DOI: http://dx.doi.org/10.26483/ijarcs.v9i1.5237

ISSN No. 0976-5697

Volume 9, No. 1, January-February 2018

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

IMPACT SCORE ESTIMATION WITH PRIVACY PRESERVATION IN INFORMATION RETRIEVAL

Kinjal Sheth
Research Scholar, Dept. of Computer Science
C. U. Shah University
Wadhwan (Surendranagar), India

Dr. Harshad Bhadka
Dean, Computer Science Department,
C. U. Shah UniversityWadhwan (Surendranagar),
Gujarat, India

Dr. Ashish Jani
Associate Professor, School of Engineering,
P. P. Savani University,
Surat, Gujarat, India

***Abstract*:** Nowadays, Information Retrieval (IR) is becoming more popular technique due to the tremendous growth of resources on the internet. However, the present information retrieval techniques have several limitations such as lack of semantic keyword, more time consumption and vague user's query, etc. To mitigate these issues, this paper proposed a novel Information Retrieval (IR) framework to achieve effective data access which is available in online. The proposed IR system includes five major steps, at first the documents which are shared as the resources are pre-processed, and domain analysis is made to find the category of the document. Secondly, the keywords are extracted using semantic keyword extraction and indexing, and impact score estimation is obtained to determine the importance of the keyword in each document. Thirdly, the document similarity is estimated using novel similarity estimation algorithm for clustering the documents based on the attained score. Fourth, the documents are ranked based on the similarity score and the impact score of the keywords in the query. Finally, the user needs to register their personal information based on the novel privacy preservation algorithm to maintain the privacy of the querying users. The simulation results of proposed framework achieved significant improvement than existing approaches in terms of average precision, recall, mean average precision and execution time.

Keywords: Information Retrieval (IR), semantic keyword extraction, impact score estimation, novel similarity estimation, Levenshtein distance.

1. INTRODUCTION

Information retrieval is a sub-domain of computer science field and refers to the process where the collection of data is represented, gathered and searched for the application of knowledge discovery as a response to a query [1, 2]. This process consist of various steps start with representing data and finishing with returning relevant information to the user, and the intermediate steps include filtering, searching, matching and ranking operations [3]. The main objective of information retrieval system is to find the relevant information and a document that fulfils the information requirements. In information retrieval (IR), queries and documents are denoted by term vectors in which every term is a content word [4]. The similarity of a query and document is considered as a dot product, and the traditional information retrieval system frequently fails to find the relevant documents relevant documents when synonymous words are used in a dataset. One of the best solutions of this problem is to use query expansion [5, 6], but these replacements still depend on the same idea that is queries and documents should share the same words.

Query expansion aims at improving the efficiency of information retrieval operations by providing more relevant documents as per the query search. The method of reformulating a seed query is used here where expansion terms based on original query terms are included [7]. The traditional methods of query expansion are classified into two types such as automatic relevance feedback which includes pseudo relevance feedback (PRF) [8, 9] and the other is log-based QE [10, 11]. The performance of information retrieval (IR) was found to be improvised by employing Query expansion technique. Nevertheless, these approaches consider the interactions between terms in a single local corpus in the process of query expansion [7]. Hence, several information retrieval systems suffer from two issues in query expansion in low retrieval performance. (a) Term relationships are restricted, and some terms may be false expansion terms and result in topic drift, (b) there are several terms only existing in the query set but not in the document set for a specified corpus.

However, on considering the algorithm that was used in the retrieval model, it was observed that selection techniques utilized in the conventional query expansions were either association or corpus statistics based concerning automatic query expansion. Consequently, diverse conventional term selection approaches were derived considering the rank, semantic filtering based rank and score combination [17]. These techniques includes Okapi BM25 [12], Co-occurrence Information [16], Kullback-Leibler Divergence (KLD) [21], Robertson Selection Value (RSV) [18], and Binary Independence Model (BIM) [19]. Furthermore, multiple term ranks were combined by distinct approaches such as Borda [22], Condorcet [23], Reciprocal [23], and the

SumScore [24] which are obtained from Co-occurrence, KLD, BIM and RSV techniques [24].

In this paper, a Novel Information Retrieval Framework is proposed to improve the effective access to the data which is available in online. Semantic Keyword Extraction& Indexing are used to extract keywords and an additional updating. Then, the Impact Score Estimation is achieved for the keyword in the document to determine the importance of the keyword in each document. Finally, the document similarity is computed based on a Novel Similarity Estimation Algorithm to cluster the documents based on the computed score. .This paper is systematized as follows: The related works will be reviewed in Section 2. Section 3 presents the proposed methodology. In Section 4, simulation results are discussed and compared with the other existing techniques. Finally, section 5 concludes the paper.

2. RELATED WORK

The Probabilistic Relevance Framework (PRF) is the method for document retrieval which is developed based on the text-retrieval algorithms called BM25 [12]. In this, the retrieval models considered the meta-data which includes structure and link-graph information that provides powerful web-search and corporate algorithms (BM25). An adaptation of BM25 weighting approach is developed for two non-textual modalities ratings as well as geographical coordinates [13]. An integration of term proximity scoring into Okapi BM25 is presented [14] which provides the maximum gain. An enhanced Okapi’s BM25 information retrieval model is proposed using the term proximity evidence [15]. The proximity between query terms includes the window-based N-gram Counting method, various statistics such as the Poisson process, an empirical function, and an exponential distribution. The results demonstrate that the enhanced BM25 improves the retrieval efficiency based on the term proximity.

An enhancing the efficiency of an automatic retrieval model is developed based on the co-occurrence statistics [16]. The relationship between the terms in the query and corpus is estimated by utilizing the co-occurrence statistics. Also, a number of estimation rules are provided and the feasibility of the computations needed for a nonlinear weighting method. of the performance of Pseudo-Relevance Feedback (PRF) is enhanced by considering hybrid approaches by the integration of co-occurrence and semantic information based on query expansion was proposed [17]. An optimal combination of query terms is selected by considering corpus-based term co-occurrence approach and are generally attained using PRF. Then, semantic similarity method is used to rank the query expansion terms which is achieved from top feedback documents. The simulation results demonstrate that the technique achieved significant performance improvement in terms of precision and recall.

The Robertson Selection Value (RSV) was ascertained as an Information Retrieval (IR) system based on a Swets model [18]. The connections between queries were considered for ranking the terms retrieved by utilizing this model. Furthermore, Swets theory was generally employed to distribute the values that are matching the functions related to the collection of documents. The relevant and non relevant distributions are considered in this approach From tis model, it was ascertained that the a higher match function values were observed for the relevant documents in comparison to the non relevant documents. The simple estimation of the document and query probability were determined by using Binary Independence Model (BIM) [19]. This method was found to consider the keywords that were significant for both the relevant and nonr elevant class (NR) based on documents that are statistically independent.

