OntDR: An Ontology-based Augmented Method for Document Retrieval

Poonam Yadav
Assistant Professor
D.A.V College of Engineering & Technology,
Kanina (Mohindergarh)
Haryana, India

R. P. Singh, PhD.
Professor&Director
Bimla Devi Educational Society's Group of Institutions
JB Knowledge Park, Faridabad – 121001
Haryana, India

ABSTRACT
The document retrieval is one of the fast growing and complex research area in the field of information retrieval. An effective Information retrieval can be obtained only under strong document retrieval algorithm. As compared to the information retrieval, document retrieval is also a tedious process. The accurate retrieval of a document needs highly precise and mathematically vibrant methods. A number of researches have been targeted for the document retrieval, which yielded expected result within their boundaries. In this paper, we proposed an ontology-based augmented method for document retrieval. The ontology defined in our proposed approach gives extra freedom to choose between the documents and thus give an accurate retrieval of the documents. The mutual association (MA) value specifies the interrelated documents in the problem space. The array index values, which we provide, give accurate distinction between each document. The results and analysis of our proposed method showed expected results and a comparative analysis was subjected for analyzing the proposed method with an existing algorithm. The F-measure comparison showed the performance improvement of the proposed method with respect to the existing method.

Keywords
Information retrieval, Document retrieval, Ontology, mutual association, array index, recall, precision, f-measure.

1. INTRODUCTION
The field of searching possesses text and documents retrieval as key features. It is very hard to obtain a specific content from the internet database, so for the ease of that process searching is taken in concern. In order to fetch the whole document, the document retrieval methods are used. Document Retrieval is the computerized process of trading a list of documents that are appropriate to an inquirer's request by associating the user's request to spontaneously produced index of the textual content of documents in the system. Document Recovery schemes are grounded on diverse theoretic models, which fix how matching and ranking are steered [1]. In Document Retrieval, most of the procedures take place with dynamism when the user inputs their query, while other processes take place off-line in advance and in batch mode and do not involve individual users. The methods which are selected for document retrieval is mainly depend on the nature of the user queries. As discussed above, there are a number of methods are used for the retrieval of the documents such as clustering, indexing, page ranking, etc. Most of the methods retrieve only the document that containing relevant data only, i.e. a document is fetched in respective of the user query. A meliorated retrieval system is introduced by including ontology with the document retrieval system [20].

Use of ontology enables to define concepts and relations representing knowledge about a particular document in domain specific terms. In order to express the contents of a document explicitly, it is necessary to create links (associations) between the document and relevant parts of a domain model, i.e. links to those elements of the domain model, which are relevant to the contents of the document [2]. With the help of ontology, there is lot of advantages such as improved access to documents. The search mechanism must exploit semantic characteristics of search queries and documents, and be able to find relevant documents that would not be found by a simple full-text search [3, 4]. Most of the system includes domain ontology for the retrieval of the documents [5]. The main reason behind the use of ontologies is to overtake the limits of keyword-based search [4], i.e. in ontology based systems, instead of keywords, a concept is extracted. The main difference between a concept and keyword is that the concept carries information about a specific part of the document, while keyword doesn't contain any information like that. With the help of the concepts, we can describe a document. The concepts have a name and/or description in each of the addressed languages, which can be used for presenting them in the user interface. The documents are annotated with concepts. In the simplest case, for each document, there is a set of concepts that are relevant for the contents of the document [5].

The ontology-based system provided quick access to documents and information with the help of taxonomy created from the concepts, which is called a concept map. So, from the concept map, we can easily form the document retrieval system. But, even though it provides quick access and execution, error may occur if it is not handled properly [6]. The concept map creation is the most tedious process. The main difficulty concerning to the concept map is that the automatically generated conception map will not provide an effective result, so we have manually generated concept maps becomes a necessity [7]. Although many ontology based applications have been developed, all of them require the users to include some forms of semantic annotations explicitly in their queries. Khan *et al.* [8] required the user to write SQL-like queries wherein the exact concepts are incorporated. These applications are thus not suitable for typical information users as it is usually not straightforward to identify the matching concepts of a query from domain ontology. Contreras *et al.* [9] enabled a user to submit queries in natural language by using Natural Language Processing (NLP) tool to extract concepts and instances from the queries. However, the performance of their application heavily depended on the quality of the NLP tool. Thus a method with improved abilities has became obvious.

