ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 4, April 2018

Ontological Information on Tumor Knowledge Management Using Text Mining Techniques

P. Jyotsna[1], Prof. P. Govindarajulu[2]

Research Scholar, Department of Computer Science, Sri Venkateswara University, Tirupati, Andhra Pradesh, India [1]

Professor, Department of Computer Science, Sri Venkateswara University, Tirupati, Andhra Pradesh, India[2]

ABSTRACT: Large numbers of documents are grouped according to their similarities. The Text Mining methods have been proposed to solve the problem by automatically classifying text documents. An ontology based text mining approach is to cluster the documents based on their similarities. For ontology making in text based medical documents. This paper focuses on an annotation scheme that combines the semantic similarity measures and indexing in the document. Text mining classification algorithms and Computational methods we can provide the awareness of patients and quality of healthcare by having, problem solving and decision-making systems. Tumor based Information systems can help in supporting clinical care in addition to helping Stages of Tumor Grades. The approach can be seen as evolution of the keyword indices are replaced by ontology knowledge representative and a Pre-automatic document annotation weighting procedure that improves the retrieval process. Evaluation results show that the proposed approach can easily outperform the traditional approach.

KEYWORDS : Text Mining, Ontology, similarity, indexing, and Classification,

I. INTRODUCTION

The text mining techniques as become a widespread approach to identify and extract information from unstructured text. Text mining is used to extract facts and relationships in a structured form that can be used to annotate specialized databases, to transfer knowledge between domains and more generally within business intelligence to support operational and strategic decision-making. Bio-medical text mining is concerned with the extract of information regarding medical records from a variety of sources. Text classification is the task of automatically sorting a set of documents into categories from a pre-defined set [9]. The resources of unstructured and structured information include the www, electronic medical repositories, biological databases, etc. Therefore, proper classification and knowledge representative from these resources is an important area for research.

Text Mining is the process of Analyzing collections of text resources in order to capture key concepts and themes and uncover hidden relationships and trends without requiring that you know the precise words or terms that authors have used to express those concepts. although they are quite different, text mining is an interdisciplinary field that enormous challenge, the extract and management of quality content, terminology, and additionally, combines methodologies from various other areas such as Information Extract (IE), Information Retrieval (IR), Computational Linguistics', Categorization, Topic Tracking and Concept Linkage(TCL).

Linguistic Processing and NLP

Linguistics-based text mining, on the other hand, applies the principles of natural language processing (NLP)—the computer-assisted analysis of human languages—to the analysis of words, phrases, and syntax, or structure, of text. A system that incorporates NLP can intelligently extract concepts, including compound phrases. Moreover, knowledge of the underlying language allows classification of concepts into related groups, such as products, organizations, or people, using meaning and context. Data Mining and Machine Learning techniques work together to automatically classify and discover database patterns from the electronic medical reports. The main goal of text mining techniques is to enable users to extract information from textual resources and deals with operations like Searching and Indexing, Classification (supervised,unsupervised)and summarization.

However how these reports can be properly annotated, presented and classified. So it consists of several challenges, like proper annotation to the documents, appropriate document representation, dimensionality reduction to

DOI: 10.15680/IJIRCCE.2018.0604037

handle algorithmic issues [1], and an appropriate classifier function to obtain good generalization and avoid over-fitting.

The need of automatically retrieval of useful knowledge from the huge amount of textual data in order to assist the Key word and tumor ontology based searching and indexing For these purpose state-of-the-art approaches to text classifications are presented in [6], So constructing a data structure that can represent the medical reports, and constructing a classifier that can be used to predicate the class label of a document with high accuracy, are the key points in text classification. One of the purposes of research is to review the available and known work, so an attempt is made to collect what's known about the documents classification and representation.

This paper covers the overview of Keyword based Searching and Indexing matters, medical informatics ontology focused on the different techniques for text classification using the existing literature. The motivated perspective of the related research areas of text mining are: Information Extract(IE) methods is aim to extract specific information about tumor from text based on historical medical reports . This is the first approach assumes that text mining essentially corresponds to information extract about tumors.

The widespread adoption of electronic medical reports has resulted in increased availability of free text clinical data, usually in the form of plain text reports dictated or typed by clinicians, for secondary use. Because free text clinical data must be converted to annotated information to realize its full value, analyzing and extracting pertinent information from unstructured clinical text has become an increasingly important activity within the healthcare industry.

The keyword based Searching and Indexing systems are available for accessing the information but the retrieval of relevant information is still a problem. One current problem of information retrieval is that it is not really possible to extract relevant documents automatically. An information retrieval system uses indexing and the system's performance depends on the quality of the indexing. The two main challenges in indexing are to create representative internal descriptions of documents and to organize these descriptions for fast retrieval. Descriptions of documents in information retrieval are supposed to reflect the documents content and establish the foundation for the retrieval of information when requested by users. The process of assigning descriptions to documents in an information retrieval system is called indexing.