Text classification and retrieval or filtering are achieved based on the Kullback-Leiblerdistance which is proposed in [20]. Consequently, this distance is calculated between the probability distribution of the document and the relevant category. Based on the similar representation of categories, these experiments demonstrate a significant improvement. The Kullback-Leibler Divergence (KLD) was defined based on the information theory [21]. This approach was further integrated with statistical language modelling in speech processing and natural language applications in information retrieval (IR). KLD is further developed to term-scoring function considering the difference in the keywords selected among the top retrieved documents and the overall document collection.

Based on the Borda rank combining method, each voter has its partiality list of candidates [22]. In this, the top first candidate attains k points, the top second candidate acquires k-1 points, and the top third candidate achieves k-2 points for each voter. The total value obtained by each voter is used to provide the final points to the candidate. Besides, the remaining points will be provided to the unranked candidate. Therefore, the candidate having the highest point will win. Condorcet ranking combination approach was developed in order to rank the candidate based on the idea that the candidate with the highest majority will beat the other candidates in pair-wise assessment [22, 23]. If there are more than one candidates unranked then all the unranked candidate's connection with each other. However, the major challenges observed in these methods were in estimating the individual query. To overcome this issue, the genetic algorithm is used in both Borda rank combining method and Condorcet ranking combination method [24].

A simple ranks combination approach is the reciprocal ranking technique which is developed based on the rank scores values. Using this method, first top candidate term attain score 1, the second top candidate term attain score 1/2 and third top candidate term attain score 1/3 for every voter [23]. A non-ranked candidate term of a voter is not used in the calculation of this voter. Therefore, all the candidates were found to be ranked considering their final score. The combination of similarity score of each candidate term is generally defined by the summation of the similarity score obtained while considering the query selection approaches. In order process the similarity score, diverse query expansion term are normalized before the integration process [24]. It is provided that the SumScore ranks score combination approach outperforms other score combination methods. Thus, the Sum Score method is employed to rank the highest candidate term based on the query made by the user. A novel query expansion approach is proposed [25] which aims to develop a tree of associational semantics model and select candidate keywords from the tree. Initially, a set of initial semantic trees for original keywords are created using WordNet thesaurus. Then, noise nodes on the trees are detached by estimating the similarities between words. At last, the nodes on the integrated tree are filtered and augmented based on Mutual Information. The results show the better performance in terms of precision.

3. RESEARCH METHODOLOGY

In recent years, the cloud storage services have grown hugely due to the application of internet services and the availability of the massive storage services. The size of the data upload and download has enhanced enormously which leads to the delivery of huge information resources to the internet users. The data are any kind like text documents, images albums, videos and other kinds of multimedia files. These information sources are freely available as a shared resource to the internet users which can be retrieved by the users with the help of the querying. Hence referred to information retrieval. Information retrieval (IR) can be formally defined as the process of relevant information extraction from the vast distributed information resources which are available online. IR model is mainly based on the levels of searching which is done. It can be categories into document retrieval (DR) (get the relevant document), Question answering (QA) model (designed to extract the relevant answers from the given set of information's), document summarization (DS) (create a summarized idea from a set of documents), Topic identification/Detection (TD) from the given set of documents and Recommendation Systems (RS).IR is one of the major processing methods which are followed by the internet users as a day to day routine. The IR services are given to the users by processing of search queries which performs matching the keywords in the query with the documents which are available. The data access is given to user by the web search engines which can accept the query information from the users. Hence the processing of this document extraction is critical leading to development of automated query processing techniques.

Some of the search engines such as Google, Yahoo and other academic search engines such as IEEE and Springer web search engine provides access to massive data repository by extracting keywords from the query and matching the keywords to the database which are based on the similarity estimation techniques and the most relevant document is retrieved by the rank based computation. Though IR is in practice, there is no assurance of the 100% accuracy in retrieval due to some of the issues like, the data which are retrieved are based on the exact keyword matching methods, indexing based matching methods, semantic matching is not implemented, the matching and the exaction of information is highly time consuming due to the huge amount of the data available, the topic identification and classification of the available information is not perfectly maintained.

The document grouping and the data extraction is difficult. Hence, a novel information retrieval framework is proposed to solve these issues and provide efficient access data which are available in online. The block diagram of proposed novel information retrieval framework is shown in Fig.1.

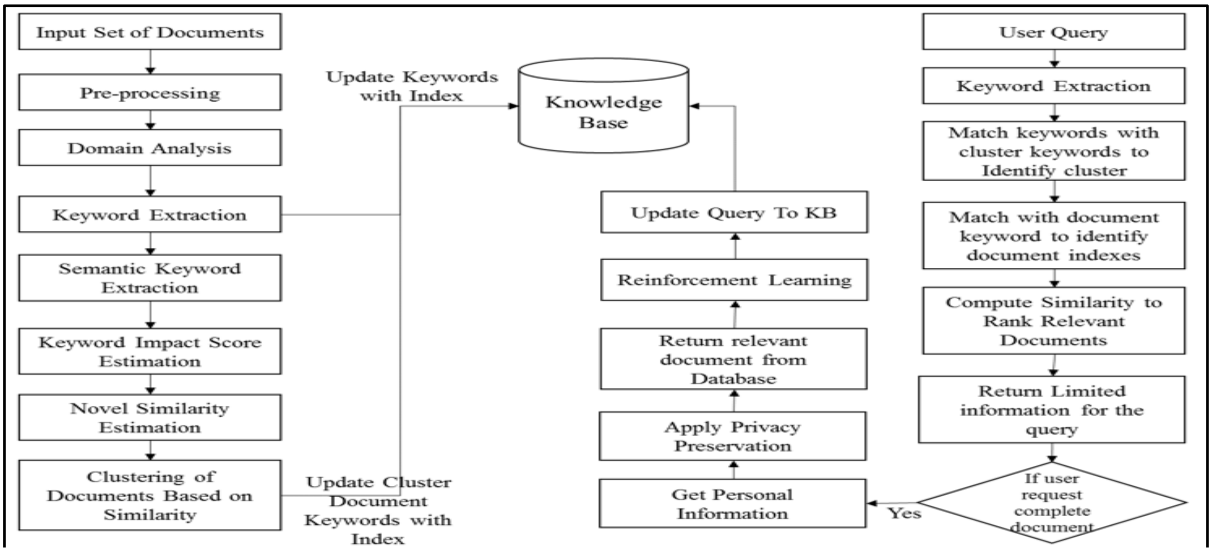
Input Set of Documents
Pre-processing
Domain Analysis
Keyword Extraction
Semantic Keyword Extraction
Keyword Impact Score Estimation
Novel Similarity Estimation
Clustering of Documents Based on Similarity
Update Keywords with Index
Knowledge Base
Update Cluster Document Keywords with Index
Update Query To KB
Reinforcement Learning
Return relevant document from Database
Apply Privacy Preservation
Get Personal Information
Yes
User Query
Keyword Extraction
Match keywords with cluster keywords to Identify cluster
Match with document keyword to identify document indexes
Compute Similarity to Rank Relevant Documents
Return Limited information for the query
If user request complete document

Fig.1 Representation of the proposed Information Retrieval Framework.