In this paper a document retrieval system incorporating ontology with array indexing. The system is adaptive because it will retrieve a most relevant document as well as documents

which are close to the user's queries. The array indexing is the key part of our concept, because the array indexing helps in obtaining an inter-relation between the documents. The array index is compared with a value that generated from the ontology defined over the concept and documents. This process over comes the difficulty caused by the concept maps, i.e. we can obtain an effective inter related concepts in the concept map. The array indexing also ensures the redundancy of concepts or documents in the database. Our method is mainly based on four steps, Concept Extraction, Ontology Definition, 3) Indexing, and 4) Retrieval. Prior to these four steps, there are some preprocessing step, which includes the most common document formatting methods like stopword removal and the stemming. After the preprocessing, the document is subjected to concept extraction phase, which extracts the most frequent concepts from the documents. The third phase is defined as Indexing, which is pointed on giving index to each of the documents and the indexing is named as array indexing. The fourth and the final step constitute the retrieval of the documents, the retrieval is based on ontology definition and the array index of each documents.

The paper is organized mainly into the following sections; the 2nd section includes the related researches that are conducted. The 3rd section describes the motivation behind the proposed method and it the algorithm, which give way to our proposed method. The 4th section consists of the detailed description of proposed method, which includes the preprocessing and the four phases of the proposed method. The 5th section described the result evaluation and the comparative analysis of the proposed method. We conclude the proposed method with the 6th section.

2. LITERATURE SURVEY

The document retrieval becomes a hot research area in the field of information and data retrieval system. So, there has been a lot of research conducted to the quest for an effective document retrieval system. Let us discuss some of the prior researches.

Shao Fen Liang *et al*. [10] has proposed an analysis of the type of information required for such a task has given rise to four main areas of research: information retrieval, document annotation, summarization and visualization. The first stage of the research has focused on information retrieval, and an algorithm, "Windmill Expansion" (WE) has been proposed to do this. The algorithm uses retrieval feedback techniques for automated query expansion in order to improve the effectiveness of information retrieval. WE is based on the extraction of human–generated written phases of automated query expansion. Top and Second Level expansion terms have been generated and their usefulness evaluated. The evaluation has concentrated on measuring the degree of overlap between the retrieved URLs. The less the overlap, the more useful the information provided. The Top Level expansion terms were found to provide 90% of useful URLs, and the Second Level 83% of useful URLs. Although there was a decline of useful URLs from the Top Level to the Second Level, the quantity of relevant information retrieved has increased. The originality of SEMIOTIKS lies in its use of the WE algorithm to help non–domain specific experts automatically explore domain words for relevant and precise information retrieval.

Dolf Trieschnigg *et al.* [11] 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) provides an efficient way of accessing and organizing biomedical information by reducing the ambiguity inherent in free-text data. Different methods of automating the assignment of MeSH concepts have been proposed to replace manual annotation, but they are either limited to a small subset of MeSH or has only been compared to a limited number of other systems. They compared the performance 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 reproducing and complementing manual MeSH annotations. A K Nearest Neighbor system clearly outperforms the other published approaches and scales well with large amounts of text using the full MeSH thesaurus. Their measurements demonstrate to what extent manual MeSH annotations can be reproduced and how they can be complemented by automatic annotations. They also showed that a statistically significant improvement can be obtained in information retrieval (IR) when the text of a user's query is automatically annotated with MeSH concepts, compared to using the original textual query alone.

Rong Zhao *et al*. [12] have developed a technique for content-based web document retrieval, using both keywords and image features to represent the documents. Two different approaches to image feature representation, namely, color histograms and color anglograms, are adopted and evaluated. Experimental results showed that LSI, together with both textual and visual features, is able to extract the underlying semantic structure of web documents, thus helping to improve the retrieval performance significantly, even when querying is done using only keywords. Anne Kathrin Bartsch *et al.* [13] used a GeneReporter, which is a web tool that reports functional information and relevant literature on a protein-coding sequence of interest. Its purpose is to support both manual genome annotation and document retrieval. PubMed references corresponding to a sequence are detected by the extraction of query words from UniProt entries of homologous sequences. Data on protein families, domains, potential cofactors, structure, function, cellular localization, metabolic contribution and corresponding DNA binding sites complement the information on a given gene product of interest.

