ABSTRACT: The aim of this paper is to improve the search query performance of the biomedical literature by expanding the queries with most significant terms. Methods: In this article, an enhanced BM25 mathematical approach is proposed to retrieve the most query relevant literature from clustered Biomedical literature bank with query expansion from MeSH. The clustered biomedical topics are analyzed with different pre-processing methods and term-weighting functions and found the best values for the tuning parameters $K_1$, $b$, $K_3$ and the right combination of pre-processing and term weighting functions for improving the query performance in terms of Average Precision, Mean Average Precision and R-precision. Novelty/Improvements: In this approach, the existing best match retrieval technique is normalized with the calibrate constants $K_1 = 1.3$, $b = 0.75$, $K_3 = 1.2$ and the significant terms are identified with comparison of MeSH for query expansion. The retrieval performance is improved in terms of Mean Average Precision and R-precision.

1. Introduction

Due to the enormous innovations in Biomedicine, the size of Biomedical Literature is increasing at a faster rate, and the task of literature retrieval is becoming complex. Information retrieval is related to the representation of documents as term vectors with frequency and modeling of the terms as topics, where topics are used for representing articles and to access the literature [1]. Term vectors are helpful to the researchers to have informal access to specific literature, and researchers should present a query to access the articles. The query is processed by Information retrieval applications. The terms in a query are taken as index topics to be searched in the database and summarize the user-interested specific information and rank the documents according to their relevance.

The key problem with Information retrieval system is to search with correct query terms which function as the accurate descriptors of the underlying content of a document. In other words, information regarding a specific concept can be represented with different terms which would retrieve the documents not related to the query concept, resulting false retrieval. False retrieval is characterized by many factors of which the major cause is the use of an identical term for different concepts. The actual concept behind the articles may be different from the way it is queried and accessed.

The user query may represent some part of the document and other relevant terms of the document which describe the overall topic of the document may not appear in query. To overcome this problem the user query should be
refined with high weighting index terms or with the accurate descriptors of the document to improve the retrieval efficiency and to obtain relevant articles. Enrichment of query with relevant document terms is called as Query expansion by where the initial user query is expanded with significant terms and reweighting the added terms will improve query performance. Query expansion techniques are often used for impressing the efficiency of document retrieval process, to overcome word ambiguity issues which have the same meaning with different context. This research is to work on noisy words removal in query and document in order to include the overall content of document and representing the documents with significant term vectors for improving the relevancy parameters in document retrieval with the given query and query expansion. Figure 1 shows the Document Retrieval System with Enhanced BM25 approach which we proposed to use it and the query expansion.

This paper is organized, as chapter 2 describes the Document Retrieval Process, i.e., the Preprocessing Techniques, and chapter 3 discusses about Enhanced BM25 Retrieval Techniques and various term weighting approaches and chapter, 4 discusses the query expansion techniques and relevance feedback from MeSH. Chapter 5 discusses the approaches and experiment results of query expansion.

Query expansion is one of the crucial factors in literature retrieval system. Many different methods are proposed by researchers to perform Query Expansion. Few approaches concentrated on expansion of terms using topic models and other approaches mostly on ontologies (structured data).

Jitendra Jonnagaddala et al [37] proposed the CRF classifier for disease named entity recognition and normalization. In this approach for query expansion they used dictionary Look Up approach on Bio creative V CDR Track data set. Yanshan Wang et al [38] discussed about leveraging the unstructured text, state of art information retrieval systems, Medical Named Entity Recognition and

![Figure 1. Literature Retrieval System with Query Expansion Techniques](image-url)
Query expansion with deep learning based word Embeddings and re-ranking strategy to enhance the retrieval performance. Jose Perez-aguera et al [2] given the comparison of different query expansion techniques on unstructured text documents and given an analysis of the similar terms identification in different documents using \textit{tanimoto}, dice and cosine similarity measures. They used the \textit{Kullback- Liebler Divergence} to compute the divergence between to probability distributions of terms in the whole collection and in the top ranked documents and obtained for a first pass retrieval using the original query user. The most likely terms to expand the query are those with a high probability in the top ranked set and low probability in the whole collection.

Yuanhua Lv et al [3] proposed a positional relevance model (PRM) to address this problem in a unified probabilistic way. The proposed PRM is an extension of the relevance model to exploit term positions and proximity to assign more weights to words closer to query words based on the intuition that words closer to query words are more likely to be related to the query topic. Experimental results on two large retrieval datasets show that the proposed PRM is effective and robust for pseudo-relevance feedback, significantly outperforming the relevance model in both document-based feedback and passage-based feedback. Alipanah et al [4] proposed a novel weighting mechanism called basic expansion techniques (BET) and new expansion terms for structured documents. They considered each individual ontology and user query keywords to determine the Basic Expansion Terms (BET) using many semantic measures including Betweenness Measure (BM) and Semantic Similarity Measure. Rivas et al [5] have proposed preprocessing techniques for query expansion and retrieving documents in several fields of biomedical documents belonging to the corpus of Medline and disease, the cystic fibrosis. They conducted experiments showing the different results and benefit of using stemming and stop words in the pre-processing of documents and queries. Their studies and experiments were conducted to compare the weighting algorithms Okapi BM25 and TF-IDF available in the Lemur tool, concluding that the TF-IDF with TF formula given by BM25 approximation provided superior results. Hongfei Lin et al [40] presented a new technique to refine information retrieval searches to better represent the user’s information need to enhance the performance of information retrieval by using different query expansion techniques and apply a linear combination between them, where the combinations were linearly between two expansion results at one time.

