Multi-View Point based Similarity Measure for Hierarchical Clustering

Meena S. U.[1], P. Parthasarathi[2]

[1]PG Scholar, Sri Krishna College of Technology, Thiruvananthapuram, Kerala, India

[2]Assistant Professor, CSE Department, Sri Krishna College of Technology, Thiruvananthapuram, Kerala, India

Abstract: ***Clustering is a fundamental operation used in unsupervised document organization and information retrieval. It aims to find intrinsic structures in data, and organize them into meaningful subgroups. It groups data instances that are similar to each other in one cluster and data instances that are very different from each other into different clusters. Hierarchical clustering is used to find the cluster relationship between data objects in the data set. A novel multi-viewpoint based similarity measure and two related clustering methods are proposed. The main difference of the novel method from the existing one is that it uses only single view point for clustering and where as in Multi-Viewpoint Based Similarity Measure uses many different viewpoints, which are objects and are assumed to not be in the same cluster with two objects being measured. Using multiple viewpoints, more informative assessment of similarity could be achieved. 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. This is called as Multiviewpoint-based Similarity, or MVS. Based on this novel method two criterion functions are proposed for document clustering. We compared this clustering algorithm with other measures in order to verify the performance of multiviewpoint clustering.***

Keywords: Multiview-point clustering, Document Clustering, Hierarchical Clustering, Information retrival

1. Introduction

Clustering is a process of grouping a set of objects into classes of *similar* objects and is a most interesting concept of data mining in which it is defined as a collection of data objects that are similar to one another. Purpose of Clustering is to group fundamental structures in data and classify them into meaningful subgroup for additional analysis. Many clustering algorithms have been published every year and can be proposed for developing various techniques and approaches. The k-means algorithm has been one of the top most data mining algorithms presently used. Even though it is a top most algorithm, it has a few basic drawbacks when clusters are of differing sizes. Irrespective of the drawbacks is simplicity, understandability, and scalability is the main reasons that made the algorithm popular. K-means is fast and easy to combine with other methods in larger systems. A common approach to the clustering problem is to treat it as an optimization process. An optimal partition is found by optimizing a particular function of similarity among data. Basically, there is an implicit assumption that the true intrinsic structure of data could be correctly described by the similarity formula defined and embedded in the clustering criterion function. An algorithm with adequate performance and usability in most of application scenarios could be preferable to one with better performance in some cases but limited usage due to high complexity. While offering reasonable results, k-means is fast and easy to combine with other methods in larger systems. The original k-means has sum-of-squared-error objective function that uses Euclidean distance. In a very sparse and high-dimensional domain like text documents, spherical k-means, which uses cosine similarity (CS) instead of Euclidean distance as the measure, is deemed to be more suitable [5], [9]. The nature of similarity measure plays a very important role in the success or failure of a clustering method. Our first objective is to derive a novel method for measuring similarity between data objects in sparse and high-dimensional domain, particularly text documents. From the proposed similarity measure, we then formulate new clustering criterion functions and introduce their respective clustering algorithms, which are fast and scalable like k-means, but are also capable of providing high-quality and consistent performance.

2. Related Works

Document clustering is a fundamental operation used in unsupervised document organization, automatic topic extraction and information retrieval. The similarity-measure based method focuses on detecting the intrinsic structure between nearby documents rather than on detecting the intrinsic structure between widely separated documents. Since the intrinsic semantic structure of the document space is often embedded in the similarities between the documents. Each document in a corpus corresponds to an m-dimensional vector d, where m is the total number of terms that the document corpus has. Document vectors are often subjected to some weighting schemes, such as the standard Term Frequency-Inverse Document Frequency (TF-IDF), and normalized to have unit length. Clustering is to arrange data objects into separate clusters such that the intra cluster similarity as well as the inter cluster dissimilarity is maximized. The clustering approaches that do not employ any specific form of measurement based on probabilistic model method, non-negative matrix factorization. The primarily focus on methods is to utilize the specific measures. In the literature, Euclidean distance is one of the most popular measures

$$Dist(d_i, d_j) = \|d_i - d_j\| \qquad (1)$$

It is used in the traditional k-means algorithm. The objective of k-means is to minimize the Euclidean distance between objects of a cluster and that cluster's centroid

$$min \sum_{r=1}^{k} \sum_{d_i \in S_r} \|d_i - C_r\|^2 \qquad (2)$$

The similarity of two document vectors is defined as the cosine of angle between them.

