Journal of Engineering and Applied Sciences 13 (11): 4073-4079, 2018
ISSN: 1816-949X

Ontology Based Text Document Clustering for Sports

[1]A. Sudha Ramkumar, [2]B. Poorna and [3]B. Saleena
[1]Bharathiar University, Coimbatore, India
[2]SSS Jain College, Chennai, India
[3]VIT, Chennai, India

Abstract: Text document clustering is used to group a set of documents based on the information it contains and to provide retrieval results when a user browses the internet. Majority of the text document clustering algorithms clusters the documents based on the terms and the frequency of occurrence of those terms and do not consider the meaning among terms because of this clustering performance decreases in terms of precision and recall. To overcome this problem, this research proposes ontology based text document clustering in which documents related to sports have been clustered semantically using sports domain. In this study, sports domain ontology along with WordNet ontology, the lexical database has been used to improve the quality of clustering. With the help of WordNet ontology, the terms and their relevant terms has been retrieved by the synonym retrieval algorithm. This study proposes how these terms along with the relevant terms when it applied to k-means clustering algorithm will improve the performance of the clustering process. Experimental evidence has been shown to prove how the ontology based clustering approach significantly improves the performance of clustering over traditional k-means approach and k-means with dimension reduction technique in terms of precision, recall and accuracy for the bbc dataset.

Key words: Clustering, domain ontology, WordNet, semantic, traditional, dimension

INTRODUCTION

Clustering is a common form of unsupervised learning and has been an active research area for many decades (Jajoo, 2008). It helps a lot in improving the search results by organizing large document sets into relevant documents as clusters. Domain ontology represents concepts, subconcepts and relationship between them of a particular domain. Ontology can be applied to the document clustering to get better retrieval results. Many researches proposed how ontology can improves the clustering results. WordNet is the lexical database for the English language developed at Princeton University and can be interpreted and used as ontology in the computer science. It is an online database which includes nouns, verbs, adjectives and adverbs grouped into sets of synonyms called synsets. Many researchers proposed that WordNet is widely used to compute the semantic similarity measure between the concepts and is not only reduces the dimensionality but also improves the clustering purity. This study proposes that text document clustering based on sports domain ontology along with synonyms retrieved from WordNet to improve the clustering performance in terms of precision, recall and accuracy.

The performance of clustering will decreases dramatically due to the problem of high dimensionality and sparsity of data in the dataset (Liu *et al*., 2003). In order to improve the clustering performance, the sparsity of data should be reduced. Experimental evidences shows thatfeature selection of dimension reduction techniques improves the clustering performance in terms of evaluation metrics (Liu *et al*., 2005). As a first step of this study starts with the term extraction, followed with the sports domain ontology creation, performing term-concept match, implemented the synonym retrieval algorithm to retrieve synonyms of all the selected terms using WordNet ontology and finally, the similarity measure calculated to select only the most relevant synonym of the terms. This study proves experimentally the performance of the clustering in k-means with the ontology based approach.

Literature review: Text document clustering is used to group a set of documents based on the information it contains and to provide retrieval results when a user browses the internet. This study describes some of the ontology based approach towards the information retrieval, data mining and clustering process and some approaches proved experimentally, that the use of

Corresponding Author: A. Sudha Ramkumar, Bharathiar University, Coimbatore, India

ontology improves the performance of clustering or information retrieval when compared to traditional keyword approach.

Dou *et al.* introduced general concept of semantic data mining and investigates why ontology has the potential to help semantic data mining and how formal semantics in ontologies can be incorporated into the data mining process and provided the state of art of ontology based approaches for the use of ontology from data pre-processing to mining results. Huang *et al.* (2008) compared and analyzed the effectiveness of variety of distance functions and similarity, measure in partitional clustering for five different text document dataset. Prabha *et al.* (2014) compared the performance of different clustering techniques for five different dataset and experimentally, proved the clustering performance in terms of precision, recall and recall.

Xinhua and Xutang (2012) proposed the concept of technique preparation domain ontology and its structure design. This study also proposed the similarity measure between the concepts of domain ontology and how it is applied in technique preparation process. The experimental evidences proved the efficiency of the use of domain ontology in information retrieval process. Swe (2011) proposed a model to solve the problems on keyword searching by using ontology based approach and metadata casebase. This model consists of identifying domain concepts in user's query and applying query expansion. This study also, proves experimentally, that domain ontology can be effective for query expansion in terms of precision with the other approaches.