Pre-processing

Initially, the documents which are shared as the resources are considered. The pre-processing is a fundamental step to load the text into the proposed framework that increases the accuracy of the system by differentiates similar words. To achieve this, the pre-processing step includes the stop word removal, word tagging and stemming. Stop word removal is a procedure to make a filter for those words that remove them from the document which reduces the size of the candidate keywords. Word tagging is the method of assigning Part-of-Speech (POS) to each word in a sentence to provide word class. Tokenization is an essential element of IR systems, includes pre-processing of specified documents and makes respective tokens. Also, it is a critical activity in IR systems, which simply segregates all the words, numbers, and their characters from given document. These recognized words, numbers, and other characters are known as tokens. Besides, this process estimates the frequency value of all these tokens existing in the input documents. Stemming is the process of removing suffixes which are use fulin keyword extraction and information retrieval. The proposed information retrieval framework uses the Porter stemming algorithm with the enhancements on its rules for stem. This is achieved by removing the various suffixes such as -ED, -ING, -ION, IONS. This process reduces the number of terms as well as reduces the size and complexity of the data in the IR system.

Domain Analysis

Domain Analysis is defined as the procedure of analyzing number of related systems providing a method to make a conceptual structure. This structure is capable of determining, understanding and using data structures, techniques and algorithms effectively. A domain analysis structure is developed [26] which is nothing but a system analysis for multiple related systems. The crucial steps of domain analysis include the two activities. One is to determining the important concepts and vocabulary in the domain and other is to define these important concepts and vocabulary.

Semantic keyword extraction

In the proposed methodology, the semantic keywords are extracted using WordNet and Synset. WordNet is widely used and largest lexical databases of English. Generally, WordNet includes some specific terms from every subject related to their terms. It characterises all the stemmed words from the standard documents into requires lexical groups. Nouns, verbs, adjectives and adverbs are organized into sets of conceptual synonyms, each conveying a different concept. WordNet groups English words into sets of synonyms called synsets which offers short, general definitions, and the various semantic relations between these synonym sets. Synsets are related to the conceptual-semantics and lexical groups. The computation of the linguistics and natural language will be processed by utilizing the WordNet's framework by grouping the words randomly based on their meanings. There are two main purpose of WordNet are, to create a combination of dictionary and thesaurus and to provide automatic text analysis. WordNet is used several different application in information systems includes information retrieval, word sense disambiguation, automatic text classification and so on. Fig.2 shows that the process of the semantic keyword extraction.

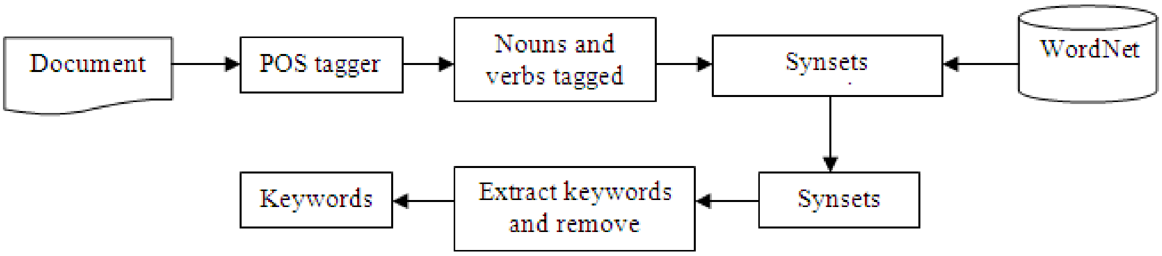

Fig.2 Representation of the process involved in the Semantic Keyword Extraction

Initially, the text of a document is tagged with POS tagger and achieves the synset of nouns and verbs using WordNetontology. Then, to add only the word, without POS andsynsets information and to remove repeated words in thesynsets. At last, all words in the synsets are included at the end of the original text. Followed by the Semantic Keyword extraction, the Impact Score Estimation is done for the keyword in the document to determine the importance of the keyword in each document.

Novel Similarity estimation

As the next step the document similarity is computed based on a Novel Similarity Estimation Algorithm to cluster the documents based on the computed score. Here the document clustering is implemented as an added advantage to reduce the processing time and the information retrieval time for a particular query. Each cluster will be represented by the main keywords in the cluster based on the estimated score.

Euclidean distance is one of the mostly used standard metric for geometrical problems. It is calculated as the normal distance between two points and it is easily measured with a ruler in two or three-dimensional space. Euclidean distance is widely used in clustering problems, including clustering text. In the proposed IR framework, clustering the documents based on the estimation of minimum Euclidean distance. The minimum Euclidean distance between two points (x, y) and (a, b) are expressed as,

$$E_{dist} = \sqrt{(x-a)^2 + (y-b)^2}$$

In information retrieval, the Levenshtein distance is a string metric for computing the amount of difference between two sequences. The term edit distance is frequently used to refer particularly to Levenshtein distance. The number of edits that is required for the transformation of one string to the other is defined by the distance between the two strings. This operation permits edit operations such as insertion, deletion, substitution of a single character. The Levenshtein distance is also called as edit distance which is widely used metric. The distance is calculated based on the Wagner-Fischer method. The Levenshtein distance is mathematically expressed as,

$$D_L(i,j) = \begin{cases} \max(i,j) & if\ \min(i,j) = 0 \\ min \begin{cases} L(i-1,j)+1 \\ L(i,j-1)+1 \\ L(i-1,j-1)+1 \end{cases} & otherwise \end{cases}$$