This experiment shared the different results and advantages of using stop words and stemming methods in pre-processing stage of documents and queries. They are compared with Enhance BM25 approach and the top retrieved documents are used for relevance feedback in query expansion.

In this paper, we have proposed an enhancement for BM25 algorithm and discuss the diseases brain tumor. The data bases are considered from NCBI (National Center for Biotechnological Information) [39].

2. Pre-Processing of Clustered Biomedical Topics

The National Center for Biotechnology Information in India has the largest collection of biomedical articles. Every year they refine their literature search tools to manage the voluminous data. Over 30 million documents are available in their web dataset and PUBMED is the tool used to retrieve biomedical data from NCBI. To improve the efficiency of biomedical literature access and retrieval procedure, preprocessing of literature is one of the vital techniques. Document retrieval process is the key for article extraction. Documents are retrieved from PUBMED using a query, in a collection of indexed terms which describes the underlying content of documents. Query engine processes the query and optimizes it to get the best matching’s from the data base.

Preprocessing step in the figure 1 contains two sub-processes, viz., stop word removal and stemming process. Preprocessing is a procedure to obtain index of terms to represent the document, In the process of arriving at the index of terms it is significant to stem same meaning words to their root. It will be done by the stemming. The words such as ‘it’, ‘so’, ‘what’, ‘this’, ‘that’ etc., are the repeated words called as stop words, which need to be removed for obtaining the specific relevant topics of the documents.

In [7], the biomedical literature is retrieved from PUBMED using a search query and by using the pre-processing approach, the retrieved query-relevant articles from PUBMED are cleaned. Retrieval techniques are applied on cleaned documents for relevant article retrieval for the base query, and clustering techniques are applied on relevant articles for the relevant topic grouping. Using the \textit{Multi-Kernel Fuzzy C means} (MK-FCM) technique relevant articles are clustered under relevant topic of the query. In this approach different kernel functions such as \textit{Linear}, \textit{Quadratic} and composite kernel function are analyzed for relevance measuring metrics such as \textit{Precision}, \textit{Recall}, and \textit{F-Measure}. The query used to retrieve the articles from PUBMED databases are on the topics such as ‘brain tumor’, ‘breast cancer’, ‘kidney stone’ and ‘neovascularization’. This work deals with different kernel functions analysis in FCM, which shows that the proposed composite function (hybrid kernel) of FCM out-performed the others. Eleven performance measures are taken in this process for analyzing four topics relevant articles. The proposed algorithm was implemented on MATLAB-16a. [7].

In this approach, we have used the clustered articles under the topic brain tumor [7] for our analysis and on that cluster different pre-processing techniques, term weighting functions are analyzed to find the best combination to be used for the improved retrieval performance. Initial query
results on this clustered document are compared with MeSH indexed terms and relevance feedback is taken to find the most significant terms which are the accurate descriptors of the underlying content of document. After finding the most relevant topics for the documents, the initial queries are refined till the retrieval accuracy is increased.

When the first answer sets are obtained, the further procedure is refining the query to get most relevant documents. The refinement of query from top retrieved documents terms is called Query Expansion. The relevant keywords retrieved are the accurate descriptors of documents in first step and can be added to the query to re-rank the documents. This procedure is called ‘Relevance Feed Back’. The retrieval process is further enhanced by modifying the terms in the query [5] respective to the relevant documents.

This study used Bioinformatics tool box in MATLAB for analyzing the pre-processing methods, different term weighting schemes and comparison methods and query expansion of search query to retrieve relevant articles and to increase retrieval accuracy.

Most of the public biomedical documents belong to PUBMED and MEDLINE of NLM (National Library for Medicine). In this paper, we have analyzed the ‘Disease Brain Tumor data base’ from PubMed (www.ncbi.nlm.nih.gov/PubMed) [5]. It consists of 192584 documents for Brain tumor published between from 1964 to 2017. All these documents are composed of Title (TI), Abstract (AB), and related articles manually assigned to MeSH (Medical Sub Headings) www.ncbi.nlm.nih.gov/MeSH/[5] of MeSH thesaurus. Table 1 shows the sample document.

