts that our combined approach outperforms Khennak et al. [35] proposition and Sharma et al. [53] proposition, but it lacks behind from Gupta et al. [25] approach. This method achieved 0.287 MAP value for the CACM dataset,0.2546 MAP Value for CISI, and 0.3027 for the TREC-3 dataset. This method improved 0.31% in the case of the CACM dataset,1.19% improved in the CISI dataset,1.71% MAP value reduced in the Trec-3 dataset in comparison to Gupta et al. [25]. Our proposed method works well on CACM and CISI because These datasets contain less number of documents, while TREC-3 contains a large number of documents, so it is not useful there.

We present a precision-recall curve to compare the overall performance of different approaches, as shown in Figs. 23, 24 and 25 for all the three datasets. Figure 23 shows the improvement in performance in QE using the combined effect of all techniques as compared to Baseline method approaches, Khennak et al. [35], Sharma et al. [53], and Gupta et al. [25] used over CACM dataset. Figure 24 presents the comparison for the CISI dataset and clearly shows

Table 7 Comparison of MAP values obtained by various approaches

Dataset	Baseline model	Khennak et al. query expansion proposal [35]	Sharma et al. query expansion method [53]	Gupta et al. query expansion approach [25]	Combined techniques based Query Expansion
CACM	0.2617	0.2803	0.2816	0.2818	**0.2827**
CISI	0.2261	0.2514	0.2534	0.2516	**0.2546**
TREC-3	0.2708	0.2893	0.2908	**0.3079**	0.3027

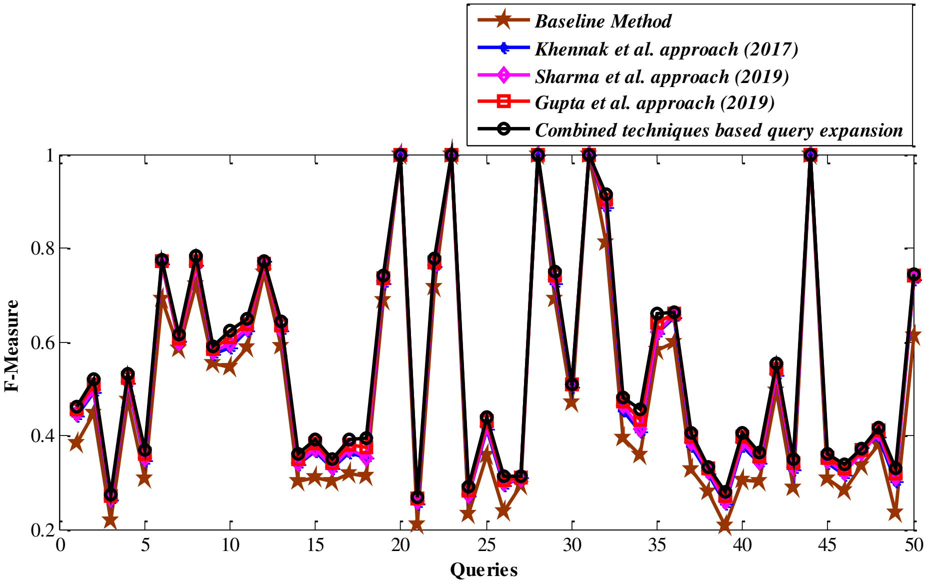

Fig. 20 Comparison of combined techniques based QE with baseline method, Khennak et al. [35], Sharma et al. [53] and Gupta et al. [25] as to F-Measure for CACM

the improvement in QE performance. Similarly, Fig. 25 shows the results, of using the combined effect of all techniques as compared to Baseline method over TREC-3 dataset and also depicts that our proposed approach enhances the overall performance of QE as compared to approaches Khennak et al. [35] and Sharma et al. [53], but it lags behind Gupta et al. [25] proposal.