Percentage for two strings can be calculated by,

$$Percentage = (100 - D_L(\mathrm{i,j}) * 100/(i.length + j.length))$$

Where, i, j are two strings from documents.

```
Algorithm: Estimation of Levenshtein distance
Step 1:Set n to be the length of s.
Set m to be the length of t.
If n = 0, return m and exit.
If m = 0, return n and exit.
 Construct a matrix containing 0 to m rows and 0 to n columns.
Step 2: Initialize the first row to 0 to n and the column to 0 to m.
Step 3: Examine each character of s (i from 1 to n).
Step 4: Examine each character of t (j from 1 to m).
Step 5: If s[i] equals t[j], the cost is 0.
If s[i] doesn't equal t[j], the cost is 1.
Step 6:Set cell D(i, j) of the matrix equal to the minimum of:
a. The cell immediately above plus 1: d(i-1, j) + 1.
b. The cell immediately to the left plus 1: d(i, j-1) + 1.
c. The cell diagonally above and to the left plus the cost: d(i-1, j-1) + cost.
Step 7:Repeat step 3 to 6,the distance is found in cell D(n, m).
```

Novel Privacy Preservation Algorithm

Once user inputs the query for search, the keywords of the query are extracted and matched with the indexed keywords of the clustered documents to choose the best cluster to return the relevant document. Once the best cluster is identified the document keywords are matched with the query keywords based on the similarity estimation. Based on the similarity score the documents are ranked based on the impact score of the keywords in the query. A limited amount of information is retrieved from the document as the query result to the user. If the user is in need of accessing the whole document from the resources, then the user needs to register with this personal information, where a Novel Privacy Preservation Algorithm is implemented to maintain the privacy of the querying users. The user history and relevant search are recorded for user specific effective relevant document retrieval. The best results for any user query is also updated by the feedback based score evaluation. Hence the proposed Information Retrieval framework is expected to be outperforming while comparing with the existing IR models. The Proposed state-of-art techniques will be compared with the existing techniques on the scale of Accuracy, Mean Absolute Error, Precision, Recall, Sensitivity, specificity and other metrics to evaluate the results for proving the effective of the model.

4. RESULTS AND DISCUSSION

The performance of the proposed information retrieval system is analysed based on the two evaluation parameters such as Recall and Precision. Recall is defined as the set of relevant documents retrieved divided by the set of all relevant documents and it is expressed as,

$$Recall = \frac{|D_R|}{|D_{AR}|}$$

Where, D_R is the set of relevant retrieved documents and D_{AR} is the set of all relevant documents.

Precision is defined as the ratio of set of relevant documents retrieved to retrieved documents set and it is expressed as follows,

$$Precision = \frac{|D_R|}{|S_R|}$$

Where,S_R - Retrieved documents set.

The average precision rate of the documents is obtained in terms of the mean precision of the relevant documents. It is expressed as,

$$Average\ Precision = \frac{1}{n}\sum_{i=1}^{n} Precision(D_i)$$

Where, D_i is the relevant document set.

a) Techniques employed in Query expansion term selection

In the proposed research, the performance evaluation of the proposed query expansion approach achieved significant improvement over other approaches such as Okapi-BM25, Co-occurrence Based Query Expansion (CBQE), Robertson Selection Value Based Query Expansion (RSVBQE), Binary Independent Model Based Query Expansion (BIMBQE)and Kullback-Leibler Divergence Based Query Expansion (KLDBQE).Thus, the comparisons of diverse query expansion approaches are tabulated based on average precision and recall as shown in Table 1. The distinct query expansion selection approaches have been compared and demonstrated in Fig.3.. From this, the performance of proposed query expansion term selection approach is the greater than other approaches in both top 10 and overall retrieved documents set.

Table 1. The performance of diverse query expansion term selection approaches has been compared

Methods	Top 10 Documents		Overall Documents	
	Average Precision	**Average Recall**	**Average Precision2**	**Average Recall2**
Okapi-BM25	0.2378	0.1172	0.1955	0.3012
CBQE	0.2381	0.1195	0.1996	0.3165
RSVBQE	0.2435	0.1193	0.2098	0.3159
BIMBQE	0.2477	0.1267	0.2273	0.3285
KLDBQE	0.2525	0.1295	0.2395	0.3301
Proposed Method	0.2653	0.1368	0.2597	0.3605

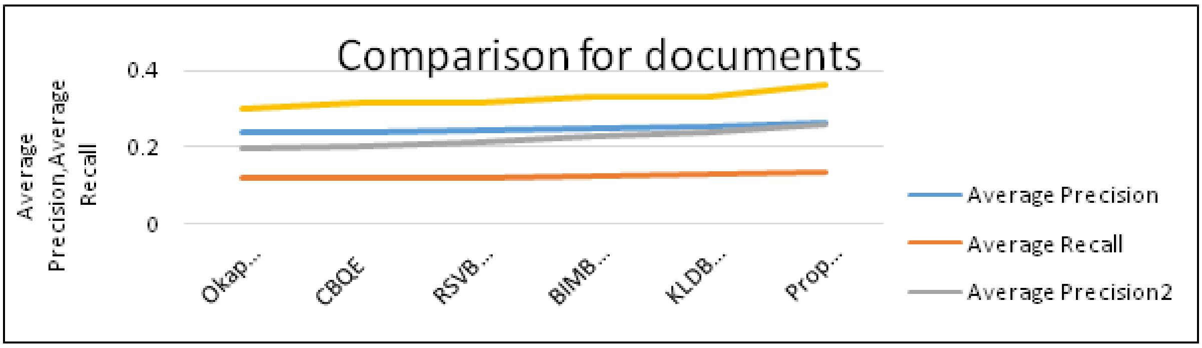

Fig.3 Discretequery expansion term selection techniques are compared s.

b) Rank-based approaches

Table 2 shows that the average precision and recall performance of different exiting and proposed rank combination approaches. In this, Aguera et al.'s as curtained the integration of three query expansion term selection approach. The retrieval performance proposed rank combination approaches SumScore Based

Query Expansion (SSBQE), Reciprocal Based Query Expansion (RBQE), Aguera et al.'s model, Condorcet Based Query Expansion (CNBQE) and Borda Based Query Expansion (BBQE) obtained considerable improvement over existing approaches. Moreover, it was noticed that the on comparison with distinct ranking based approaches, the retrieval performance of the CNBQE and BBQE approaches was found to be enhanced. Fig.4 depicts that the comparison of different conventional and proposed ranking based approaches where the proposed approaches achieved better retrieval performance.

