Journal of Theoretical and Applied Information Technology
20th August 2013. Vol. 54 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN EFFICIENT CROSS ONTOLOGY-BASED SIMILARITY MEASURE FOR BIO-DOCUMENT RETRIEVAL SYSTEM

[1]D.JAYASRI, [2]DR. D. MANIMEGALAI

[1]Associate Professor, Department of Mathematics
ULTRA College of Engineering & Technology for Women, Madurai
[2]Professor and Head, Department of Information Technology,
National Engineering College, Kovilpatti. Tamilnadu, India
E-mail: [1]djayasriphd@gmail.com

ABSTRACT

In Biomedical research, retrieving documents that match an interesting query is a task performed quite frequently. Typically, the set of obtained results is extensive containing many non-interesting documents and consists in a flat list, i.e., not organized or indexed in any way. In this paper, we have presented an efficient bio-medical document retrieval system with the proposed cross-ontology based semantic similarity measure. Here, we have utilized the WordNet and MeSH ontologies for matching the input query keyword. As well, we have designed a novel cross-ontology based semantic similarity measure for the query keywords. The proposed system runs with three major processes, which includes 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. Finally, the relevant documents are retrieved from the document repository using the matching result. The experimentation process is carried out with the aid of the different set of medical documents and hence achieved the results. The performance analysis of the proposed retrieval system is evaluated by comparing the existing similarity measures along with evaluation metrics such as precision, recall and F-measure by achieving more than 95% accuracy in most cases.

Keywords: *Biomedical documents, Ontology, Wordnet, MeSH (Medical Subject Headings), TF-IDF Similarity measure, Query refinement, Rabin Fingerprint algorithm.*

1. INTRODUCTION

Nowadays, due to the rapid advancement of Internet, the amount of information on it is increasing considerably. Most importantly, an information retrieval system should have ability to facilitate the user to achieve the useful information they requires. A user profile is very essential to provide the information that the user actually desires [10]. On the other hand, the users are discontented with the low precision and recall. The large volume of machine comprehensible information on the Semantic Web has provided some opportunities for the improvement of traditional search. Some semantic search techniques [11] have been developed to enhance the traditional search technology. Since the ranking of documents is an important part of today's search engines, the ranking of relationships will also be crucial for future semantic search engines that would support detection and mining of the Semantic Web [12]. It is believed that only 20% of formal information can be extracted from the data repository containing numeric data only, and the residual 80% of information are concealed in the documents [13]. A similar surveillance has been made by Feldman [14], who states that 80% of precise knowledge in an enterprise can be found in documents. Hence, the document management has been recognized as an important topic in the information and knowledge management [15], which has been a well-established research field and triumphant for several applications in many areas.

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. To devise a search expression that gives the desired content, we can select the keywords (e.g., benchmark, performance) and phrases (e.g., document retrieval, full-text retrieval, information retrieval) which would likely be found in suitable documents. The intent of document retrieval systems is to return the appropriate documents to a

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

user based on their query, where the query is a collection of keywords. A document is considered as pertinent when its content is related to the query [32]. The document that has to be identified should be taken as a query in order to compare it with many or all of the document data in the digital library. The compressions or matching between the query document and each document in the digital library are carried out using the different attributes of the documents. Document Retrieval is a computerized process, which produce a relevance ranked list of documents according to an inquisitor's request by comparing their request to an automatically formed index of the documents in the system. Today, each one is utilizing such systems in the form of web-based search engines. The document processor, query analyzer, and matching function are the three main components of document retrieval system [26].

Document Retrieval is usually called as Information Retrieval, where the user's request are compared with an automatically created index of the textual content of documents in the system for generating a record that contains the list of documents in response to the request. Then, these documents can be accessed by the user in the same system. Good similarity measures are vital for techniques namely retrieval, matchmaking, clustering, data-mining, ontology translations, automatic database schema matching, and simple object comparisons. However, measures performed using complex or aggregated objects in ontologies are rare, they are essential for semantic web applications [3, 8, 16]. As compared to other measures, ontology-based similarity measure has some benefits: i) ontology is manually formed by human beings for a domain and so it was more exact; ii) it is much more computational efficient when compared to other techniques such as latent semantic indexing; iii) it helps to include domain knowledge into the data mining process. Generally, comparing two terms in a document by means of ontology information exploits the truth that their corresponding concepts in the ontology normally contain the properties in the form of attributes, level of generality or specificity, and their relationships with other concepts [1, 2, 17].

Ontologies are employed extensively in numerous fields such as knowledge engineering, artificial intelligence and applications related to knowledge management, information retrieval and the semantic web. Ontology defines "the basic terms and relations representing the vocabulary of topic areas and the rules for integrating terms and relations to specify extensions of the vocabulary" [5, 6, 7]. In this paper, we have utilized the WordNet and MeSH ontologies. A system that takes as input a list of keywords provided by the user and discovers their possible meanings by consulting the knowledge represented by these ontologies. These keyword senses are semantically enriched with the synonym terms found during the ontology matching process. Semantic similarity is concerned about to determine relation between two terms or concepts. Rodriguez M.A. and Egenhofer M.J [41] have utilized the WordNet and SDTS ontologies to retrieve the appropriate document using word matching. Euripides G.M. Petrakis *et al.* [42] make use of the MeSH ontology to determine the accurate document using X-similarity measure. By considering these ideas, we have designed an efficient bio-document retrieval system using cross ontology based similarity measure.

