International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 13 (2015)

Bio-Document Retrieval System With Re-Ranking And Cross Ontology

[1]D. Jayasri and [2]Dr. D. Manimegalai

[1] *Associate Professor and Head, Department of Mathematics*
ULTRA College of Engineering & Technology for Women,
Madurai,625104 Tamilnadu, India
[2] *Professor and Head, Department of Information Technology,*
National Engineering College, Kovilpatti. 628503. Tamilnadu, India
E-mail: [1] jaysri6@yahoo.co.in.

ABSTRACT
Nowadays, quickly growing volume of publications in the biomedical, finding associated work is an ever more difficult task. Common solutions to the document search problems are hard since biomedical science is very different for the articles most related to the readers the relevancy may differ. A competent biomedical document retrieval (BDR) system for user-defined queries is frequently necessary to the developing body of research on Biomedical Text Mining, which goals at automatically classifying valued information. A competent bio-medical document retrieval system with the proposed Re-ranking model (RRM) is suggested in this document. In RRM, to compute the similarity between two documents regarding a feature, a measure is presented, named SMTP, which takes the next three cases into account: i) The feature appears in both documents, ii) the feature appears in only one document, and iii) the feature appears in none of the documents. The suggested system runs with six main processes, which comprises 1) dataset preparation, 2) Preprocessing, 3) Feature extraction, 4) Indexing, 5) Re-ranking model (RRM) 6) cross ontology measure and 7) Retrieval process. Finally, the related documents are recovered from the document repository by means of the matching result. The simulation effects display that the presentation attained by the suggested BDR system is better than that attained by other BDR system in terms of accuracy, recall and F-measure.

Keywords: Biomedical document retrieval, Re-ranking model, similarity measure, Word Net, MeSH

1. INTRODUCTION

Data mining is the process of automating information (knowledge) discovery. In Databases (KDD) the process of getting high-level knowledge from low-level data [1] is Knowledge Discovery. The common functions in current data mining practice include Classification, Regression, Clustering, Rule generation, Discovering association rules, summarization, dependency modeling, information retrieval and sequence study. Among this Document retrieval system also one of the data mining techniques [2]. The technology of automatic document retrieval is growing -up and may present a solution to the information overload problem [3, 4]. Presently, document retrieval executes a very important position in information retrieval (IR). With a huge volume of documents, presenting the user with a summary of each document greatly makes easy the task of discovering the wanted documents [5]. The objective of document retrieval systems is to revisit the appropriate documents to a user based on their query, where the query is a collection of keywords. A document is considered as related when its substance is linked to the query [6]. The document that has to be acknowledged should be taken as a query in order to compare it with several or all of the document data in the digital library. Using the different attributes of the documents the firmness or matching among the query document and each document in the digital library are carried out.

Document Retrieval is a computerized process, which formulate an importance ranked list of documents according to an inquisitor's application by comparing their demand to an automatically formed index of the documents in the system. Presently, each one is using such systems in the form of web-based search engines. The three major parts of document retrieval system [7] are the document processor, query analyzer, and matching function. To expand the presentation of IR systems Word Net [8, 9] has been used in many capacities. Word Net can be used to solve the research problems in IR. An ontology based representation has been recently proposed [10] to overcome the faults of term-based representation that is found in the conventional IR approaches, which employs the hierarchical is-a relation among ideas, that is., the meanings of words. Alternatively, measures performed using difficult or aggregated objects in ontology are unusual, they are essential for semantic web applications [11, 12, and 13]. Ontology-based similarity measure has more than a few benefits as compared to other measures: i) ontology is physically created by human beings for a domain and so it was more accurate; ii) it is much more computational proficient when compared to other techniques such as latent semantic indexing; iii) it helps to include domain knowledge into the data mining process. Frequently, comparing two terms in a document employing ontology information employs the truth that their associated ideas in the ontology normally include the properties in the form of features, level of generality or specificity, and their relationships with other ideas [14, 15, and 16].

Additionally, document retrieval refers to finding associated documents for a given user's query. A user's query can be arranged from a full explanation of a document to a few keywords. A good number of the broadly used retrieval approaches are keywords based searching methods [17].

Another type of document retrieval is to use a query context by applying language modeling, to put together many contextual factors so that document ranking will be adapted to the exact query contexts [18]. Using an entire document as a query implements well in improving retrieval precision, on the other hand it is more computationally persisting compared with the keywords based method [19]. Offered document retrieval systems use statistical methods [20] and natural language processing (NLP) [21] approaches joined with different document representation and query structures. Document retrieval [22] has produced several interests in the information retrieval (IR) community. Additionally, retrieval by classical information retrieval models (e.g., Vector Space, Probabilistic, Boolean) [23] is based on lexicographic term matching. Alternatively, two terms can be semantically associated (e.g., can be synonyms or have similar meaning) although they are lexicographically different. Therefore, retrieval by classical retrieval methods will not exceed to retrieve documents with semantically associated terms. Many IR process models, such the Boolean [24], the vector space [25] and the probabilistic models [26] have been proposed to cover the activities and technical user queries with storage and recovery of information items from unstructured sources.

2. REVIEW OF RELATED WORKS

In the literature, there are formerly several benchmarking tools which standardize the procedure of obtaining back the documents using dissimilar techniques. Some of the novel points of reference works are symbolized here. Comfort T. Akinribido *et al.* [27] have presented a fuzzy-ontology based information retrieval system that detect the semantic equivalence among terms in a query and terms in a document by attaching the synonyms of query terms with those of document terms. As a result, documents could be recuperated based on the meaning of query terms. The argument had been that surface form did not sufficiently retrieve associated documents to user's query. Alternatively, the results presented demonstrated that the Fuzzy-Ontology Information Retrieval system successfully recuperated associated documents to user's query. This was irrespective of different meaning and varieties of domain. With different meanings the System was tested on words and some set of user's query from different domains.

A novel feature based approach was suggested by David Sanchez *et al*. [28]. They have learned and gathered the bulk of the ontology-based methodologies formulated with a particular end goal further bolstering calculate their good chance and limitations and stare at their normal execution both from hypothetical and positive viewpoints. They have drawn and made cleared the greater part of the ontology-based methodologies produced keeping in mind the end goal further bolstering calculate their good chance and limits and examine their normal execution both from hypothetical and handy sights. His assess just relies on taxonomic ontological learning, insufficient of corpora-reliance or parameter tuning. The semantic similarity was not precise comparing with substitute works.

Assessing measure of semantic similarity and relatedness to disambiguate text in biomedical text was proposed by Bridget T. Mclnnes and Ted Pedersen [29]. In this approach they have produced a system that has disambiguate terms in biomedical content by using wrongly similarity and relatedness data extrapolated from UMLS and calculated the sufficiency of similarity and relatedness measures. They have employed Umls::sense relate on the biomedical dataset (MSH-WSD) that enclose 203 questionable term and acronyms. This methodology demonstrates that the data substance based measure discovered from either a corpus or scientific categorization get higher disambiguation precision than way based measure or relatedness measure on the MSH-WSD dataset. The objective of the most genuine sense is uttered by a simple averaging of the weighted similarity scores of the neighboring terms. Therefore for aggregating vague information with vague weights is unfeasible.

Shi-Jay Chen and Hung-Chin Chu [30] have proposed an extended fuzzy concept networks based approach for fuzzy query processing of document retrieval. In order to suggest the extended fuzzy concept networks an importance matrix and relation matrix have been employed. Currently, a satisfaction matrix has been reached by the proposed approach by linking the document descriptor relevance matrix described by the capable with the user's query descriptor based on varied weights. Next, an AND operator of the quadratic-mean averaging operators has been used for calculating all the elements in each row of the satisfaction matrix. Finally, the user required associated documents has been reached by grading the degrees of fulfillment of each satisfaction matrix.