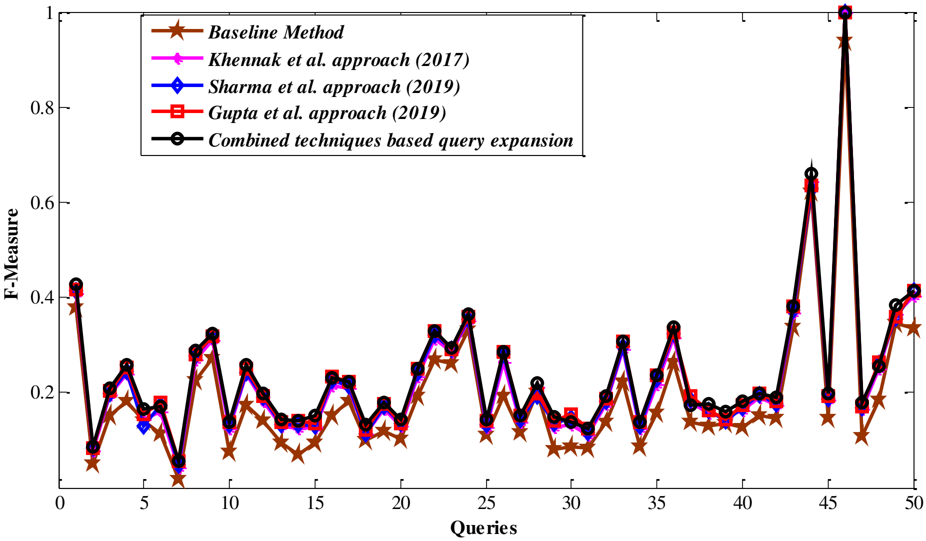

Fig. 21 Comparison of combined techniques based QE with baseline method, Khennak et al. approach [35], Sharma et al. [53] and Gupta et al. [25] as to F-Measure for CISI

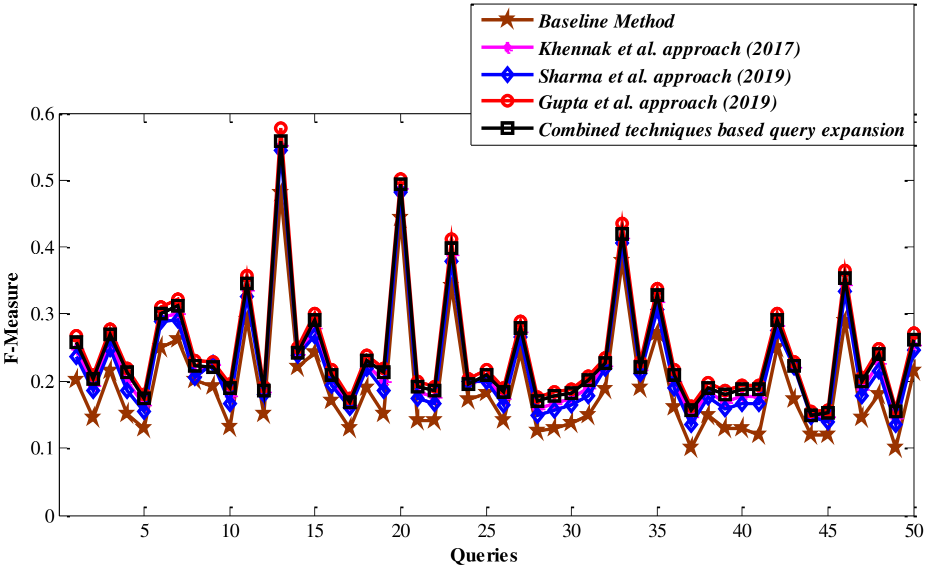

Fig. 22 Comparison of combined techniques based QE with baseline method, Khennak et al. approach [35]), Sharma et al. [53] and Gupta et al. [25] as to F-Measure for TREC-3

The above results show that our proposed combined techniques based QE approach improves the performance significantly for small datasets like CACM and CISI. However, for large datasets as TREC-3, the performance is not improved that much. The cause of enhancement in the performance of proposed combined techniques based on QE in comparison to other approaches is that this approach takes the advantages of each method. The abbreviation method decreases the performance; therefore, it is not included in the proposed combined techniques based approach.

For text categorization, There is a possibility to use combined approach. This method can be applied in image retrieval, Cross-Language Information Retrieval(CLIR), Multimedia