The main contributions of this research work are as follows,

- We have designed an effective method for retrieving the bio-medical document from the document repository.
- We have designed a novel cross-ontology measure.
- We have utilized two ontologies like WordNet and MeSH for matching the input query keyword.
- We have designed an effective query refining schema for matching the results and retrieving the documents.
- We have carried out the experimentation results with different set of bio-medical documents which satisfy the Wordnet and the MeSh terms.
- We have made a comparative analysis with an existing research and achieved better results in terms of evaluation metrics like precision and recall.

The rest of the paper is organized as follows: a brief review of some of the literature works in document retrieval system is presented in Section 2. The basic information about ontology utilized in our proposed technique is given in Section 3. Section 4 explains the designing procedure of the cross ontology measure. The proposed methodology for bio-document retrieval system is detailed in Section 5. The experimental results and performance analysis discussion is provided in Section 6. Finally, the conclusions are summed up in Section 7.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2. LITERATURE SURVEY

In the literature, there are already several benchmarking tools, which standardize the process of retrieving the documents using various techniques. Some of the recent points of reference works are portrayed here.

Dolf Trieschnigg *et al.* [34] 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) provide an efficient way of accessing and organizing biomedical information by reducing the ambiguity inherent to 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 have 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.

Shi-Jay Chen and Hung-Chin Chu [35] have proposed an extended fuzzy concept networks based approach for fuzzy query processing of document retrieval. A relevance matrix and relation matrix have been employed to design the extended fuzzy concept networks. Here, a satisfaction matrix has been obtained by the proposed approach by combining the document descriptor relevance matrix defined by the expert with the user's query descriptor based on diverse weights. Then, an AND operator of the quadratic-mean averaging operators has been used for computing all the elements in each row of the satisfaction matrix. Finally, the user desired relevant documents has been obtained by ranking the degrees of satisfaction of each satisfaction matrix.

Ali *et al.* [36] have proposed an approach, where the depiction of the documents has the advantage of including the information in its model of the neighborhood of terms. They have analyzed the performance of their approach in terms of research relevance and also the time of indexing and research. The obtained results have shown a substantial improvement in the relevance due to the use of the neighborhood of the terms, and this hasn't influence on the indexing and research time that stay so quick. As well, another author Dang Tuan Nguyen [37] has constructed a document retrieval system with three main features: 1) processing English queries of users, 2) cooperating with users to correct the wrong syntax queries, 3) giving results of the queries. Moreover, an important semantic rendition has been introduced for natural language queries. Their research has been limited in some specific applications such as searching e-books in e-libraries with some information about e-books. In these applications, information about application fields, data structures and more has been clearly understood.

Rong Zhao *et al.* [38] examined the use of this 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.* [39] used a Gene-Reporter, 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.* [43] 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

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

terms. In the proposed document retrieval system, we extract 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.

3. BACKGROUND INFORMATION

3.1 Wordnet

WordNet is an online lexical database of English, developed under the guidance of Miller at Princeton University [4]. Here, a set of cognitive synonyms called synsets, each representing a different concept, are formed by grouping the nouns, verbs, adjectives and adverbs. Synsets are created by using conceptual semantic and lexical relations. WordNet can also be seen as ontology for natural language terms. It has more than 100000 words, organized into taxonomic hierarchies. Nouns, verbs, adjectives and adverbs are grouped into synonym sets (synsets). The synsets are also grouped into senses i.e., diverse meanings of the same word or concept. The synsets (or concepts) have a connection to other synsets higher or lower in the hierarchy by diverse types of relationships. Hyponym/Hypernym (i.e., Is-A relationships), and the Meronym/Holonym (i.e., Part-Of relationships) are the two most common relationships. Hyponym and Hypernym are the secondary organizing principle. If a word is the hyponym of another word, then the first word has a narrower definition than the second. Inversely, the second word is the hypernym of the first word. They are both transitive relations and are their respective inverse relations. There are, nine noun and several verb Is-A hierarchies, but the adjectives and adverbs are not organized into Is-A hierarchies [9]. Same as the Open Directory, the synset ids are altered when new versions of the ontology are published, however a backward compatibility utility program is used to map synsets between the versions [18].

3.2 MESH

Medical Subject Headings (MESH) is the National Library of Medicine's vocabulary thesaurus. MeSH contains a collection of words representing descriptors in a hierarchical structure. MeSH [19, 20] is a taxonomic hierarchy of medicinal and biological terms suggested by the U.S National Library of Medicine (NLM). NLM has utilized the Extensible Markup Language (XML) as the description language for MeSH. The MeSH vocabulary file is available in XML format. All words in MeSH are placed in a hierarchy with most common words such as "Chemicals and Drugs", higher in the nomenclature than the most specific words such as "Aspirin". There are 21,973 main headings, termed descriptors, in MeSH (22,568 in 2004). Furthermore, MeSH is a hierarchical tree like structure, in which a term can emerge in different sub-trees. There are 15 tree hierarchies i.e., sub-trees in the MeSH ontology and the type of relationship between nodes in each sub-tree is IS-A relationship [21, 22].

4. DESIGNING OF CROSS ONTOLOGY MEASURE

Cross ontology measures compares the words from diverse ontologies such as WordNet and MeSH. The cross ontology approaches often requires hybrid or feature based measures, because the structure and information content between diverse ontologies cannot be compared directly. For instance, two terms are alike if they have same spelling or meaning, or they are related with other terms that are alike. Several intelligent knowledge-based applications have techniques for computing semantic similarity between the terms. Most of the existing semantic similarity measures have used ontology structure as their key source, but they cannot calculate the semantic similarity between words and concepts using several ontologies.