Medical ontology has a good conceptual structure representation and it can be combined with knowledge representation. This model makes use of annotation and indexing. The ontology model depends on the semantic index terms but the vector space model depends on the keyword index. The semantics of the concepts are used to build a concept term representation. The medical ontology similarity measure improves the concept relevance score.

The semantically related terms gain more weights and it will improve the term importance in indexing process. The semantic analysis should recognize concepts in the documents and then map them into the medical ontology. The indexing process maps information found in documents into the ontology, identifying concepts and their positions in the medical reports in queries can similarly be mapped into the ontology and thus in addition to retrieving the exact match, the structure of the ontology can be used to retrieve semantically related documents. Document is composed of many terms and important words are present in documents. Document should be preprocessed to obtain semantic annotations and indexing of the document should be done. Most indexing systems depend on frequency of term within a document. The information retrieval can be improved by combining the ontology similarity and ontology indexing method. In this model, the weight of the concepts is computed using their semantic similarity to other concepts in the document. The concept vector is generated in the document annotation process and the concept index is built.

This approach aims to enhance the semantic annotation and indexing of document units by enhancing both syntactic and semantic matching. The indexing process is divided into two parts, the conceptual analysis and a transformation process, where the former extracts information from documents and the latter creates descriptions in accordance with the description notation. To improve the recognition of important indexing terms, it is possible to weigh the concepts of a document in different ways (Valkeapaa et al 2007).

Indexing can be performed either manually or automatically by computers. In manual indexing, experts assign the descriptions, while automatic indexing is performed by computers. The two main challenges in utilizing ontologies in information retrieval are

(i) To map the information in documents and queries into the Medical reports.

(ii) To improve retrieval by using knowledge about relations between concepts in the medical manuscript.

1.1 Text Classification Process

1.1.1. Documents Collection

This is first step of classification process in which we are collecting the different types (format) of document like .html, .pdf, .doc, web content etc.

1.1.2. Pre-Processing

The first step of pre-processing which is used to presents the text documents into clear word format. The documents prepared for next step in text classification are represented by a great amount of features. Commonly the steps taken are:

- Tokenization: A document is treated as a string, and then partitioned into a list of tokens.
- Removing stop words: Stop words such as "the", "a", "and", etc. are frequently occurring, so the insignificant words need to be removed.
- Stemming word: Applying the stemming algorithm that converts different word form into similar canonical form.

This step is the process of conflicting tokens to their root form, e.g. connection to connect, computing to compute.

1.1.3. Indexing

The documents representation is one of the pre-processing technique that is used to reduce the complexity of the documents and make them easier to handle, the document have to be transformed from the full text version to a document vector that Perhaps most commonly used document representation is called vector space model (SMART) vector space model, documents are represented by vectors of words. Some of limitations are: high dimensionality of the representation, loss of correlation with adjacent words and loss of semantic relationship that exist among the terms in a document. To overcome these problems, term weighting methods are used to assign appropriate weights to the term.

1.1.4. Feature Selection

After pre-processing and indexing the important step of text classification, is feature selection to construct vector space, which improves the scalability, efficiency and accuracy of a text classifier. The main idea of Feature Selection (FS) is to select subset of features from the original documents. FS is performed by keeping the words with highest score according to predetermined measure of the importance of the word. Because of for text classification a major problem is the high dimensionality of the feature space. Many feature evaluation metrics have been notable among which are information gain (IG), term frequency, Chi-square, expected cross entropy, Odds Ratio, the weight of evidence of text, mutual information, Gini index.

1.1.5 Classification

The automatic classification of documents into predefined categories has observed as an active attention, the documents can be classified by three ways, unsupervised, supervised and semi-supervised methods [1]. From last few years, the task of automatic text classification have been extensively studied and rapid progress seems in this area, including the machine learning approaches such as Bayesian classifier, Decision Tree, K-nearest neighbor(KNN), Support Vector Machines(SVMs), Neural Networks, Rocchio's.

TEXT DOCUMENT CLASSIFICATION PROCESS

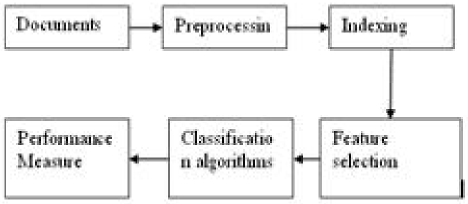


Figure : 1 Document Classification Process

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 4, April 2018

II. RELATED LITERATURE WORK

The Ontology based text mining framework is used to increase the effectiveness of the project, which is described in [8]. The Ontology based text mining method and Information Extraction is used for biological domain is explained in [9]. In pattern discovery for text mining [10] and the Ontology based concept weighing [11], different concepts are compared. The Ontology based text mining techniques uses different clustering methods. The clustering methods to cluster the text documents are discussed in [12].