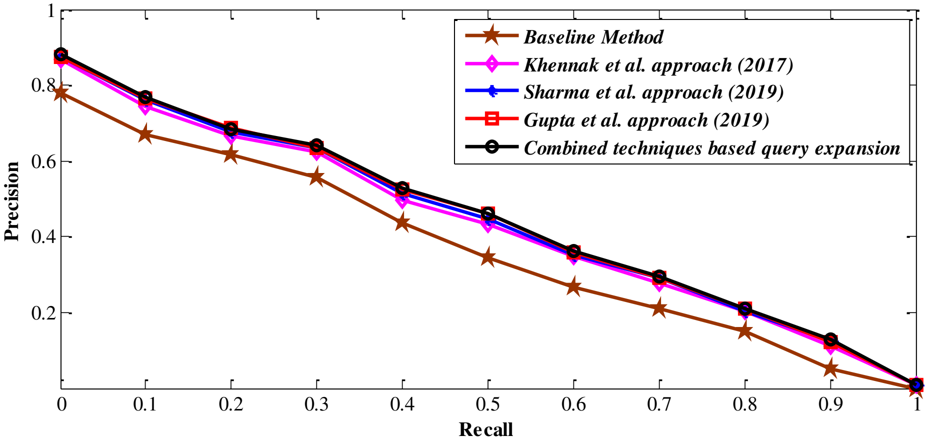

Fig. 23 Comparison of Precision-Recall curves for various proposals on CACM dataset

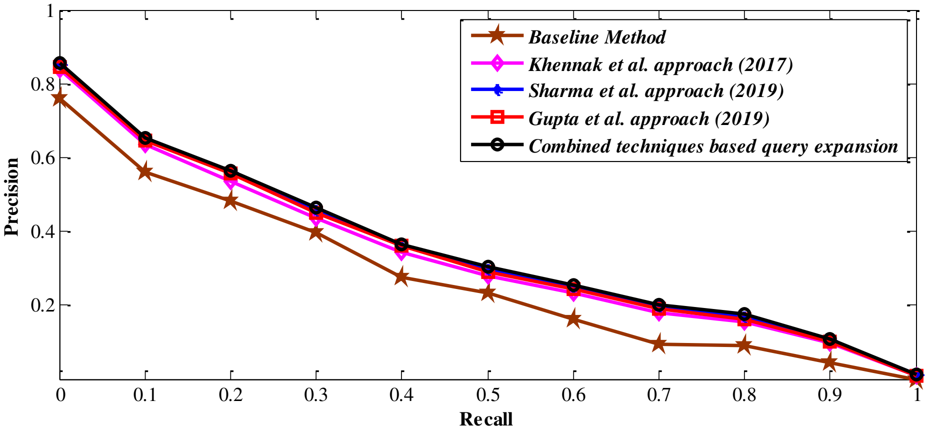

Fig. 24 Comparison of Precision-Recall curves for various proposals on CISI dataset

retrieval, Mobile Search, Information Filtering, and medical field also for query expansion. It can be further used in other applications such as blog searching, a recommendation system.

6 Conclusion

This paper is a step to explore various essential conditions for the optimal query using a QE technique. Our proposed technique uses knowledge sources and other IR parameters. AQE based method can be used in the neural network and bio and health field. Our combined approach can be utilized in the medical field. The proposed work has tested the effect of sources of term expansion, pseudo-relevance feedback, and ontologies on QE techniques.. This work also leads to a new QE framework, i.e., combined techniques based QE for improving IR performance. The developed combined model is evaluated and verified using three standard datasets CACM, CISI, and TREC-3. Comparing this combined model with freshly proposed approaches in the similar field like Khennak et al. [35], Sharma et al. [53] and Gupta et al. [25] QE approaches. The current work concludes that the proposed combined

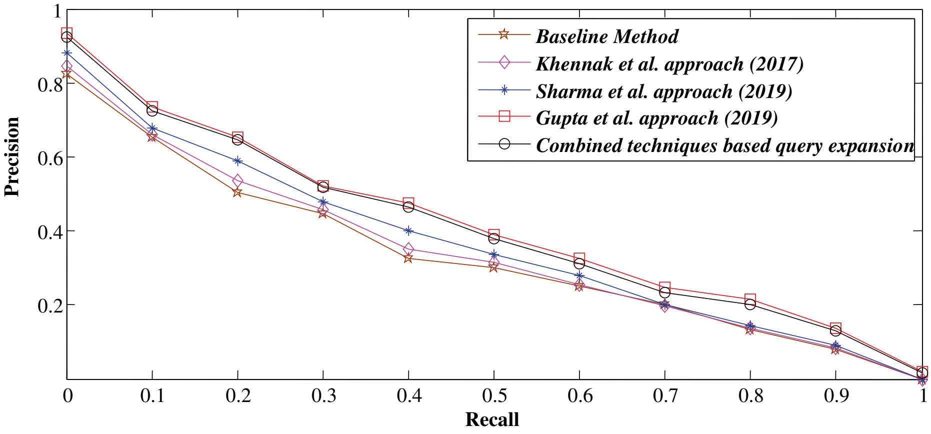

Fig. 25 Comparison of Precision-Recall curves for various proposals on TREC-3 dataset