4.1 Extracting Set of Relevant Definitions, Features, Synsets, Neighbors from both Ontologies

In general, ontologies can be distinguished into domain ontologies, representing knowledge of a particular domain, and generic ontologies representing common sense knowledge about the world. There are several examples of general purpose ontologies available including WordNet attempts to model the lexical knowledge of a native speaker of English. English nouns, verbs, adjectives, and adverbs are organized into synonym sets, called synsets, each representing a concept. As well, one of the domain specific ontology designed for medical concepts includes MeSH. Based on the relevant input query keyword, the set of appropriate definitions, features (Hypernyms), synset, neighbors (Hyponyms) are extracted from both the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ontologies, WordNet and MeSH. The sample XML descriptions about the query keywords from both ontologies with the given bio-medical term are shown below.

Table 1. 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>

4.2 Finding Cross Ontology Measure for the Input Query

In order to find the cross ontology measure for the input query, we have found out the semantic similarity measures of the extracted feature sets, synsets, neighborhoods and the definitions of the two different ontologies. The similarity between two different terms is computed as a weighted sum of similarities between synonym sets (synsets), features, neighborhoods and their definitions. Consider the WordNet O_1 and MeSH O_2 ontologies, in which the Query keyword Q consists of Features F, Synsets S, Neighborhoods N and Definitions D obtained from both the ontologies. In addition, we have combined all the chosen features together in a vector named as A_s. Based on the input query, we have to find out the cross ontology measure for every set of features, synsets, neighborhoods and definitions obtained from the ontologies. The set of features, synsets, neighborhoods and definitions obtained from the ontologies O_1 and O_2 are represented as follows,

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

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

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

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

$$A_s = \{F, S, N, D\}$$

The similarity measure $Sim(Q_1, Q_2)$ of the input query keywords Q_1 and Q_2 from ontologies O_1 and O_2 respectively is computed with the aid of the set of features, synsets, neighborhoods and the definitions extracted from both the ontologies. The formula utilized for computing the similarity measure of the corresponding query keyword from the Wordnet and MeSH is given as follows,

$$Sim(Q_1, Q_2) = \sqrt{\frac{\alpha S_f^2(Q_1, Q_2) + \beta S_s^2(Q_1, Q_2) + \gamma S_n^2(Q_1, Q_2) + \delta S_d^2(Q_1, Q_2)}{4}}$$

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

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

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

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

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

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

Also, $S_f(Q_1,Q_2)$, $S_s(Q_1,Q_2)$, $S_n(Q_1,Q_2)$ and $S_d(Q_1,Q_2)$ are the individual similarity measures of the every feature set, synsets, neighborhoods and definitions respectively. Here, the formula for finding the similarity of every set of terms by means of their common universal set of all terms with features, synsets, neighborhoods and the definitions is given in detail.

$$S_f(Q_1,Q_2) = \left(\frac{\left|f^{(1)} \cap f^{(2)}\right|}{\left|f^{(1)}\right| * \left|f^{(2)}\right|}\right) + \left(\frac{\left|\sim f^{(1)} \cap \sim f^{(2)}\right|}{\left|\sim f^{(1)}\right| * \left|\sim f^{(2)}\right|}\right) - \left(\frac{\left|f^{(1)} \cap \sim f^{(2)}\right|}{\left|f^{(1)}\right| * \left|\sim f^{(2)}\right|}\right) - \left(\frac{\left|\sim f^{(1)} \cap f^{(2)}\right|}{\left|\sim f^{(1)}\right| * \left|f^{(2)}\right|}\right)$$

$$S_s(Q_1,Q_2) = \left(\frac{\left|s^{(1)} \cap s^{(2)}\right|}{\left|s^{(1)}\right| * \left|s^{(2)}\right|}\right) + \left(\frac{\left|\sim s^{(1)} \cap \sim s^{(2)}\right|}{\left|\sim s^{(1)}\right| * \sim \left|s^{(2)}\right|}\right) - \left(\frac{\left|s^{(1)} \cap \sim s^{(2)}\right|}{\left|s^{(1)}\right| * \left|\sim s^{(2)}\right|}\right) - \left(\frac{\left|\sim s^{(1)} \cap s^{(2)}\right|}{\left|\sim s^{(1)}\right| * \left|s^{(2)}\right|}\right)$$

$$S_n(Q_1,Q_2) = \left(\frac{\left|n^{(1)} \cap n^{(2)}\right|}{\left|n^{(1)}\right| * \left|n^{(2)}\right|}\right) + \left(\frac{\left|\sim n^{(1)} \cap \sim n^{(2)}\right|}{\left|\sim n^{(1)}\right| * \sim \left|n^{(2)}\right|}\right) - \left(\frac{\left|n^{(1)} \cap \sim n^{(2)}\right|}{\left|n^{(1)}\right| * \sim \left|n^{(2)}\right|}\right) - \left(\frac{\left|\sim n^{(1)} \cap n^{(2)}\right|}{\left|\sim n^{(1)}\right| * \left|n^{(2)}\right|}\right)$$

$$S_d(Q_1,Q_2) = \left(\frac{\left|d^{(1)} \cap d^{(2)}\right|}{\left|d^{(1)}\right| * \left|d^{(2)}\right|}\right) + \left(\frac{\left|\sim d^{(1)} \cap \sim d^{(2)}\right|}{\left|\sim d^{(1)}\right| * \sim \left|d^{(2)}\right|}\right) - \left(\frac{\left|d^{(1)} \cap \sim d^{(2)}\right|}{\left|d^{(1)}\right| * \sim \left|d^{(2)}\right|}\right) - \left(\frac{\left|\sim d^{(1)} \cap d^{(2)}\right|}{\left|\sim d^{(1)}\right| * \left|d^{(2)}\right|}\right)$$