Stavrianou et al. (2007)[15] present a survey of semantic issues of text mining, which are originated from natural language particularities. This is a good survey focused on a linguistic point of view, it discuss es some existing text representation approaches in terms of features, representation model, and application task. also present the relation between ontologies and text mining. Ontologies can be used as background knowledge in a text mining process, and the text mining techniques can be used to generate and update ontologies. It conclude the survey stating that text mining is an open research area and that the objectives of the text mining process must be clarified before starting the data analysis, since the approaches must be chosen according to the requirements of the task being performed.

Vandana Korde et al (2012)[1] discussed that the text mining studies are gaining more importance recently because of the availability of the increasing number of the electronic documents from a variety of sources which include unstructured and semi structured information. The main goal of text mining is to enable users to extract information from textual resources and deals with the operations like, retrieval, classification (supervised, unsupervised and semisupervised) and summarization, Natural Language Processing (NLP), Data Mining, and Machine Learning techniques work together to automatically classify and discover patterns from the different types of the documents

Wimalasuriya and Dou (2010)[17] present a detailed literature review of ontology-based information extraction. The authors define the recent information extraction subfield, named ontology-based information extraction (OBIE), identifying key characteristics of the OBIE systems that differentiate them from general information extraction systems. Besides, they identify a common architecture of theOBIE systems and classify existing systems along with different dimensions, as information extraction method applied, whether it constructs and updates the ontology, components of the ontology.

Grobelnik M (2011)[16] presents, briefly but in a very clear form, an interesting discussion of text processing in his paper. It organizes the field in three main dimensions, which can be used to classify text processing approaches: representation, technique, and task.

Y. Megan Kong a, Carl Dahlke b, et., al., (2011)[14] toward an ontology-based framework for clinical research databases were developed the Ontology-Based extensible data model (OBX) to serve as a framework for clinical research data in the Immunology Database and Analysis Portal (ImmPort). By designing OBX around the logical structure of the Basic Formal Ontology (BFO) and the Ontology for Biomedical Investigations (OBI), they have found that a relatively simple conceptual model can represent the relatively complex domain of clinical research. In addition, the common framework provided by BFO makes it straightforward to develop data dictionaries based on reference and application ontologies from the OBO Foundry.

Seyed Abbas Mahmoodi1*, Kamal Mirzaie2 and Seyed Mostafa Mahmoudi3(2016)[19] article indicates a new method to discover association rules using ontology to solve the expressed problems. This paper reports a data mining based on ontology on a medical database containing clinical data on patients referring to the Imam Reza Hospital at Tabriz. The proposed data mining algorithm based on ontology makes rules more intuitive, appealing and understandable, eliminates waste and useless rules, and as a minor result, significantly reduces Apriori algorithm running time. The experimental results confirm the efficiency and advantages of this algorithm.

R.Savitha, Dr.R.Porkodi(2015)[13] presents An Overview of Text Mining Techniques and Methodologies Used in Bioinformatics Domain reviews various techniques proposed by various researchers such as new techniques, algorithms, tools and methodologies like My Med, Disco TEX, dictionaries. These techniques are employed to lessen the burden of information overload by applying it to the vast data source.

III. MEDICAL ONTOLOGY ON SIMILARITY AND INDEXING

In order to generate the particular medical reports efficiently, the constructing process contains three main steps are term similarity processing, document analysis and concept clustering.

Step 1: Similarity processing, the synonymous relation among keywords is extracted. The synonymous relation is used to describe the semantic concepts of documents.

Step 2: The degree of synonymous relation can be measured by semantic similarity measure method. In this step, the word similarity for keywords is calculated by similarity measure of WordNet.

Step 3: Document analysis first selects the significant keywords from documents as the keyword space.

Step 4: The annotation process of a web page is done with concepts of the ontology. Then, relations between individuals are discovered and instances are added.

The document is annotated by the following steps and is shown in Figure 2.

Step 1: The document should be preprocessed to obtain semantic annotations.
Step 2: The annotated instance location in the document is stored.
Step 3: The dimension reduction techniques are applied for feature extraction
Step4: The annotation weights are computed, by combining the frequency and concept weights.
Step 5: The documents in the corpus are indexed and
Step 6: Retrieval of documents related with query and ranking the document is done by the relevance of the query.

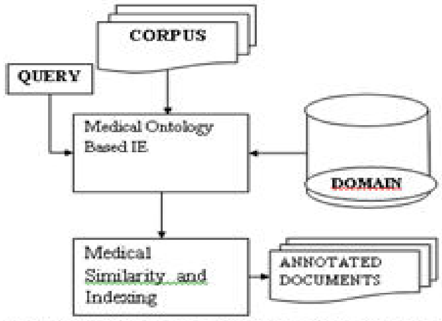


Fig: 2 Medical Ontology-based similarity and Indexing model

3.1. Knowledge Management

The documents are annotated with concept instances from the Knowledge representative by creating instances of the annotation class. The semantic information retrieval Knowledge representative has been built and associated to the information document base by using domain ontology's that describe the concepts. The knowledge engineering of a document defines the concepts and relations. The medical ontology model includes those details that are relevant to the content of the document. The similarity measure considers the conceptual structure of the model. The words in the documents are mapped to the corresponding domain concepts. Resource Description Framework (RDF) is a specification describing and interchanging semantic metadata representation on the document. The document is represented using the ontology indexing method given in onto Search system (Jiang and Tan 2006). Annotation of the documents represents the weight and the importance of the concept.