Sonakneware *et al.* described the information retrieval system in which a user gives a natural language query which the system interprets as input query and extracts the semantic information by using ontology based approach.. For conceptual search, user's query is expanded to SPARQL query and this query is fired on the knowledge base which is in the RDF data format thus, retrieving the relevant answer. The information retrieval system finds out semantic query terms for query expansion by using WordNet. Shivakumar and Bhuvaneswarir (2012) provided a unified framework that integrates diverse source of genomic data from various repositories such as ontologies and databases using ontology based approach. The proposed ontology design converts all possible relations for genomic data and experimental result shows that the ontology is designed with well defined classes and conceptual relationships.

Shivakumar and Bhuvaneswari (2012) analyzed the set of keywords can be represented as features for grouping documents and the use of gene ontology terms to group documents semantically. This study focuses on clustering genomic documents based on both syntactic and semantic. Experimentally, proved GO terms based approach grouped large number of documents without considering any other keywords this is semantically relevant which results in reducing the complexity of the attributes considered. Punitha *et al.* (2014) proposed a method called hierarchical agglomerative clustering which manages clusters as tree like structure that makes possible for browsing. In this clustering method, the nodes in the tree can be viewed as parent-child relationships, i.e., topic-suptopic relationship in a hierarchy. Experimental results show that this method is best when compared to k-means, EM and TCFS method.

Banchs (2012) presented a survey of 23 papers and identified 4 major areas under semantic clustering such as latent semantic indexing based, graph based, ontology based and lexical chain based. This study concludes that semantic approach is better than traditional keyword approach in terms of accuracy and quality of clusters. Salih and Qasim (2013) proposed the use of domain ontology in the knowledge extraction process. They proposed algorithm to build semantic relations which captures the semantic similarity among sentences based on WordNet semantic dictionary.

Tar and Nyaunt (2011) investigated how ontology can be applied to clustering process and present the concept weight for text clustering system based on the ontology. It also, gives experimental results of clustering dissertation papers from Google search engine and compared the ontology based k-means with the traditional k-means in terms of performance measures such as precision, recall and F-measure. Zhang *et al.* (2008) evaluated the effects of nine semantic similarity measures with a term re-weighting method on document clustering of PubMed document sets. They proved the use of domain ontology not only yields better clustering results but also reduces the dimension of clustering space and computational complexity.

MATERIALS AND METHODS

The proposed methodology consists of 6 phases, preprocessing, ontology creation, term-concept match, synonyms retrieval and semantic similarity measure calculation, ontology based clustering and cluster quality. The proposed methodology is shown in Fig. 1:

Preprocessing: The BBC sport document collection is taken for the experiment. In order to form mathematical data model that the computer can deal with preprocessing techniques are applied to the document collection which greatly influence the outcome of any clustering process. Given a collection of text documents, the first step is to

parse the documents which yields mathematical model which is easy for analyzing in subsequent steps. The parsing task involves a tokenization followed by stemming and stopword removal. Tokenization transforms the contents of a document into a set of terms. Stemming is used to reduce the number of unique terms by stemming the terms to their roots like "jumping" into "jump". Stop-word removal is used to remove the basic words like "also" and "about" which occurs frequently in documents but do not have any meaning and are considered to be noise. This study uses snowball stemming algorithm.

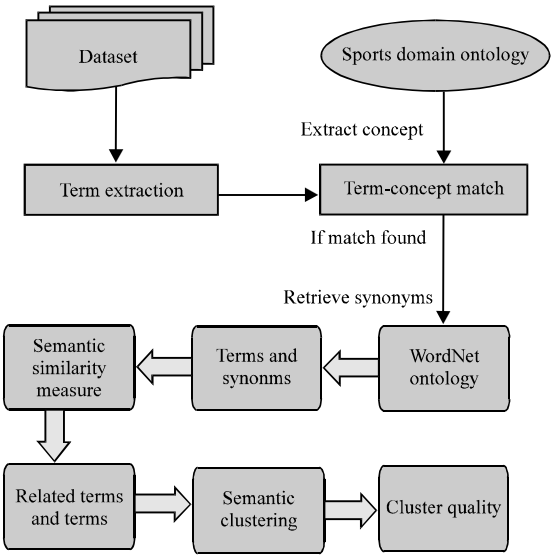

Fig. 1: Methodology of ontology based clustering preprocessing

Ontology creation: In this study, sports domain ontology is created as a lightweight ontology which is a structured representation of knowledge where the concepts are arranged in a hierarchy with a simple relationship between them. It consists of five games such as athletics, cricket, football, rugby and tennis developed using protege tool. The sports domain ontology is created with the help of domain expert knowledge and the concepts and subconcepts of this ontology were called through Jena using eclipse IDE. Figure 2 shows the excerpt of sports ontology in graphical representation.

