International Journal of Scientific Research in Computer Science and Engineering **Research Paper**
Vol.5, Issue.4, pp.25-30, Aug (2017) **E-ISSN:** 2320-7639

Semantics Based Document Clustering

Apurva Dube[1*], Pradnya Gotmare [2]

[1*]Dept.of Computer Engineering, K.J.Somaiya College of Engineering, Mumbai, India
[2]Dept.of Computer Engineering, K.J.Somaiya College of Engineering, Mumbai, India

Corresponding Author: apurva.dube@somaiya.edu

Available online at: www.isroset.org

Received 06th Jun 2017, Revised 16th Jul 2017, Accepted 10th Aug 2017, Online 30th Aug 2017

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.

Keywords- *Document Clustering, Ontology-based Clustering, eLearning, Ontology Generation, Semantic Relation, eLearning Concept.*

I. INTRODUCTION

In recent years there has been explosive growth in the volume of data. There is a need to automatically explore such large collection of data. 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]

Normal text document clustering approach neglects subjectivity and explainability. Firstly document clustering is seen as objective method. It is expected to deliver one clearly defined result. Secondly document clustering is a form of machine learning taking place in high dimensional space of words. Thirdly document clustering is not that useful unless it is combined with the explanation of why particular documents are organized into a particular cluster. [2] Hence ontology based clustering algorithm is used so that documents are organized in semantic way. Semantic importance of nodes and their corresponding relations in ontology are represented through scores and weights we assign to them. [1]

The domain of E-learning requires specialized texts to be clustered in a meaningful way. The clustering results would be helpful to many systems such as education systems, content management systems, recommender systems. Their inputs and outputs are related and also they have common entities in them. The challenges in implementing this system is that it requires to develop and utilize an E-learning domain ontology. Domain specific ontologies are ontology for particular domain of interest. Another challenge is the terms may be expressed differently in each document it will be difficult to align it. For example "e-Learning" might be written as "eLearning" in some documents while in some others it will be written as "electronic learning".

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. The reuse of this system would definitely manage this high cost. Also, using the proposed system in learning systems 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. The proposed system would not only do so but also serve more appropriate results in a semantic way.

The objective of the system is to use e-learning domain ontology along with two-stage clustering for first retrieving the most meaning and appropriate documents and then clustering them. The user will enter his/her query to retrieve documents. The corresponding documents will be retrieved based on the keywords entered by the user.

Construction of e-learning domain based ontology is done in following two phases

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

II. 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. Partitioning algorithms are well suited for clustering large document datasets. They have low computational requirements. Their time complexity is linear. K-means algorithm is also widely used for document clustering. In K-means algorithm 'k' is positive integer which represents the number of clusters. K-means algorithm was first proposed by J.B. MacQueen. It is one of the classical clustering algorithms. The main idea behind the classical K-means algorithm is the objects will be placed into clusters according to their distance with the cluster center. The cluster center will be recalculated after every classification.

The following is the algorithm:

(1) Choosing k objects from n data objects as the initial cluster centers.
(2) According to the center of each cluster object, the distance between each object and the center object is calculated; the object can be reclassified according to minimum distance.
(3) Recalculate the center point of each cluster.
(4) Calculation of the standard measurement function, if certain conditions are met, the algorithm will be terminated; If the condition is not met then back to the step (2). [10]
The steps (2) and (3) when there is no further variation in value of center point.

- **Second stage :**

In the second stage, a hierarchical agglomerative clustering procedure is performed on clusters obtained from first stage to form homogeneous clusters. Agglomerative hierarchical clustering is one of popular clustering techniques. This method is not good at handling huge data sets because of the computational complexity. To solve this problem, we proposed a two-stage clustering in which the first stage uses K-means. In the second stage, agglomerative hierarchical clustering handles only centers obtained from the first stage.

The following is the algorithm:
AHC1: For initial clusters derived from the first stage, calculate (G,G′) for all G,$G' \in \mathcal{G}$.

AHC2: Merge the pair of clusters of minimum dissimilarity:

(Gq,Gr) = $min_{\mathrm{G,G'}}\, d(\mathrm{G}, \mathrm{G}')$

Add $\hat{\mathrm{G}}$= $Gq \cup \mathrm{G}r$ to G and remove Gq, Gr from G.

$\mathcal{C}$= $\mathcal{C}$ − 1. If $\mathcal{C}$= 1, stop.

AHC3: Calculate d($\overline{\mathrm{G}}$, G') for all other $G' \in$ G.

