# Survey on Document Clustering Using Fuzzy Logic

Vivek Joshi[1], Pranav Naik[2], Akshay Shinde[3], Girish Shinde [4] and Manisha Mali[5]

[1] Vivek Joshi, Department of Computer Engineering, Vishwakarma Institute of Information Technology, Pune.
e-mail id : vivek.joshi@viit.ac.in
[2] Pranav Naik, Department of Computer Engineering, Vishwakarma Institute of Information Technology, Pune.
e-mail id : pranav.naik@viit.ac.in
[3]Akshay Shinde, Department of Computer Engineering, Vishwakarma Institute of Information Technology, Pune.
e-mail id : akshay.shinde@viit.ac.in
[4]Girish Shinde, Department of Computer Engineering, Vishwakarma Institute of Information Technology, Pune.
e-mail id : girish.shinde@viit.ac.in
[5] Prof. Manisha Mali, Department of Computer Engineering, Vishwakarma Institute of Information Technology, Pune.
e-mail id : manisha.mali@viit.ac.in

## ABSTRACT

Handling large and increasing amount of unstructured digital data is very difficult. Document clustering is very efficient way to handle such data. There are various methods used for clustering the documents viz., partitioning and hierarchical method, centroid based, distribution based, density based. This paper demonstrates a survey on the existing methods.

**Keywords**— Pre-Processing, Fuzzy Logic, Clustering.

## 1. INTRODUCTION

Document clustering is a technique aimed toward grouping similar documents in clusters. The main aim is to make a system that clusters similar sort of documents efficiently. Document clustering is very useful in information retrieval, data mining, web mining, text mining etc. The approach is to extract words from documents and then to compare documents with each other, documents with higher degree of matching will be clustered together. This approach can give us accuracy and the degree to which documents are similar to each other. Everyday documents are scaling at very high level, as tremendous amount of documents are generated it becomes difficult to search a particular or a group of file(s) / document(s). Goal is to create clusters such that documents in the same clusters should be as similar as possible whereas, documents in one cluster should be as dissimilar as possible from documents in other clusters. As we know partitioning methods consist of different algorithms such as k-means, k-medoids etc. In k-means first we decide number of clusters to be formed then we take arbitrary cluster centroid from given data, then assign the values which are similar to the centroid then we update the centroid and repeat the process. k-medoid relies on center of mass (or medoids) hard by minimizing absolutely the distance between the points, rather than minimizing the square distance. As a result, it's more robust to noise and outliers than k-means.

## 2. LITERATURE SURVEY

A) By Amy J. C. Trappey, Charles V. Trappey, Fu-Chiang su, and David W. Hsiao [2] "A Fuzzy Ontological Knowledge Document Clustering Methodology". Documents which provide different views and details are clustered using ontology, which carries meanings and relations. This technique of ontological structure based extracts representatives' information rather than extracting general phrases as in key-phrase based k-means approach. It combines technique of ontological knowledge representation with fuzzy logic control on linguistic expressions and similarity measures are derived among the patents documents for clustering. It designs a coaching set of patents employing a linguistic communication process and tagging tool known as MontyLingua. Then the probabilities of the concepts from the documents are calculated. The calculated concept probabilities in any given patent document are then used for clustering the patents with fuzzy logic inferences. It also uses knowledge based RDF editing tool called Protégé which defines ontology schema with the help of graphical interface. This approach is then tested on 3 cases of chemical mechanical polishing (CMP) patents, patent news content, and radio-frequency identification (RFID) patents.

B) By Anuj Sharma, Renu Dhir [3] "A Wordsets Based Document Clustering Algorithm for Large Datasets". Document clustering is a crucial and great tool for applications like search engines. It provides the user read of the data contained within the documents. It gives the user view of the information contained in the documents. This paper proposes a wordsets-based document clustering (WDC) approach for clustering. Here instead of comparing documents and clustering them based on common words, clustering the document based on closed word sets is done. It first searches frequent closed word sets using association rule mining and then forms initial clusters of the documents, with each cluster representing a single closed wordset. Here Documents can be clustered according to meaning of word. Also fast implementation of Apriori algorithm should be done. So drawback is speed of implementation of algorithm is slow.

C) By Chengzhi ZHANG [4] "Document Clustering Description Based on Combination Strategy". The traditional algorithm of clustering can cluster the documents, but they cannot give concept description to the clustered results. This paper focuses on labeling the clustered set of documents. Here Description comes First (DCF) method is used to generate document clustering description. But there is a semantic interval between clustering description and cluster central vector. This decreases readability of clustering description. So, after that a combination strategy of DCF and Description comes last is used for solving the problem of weak readability of clustering description. Here relevance of the cluster can be determined. The method proposed in this paper is based on integrated learning method. One of major drawback of strategy is effective evaluation of search results of clustering is not done.