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TI</td>
<td>The Successive projection algorithm as an Initialization method for brain tumor segmentation using non-negative matrix factorization</td>
</tr>
<tr>
<td>MJ</td>
<td>Brain Tumor: fg Successive Projection: fg</td>
</tr>
<tr>
<td>MN</td>
<td>SPA, NMF, Brain Tumor, Segmentation, MRI</td>
</tr>
<tr>
<td>AB</td>
<td>Non Negative Matrix Factorization NMF</td>
</tr>
</tbody>
</table>

Table 1. A Sample PubMed Document [5]

MeSH thesaurus is a controlled Vocabulary used for indexing, cataloging, identifying and searching biomedical database. MeSH thesaurus contains approximately -26 plus million terms and is updated time-to-time to reflect changes in medical terminology [5]. MeSH has a hierarchical structure with set of terms and descriptors [5]; naming that allows various levels of searching. It will allow retrieving the document where the same concept is explained with different terminology. On an average 5 to 20 subject headings are assigned by document at which 2 to 4 are the major (MJ) subjects and other are minor (MN) subjects [5]. Major MeSH terms describe the main contents of the document and minor MeSH terms provide more details. MeSH headings are the key terms that are in documents and act as descriptors for the document, and they are used to search for the key terms in the retrieval process.

| QU How Brain Tumor Leads to Brain Hemorrhage |
| RD 228 624 826 887 780 799 697 626 540 496 442 352 339 334 294 554 236 |
| 211 191 162 122 119 131 91 70 52 40 58 37 16 24 18 20 13 12 9 10 |
| 12 8 36 1 2 111 |

Table 2. A sample of query with its relevant document and relevant scores (Query) from 1964 to 2017 [5]

The brain Tumor collection also contains query corpus with more than 100 queries and the documents relevant to each query [1]. The Table 2 shows the retrieved relevant documents from 1964 to 2017.

2.1 Stop Words Removal
In information retrieval process (IRP), a document is indexed by significant words. Statistical analysis at this procedure shows that few words have high frequency, and
more have low frequency [10]. For example, the words ‘or’, ‘and’, ‘not’‘the’ and ‘if’ appears more number of times in the documents without much significant content. These words often referred in information retrieval as stop words. The non-significant terms need to be removed from the indexing process and this procedure also reduces the size of indexing and boosts up the efficiency and accuracy of retrieval performance. The NLM stop word list [http://www.netautopsy.org/umlsstop.html] and the other stop word lists are prepared by GceraIsdston and Chris Buckley for experimentation. For SMART stop words refer [http://www.lextek.com/manuals/onix/stopwords2.html][5].

2.2 Stemming Process
Stemming is the process of adding the inflectional alternatives of the Si to its stem, or root. In document retrieval process, we have reviewed three stemmers namely HUSK stemmer, Krovert Stemmer and Porter Stemmer.

Porter stemmer is a procedure for reducing suffixes from words such as plurals and gerunds and replacing inflectional ends [8]. It is composed of few rules, each of which deals with a specific suffix and has certain conditions to satisfy. The suffixes of words are checked with each rule sequentially until it matches one; the conditions in the rule are then tested, which will result in a suffix removal or modification [5].

Alternatively, Paice/Husk stemmer proposed by Paice [9] a group office given as input to the programs. These files contain list of words(terms), alphabetically stored and any terms that are considered by the evaluator to be semantically equivalent are formed in the concept group [8].

3. Term Weighting Functions and Similarity Score

3.1 Enhanced OKAPI BM25 Weighting Function(E-BM25)
Best Matching is used to find the frequency and importance of words in corpus to retrieve the documents which are relevant. Many Information Retrieval processes use BM25 Weighting Function on Textual Databases to retrieve relevant articles. It is a probabilistic approach, the weight of key terms is calculated based on number of times it is present in a document and the inverse Document Frequency of the term. BM25 Weighting Function is defined as [11][12]

\[
 w_{qi} = \frac{n(r+0.5)(r-R+0.5)}{(n-r+0.5)(N-n-R+r+0.5)} \times \frac{(k_3+1)f_{qi}}{f_{qi} + k_1(b_0/(avgl)+0.5)} \times \frac{(k_2+1)f_{qi}}{f_{qi} + k_2} + \delta
\]

Where \(K1, b, and k3\) are calibrate constants for document and Query Length and Weight, \(f_{qi}\) is frequency of term \(i\) in the document \(d\). \(f_{qi}\) is frequency of the term \(i\) in query \(Q\). \(ds\) and \(ads\) are respectively the document length and the average document length in the database. \(\delta\) is an additional parameter with default value of 1.0 in the absence of training Data. Experimented this parameter with a value of \(\delta = 0.5\) to 1.0.

Where \(r\) represents relevant articles that have term \(x\) and \(R\) represents the relevant documents to significant topic. \(N\) defines number of documents of the corpus and then represents number of articles containing the term \(x\).

If a query term, \(q\) appears in \(n\) document, it is \(n(q)\). If any documents \(D\) picked randomly from \(n(q)\) with a probability \(n(q)/N\).