S. Siva Sathya *et al.* [14] have proposed a document crawler is used for gathering and extracting information from the documents available from online databases and other databases. Since search space is too large, Genetic Algorithm (GA) is used to find out the combination terms. In the proposed document retrieval system, they extracted the keywords from the document crawler and with these keywords GA generates combination terms. The proposed work is having three main features: First is to extract keywords and other information from the database by a document crawler. Second is to generate the combination terms using genetic algorithm. Third, results generated from the GA are applied to information retrieval system to generate better results. From the results obtained, the relevance of the documents is verified using evaluation measures namely precision and recall. Jayapal R *et al.* [15] provided a survey of methods developed by researchers to access document images. The survey included papers covering the current state of the art of the research in document image retrieval based on images such as signature, logo, machine-print, different fonts etc.

3. MOTIVATIONS AND CONTRIBUTIONS OF THE RESEARCH

The scenarios that discussed above points out that a detailed information retrieval can only be achieved with the help of an efficient document retrieval system. So this research is subjected to design an efficient document retrieval method, which can achieve a good information retrieval strategy. The

new method is equipped with some improved capabilities such as mutual association and array indexing. The improved capabilities differentiate our method from the other methods. The ontology defined over the main concept give emphasizes to our method, and our method attains an adaptive searching capacity. The main motivation to our method is to propose an adaptive document retrieval system which can retrieve relative documents with less redundancy and with less retrieval time. So the proposed method is organized as per motivating criteria that we discussed.

Motivating Algorithm: The motivation behind proposing a method for effective document retrieval is furnished in the above section. The document retrieval is a process which includes process like indexing, grouping and retrieving. According to this, Rong Zhao *et al.* [12] have proposed a document retrieval, which is based on text in the documents. Their algorithm is purely based on the frequency of the text, i.e., the words in each document. In addition to, Raymond Y.K *et al.* [19] have proposed an e-learning algorithm which, is related to the ontology. They have developed a concept extraction under ontology. So, our research is planned in accordance with the algorithms proposed by Rong Zhao *et al.* [12] and Raymond Y.K *et al.* [19].

4. PROPOSED ONTOLOGY-BASED AUGMENTED APPROACH FOR DOCUMENT RETRIEVAL

Document retrieval is one of the most demanding area in the field of information retrieval. There are a number of algorithms for carrying out the document retrieval process. Here, in this paper we have proposed a document retrieval technique, which is an ontology based technique. The processes can be explained though a block diagram that is shown Figure 1.

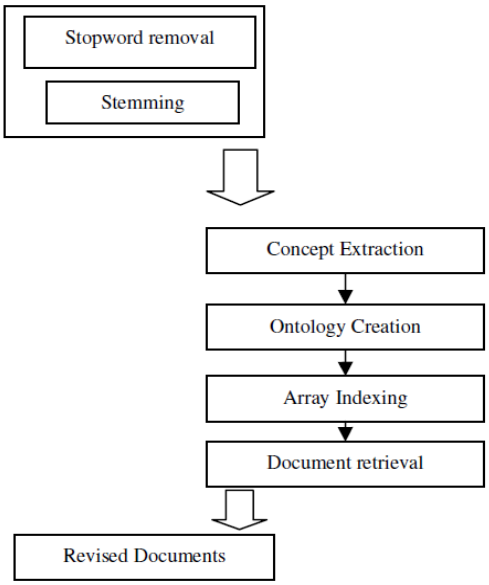

Fig.1. Block diagram of the proposed document retrieval method

4.1 Preprocessing

Preprocessing is done prior to the main processes of our proposed method. Preprocessing converts a document into a set of keywords. The preprocessing consists of stop word removal and stemming. The selected documents are subjected to the stop word removal process, where connecting words such as “is”, “as”, etc. are removed. The words in the document are converted into its basic form by a process, called stemming. The preprocessed documents are given for further processing.

4.1.1 Stop Word Removal

In a document, there are commonly utilized words that carry less important meaning than keywords hence it is necessary and beneficial to remove these words. Most of the search engines remove the commonly utilized words or rather known as ‘stop words’ from a keyword phrase to return the most relevant result. In searching, all stop words, for instance, most used words like 'a' and 'the', are detached from multiple word queries for increasing search performance [16]. Stop words like “it", "can”, "an", "and", "by", "for", "from", "of", "the", "to", "with" are the common stop words. Stop word removal is done while parsing a document to obtain information about the content or while scoring fresh URLs that the page recommends.