Table 2. Comparison of different rank combination approaches in terms of average precision and recall

Methods	Existing System		Proposed System	
	Average Precision	**Average Recall**	**Average Precision2**	**Average Recall2**
SSBQE	0.2639	0.1324	0.2735	0.145
RBQE	0.2747	0.1369	0.285	0.1472
Aguera et al.'s model	0.2986	0.1397	0.3052	0.1523
CNBQE	0.3075	0.1485	0.3281	0.1523
BBQE	0.3183	0.1483	0.3353	0.1538

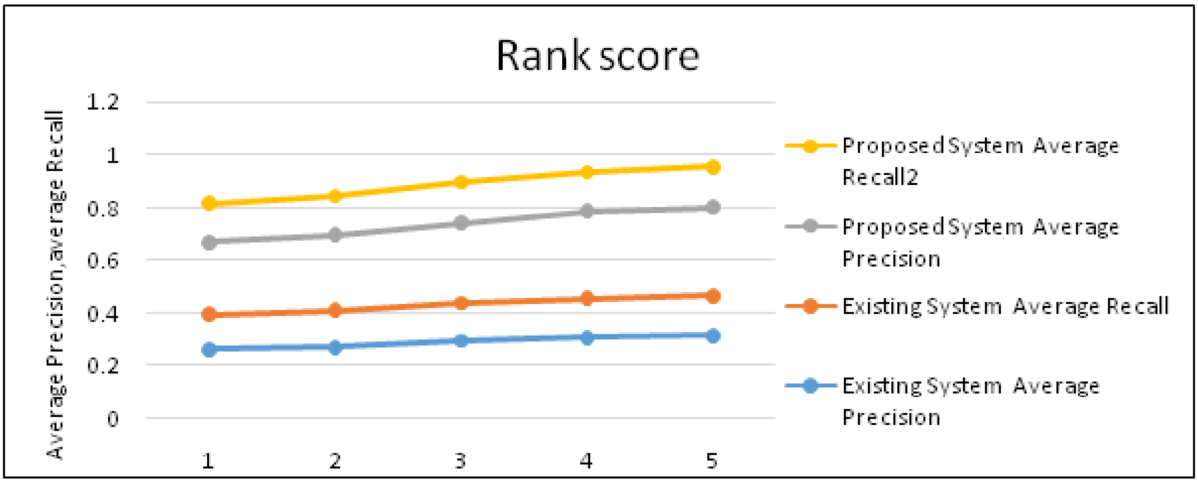

Fig.4 Diverse rank combination approaches are compared.

c) Semantic rank score-based approaches

The retrieval performance of various traditional and proposed semantic based rank combination approaches using average precision and recall parameters for both top 10 and overall documents are shown in Table 3. A semantic based query expansion technique has been developed by Zhang et al.'s. The proposed semantic rank based information retrieval approaches are attained better precision and recall performance than other semantic based approaches such as SumScore and Semantic GeneticBased Query Expansion (SSSGBQE), Reciprocal and Semantic Based Query Expansion (RSBQE), Zang et al.'s model, Borda and Semantic Genetic Based Query Expansion

(BSGBQE) and Condorcet and Semantic Genetic Based Query Expansion (CNSGBQE).The diverse stages used in the semantic rank based technique based on the average precision and recall is shown in Fig.5.

Table 3. Comparison of different semantic rank based approaches in terms of average precision and recall

Methods	Top 10 Documents		Overall Documents	
	Average Precision	**Average Recall**	**Average Precision2**	**Average Recall2**
SSSBQE	0.2583	0.1259	0.2154	0.3163
RSBQE	0.2631	0.1367	0.2214	0.3228
Zang et al's Model	0.2722	0.1303	0.2456	0.3359
CNSBQE	0.2956	0.1474	0.2523	0.3506
BSBQE	0.3029	0.1491	0.2799	0.3513
Proposed Method	0.3236	0.1633	0.3059	0.4064

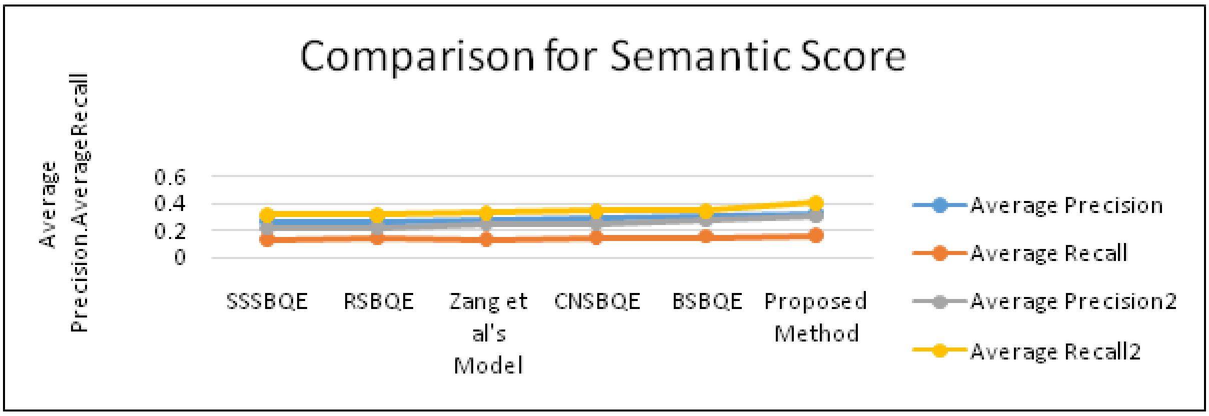

Fig 5. The semantic rank based approaches are compared

d) Keyword retrieval rate

The rate of keywords retrieval in conventional and proposed approaches based on the number of queries is shown in Table 4. In this, the queries include both simple query as well as complex query. It depicts that the keyword retrieval rate is increased when increasing the number of queries. Also, the proposed system achieved greater retrieval rate when compared to the existing system. Fig.10demonstrates that the plot between number of queries and information retrieval rate.

Table 4. Number of Queries versus rate of keywords retrieval.