Go to **AHC2**. [11].

III. RELATED WORKS

Nadana Ravishankar T. and Shiraram R. have developed a system for ontology based information retrieval and clustering. This work performs preprocessing of documents for extracting meaningful keywords. From the domain ontology semantically related words are obtained. Together the keywords as well as the semantically related words are used to build the decision tree. K-means algorithm is used for clustering the documents. Precision and recall values are calculated with non-ontology based clustering and ontology based clustering. [2]

Dorian Kokoshi and Betim Çiço have stated the importance of integration semantic web and e-learning. There is not only the need of a suitable content of the learning material, but also a powerful mechanism for organizing such material. eLearning is replacing old-fashioned time/place/content

predetermined learning with a just-in-time/at work-place/customized/on-demand process of learning. The Semantic Web technology can be used for realizing eLearning requirements. Semantic browsing first locates metadata, and then assembles point-and-click interfaces from a combination of relevant information. Semantic search enhances current search engines with semantics; it goes beyond normal keyword matching by adding semantic information. It thus allows easy removal of non-relevant information from the result set. [8]

Elizabeth D. Liddy has described the automatic document retrieval system. The paper gives the walkthrough of the automatic document retrieval system. It also gives the detail explanation of various theoretical models of document retrieval out of which our proposed system uses the Vector space model.[13]

IV. THE VECTOR SPACE MODEL

Information Models are used to define a way to represent the document text and the query. It generates weighted term vectors for each document in the collection, and for the user query. The required document is retrieved based on the similarity between the query vector and document vectors. The output documents are ranked according to this similarity.[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.

V. SYSTEM OVERVIEW

A. System Architecture:

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

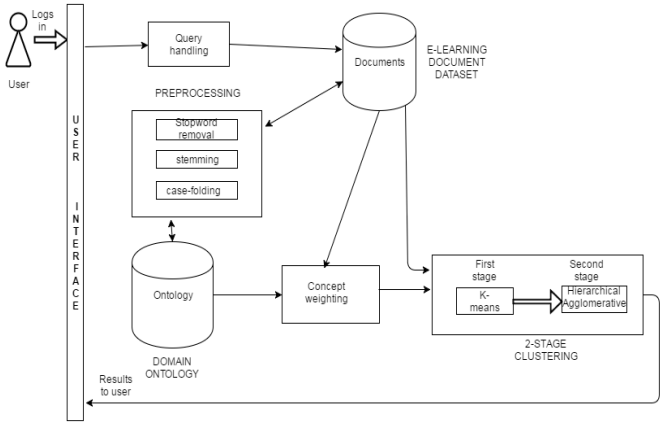

Fig. 1 System Architecture

The architecture diagram is depicted in the figure 1. The system architecture will consist of following blocks:-

- 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.

- 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.

B. Methodology

The methodology for ontology generation and clustering is as follows:

Ontology generation

1) Preprocessing

In this step, all textual data from the documents will be extracted. The main objective is to obtain key terms that will describe each document. The general procedures include: case folding, removing stop-words, stemming the keywords, and merging synonyms and complex words. In case folding, all characters of the words will be converted to either lower or upper case. Here, it will be converted them to lower case. A document may contain many words which are irrelevant to the main subject, namely verbs, identifiers and propositions. These are called stop-words. List of stop-words is available on various websites. Stemming is the process for reducing words to their root forms. We will be using Porter's Stemmer Algorithm.

2) Creating graph

Ontology is a graph of words. Each concept is represented as a node and the corresponding relations with other words would be the edges. Fig. 2 shows a representation of such graph. All words obtained from the previous step are potentially regarded as key words. They represent the core concepts of the documents. However, only few of them will be placed i.e. title words, as nodes in our ontology. This is because title words are semantically more related to the domain.

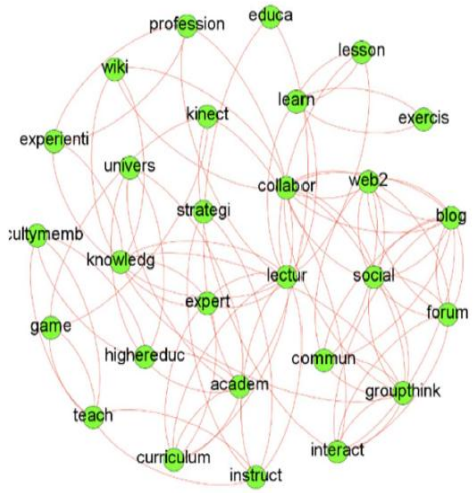

Fig. 2 Domain Ontology

Ontology based clustering

1) Concept weighting

The concept of "concept weighting" is applied before actual clustering algorithms are used. The words obtained from the preprocessing step are the key words as they represent the core concepts of the document. These words should be assigned weights. Thus each document should be converted to a vector of key word weights. In the proposed system, we will use TF-IDF along with some information from the domain ontology (the e-Learning domain ontology) to define the weighting scheme.