4.1.2 Stemming

In many cases, morphological variants of words have similar semantic interpretations and can be considered as equivalent for the purpose of many applications. For this reason, a lot of stemming Algorithms or stemmers has been developed to reduce a word to its stem or root form. The stems are used to represent the key terms of a query or document instead of the original word. Lemmatization [17] is an algorithm which attempts to convert a word to its linguistically correct root which ultimately facilitates the reduction of all words, possessing an identical root to a single one. This is obtained by eliminating each of the word of its derivational and inflectional suffixes [18]. For instance, "orient," "oriented" and "orientation" are all condensed to "orient", which is the base form, similarly “runs”,” run “ are all condensed into “run”.

4.2 Concept Extraction from the Documents

It is the initial step in our proposed method, which deals with extraction concepts from documents. The concepts are the keyword which plays a key role in the document. In order to find the most relevant documents from the database, we should group the documents in the increasing order of the weights of the documents. For finding the weight of a single document, we need to find the weights of the concepts which are present in the document. So, the concept extraction plays a crucial role in our proposed method. We select document from our database and extract all the words from the document. We find the concept from the extracted keywords, by assessing the relation between one keyword to another.

$$D_1 = \{k_1, k_2, k_3, \mathrm{K}\ , k_n\}$$

Where, the D_1 represents the first document in the database and the $k_1, k_2, k_3, \mathrm{K}\ , k_n$ are the keywords present in the document. The equation represents a document and its set of keywords. We have to find the relation between the keywords in order to extract the concepts. The concept can be extracted from the *mutual association* [19] of a keyword to another. The mutual association can be given by the following formula.

$$MA(k_i, k_j) = \log_2 \frac{P(k_i, k_j)}{P(k_i).P(k_j)}$$

We consider a text window here, it contains the keywords k_j and k_j, we find the probability $P(k_i)$ by the frequency with which k_i occur in the text window to the total frequency with which the k_i occur on the whole document. This is given by,

$$P(k_i) = \frac{|f_w(k_i)|}{|f_d(k_i)|}$$

Similarly, we can find the $P(k_j)$ value of the keyword k_j from the whole document. Where $P(k_i, k_j)$ is the conditional probability of both the keywords to be occurring in the text window to their occurrence in that whole document. But, we have to consider presence and absence of both keywords present in the document, so for that we have adopted *Optimized mutual association (OMA)* from [19].

$$OMA(k_i,k_j) = \left[\begin{array}{l} P(k_i,k_j).\log_2 \frac{P(k_i,k_j)+1}{P(k_i).P(k_j)} + \\ P(\neg k_i,\neg k_j).\log_2 \frac{P(\neg k_i,\neg k_j)+1}{P(\neg k_i).P(\neg k_j)} \end{array} \right] - \left[\begin{array}{l} P(k_i,\neg k_j).\log_2 \frac{P(k_i,\neg k_j)+1}{P(k_i).P(\neg k_j)} + \\ P(\neg k_i,k_j).\log_2 \frac{P(\neg k_i,k_j)+1}{P(\neg k_i).P(k_j)} \end{array} \right]$$

OMA is more effective in extracting the concepts from the document, we extract concept from the document with aid of the OMA value refined under a threshold value

i.e. $$\begin{array}{l} if\ OMA(k_i,k_j) \geq th, C_i = true \\ else C_i = false \end{array}$$

Thus, concepts are created according to the OMA values and we generate concept of all the documents in the database and these concept are subjected for the ontology creation.