3.2.Semantic similarity using 'Wordnet'

Word Net is a semantic network, which is organized in such a way that synsets and word senses are the nodes of network, and relations among the synsets and wordsenses are the edges of the network. (Fellbaum 1998). In WordNet, each meaning of a word is represented by a unique wordsense of the word, and a synset (stands for "synonym set") is consisting of a group of wordsenses sharing the same meaning. More than two thirds of the nodes in WordNet are synsets. Hyponym Of is the key relationship for noun synsets in WordNet, which has been widely used to estimate the semantic relatedness among nouns.

Semantic similarity represents the concept similarity between the two words. The semantic similarity method depends on ontology. Using medical ontology's prevents synonyms and misspelling problems during document annotation. Ontology similarity measures can be defined using Wordnet synsets. Euzenat and Shvaiko (2007) proposed the method for using WordNet as a resource for matching terms. The terms are similar if they belong to some common synset. The distance measure evaluates the distance between the synset and the terms. The edge count method proposed by Wu and Palmer (1994) is used to lengths of all edges on the shortest path are accumulated to quantify the semantic similarity.

The semantic similarity measure uses the ontology to identify the relations between concepts. The concept frequency is calculated and the weights are assigned to concepts in the document. The documents are represented by the concept weight vector. Wu and Palmer (1994) proposed the edge-based method using WordNet to define the similarity between two concepts. The similarity between the concepts C_1 and C_2 which is represented as sim(C_1, C_2), is calculated using the following equation.

$$\text{sim}\,(c1, c2) = \frac{2 * d\,(Least\ common\ factor)}{d(c1) + d\,(c2)}$$

The synset from Word Net taxonomical hierarchy are taken to compute the similarity. The depth from the root node to the word 'a' is represented by depth(a). LCF is the least common factor concept of 'c1' and 'c2'. The importance of ontological concepts assigns the weights to concepts and relations. The weight calculated reflects how the concept is relevant to the annotated document. The concept-document matrix is created using medical ontology.
The steps followed in the information retrieval process are as follows:
Step 1: To identify sources of information relevant to the areas of the target user community
Step 2: To analyze the contents of the sources and contents of analyze sources that will match queries
Step 3: Analyze user queries that will match with database.
Step 4: Semantic similarity measure is used to recalculate the weight measure of each term by using Wu and Palmer (1994).
Step 5: Retrieve the information that is relevant
Step 5: Indexed documents are returned.

This approach combines the use of the class hierarchy in the medical ontology, the terms frequency in the collection and the relationships in the ontology to compute the similarity. In this algorithm the similarity of each term in the index with the query is computed, represented by the weighted average of the similarity with all query terms. Different methods are used in the term similarity algorithm. The Get paths method returns all the paths and their lengths. The Getsynset returns the synset of given terms. The Get class returns the class of the term. The common parent returns the common parents between the class. The Least Common factor concept is returned by using get LCF. The depth method returns the depth of the word from the root of the ontology. The similarity between two terms in the ontology is computed through the algorithm shown as below

```
Algorithm : sim (c1, c2)
Input : Medical ontology and terms
1. Find the synset for the term c1 and c2
   if c1 ==c2 Returns 1.0
   $Syset_1$ = getsynset(c1)
   $Syset_2$ = getsynset (c2)
2. For every sysnet and term identify the class
   cls1 = getclass(c1)
   cls2 = getclass(c2)
3. if cls1 or cls2 is empty return 0.0
4. if intersection (cls1, cls2) is not empty Return 1.0;
5. Get the path for the class
   path class 1= get paths(cls1)
   path class 2 = get paths(cls2)
6. Find the common parent of the class
   cls_fac = commonparent (cls1, cls2);
7. Calculate the similarity of the term
   LCF =getLCF(cls1, cls2)
   if cls_fac is not empty {
   sim = ((2*LCF/d_cls1 + d_cls2))
   Output : similarity of terms.
```

DOI: 10.15680/IJIRCCE.2018.0604037

3.3. Reduction dimension

The Singular value decomposition (SVD) program calculates the best reduced dimension approximation for the transformed term-document matrix. This reduced dimensional representation is used for determining the appropriate documents. This approach measures the similarity between the search and the document collection by the weighted inner product of overlapping terms. Vector Space model, however, has certain inherent restrictions as it does not account for the order and association between terms. The concept-document matrix created using ontology and dimension algorithm is applied on the matrix.

The index structure maintains document and concept tables. The document table contains the document identifier and the list of concepts that occur in the documents. The concepts are listed according to an order based on the similarity relevance measure. The concept table contains the concept identifier and list of documents identifiers in which the concept appears. Each index entry gives the word and a list of documents in which the word occurs.