Semantic similarity of short text in language with a poor natural language processing support was distributed by Bojan furlan et al [31]. They have symbolized an approach, which was used to form a product framework that chooses the semantic similarity of two given short messages called as the Linstss. The main characteristic of their proposed framework building approach was its uniqueness that permitted other asset to competently adapt to the purpose of their own language. For semantic information, the precision of co-occurrence lattice was small in light of the more modest content corpus. This methodology is helpful for dialects where other option framework building techniques are not applicable.

Dolf Trieschnigg *et al*. [32] have proposed An Effective MeSH Text Classification for Improved Document Retrieval for Controlled vocabularies such as the Medical Subject Headings (MeSH) thesaurus and the Gene Ontology (GO) present a proficient way of accessing and organizing biomedical information by diminishing the ambiguity intrinsic to free-text data. Different methods of automating the task of MeSH ideas have been proposed to hold up manual annotation, on the other hand they are either restricted to a small subset of MeSH or have only been contrasted to a limited number of other systems. They compared the appearance of 6 MeSH classification systems (MetaMap, EAGL, a language and a vector space model based approach, a K-Nearest Neighbor approach and MTI) in terms of generating and going with manual MeSH annotations. A K Nearest Neighbor system clearly outperforms the other circulated approaches and scales well with vast amounts of text using the total MeSH thesaurus. They also demonstrated that a statistically significant improvement can be reached in information retrieval (IR) when the text of a user's query is regularly interpreted with MeSH ideas, compared to using the unique textual query merely.

3. PROBLEM IDENTIFICATION

The amount of information on the internet has always been on the rise and hence, it calls for an information retrieval system which has the ability to facilitate the user in achieving the useful information. Most importantly, an information retrieval system should have ability to facilitate the user to achieve the useful information they requires. Informational retrieval has been applied to large areas especially in biomedical research. To the input query, normally a system outputs extensive results in a non-indexed format and the retrieved results will also feature many non-interesting documents. But the existing and traditional key word based search techniques have the drawback of retrieving false information and struggle to work on the large volume of information available. Due to the hugeness of data, a lot of time gets wasted for the user for browsing the Internet as well as searching for the information they needs. This makes the tasks of searching, accessing, displaying, integrating and preserving the data more difficult. The retrieval process often comes up with undesired information and the users are discontented with the low precision and recall. There has been large research works put into overcoming this scenario but fails to find the complete answer. Semantic web, information knowledge management systems and ontology has been applied in recent years to bring more of user desired results. As compared to other measures, ontology-based similarity measure has some benefits like ontology is manually formed by human beings for a domain and so it was more exact. It is also more computational efficient and also helps to include domain knowledge into the data mining process.

4. A SYSTEM FOR BIO-DOCUMENT RETRIEVAL (BDR)

The quantity of information on the internet has permanently been on the increase and therefore, it demands for an information retrieval system which has the capacity to make easy the user in attaining the beneficial information. Informational retrieval has been employed to huge areas particularly in biomedical research. To the input probe, generally a system yields general effects in a non-indexed format and the recovered effects will as well feature several non-interesting documents. In our earlier work, according to this, a competent bio-medical document retrieval system [33] with the cross-ontology based semantic similarity measure was proposed. The technique applied the WordNet and MeSH ontologies for matching the input query keyword and as well planned a new cross-ontology based semantic similarity measure for the query keywords. The system contended with three main processes, which comprised 1) Extracting features from the documents based on TF-IDF similarity, 2) Indexing of documents by Rabin Fingerprint algorithm and 3) Retrieving the relevant documents based on distance measure. Lastly, by means of the matching result the related documents were recovered from the document repository. To develop the retrieval process, we plan a system namely Re-ranking model (RRM), in which a competent measure is presented to measure the similarity among two documents. Generally, the suggested system contains three phases like DPFI, RRM, and BDR. The phase 1 contain four steps comprises data preparation, pre-processing, feature extraction and indexing. In phase 2 the suggested system named as Re-ranking model (RRM) is comprised. The phase 3 contains cross ontology measure and retrieval process. In figure 1 the general architecture of the suggested BDR system is demonstrated.

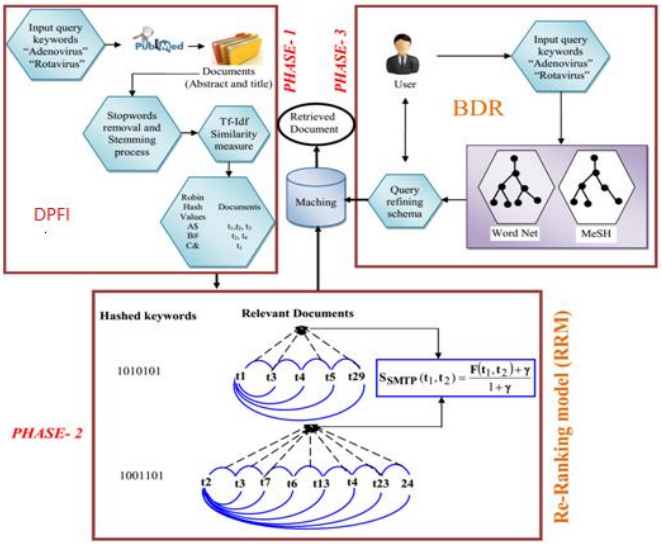

Figure 1: Architecture of the proposed BDR system

4.1 PHASE 1: DPFI PROCESS

A. Data sets and tools for BDR system

In this paper, the database is collected from Pubmed search engine [34, 35] to evaluate the proposed system. In terms of information retrieval systems, PubMed is one of the best known biomedical databases and it contains more than 20 million citations on biomedical articles from MEDLINE and life science journals, which provides a convenient web-based search portal for users as well as an application program interface for developers. PubMed is a free search engine to search about medicine and biomedical journal literature. It searches several databases and interfaces Medline, directly. This search engine maps user's search terms to the Medical subject heading (Mesh) and text words in Medline records and then searching.

B. Pre-processing

Pre-processing stage contains three steps, as demonstrated in figure 2. As exposed in figure 2, a document or documents may be specified as input of this stage which is concatenated into a full text. The combined full text turns out to be the goal for the tokenization. By a white space or a punctuation mark the full text is tokenized into tokens. Hence, the output of this step is a list of tokens.

As demonstrated in figure 2 the following step to the concatenation & tokenization is the stemming & exception handling. In this stage, each token is changed into its root form. Rules of stemming and exception handling are saved into a file before doing that. When the program encoding documents is stimulated, the rules are loaded into memory and the analogous rules are used to each token. The production of this stage is a list of tokens changed into their root forms. The previous step

of abstracting feature candidates from a corpus is to eliminate stop words as demonstrated in figure 2. This stage can as well be the second stage i.e. either we do stop words elimination first then stemming or stemming tracked by stop word removal. Now, stop words are described as words which function only grammatically without their significance to content of their document; articles (a, an, or the), prepositions (in, on, into, or at), pronoun (he, she, I, or me), and conjunctions (and, or, but, and so on) belong to this kind of words. For more proficient processing it is essential to eliminate this kind of words. Frequencies of left over words are counted after removing stop words. Hence, a list of the left over words and their frequencies is produced as the last output from the process demonstrated in figure 2.