4.3 Defining Ontology over the Concepts

The ontology is defined for the accurate retrieval of the documents from the database. In the case of defining the ontology over the concept, initially we have to define a concept map. When we consider our proposed method, we use a manually created concept map. The concept maps are the key functional unit for defining the ontology. The main reason behind defining the ontology is to obtain a relationship between the documents in the database. We use the generated concept to getting more precise output for our proposed method. The concept map used is designed by domain experts. We find the relation between the documents by deriving the affinity value of the concept to the document related to it. Consider the simple concept map,

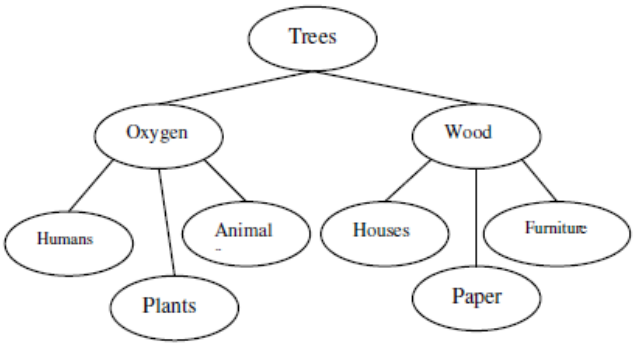

Fig: 2. Concept map: Tree

Here we can see different concepts associated with each other and a list of documents to which these concepts are belonging to and by analyzing that list, the affinity value of each concept to the respective documents are obtained. This affinity value is used to group the documents in the third step, the indexing phase. The list obtained is represented as a table, which is given below,

Table.1. Concept document mapping.

Concepts	Documents
Tree	[D_1, D_4, D_5, D_6]
Oxygen	[D_1, D_4, D_5]
Wood	[D_1, D_5, D_6]
Humans	[D_4, D_3, D_5]
Plants	[D_5, D_2, D_3]
Animals	[D_2, D_4, D_5]
House	[D_6, D_1, D_8]
Paper	[D_8, D_1, D_9]
Furniture	[D_9, D_8, D_6]

Here, 10 documents are selected for the concept map creation, $D_1, D_2, ..., D_n$ are the documents in the database. The affinity value can be defined as,

$A_v(C_i) = \frac{\sum D_X}{|D|}$, $Av(C_i)$ is the affinity value of concept C_i, D_X is the number of documents that contains the concept C_i and $|D|$ is the total number of documents. Thus $Av(tree) = \frac{4}{10} = 0.5$, so the concept *tree* has affinity value of 0.5 for the documents D_1, D_4, D_5 *and* D_6. In the similar way we find the affinity value for all the concepts generated for our proposed method in the concept map. This affinity value provides ontology over the concepts.

4.4 Indexing the Documents

The documents are arranged irregularly in the database, we have to provide an ordered arrangement for the documents. The ordered arrangements help us in the retrieval of the documents. We adopt the indexing technique for the arrangements of the documents; the forwarding index gives better choices for our proposed method. We extract the entire concept from the documents and from the extracted documents we select top N frequent documents.

$$D_i = [c_1, c_2, ..., c_n], n \in N$$

After that, we group all the documents in a relation table in order to execute the indexing process. We define an array index, I_e for the marking the index value.

$$I_e = [i_1, i_2, ..., i_n]$$

After defining the array index for every document, we execute a concept comparison between the documents. The number of keywords that matches between one document to other document is considered as its index value,

$$D_1 = [c_1, c_3, c_4, c_5], \ D_2 = [c_1, c_2, c_4, c_5, c_6]$$

From D_1 *and* D_2, we get the index as

$$D_1 \cap D_2 = [c_1, c_4, c_5]$$
$$I(D_1, D_2) = 3$$

$I(D_1, D_2)$ is the index value of D_1 to D_2, similarly we find the index value for D_1 to all other documents, so we obtain an array index for the document D_1, the array index contains all the index value of D_1 to all other documents.

$$I_{D_1} = \sum_{i=1}^{n} I(D_1, D_i)$$

In similar way, we set the index array for all the documents, we define this index array in order to retrieve the similar documents from the database. Thus, generally the array index can be defined as I_{D_i} which is given by,

$$I_{D_i} = \sum_{i=1}^{n}\sum_{j=1}^{n} I(D_i, D_j)$$

4.5 Document Retrieval from the Documents

Document retrieval is the main stage of our proposed method. When we consider the processing of our proposed method, it outperforms existing retrieval methods, due to the fact that the requested documents as well as its related documents are also retrieved by our proposed method, since the proposed method is an adaptive document retrieval system. The effectiveness of the system is based on two tables, that we are discussed in the above sections. The tables consist of the affinity value and the index values respectively, that means the retrieval is mainly depending on the values from the above mentioned tables. The retrieval method can be explained as following,

Input a concept c_i.

if $c_i \in D(set\ of\ documents)$

$select D_i$, $where c_i \in D_i$.

