ISSN 2229-5518

# Document Clustering by using Semantics

Mr. Sumit D. Mahalle, Dr. Ketan Shah

**Abstract**— Previous methods of clustering mainly uses matching key words of text, However it does not capture the meaning behind the words which is bad side of traditional method to mine the text. The paper is based on Semantic based approach for document clustering which is mainly based on semantic notations of text in documents. In the Semantic document clustering we can parse the web documents into two way, first is syntactically and second is semantically. Syntactical parsing can ignore the less important data from documents so that we can have proper data to pass into next step. Then in next step i.e. Semantic parsing can apply on the parsed syntactic data which give can cluster the documents properly and give the needed response to user at the time of data mining which is not accurately in traditional methods.Basically we are taking n number of IEEE papers from IEEE.org website as a dataset of web documents. Then we applied Semantic Clustering Algorithm, In first step of Syntactic parser gives the proper data in the text file format by removing an useless data from web documents of each IEEE paper. Then in next step these text files will be pass into semantic clustering, here we will get the membership value of each text file. So finally we will get clusters of text files which can be calculated by comparing its membership values with each other. "Document Clustering by using semantics" is a technique which is directly work on textual part of web documents in our database, there are very few technique present which are based on textual data clustering. As searching space is small after clustering with semantic approach, we need very less time to search through billions of web pages or documents in fraction of seconds or less.All the experimental values are the result of two words "data" and "mining" from documents which go through semantic clustering.

**Index Terms**— Document clustering, Semantic clustering, Parse document, Similarity estimator, Text parser, Membership values, Parse tree.

## 1 INTRODUCTION

The growth of the World Wide Web has enticed many researchers to attempt to devise various methodologies for organizing such a huge information source. Scalability issues come into play as well as the quality of automatic organization and categorization. Documents on the web have a very large variety of topics, they are differently structured, and most of them are not well-structured. The nature of the sites on the web varies from very simple personal home pages to huge corporate web sites, all contributing to the vast information repository. Search engines were introduced to help find the relevant information on the web, such as Google, Yahoo!, and AltaVista. However, search engines do not organize documents automatically; they just retrieve related documents to a certain query issued by the user. While search engines are well recognized by the Information Retrieval community, they do not solve the problem of automatically organizing the documents they retrieve.


- *Mr. Sumit D. Mahalle pursued Masters in Technology degree in computer science engineering in MPSTME, NMIMS University Mumbai, Maharashtra,India, PH- +91 9096745266. E-mail: sumitmahalle@gmail.com*
- *Dr. Ketan Shah is currently Associate Professor in MPSTME,NMIMS University Mumbai, Maharashtra,India, PH- +91 9892793545 E-mail: ketanshah@nmims.edu*


The work will focus on the problem of mining the useful information from the collected web documents using **semantic based approach of document clustering** of the text, from the downloaded web documents.[1]

This project work will try to achieve some or all of the following objectives.

- To collect the web pages related to application domain.
- To generate various rules as per selected domain.
- To implement semantic based approach of document clustering in web text mining.
- To retrieve (mine) relevant information to the user from collected web pages.
- To analyze the retrieved result.[1]

**Problems:**

1. Conventional systems mainly use the presence or absence of keywords to mine texts.
2. Only deal with simple word counting and frequency distributions of term appearances.
3. Information overload problem.
4. Do not capture the meaning behind the words, which results in limiting the ability to mine the texts.

Therfore we need to cluster the document properly,the user to get proper data who he want to search in WWW world [8].

Our work proposes a technique to automatically cluster these documents into the related topics. Clustering is the proven technique for document grouping and categorization based on the similarity between these documents which is main aim to do it in Semantic way.

The work will focus on the problem of mining the useful information from the collected web documents using semantic based approach of document clustering of the text, from the downloaded web documents and the aim of the paper is to try and measure the efficiency of Document clustering with using Semantic approach to enhance web mining.

The proposed system is based on implementation of the semantic based approach of document clustering technique to enhance web mining. In this work, focus is on the problem of mining the useful information from the collected web documents by semantically document clustering of the text, from the downloaded web documents. [8].