The information content at the related query i.e., \(D\) has \(q\) terms is:

\[
 \log \frac{n(q)}{N} = \log \frac{N}{n(q)}
\]

If multiple terms are there like \(q1, q2,\ldots, p\) terms in a query and the information content for the query is

\[
 \sum_{i=1}^{p} \log \frac{N}{n(q_i)}
\]

For best matching, the existing BM25 has taken values as \(k1 = 1.2, b = 0.75,\) and \(k3 = 7\). In enhanced BM25 (E-BM25) for best matching, i) changed the constants to the lower value than the existing values. (i.e., from 1 to 0.5 or 0.75 only) ii) changed the \(k1, b, k3\) values in such a way that \(k1\) varies with respect to \(k3\) and keeping \(b\) as constant. iii) \(b\) varies with respect to \(k1\) and \(k3\) by keeping \(k1\) & \(k3\) as same values.

The Similarity Score between query terms and documents are calculated as [41]

\[
 \text{score}(D,Q) = \sum_{i=1}^{n} \frac{\text{IDF}(q_i) \times (f_{qi} \times k_1 + 1)}{f_{qi} + k_1 \times \left[1-b+b \times \left(\frac{dl}{avgl}\right)\right] + \delta}
\]

\(\delta\) - Additional Free parameter
\(\delta = 1.0\) (default value)

\[
 \text{IDF}(q_i) = \frac{\log N - n(q_i) + 0.5}{n(q_i) + 0.5}
\]

\(K1, b, k3\) values in equations varies in such a way, \(K1 \in [0.5, 2.0], b \in [0.75], K3 \in [1.2]\)

3.2 TF-IDF Weighting Function
TF-IDF is frequently used in information retrieval and text mining [5]. Term frequency measure is used to find the frequency occurrence of term in a document and inverse
document frequency is used to measure the absence term in the number of documents. The scalar product of this will give the importance of the term in the document collection. TF-IDF weighting schemas are often used by search engine as a tool in scoring and ranking the document.

\[
tf_{xd} = \frac{f_{xd}}{\text{max} \times f_{xd}}
\]  

(6)

relevance for a given query [5] [13, 14].

Term frequency in document collection is calculated as follows

\[
tf_{sq} = f_{sq}
\]  

(7)

Where \( \text{Max} \times f_{xd} \) is the frequency of most common term in document \( d \).

Term Frequency with logarithmic approach is defined as Log TF formulae (Equation8). Logarithmic approach is used to normalize the frequency of the terms so that finding the most relevant term is easier, because sometimes the term which high most may not be relevant to the user query.

\[
tf_{xd} = \log (\text{raw} \times TF + 1) \quad \text{and} \quad tf_{sq} = \log (\text{raw} \times TF + 1)
\]  

(8)

Table 3. Parameters of the BM25 weighting and okapi TF[5]

<table>
<thead>
<tr>
<th>Term Frequency in Document D</th>
<th>E-BM25</th>
<th>Okapi TF/BM 125</th>
</tr>
</thead>
<tbody>
<tr>
<td>K1 for Document length normalization</td>
<td>K1 for Document length normalization</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Term Frequency in Document and Query</th>
<th>b for Term Weight Normalization</th>
<th>b for Term Weight Normalization</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Query Term Frequency</th>
<th>k3 for Query Term Normalization</th>
<th>K1 for Query Term Normalization</th>
</tr>
</thead>
</table>

The inverse document frequency is calculated as follows

\[
idf = \log \left( \frac{n}{N + 1} \right)
\]  

(11)

The importance of the term is known from its term frequency and inverse document frequency as follows and in this we have proposed to include the search query term weight as also for scalar product [5]

\[
w_{xi} = tf_{xd} \times tf_{sq} \times idf^2
\]  

(12)

4. Query Enrichment

Query enrichment is expanding the query with most weighted and important terms for query retrieval process, since three decades the query expansion techniques are in use. The query expansion methods include expanding words which can be identified from additional terms of relevant documents which are retrieved by an explicit query. This query expansion is based on pseudo-relevance feedback system. The expanding terms for the query enhancement will be taken from user recognized relevant document terms. This procedure requires user feedback on relevant documents. [5], [16, 17].

The expanding terms may be taken from corpus and it can be corpus-specific. Query enhancement terms are taken from the concept analysis of corpus, from thesaurus such as MeSH and group words together based on the occurrences of terms at a document level [18, 19] and can be considered for expansion. MeSH is used for Automatic Query Expansion [5].

The expanding terms are taken form language models. Terms are found from general online thesauri that are not tailored for any specific text collection [5]. Liddy and Myaeng [20] used the Long Man’s Dictionary of Contemporary English; it is a semantic coded dictionary. Voorhees [21] used Word Net, a manually constructed lexical related network. Borrajo et al. [22] studied the use of dictionaries in the classification of biomedical texts with the [5] three dictionaries (Bio creative [23], NLPBA [24] and Uni Prodata base [5], [25]).
In this work, we have implemented query expansion based on corpus specific term approach for finding related terms. This thesaurus search approach commonly used to add important words to the query. Previous approaches provide relevance feedback from user. In this case, unsighted feedback is used to expand query. This procedure takes the pseudo relevance feedback from the MeSH and the relevance feedback from user is not compulsory. In our approach given a query, many documents are retrieved; considering all documents as relevant without user feedback [26]. The retrieved documents are assessed and are used towards altering weights of terms present in user query and expanding the query with most weighted terms. The new query is reformulated towards relevant documents and away from the non-relevant ones [5], [27, 28].