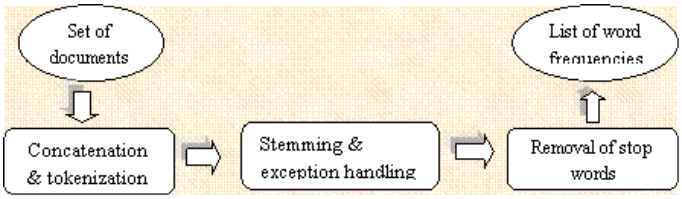

Figure 2: Steps in pre-processing

B. Feature extraction

We find the similarity measure of all keywords abstracted from the document repository after the pre-processing steps. The similarity measure we have employed here is TF-IDF similarity as specified in equation 1. After that, based on the similarity measure, we have taken the set of keywords with maximum score.

***TF-IDF similarity measure*:**

The word frequency–inverse document frequency (TF–IDF) is a weight frequently used in information retrieval and text mining. This TF-IDF weight is a mathematical measure employed to compute how dynamic a word is to a document in a group or corpus. The significance rises uniformly to the number of times a word happens in the document however is equalized by the frequency of the word in the collection. Variations of the TF–IDF weighting scheme are frequently employed by search engines as an important tool in scoring and ranking a document's significance specified a user query. TF–IDF can be competently employed for stop-words filtering in diverse subject areas such as text summarization, classification etc. [37]. By means of the TF-IDF weighting scheme [36], t_x is explained as,

$$t_x = (TF_{t,x}) * (IDF_x) \tag{1}$$

Where $TF_{t,x}$ is the number of times that term x happens in the document symbolized by t, $IDF_x = N/n_x$, N is the total number of documents in the database, and n_x is the total number of documents in the database that have the term x.

The document repository T consists of a set of bio-medical documents based on the input query keyword.

$$T = \{t_1, t_2, ..., t_n\} \tag{2}$$

Each document comprise of set of extracted keywords K by completing the pre-processing steps.

$$t = \{k_1, k_2, ..., k_n\} \tag{3}$$

Subsequently, the TF-IDF similarity measure is computed for all the extracted keywords. Then, sort the keywords based on their corresponding similarity measures. The similarity measure with the highest score is considered to be the significant features I_f from the corresponding documents.

$$S = t_x(K) \tag{4}$$

$$I_f = S > \min\_threshold \tag{5}$$

D. Indexing

By indexing the documents and formulating the queries the document retrieval system arranges for retrieval, effecting in document representations and query representations correspondingly. Automatic indexing initiates with raw feature extraction, such as removing all the words from a text, tracked by refinements in accordance with the conceptual schema. Now, the indexing is prepared with the help of the Rabin's Fingerprint hashing algorithm so that the matching process can be completed without difficulty. Undertake that the character string B is a bit string having m bits $[b_1, \cdots, b_m]$ and it is related to a polynomial of degree $(m-1)$ in indeterminate t as follows [38]:

$$B(y) = b_1 y^{m-1} + b_2 y^{m-2} + ... + b_{m-1} y + b_m \tag{6}$$

Then, a polynomial $P(y)$ of degree k is represented as,

$$P(y) = a_1 y^k + a_2 y^{k-1} + \cdots + a_{k-1} y + a_k \tag{7}$$

In Rabin's fingerprinting technique, an irreducible polynomial is used for $P(y)$. As we are dealing with bit strings, all the coefficients of $A(y)$ are in Z_2. Hence $P(y)$ will be selected by using $a_i's$ in Z_2. Then, the fingerprinting function for a given character string B is defined as,

$$f(B) = B(y) \bmod P(y) \tag{8}$$

We have computed the fingerprint value F_{val} for the chosen features I_f of the documents with the similarity measure by means of the Rabin's fingerprint algorithm.

$$F_{val} = Hash(I_f)_{Rabin} \quad (9)$$

Consequently, the indexing process of the documents is performed based on the fingerprint values of every feature sets. Now, we have employed the inverted index method. An inverted index is an indexing information structure storing a mapping from content, such as words or numbers, to its positions in a database file, or in a document or a set of documents. The determination of an inverted index is to allow fast full text searches, at a cost of enlarged processing when a document is increased to the database. The capsized file may be the database file itself, rather than its index. In this, every feature is chosen and all the consequent documents containing the specific features are being indexed I_T. For the meantime, the keywords are indexed by their own mess values F_{val}.

$$I_T = \{ \in \} hash(I_f) \quad (10)$$

The sample hashing process is given in the following table 1.

Table 1: Indexing of documents

Hashed keywords	Relevant Documents
1010101	t1, t3, t4, t5, t29
1001101	t2, t3, t7, t6, t13, t4, t23, t24
1011100	t8, t10, t17, t18
1000001	t1, t21, t30, t19, t11, t27
1111101	t26, t22, t18, t16,t28

4.2 PHASE 2: RRM PROCESS

A. Re-ranking model (RRM)

Once the indexing process is constructed via Rabin hash algorithm, the indexed documents are re-ranked to improve the retrieval results using the RRM model. This section explains how to measure the similarity between two documents in RRM. In RRB model, a similarity measure is presented namely SMTP (similarity measure for text processing) for the two documents. Several characteristics are embedded in this measure. It is a symmetric measure. The following properties, among other ones, are preferable for a similarity measure between two documents:

- The presence or absence of a feature is further important than the difference between the two values associated with a present feature. Consider two features f_i and f_j and two documents t_1 and t_2. Suppose f_i does not appear in t_1 but it appears in t_2. Then f_i is considered to have no relationship with t_1 while it has some relationship with t_2 .In this case, t_1 and t_2 are dissimilar in terms of f_i If f_j appears in both t_1 and t_2. Then f_j has some relationship with t_1 and t_2 simultaneously. In this case, t_1 and t_2 are similar to some degree in terms of f_j. For the above two cases, it is reasonable to say that f_i carries more weight than f_j in determining the similarity degree between t_1 and t_2 .For exam-ple, assume that f_i is absent in t_1, i.e., $t_{1i} = 0$, but appears in t_2, e.g., $t_{2j} = 2$, and f_j appears both in t_1 and t_2, e.g., $t_{1i} = 3$ and $t_{2j} = 5$. Then f_i is considered to be more essential than f_j in determining the similarity between t_1 and t_2, although the differences of the feature values in both cases are the same.
- The similarity degree should increase when the difference between two non-zero values of a specific feature decreases. For example, the similarity involved with $t_{13} = 2$ and $t_{23} = 20$ should be smaller than that involved with $t_{13} = 2$ and $t_{23} = 3$.
- The similarity degree should decrease when the number of presence-absence features increases. For a presence-absence feature of t_1 and t_2, t_1 and t_2 are dissimilar in terms of this feature as commented earlier. Therefore, as the number of presence-absence features increases, the dissimilarity between t_1 and t_2 increases and thus the similarity decreases. For example, the similarity between the documents $<1,0,1>$ and $<1,1,0>$ should be smaller than that between the documents $<1,0,1>$ and $<1,0,0>$.
- Two documents are least similar to each other if none of the features have non-zero values in both documents. Let $t_1 = <t_{11}, t_{12}, \ldots, t_{1m}>$ and $t_2 = <t_{21}, t_{22}, \ldots, t_{2m}>$. If $t_{1i} t_{2i} = 0$, $t_{1i} + t_{2i} > 0$ for $1 \le i \le m$, then t_1 and t_2 are least similar to each other. As mentioned earlier, t_1 and t_2 are dissimilar in terms of a presence-absence feature. Since all the features are presence-absence features, the dissimilarity reaches the extremity in this case. For example, the two documents $<x,0,y>$ and $<0,z,0>$, with x, y, and z being non-zero numbers, are least similar to each other.
- The similarity measure should be symmetric. That is, the similarity degree between t_1 and t_2 should be the same as that between t_2 and t_1.
- The value distribution of a feature is considered, i.e., the standard deviation of the feature is taken into account, for its contribution to the similarity between two documents. A feature with a larger spread offers more contribution to the similarity between t_1 and t_2.