Term-concept match: After the preprocessing of document collection is done. The document term matrix is constructed with the table of frequencies of occurrences of terms in each document. Local and global weighting functions are applied to estimate the relative importance of a term within the document and within the whole collection. Since, each document contains different words, this table is a high dimensional sparse m×n matrix, m is the number of unique terms in the document collection and n is the number of documents in the collection. The high-dimensional and sparse features bring great noise to the text clustering and make it difficult for clustering algorithms to effectively cluster similar documents. In order to reduce the dimension of this matrix, this study used the InfoGain method. The reduced terms are then extracted and mapped with sports domain ontology concepts.

Synonyms retrieval and similarity measure calculation: If an extracted term is equal to a concept of sports domain

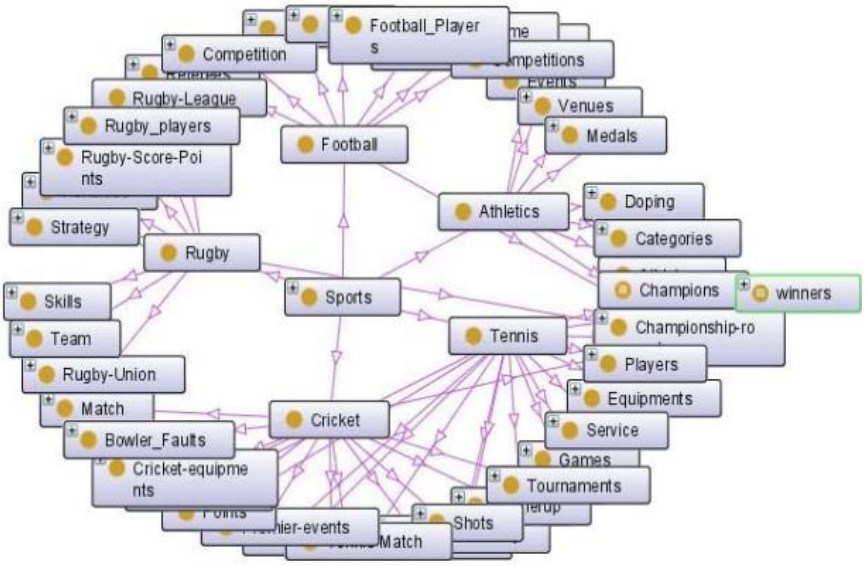

Fig. 2: Sports domain ontology

ontology then it is searched in WordNet ontology for the synonyms using synonyms retrieval algorithm which is shown in Algorithm 1 and is implemented in Java program. Retrieved synonyms consist of a set of words and are called as WordForms. In this study, semantic similarity measure for each and every WordForms has been calculated with the help of Wu and Palmer (1994) measure. Wu and Palmer (1994) proposed a path based measure that takes into account the depth of the concepts in the hierarchy. In this study, the synonyms are verified for their relevance with the help of Wu and Palmer (1994) measure using WS4J (WordNet similarity for Java) is a Java API which measures the semantic similarity between words of synsets. The semantic similarity value lies between 0 and 1. By Wu and Palmer (1994) calculates the similarity value by considering the depths of the two synsets in the WordNet, along with the depth of the least common Subsumer. If the value is 1, then more relevant the term is. Therefore, the similarity measure s greater than the threshold value of 0.8 is taken for the clustering process. By Wu and Palmer (1994) similarity measure $S_{W \text{ and } P}$ is calculated by using the following Eq. 1:

$$S_{W \text{ and } P} = \frac{2 * \text{Depth}(LCS)}{\text{Depth}(S_1) + \text{Depth}(S_2)} \quad (1)$$

Terms T = {t1, t2, ..., tm} where ti are terms extracted from the dataset where i = 1, 2, 3, ..., m terms. Concepts = {c1, c2, ..., cn} where cj-concepts and sub-concepts of domain ontology where j = 1, 2, ..., n concepts.

Count-numbers of matched terms with concepts. WordForms is a set of words W = {w1, w2, ..., wk} where k = number of synonyms of a word. Similarity, measure is used to check for relevancy between WordForms.