Elastic search is used in the implementation for pre-processing and Bio toolkit is used for retrieval and expansion. First it applies the standard retrieval model [5] to retrieve \( d_1 \ldots d_m \) for a given query \( q \) [5].

From the collection of retrieved documents, the most weighted terms are assessed with document descriptor collection and are used as following for the query expansion [5]

\[
q' = q + \frac{\alpha}{M} \sum_{i=1}^{M} d_i + \frac{\beta}{T} \sum_{j=1}^{T} \sum_{i=1}^{c} C_{ij} \tag{13}
\]

Where \( \alpha, \beta \) are weight of the relevant terms of the query to be added. \( M \) indicates the number of terms in collection. \( T \) represents the number of topics in the cluster \( j \).

\( d_i \) is document collection with \( M \) terms

\( C_{ij} \) is cluster \( j \) with \( T \) topics

\( q \) is given search query

### 4.1 Relevance Measures and Corpus Used

In this experimentation, the MATLAB is used with the dataset, MEDLINE Brain Tumor Corpus. A search query “Brain Tumor” and the publication date was given from 2011 to till date, then we found 41009 articles from the MEDLINE database with the following URL: https://www.ncbi.nlm.nih.gov/sites/entrez?cmd=search. We also searched for 2007 to 2016 publications on “Brain Tumor” query search and we found 73807 articles. These articles extracted and structured with the terms “PubMed ID”, “Publication Date”, “Title”, “abstract”, “Authors”, “Citation”. On the extracted data, we performed the Fuzzy C Means Clustering and generated 10 clusters of diverse topics of brain tumor search query. Each cluster of 1000 articles are considered for our experimentation.

In this study, we have analyzed various pre-processing techniques on the clustered topics and considered 1000 document set from this cluster. Various combinations of filtering techniques and term weighting functions are analyzed, and the result set is measured with Average Precision. Precision is the ratio of articles retrieved to its relevant where as average precision is the mean of retrieved articles at specific rank for different queries. The methodology is measured with Mean Average Precision where it is the sum of average precision with respect to the retrieved documents on different queries.

**R-Precision** is calculated on the result set retrieved at a specific rank and is divided with the total number of relevant articles in the collection. Overall the 1000 document set corpus is analyzed with various techniques mentioned in the paper and measures like average precision, Mean Average Precision, R-Precision are calculated for different queries.

The \( q \) in the above Equation 9 and the \( q' \) in each iteration of results with the query expansion includes different queries.

The queries considered after every query expansion includes such as “How Brain Tumor leads to Brain Hemorrhage”, “How Malignant Brain Tumors leads to Brain Hemorrhage”, “How loss of Function mutation of gene VPS13A leads to neuro degenerative disorder”, “How Glioblastoma Multiforme (GBM) comprises the most common and aggressive intracranial tumor”, “The role of CXCR3LRP1 cross-talk in the invasion of primary brain tumors,” “what are the Diverse Functions of Chorein”, “how AJAP1 expression modulation to glioma cell motility increase the tumor growth”, “how noncoding micro RNAs Leads to DNA damage and Tumor Growth” etc. These are some of the expanded queries in our experimentation after each query performed on clustered 1000 documents related to “Brain Tumor” Topic. The top 50 retrieved articles are considered for the average precision and R-precision calculation.

### 4.2 Average Precision

Average Precision averages the precision of the retrieved documents of each query at a specific rank. The Average Precision is calculated as follows [5].

\[
AvP = \frac{\sum_{r=1}^{Rd} \left( \frac{P \times relevace(r)}{D_r} \right)}{R}
\tag{14}
\]

\( P \) represents the precision at the given rank, \( r \) defines the rank; \( R \) represents the number of documents retrieved, and relevance(\( r \)) represents a 0 or 1;

The below table 4 illustrates the use of different pre-processing techniques in combination with different term weighting functions and the average precision measure for each combination techniques at Rank 50 of the retrieved documents based on the query “How Brain Tumor Leads to Brain Hemorrhage”. The table also shows that the result of Average Precision on clustered “Brain Tumor” Topics using Husk Stemma with stop word removal combination with E-best matching 25 approach is high compared to all other methods.

The average precision for top 50 retrieved documents based on the base query “How brain Tumor Leads to Brain
Combinations & TF-IDF & BM25 & E-BM25 & Log TF & RAW TF & D
---
Porter stemmer PUBMED Stopwords & 0.213 & 0.231 & **0.254** & 0.125 & 0.112 & 1000

HUSK Stemmer PUBMED Stopwords & 0.215 & 0.242 & **0.256** & 0.132 & 0.121 & 1000

Porter stemmer SMART Stopwords & 0.224 & 0.241 & **0.253** & 0.142 & 0.132 & 1000

HUSK Stemmer SMART Stopwords & 0.231 & 0.254 & **0.268** & 0.154 & 0.142 & 1000

Table 4. Average Precision@50 for the query “How Brain Tumor Leads to Brain Hemorrhage” on 1000 Clustered Brain Tumor Topic Document Set

4.3 R-Precision
R-Precision is defined as $R-Prec = \frac{r}{R}$, which is the ratio between all the relevant documents retrieved until the rank that contemporaries the number of relevant documents that are in collection in total ($r$), to the total number of relevant documents in the collection $R$ [5].