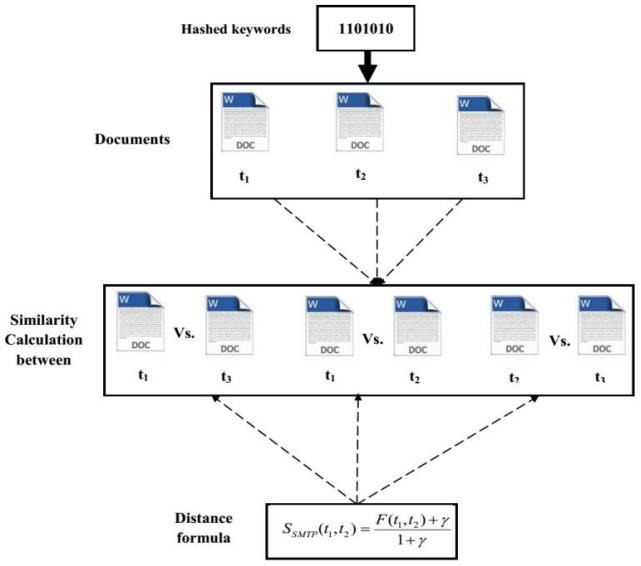

Figure 3: Similarity computation process between documents (after Rabin hash algorithm)

Considering above properties, to compute the similarity between two documents with respect to a feature, the proposed measure takes the following three cases into account: (a) The feature appears in both documents, (b) the feature appears in only one document and (c) the feature appears in none of the documents. Based on the preferable properties mentioned above, we present a similarity measure, S_{SMTP} for t_1 and t_2 is,

$$S_{SMTP}(t_1,t_2)=\frac{F(t_1,t_2)+\gamma}{1+\gamma} \quad (11)$$

Where, F is the function, which is defined for the two documents $t_1 =< t_{11},t_{12},...,t_{1m} >$ and $t_2 =< t_{11},t_{12},...,t_{1m} >$ and as follows:

$$F(t_1,t_2)=\frac{\sum_{j=1}^{l} N*(t_{1j},t_{2j})}{\sum_{j=1}^{l} N_{\cup}(t_{1j},t_{2j})} \quad (12)$$

Where,

$$N*(t_{1j},t_{2j})=\begin{cases} 0.5\left(1+\exp\left\{\left(\frac{t_{1j}-t_{2j}}{\sigma_j}\right)^2\right\}\right) & ,if\ t_{1j}t_{2j}>0 \\ 0 & ,\ if\ t_{1j}=0\ and\ t_{2j}=0 \\ -\gamma & ,\ \ Otherwise \end{cases} \quad (13)$$

$$\sum_{j=1}^{l} N_{\cup}(t_{1j},t_{2j})=\begin{cases} 0 & ,if\ t_{ij}=0\ and\ t_{2j}=0 \\ 1, & otherwise \end{cases} \quad (14)$$

For the first case, we set a lower bound 0.5 and decrease the similarity as the difference between the feature values of the two documents decreases, scaled by a Gaussian function as shown in equation (13). Where, σ_j is the standard deviation of all non-zero values for feature z_j in the training dataset. For the second case, we set a negative constant $-\gamma$ disregarding the magnitude of the non-zero feature value. For the last case, the feature has no contribution to the similarity. After In RRM process, the similarity measurement between two documents is computed as follows:

4.3 PHASE 3: BIO-DOCUMENT RETRIEVAL (BDR) PROCESS

Once RRM process is completed, the testing documents are tested on phase 3. Based on the user query, the system finds the cross ontology similarity measure for the query keywords using the features extracted from the ontologies. In this section, we describe cross ontology measure and retrieval process.

A. Extracting set of relevant definitions, features, synsets, and neighbors from both ontologies

In common, ontologies can be differentiated into domain ontologies, denoting knowledge of a specific domain, and generic ontologies denoting public sense knowledge about the world. There are more than a few examples of common purpose ontologies accessible together with WordNet [39] attempts to perfect the lexical knowledge of a native speaker of English. English nouns, verbs, adjectives, and adverbs are arranged into synonym sets, called synsets, each denoting an idea. In addition, one of the domain specific ontology planned for medical ideas comprises MeSH [40] [41]. Based on the related input query keyword, the set of suitable definitions, features (Hypernyms), synset, neighbors (Hyponyms) are removed from both the ontologies, WordNet and MeSH. The sample XML explanations about the query keywords from both ontologies with the specified bio-medical term are publicized beneath.

Table 2: XML descriptions taken from the Wordnet and MeSH ontology

WordNet: Adenovirus	**MeSH:** Rotavirus
<Term>Adenovirus <Definition>any of a group of viruses including those that in humans cause upper respiratory infections or infectious pinkeye,</Definition> <Synset>adenovirus,</Synset> <Hypernyms>animal_virus,,</Hypernyms> <Hyponyms>parainfluenza_virus,,</Hyponyms> </Term>	<Term>rotavirus enteritis <Definition>A viral infectious disease that results_in inflammation located_in stomach and located_in intestine, has_material_basis_in Rotavirus, which is transmitted_by ingestion of contaminated food or water, or transmitted_by fomites. The infection has_symptom fever, has_symptom vomiting, has_symptom diarrhea, and has_symptom abdominal pain.</Definition> <Synset>rotavirus enteritis, Enteritis due to rotavirus (disorder),</Synset> <Hypernyms>Nil</Hypernyms> <Hyponyms>rotavirus enteritis</Hyponyms> </Term>

B. Finding cross ontology measure for the input query

We have discovered the semantic similarity measures of the abstracted feature sets, synsets, neighborhoods and the definitions of the two diverse ontologies in order to discover the cross ontology measure for the input query. The similarity between two different terms is calculated as a weighted sum of similarities among synonym sets (synsets), features, neighborhoods and their definitions. Reflect the WordNet O_1 and MeSH O_2 ontologies, in which the Query keyword Q contains Features F, Synsets S, Neighborhoods N and Definitions D attained from both the ontologies. Moreover, we have associated all the selected features together in a vector named as A_s. Based on the input query, we have to discover the cross ontology measure for every set of features, synsets, neighborhoods and definitions attained from the ontologies. The set of features, synsets, neighborhoods and definitions attained from the ontologies O_1 and O_2 are symbolized as follows,

$$F = \left\{ f_i^{(1)}, f_i^{(2)} \right\} f_i^{(1)} \in O_1, f_i^{(2)} \in O_2 \;; 1 \le i \le m \quad (15)$$

$$S = \left\{ s_i^{(1)}, s_i^{(2)} \right\} s_i^{(1)} \in O_1, s_i^{(2)} \in O_2 \;; 1 \le i \le m \quad (16)$$

$$N = \left\{ n_i^{(1)}, n_i^{(2)} \right\} n_i^{(1)} \in O_1, n_i^{(2)} \in O_2 \;; 1 \le i \le m \quad (17)$$

$$D = \left\{ d_i^{(1)}, d_i^{(2)} \right\} d_i^{(1)} \in O_1, d_i^{(2)} \in O_2 \;; 1 \le i \le m \quad (18)$$

$$A_s = \{ F, S, N, D \} \quad (19)$$

With the help of the set of features, synsets, neighborhoods and the definitions abstracted from both the ontologies the similarity measure $Sim(K_1, K_2)$ of the input query keywords K_1 and K_2 from ontologies O_1 and O_2 correspondingly is calculated. The formula used for computing the similarity measure of the analogous query keyword from the Wordnet and MeSH is specified as follows,

$$Sim(K_1, K_2) = \sqrt{\frac{\alpha S_f^2(K_1, K_2) + \beta S_s^2(K_1, K_2) + \gamma S_n^2(K_1, K_2) + \delta S_d^s(K_1, K_2)}{4}} \quad (20)$$