Queries	Simple Query	Complex Query	Rate in % (Existing System)	Rate in % (Proposed system)
10	5	5	20	20
15	3	12	30	40
20	10	10	45	48
30	20	10	49	50
40	25	15	56	60
50	26	24	58	64

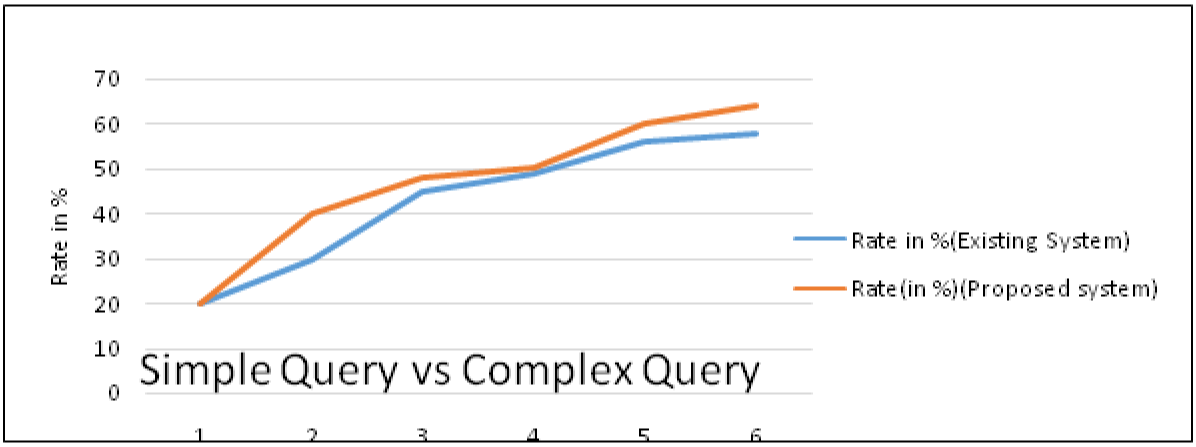

Fig.6 Number of Queries versus rate of keywords retrieval.

e) Similarity Estimation

The performance of proposed novel similarity estimation is obtained by estimating the value of mean average precision. The average of documents is taken to determine the mean average precision based on the documents that are retrieved. The mean average precision is expressed as,

$$Mean\ Average\ Precision = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{Q_j}\sum_{i=1}^{Q_j} Precision(D_i)$$

Where, Q_j is the number of relevant documents for query j, n is the number of queries and $Precision(D_i)$ is the precision at ith relevant document. In this approach, every document which are not retrieved will utilize a a zero precision.

The comparison of various similarity estimation approaches in terms of mean average precision is shown in Table 5. From this, the mean average precision performance of Leventein’s based approach obtained better results than other approaches such as Tf-Idf, BM25, CENTROID, SEM. Fig.7 show that the Mean average precision performance comparison of different similarity estimation approaches.

Table 5. Comparison of different similarity estimation techniques in terms of mean average precision

Methods	Existing System	Proposed System
Tf-Idf	0.3018	0.2375
BM25	0.3136	0.2463
CENTROID	0.2363	0.2459
SEM	0.3732	0.2601
Leventein'sDistance	0.3965	0.285

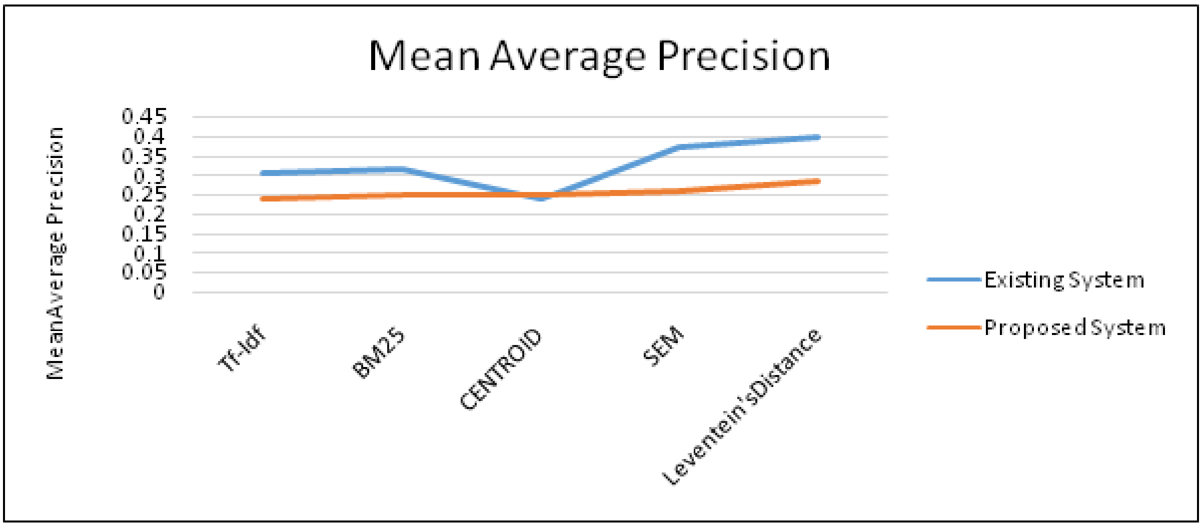

Fig.7 Mean average precision performance comparison of different similarity estimation approaches.

f) Clustering Vs Time

In the proposed research, minimum Euclidean distance based technique is used for document clustering. This method achieves more efficiency than other methods and it has consumed minimum time that is 350 seconds. But it is more than the K-means clustering techniques. Table 6 shows that the comparison of various traditional and proposed document clustering techniques. The proposed minimum Euclidean distance based clustering is consumed less time than spectral, active K-means, activespec and Heurspec.

Fig.8 demonstrates that the comparison of different clustering techniques using time consumption.

Table 6. Comparison of different document clustering techniques in terms of time consumption

Clustering Methods	**Time(s)**
HeurSpec	1350
ActiveSpec	450
Active K-means	420
K-means	160
Spectral	5660
Proposed Method	350

