Abstract
The main task of Clustering technique in data mining and text mining is to analyze datasets by dividing it into meaningful groups. Certain relationship exists between the objects in the dataset. The similarity between the data objects can be described either explicitly or implicitly. In the existing algorithm the similarity and dissimilarity between the objects is measured using single view point, which is the origin. The drawback is that the clusters can’t exhibit the entire set of relationships among objects. To overcome the conflict, in this paper a new measure of similarity called multiview point based similarity measure is proposed. This approach makes use of different viewpoints in clustering the web documents which show all relationships among objects where more informative assessment of similarity could be achieved. Analysis and experimental study are conducted in support of this approach. Several well known clustering algorithms that use other popular similarity measures on various document collections are compared to verify the advantages of our proposal.

Keywords
Multi View Point, Incremental Clustering, Data Mining, Clustering Algorithms

I. Introduction
Data mining is the gathering of information from pre-existing data stored in a database, the process of discovering interesting and useful patterns and relationships in large volumes of data. Clustering can be considered the most crucial topic in data mining. The process of clustering is to organise and group the similar objects together and to catch fundamental structures in data and classify them into meaningful subgroup for additional analysis. Clustered groups make search mechanisms easy and reduce the more number of operations and computational cost. Many clustering algorithms have been introduced every year and can be proposed for different research fields with different techniques. A new concept of novel similarity measure which works faster and provide consistent, high quality performance in the process of clustering high dimensional domain particularly text documents is introduced called “multiview point similarity measure with incremental clustering”.

II. Referred Information
One of the simplest unsupervised learning algorithm which solves the well known clustering problem and still one of the top data mining algorithms is the k-means algorithm. With a large number of variables, K-Means may be computationally faster than the other algorithms. K-Means easily partition the documents by calculating the euclidean distance and may produce tighter clusters than other clustering approaches, especially if the clusters are globular. Dist(di, dj) = \|d_i - d_j\|. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters, assume k clusters fixed a priori. In Fixed number of clusters it is difficult to predict what K should be, because different locations causes different results. As the k-algorithm measures only vector directions, it is sensitive regarding outliers. There is a difficulty in comparing quality of the clusters and does not work well with non-globular clusters. Another approach is the spherical k-means which measures the cosine similarity instead of the Euclidean distance which maximizes the similarity between the documents especially for hi-dimensional and sparse document clustering, \text{sim}(d_i, d_j) = \frac{1}{1 - \|d_i - d_j\|}. But does not follow sequential order of selecting clusters and does not cluster the entire document. Reminding up on the above conflicts in this paper an algorithm is implemented from the above and similar research findings which appears to us that the nature of similarity measure plays a very important role in the success or failure of a clustering method.

III. Multiview Point Similarity Measure With Incremental Clustering
Our goal is derive an algorithm to find out the similarity between the documents by performing the clustering accurately by more than one point of reference while the existing algorithms used single view point determining the angle between the two points viewing from the origin. As per our approach the assessment of how close or distant a pair of points is done, by observing at them from many different viewpoints. The description of this method is, Consider two points di and dj in cluster sr. The similarity between those two points is viewed from a point dy, which is outside the cluster the directions and distances to di and dj are indicated respectively by the difference vectors (di-dy) and (dj-dy). The similarity between the two documents is defined as Sim(di, dj) = 1/n-\text{nr} \sum (di-\text{dy}, dj-\text{dy}). An assumption on which this definition is based on is dy is not the same cluster as di and dj. The two objects to be measured must be in the same cluster, while the points from where to establish this measurement must be outside of the cluster. We call this proposal multiview point similarity measure with incremental clustering.

A. MVS Matrix
The validity test is designed as following: For each type of similarity measure, a similarity matrix A = \{a_{ij}\}_{n*n} is created. For CS, this is simple, as a_{ij} = |d_i - d_j|. The procedure for building MVS matrix is described. First, the outer composite w.r.t. each class is determined. Then, for each row ai of A, i = 1, . . . , n, if the pair of documents di and dj; j = i, 1, . . . , n are in the same class, a_{ij} is calculated. Otherwise, d_i is assumed to be in di’s class, and then a_{ij} is calculated.

B. Validity Test
After matrix A is formed, the procedure is used to get its validity score. For each document di corresponding to row ai of A, select qr documents closest to di. The value of qr is chosen relatively as percentage of the size of the class r that contains di, where percentage \epsilon(0,1) Then, validity w.r.t. di is calculated by the fraction of these qr documents having the same class label with di. The final validity is determined by averaging over all the rows of A. It is clear that validity score is bounded with in 0 and 1. The higher validity score a similarity measure has, the more suitable
it should be for the clustering task.