5. PROPOSED SYSTEM FOR BIO-MEDICAL DOCUMENT RETRIEVAL

Today, search engines are being extensively used for retrieving information from several resources all over the world. Where, most of the searches are based on the area of biomedical for obtaining relevant documents from different biomedical databases. Nowadays, search engines are inefficacious in document clustering and representing the relevant level of the documents extracted from the databases. Currently, there is an enormous expansion in the development of new technologies that are being established in each and every field including bioinformatics. The area of bioinformatics found to be lacking in development previously but now, this area found to have an extraordinary expansion comparatively to other areas [27, 28, 29]. Nowadays, due to hastily increasing volume of publications in the biomedical, finding related work is more and more a complicated challenge. Since the biomedical science is very diverse, the solutions for the document search problems are complex and the articles most pertinent to one reader may not be relevant to another. [30,31].

In Biomedical research, the ability to extract the sufficient information from the sprouting literature is an extremely significant asset. Scientific publishing grows at a constant rate and research goals are becoming increasingly focused and intricate. The urge for automatic curation techniques and tools is now greater than ever and the competence to retrieve the proper set of documents about a particular problem is decisive. Biomedical information retrieval is often supported by bibliographic databases and open-access journals. At present, PubMed maintains the largest life science and biomedical bibliographic database, comprising more than 17 million records. Even though providing an excellent service, PubMed search engine is based on user-specified queries, i.e., sets of keywords that the user considers to best represent the query. Achieving a sufficient formulation of a query is not easy. Users may select

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

common words or address broad-scope problems (e.g. a search on "leukemia"). While searching for pertinent documents through such a process, several partly related and unrelated documents will be retrieved additionally. Every document that matches the posted keywords in any of the requested search fields is considered as a candidate. But, it is insignificant for the user to pose its query in such a way that the keywords do not bring attention over documents that are not connected to the subject of their interest [33]. The overall architecture of the proposed bio-document retrieval system is given in figure 1.

