Clustering With Multiview Point Based Similarity Measure

[1]Ms. A. B. Jadhav, [2]Prof. Dr. S. P. Abhang

Chh. Shahu College Of Engineering Kanchanwadi, Paithan Road, Aurangabad, Maharashtra, India.

[1]*jadhavamruta19@gmail.com* , [2]*tapadevrushali@gmail.com*

Abstract: All clustering methods have to assume some cluster relationship among the data objects that they are applied on. Similarity between a pair of objects can be defined either explicitly or implicitly. In this paper, we introduce a novel multiviewpoint-based similarity measure and two related clustering methods. The major difference between a traditional dissimilarity/similarity measure and ours is that the former uses only a single viewpoint, which is the origin, while the latter utilizes many different viewpoints, which are objects, assumed to not be in the same cluster with the two objects being measured. Using multiple viewpoints, more informative assessment of similarity could be achieved. Theoretical analysis and empirical study are conducted to support this claim. Two criterion functions for document clustering are proposed based on this new measure. We compare them with several well-known clustering algorithms that use other popular similarity measures on various document collections to verify the advantages of our proposal.

Keywords: Clustering, Multiviewpoint, Streaming.

I. INTRODUCTION

Clustering is one of the most interesting and important topics in data mining. Clustering is the grouping of a particular set of objects based on their characteristics, aggregating them according to their similarities. Regarding to data mining, this methodology partitions the data implementing a specific join algorithm, most suitable for the desired information analysis. The aim of clustering is to find intrinsic structures in data, and organize them into meaningful subgroups for further study and analysis [1]. This clustering analysis allows an object not to be part of a cluster, or strictly belong to it, calling this type of grouping hard partitioning. In the other hand, soft partitioning states that every object belongs to a cluster in a determined degree. More specific divisions can be possible to create like objects belonging to multiple clusters, to force an object to participate in only one cluster or even construct hierarchical trees on group relationships [2].

A cluster of data objects can be treated as one group. While doing cluster analysis, we first partition the set of data into groups based on data similarity and then assign the labels to the groups. The main advantage of clustering over classification is that, it is adaptable to changes and helps single out useful features that distinguish different groups. Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing [4]. Clustering can also help marketers discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns [5]. In the field of biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionalities and gain insight into structures inherent to populations. Clustering also helps in identification of areas of similar land use in an earth observation database. It also helps in the identification of groups of houses in a city according to house type, value, and geographic location. Clustering also helps in classifying documents on the web for information discovery. Clustering is also used in outlier detection applications such as detection of credit card fraud. As a data mining function, cluster analysis serves as a tool to gain insight into the distribution of data to observe characteristics of each cluster.

Intra the inter-class similarity is low. The quality of a clustering result also depends on both the similarity measure used by the method and its implementation. The quality of a

clustering method is also measured by its ability to discover some or all of the hidden patterns[6].

There are certain Requirements for Clustering in Data Mining and are as follows.

Scalability − we need highly scalable clustering algorithms to deal with large databases.

Ability to deal with different kinds of attributes − Algorithms should be capable to be applied on any kind of data such as interval-based (numerical) data, categorical, and binary data.

Discovery of clusters with attribute shape − the clustering algorithm should be capable of detecting clusters of arbitrary shape. They should not be bounded to only distance measures that tend to find spherical cluster of small sizes.

High dimensionality − the clustering algorithm should not only be able to handle low-dimensional data but also the high dimensional space.

Ability to deal with noisy data − Databases contain noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters.

Interpretability − the clustering results should be interpretable, comprehensible, and usable.

II. LITERATURE SURVEY

Duc Thang Nguyen, Lihui Chen, Chee Keong Chan, propose a Multiviewpoint-based Similarity measuring method, named MVS. Theoretical analysis and empirical examples show that MVS is potentially more suitable for text documents than the popular cosine similarity. Based on MVS, two criterion functions, IR and IV, and their respective clustering algorithms, MVSC-IR and MVSC-IV, have been introduced. Compared with other state-of-the-art clustering methods that use different types of similarity measure, on a large number of document data sets and under different evaluation metrics, the proposed algorithms show that they could provide significantly improved clustering performance.[1]

Daewon Lee and Jaewook Lee propose a novel dissimilarity measure based on a dynamical system associated with support estimating functions. Theoretical foundations of the proposed measure are developed and applied to construct a clustering method that can effectively partition the whole data space. Simulation results demonstrate that clustering based on the proposed dissimilarity measure is robust to the choice of kernel parameters and able to control the number of clusters efficiently. Despite their advantages over other clustering methods, the existing support-based clustering algorithms have some drawbacks. First, out-of-the sample points outside of the generated cluster boundaries cannot directly be assigned a cluster label. Second, the clustering results are very sensitive to the choice of kernel parameters used for a support estimate since the boundaries can show highly fluctuating behavior caused by small changes of the kernel parameters. Finally, it is difficult to control the number of clusters when they are applied to clustering problems with a priori information of the cluster numbers. [2]