Where, $\alpha, \beta, \gamma, \delta$ are the set of the similarity parameters and these parameters are recognized as belows:

$$\alpha = \frac{\left| f^{(1)} \cap f^{(2)} \right| + \left| \sim A_s^{(1)} \cap \cup A_s^{(2)} \right|}{\left(f^{(1)} \cap f^{(2)} \right) + \left(s^{(1)} \cap s^{(2)} \right) + \left(n^{(1)} \cap n^{(2)} \right) + \left(d^{(1)} \cap d^{(2)} \right) + \left(\sim A_s^{(1)} \cap \cup A_s^{(2)} \right)} \quad (21)$$

$$\beta = \frac{\left| s^{(1)} \cap s^{(2)} \right| + \left| \sim A_s^{(1)} \cap \cup A_s^{(2)} \right|}{\left(f^{(1)} \cap f^{(2)} \right) + \left(s^{(1)} \cap s^{(2)} \right) + \left(n^{(1)} \cap n^{(2)} \right) + \left(d^{(1)} \cap d^{(2)} \right) + \left(\sim A_s^{(1)} \cap \cup A_s^{(2)} \right)} \quad (22)$$

$$\gamma = \frac{\left| n^{(1)} \cap n^{(2)} \right| + \left| \cup A_s^{(1)} \cap \cup A_s^{(2)} \right|}{\left(f^{(1)} \cap f^{(2)} \right) + \left(s^{(1)} \cap s^{(2)} \right) + \left(n^{(1)} \cap n^{(2)} \right) + \left(d^{(1)} \cap d^{(2)} \right) + \left(\sim A_s^{(1)} \cap \cup A_s^{(2)} \right)} \quad (23)$$

$$\delta = \frac{\left| d^{(1)} \cap d^{(2)} \right| + \left| \cup A_s^{(1)} \cap \cup A_s^{(2)} \right|}{\left(f^{(1)} \cap f^{(2)} \right) + \left(s^{(1)} \cap s^{(2)} \right) + \left(n^{(1)} \cap n^{(2)} \right) + \left(d^{(1)} \cap d^{(2)} \right) + \left(\sim A_s^{(1)} \cap \cup A_s^{(2)} \right)} \quad (24)$$

Likewise, $S_f(K_1, K_2)$, $S_S(K_1, K_2)$, $S_n(K_1, K_2)$ and $S_d(K_1, K_2)$ are the individual similarity measures of the every feature set, synsets, neighborhoods and definitions correspondingly. Now, the formula for discovering the similarity of every set of terms using their communal universal set of all terms with features, synsets, neighborhoods and the definitions is specified in detail.

$$S_f(K_1, K_2) = \left(\frac{\left| f^{(1)} \cap f^{(2)} \right|}{\left| f^{(1)} \right| * \left| f^{(2)} \right|} \right) + \left(\frac{\left| \sim f^{(1)} \cap \sim f^{(2)} \right|}{\left| \sim f^{(1)} \right| * \left| \sim f^{(2)} \right|} \right) - \left(\frac{\left| f^{(1)} \cap \sim f^{(2)} \right|}{\left| f^{(1)} \right| * \left| \sim f^{(2)} \right|} \right) - \left(\frac{\left| \sim f^{(1)} \cap f^{(2)} \right|}{\left| \sim f^{(1)} \right| * \left| f^{(2)} \right|} \right) \quad (25)$$

$$S_s(K_1,K_2)=\left(\frac{\left|s^{(1)}\cap s^{(2)}\right|}{\left|s^{(1)}\right|*\left|s^{(2)}\right|}\right)+\left(\frac{\left|\sim s^{(1)}\cap\sim s^{(2)}\right|}{\left|\sim s^{(1)}\right|*\sim\left|s^{(2)}\right|}\right)-\left(\frac{\left|s^{(1)}\cap\sim s^{(2)}\right|}{\left|s^{(1)}\right|*\sim\left|s^{(2)}\right|}\right)-\left(\frac{\left|\sim s^{(1)}\cap s^{(2)}\right|}{\left|\sim s^{(1)}\right|*\left|s^{(2)}\right|}\right) \quad (26)$$

$$S_n(K_1,K_2)=\left(\frac{\left|n^{(1)}\cap s^{(2)}\right|}{\left|n^{(1)}\right|*\left|s^{(2)}\right|}\right)+\left(\frac{\left|\sim n^{(1)}\cap\sim n^{(2)}\right|}{\left|\sim n^{(1)}\right|*\sim\left|n^{(2)}\right|}\right)-\left(\frac{\left|n^{(1)}\cap\sim n^{(2)}\right|}{\left|n^{(1)}\right|*\sim\left|n^{(2)}\right|}\right)-\left(\frac{\left|\sim n^{(1)}\cap n^{(2)}\right|}{\left|\sim n^{(1)}\right|*\left|n^{(2)}\right|}\right) \quad (27)$$

$$S_d(K_1,K_2)=\left(\frac{\left|d^{(1)}\cap d^{(2)}\right|}{\left|d^{(1)}\right|*\left|d^{(2)}\right|}\right)+\left(\frac{\left|\sim d^{(1)}\cap\sim d^{(2)}\right|}{\left|\sim d^{(1)}\right|*\sim\left|d^{(2)}\right|}\right)-\left(\frac{\left|d^{(1)}\cap\sim d^{(2)}\right|}{\left|d^{(1)}\right|*\sim\left|d^{(2)}\right|}\right)-\left(\frac{\left|\sim d^{(1)}\cap d^{(2)}\right|}{\left|\sim d^{(1)}\right|*\left|d^{(2)}\right|}\right) \quad (28)$$

C. Retrieval process

The features of the input query words, $A_S=\{F,S,N,D\}$ is attained from WordNet and MeSH ontologies when the user offers the input query keywords. After that, the system discovers the cross ontology similarity measure for the query keywords by means of the features abstracted from the ontologies. The query refining process is prepared if the similarity measure is less than the user indicated threshold, means that the user have to verify or give another related keywords. The input query is messed and matched with the indexed document's hash values if the similarity measure is beyond the user indicated threshold. If the hash value in the repository matches with the hash value of the input keyword, then we can recover the necessary number of bio-documents related to the query keyword.

Algorithm: Proposed BDR system

Input : Query keyword, $k_1 \in O_1$;
Query keyword, $k_2 \in O_2$

Assumptions:

$F \rightarrow$ *Features*
$S \rightarrow$ *Synsets*
$N \rightarrow$ *Neighborhoods*
$D \rightarrow$ *Definitions*
$S_m \rightarrow$ *Similarity measure*
$I(H,T)_{RRM} \rightarrow$ *Indexed documents (H refers to hashed keyword and D refers to documents)based Re-ranking Model (RRM)*
$S_{SMTP} \rightarrow$ Similarity measure in RRM process

Process :

Begin
Get query, $K = k_1, k_2$
obtain feature vectors, $F, S, N, D \in K$
$S_m = Sim(K_1, K_2)$
if $S_m < thresh$
promt the user to check K
else
hashing, $H(k_1)$ *and* $H(k_2)$ *based* Re–*ranking* mod
Compute similarity score S_{SMTP} b/w $T_1(H(k_1)$ *and*
If $T_1 > Th$
$R_T << I(H,T)_{RRM} \in H(k_1)$
end if
If $T_2 > Th$
$R_T << I(H,T)_{RRM} \in H(k_1)$
end if
end

Output : Relevant document R_T

5. RESULT AND DISCUSSION

We have offered the results of our suggested methodology and have examined their presentation in this part. The proposed bio-document retrieval system is implemented in the JAVA program and the retrieval process is experimented with the Pub Med database and the result is compare with existing techniques [33]. We have implemented our proposed bio-document retrieval system using Java (jdk 1.6) and a series of

experiments were performed on a PC with Windows XP Operating system at 2 GHz dual core PC machine with 2 GB main memory running a 64-bit version of Windows 2007.