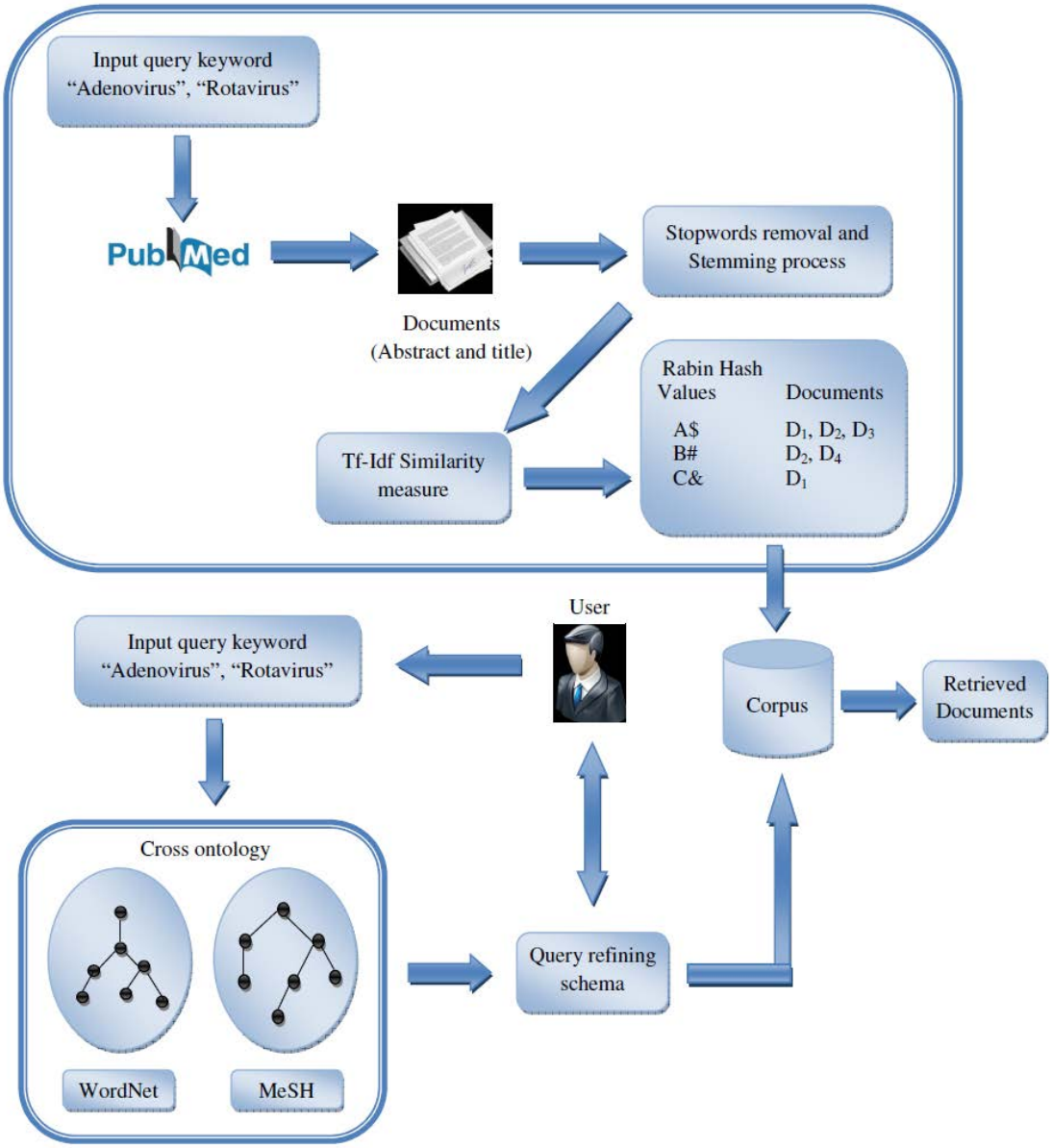

Figure. 1 Block Diagram Of The Proposed Document Retrieval System

5.1 Extracting Feature Vectors from the Documents of Repository

The initial process of the document retrieval system mainly deals with the extensive empirical runs which showed the value of the simple natural language processing techniques of stemming, deletion of stop words, finding similarity measure for extracting the feature vectors. In this, the set of keywords are extracted from the documents as the outcomes of the pre-processing steps. The pre-processing step mainly integrates the stop words

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

removal and the stemming process which results the set of keywords K of all documents in the repository. Subsequently, the features are extracted based on the similarity measure as computed with the aid of the TF-IDF similarity measure.

5.1.1 Pre-Processing

The pre-processing mainly carries out with two steps which includes, 1) Deletion of stop words and 2) Stemming process. After the completion of these processes, the set of significant keywords results as an output from every document by removing the stop words and stemmed some of the words.

a) Deletion of Stop Words: This method filters the document's promising indexing elements against a Stop Word list in order to remove the words, which are considered to be trivial in determining a document's relevance to a user's request. The main purpose of deleting stop words is to preserve the system resources by removing those terms that contain small value for retrieval performance. The general word classes that are marked as stop words comprise the function word classes and a few more (i.e. articles, conjunctions, interjections, prepositions, pronouns, and 'to be' verb forms) [26].

b) Stemming Algorithm: Stemming is the process of acquiring the root words from the derived words that are present in the filtered tokens. The function of stemming is to diminish the storage requirements of the inverted index file via minimizing the number of unique terms. However, stemming has remained in use even today when storage is not a problem, because it improves recall of pertinent documents. For instance, if a query includes the word *study*, the user may desire documents that contain the words *studies, studying, or studied* [40].

5.1.2 Extracting Features from the Documents Based On TF-IDF Similarity

After the pre-processing steps, we find the similarity measure of all keywords extracted from the document repository D. The similarity measure we have utilized here is TF-IDF similarity as given in equation 1. Then, based on the similarity measure, we have taken the set of keywords K with highest score.

a) TF-IDF similarity measure

The term frequency–inverse document frequency (TF–IDF) is a weight usually employed in information retrieval and text mining. This TF-IDF weight is a mathematical measure used to calculate how vital a word is to a document in a group or corpus. The importance increases proportionally to the number of times a word occurs in the document but is equalized by the frequency of the word in the collection. Deviations of the TF–IDF weighting scheme are usually used by search engines as an essential tool in scoring and ranking a document's relevance given a user query. TF–IDF can be efficiently used for stop-words filtering in different subject areas such as text summarization, classification etc [24]. Using the TF-IDF weighting scheme [23], d_t is described as,

$$d_t = (TF_{d,t}) * (IDF_t)$$

Where $TF_{d,t}$ is the number of times that term t occurs in the document represented by d, $IDF_t = N / n_t$, N is the total number of documents in the database, and n_t is the total number of documents in the database that contain the term t.

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

$$D = \{d_1, d_2, ..., d_n\}$$

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

$$d = \{k_1, k_2, ..., k_n\}$$

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 f_v from the corresponding documents.

$$S = d_t(K) \;\; f_v = S > \min\_threshold$$

5.2 Indexing Of Documents

The document retrieval system prepares for retrieval by indexing the documents and formulating the queries, resulting in document representations and query representations respectively. Automatic indexing begins with raw feature extraction, such as extracting all the words from a text, followed by refinements in accordance with the conceptual schema. Here, the indexing is done with the aid of the Rabin's Fingerprint hashing algorithm so that the matching process can be done easily.

Journal of Theoretical and Applied Information Technology
20th August 2013. Vol. 54 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

5.2.1 Indexing using Inverted Indices and fingerprint value using Rabin fingerprint algorithm

a) Rabin fingerprint algorithm

Assume that the character string A is a bit string having m bits $[b_1,\cdots,b_m]$ and it is associated to a polynomial of degree $(m-1)$ in indeterminate t as follows [25]:

$$A(t) = b_1 t^{m-1} + b_2 t^{m-2} + \cdots + b_{m-1}t + b_m$$

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

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

In Rabin's fingerprinting technique, an irreducible polynomial is used for $P(t)$. As we are dealing with bit strings, all the coefficients of $A(t)$ are in Z_2. Hence $P(t)$ will be selected by using $a_i's$ in Z_2. Then, the fingerprinting function for a given character string A is defined as,

$$f(A) = A(t) \bmod P(t)$$

Using the Rabin's fingerprint algorithm, we have calculated the fingerprint value F_{val} for the selected features f_v of the documents with the similarity measure.

$$F_{val} = Hash(f_v)_{Rabin}$$

Subsequently, the indexing process of the documents is carried out based on the fingerprint values of every feature sets. Here, we have applied the inverted index method. An inverted index is an indexing data structure storing a mapping from content, such as words or numbers, to its locations in a database file, or in a document or a set of documents. The purpose of an inverted index is to allow fast full text searches, at a cost of increased processing when a document is added to the database. The inverted file may be the database file itself, rather than its index. In this, every feature is selected and all the corresponding documents having the particular features are being indexed I_D. Meanwhile, the keywords are indexed by their own hash values F_{val}.

$$I_D = \{ f_v \in [D] \mid hash(f_v) \}$$

The sample hashing process is given in the following table 2.