D) By Manolis Wallace, Giorgos Akrivas and Giorgos Stamou [6] "Automatic Thematic Categorization of Documents Using a Fuzzy Taxonomy and Fuzzy Hierarchical Clustering". This paper defines the matter of automatic detection of thematic categories throughout a semantically indexed document, and confirm the foremost obstacles with this method. What is more, it explains however detection of thematic classes is achieved, with the use of a fuzzy quasi-taxonomic relation.

E) By Tatiane M. Nogueira, Solange O. Rezende Heloisa A. Camargo [10] "On the Use of Fuzzy Rules to Text Document Classification". In this paper they have aimed to bring a group of texts closer with respect to relevance related to a particular topic. This is done so that it would be easy for a user to organize and classify a particular document and be a support for the decision making in the future. Then in the pattern extraction step clustering algorithm is applied, fuzzy clustering, where the document can belong to more than one domain with varying degrees of relevance. The relevance of the documents can be represented as a document can belong 'very much' or a 'bit' to a particular domain. First the fuzzy C-means algorithm is applied to cluster text documents into groups so all documents belong to any or all groups with completely different degrees of relevance. Second they have generated fuzzy rules, to be used to classify the documents. The main drawback is that preprocessing of the documents is needed because it directly interferes on the results. Preprocessing is needed so that most relevant terms are selected to make the set of terms more concise but not less representative than the whole document.

F) By S. C. Punithaa and M. Punithavallib [8] "Performance Evaluation of Semantic Based and Ontology Base Text Document Clustering Techniques". As the volume of information continues to increase, there is growing interest in helping people better find, filter and manage these resources. Text clustering, which is the process of grouping documents having similar properties based on semantic and statistical content, is an important component in many information organization and management tasks. In the present research work two novel approaches to document clustering was considered and their methods and performance were analyzed.

The first approach, HSTC, uses a hybrid approach to mix pattern recognition algorithms with linguistics driven processes. The second approach, TCFS, used ontology based feature selection for clustering. Experiments proved that both techniques were efficient in cluster method, however the performance of TCFS was slightly higher in terms cluster quality, but slow. In future, both these methods can be combined to take advantage of quality clustering in a fast manner.

G) By S.Santhana Megala, Dr.A.Kavitha and Dr. A.Marimuthu[7] “International Journal of Advanced Research in Computer Science and Software Engineering”. In this paper an improvised stemming algorithm is given which is used to produce a clear and meaningful stem. Word Stemming is a method of reducing the words to their root word by removing the attached suffixes and prefixes before indexing, to combine the words. It simply reduces the grammatical or inflectional or derivational form of a word to a common base form called stem. The proposed algorithm is based on Porters stemming algorithm. A Slight modification in Porters algorithm is done without compromising the efficiency and simplicity. Here size of the output is slightly larger than the other algorithms, but it is negligible when compared to the Meaningful output, because if the stem word isn’t a meaty word within the Preprocessing stage then it needs a manual correction at the post processing which leads to a time delay. The experimental result shows that this improvised algorithm shows a better accuracy in generating a meaning full stems comparing to standard porters algorithm. Due to this, the error rate is reduced and algorithm becomes more efficient. This stemming algorithm is further applied to the research work on summarization and classification of textual data to utilize the efficiency and simplicity. The only drawback of this algorithm is it produces large stem words.

H) By Sumit Goswami and Mayank Singh Shishodia [9] “A Fuzzy Based Approach To Text Mining And Document Clustering”. This paper presents use of fuzzy logic in text mining to cluster documents by taking an example where the documents were clustered into two topics :- sports and politics. The advantage of using fuzzy logic over probabilitywas that in the former, by calculating the degree to which a given document belonged to either categories- sports as well as politics. This is not possible in the probability model. In different words, rather than merely speaking whether or not the document belonged to sports or politics, it also tells the degree to which the document characteristics resembled that of a sports-related document as well as a politics-related document. By doing this for all documents in the data-set, and also compares two documents and tell which one belongs more to which topic. Also, because every document will have some membership values in each of the clusters, no useful document will ever excluded from search results.