## 2 DOCUMENT CLUSTERING

First the Clustering is one of the techniques to improve the efficiency in information retrieval for improving search and retrieval efficiency. It is a data mining tool to use for grouping objects into clusters. Clustering divides the objects (Documents) into meaningful groups based on similarity between objects. Documents within one cluster have high similarity with each other, but low similarity with documents in other clusters [7].

### 2.1 SEMANTIC APPROACH FOR DOCUMENT CLUSTERING

The proposed system is based on implementation of the semantic based approach of document clustering technique to enhance web mining. In this work, focus is on the problem of mining the useful information from the collected web documents by semantically document clustering of the text, from the downloaded web documents [5].

The three main parts in this approach [2] -

- Text Parser.
- Similarity Estimator.
- Mining Process.

Test parser is responsible to convert the text into tree like structure called as parse tree which is based on semantic notations, Similarity Estimator: Similarity estimator is used to measure percentage of similarity between two tress which are responsible for clustering the documents. [6]

Finally last step is mining the data for user. Mining process will find out the proper text for the user by using its membership values.

## 3 ALGORITHM

1. Take data set of 50 web pages i.e. IEEE paper web pages.
2. Apply the extraction technique on these web pages and parse those documents.
3. Find the membership values of each keyword of each data set file in database.
4. Apply the Semantic Clustering algorithm to these extracted web pages of parsed documents and cluster it to particular folder depending on membership values.
5. Experimental Result with final values of work done.

We are doing this project based on some web documents. By applying the technique of Semantic based document clustering on 50 document set i.e. 50 IEEE paper web pages.

**Disadvantages in making of Parse Tree:**

Tree like structure to show the parse document is too much overhead for large documents. Syntactic restrictions are not well-defined, hence difficult to produce trees.

This, however, is not such a practical method because the process is simply too time consuming and computationally costly. So in place of trees we can make the simple text file to store the parse document, and can remove this ambiguity. [1]

## 4 SELECTION AND MEANING CALCULATIONS

Here is the coding part of in dot net language which is showing the selection part of the two clusters in which all files in data set are going to clusters.

We can cluster as many of clusters as we want just by typing here with separation of pipe symbol (|). [1]

Meaning of cluster can be finding by giving its synonyms, in above example it gives the synonyms for cluster "mining" which are 'retrieval', 'extraction', 'exploit' etc., we can give as many synonyms as we want.

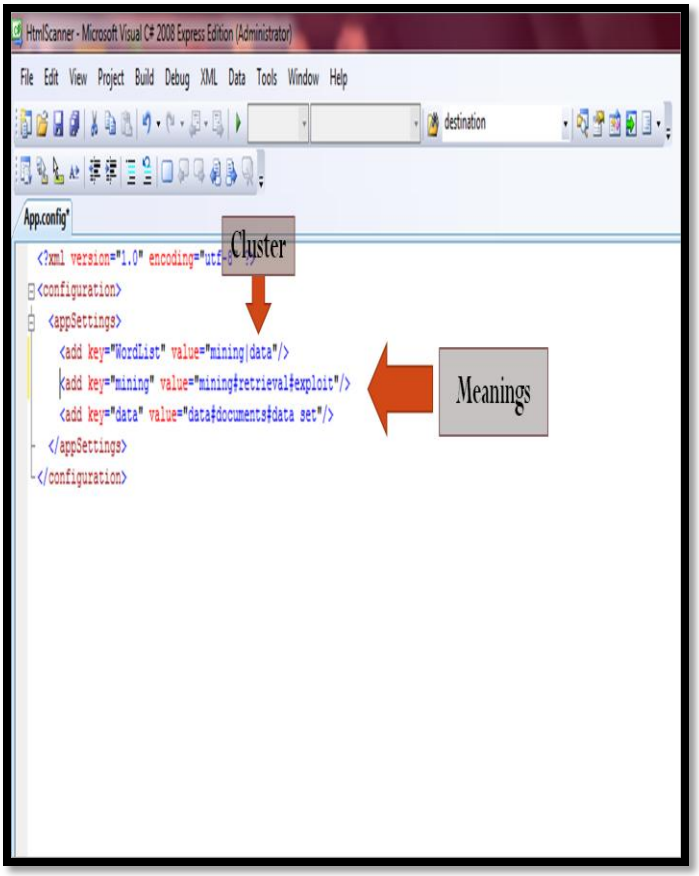



**Figure 1:** Meaning Calculation of Cluster

These synonyms will affect on counter, counter will affect on membership value, membership value will affect on clustering. In this way by adding this synonym we are finding the meaning of clusters which is major and important part of semantic clustering for populate the accurate result in terms of clustering.