Table 2. Indexing Of Documents

Hashed keywords	Relevant Documents
1010101	d1, d3, d4, d5, d29
1001101	d2, d3, d7, d6, d13, d4, d23, 24
1011100	d8, d10, d17, 18
1000001	d1, d21, d30, d19, d11, d27
1111101	d26, d22, d18, d16,d28

5.3 Retrieving the Relevant Documents

This section describes the retrieval procedure of bio-documents from the input database. When the user provides the input query keywords, the features of the input query words, $A_s = \{F, S, N, D\}$ is obtained from WordNet and MeSH ontologies. Then, the system finds the cross ontology similarity measure for the query keywords using the features extracted from the ontologies. If the similarity measure is less than the user specified threshold, the query refining process is done, means that the user have to check or give alternative relevant keywords. If the similarity measure is above the user specified threshold, the input query is hashed and matched with the indexed document's hash values. If the hash value in the repository matches with the hash value of the input keyword, then we can retrieve the required number of bio-documents relevant to the query keyword.

The pseudo code of the proposed document retrieval system is given as below.

Pseudo code

Input: *Query keywords,* $q_1 \in O_1$; $q_2 \in O_2$

Output: *Relevant Documents* R_D

Assumptions

F → Features
S → Synsets
N → Neighborhoods
D → Definitions
S_m *→ Similarity measure*
I(H,D) → Indexed documents (H refers to hashed keyword and D refers to documents)

Pseudo code

Begin
Get query, $Q = \{q_1, q_2\}$
obtain f eature vectors, $\{F, S, N, D\} \in Q$
$S_m = Sim(Q_1, Q_2)$
if $S_m < thresh$
prompt the user to check Q
else
hashing, $H(q_1)$ *and* $H(q_2)$

Journal of Theoretical and Applied Information Technology
20th August 2013. Vol. 54 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

finding distance, $D_1(H(q_1), I(H,D))$ *and* $D_2(H(q_2), I(H,D))$

if $D_1 > thresh$

$R_D << I(H,D) \in H(q_1)$

end if

if $D_2 > thresh$

$R_D << I(H,D) \in H(q_1)$

end if

end

6. RESULTS AND DISCUSSION

The results obtained from the experimentation of the proposed cross ontology-based similarity measure for bio-document retrieval system is presented in this section. We have implemented our proposed bio-document retrieval system using Java (jdk 1.6). The dataset utilized in our experimental results are bio-medical documents obtained from the PubMed database.

6.1 Experimental Environment And Dataset Description

This experimental environment of proposed bio-document retrieval system is Windows XP Operating system at 2 GHz dual core PC machine with 2 GB main memory running a 64-bit version of Windows 2007.

Dataset Description: We have tested our algorithm in different documents obtained from PubMed database which satisfies the WordNet and MeSH ontologies. PubMed is the National Library of Medicine's search service that provides access to over 11 million citations in MEDLINE. MEDLINE is the premier bibliographic database covering the fields of medicine, nursing, dentistry, veterinary medicine, the health care system, and the preclinical sciences. It contains more than 11 million references and abstracts from over 4000 biomedical journals. From that database, we have chosen only 180 medical for 6 different medical keywords. As well, we have taken 30 documents for every keyword of both the ontologies. The chosen sample keywords of WordNet and MeSH ontologies are shown in table 3.

Table 3. Wordnet And Mesh Terms

WordNet	**MeSH**
Adenovirus	Rotavirus
Anemia	Appendicitis
Pneumonia	Asthma
Carcinoma	Neoplasm
Hypothyroidism	Hyperthyroidism
Pain	Ache

6.2 Evaluation Metrics

An evaluation metric is used to evaluate the effectiveness of document retrieval systems and to justify theoretical and practical developments of these systems. It consists of a set of measures that follow a common underlying evaluation methodology. Some of the metrics that we have chosen for our evaluation purpose are Recall, Precision and the F-measure.

Precision,

$$P = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|}$$

Recall,

$$R = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{relevant\ documents\}|}$$

F- Measure, $F = \frac{2PR}{(P+R)}$

As suggested by above equations in the field of Document retrieval, ***Precision*** is the fraction of retrieved documents that are relevant to the search, ***Recall*** is the fraction of the documents that are relevant to the query that are successfully retrieved and the ***F-measure*** that combines precision and recall is the harmonic mean of precision and recall.

6.3 Performance Analysis Over The Proposed Cross Ontology Based Similarity Measure

The performance of the proposed retrieval system is analyzed over by the proposed cross ontology based similarity measure along with the X-similarity measure [42] and the Rodriguez M.A's [41] similarity measure. Here, we the similarity measures of the [41, 42] are taken from the semantic similarity system intelligence laboratory. They have analyzed by their own similarity measures with the aid of the WordNet and the MeSH ontology terms. In this, some of the medical terms fail to reach the similarity values of the existing ones, in which our proposed cross ontology based similarity measure performs better results. Table 4 lists the comparative values obtained by the proposed similarity measure and the existing works.