I) By Khaled M. Hammouda and Mohamed S. Kamel [5] “Efficient Phrase-Based Document Indexing for Web Document Clustering”. This paper presents to parts of document clustering. The first part is phrase based document index model. The second half associates’ progressive document cluster algorithmic program supported maximizing the tightness of clusters by rigorously looking the pairwise distribution in aspect clusters. For this paper they need thought about the document cluster supported phrase analysis with single word analysis i.e. similarity between documents should be based on matching phrases. They have planned a system for agglomeration supported two key ideas. The first, similarity between documents are supported the matching phrases and their weights. The second thought is that the progressive agglomeration of documents employing a bar graph primarily based methodology to maximize the tightness of clusters by care- totally looking the similarity distribution within every cluster.. The system consists of four components: 1). Document restructuring scheme that identifies different document parts and assigns levels of significance to the parts. 2).A phrase based document indexing model, it constructs a Document Index Graph that captures the structure of sentences rather than single words only. 3) A phrase based similarity measure for scoring the similarity between two documents according to matching phrases is implemented. 4).To maintain high cluster quality a concept called “similarity histogram” is implemented. The main drawback is that there is need for cluster similarity accuracy. The similarity calculation between documents can be improved by applying different similarity calculation strategies.

J) By A Muralidhar, V Pattabiraman [1] “An Efficient Association Rule Based Clustering of XML Documents”. It describes the method for an efficient association rule based clustering using XML documents. The unstructured information from the web is well represented in terms of extensible markup language (XML). XML is also used for storage, data representation and exchange the data across sites.

The algorithms for this mining of association rules from relational data are trained considerably with the help of several database query languages. Here hybrid technique of parallel Apriori and K-means is used, where similarity features between the frequent XML documents are found out using Euclidean distance to acquire the frequent documents and the resulting clusters are compared with Dunns Index along with the hadoop technology. The major drawback is frequent pattern mining becomes problematic in big data along with the expensive computational cost and memory use.

## 3. PROPOSED SYSTEM

We are proposing a method which will cluster the documents using generalized fuzzy logic. Initially the documents will go under pre-processing. Then features like title sentence, proper noun, numerical data and top words will be extracted. We will compare features extracted from each document with every other document. After that a weighted matrix will be generated on the basis of comparison of features. Then using fuzzy logic, documents with high degree of matching will be clustered together.

## 4. CONCLUSIONS

Handling large and increasing amount of unstructured data is very difficult. So to overcome this flaws this paper studied many of the systems in depth as mentioned in literature survey (section 2), so as our future work to bring innovative methodology we are using fuzzy based approach for document clustering.

## REFERENCES

[1] A. Muralidhara and V. Pattabiramanb, "An Efficient Association Rule Based Clustering of XML Documents", 2nd International Symposium on Big Data and Cloud Computing (ISBCC), 2015.

[2] Amy J. C. Trappey, Charles V. Trappey, Fu-Chiang Hsu, and David W. Hsiao, "A Fuzzy Ontological Knowledge Document Clustering Methodology" ,IEEE Transactions on Systems, Man, and Cybernetics- Part B:Cybernetics, vol.39, No. 3, June 2009.

[3] A. Sharma and R. Dhir, "A Wordsets Based Document Clustering Algorithm for Large Datasets", in International Conference on Methods and Models in Computer Science, 2009.

[4] C. ZHANG, "Document Clustering Description Based on Combination Strategy", Fourth International Conference on Innovative Computing, Information and Control, 2009.

[5] K. M. Hammouda and M. S. Kamel, "Efficient Phrase-Based Document Indexing forWeb Document Clustering", IEEE Transactions on knowledge and data engineering, vol.16, No.10, Oct 2014.

[6] G. A. Manolis Wallace and G. Stamou, "Automatic Thematic Categorization of Documents Using a Fuzzy Taxonomy and Fuzzy Hierarchical Clustering", in The IEEE International Conference on Fuzzy Systems, 2010.

[7] S.Santhana Megala, Dr.A.Kavitha and Dr. A.Marimuthu, "Improvised Stemming Algorithm – TWIG", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, Issue 7, 2013.

[8] S. C. Punithaa and M. Punithavalli, "Performance Evaluation of Semantic Based and Ontology Based Text Document Clustering Techniques", International Conference on Communication Technology and System Design, 2011.

[9] S Goswami and M. S. Shishodia, "A Fuzzy Based Approach To Text Mining and Document Clustering", arXiv, 2013

[10] Tatiane M. Nogueira, Solange O. Rezende, Heloisa A. Camargo, "On the Use of Fuzzy Rules to Text Document Classification", 10th International Conference on Hybrid Intelligent Systems, 2010.