International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 04 | Apr 2019 www.irjet.net p-ISSN: 2395-0072

SEMANTICS BASED DOCUMENT CLUSTERING

Shraddha Shinde[1], Pratiksha Gurav[2], Vaishnavi Tandel[3], Prof. Amol P. Pande[4]

[1,2,3]*Student, Department of Computer Engineering, DMCE, Maharashtra, India.*
[4]*Head of Department, Department of Computer Engineering, DMCE, Maharashtra, India.*

Abstract - *Document clustering is a technique used to organize large datasets of documents into meaningful groups. The associated documents are described by the relevant words which serve as cluster labels. The traditional approach for document clustering uses bag-of-words representation. This representation often ignores the semantic relations between the words. Therefore ontology-based document clustering is proposed. One of the ways to deal with reusability and remix of learning objects in context of e-learning is via the use of appropriate ontologies. The more appropriate use of ontology the better will be the annotation of learning material. To couple document clustering with ontology will help in producing better clusters which will not ignore the semantic relation between the words. The proposed system uses "an ontology-based document clustering" approach based on two-step clustering algorithm. Since it is two step clustering, it uses both partitioning as well as hierarchical clustering algorithms. Ontology is introduced through defining a weighting scheme. This weighing scheme integrates traditional scheme of co-occurrences of words paired with weights of relations between words in ontology. The algorithm used from partition clustering technique is K-means whereas from hierarchical clustering technique is hierarchical agglomerative algorithm. Thus we can say that the clustering approach that uses the semantics of the documents for term weighting produces better results than the approach without semantics.*

Key Words*: *Document Clustering, Ontology-based Clustering, eLearning, Ontology Generation, Semantic Relation, eLearning Concept.

1. INTRODUCTION

In recent years there has been explosive growth in the volume of data. As there is increase in volume of data it is very difficult to retrieve useful information from such large volume. . There is explore such a need to automatically large collection of data. Information Retrieval is the process of locating material (or documents) of an unstructured nature (generally text) from large collections (usually stored on computers).For this purpose unsupervised clustering algorithm is the best option. These algorithms are fast and scalable. They require no prior understanding of data. They do not need any costly graph building or association rule preprocessing. Clustering means dividing collection of objects into number of clusters. The main aim behind clustering is to find structure in data object and then reflecting this structure as group. The objects within the group will possess large degree of similarity. This similarity should be minimum outside the cluster groups. [9]

The E-learning domain ontology will be reused or combined/merged with their own ontologies in following systems:-

- Educational Systems.
- Content management systems.
- Recommender systems.
  The clustering results produced will be valuable to all of the above systems. The cost of content generation and classification is high. Using the proposed system in learning systems will be able to serve more appropriate results to users in a semantic way. Also there has been increase a large number of documents. Construction of e-learning domain based ontology is done in following two phases:

  • Ontology generation- the retrieved text documents will be preprocessed first. Then their semantic importance of nodes and their corresponding relation will be represented.
  • Clustering- Concept weighting will be performed. Then clustering will be performed and results will be presented to the user.

1.1 Clustering

To cluster documents Two-step clustering is used. The algorithm is based on a two-stage approach.

- **First stage :**

In the first stage, K-means is applied on the input data. One of the best known partitioning algorithms is K-means. K-means is a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. They have low computational requirements. Their time complexity is linear i.e O(n). K-means algorithm is also widely used for document clustering. K-means algorithm was first proposed by J.B. MacQueen. This algorithm works in these 5 steps:

1.Specify the desired number of clusters K.
2.Randomly assign each data point to a cluster .
3.Compute cluster centroids .

4.Re-assign each point to the closest cluster centroid.
5.Re-compute cluster centroids.
6.Repeat steps 4 and 5 until no improvements are possible
Similarly, we'll repeat the 4th and 5th steps until we'll reach global optima.

1.2 Second stage

In the second stage, a hierarchical agglomerative clustering procedure is performed on clusters obtained from first stage to form homogeneous clusters. This method is not good at handling huge data sets because of the computational complexity i.e.O(n^2). Then two nearest clusters are merged into the same cluster. Agglomerative clustering works in a "bottom-up" manner. Steps to agglomerative hierarchical clustering:
1.Preparing the data
2.Computing (dis)similarity information between every pair of objects in the data set.
3.Using linkage function to group objects into hierarchical cluster tree, based on the distance information generated at step1.

1.2 . The Vector Space Model

Information Models are used to define a way to represent the document text and the query. Vector space model is an model for representing text document as vectors of identifiers, such as for example, index terms. It is used in information filtering, information retrieval. [6]

- TF-IDF Model

Term Frequency–Inverse Document Frequency is a numerical statistic that reflects how important a word is to a document in a corpus. It is used as a weighting factor in information retrieval.