Algorithm 1; Synonyms retrieval algorithm:

(To retrieve the synonym for a given word and check for relevancy. Relevant synonym is added with the terms. The terms and relevant terms are applied to k-means clustering) input: extracted terms, ontology concepts and subconcepts output: synonyms

1. Set count = 0, threshold = 0.8
2. Load WordNetDatabase and create object
3. For all terms and concepts
4. Check if term = concept
5. Increment count
6. Get Synsets of term
7. If length of Synsets >0
    - 7.1 Get WordForms
    - 7.2 For each pair of WordForms
    - 7.3 Calculate similarity measure S
    - 7.4 If S>threshold
    - 7.5 Add synonym with terms
    - 7.6 Else ignore
    - 7.7 Repeat step 7.2 until WordForms exists
8. Add terms
9. Repeat step 3 until terms and concepts exists
10. Count is returned
11. Terms and synonyms are returned

The similarity measure is calculated for each and every pair of words of synset and the word whose similarity value greater than the threshold value of 0.8 will be considered to be the most relevant. The synonyms retrieval algorithm is implemented in Java using eclipse IDE. Figure 3 shows the output for the synonym retrieval algorithm for the term "centuries" in Eclipse. The

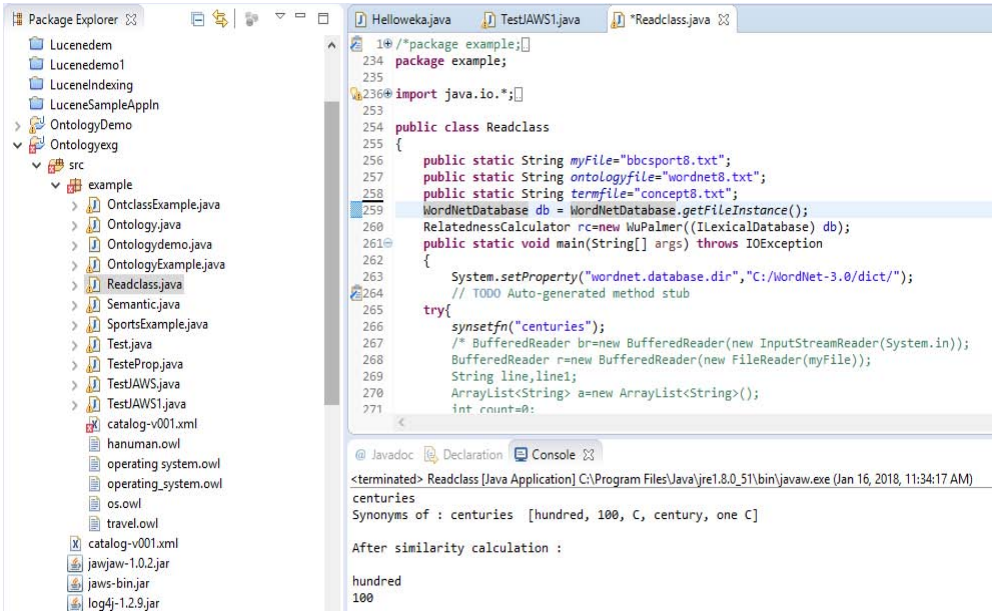

Fig. 3: Output of synonym retrieval algorithm for the term "centuries"

synonyms returned for the term "centuries" are {hundred, 100, century, one century} and according to this algorithm for each pair of synonyms whose similarity measure >0.8 will be considered to be the most relevant for the term. So, the synonyms hundred, 100, century, one century are given as input to the wordsimilarity function. The word similarity function will calculate the semantic similarity for each and every pair of words of synset. The synonyms returned are hundred, 100, century, one century and according to this algorithm among these five words of synset, hundred and 100 are returned as the outpu.

Ontology based clustering: The k-means clustering algorithm is an efficient unsupervised learning algorithm developed by MacQueen to solve the well known clustering problem. The k-means algorithm aims to partition a set of objects based on their features into k clusters where k is a predefined constant. The main idea is to define k-centroids, one for each cluster. The centroid of a cluster is formed in such a way that it is closely related in terms of similarity functions in which similarity is calculated by euclidean distance to all objects in that cluster. Ontology based clustering process involves the use of terms along with the most relevant terms retrieved with the help of synonyms retrieval algorithm. This ontology based clustering has some advantages over traditional k-means and k-means using dimension reduction technique, they are as follows:

- Sports domain ontology is created with the help of domain experts and it is more accurate
- The number of terms is significantly reduced with the help of domain ontology and it yields the document term matrix with ease
- Ontology based clustering uses the similarity measure based on semantics and it helps to incorporate the domain knowledge into the mining process