The algorithm for indexing process is given as follows:

3.4. Algorithm: OntoIndex

```
Input: Corpus
1.      Enter the document identifier from the document collection into the index
2.      For each token in the document
            1 Count the occurrence of tokens in the document
            2 Count the occurrence of concepts in the document
            3 Compute the concept weight of the token
            4 Add the concept weight and token frequency to the index structure
3.      Add the document information to index structure
Output: Index structure
```

3.5. Ontology-base Term Unstructured Text

The computational approaches for automation have gained popularity in the biomedical field from decades. These procedures need to extract data efficiently, aggregate, annotate and store information from these unstructured texts. However, unstructured data presents significant challenges for computational methods. Author represent results in natural language using different word choices and different sentence structures. Algorithms need to understand and link synonyms for named entities like Tumors, TNM staging, etc. Computational approaches need to overcome differences in sentence structures (like “X decreases Y”, “Y is decreased by X”) and word choices (like ‘decrease’, ‘reduce’) which leads to a formal representation of text for better analysis. Moreover, the scale of the unstructured data available for analysis is increasing rapidly which lays the emphasis on scalability issues. Performance and scalability are important features for these computational approaches. The word `Concept' is associated with many controversies related to its representation and relevance in the biomedical domain. In our research, we refer concept as any ontologically defined term. Throughout the research, we use different terms like entities, nominal’s and concepts which all mean the same. Ontologies are a formal way of representing knowledge in which terms have a clear, unambiguous meaning which makes them suitable for representing semantics in an ambiguous unstructured text.

IV. RESULTS AND ANALYSIS

The test collection and experimental results are discussed as follows:

4.1 Test Data

The semantic similarity analysis collects the set of keywords or terms that occur frequently together and then finds the relationship among them. The semantic information retrieval Knowledge management has been built and associated to the information document base by using domain ontology that describes the concepts. The query model can find and manipulate the needful data from the annotated documents. The performances of the proposed methods are evaluated using database documents collected from tumor domain. According to tumor domain the major kinds of topics are primary tumor, (T),Regional lymph nodes (N) and distant metastasis (M) impacts. The concepts will be defined as the class and some attributes, which describe the document information. The predefined base ontology described using Scientific directory and provides the basis for the semantic indexing of documents with nonembedded annotations. The document annotation creates the instance of the class.

Once the experimental setting has been set up, the retrieval is tested with Information Retrieval functionality in Natural Language Processing (NLP). NLP comes with a full-featured information retrieval subsystem. The NLP architecture has enabled us

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 4, April 2018

not only to develop a number of successful applications for various language processing tasks, but also to build, to annotate corpora and carry out evaluations on the applications generated. In NLP information retrieval the documents can be retrieved from the corpora not only by their textual content but also according to their features or annotations.

Table 1 Sample Queries

In situ	Abnormal cell are present
Localized	Tumor is limited to the place, no sign that it has spread.
Regional	Tumor has spread to nearby lymph nodes, tissues, or organs
Distant	Tumor has spread to distant parts of the body.
Unknown	Enough information to figure out the stages.

Table 2 Top-5 search Ratings for 5 queries

Keyword Queries	Keyword Similarity	Ontology Similarity
In situ	2.20	3.13
Localized	2.45	3.70
Regional	1.60	3.45
Distant	2.91	3.65
Unknown	2.45	3.56

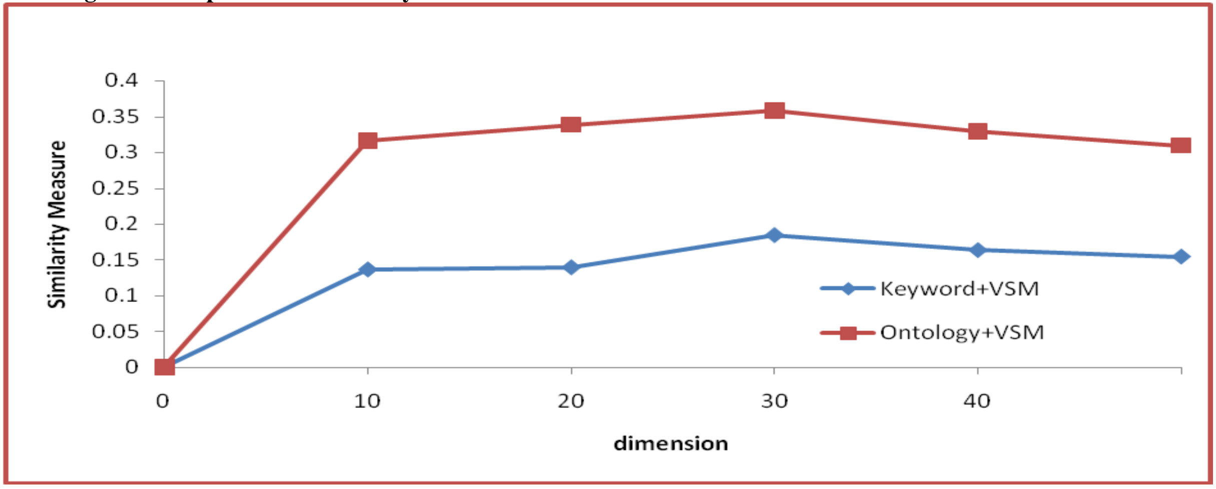
Table 1 shows the set of sample queries. The documents have been annotated and stored. Table 2 shows the search rating for the queries. The similarity degree of each document was evaluated. The average time for responses and comparison of average precision is given in Table 3. The result shows the performance of semantic information retrieval combined with keyword-based retrieval.