$D_i \Rightarrow IndexTable$

$select I_D(D_i)$.

$then c_i \Rightarrow AffinityTable$

$select A_V(c_i)$.

$$I_D(D_i) : A_V(c_i) = \begin{cases} 1, retrieve v D_i \ w.r.t \ I_{D_i} . \\ 0, reconstruct \ IndexTable \end{cases}$$

else

exit.

Here, if the input concept is presented in the database, then the document that possesses least distance to the concept is selected. Then, the document is pointed to the index table, which consists of the index values. The index of the document D_i is selected. The concept selects the affinity value A_v of the concept. The values I_{D_i} and $A_V(C_i)$ are compared, if the comparison results are matching then the document D_i is retrieved in accordance with its I_D value.

5. RESULTS AND DISCUSSION

The experimental results of the proposed method to document retrieval are presented in this section. The proposed method has been implemented in java (jdk 1.6) and the experimentation is performed on a 3.0 GHz Core 2duo PC machine with 2 GB main memory. The evaluation of our proposed method is executed by evaluating three databases; each database is having 1000 documents which are different in their content.

5.1 Dataset Description

We have selected three datasets for evaluating the proposed method under different criteria. The datasets are from different knowledge domains namely, Data mining, Software engineering and mobile communication. We used a web crawler algorithm for retrieving the documents from the web. The algorithm is designed in such a way that it only grabs the document which is related to the user's search criteria. Thus the web crawler acquired 100 documents for each of the datasets. Each of the dataset containing 100 documents and each of these 100 documents are different in their contents. The different in content will give a challenging task our adaptive document retrieval method for retrieving documents according to the contents present in the documents.

5.2 Evaluation Metrics

The proposed method is ontology based adaptive document retrieval, which consists of four phases including concept Extraction, ontology Definition, Indexing and retrieval. The Documents in the database are subjected to undergo these phases and obtained the results according to the algorithm defined in our proposed method. The result from our proposed method is evaluated to find the effectiveness of the proposed method. The evaluation is based on mainly three parameters, 1) Recall, 2) Precision and 3) F-measure. Based on these evaluating factors, we assess the performance of our proposed method.

$$\mathrm{Re}call_{doc} = \frac{|Doc_{\mathrm{Re}l} \cap Doc_{retrieved}|}{|Doc_{\mathrm{Re}l}|}$$

$Doc_{\mathrm{Re}l}$ and $Doc_{\mathrm{Re}trieved}$ are representing the total number of relevant documents and the total number of retrieved documents in the problem space. The Recall parameter for the documents is calculated using the above represented formulae. In the similar manner, we can obtain the precision parameter of the documents with help of the following condition.

$$\mathrm{P}recision_{doc} = \frac{|Doc_{\mathrm{Re}l} \cap Doc_{retrieved}|}{|Doc_{\mathrm{Re}tieved}|}$$

The precision values and the recall values are considered for finding the F-measure value for the total dataset. Thus the F-measure can be expressed as,

$$F - measure = \frac{2.\mathrm{Re}call_{doc} \times \mathrm{P}recision_{doc}}{\mathrm{Re}call_{doc} + \mathrm{P}recision_{doc}}$$

5.3 Performance Evaluation

The performance of the proposed method is subjected to action on the selected three datasets. The performance evaluation graphs are plotted on the basis of the three evaluation factors such as $\mathrm{Re}call_{doc}$, $\mathrm{P}recision_{doc}$ *and* $F - measure$. After evaluating all the datasets with our proposed method, we compare our method with an existing method [21] for assessing the performance of our methods. The figure 3, 4 and 5 describe the Recall, Precision and F-measure evaluations of different databases through our proposed method. Here, we represent the document related with Data mining as dataset1, Software engineering as dataset2 and mobile communication as Dataset3. Thus, we find the recall parameter for the three set of documents for the three datasets, which are in concern. The Values are mapped in the graph Fig.3. In the similar way, we plot the precision parameter also, which plotted in the graph Fig.4. After evaluating the values of recall values and the precision values, we measure the F-measure value for the three set of documents that we select from the three datasets. The F-measure chat is plotted in Fig.5.