Journal of Theoretical and Applied Information Technology
20th August 2013. Vol. 54 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 4. Cross Ontology Based Similarity Measure Comparison

Query keyword		X-similarity measure [42]	Rodriguez M.A [41]	Proposed similarity measure
WordNet	MeSH			
Adenovirus	Rotavirus	0.16	0.018666667	0.03406453
Anemia	Appendicitis	0	0	0.02938514816
Pneumonia	Asthma	0.07	0.0119	0.01566590728
Carcinoma	Neoplasm	0.17	0.04	0.0569153419
Hypothyroidism	Hyperthyroidism	0.387	0	0.0871796247
Pain	Ache	1	0.021666667	0.04950827
Dementia	Atopic Dermatitis	0	0	0.044338768
Malaria	Bacterial Pneumonia	0.113	0	0.04502309
Osteoporosis	Patent Ductus Arteriosus	0.122	0	0.2681062
Sinusitis	Mental Retardation	0	0	0.075982524
Urinary Tract Infection	Pyelonephritis	0.03	0.01	0.1153284073
Iron Deficiency Anemia	Sickle Cell Anemia	0.14	0.01166667	0.060882246

6.4 Performance Analysis Over The Proposed Bio-Document Retrieval System Using Evaluation Metrics

The performance of the proposed retrieval system is evaluated based on the input query keywords of WordNet and MeSH ontologies using the Precision, recall and F-measure. Here, we have utilized six 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. The table 5 lists the obtained values for the evaluation measures with different keywords and the relevant documents as 30. It reveals that the proposed system works fine in the medical document retrieving process.

Table 5. Precision, Recall And F-Measure For Different Keywords

Query keyword		Relevant documents	Retrieved documents	Precision	Recall	F-measure
WordNet	MeSH					
Adenovirus	Rotavirus	30	31	0.967742	1	0.983607
Anemia	Appendicitis	30	31	0.967742	1	0.983607
Pneumonia	Asthma	30	31	0.967742	1	0.983607
Carcinoma	Neoplasm	30	38	0.789474	1	0.882353
Hypothyroidism	Hyperthyroidism	30	30	1	1	1
Pain	Ache	30	28	0.607143	0.566667	0.586207

In addition, the results are obtained 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 2, 3 and 4 respectively. By analyzing the graphs, when the threshold is fixed as 0.4, the proposed system achieved maximum precision compared with other threshold values.

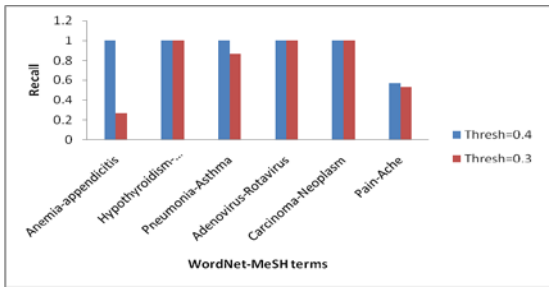

Figure. 3 Recall Graph Plotted For Different Similarity Thresholds

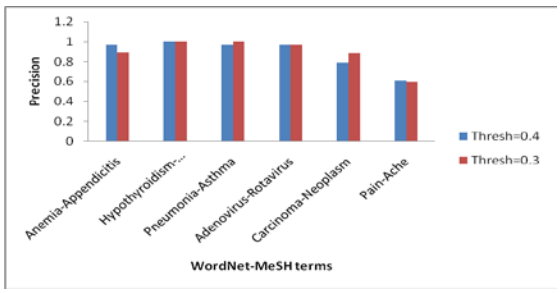

Figure. 2 Precision Graph Plotted For Different Similarity Thresholds

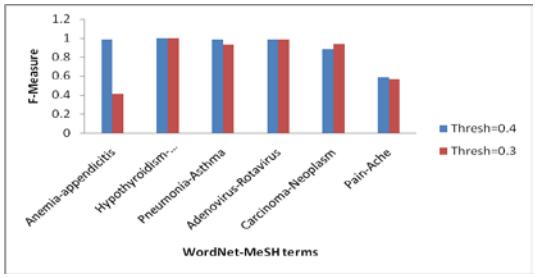

Figure. 4 F-Measure Graph Plotted For Different Similarity Thresholds

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

7. CONCLUSION

In this paper, we have presented an effective bio-medical document retrieval system with cross ontology based similarity measure. As well, the experimentation is carried out with the aid of the PubMed database documents. The performance of the proposed retrieval system is analyzed by means of the two existing similarity measure with the proposed cross ontology based similarity measure for different medical terms. As well, we have evaluated our proposed system with standard evaluation metrics like Precision, recall and F-measure and achieved more than 95% accuracy for most of the medical terms.

REFERENCES

[1] Pedersen, T., Pakhomov, S., Patwardhan, S., and Chute, C, "Measures of semantic similarity and relatedness in the biomedical domain", *Journal of Biomedical Informatics*, Vol: 40, No: 3, 288-299, 2007.

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

[3] 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.

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

[5] Neches, R.E. Fikes, T. Finin, T.R. Gruber, T. Senator and W. R. Swartout, "Enabling Technology for Knowledge Sharing", *AI Magazine*, Vol: 12, No: 3, pp. 36-56, 1991.