techniques based QE approach enhances the performance of IR significantly for small datasets like CACM and CISI. However, for large datasets like TREC-3, the improvement is not significant. So, In the future, there is a need to consider other large datasets also. Our proposed method does not consider abbreviation for expanding the query, but this is also additional features to work in future. One mentionable point for the research work presented in this paper can be summarized in terms that it does not consider any such external environmental factors like sentiments, the mood of the user to retrieve documents. Hence, such external factors can be used as a basis or inclusions for future research works.

References

1. Anand R, Kotov A (2015) An empirical comparison of statistical term association graphs with dbpedia and conceptnet for query expansion. In: Proceedings of the 7th forum for information retrieval evaluation, pp 27–30
2. Azad HK, Deepak A (2019) Query expansion techniques for information retrieval: a survey. Inf Process Manag 56(5):1698–1735
3. Azad HK, Deepak A (2019) A new approach for query expansion using Wikipedia and WordNet. Inf Sci 492:147–163
4. Azad HK, Deepak A (2019) A novel model for query expansion using pseudo-relevant web knowledge. arXiv preprint arXiv:1908.10193
5. Bendersky M, Metzler D, Croft WB (2012) Effective query formulation with multiple information sources. In: Proceedings of the fifth ACM international conference on web search and data mining, pp 443–452
6. Bhogal J, MacFarlane A, Smith P (2007) A review of ontology based query expansion. Inf Process Manag 43(4):866–886
7. Bouchoucha A, He J, Nie JY (2013) Diversified query expansion using conceptnet. In: Proceedings of the 22nd ACM international conference on Information & Knowledge Management, pp 1861–1864
8. Bounhas I, Soudani N, Slimani Y (2019) Building a morpho-semantic knowledge graph for Arabic information retrieval. Inf Process Manag 102124
9. Carpineto C, Romano G (2012) A survey of automatic query expansion in information retrieval. ACM Computing Surveys (CSUR) 44(1):1–50
10. Carpineto C, De Mori R, Romano G, Bigi B (2001) An information-theoretic approach to automatic query expansion. ACM Transactions on Information Systems (TOIS) 19(1):1–27
11. Chandra G, Dwivedi SK (2019) Query expansion for effective retrieval results of hindi–english cross-lingual IR. Appl Artif Intell 33(7):567–593
12. Chang YC, Chen SM, Liau CJ (2007) A new query expansion method for document retrieval based on the inference of fuzzy rules. J Chin Inst Eng 30(3):511–515
13. Chaudhary C, Goyal P, Goyal N, Chen YPP (2020) Image retrieval for complex queries using knowledge embedding. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 16(1):1–23
14. Chen H, Furuse K, Yu JX, Ohbo N (2001) Support IR query refinement by partial keyword set. In: Proceedings of the second international conference on web information systems engineering, vol 1. IEEE, pp 245–253
15. Cooper JW, Byrd RJ (1998) OBIWAN-A visual interface for prompted query refinement. In: Proceedings of the thirty-first Hawaii international conference on system sciences, vol 2. IEEE, pp 277–285
16. Dahab MY, Alnofaie S, Kamel M (2018) A tutorial on information retrieval using query expansion. In: Intelligent natural language processing: trends and applications. Springer, Cham, pp 761–776
17. Dalton J, Naseri S, Dietz L, Allan J (2019) Local and global query expansion for hierarchical complex topics. In: European conference on information retrieval. Springer, Cham, pp 290–303
18. Di Marco A, Navigli R (2013) Clustering and diversifying web search results with graph-based word sense induction. Computational Linguistics 39(3):709–754
19. Esposito M, Damiano E, Minutolo A, De Pietro G, Fujita H (2020) Hybrid query expansion using lexical resources and word embeddings for sentence retrieval in question answering. Inf Sci 514:88–105
20. Fang F, Zhang BW, Yin XC (2018) Semantic sequential query expansion for biomedical article search. IEEE Access 6:45448–45457
21. Fattahi R, Wilson CS, Cole F (2008) An alternative approach to natural language query expansion in search engines: text analysis of non-topical terms in web documents. Inf Process Manag 44(4):1503–1516