Cluster quality: The BBC sport dataset is preprocessed with the help of StringToWordVector filter of the WEKA tool. Preprocessing step involves tokenization, stemming and stop-word removal. Then, the k-means clustering is applied on the dataset which takes the euclidean distance to measure the similarity between the documents. It is considered that the closer two texts have high similarity between them. Classes to clusters evaluation method of WEKA tool is used which generates an output in the form of confusion matrix. A confusion matrix is a table that allows the visualization of the performance of an algorithm and in unsupervised learning it is called as matching matrix as shown in Table 1 (Banchs, 2012). In the confusion matrix, all the diagonal elements are true positives and it is the relevant document to that particular class where as the number of documents retrieved are true negatives and true positives. To evaluate the performance of the clustering, precision, recall and accuracy metrics are considered in this study:

Table 1: Confusion matrix

	Predicted class	
Actual class	Yes	No
Yes	TP	FN
No	FP	TN

$$\text{Precision} = \frac{\text{Number of relevants documents ritrieved (TP)}}{\text{Number of documents ritrieved (TP+FP)}}$$

$$\text{Recall} = \frac{\text{Number of documents ritrieved (TP)}}{\text{Number of relevant documents (TP+FN)}}$$

RESULTS AND DISCUSSION

To prove the significance of ontology based clustering approach, we compared the ontology based approach with k-means approach and k-means using dimension reduction technique. In our experiment, the BBC sport dataset has been used for all the approaches and is downloaded from the BBC website. The confusion matrix generated from the classes to clusters evaluation method is used for the comparison. The BBC sport dataset consists of 737 text documents and 5 natural classes: athletics, cricket, football, rugby and tennis. In k-means approach, the total numbers of incorrectly clustered documents are 123 out of 737 documents. The overall recall is 83% and overall precision is 85%. The accuracy is 83%. The confusion matrix for the k-means approach is shown in Table 2.

In k-means with dimension reduction approach, the information gain feature selection method is used to reduce the dimension. For this approach, the total numbers of incorrectly clustered documents are 69 out of 737 documents. The overall recall is 89.2% and overall precision is 93.4%. The accuracy is 90.7%. The confusion matrix for the k-means with dimension reduction method is shown in Table 3.

In ontology based approach, the terms mapped with concepts of sports domain ontology along with the synonyms is applied to k-means clustering algorithm. In this approach, the total numbers of incorrectly clustered documents are 43 out of 737 documents which yield 5.8% incorrect. The confusion matrix for the ontology based k-means clustering is shown in Table 4.

For this reduced data, overall recall is 93.7% and overall precision is 96.6%. The accuracy is 94.2%.

Table 2: Confusion matrix for k-means approach

Games	Athletics	Cricket	Football	Rugby	Tennis	Predicted
Athletics	100	1	6	2	24	133
Cricket	0	112	34	14	0	160
Football	0	7	221	14	11	253
Rugby	1	4	4	117	1	127
Tennis	0	0	0	0	64	64
Original classes	101	124	265	147	100	737

Table 3: Confusion matrix for k-means with dimension reduction

Games	Athletics	Cricket	Football	Rugby	Tennis	Predicted
Athletics	101	0	0	0	0	101
Cricket	0	67	0	0	0	67
Football	0	2	260	4	2	268
Rugby	0	55	5	143	1	204
Tennis	0	0	0	0	97	97
Actual	101	124	265	147	100	737

Table 4: Confusion matrix for ontology based k-means clustering

Games	Athletics	Cricket	Football	Rugby	Tennis	Predicted
Athletics	97	0	2	0	4	103
Cricket	0	119	0	0	1	120
Football	1	2	249	2	10	264
Rugby	3	2	13	145	1	164
Tennis	0	1	1	0	84	86
Actual	101	124	265	147	100	737

Table 5: Comparison of ontology based with k-means and k-means using dimension reduction

Metrics	k-means clustering reduction (%)	k-means with dimension (%)	Ontology based clustering (%)
Precision	83.0	93.4	96.6
Recall	85.0	89.2	93.7
Accuracy	83.4	90.7	94.2

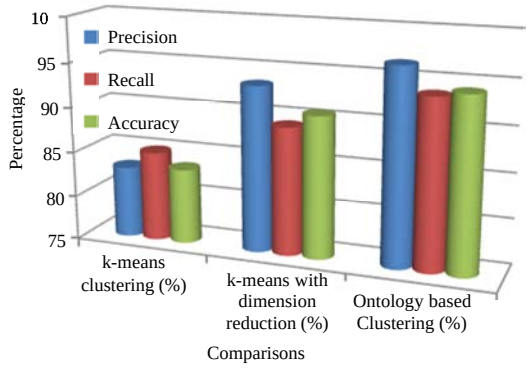

Fig. 4: Comparison of k-means with ontology based approach

Comparison of ontology based approach with k-means approach and k-means using dimension reduction in terms of precision, recall and accuracy is shown in Table 5.