[6] T.R. Gruber, "A Translation Approach to Portable Ontology Specification", *Knowledge Acquisition*, Vol: 5, No: 2, pp. 199-220, 1993.

[7] A. Gomez-Perez, M. Fernandez-Lopez and O. Corcho, "Ontological Engineering", *Berlin: Springer-Verlag*, 2004.

[8] 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.

[9] Giannis Varelas, Epimenidis Voutsakis, Paraskevi Raftopoulou, Euripides G.M. Petrakis, Evangelos E. Milios, "Semantic Similarity Methods in WordNet and their Application to Information Retrieval on the Web", *Proceedings of the 7th annual ACM international workshop on Web information and data management*, 2005.

[10] Lixin Han, Guihai Chen, "A fuzzy clustering method of construction of ontology-based user profiles", *Journal of Advances in Engineering Software*, Vol: 40, No: 7, 2009.

[11] Heflin J, Hendler J. "Searching the web with SHOE". *AAAI-2000 workshop on AI for Web search, California: AAAI Press*; 2000.

[12] Aleman-Meza B, Halaschek C, Arpinar IB, Sheth A. "Context-aware semantic association ranking", *Proceedings of the Semantic web and databases workshop, Berlin, Germany*, September 7–8, 2003.

[13] F.S.C. Tseng, "Design of a multi-dimensional query expression for document warehouses", *Information Sciences*, Vol: 174, No: 1–2, pp: 55–79, 2005.

[14] R. Feldman, "Text mining: theory and practice", Proceedings of the Fourth World Congress on Expert Systems-Application of Advanced Information Technologies, *ITESM Mexico City Campus*, 1998.

[15] D.A. Guerra-Zubiaga, "A manufacturing model to enable knowledge maintenance in decision support systems", *PhD Thesis, Wolfson School of Mechanical and Manufacturing Engineering, Loughborough University, UK*, 2004.

[16] 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, Pages 1-12, 2006.

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

[18] Stephen L. Reed and Douglas B. Lenat, "Mapping Ontologies into Cyc", *ACM*, Vol: 38, No: 11, pp: 33- 38, 2002.

[19] 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.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[20] 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.

[21] Angelos Hliaoutakis, "Semantic Similarity Measures in MeSH Ontology and their application to Information Retrieval on Medline", *Technical report*, 2005

[22] Zharko Aleksovski, Warner ten Kate, and Frank van Harmelen, "Ontology matching using comprehensive ontology as background knowledge", *In P. Shvaiko et al., editor, Proceedings of the International Workshop on Ontology matching at ISWC*, pp: 13-24, 2006.

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

[24] 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.

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

[26] Elizabeth D. Liddy, "Document Retrieval, Automatic", Encyclopedia of Language & Linguistics, *2nd Edition. Elsevier*, 2005.

[27] Jayanthi Manicassamy and P. Dhavachelvan, "Metrics based performance control over text mining tools in bioinformatics", *ACM Portal*, pp: 171-176, 2009.

[28] Jayanthi Manicassamy and P. Dhavachelvan, "Based Accuracy Perpetuation for Bioinformatics Sequence Analysis Tools", *International Journal of Recent Trends in Engineering (IJRTE)* - Finland, pp: 550-555, May 2009.

[29] Marta Sabou, Chris Wroe, Carole Goble, Gilad Mishne, "Learning domain ontologies for web service descriptions: An experiment in bioinformatics", *citeseer*, 2005.

[30] Richard Tzong-Han Tsai, Shih-Hung Wu, Wen-Chi Chou, Yu-Chun Lin, Ding He, Jieh Hsiang, Ting-Yi Sung and Wen-Lian Hsu, "Various criteria in the evaluation of biomedical named entity recognition", *PubMed*, pp: 7-92, 2006.

[31] Jung-jae Kim, Piotr Pezik and Dietrich Rebholz-Schuhmann, "MedEvi: Retrieving textual evidence of relations between biomedical concepts from Medline", *ACM portal*, pp: 1410–1412, March, 2008.

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

[33] Martijn J. Schuemie, Jan A. Kors, and Barend Mons. "Word sense disambiguation in the biomedical domain: an overview", *Journal of Computational Biology*, Vol: 12, No: 5, pp: 554–565, 2005.

[34] 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.

[35] 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.

[36] Ali, B.; Abdelkrim, B.; Mebarek, S., "Term proximity in document retrieval systems", Proceedings of the 2011 *IEEE International Conference on Computer Science and Automation Engineering (CSAE)*, Shanghai, pp: 267 - 271, 2011.

[37] Dang Tuan Nguyen, "Interactive document retrieval system based-on natural language query processing", *Proceedings of the International Conference on Machine Learning and Cybernetics, Baoding*, pp: 2233 - 2237, 2009.

[38] 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.

[39] 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.

[40] Willett P, "The Porter stemming algorithm: then and now", *Program: electronic library and information systems*, Vol: 40, No: 3, pp. 219-223, 2006.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[41] Rodriguez M.A. and Egenhofer M.J, "Determining Semantic Similarity among Entity Classes from Different Ontologies," *IEEE Transaction on Knowledge and Data Engineering*, vol. 15, no. 2, pp. 442-456, 2003.

[42] Euripides G.M. Petrakis, Giannis Varelas, Angelos Hliaoutakis and Paraskevi Raftopoulou "X-Similarity: Computing Semantic Similarity between Concepts from Different Ontologies," *Journal of Digital Information Management*, vol.6, 2006.

[43] 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.