## 5 EXPERIMENTAL VALUES

Following are the membership values of all 50 web documents which were extracted before applying to get these values, which is responsible for clustering.

This figure is membership values of word 'data' and its synonym, means it counts the total 'data' and its synonyms 'document', 'dataset' and count its membership values.

Similarly it counts words like 'mining' with synonyms 'retrieval', 'exploit' from all web documents i.e. 50 IEEE papers.

so as if program will find the instance of word 'retrieval' in any of files, it will increase the 'mining' counter by one, as counter increase it will affect on membership value, and membership value will be decide the which cluster that file is belong.

1) As documents preprocessed for removing frequently used stop words and then in main algorithm rules are generated as per the association in the words of query provided by user and these rules will be added into rule base and at the last step number of clusters will be created from the processed text using semantic parsing document clustering. The clusters will be selected as per membership values of word and then file will be cluster to similar folder in which that file belongs with higher number of membership values.
2) As we are having Membership values of each data file, by defining some threshold membership values, file will be cluster to those folders with having greater membership value of that folder or word. By calculating membership it is easy to find out the nearest cluster of data. If membership values of comparing topic/words or folder will be equal then that file will be cluster in all the folders.

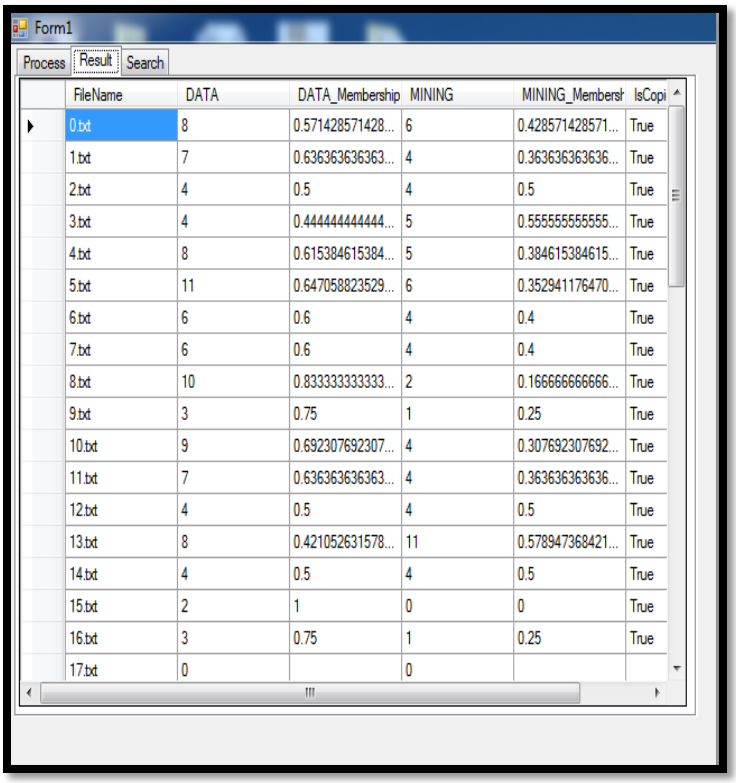
**Membership = (Number of instances of that word) /**
**(Total number of instances of all words.)**