**Proposed Algorithm:**

**Procedure INITIALIZATION**

Select k seeds \( s_1, \ldots, s_k \) randomly

Cluster\([d_i] \leftarrow p = \arg\max_r \{s_r d_i\}, \forall i = 1, \ldots, n\)

\[ D \leftarrow \sum_{d_i, n_r} \Delta s, n_r \leftarrow |S_i|, \forall r = 1, \ldots, k \]

End procedure

**procedure** REFINEMENT

repeat

\{V[1:N]\} \leftarrow \text{random permutation of } \{1, \ldots, n\}

for \( j \leftarrow 1: n \) do

\( i \leftarrow v[j] \)

\( p \leftarrow \text{cluster}[d_i] \)

\( \Delta I_p \leftarrow I(n_p - 1, D_p - d_i) - I(n_p, D_p) \)

\( Q \leftarrow \arg\max_r \{I(n_r + 1, D_r + d_i) - I(n_r, D_r)\} \)

If \( \Delta I_p + \Delta I_q > 0 \) then

Move \( d_i \) to cluster \( q \): cluster\([d_i] \leftarrow q \)

Update \( D_p, n_p, D_q, n_q \)

End if

End for

Until no moves for all \( n \) documents

End procedure

At Initialization, \( k \) arbitrary documents are selected to be the seeds from which initial partitions are formed. Refinement is a procedure that consists of a number of iterations. During each iteration, the \( n \) documents are visited one by one in a totally random order. Each document is checked and is moved to another cluster if it has more similarity. Else if the document is matched more with current cluster then it resides in the same cluster and doesn’t move. This process terminates when an iteration completes without any documents being moved to new clusters. The clustering algorithm updates immediately whenever each document is moved to new cluster.

**V. Clustering Criterion Functions**

Having defined our similarity measure, we now formulate our clustering criterion functions. The first function, called IR, is the cluster size-weighted sum of average pairwise similarities of documents in the same cluster. First, express this sum in a general form by function \( F \).

\[ F = \sum_{r=1}^k \left( \frac{1}{n_r} \frac{n+n_r}{n-n_r} ||D_r||^2 - \frac{(n+n_r/n-n_r-1)D_r D^t}{n} \right) \]

To transform this objective function into some suitable form such that it could facilitate the optimization procedure to be performed in a simple, fast and effective way.

\[ F = \sum_{r=1}^k \left( \frac{1}{n_r} \frac{n+n_r}{n-n_r} ||D_r||^2 - \frac{(n+n_r/n-n_r-1)D_r D^t}{n} \right) \]

If comparing \( F \) with the min-max cut in both functions contain the two terms \( ||D||^2 \) and \( D^t D \). The objective of min-max cut is to minimize the inverse ratio between these two terms, the aim here is to maximize their weighted difference.

\[ I_R = \sum_{r=1}^k \frac{1}{n_r} \left[ \frac{n+n_r}{n-n_r} ||D||^2 - \frac{(n+n_r/n-n_r-1)D^t D}{n} \right] \]

It appears that IR’s performance dependency on the value of \( \alpha \) is not very critical. The criterion function yields relatively good. Clustering results for \( \alpha \epsilon (0,1) \). In the formulation of IR, a cluster quality is measured by the average pairwise similarity between documents within that cluster.

**VI. Frame Work**

In the proposed system clustering with multi-view based similarity measure the modules defined are

- HTML Parser
- Cumulative Document
- Document Similarity
- Clustering

**A. HTML Parser**

In Parsing first the document enters the process state. Parsing is the process of division or identification of meta tags in a HTML document.

**B. Cumulative Document**

In cumulative document it contains sum of all the documents, containing meta-tags from all the documents. The references in the input base document are found then read other documents and find references in them and so on.

**C. Document Similarity**
the documents.
The weights in the cosine-similarity are found from the TF-IDF measure between the phrases (meta-tags) of the two documents.

D. Clustering

Fig. 3:
Clustering is a division of data into groups of similar objects. Representing the data by fewer clusters necessarily loses certain fine details, but achieves simplification.

VII. Results
‘N’ number of documents consisting of its own Phrases, Nodes and Edges.

Fig. 4:
Cumulative document is analyzed. Document similarity is analysed and find the Overlapping Rate (OLP Rate).

Fig. 5:
Formation of clusters is done. Thus the Document clustering using Incremental Clustering is done and the causes are documented.

VIII. Conclusion
In this paper we proposed a new similarity measure known as MVS (Multi-Viewpoint based similarity). When it is compared with cosine similarity, MVS is more useful for finding the similarity of text 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.

References
[1] Clustering with Multiviewpoint-Based Similarity Measure Duc Thang Nguyen, Lihui Chen, Senior Member, IEEE, and Chee Keong Chan, IEEE transactions on knowledge and data engineering, Vol. 24, No. 6, June 2012.