$$d_i \ and \ d_j, Sim(d_i, d_j)$$

$$Sim(d_i, d_j) = \cos(d_i, d_j) = d_i^t d_j \qquad (3)$$

Cosine measure is used in a variant of k-means called spherical k-means [5]. While k-means aims to minimize euclidean distance, spherical k-means intends to maximize the cosine similarity between documents in a cluster and that cluster's centroid

$$max \sum_{r=1}^{k} \sum_{d_i \in S_r} \frac{d_i^t C_r}{\|C_r\|} \qquad (4)$$

There are many other graph partitioning methods with different cutting strategies and criterion functions, such as Average Weight [4] and Normalized Cut [7], all of which have been successfully applied for document clustering using cosine as the pair wise similarity score [6], [12]. In [13], an empirical study was conducted to compare a variety of criterion functions for document clustering. Ahmadand Dey [1] proposed a method to compute distance between two categorical values of an attribute based on their relationship with all other attributes. Subsequently, Ienco [2] introduced a similar context-based distance learning method for categorical data. However, for a given attribute, they only selected a relevant subset of attributes from the whole attribute set to use as the context for calculating distance between its two values.

Document partitioning methods decompose a document corpus into a given number of disjoint clusters which are optimal in terms of some predefined criteria functions. Partitioning methods can also generate a hierarchical structure of the document corpus by iteratively partitioning a large cluster into smaller clusters. Typical methods in this category include K- Means clustering [8], probabilistic clustering using the Naive Bayes or Gaussian mixture model. K-Means produces a cluster set that minimizes the sum of squared errors between the documents and the cluster centers, while both the Naive Bayes and the Gaussian mixture models assign each document to the cluster that provides the maximum likelihood probability. The common drawback associated with these methods is that they all make harsh simplifying assumptions on the distribution of the document corpus to be clustered. K-Means assumes that each cluster in the document corpus has a compact shape, the Naive Bayes model assumes that all the dimensions of the feature space representing the document corpus are independent of each other, and the Gaussian mixture model assumes that the density of each cluster can be approximated by a Gaussian distribution. Document clustering using the latent semantic indexing method (LSI) [10], This method basically projects each document into the singular vector space through the SVD, and then conducts document clustering using traditional data clustering algorithms (such as K-means) in the transformed space. Although it was claimed that each dimension of the singular vector space captures a base latent semantics of the document corpus, and that each document is jointly indexed by the base latent semantics in this space, negative values in some of the dimensions generated by the SVD.

Chim and Deng [3] proposed a phrase based document similarity by combining suffix tree model and vector space model. They then used Hierarchical Agglomerative Clustering algorithm to perform the clustering task. However, a drawback of this approach is the high computational complexity due to the needs of building the suffix tree and calculating pair wise similarities explicitly before clustering. There are also measures designed specifically for capturing structural similarity among XML documents [11]. They are essentially different from the document-content measures that are discussed in this paper.

3. Proposed Methodology

Data Preprocessing

In this module the preprocessing of database is done. Preprocessing is the phase to remove stop words, stemming and identification of unique words in document. Identification of unique words in the document is necessary for clustering of document with similarity measure. And after that we remove the stop words that is the non informative word for example the, end, have, more etc. We need to eliminate those stop words for finding such similarity between documents.

A stemming algorithm is a process of linguistic normalization, in which the variant forms of a word are reduced to a common form, for example,

- Removal of suffix to generate word stem
- Grouping words
- Increase the relevance

Example: connection, connections, connective ---> connect (root word)

Suffix-Stripping algorithm

Suffix stripping algorithms do not rely on a lookup table that consists of inflected forms and root form relations. Instead, a typically smaller list of "rules" is stored which provides a path for the algorithm, given an input word form, to find its root form. Some examples of the rules include:

- if the word ends in 'ed', remove the 'ed'
- if the word ends in 'ing', remove the 'ing'
- if the word ends in 'ly', remove the 'ly'

Suffix stripping approaches enjoy the benefit of being much simpler to maintain than brute force algorithms, assuming the maintainer is sufficiently knowledgeable in the challenges of linguistics and morphology and encoding suffix stripping rules. Finally term weighting is to provide the information retrieval and text categorization. In document clustering groups together conceptually related documents. It also provides metadata characterization the content of given document cluster.