Term Frequency (t) = Number of times term t appears in a document ...(1)

Inverse Document Frequency measures how important a term is.

Inverse Document Frequency (t) = log N / Nt ....(2)

Where N is the total number of documents and Nt is the number of documents with term t in it.

2 . System Overview

A. System Architecture:

The proposed system is based on semantics whose architecture is depicted in Fig.1.

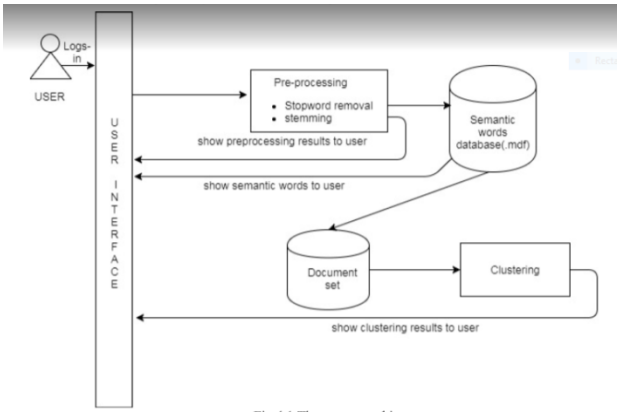

Fig. 1 System Architecture

The proposed system will have following main three modules. They are:

1. User Interface module: This module is responsible for accepting keywords from the user and retrieving the most meaningful and appropriate documents.

2. Concept weighting module: The concept weighting is done before the actual clustering is done. It defines a weighing scheme based on ontology.

3. Clustering module: This module is responsible for clustering the documents.

1. **User Interface module:**

- **User Log in:** In order to access the system the user will have to log into the system with correct values of username and password. If the user is new he/she will be asked to register first and then can log in using the credentials.

- **User interface:** If the user enters correct log in credentials he/she will be presented with the user interface. This user interface will accept query from the user and will be responsible for giving results back to the user.
- **Query handling:** This block will be responsible for query processing. The keywords entered by the user in the search will be used to retrieve documents.

2. Pre-processing

The preprocessing step will be performed two times- firstly while building the domain ontology and second time for the sake of preprocessing the document set so as to represent the document in vector form. The preprocessing step will consist of stopword removal, stemming and case folding. Porter's algorithm will be used for stemming. For case folding all the words will be converted to lower case.

- **Document set:**The document set will consist of E-learning documents.

- **Domain ontology**:The proposed system requires domain specific ontology. The domain is e-learning. The retrieved text documents will be preprocessed first. Then their semantic importance of nodes and their corresponding relation will be represented. If two nodes are semantically related then there will be an edge between these two nodes. The weights between these two edges will be determined using the formula-

$$M_{ij} = \frac{f(x_i,\ x_j)}{f(x_i) * f(x_j)}$$

The weights between the nodes will be pre-computed and stored in excel sheet.

- **Concept weighting:** The concept weighting will be performed using the formula

$$W_i' = W_i + \sum_j [-log_{10}(E_{ij}) * W_j]$$

Where W_i' is the weight of word i after reweighting by ontology.
$W_i$is the value of TF-IDF for word i.
$E_{ij}$is the weight of the edge from i to j in the ontology which will be obtained from the pre-computed excel sheet.

- **Clustering :**The clustering of the documents using two stage clustering approach. The algorithms used are k-means and hierarchical agglomerative clustering algorithm.

3. The Flow Diagram

The following is the flow of the system represented in diagram:

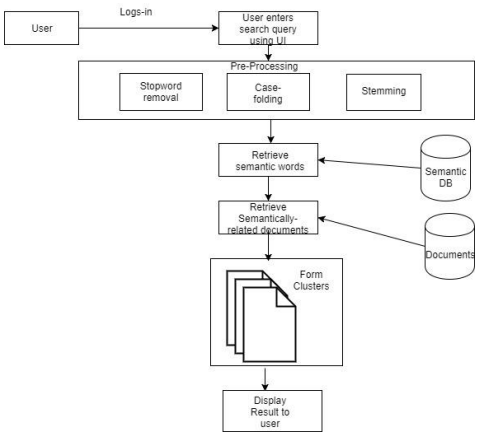

Fig 4.2 Flow Diagram

The following is the explanation of the flow diagram:

i User enters log in credentials. If the credentials are correct then user will be presented user interface
ii Using UI the user enters the search query
iii The search query will be preprocessed to get the important keywords
iv Using these keywords the relevant semantic words will be found out using the semantic database.
v Using these semantic words the relevant documents will be retrieved.
vi Upon getting the documents clusters will be formed. The semantic words will serve as cluster labels.
vii The output in the form of list of documents as well as document clusters will be presented to the user.