Hung Chim and Xiaotie Deng propose a Phrase has been considered as a more informative feature term for improving the effectiveness of document clustering. In this paper, we propose a phrase-based document similarity to compute the pairwise similarities of documents based on the Suffix Tree Document (STD) model. By mapping each node in the suffix tree of STD model into a unique feature term in the Vector Space Document (VSD) model, the phrase-based document similarity naturally inherits the term tf-idf weighting scheme in computing the document similarity with phrases. [3]

S. Zhong propose a spherical k-means algorithm, i.e., the k-means algorithm with cosine similarity, is a popular method for clustering high-dimensional text data. In this algorithm, each document as well as each cluster mean is represented as a high-dimensional unit-length vector. However, it has been mainly used in batch mode. That is, each cluster mean vector is updated, only after all document vectors being assigned, as the (normalized) average of all the document vectors assigned to that cluster. This paper investigates an online version of the spherical k-means algorithm based on the well-known Winner-Take-All competitive learning. In this online algorithm, each cluster centroid is incrementally updated given a document. We demonstrate that the online spherical k-means algorithm can achieve significantly better clustering results than the batch version, especially when an annealing-type learning rate schedule is used. [4]

Inderjit S. Dhillon, Dharmendra S. Modha propose a concept decompositions to approximate the matrix of document vectors; these decompositions are obtained by taking the least-squares approximation onto the linear subspace spanned by all the concept vectors. We empirically establish that the approximation errors of the concept decompositions are close to the best possible, namely, to truncated singular value decompositions. As our third contribution, we show that the concept vectors are localized in the word space, are sparse, and tend towards orthonormality. In contrast, the singular vectors are global in the word space and are dense. [6]

III. SYSTEM ARCHITECTURE

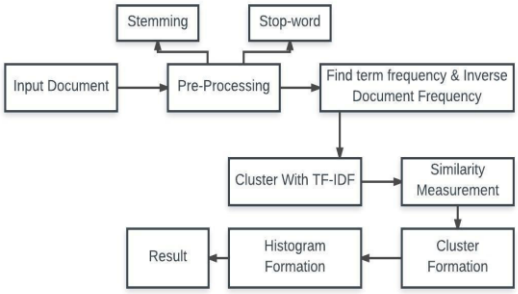

Fig. 1 Proposed System

The above Figure shows the Architecture of proposed system. We are going to pass HTML Document as a input Document. On this document we have to perform preprocessing. Preprocessing is an important task and critical step in Text mining, Natural Language Processing (NLP) and information retrieval (IR). In the area of Text Mining, data preprocessing used for extracting interesting and non-trivial and knowledge from unstructured text data. Information Retrieval (IR) is essentially a matter of deciding which documents in a collection should be retrieved to satisfy a user's need for information. The user's need for information is represented by a query or profile, and contains one or more search terms, plus some additional information such as weight of the words. Hence, the retrieval decision is made by comparing the terms of the query with the index terms (important words or phrases) appearing in the document itself. The decision may be binary (retrieve/reject), or it may involve estimating the degree of relevance that the document has to query. Unfortunately, the words that appear in documents and in queries often have many structural variants. So before the information retrieval from the documents, the data preprocessing techniques are applied on the target data set to reduce the size of the data set which will increase the effectiveness of IR System The objective of this study is to analyze the issues of preprocessing methods such as Tokenization, Stop word removal and Stemming for the text documents. There are two main processes are associated with this preprocessing. And the processes are Stemming and Stop-Word Removal.

Stemming: Stemming is the process of conflating the variant forms of a word into a common representation, the stem. For example, the words: “presentation”, “presented”, “presenting” could all be reduced to a common representation “present”. This is a widely used procedure in text processing for information retrieval (IR) based on the assumption that posing a query with the term presenting implies an interest in documents containing the words presentation and presented.

Stop Word Removal: Many words in documents recur very frequently but are essentially meaningless as they are used to join words together in a sentence. It is commonly understood that stop words do not contribute to the context or content of textual documents. Due to their high frequency of occurrence, their presence in text mining presents an obstacle in understanding the content of the documents. Stop words are very frequently used common words like ‘and’, ‘are’, ‘this’ etc. They are not useful in classification of documents. So they must be removed. However, the development of such stop words list is difficult and inconsistent between textual sources. This process also reduces the text data and improves the system performance. Every text document deals with these words which are not necessary for text mining applications. After applying preprocessing find the term frequency and inverse domain frequency of document.