5.1 Dataset description

In this paper, we tested our proposed algorithm with different documents we using Pub Med database which is contain the WordNet and MeSH ontology. Pub Med is the National Library of Medicine's pursuit service that gives access to more than 11 million references in MEDLINE. MEDLINE is the head bibliographic database covering the fields of pharmaceutical, nursing, dentistry, veterinary solution, the medicinal services framework, and the preclinical sciences. It contains more than 11 million references and digests from more than 4000 biomedical diaries. From that database, we have picked just 180 medicinal for 11 diverse restorative catchphrases. Too, we have taken 30 archives for each decisive word of both the ontologies. Here we use 14 keywords to keywords to explain the proposed approach. The picked specimen pivotal words of WordNet and MeSH ontologies are indicated in table 3.

Table 3: WordNet and MeSH terms

WordNet	MeSH
Adenovirus	Rotavirus
Anemia	Appendicitis
Pneumonia	Asthma
Carcinoma	Neoplasm
Hypothyroidism	Hyperthyroidism
Pain	Ache
Dementia	Atopic Dermatitis
Malaria	Bacterial Pneumonia
Osteoporosis	Patent Ductus Arteriosus
Sinusitis	Mental Retardation
Iron Deficiency Anemia	Sickle Cell Anemia
Diarrhea	Mucosal
edema	Angioedema
ulcer	Decubitus

5.2 Evaluation metrics

The evaluation of proposed bio-document retrieval system is carried out using the following metrics as suggested by below equations,

Precision:

Precision is the fraction of retrieved documents that are relevant to the search which is given in equation (1).

$$P = \frac{\left|\{relevant\ documents\} \cap \{retrieved\ documents\}\right|}{\left|\{retrieved\ documents\}\right|} \quad (29)$$

Where; P represents precision

Recall:

Recall is the fraction of the documents that are relevant to the query that are successfully retrieved which is given in equation (2).

$$R = \frac{\left|\{relevant\ documents\} \cap \{retrieved\ documents\}\right|}{\left|\{relevant\ documents\}\right|} \quad (30)$$

Where; R represents Recall

F-measure:

F-measure is defined as the harmonic mean of information retrieval precision and recalls metrics which is given in equation (3).

$$F = \frac{2PR}{(P+R)} \quad (31)$$

Where;
P →precision
R →Recall
F →F-measure

5.3 Performance analysis of the proposed approach:

The performance of the proposed document retrieval system is evaluated based on the input query keywords to the WordNet ontology using the Precision, recall and F-measure. Here, we have utilized four query keywords and the corresponding documents are obtained from the document repository. We have analyzed our proposed system with different keywords with the relevant and retrieved documents. It reveals that the proposed system works fine in the document retrieving process. The basic idea of our research is to efficient bio-document retrieval system using robin fingerprint and re-ranking model (RFRRM-BDR). Here, we utilized the Word Net and Mesh ontologies for matching the input query keywords. The performance of the proposed approach carried out by varying the similarity threshold in matching the hashed query keyword with the indexed hash values. The obtained results are used to measure the precision, recall and F-measure values that are plotted as a graph and shown in figure 4, 5 and 6 respectively.

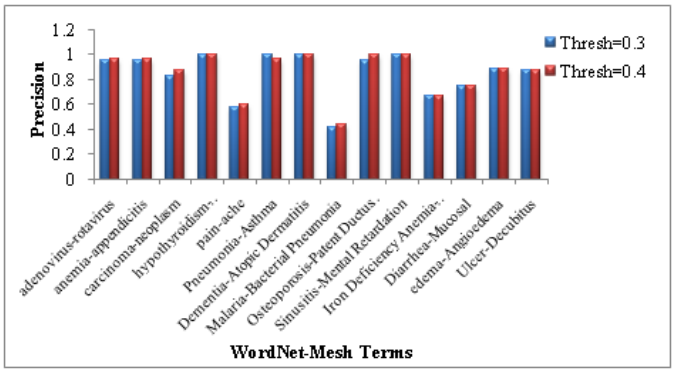

Figure 4: Precision graph plotted for different similarity thresholds

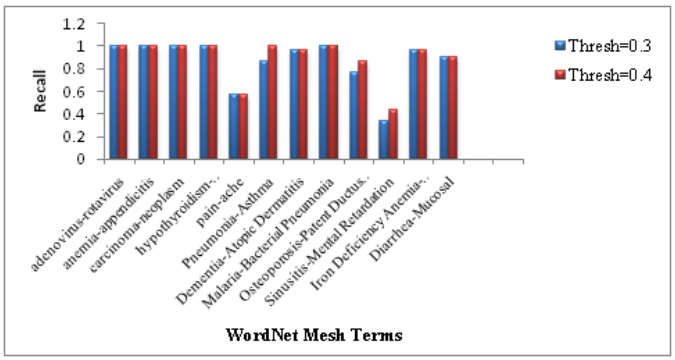

Figure 5: Recall graph plotted for different similarity thresholds

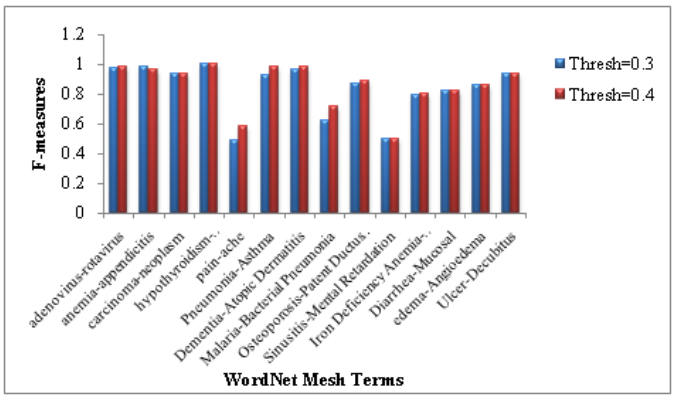

Figure 6: F-measure graph plotted for different similarity thresholds

When analyzing figure 4, the threshold value 0.3 and 0.4 we obtain the maximum precision value for Hypothyroidism, Dementia and Sinusitis WordNet keywords which corresponding MeSH terms are Hyperthyroidism, Atopic Dermatitis and Mental Retardation. Here, we obtain the minimum precision value of 0.4177 for Malaria- Bacterial Pneumonia WordNet Mesh term. Moreover, figure 5 illustrate the performance of the proposed approach using Recall measures. Here, the recall value is mostly similar for the both the threshold values. In figure 6, shows the performance of proposed approach for F-measure values. This graph also shows our proposed approach having the better performance for the threshold value 0.3 and 0.4. Moreover, table 4 lists the obtained values for the evaluation measures with different keywords and the relevant documents. Here we have utilized eleven set of medical keywords and the corresponding medical documents obtained from the PubMed database. We have analyzed our proposed system with different keywords with the relevant and retrieved documents. When analyzing table 4 we obtain the relevant documents as 30. It reveals that the proposed system works fine in the medical document retrieving process.