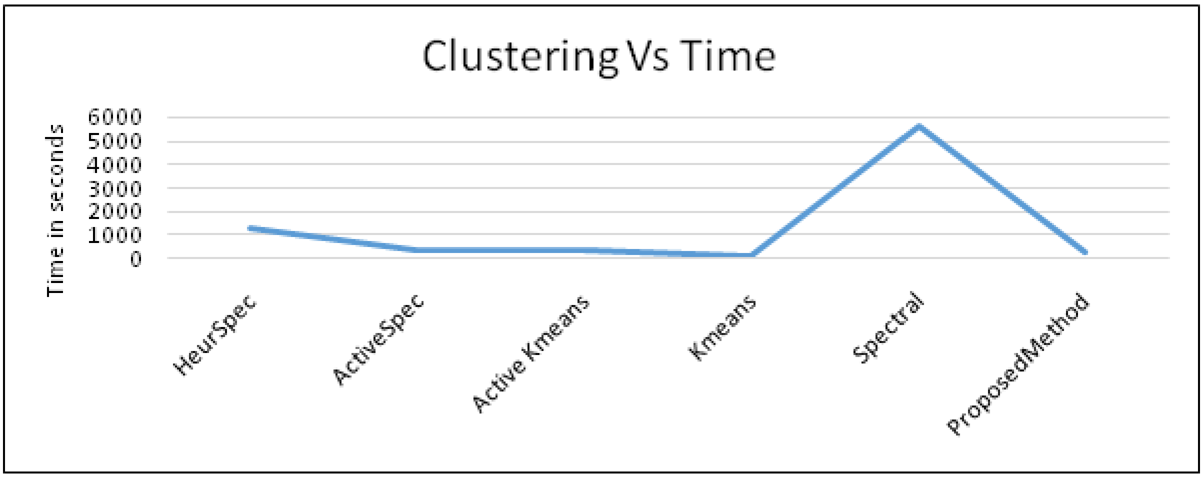

Fig.8 Comparison of different clustering techniques using time consumption

g) No of Keywords Vs Document Retrieval

Table 7 shows that the percentage of document retrieval in conventional and proposed approaches based on the number of keywords. From this, the percentage of document retrieval is increased when increasing the number of keywords. In addition, the percentage of document retrieval from the proposed approach is better than the conventional approaches. Fig.9 shows that graph of the number of keywords versus document retrieval.

Table 7. Number of keywords versus document retrieval

Keywords	**Existing System**	**Proposed System**
	Document Retrieval in %	**Document Retrieval in %2**
100	20	20
400	28	30
500	30	33
1500	36	38
2000	40	45
2500	46	50
3000	65	69

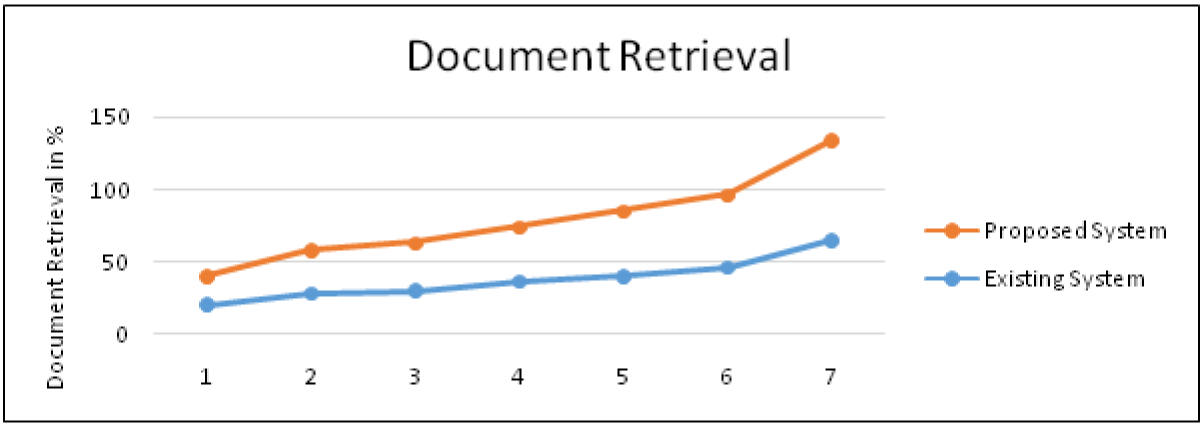

Fig.9 Number of keywords versus document retrieval.

h) Keywords Vs Execution Time

In the IR system, execution time is one the important parameters which is need to be estimated for verifying efficiency of the system. The execution time of document retrieval process in conventional and proposed approaches based on the number of keywords is shown in Table 8. It depicts that the execution time is increased when increasing the number of keywords. Also, the proposed IR system consumes less time when compared to the existing system.Fig.8 shows that the plot between number of keywords and execution time.

Table 8. Number of keywords versus execution time.

Keywords	Existing System	Proposed System
	Execution Time(ms)	Execution Time(ms)2
100	10	10
400	14	12
1000	19	17
1500	22	18
2500	24	20
3500	26	23
5000	36	35

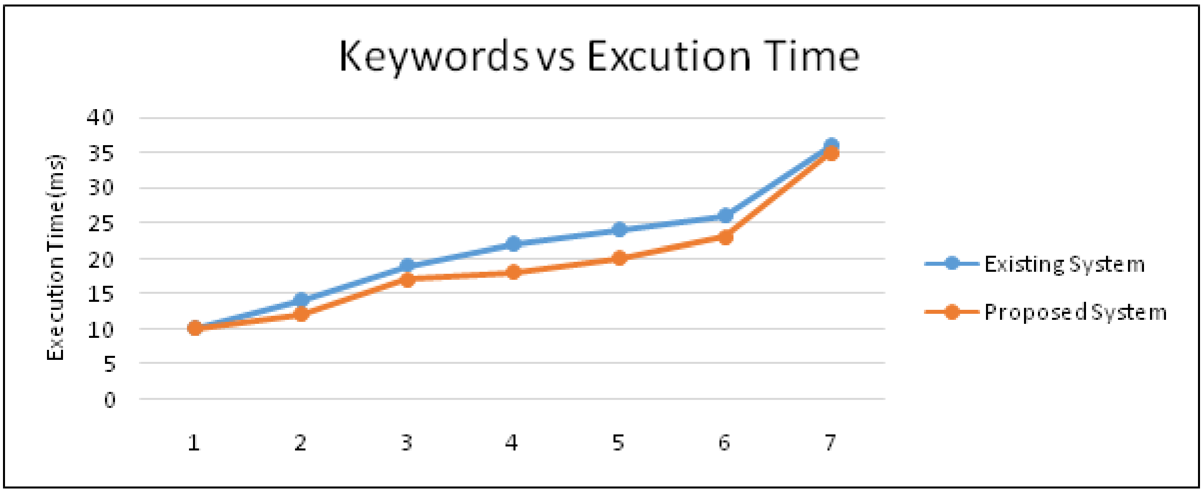

Fig.10Number of keywords versus execution time.

i) Query Vs Execution Time

The execution time of information retrieval system in conventional and proposed approaches based on the number of keywords is shown in Table 9. It demonstrates that the execution time is increased when increasing the number of queries. Besides, the proposed IR system takes less time when compared to the existing system. Fig.11demonstrates that the plot between number of keywords and execution time.

Table 9. Number of queries versus execution time.