Table 3 Average time for response

Number of Keywords	2	3	4
Time (msec)	3.21	5.3	8.1

The document is represented by the dimensional concept vectors. The concept vector specifies the occurrence of concepts in the document. Figure 3 depicts the dependency between the number of dimension and similarity measure. The selected concepts may be used to indicate to the user, which concepts were most relevant for the particular domain.

Figure:3 Comparison of similarity measure for the dimension



DOI: 10.15680/IJIRCCE.2018.0604037

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 4, April 2018

Table 4 Comparisons of Average Precision

Average Precision	Keyword similarity	Ontology similarity	Ratio
Top 5	0.246	0.334	1.36
Top 10	0.188	0.201	1.07

Table 4 shows the comparison of the keyword and ontology retrieval. The Table 4 shows that the method can improve precision by 10% from 0.246 to 0.334 in relevant measure. The performance of the two methods is compared using precision and recall. That the F measure is defined using precision and recall shows the accuracy of the methods. Table 5 shows the F-measure value of the similarity measure and the ontology-based similarity has improved results than keyword-based similarity.

Table 5 Comparisons of keyword and ontology distance measure

Methods	Precision	Recall	F-Measure	Time in Minutes
Keyword Similarity	0.663	0.552	0.595	12.22
Ontology Similarity	0.674	0.629	0.646	15.34

Figure : 4 Comparison of precision and recall for similarity measure

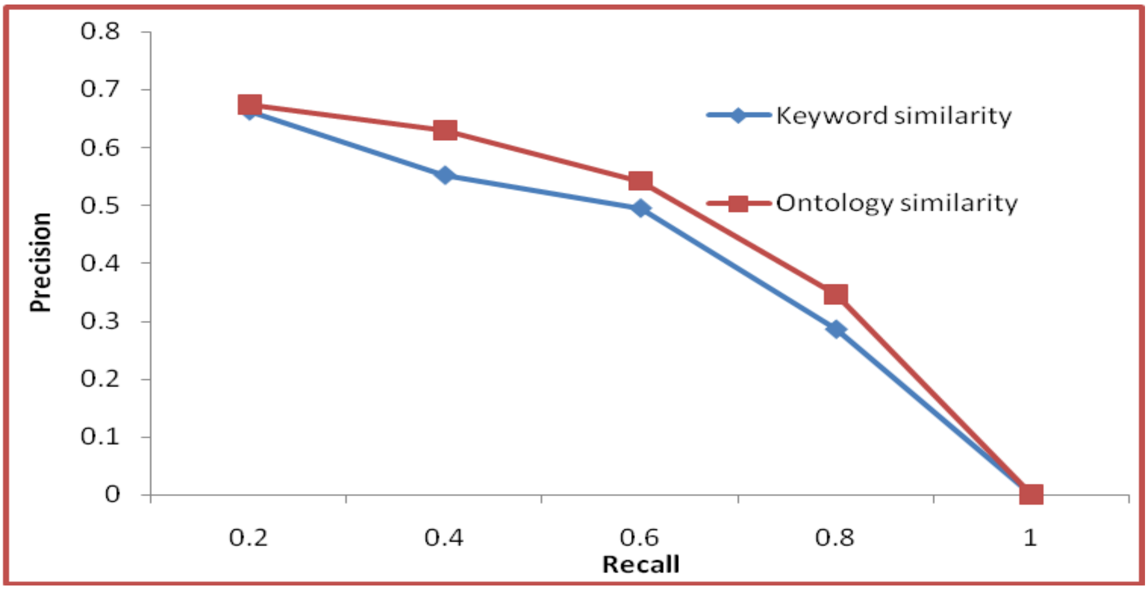


Figure 4 shows that the performance of retrieval on document annotation and without annotation. Instead of simple keyword index lookup, the semantic search system processes a semantic query against the KB, which returns the relevant document. Semantic retrieval is achieved by implementing a document ranking for Lucene indices so that documents containing ontological information get higher rates. Better precision is achieved by using structured document annotation weight and the average precision for the top 5 and 10 documents is shown in Table 4. Table 5 concludes the ontology distances are more accurate compared to keyword and the F-measure value is improved from 0.595 to 0.674 in relevant measure.