22. Gong Z, Cheang CW (2006) Multi-term web query expansion using WordNet. In: International conference on database and expert systems applications. Springer, Berlin, pp 379–388
23. Gupta Y, Saini A (2017) A novel fuzzy-PSO term weighting automatic query expansion approach using combined semantic filtering. Knowl-Based Syst 136:97–120
24. Gupta Y, Saini A (2019) A new swarm-based efficient data clustering approach using KHM and fuzzy logic. Soft Comput 23(1):145–162
25. Gupta Y, Saini A (2019) A novel term selection based automatic query expansion approach using PRF and semantic filtering. International Journal of Engineering and Advanced Technology 8(C):130–137
26. Gupta Y, Saini A, Saxena A (2013) A review on important aspects of information retrieval. International Journal of Computer, Information science and Engineering 7(12):940–948
27. Gupta Y, Saini A, Saxena A (2014) Fuzzy logic based approach to develop hybrid similarity measure for efficient information retrieval. J Inf Sci 40:846–857
28. Gupta Y, Saini A, Saxena AK (2015) A new fuzzy logic based ranking function for efficient information retrieval system. Expert Syst Appl 42(3):1223–1234
29. Horng JT, Yeh CC (2000) Applying genetic algorithms to query optimization in document retrieval. Inf Process Manag 36(5):737–759
30. Hsu MH, Tsai MF, Chen HH (2008) Combining WordNet and ConceptNet for automatic query expansion: a learning approach. In: Asia information retrieval symposium. Springer, Berlin, pp 213–224
31. Htun NN, Halvey M, Baillie L (2018) Beyond traditional collaborative search: understanding the effect of awareness on multi-level collaborative information retrieval. Inf Process Manag 54(1):60–87
32. Huang Q, Yang Y, Zhan X, Wan H, Vakis G (2018) Query expansion based on statistical learning from code changes. Software: Practice and Experience 48(7):1333–1351
33. Keyword (2020) Query size by country. https://www.keyworddiscovery.com/keyword-stats.html
34. Khan L, Luo F (2002) Ontology construction for information selection. In: 14th IEEE international conference on tools with artificial intelligence, 2002. (ICTAI 2002). Proceedings. IEEE, pp 122–127
35. Khennak I, Drias H (2017) An accelerated PSO for query expansion in web information retrieval: application to medical dataset. Appl Intell 47(3):793–808
36. Khennak I, Drias H (2020) A novel hybrid correlation measure for query expansion-based information retrieval. In: Critical approaches to information retrieval research. IGI Global, pp 1–19
37. Kotov A, Zhai C (2012) Tapping into knowledge base for concept feedback: leveraging conceptnet to improve search results for difficult queries. In: Proceedings of the fifth ACM international conference on web search and data mining, pp 403–412
38. Krovetz R, Croft WB (1992) Lexical ambiguity and information retrieval. ACM Transactions on Information Systems (TOIS) 10(2):115–141
39. Kumar R, Bhanodai G, Pamula R (2019) Book search using social information, user profiles and query expansion with Pseudo relevance feedback. Appl Intell 49(6):2178–2200
40. Lafourcade M, Zarrouk M, Joubert A (2014) About inferences in a crowdsourced lexical-semantic network. In: Proceedings of the 14th conference of the European chapter of the Association for Computational Linguistics, pp 174–182
41. Latiri C, Haddad H, Hamrouni T (2012) Towards an effective automatic query expansion process using an association rule mining approach. J Intell Inf Syst 39(1):209–247
42. Li H, Xu J (2014) Semantic matching in the search. Foundations and Trends® in Information Retrieval 7(5): 343–469
43. Macdonald C, Ounis I (2007) Expertise drift and query expansion in expert search. In: proceedings of the sixteenth ACM conference on conference on information and knowledge management, pp 341–350
44. Mahler D (2004) Holistic query expansion using graphical models. New Directions in Question Answering 2004:203–227
45. Nasir JA, Varlamis I, Ishfaq S (2019) A knowledge-based semantic framework for query expansion. Inf Process Manag 56(5):1605–1617
46. Nowacka K, Zadrozny S, Kacprzyk J (2008) A new fuzzy logic based information retrieval model. In: 12th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2008), Malaga, Spain
47. Pérez F, Font J, Arcega L, Cetina C (2019) Collaborative feature location in models through automatic query expansion. Autom Softw Eng 26(1):161–202
48. Raza MA, Mokhtar R, Ahmad N (2018) A survey of statistical approaches for query expansion. Knowl Inf Syst:1–25
49. Robertson SE, Walker S, Jones S, Hancock-Beaulieu MM, Gatford M (1995) Okapi at TREC-3. Nist Special Publication Sp, 109, 109
50. Robertson SE, Walker S, Beaulieu M, Willett P (1999) Okapi at TREC-7: automatic ad hoc, filtering, VLC and interactive track. Nist Special Publication SP, (500), 253–264