Queries	Existing System	Proposed System
	Execution Time(ms)	**Execution Time(ms)2**
10	13	12
15	19	17
20	26	19
25	29	23
50	32	25

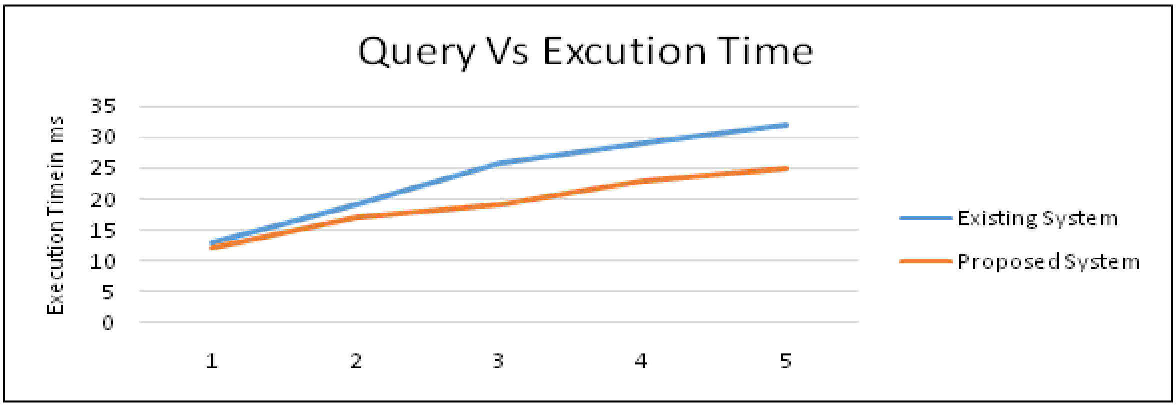

Fig.11Number of queries versus execution time.

5. CONCLUSION

In this paper, a novel Information Retrieval (IR) framework is proposed for attaining efficient data access on the internet. The framework includes semantic keyword extraction and indexing, and impact score estimation for extracting the keywords and determining the importance of the keyword in each document respectively. Then, novel similarity estimation algorithm is proposed for clustering the documents based on the attained score. In the proposed method, the keywords specified in the query and similarity score are utilized to rank the documents. Besides, a novel privacy preservation algorithm is used to maintain the privacy of the querying users by using their personal information. The simulation results demonstrate that the semantic keyword retrieval performance of the proposed IR technique outperforms other existing approaches in terms of average precision, recall, mean average precision and execution time. In future research, the performance of the proposed Information Retrieval (IR) framework is to be further improved by using metaheuristic algorithms.

REFERENCES

[1] Sy, M. F., Ranwez, S., Montmain, J., Regnault, A., Crampes, M., &Ranwez, V. (2012). User centered and ontology based information retrieval system for life sciences. BMC bioinformatics, 13(Suppl 1), S4.

[2] Sagayam, R., Srinivasan, S., &Roshni, S. (2012). A survey of text mining: Retrieval, extraction and indexing techniques. International Journal of Computational Engineering Research, 2(5).

[3] Wu, Q., Burges, C. J., Svore, K. M., &Gao, J. (2010). Adapting boosting for information retrieval measures. Information Retrieval, 13(3), 254-270.

[4] Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information processing & management, 24(5), 513-523.

[5] Carpineto, C., & Romano, G. (2012). A survey of automatic query expansion in information retrieval. ACM Computing Surveys (CSUR), 44(1), 1.

[6] Roy, D., Paul, D., Mitra, M., &Garain, U. (2016). Using word embeddings for automatic query expansion. arXiv preprint arXiv:1606.07608.

[7] Gan, L., & Hong, H. (2015). Improving query expansion for information retrieval using wikipedia. International Journal of Database Theory and Application, 8(3), 27-40.

[8] Cao, G., Nie, J. Y., Gao, J., & Robertson, S. (2008, July). Selecting good expansion terms for pseudo-relevance feedback. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (pp. 243-250). ACM.

[9] Lavrenko, V., & Croft, W. B. (2001, September). Relevance based language models. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 120-127). ACM.

[10] Gao, J., &Nie, J. Y. (2012, October). Towards concept-based translation models using search logs for query expansion. In Proceedings of the 21st ACM international conference on Information and knowledge management (p. 1). ACM.

[11] Riezler, S., & Liu, Y. (2010). Query rewriting using monolingual statistical machine translation. Computational Linguistics, 36(3), 569-582.

[12] Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. Foundations and Trends® in Information Retrieval, 3(4), 333-389.

[13] Imhof, M., &Braschler, M. (2017). A study of untrained models for multimodal information retrieval. Information Retrieval Journal, 1-26.

[14] Büttcher, S., Clarke, C. L., &Lushman, B. (2006, August). Term proximity scoring for ad-hoc retrieval on very large text collections. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 621-622). ACM.

[15] He, B., Huang, J. X., & Zhou, X. (2011). Modeling term proximity for probabilistic information retrieval models. Information Sciences, 181(14), 3017-3031.

[16] Van Rijsbergen, C. J. (1977). A theoretical basis for the use of co-occurrence data in information retrieval. Journal of documentation, 33(2), 106-119.

[17] Singh, J., &Sharan, A. (2015, February). Co-occurrence and Semantic Similarity Based Hybrid Approach for Improving Automatic Query Expansion in Information Retrieval. In ICDCIT (pp. 415-418).

[18] Robertson, S. E. (1990). On term selection for query expansion. Journal of documentation, 46(4), 359-364.

[19] Carpineto, C., & Romano, G. (2012). A survey of automatic query expansion in information retrieval. ACM Computing Surveys (CSUR), 44(1), 1.

[20] Bigi, B. (2003, April). Using Kullback-Leibler distance for text categorization. In European Conference on Information Retrieval (pp. 305-319). Springer, Berlin, Heidelberg.

[21] Pérez-Agüera, J. R., & Araujo, L. (2008). Comparing and combining methods for automatic query expansion. arXiv preprint arXiv:0804.2057.

[22] Shaw, J. A., & Fox, E. A. (1995). Combination of multiple searches. NIST SPECIAL PUBLICATION SP, 105-105.

[23] Wei, Z., Gao, W., El-Ganainy, T., Magdy, W., & Wong, K. F. (2014, July). Ranking model selection and fusion for effective microblog search. In Proceedings of the first international workshop on Social media retrieval and analysis (pp. 21-26). ACM.

[24] Singh, J., &Sharan, A. (2017). Rank fusion and semantic genetic notion based automatic query expansion model. Swarm and Evolutionary Computation.

[25] Huang, G., Wang, S., & Zhang, X. (2011). Query expansion based on associated semantic space. Journal of Computers, 6(2), 172-177.

[26] Prieto-Diaz, R., &Arango, G. (1991). Domain analysis and software systems modeling. IEEE Computer Society Press.