Figure: 5 Average execution time

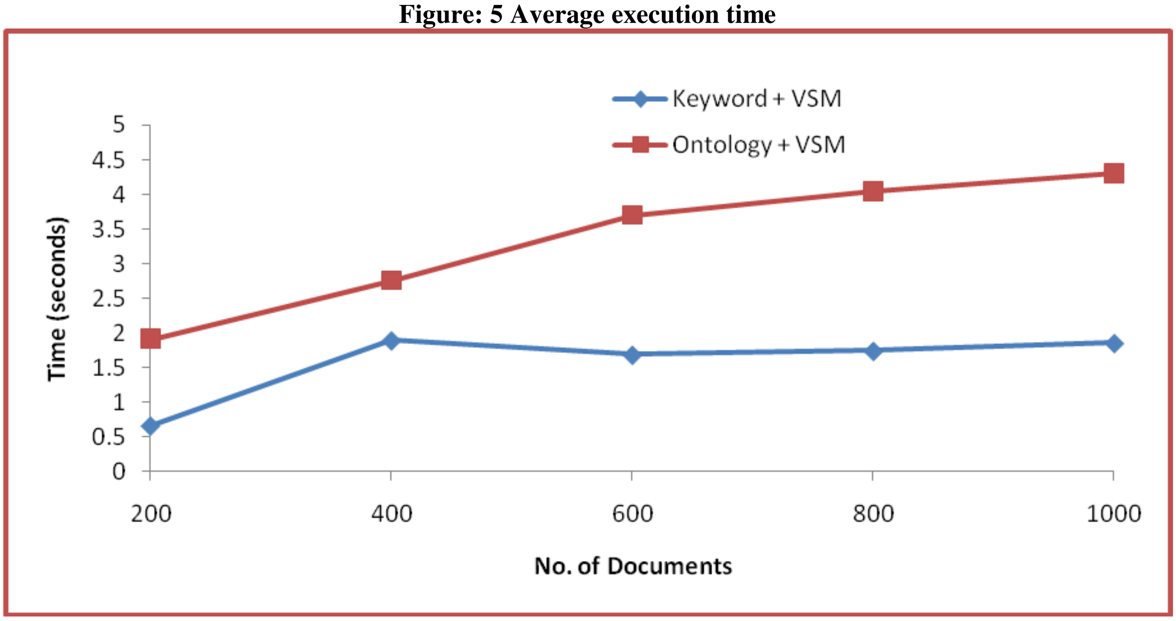
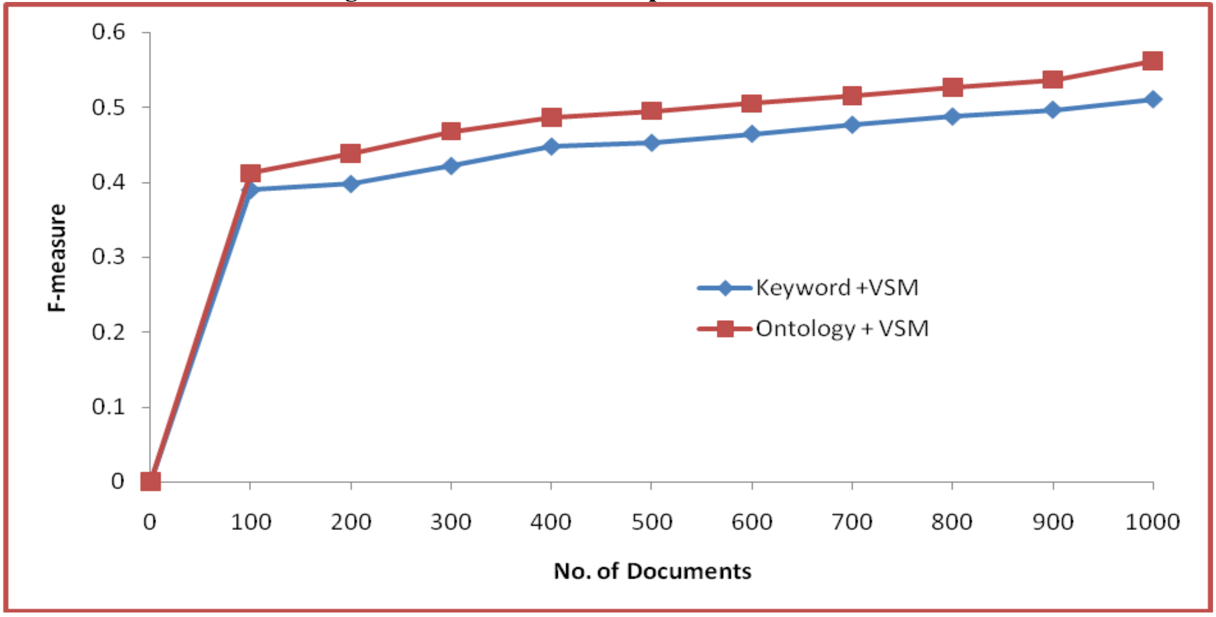


Figure 5 shows that the average execution time depends on the number of documents. Figure 6 shows the comparison of F-measure based on the number of document and similarity measures.