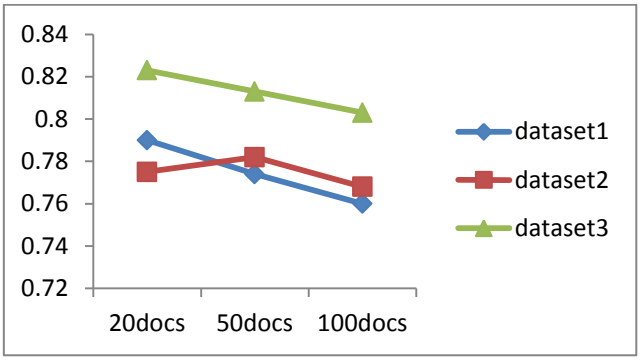

Fig. 3. Recall values of different databases

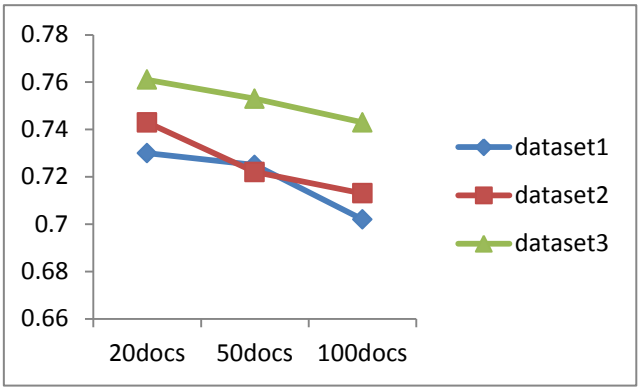

Fig. 4. Precision values of different databases

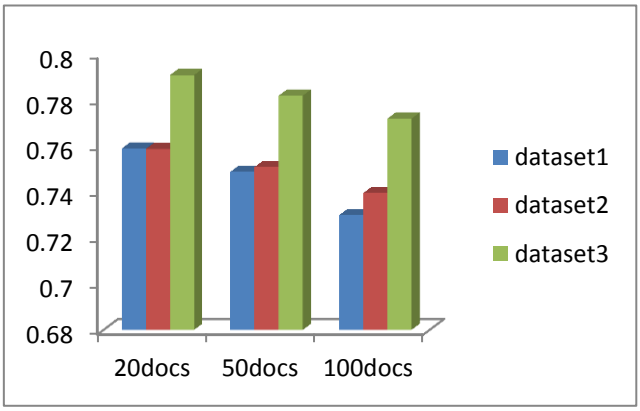

Fig.5. F-measure values of different datasets

The evaluation results showed that our result provides encouraging results. Now, we compare our method with an existing method [21] using the values obtained by the existing method. From the values, we find that our proposed method, ontDR, shows advantage over the existing method, bioDR. From the evaluations and assessments, we find that our method outperformed the other methods.

Table.2.Comparison mapping

dataset	bioDR recall	bioDR precision	bioDR F-measure	OntDR recall	OntDR precision	OntDR F-measure
1	0.60	0.65	0.624	0.76	0.702	0.72985
2	0.77	0.63	0.647	0.768	0.713	0.739479
3	0.61	0.54	0.709	0.803	0.743	0.771836

6. CONCLUSION

We have proposed ontology based augmented method for the accurate and precise selection of the relevant information. The proposed method performed well under different test criteria and showed expected results. The key parts in the method are the mutual association value and the array index values. We have tested the proposed method with three datasets and the results obtained were up to the marks.. From the results and comparative analysis, it is proved that the proposed method performs well under different test criteria. The future enhancements to the system can be done by improving the calculation strategy of the mutual association value and by creating more dynamic array indices.

REFERENCES

[1] R. Meersman and Z. Tari, “Ontology Learning for Search Applications”, International transactions from Springer, pp. 1050-1062, 2007.

[2] Jan Paralic and Ivan Kostial, “Ontology-based Information Retrieval”, Information and Intelligent System, Croatia, pp. 23-28, 2003.

[3] Eelco Mossel, “Crosslingual Ontology-Based Document Retrieval”, In Proceedings of the RANLP 2007 workshop of Natural Language Processing and Knowledge Representation for eLearning Environments, Borovets, Bulgaria, 2007.

[4] D. Vallet, M. Fernández and P. Castells. “An Ontology Based Information Retrieval Model.” In Proceedings of the 2nd European Semantic Web Conference on the Semantic Web Research and Applications, pp. 455-470, 2005.