Combinations & TF-IDF & BM25 & E-BM25 & Log TF & RAW TF & D
---
Porter stemmer PUBMED Stopwords & 0.201 & 0.215 & **0.245** & 0.143 & 0.112 & 1000

Porter stemmer SMART Stopwords & 0.212 & 0.225 & **0.264** & 0.132 & 0.152 & 1000

HUSK Stemmer PUBMED Stopwords & 0.221 & 0.250 & **0.267** & 0.156 & 0.143 & 1000

HUSK Stemmer SMART Stopwords & 0.227 & 0.231 & **0.278** & 0.167 & 0.121 & 1000

Table 5. R-Precision@50 for brain tumor Query on Top-50 Retrieved Documents

4.4 Mean Average Precision(MAP)
The Mean Average Precision is calculated from average precision and it summarizes the precision of all the queries [5]

$Map = \frac{\sum_{q=1}^{Q} AvP_q}{Q}$

(15)

In Table 9, we can see that the best MAP results are obtained using E-BM25 Approach, so we continue our study with this approximation. (brain tumor)

5. Approaches and Results
In our proposed approach, we have used query expansion from the medical topic descriptors of document and the experimentation results are mentioned below. And the results of preprocessing techniques and term weighting functions are discussed below with various tuning Parameters. In this experimentation articles from PUBMED are extracted Using search query "How Brain Tumor Leads to brain hemorrhage" and are clustered under query topic "Brain Tumor" and this database of 1000 articles are used for analysis of pre-processing techniques, Term Weight-
5.1 Weighting Function for Key Terms
Before performing weighting function on documents, the retrieved documents must be pre-processed. In this work, we have analyzed various stemming and stop word removal methods on clustered document corpus and retrieved MEDLINE query Corpus. This work studies the impact of various stemming algorithms such as Husk and Porter Stemmer and investigates the impact of various stop word lists such as PUBMED and SMART in retrieval of the relevant documents from PUBMED, abstracts related to brain tumor. The proposed enhanced E-BM25 increment algorithm is used with default parameters ($k_1 = 1.3, 5$, $b = 0.75$, and $k_3 = 1.2$) and is applied to the title, abstract and MeSH fields.

We experimented with different values of $K_1, b, K_3$ to analyze the optimal values which can be used in the query performance improvement in terms of retrieval efficiency and below is the tables 6, 7, 8 which illustrates the use of different standards for the calibrate parameters. The Mean Average Precision is tested for set of queries extended from the initial query “How Brain Tumor Leads to brain hemorrhage” with different calibrate parameters values of $K_1, b, K_3$.

<table>
<thead>
<tr>
<th>Table 6. MAP values for Different Preprocessing and term weighting Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combinations</td>
</tr>
<tr>
<td>Porter stemmer PUBMED Stopwords</td>
</tr>
<tr>
<td>Porter stemmer SMART Stopwords</td>
</tr>
<tr>
<td>HUSK Stemmer PUBMED Stopwords</td>
</tr>
<tr>
<td>HUSK Stemmer SMART Stopwords</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 7. MAP values for Different Preprocessing and term weighting Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combinations</td>
</tr>
<tr>
<td>Porter stemmer PUBMED Stopwords</td>
</tr>
<tr>
<td>Porter stemmer SMART Stopwords</td>
</tr>
<tr>
<td>HUSK Stemmer PUBMED Stopwords</td>
</tr>
<tr>
<td>HUSK Stemmer SMART Stopwords</td>
</tr>
</tbody>
</table>

The table 7 has the different values of $K1$ where $b = 0.75$, and $K3 = 1.2$ with the search term brain tumor, for the query “How brain tumor leads to brain hemorrhage” [5]

The above Table 7 illustrates the use of different combinations of filtering mechanisms and the MAP measure on the use of different values for calibrate parameters in different term weighting functions on different expanded queries. It is measured that the use of $K1=1.3$ value with the combination of $b = 0.75$, $K3 = 1.2$ is measured high for various combination of filtering techniques and Term Weighting functions.