Figure 4 significant improved performance in terms of recall, precision and accuracy for the ontology based approach with k-means algorithm and k-means using dimension reduction technique.

CONCLUSION

This study presents methodology of ontology based clustering for sports with the phases of preprocessing, domain ontology creation, term-concept match, synonyms retrieval and ontology based clustering advantages over k-means approach and k-means using dimension reduction technique. It proposes ontology based clustering approach for sports domain which is based on the term-concept mapping and synonyms retrieved from WordNet. The experimental evidences show the significance of ontology based approach over other two approaches. This system also, provides increased recall, precision and accuracy and faster retrieval of documents when compared to other 2 approaches.

RECOMMENDATIONS

In future, we can fine tune the performance of ontology based clustering by updating the concepts of sports domain ontology and this methodology can be applied to other domain ontologies and dataset also.

REFERENCES

Banchs, R.E., 2012. Text Mining with MATLAB. Springer, Berlin, Germany, ISBN:978-1-4614-4150-2, Pages: 353.

Huang, A., 2008. Similarity measures for text document clustering. Proceedings of the Sixth New Zealand Conference on Computer Science Research Student (NZCSRSC2008), April 14-18, 2008, University of Canterbury, Christchurch, New Zealand, pp: 49-56.

Jajoo, P., 2008. Document Clustering. Indian Institute of Technology Kharagpur, Kharagpur, India,.

Liu, L., J. Kang, J. Yu and Z. Wang, 2005. A comparative study on unsupervised feature selection methods for text clustering. Proceedings of the 2005 International Conference on Natural Language Processing and Knowledge Engineering, October 30-November 1, 2005, IEEE, Beijing, China, ISBN: 0-7803-9361-9, pp: 597-601.

Liu, T., S. Liu, Z. Chen and W.Y. Ma, 2003. An evaluation on feature selection for text clustering. Proceedings of the 20th International Conference on Machine Learning, (ICML'03), Washington, DC., pp: 415-424.

Prabha, M.S., K. Duraiswamy and M.M. Sharmila, 2014. Analysis of different clustering techniques in data and text mining. Intl. J. Comput. Sci. Eng., 3: 107-116.

Punitha, S.C., P.R.J. Thangaiah and M. Punithavalli, 2014. Performance analysis of clustering using partitioning and hierarchical clustering techniques. Intl. J. Database Theory Appl., 7: 233-240.

Salih, A. and M. Qasim, 2013. Towards exploring semantic similarity based on WordNet semantic dictionary. Intl. J. Comput. Appl., 66: 24-28.

Shivakumar, B.L. and V. Bhuvaneswari, 2012. Semantic clustering of genomic documents using go terms as feature set. Global J. Comput. Sci. Technol., Vol. 17, 10.17406/gjcst

Shivakumar, B.L. and V. Bhuvaneswarir, 2012. Ontology design for gene integration and feature representation-OGFR. IOSR. J. Eng., 2: 1188-1195.

Swe, T.M.M., 2011. Intelligent information retrieval within digital library using domain ontology. Comput. Sci. Inf. Technol., 1: 363-373.

Tar, H.H. and T.T.S. Nyaunt, 2011. Enhancing traditional text documents clustering based on ontology. Intl. J. Comput. Appl., 33: 38-42.

Wu, Z. and M. Palmer, 1994. Verbs semantics and lexical selection. Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics, June 27-30, 1994, Las Cruces, New Mexico, USA., pp: 133-138.

Xinhua, L. and Z. Xutang, 2012. A domain ontology-based information retrieval approach for technique preparation. Phys. Procedia, 25: 1582-1588.

Zhang, X., L. Jing, X. Hu, M. Ng and J. Xia *et al*., 2008. Medical Document Clustering using Ontology-Based Term Similarity Measures. In: Strategic Advancements in Utilizing Data Mining and Warehousing Technologies, David, T. (Ed.). IGI Global, Hershey, Pennsylvania, ISBN:978-1-60566-717-1, pp: 2233-2243.