The weighting scheme will be defined using the following 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 (i.e. the weight of word i before reweighting)

E_{ij} is the weight of the edge from i to j in the ontology. [1]

2) Clustering based on concept weights

To cluster documents the proposed system will use Two-step clustering. The algorithm is based on a two-stage approach. In the first stage, K-means is applied on the input data. The inputs are vectors of concept weights calculated using Eq. 1 in step 2. In the second stage, a hierarchical agglomerative clustering is performed on pre-clusters to form homogeneous clusters.

VI. IMPLEMETATION AND RESULTS

Basic UI has been developed for retrieving documents. The documents are retrieved based on the keyword entered in the search field. The documents containing the keyword will be retrieved and displayed in the listbox where user can select and open the required document.

The user is first presented with the log in page:

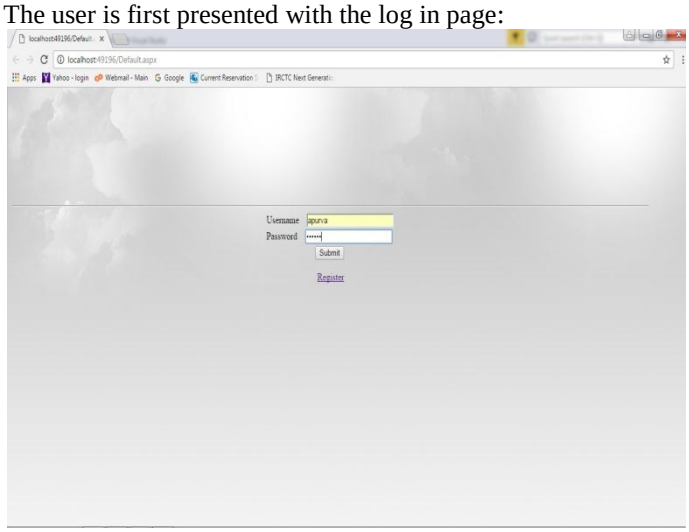

Fig. 3 Log in form

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

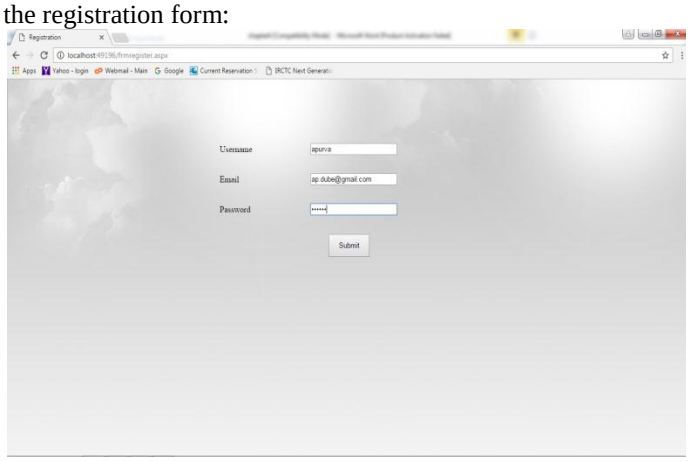

Fig. 4 Registration Form

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:

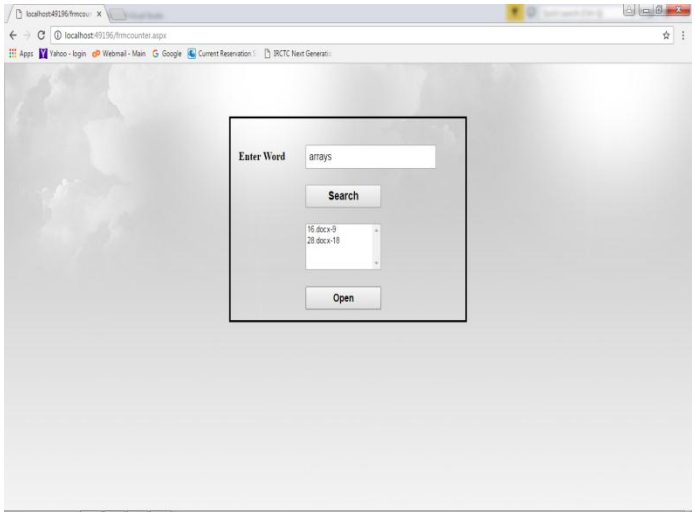

Fig. 5 UI of Document Retrieval

By selecting the document and clicking on open button the document will be opened:

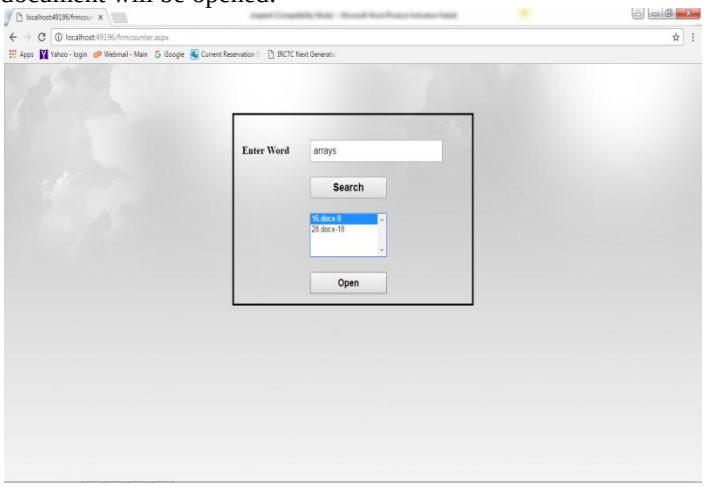

Fig. 6 Clicking on open button

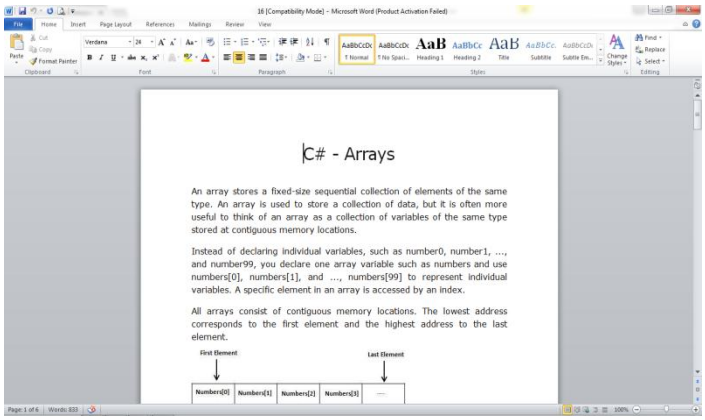

Fig. 7 Selected Document opened

VII. CONCLUSION

This proposed system introduced an ontology-based approach for e-Learning documents clustering. It uses term re-weighting and a clustering algorithm, namely two-step algorithm document set. First the ontology will have to be generated based on document set. Then in the clustering procedure, after stop words removal, stemming, words unification, concept weights will have to be calculated. The term weights vectors for these documents were then clustered using the clustering algorithm. The experimental results (as in base paper) based on ontology will enhance precision, recall and f-measure values on dataset. We believe that the proposed system will be helpful to learning and content management systems. Also, using the proposed 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. The proposed 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.

[8] Dorian Kokoshi and Betim Çiço, "*Integration of Semantic WEB in an eLearning Environment*", Fourth Balkan Conference in Informatics, 2009 IEEE.

[9] Jaskaranjit Kaur and Harpreet Singh "*Performance Evaluation of a Novel Hybrid Clustering Algorithm using Birch and K-Means*", IEEE, 2015.

[10] Li Jun Tao, Liu Yin Hong and Hao Yan "*The Improvement and Application of a K-Means Clustering Algorithm*", International Conference on Cloud Computing and Big Data Analysis, IEEE, 2016.

[11] 11.Yusuke TAMURA and Sadaaki MIYAMOTO, "*A Method of Two Stage Clustering Using Agglomerative Hierarchical Algorithms with One-Pass k-Means++ or k-Median++*", IEEE International Conference on Granular Computing (GrC), 2014 IEEE.

[12] I. Bedini and B. Nguyen, "*Automatic Ontology Generation: State of the Art*," PRiSM Laboratory Technical Report, University of Versailles, Versailles, 2007.

[13] Elizabeth D. Liddy, "*Document Retrieval Automatic*", Syracuse university, 2005.