Figure : 6 Performance Comparison of F-Measure



The experimental results show that the reweighting improves the document and illustrates that the ontology-based VSM has the higher F-measure value than the traditional VSM. Higher F-measure shows high accuracy. The traditional VSM cannot identify the semantic relationship between words, which is important in representing the text documents. The improvement of 10% is achieved by VSM+ Ontology in comparison with VSM+ keyword search.

V. CONCLUSIONS

As an extension of the current document, Semantic annotations provide a structured data and knowledge representation framework for information extraction. Semantic provides a structured data and knowledge representation framework for information extraction. The approach can be seen as an evolution of the keyword indices are replaced by ontology knowledge representation and pre-automatic document annotation weighting procedure that improves the retrieval process. The semantic indexing and retrieval framework includes all the aspects of Semantic document, namely, ontology development, information extraction, ontology population, inference, semantic rules, semantic indexing and retrieval. When these technologies are combined with the comfort of keyword-based search interface, high performance and scalable semantic retrieval system is obtained. Evaluation results show that the proposed approach can easily outperform the traditional approach.

REFERENCES

1) Vandana Korde et al Text classification and classifiers:" International Journal of Artificial Intelligence&Applications (IJAIA), Vol.3, No.2, March 2012".
2) Zakaria Elberrichi,Abdelattif Rahmoun, and Mohamed Amine Bentaalah"Using WordNet for Text Categorization"The International Arab Journal of Information Technology, Vol. 5, No. 1, January 2008".
3) Aurangzeb Khan, Baharum Baharudin, Lam Hong Lee, Khairullah khan"A Review of Machine Learning Algorithms for Text-Documents Classification" journal of advances in information echnology, vol. 1, no. 1, February 2010".
4) Fellbaum, C. " Wordnet: an electronic lexical databases", Cambridge MA: MIT Press, 1998.
5) Euzenat, J. and Shvaiko, P. "Ontology matching", Berlin Heidelberg (DE), Springer-Verlag, 2007.
6) Wu, Z. and Palmer, M. "Verb semantics and lexical selection", Proceeding of the 32nd Annual Meeting of the Association for Computational Linguistics, pp. 133-138, 1994.
7) Jiang, X. and Tan, A. "OntoSearch: a full -text search engine for the semantic web", Proceedings of the 21st National Conference on Artificial intelligence, pp. 1325-1330, 2006.
8) S. Bloehdorn , P. Cimiano1, A. Hotho, and S.Staab"An Ontology-based Framework for Text Mining", July 28, 2004 .
9) Khaled Khelif, Rose Dieng-Kuntz, and Pascal Barbry "An Ontology-based Approach to Support Text Mining and In-formation Retrieval in the Biological Domain", Journal of Universal Computer Science, vol. 13, no. 12 (2007), 1881- 1907.
10) Ning Zhong, Yuefeng Li, and Sheng-Tang Wu," Effective Pattern Discovery for Text Mining" IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 1, January 2012.
11) Hmway Hmway Tar , Thi Thi Soe Nyaunt "Ontology-based Concept Weighting for Text Documents" World Academy of Science, Engineering and Technology 57 ,2011.
12) S.C. Punitha, K. Mugunthadevi, and M. Punithavalli ,"Impact of Ontology based Approach on Document Cluster-ing" International Journal of Computer Applications (0975 – 8887) Volume 22– No.2, May 2011.
13) R. Savitha*, Dr. R. Porkodi," An Overview of Text Mining Techniques and Methodologies Used in Bioinformatics Domain" International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-4, Issue-3)march 2015.
14) Y. Megan Kong a, Carl Dahlke b, Qun Xiang a, Yu Qian a, David Karp d, Richard H. Scheuermann ",Toward an ontology-based framework for clinical research databases" J Biomed Inform. 2011 February NIH Public Access
15) Stavrianou et al. "Overview and Semantic Issues of Text mining " SIGMOD Record, September, 2007. (vol.36,No.3)
16) Grobelnik M Many faces of text processing. In: WIMS'11: Proceedings of the International ConferenceonWebIntelligence,Mining and Semantics. ACM. p 5 2011
17) D. C. Wimalasuriya and D. Dou. Ontology-based information extraction: An introduction and a survey of current approaches. Journal of Information Science, 2010, 36(3): 306.
18) H. Cunningham, K. Bontcheva, V. Tablan, and D. Maynard, General Architecture for Text Engineering (GATE) (2003). Available at: http://www.gate.ac.uk (accessed 25 June 2009)
19) Seyed Abbas Mahmoodi1*, Kamal Mirzaie2 and Seyed Mostafa Mahmoudi3 "A new algorithm to extract hidden rules of gastric cancer data based on ontology" SpringerPlus 2016
20) D.C. Wimalasuriya and D. Dou, Using multiple ontologies in information extraction. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, (ACM, New York,2009)