If word "data" found 8 times in all one parse file,

Word "mining" found 6 times in same parse file which was for word "data",

Then, **Membership for word "data" =>** 8/14 = 0.571428

**Membership for word "mining"=>** 6/14= 0.428571….shown in Fig.2 at 0.txt.




| FileName | DATA | DATA_Membership | MINING | MINING_Membersh | IsCopi |
|---|---|---|---|---|---|
| 0.txt | 8 | 0.571428571428... | 6 | 0.428571428571... | True |
| 1.txt | 7 | 0.636363636363... | 4 | 0.363636363636... | True |
| 2.txt | 4 | 0.5 | 4 | 0.5 | True |
| 3.txt | 4 | 0.444444444444... | 5 | 0.555555555555... | True |
| 4.txt | 8 | 0.615384615384... | 5 | 0.384615384615... | True |
| 5.txt | 11 | 0.647058823529... | 6 | 0.352941176470... | True |
| 6.txt | 6 | 0.6 | 4 | 0.4 | True |
| 7.txt | 6 | 0.6 | 4 | 0.4 | True |
| 8.txt | 10 | 0.833333333333... | 2 | 0.166666666666... | True |
| 9.txt | 3 | 0.75 | 1 | 0.25 | True |
| 10.txt | 9 | 0.692307692307... | 4 | 0.307692307692... | True |
| 11.txt | 7 | 0.636363636363... | 4 | 0.363636363636... | True |
| 12.txt | 4 | 0.5 | 4 | 0.5 | True |
| 13.txt | 8 | 0.421052631578... | 11 | 0.578947368421... | True |
| 14.txt | 4 | 0.5 | 4 | 0.5 | True |
| 15.txt | 2 | 1 | 0 | 0 | True |
| 16.txt | 3 | 0.75 | 1 | 0.25 | True |
| 17.txt | 0 | | 0 | | |

**Figure 2:** Experimental values of parsed data.

**Experimental Result with final values of work done**:
These are the membership values of all 50 web documents which were extracted before applying to get these values, which is responsible for clustering.

## 6 DISTRIBUTION OF FILES IN CLUSTER

Here we can see the all distribution of 50 documents set to particular cluster. As we are having 50 web documents, these web documents are going through syntactic analysis result into 50 parse text files. These parse text files are pass into Semantic clustering and then it will store into respective cluster by using and comparing its membership value.

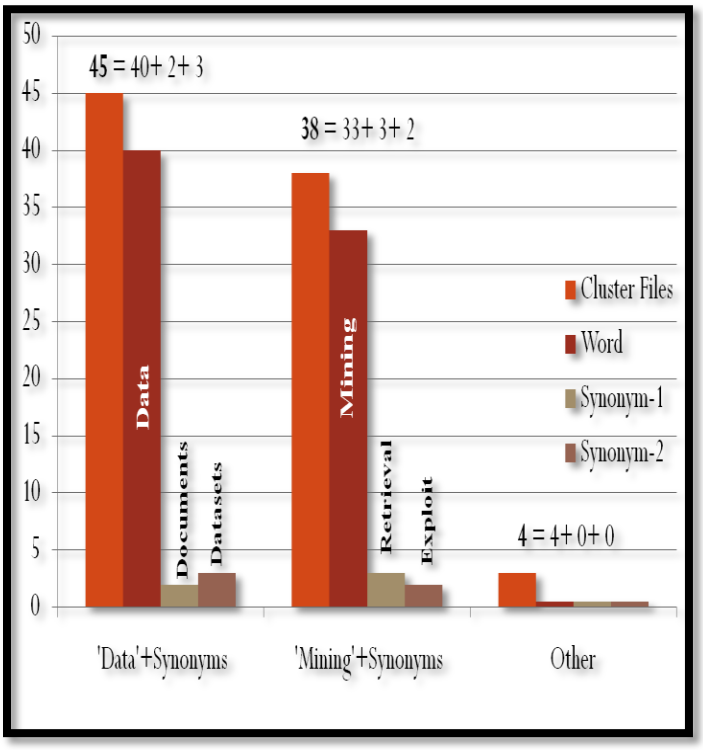



**Figure 3**: Distribution of files in cluster.

In figure Y-axis shows number of files as we are taking 50 web documents hence X-axis showing all 50 files i.e. datasets. X-axis shows the main Word-cluster with their synonym and meaning.
As shown in figure "data" cluster is the combination of all files containing word 'Data', 'Documents' and 'Datasets'. So "data" cluster is having word 'data'(40 files) + word 'Documents'(2 files) and word 'Datasets'(3 files) so total instances of word "data" and its meanings are found in 45 files which is belong to same "data" cluster.