Table: 4 Precision, Recall and F-measure for different keywords

Query keyword		Relevant documents	Retrieved documents	Precision	Recall	F-measure
WordNet	MeSH					
Adenovirus	Rotavirus	30	31	0.96774193	1	0.983607
Anemia	Appendicitis	30	31	0.96774193	1	0.983607
Carcinoma	Neoplasm	30	38	0.88235294	1	0.9375
Hypothyroidism	Hyperthyroidism	30	30	1	1	1
Pain	Ache	17	28	0.6071428	0.5666	0.586207
Pneumonia	Asthma	30	31	1	0.8666	0.928571
Dementia	Atopic Dermatitis	29	29	1	0.9666	0.983051
Malaria	Bacterial Pneumonia	22	43	0.4477611	1	0.618557
Osteoporosis	Patent Ductus Arteriosus	23	23	0.9583333	0.7666	0.867925
Sinusitis	Mental Retardation	10	10	1	0.3333	0.5
Iron Deficiency Anemia	Sickle Cell Anemia	29	43	0.6744186	0.9666	0.794521
Diarrhea	Mucosal	27	36	0.75	0.9	0.8181
edema	Angioedema	25	28	0.8928	0.833	0.8620
ulcer	Decubitus	30	34	0.8823	1	0.9375

5.4 **Comparative analysis of proposed approach**

In this section, we compare our proposed RFRRM-BDR approach with existing approach using RF-BDR [33]. The bio document retrieval performance of proposed system is assessed using the precision, recall and F-Measure plot. In bio document retrieval, first we segment the document into keywords after that we calculate the similarity measure of the Keywords. Robin fingerprint algorithm is used to indexing the keywords and Re-ranking model is used to select the efficient bio-document from the indexing. Finally the matching process is used to retrieve the bio-document based on the keywords. The performance is made up of two process training and testing process. Depending on the training and texting process, we calculate the precision and that precision value shows the efficiency of our proposed work.

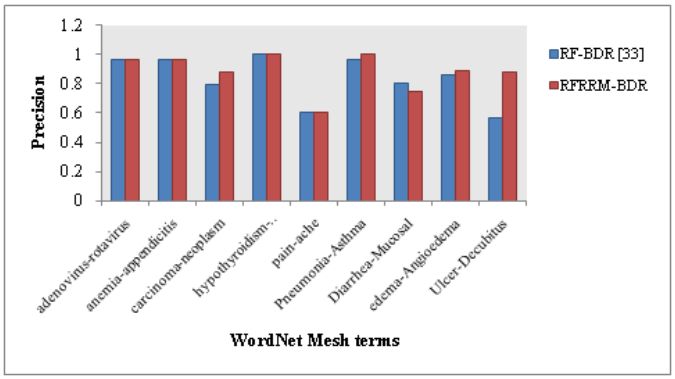

Figure 7: Performance comparison of the precision plot

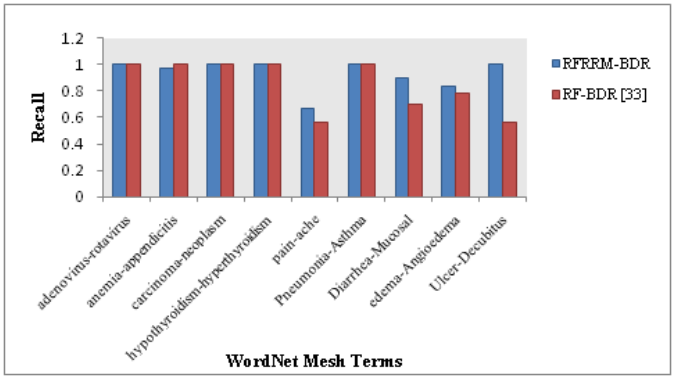

Figure 8: Performance comparison of the Recall plot

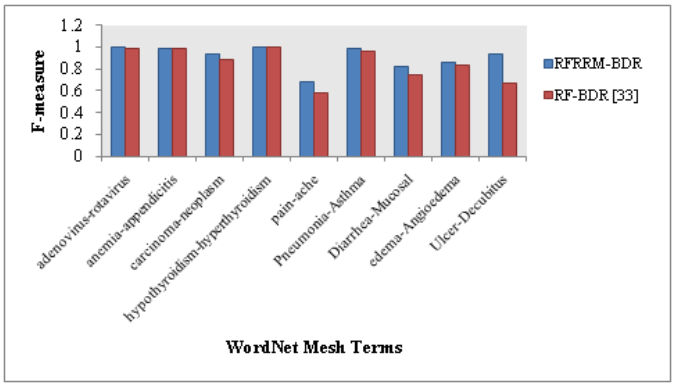

Figure 9: Performance comparison of the F-measure plot

Here, the bio- document retrieval system is achieved by using Robin Fingerprint with Re-Ranking Model (RFRRM) algorithm. When analyzing figure 7, we obtain the maximum precision value of proposed approach using RFRRM-BDR, this value is high compare to existing approach of RF-BDR [33]. When we give the key word hypothyroidism-Hyperthyroidism, that time we obtain the 100% precision value for both the method. In figure 8, shows the performance comparison of the RFRRM-BDR and RF-BDR using recall measure. Here, we using the Word Net- MeSH terms such as Adenovirus-Rotavirus, Hypothyroidism-Hyperthyroidism and Pneumonia-Asthma we obtain the similar recall values for both RFRRM-BDR and RF-BDR. Moreover, figure 9 shows the Performance comparison of the F-measure plot. When analyzing figure 9, we obtain the maximum f-measure value of proposed approach using RFRRM-BDR. Overall, the performance of the proposed technique is better compared to the cross ontology-based similarity measure and robin fingerprint based Bio document retrieval using Pub Med database. Table 5 shows the comparison of Precision, Recall and F-measure using different keywords

Table: 5 Comparison of Precision, Recall and F-measure for different keywords

Query keyword		Precision (%)		Recall (%)		F-Measure (%)	
WordNet	MeSH	RF-BDR [33]	RFRRM-BDR	RF-BDR [33]	RFRRM-BDR	RF-BDR [33]	RFRRM-BDR
Adenovirus	Rotavirus	96.77	96.77	100	100	98.36	98.36
Anemia	Appendicitis	96.77	96.77	96.66	100	98.360	98.36
Carcinoma	Neoplasm	78.94	88.23	100	100	88.23	93.75
Hypothyroidism	Hyperthyroidism	100	100	100	100	100	100
Pain	Ache	60.71	60.714	56.6	66.66	58.62	58.62
Pneumonia	Asthma	96.77	100	100	86.66	98.36	92.85
Diarrhea	Mucosal	80.76	75	70	90	75	81.81
edema	Angioedema	86.3	89.28	78.43	83.3	83.45	86.20
ulcer	Decubitus	56.83	88.23	56.45	100	67.43	93.75

6. CONCLUSION

At present, search engines are been most practiced widely for abstracting information's from several resources all through the world. Wherever, common of searches lies in the field of biomedical for recovering associated documents from several biomedical databases. Now search engines dearth's in related level of documents removed from the databases. We have proposed a bio-medical document retrieval (BDR) system in this study with the suggested Re-ranking model (RRM). In RRM, we have offered a similarity measure among documents. In this measure, several striking properties are implanted which takes the subsequent three cases into account: i) The feature seems in both documents, ii) the feature seems in only one document, and iii) the feature seems in none of the documents. The suggested BDR system contains into six most important steps, which comprises 1) dataset preparation, 2) Preprocessing, 3) Feature extraction, 4) Indexing, 5) Re-ranking model (RRM) and 6) cross ontology measure and 7) Retrieval process. Finally, the related documents were recovered from the document repository by means of the matching result. The experimental effects have exposed that the presentation attained by the suggested BDR system is superior to that attained by other BDR system in terms of accuracy, recall and F-measure.

REFERENCE:

[1] Michael Goebel and Le Gruenwald, "A survey of data mining and knowledge discovery software tools" ACM digital library,vol.1, no.1, 1999.