51. Schwartz AS, Hearst MA (2002) A simple algorithm for identifying abbreviation definitions in biomedical text. In: Biocomputing 2003, pp 451–462
52. Sharma DK, Pamula R, Chauhan DS (2018) A comparative analysis of fuzzy logic based query expansion approaches for document retrieval. In: International conference on advances in computing and data sciences. Springer, Singapore, pp 336–345
53. Sharma DK, Pamula R, Chauhan DS (2019) A hybrid evolutionary algorithm based automatic query expansion for enhancing document retrieval system. Journal of ambient intelligence and humanized computing:1–20
54. Sharma DK, Pamula R, Chauhan DS (2019) Soft computing techniques based automatic query expansion approach for improving document retrieval. In: 2019 Amity International conference on artificial intelligence (AICAI). IEEE, pp 972–976
55. Sharma DK, Pamula R, Chauhan DS (2019) Combined techniques based query expansion approach for document retrieval system. In: 2019 international conference on contemporary computing and informatics (IC3I). IEEE, pp 101–105
56. Singh J, Kumar R (2017) Lexical co-occurrence and contextual window-based approach with semantic similarity for query expansion. International Journal of Intelligent Information Technologies (IJIIT) 13(3): 57–78
57. Singh J, Sharan A (2017) A new fuzzy logic-based query expansion model for efficient information retrieval using relevance feedback approach. Neural Comput & Applic 28(9):2557–2580
58. Singh J, Sharan A (2018) Rank fusion and semantic genetic notion based automatic query expansion model. Swarm and Evolutionary Computation 38:295–308
59. Singh J, Sharan A, Saini M (2017) Term co-occurrence and context window-based combined approach for query expansion with the semantic notion of terms. International Journal of Web Science 3(1):32–57
60. Spink A, Wolfram D, Jansen MB, Saracevic T (2001) Searching the web: the public and their queries. J Am Soc Inf Sci Technol 52(3):226–234
61. Stokes N, Li Y, Cavedon L, Huang E, Rong J, Zobel J (2007) Entity-based relevance feedback for genomic list answer retrieval. In: TREC
62. Stokes N, Li Y, Cavedon L, Zobel J (2009) Exploring criteria for successful query expansion in the genomic domain. Inf Retr 12(1):17–50
63. Torjmen-Khemakhem M, Gasmi K (2019) Document/query expansion based on selecting significant concepts for context based retrieval of medical images. J Biomed Inform 95:103210
64. Wang Z, Qiang N (2012) Research on hybrid query expansion algorithm. International Journal of Hybrid Information Technology 5(2):207–212
65. Wang Y, Huang H, Feng C (2019) Query expansion with local conceptual word embeddings in microblog retrieval. IEEE Trans Knowl Data Eng:1
66. Wasim M, Asim MN, Ghani MU, Rehman ZU, Rho S, Mehmood I (2019) Lexical paraphrasing and pseudo relevance feedback for biomedical document retrieval. Multimed Tools Appl 78(21):29681–29712
67. Wu Y, Li Y, Xu Y (2019) Dual pattern-enhanced representations model for query-focused multi-document summarization. Knowl-Based Syst 163:736–748
68. Zhang C, Qin Z, Yan X (2005) Association-based segmentation for Chinese-crossed query expansion. IEEE Intelligent Informatics Bulletin 5(1):18–25
69. Zhou W, Clement TY, Torvik VI, Smalheiser NR (2006) A concept-based framework for passage retrieval at genomics. In: TREC vol 8, no 2, pp 8–2
70. Zingla MA, Latiri C, Mulhem P, Berrut C, Slimani Y (2018) Hybrid query expansion model for text and microblog information retrieval. Information Retrieval Journal 21(4):337–367

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.