Table 8 has different values of $b$, where $K1=1.3$, $K3 =1.2$ with the search term brain tumor, for the query “How brain tumor leads to brain hemorrhage” [5]

Table 8 illustrates the analysis of different values of calibrate parameter $b$ from 0 to 1 and the result of MAP for different combinations of pre-processing and Term Weighting Techniques. It is observed that the value of $b = 0.75$ with the combination of $K1 = 1.3$ and $K3 = 1.2$ which have resulted in high MAP compared to other values.
Table 8

<table>
<thead>
<tr>
<th>Combinations</th>
<th>$K_1$</th>
<th>$b$</th>
<th>$K_3$</th>
<th>$K_1$</th>
<th>$b$</th>
<th>$K_3$</th>
<th>$K_1$</th>
<th>$b$</th>
<th>$K_3$</th>
<th>$K_1$</th>
<th>$b$</th>
<th>$K_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Porter stemmer PUBMED Stopwords</td>
<td>0.1652</td>
<td>0.1715</td>
<td>0.1807</td>
<td>0.1810</td>
<td>0.1824</td>
<td>0.1799</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Porter stemmer SMART Stopwords</td>
<td>0.1669</td>
<td>0.1695</td>
<td>0.1813</td>
<td>0.1815</td>
<td>0.1814</td>
<td>0.1817</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HUSK Stemmer PUBMED Stopwords</td>
<td>0.1667</td>
<td>0.1706</td>
<td>0.1804</td>
<td>0.1802</td>
<td>0.1819</td>
<td>0.1795</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HUSK Stemmer SMART Stopwords</td>
<td>0.1701</td>
<td>0.1707</td>
<td>0.1824</td>
<td>0.1823</td>
<td>0.1824</td>
<td>0.1818</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 9

<table>
<thead>
<tr>
<th>Combinations</th>
<th>$K_1$</th>
<th>$b$</th>
<th>$K_3$</th>
<th>$K_1$</th>
<th>$b$</th>
<th>$K_3$</th>
<th>$K_1$</th>
<th>$b$</th>
<th>$K_3$</th>
<th>$K_1$</th>
<th>$b$</th>
<th>$K_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Porter stemmer PUBMED Stopwords</td>
<td>0.1824</td>
<td>0.1825</td>
<td>0.1824</td>
<td>0.1823</td>
<td>0.1824</td>
<td>0.1822</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Porter stemmer SMART Stopwords</td>
<td>0.1814</td>
<td>0.1813</td>
<td>0.1813</td>
<td>0.1814</td>
<td>0.1814</td>
<td>0.1814</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HUSK Stemmer PUBMED Stopwords</td>
<td>0.1811</td>
<td>0.1819</td>
<td>0.1819</td>
<td>0.1822</td>
<td>0.1819</td>
<td>0.1819</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HUSK Stemmer SMART Stopwords</td>
<td>0.1817</td>
<td>0.1822</td>
<td>0.1820</td>
<td>0.1824</td>
<td>0.1824</td>
<td>0.1824</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 9 has different values of $K_3$, where $K_1 = 1.3$, $b = 0.75$ with the search term brain tumor, for the query “How brain tumor leads to brain hemorrhage” [5].

The above Table 9 illustrates the use of different values for $K_3$ from 0 to 2 and it was observed that at value of $k_3=1.2$ with combination of $K_1=1.3$ and $b = 0.75$, the MAP is high for all filtering techniques and term weighting functions.

5.2 Similarity Measure for Retrieving Relevant Documents

In this experimentation, term weighs are calculated based on the Mathematical Approaches Raw TF and Log TF, TF-IDF, BM25 and E-BM25, and it was observed that the E-BM25 approach has shown highest Mean Average Precision. And the calibrate parameters in the E-BM25 is experimented with a value of $K_1 = 1.3$, $b = 0.75$, $K_3 = 1.2$ and the approach with this value has shown highest precision and R-Precision.

Finding the set of optimal parameters is costly to compute, since they have local maxima that are singularity values [5], [33]. Hence, we are using a simplistic optimization approach. The best values obtained in our tests with the PUBMED corpora are $K_1 = 1.3$, $b = 0.7$, and $k_3 = 1.2$ for the TF-IDF weighting algorithm with Okapi $TF$ formula [5]. The Tuning Parameter values for Different Weighting Functions examined and it was taken with values $K_1 = \{0-2\}$, $b = \{0-1\}$, $K_3 = \{0-2\}$ and it was shown improvement in Mean Average Precision for the values $K_1= 1.3$, $b = 0.75$, $K_3 = 1.2$.

After finding the weights of query terms in documents, the key terms in documents are compared with the query terms using formulae equation 4. We find the score for each query Term in the document collection and which document scores high for the query terms and considered as relevant and using Enhanced Best Matching Score the similarity between query terms and the document terms are calculated and the nearest value terms are treated as relevant and the articles are represented with that terms. In Table 9 different weighting function with various preprocessing techniques examined with MAP measure and observed that SMART stop words removal, Husk Stemming with Enhanced BM25 approach results a high MAP compared with all other combinations.
5.3 Relevance Feedback from Thesaurus for Query Expansion

The similarity between the documents and terms calculated and a vector of similarity scores are represented as term-document matrix and from this vector a set of relevant documents and are identified. After identifying the set of documents, the query terms are analyzed with the document descriptors in thesaurus and then a relevance feedback from MeSH is considered to re-weigh the terms and expand the query with high weighted terms of the retrieved relevant document collection. The quality of query expansion is depending on retrieved top relevant documents [5],[36]. To improve the relevancy of top retrieved documents, compared different term weighting functions for query terms and documents terms to know the importance of term for the query performance. Noisy and multi concepts are two major negative factors for expansion term selection. For instance, if an advertisement is embedded in a top-ranked web page at the first round, then some terms from the advertisement may be selected as expansion terms [5]. Once these terms are used to expand the query for the second-round retrieval, irrelevant web pages containing these advertisement terms could be ranked highly. Similarly, for a web page containing multiple concepts, the selected terms are also subject to this uncertainty which may decrease the retrieval performance. Therefore, it is necessary to segment a web page into semantically related units so that noisy information can be filtered out and multiple topics can be distinguished [5]. Table 10 and 11 show the expanded query results.