Similarly "mining" cluster is the combination of all files containing word 'Mining', 'Retrieval' and 'Exploit'. So "mining" cluster is having word 'mining'(30 files) + word 'Retrieval'(3 files) and word 'Exploit'(2 files) so total instances of word "mining" and its meanings are found in 38 files which belong to same "mining" cluster.
And last files in figure i.e. "other" represents the files which are belongs to neither "data" word nor "mining" word. All these clustering is depend upon its membership values.If we decide some threshold that those files having their respective word threshold more than 0.4 goes to that cluster.
Eg. if 1.txt file is having its Mining-Membership value 0.3333 and Data-Membership value 0.6666 so it will cluster in "data" folder as it is having greater membership value in terms of word 'data'.[1]

## 7 SUMMARY & DISCUSSION

Semantic based document clustering in web text mining can be started with the collection of web pages related to the application domain which is from IEEE site of particular topic i.e.. Data mining, followed by conversion of web pages to text documents this is called as a parse document. In this parsing process, documents will be preprocessed for removing frequently used stop words and then we are applying semantic approach to these parse data to find out the membership of each text or parsed file which is calculate depending upon meaning of calculation of that clusters.
We did our testing for finding words "data" and "mining", we can take as many words for clustering. After that Clusters will be created of as many words you want to divide and cluster the data, If you give 10 words for clustering, then 10 folders of that name will be created and files will be cluster to that respective folder in which that file belongs. Finally by using processed text data with membership values of that words using semantic parsing clustering algorithm, documents will be cluster.

## 8 Conclusion

Although a conclusion may review the main points of the An optimized version of the parse text generator is being developed which will be used by a clustering algorithm to group documents. Threshold will be the deciding criteria when matching the text given by the parse-text generator. So by giving proper threshold and synonyms or meaning of clusters we can get as many clusters stated by us with the proper document clustering. "Semantic based document clustering" is a technique which is directly work on textual part of web document s in our database; there are very few technique which are based on textual data clustering.

This method for document clustering is very new and advance technique which gives attention on meaning finding of clusters at the time of dividing it, till now there are very rare or no method which can base on meaning calculation with its membership values. As we can put as many synonyms in the coding part of particular cluster we will get more proper response in clustering. As searching space is small after clustering with semantic approach, we need very less time to search through billions of web pages or documents in fraction of seconds or less.

## References

[1] Mr. Sumit D. Mahalle and Dr. Ketan Shah, "Semantic Based Approach for Document Clustering", *Journal of Sci., Engg. & Tech. Mgt. Vol 4 (1), MPSTME ,Mumbai. July 2012.*

[2] Khaled B. Shaban, "A Semantic Approach for Document Clustering", Journal of software,vol.4.5,July 2009K.

[3] Anupam Joshi and Raghu Krishnapuram,2003 Anupam Joshi and Raghu Krishnapuram, " Robust Fuzzy Clustering Methods to Support Web Mining", *in Proceedings of the Workshop on Data Mining and Knowledge Discovery , SOGMOD ,1998.*

[4] Bezdek J. C,1988 Bezdek J. C, *"Pattern Recognition with Fuzzy Objective Function Algorithms*", Plenum Press, New York, 1988.

[5] *Hsinchun Chen and Michael Chau,* "Web Mining: Machine learning for Web Applications", *Annual Review of Information Science and Technology 2003.*

[6] Mr. Rizwan Ahmad and Dr. Aasia Khanum, "Document Topic Generation in Text Mining by Using Cluster Analysis with EROCK", *International Journal of Computer Science & Security (IJCSS), Volume (4) : Issue (2) Aug 2008.*

[7] M.E.S. Mendes Rodrigues and L. Sacks, "A Scalable Hierarchical Fuzzy Clustering Algorithm for Text Mining", *Department of Electronic and Electrical Engineering University College London Torrington Place, London, WC1E 7JE, United Kingdom, 2004.*

[8] Shady Shehata, Member, IEEE, Fakhri Karray, Senior Member, IEEE, and Mohamed S. Kamel, Fellow, IEEE, "An Efficient Concept-Based Mining Model for Enhancing Text Clustering", *IEEE Trans. On Knowledge and Data Engineering, Vol. 22, No. 10, Oct 2010.*