[2] Satchidananda Dehuri and Ashish Ghosh, "Revisiting evolutionary algorithms in feature selection and nonfuzzy/fuzzy rule based classification", WIREs Data Mining Knowl Discovery 2013.

[3] Udo Hahn and Inderjeet Mani, "The challenges of automatic summarization", IEEE Computer, vol.33, no. 11, 29–36, 2000

[4] Inderjeet Mani and Mark T. Maybury, "Advances in automated text summarization", Cambridge: MIT Press 1999.

[5] Yihong Gong and Xin Liu, "Creating generic text summaries", In Proceedings of the 6th international conference on document analysis and recognition, pp. 903–907, 2001

[6] Christof Monz, "Document Retrieval in the Context of Question Answering", Proceedings of the 25th European conference on IR research, 2003

[7] Elizabeth D. Liddy, "Document Retrieval Automatic", Encyclopedia of Language and Linguistics, 2nd Edition, 2005.

[8] Fellbaum, "WordNet Theory and Applications of Ontology", Computer Applications, pp. 231-243, 2010.

[9] Maria Indrawan and Seng W. Loke, "The Impact of Ontology on the Performance of Information Retrieval: A Case of WordNet", Building New Dimensions of Information Technology, 2010.

[10] Pereira, d. C., Tettamanzi, C. A.G.B.: "An ontology-based method for user model acquisition", In: Ma, Z. (ed.) Soft computing in ontologies and semantic Web", Studies in fuzziness and soft computing, pp. 211–227. Springer, Heidelberg (2006).

[11] Abraham Bernstein, Esther Kaufmann, Christoph Kiefer and Christoph Burki, "SimPack: A Generic Java Library for Similarity Measures in Ontologies", Technical report, University of Zurich, Department of Informatics, 2005.

[12] Shahrul Azman Noah, Lailatulqadri Zakaria and Arifah Che Alhadi, "Extracting and Modeling the Semantic Information Content of Web Documents to Support Semantic Document Retrieval", Proceedings of the Sixth Asia-Pacific Conference on Conceptual Modeling, Vol: 96, 2009.

[13] Sung-Shun Weng, Hsine-Jen Tsai, Shang-Chia Liu, Cheng-Hsin Hsu, "Ontology construction for information classification", Expert Systems with Applications, Vol: 31,No: 1, pp. 1-12, 2006.

[14] Pedersen, T., Pakhomov, S., Patwardhan, S., and Chute, "Measures of semantic similarity and relatedness in the biomedical domain", Journal of Biomedical Informatics, vol: 40, no: 3, pp.288-299, 2007.

[15] Xiaodan Zhang, Liping Jing, Xiaohua Hu, Michael Ng, Jiali Xia and Xiaohua Zhou, "Medical Document Clustering Using Ontology-Based Term Similarity Measures", International Journal of Data Warehousing & Mining, vol: 4, no; 1, pp: 62-73, 2008.

[16] Berners-Lee and Fischetti, "Weaving the web: The original design and ultimate destiny of the World Wide Web by its inventor". San Francisco, CA: HarperAudio, 1999.

[17] Mitra and Chaudhuri, "Information Retrieval from Documents A Survey Information Retrieval", vol. 2, pp. 141–163, 2000.

[18] Jing Bai and Jian-Yun Nie, "Adapting information retrieval to query contexts", Information Processing and Management, vol. 44, no.6, 2008

[19] Chow, T., Zhang, H. & Rahman, "A new document representation using term frequency and vectorized graph connectionists with application to document retrieval", Expert Systems with Applications, vol. 36, pp. 12023–12035, 2009.

[20] Manjunath Ramachandra, "Web-Based Supply Chain Management and Digital Signal Processing: Methods for Effective Information Administration and Transmission", Business Science Reference, pp. 182-194, 2010.

[21] Yue, X., Di, G. Yu, Y. Wang, W. and Shi, H., "Analysis of the Combination of Natural Language Processing and Search Engine Technology", 2012 International Workshop on Information and Electronics Engineering,vol.29, pp. 1636 – 1639, 2012.

[22] Liddy, "Document Retrieval, Automatic. In: Encyclopedia of Language and Linguistics (Second Edition) pp. 748–755, 2006.

[23] R.B.-Yates and B.R.-Neto, "Modern Information Retrieval" Addison Wesley Longman, 1999.

[24] Vester and Martiny, "Information retrieval in document spaces using clustering", Informatics and Mathematical Modelling, Technical University of Denmark, 2005.

[25] Liu, "On Document Representation and Term Weights in Text Classification", In: Handbook of Research on Text and Web Mining Technologies, 2010.

[26] Grossman, D. A. and Frieder, "Information Retrieval: Algorithms And Heuristics", The Springer International Series in Engineering and Computer Science. Springer

[27] Comfort T. Akinribido, Babajide S. Afolabi, Bernard I. Akhigbe and Ifiok J. Udo, "A Fuzzy-Ontology Based Information Retrieval System for Relevant Feedback," International Journal of Computer Science, Vol. 8, No.1, pp. 382 – 389, 2011.

[28] David Sanchez, Montserrat batet, David Isern and Aida valls, "Ontology based semantic similarity: A new feature-based approach," Journal of expert system with application, vol. 39, pp. 7718-7728, 2012

[29] Bridget T. Mclnnes and Ted Pedersen, "Evaluating measure of semantic and relatedness to disambiguate terms in biomedical text," Journal of Biomedical informatics, vol. 46, pp. 1116-1124, 2013

[30] Bojan Furlan, Vuk Batanovic and Bosko Nikolic, "Semantic similarity of short texts in language with deficient natural language processing support," Journal of Decision support system, vol. 55, pp. 710-719, 2013.

[31] Dolf Trieschnigg, Piotr Pezik, Vivian Lee, Franciska de Jong, Wessel Kraaij and Dietrich Rebholz-Schuhmann, "MeSH Up: Effective MeSH Text Classification for Improved Document Retrieval", Bioinformatics, vol. 25, no.11, pp. 1412-1418, 2009.

[32] Shi-Jay Chen; Hung-Chin Chu, "A new method for fuzzy query processing of document retrieval based on extended fuzzy concept networks", proceedings of the 2010 International Conference On Electronics and Information Engineering (ICEIE), Kyoto, pp: V2-370 - V2-375, 2010.

[33] D.jayasri and D. manimegalai, "An efficient cross ontology-based similarity measure for bio-document retrieval system", Journal of Theoretical and Applied Information Technology, vol.54, no.2, 2013.

[34] McEntyre J, Lipman D. PubMed: bridging the information gap. Can Med Assoc J 2001;164:1317–9.

[35] Pubmed. <http://www.ncbi.nlm.nih.gov/pubmed/>.

[36] G. Salton. "Automatic Text Processing," Addison Wesley, 1989.

[37] Spärck Jones, Karen, "A statistical interpretation of term specificity and its application in retrieval". Journal of Documentation, Vol: 28, No: 1, pp: 11–21, 1972.

[38] Calvin Chan, Hahua Lu, "CMPUT690 Term Project Fingerprinting using Polynomial (Rabin's method)", 2001.

[39] Miller. G.A, "WordNet: A lexical Database for English", Comn. ACM, Vol. 38, No. 11, pp. 39-41, 1995.

[40] W. Douglas Johnston Stuart J. Nelson and Betsy L. Humphreys. "Relationships in Medical Subject Headings (MeSH)",In National Library of Medicine, Bethesda, MD, USA, 2002.

[41] S.J. Nelson, D. Johnston, and B.L. Humphreys, "Relationships in Medical Subject Headings", In C.A. Bean and R. Green, editors, Relationships in the Organization of Knowledge, pp: 171-184, Kluwer Academic Publishers, New York, 2001.