<table>
<thead>
<tr>
<th>Combinations</th>
<th>MAP</th>
<th>R-Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>PUBMED Stopwords Porter stemmer E-BM25</td>
<td>0.245</td>
<td>0.215</td>
</tr>
<tr>
<td>PUBMED Stopwords HUSK Stemmer E-BM25</td>
<td>0.232</td>
<td>0.223</td>
</tr>
<tr>
<td>SMART Stopwords Porter stemmer E-BM25</td>
<td>0.243</td>
<td>0.232</td>
</tr>
<tr>
<td>SMART Stopwords HUSK Stemmer E-BM25</td>
<td>0.254</td>
<td>0.243</td>
</tr>
</tbody>
</table>

Table 10. MAP, R-Precision Measures Using Relevance Feedback from Thesaurus [5] Using E-BM25 approach

Table 10 describes the use of different preprocessing functions with E-BM25 approach and the relevance score with respect to Mean Average Precision and R-Precision for different queries with relevance feedback from MeSH and top retrieved Documents. This study proved that use of SMART stop word list with Husk Stemming and E-BM25 Term Weighting Function and similarity measure for base query and each iteration improved query with relevance feedback have shown improvement.

Table 11 illustrates the Mean Average Precision and R-Precision values against the use of different filtering and term weighting functions for the base query “How Bain Tumor leads to Brain Hemorrhage” and after first pass of retrieval from the relevance feedback the query terms are exchanged with significant terms from the retrieved results and the query terms are reweighted to find the relevance. The below table shows the Mean Precision for different queries and R-Precision for rank 100 documents.
5.4 Relevance Feedback from Clustered Biomedical Topics for Query Expansion

Similarity score between document and query is found using equation 4. This similarity score vectors of document terms are given as input to Multi kernel Fuzzy C Means clustering algorithm inorder to find the cluster membership for each point of the vector. The Objects or Terms of the Vector are assigned to a cluster whose centroid is near to the membership of term vector. The terms which are near to the cluster center are assigned to those clusters and the documents of this represented term vectors are grouped under most relevant overall topic. A topic is any MeSH(Medical Sub Heading) under which all the documents are relevant i.e., whose cluster centroid is near to membership of term vector.

All retrieved documents based on the search queries such as “How Brain Tumor Leads to Brain Hemorrhage” are clustered under suitable and most relevant topics, i.e., “Brain Tumor”. Retrieved Documents are based on “which gene Mutation causes Kidney stone”. The important topics related to the above diseases are clustered using Multi kernel fuzzy c means [7]. In this study most relevant terms in topic based biomedical clustered literature considered to enhance the query.

The above table shows the combined average precision of the queries implemented on MEDLINE biomedical literature at particular rank. The results set is given as relevance feedback to clustered biomedical topics in order to find the most relevant terms to the query and to expand the query with that most relevant topics. 20 iterations taken place to execute 20 expanded queries and in each iteration a rank of 50 top documents are considered for query expansion. The precision value in each iteration of the query with relevant documents is calculated and after 20 iterations the average precision and Mean Average Precision of relevant documents in all iterations with all 20 queries are calculated and results are mentioned in above Table.12.

6. Conclusion

In this work, we have analyzed different preprocessing approaches for literature mining from PUBMED Corpus such as stop word removal for the list of stop words from PUBMED and SMART. Various Stemming approaches such as HUSK stemmer with PUBMED and SMART stop words, Porter Stemmer with PUBMED and SMART stop words are used and we found that HUSK stemmer with SMART stop words combination shows considerable improvement in the Mean Average Precision. The preprocessing analysis results are analyzed with different weighting functions and it was observed that Husk Stemming with SMART stop words list removal will refine the noise terms and E-BM25 will normalize the document length and term weights and Query Length with calibrate parameter values $K_1=1.3$, $b = 0.75$, $K_3 =1.2$ and this approach has shown improvement in Mean Average Precision and R-Precision. This work preceded with taking the Relevance Feedback from R-Retrieved Articles, R-Relevant Articles, Clustered Biomedical Topics and controlled vo
caborcular for expanding the query with high weighted terms. The Query Expansion Techniques with the above sequence of approaches have shown maximum improvement in query retrieval performance. Further, this study can be applied on topic distribution models to find the relevant terms for the Query Expansion.

References