[5] L. Lemnitzer, P. Monachesi, K. Simov, A. Killing, D. Evans and C. Vertan. “Improving the search for learning objects with keywords and ontologies” In Proceedings of the Second European Conference on Technology Enhanced Learning, Crete, Greece, 2007.

[6] Pablo Castells, Miriam Fernández, David Vallet, Phivos Mylonas and Yannis Avrithis, “Self-Tuning Personalized Information Retrieval in an Ontology-Based Framework”, In Proceedings of the International Workshop on Web Semantics, pp. 977-986, 2005.

[7] Xing Jiang and Ah-Hwee Tan, “OntoSearch: A Full-Text Search Engine for the Semantic Web”, In Proceedings of the 21st National Conference on Artificial Intelligence, 2006.

[8] Khan L, McLeod D, and Hovy E, “Retrieval effectiveness of an ontology-based model for information selection”, The VLDB Journal, pp. 71–85, 2004.

[9] Contreras J, Benjamins V R, Bl´azquez M, Losada S, Salla R, Sevilla J, Navarro D, Casillas J, Mompo A, Paton D, Corcho O, Tena P, and Martos I, “A semantic portal for the International Affairs sector”, In EKAW Springer, Berlin, pp. 203–215, 2004.

[10] Shao Fen Liang, Paul Smart, Alistair Russell and Nigel Shadbolt, “Using Windmill Expansion for Document Retrieval”, The Open Information Systems Journal, Vol.3, pp.1-8, 2009.

[11] 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.

[12] Rong Zhao and Grosky W.I, “ Narrowing the semantic gap - improved text-based web document retrieval using visual features”, IEEE Transactions on Multimedia, Vol. 4 , No.2, pp. 189 – 200, 2002.

[13] Annekathrin Bartsch, Boyke Bunk, Isam Haddad, Johannes Klein, Richard Münch, Thorsten Johl , Uwe Kärst, Lothar Jänsch, Dieter Jahn and Ida Retter, "Gene-Reporter—sequence-based document retrieval and annotation", Bioinformatics, 2009.

[14] A. S. Siva Sathya and B. Philomina Simon, "A Document Retrieval System with Combination Terms Using Genetic Algorithm", IJCEE, Vol.2, No.1, pp.1-6, 2010.

[15] Jayapal R and J.K.Mendiratt, "Document Image Retrieval: An Overview", International Journal of Computer Applications, Vol. 1, No.7, pp.114–119, 2010.

[16] V.A. Narayana, P. Premchand and A. Govardhan, "Effective Detection of Near Duplicate Web Documents in Web Crawling", International Journal of Computational Intelligence Research, Vol. 5, No. 1, pp.83–96, 2009.

[17] Anton Karl Ingason, Sigrun Helgadottir, Hrafn Loftsson and Eirikur Rognvaldsson, "A Mixed Method Lemmatization Algorithm Using a Hierarchy of Linguistic Identities (HOLI)", Advances in Natural Language Processing-Lecture Notes in Computer Science, Vol. 5221, pp. 205-216, 2008.

[18] Lovins, J.B. "Development of a stemming algorithm". Mechanical Translation and Computational Linguistics, Vol. 11, pp. 22-31, 1968.

[19] Raymond Y.K. Lau, Dawei Song, Yuefeng Li, Terence C.H. Cheung, and Jin-Xing Hao, "Toward a Fuzzy Domain Ontology Extraction Method for Adaptive e-Learning", IEEE Transactions On Knowledge And Data Engineering, Vol. 21, No. 6, pp. 800-813, 2009

[20] Christos Faloutsos and Douglas W. Oard, "A survey of information retrieval and filtering methods", Technical Report on A survey of information retrieval and filtering methods, pp. 1-24, 1995.

[21] Analia Lourenço, Rafael Carreira, Daniel Glez-Pena, Jose R. Méndez, Sonia Carneiro, Luis M. Rocha, Fernando Diaz, Eugénio C. Ferreira, Isabel Rocha, Florentino Fdez-Riverola, Miguel Rocha, " BioDR: Semantic indexing networks for Biomedical Document Retrieval", Expert Systems with Applications, Vol 37, no.4, pp: 3444-3453, 2010.