IV. RESULT ANALYSIS

A. Selecting HTML Document for Input

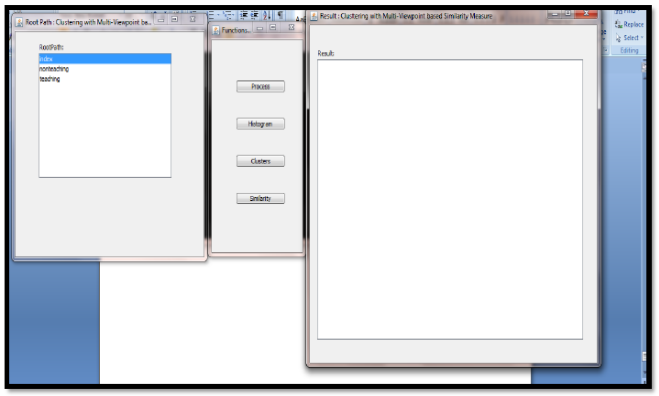

Figure 2 Selecting input Document

The input has to be taken in the html format. Html parser will read that html file and take all the meta tag in consideration.

B. Processing HTML document

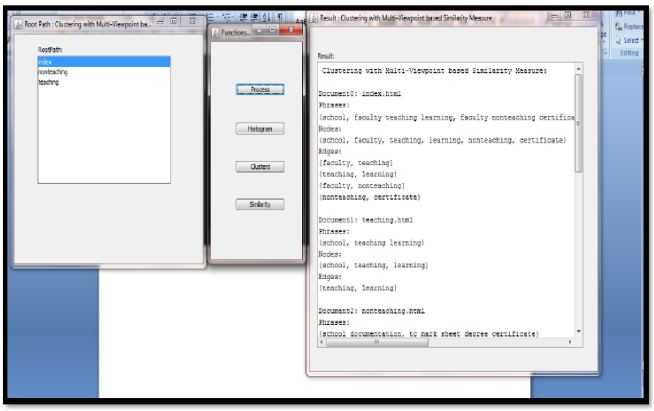

Fig. 3 Processing HTML document

This input file has to be taken for process and after processing all the values of meta tag are distributed in terms of nodes and edges. The cumulative document is the sum of all the documents, containing meta-tags from all the documents. We find the references (to other pages) in the input base document and read other documents and then find references in them and so on. Thus in all the documents their meta-tags are identified, starting from the base document.

Here is the processing of html document which results into cumulative document. It means it simply takes all the hyperlinks which are resent inside of this html document. The cumulative document is the sum of all the documents, containing meta-tags from all the documents. We find the references (to other pages) in the input base document and read other documents and then find references in them and so on.

C. Calculating similarity between documents

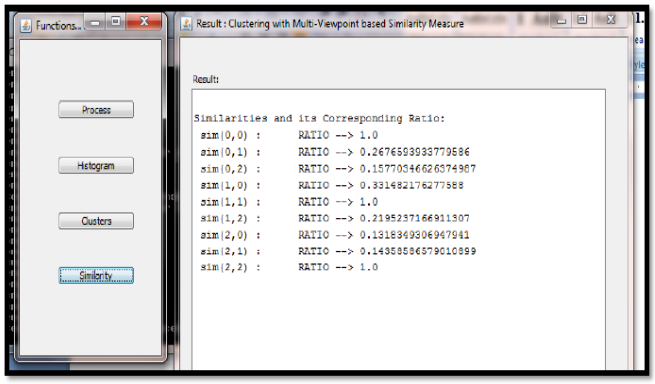

Fig. 4 Similarity Ratio

Above figure shows the similarity ratio between documents. If documents are exactly similar to each other, then their similarity ration is 1.0. If they are exactly dissimilar to each other, then their similarity is 0 and if they relatively similar to each other then it have some value in decimal.

D. Histogram Formation

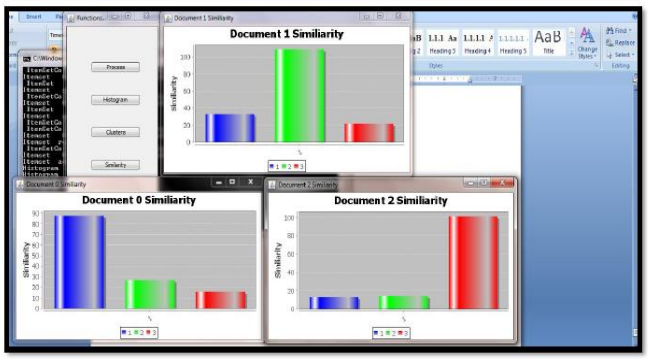

Fig. 5 Histogram

Once we have similarity ratio we can generate histogram for the same. The histogram is diagrammatic representation of similarities between documents. If the documents are relatively similar to each other, then their similarity ratio value is equal. In above figure document-0 and document-1 are relatively similar to each other and hence their similarity ratio is equal. And document-2 is not at all relatively similar with document-0 and document-2.