4. Implementation and Results:

Basic UI has been developed for retrieving documents. The documents are retrieved based on the keyword entered in the search field.

The user is first presented with the log in page:

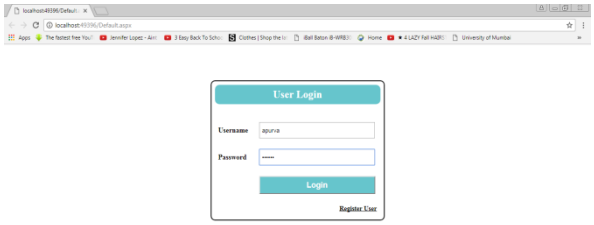

Fig. 3 Log in form

If the user is not registered user then he/she can register using the registration form:

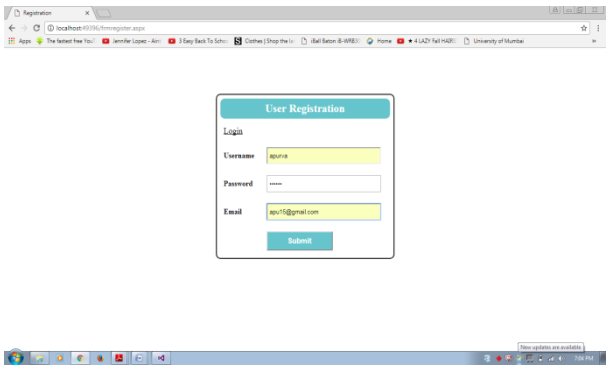

Fig. 4 Registration Form

User will be presented the UI where the user can type keyword for fetching the documents. Based on keywords typed the relevant documents will be listed in the listbox

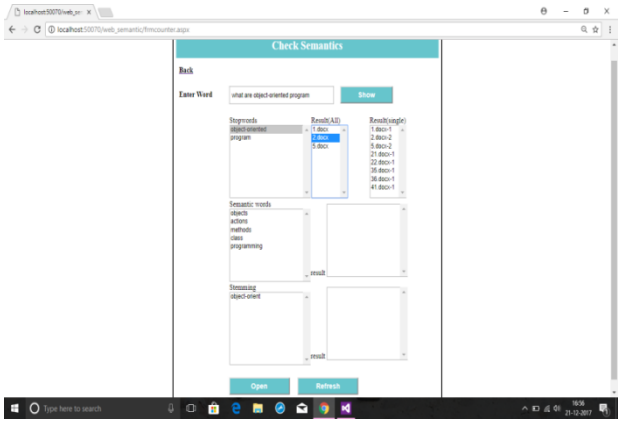

Fig. 5 UI of Document Retrieval

User can also get semantic words related to the output of stop-word selected. In the below fig 6

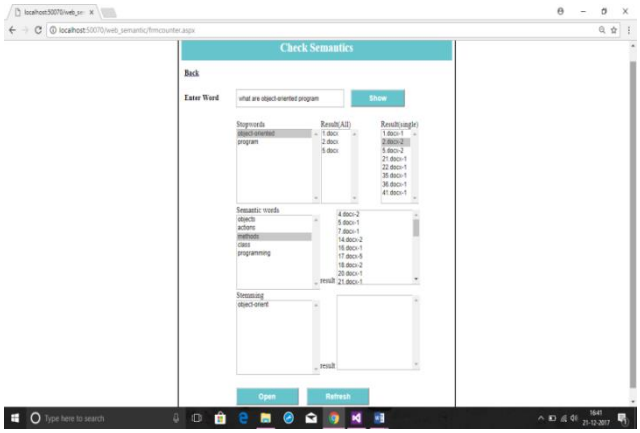

Fig 6 Semantic words output

Upon selecting the semantic word 'method' and clicking on open button it will give list of documents having the word 'method' which is shown in fig 7. User selected doc-5 which got opened as follows:

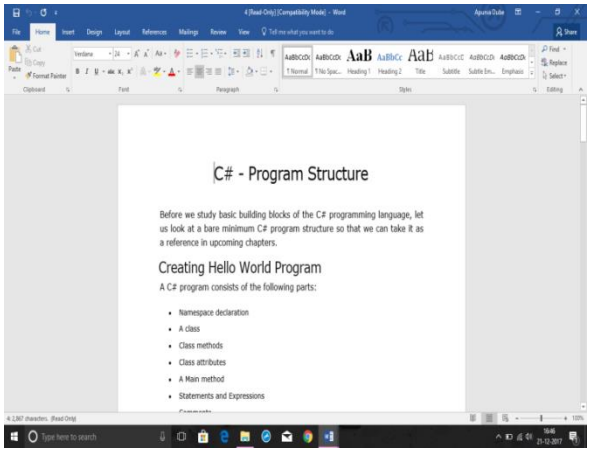

Fig 8 Semantically related document

Also clusters will be formed according to the semantic words appearing in the documents:

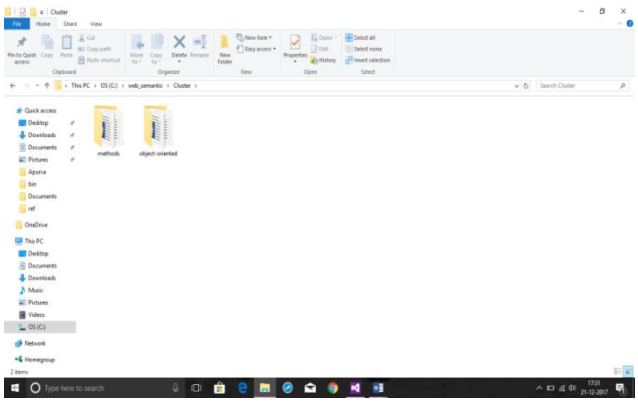

Fig 8 Document Clusters

Upon selecting the folder we can see the contents as follows:

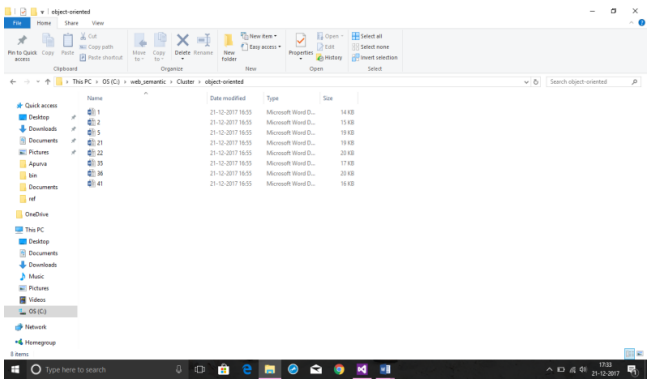

Fig 9 Contents of folder

5. Conclusion

The system introduced a semantic-based approach for documents clustering. document clusters formed using traditional clustering methods may or may not be conceptually similar to one another as semantic relationships between documents are ignored. In our system, a model for document clustering that groups documents with similar concepts together is introduced. The system will initially identify all the semantically related words against the search words entered by the user. Then all the documents having the user search words and semantic words will be retrieved and displayed to user. Also document clusters will be formed. We believe that the system will be helpful to learning and content management systems. Also, using the system will be able to serve more appropriate results to users. Also there has been tremendous increase in the number of documents. There needs to be some way to organize information in such a way that it is easy to retrieve and locate the desired documents. This system would not only do so but also serve more appropriate results in a semantic way.

References

[1] Sara Alaee and Fattaneh Taghiyareh, "A semantic ontology based document organizer to cluster E-Learning documents", 2016 Second international conference on web research(ICWR), 2016 IEEE.

[2] Nadana Ravishankar. T and Shriram. R, "Ontology based clustering algorithm for information retrieval", 4th ICCNT, July 2013, IEEE.

[3] Hongwei Yang, "A document clustering algorithm for web search engine retrieval system",2010 International conference on e-education, e-business, e-management and e-learning,2010 IEEE.

[4] XiQuan Yang, DiNa Guo, XueYa Cao and JianYuan Zhou, "Research on Ontology-based Text Clustering", 2008 Third International Workshop on Semantic Media Adaptation and Personalization, 2008 IEEE.

[5] Enrico G. Caldarola and Antonio M. Rinaldi, "An Approach to Ontology Integration for Ontology Reuse", IEEE 17th International Conference on Information Reuse and Integration, 2016.

[6] Apra Mishra and Santosh Vishwakarma, "Analysis of TF-IDF Model and its Variant for Document Retrieval", International Conference on Computational Intelligence and Communication Networks, 2015 IEEE.

[7] Sanket S.Pawar,Abhijeet Manepatil,Aniket Kadam and Prajakta Jagtap, "Keyword Search in Information Retrieval and Relational Database System: Two Class View, International Conference on Electrical", Electronics, and Optimization Techniques (ICEEOT) , 2016 IEEE.