E. Cluster Formation

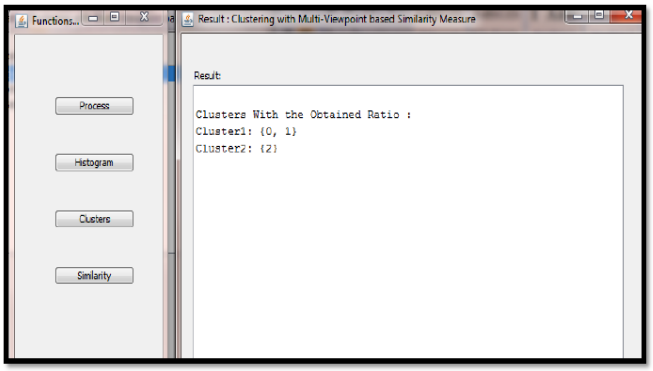

Fig. 6 Cluster Formation

If the documents are relatively similar to each other, then their similarity ratio value is equal. In figure Above document-0 and document-1 are relatively similar to each other and hence their similarity ratio is equal. And they are now part of single cluster. Document-2 is not at all relatively same with document-0 and document-2 and hence because of this document-2 is in different cluster.

V. PERFORMANCE ANALYSIS

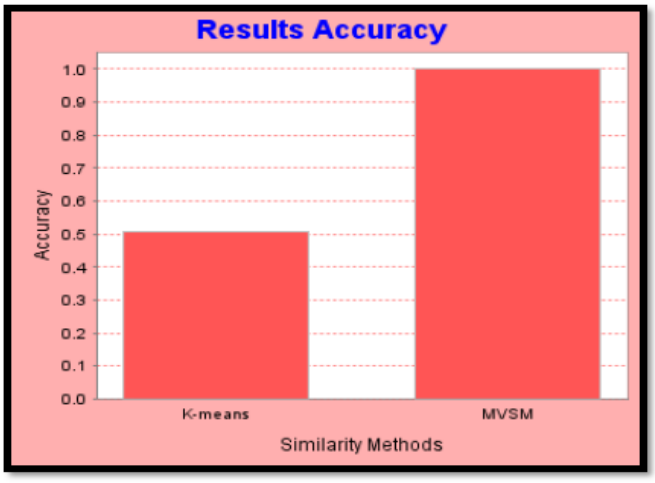

Fig. 7 Accuracy Result

The multi view point similarity measure will provide good result in high dimensional domain. According to our work more number of documents comes under high dimensional

domain. The irrelevant documentation is reduced here so that we can predict that the multi view will provide good result than the k-means algorithm.

VI. CONCLUSION

In this paper, we proposed a new similarity measure known as MVS (Multi-Viewpoint based similarity). Theoretical analysis and empirical examples show that MVS is potentially more suitable for text documents than the popular K-means similarity.

This project provided a successful implementation of a mvs with Cosine similarity measure technique derives the similarity between the documents. The weights in the cosine-similarity are found from the TF-IDF measure between the phrases (meta-tags) of the two documents. The empirical results and analysis revealed that the proposed scheme for similarity measure is efficient and it can be used in the real time applications in the text mining domain. IR and IV are the two criterion functions proposed based on MVS. This paper also concentrates on partitional clustering of documents. The key contribution of this paper is the fundamental concept of similarity measure from multiple viewpoints.

REFERENCES

[1] X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J.McLachlan, A. Ng, B. Liu, P.S. Yu, Z.-H. Zhou, M. Steinbach, D.J.Hand, and D. Steinberg, "*Top 10 Algorithms in Data Mining*," Knowledge Information Systems, vol. 14, no. 1, pp. 1-37, 2007.

[3] I. Guyon, U.V. Luxburg, and R.C. Williamson, "*Clustering:Science or Art?*," Proc. NIPS Workshop Clustering Theory, 2009.

[3] Dhillon and D. Modha, "*Concept Decompositions for Large Sparse Text Data Using Clustering*," Machine Learning, vol. 42,nos. 1/2, pp. 143-175, Jan. 2001.

[4] S. Zhong, "*Efficient Online Spherical K-means Clustering*," Proc.IEEE Int'l Joint Conf. Neural Networks (IJCNN), pp. 3180-3185, 2005.

[5] A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh, "*Clustering with Bregman Divergences*," J. Machine Learning Research, vol. 6,pp. 1705-1749, Oct. 2005.

[6] E. Pekalska, A. Harol, R.P.W. Duin, B. Spillmann, and H. Bunke,"*Non-Euclidean or Non-Metric Measures Can Be Informative*,"Structural, Syntactic, and Statistical Pattern Recognition, vol. 4109,pp. 871-880, 2006.