Multi view point Based Similarity measure calculation (MVS)

The cosine similarity in (3) can be expressed in the following form without changing its meaning

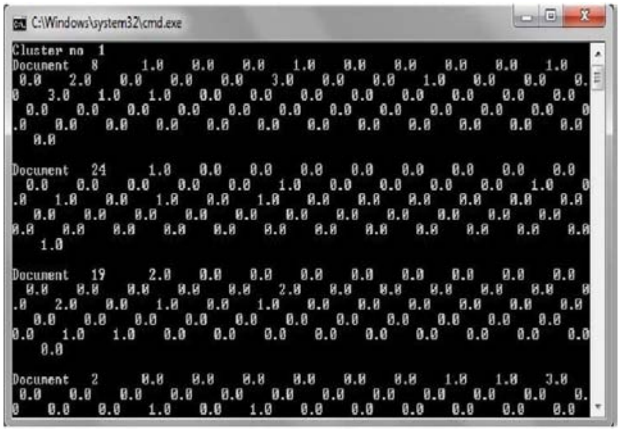
$$Sim(d_i,d_j) = \cos(d_i - 0, d_j - 0) = (d_i - 0)^t (d_j - 0)$$

where 0 is vector 0 that represents the origin point. According to this formula, the measure takes 0 as one and only reference point. The similarity between two

documents d_i and d_j is determined w.r.t. the angle

between the two points when looking from the origin. To construct a new concept of similarity, it is possible to use more than I just one point of reference. We may have a more accurate assessment of how close or distant pair of points is. if we look at them from many different viewpoints. By using the algorithm for MVS similarity matrix is formulated.

Multi view point Based Clustering by the criterion functions IR and IV



```
Cluster no  1
Document   8      1.0    0.0    0.0    1.0    0.0    0.0    0.0    0.0    1.0
 0.0    2.0    0.0    0.0    0.0    3.0    0.0    0.0    1.0    0.0    0.0    0.
0    3.0    1.0    1.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0
  0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0
.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0
   0.0

Document   24     1.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0
  0.0    0.0    0.0    0.0    0.0    1.0    0.0    0.0    0.0    0.0    1.0    0
.0    1.0    0.0    1.0    0.0    1.0    0.0    0.0    0.0    0.0    0.0    0.0
   0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0
0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0
    1.0

Document   19     2.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0
  0.0    0.0    0.0    0.0    0.0    2.0    0.0    0.0    0.0    0.0    0.0    0
.0    2.0    0.0    1.0    0.0    1.0    0.0    0.0    0.0    0.0    0.0    0.0
   0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0
0.0    1.0    1.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0
    0.0

Document   2      0.0    0.0    0.0    0.0    0.0    0.0    1.0    1.0    3.0
 0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.
0    0.0    0.0    1.0    0.0    1.0    0.0    0.0    0.0    0.0    0.0    0.0
```

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

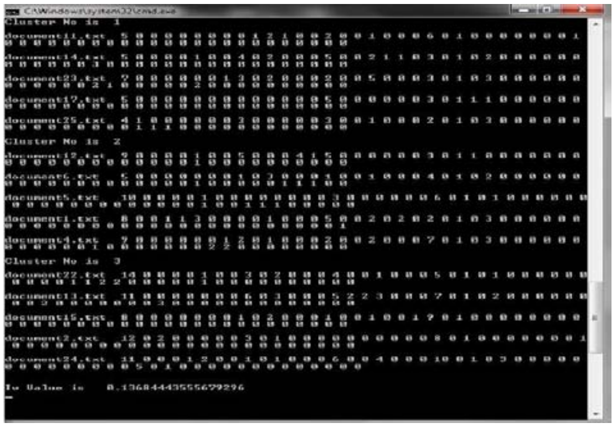
$$I_V = \sum_{r=1}^{k} \left[\frac{n+\|D_r\|}{n-n_r} \|D_r\| - \left(\frac{n+\|D_r\|}{n-n_r} - 1\right) \frac{D_r^t D}{\|D_r\|} \right]$$

The first function, called IR, is the cluster size-weighted sum of average pair wise similarities of documents in the same cluster.

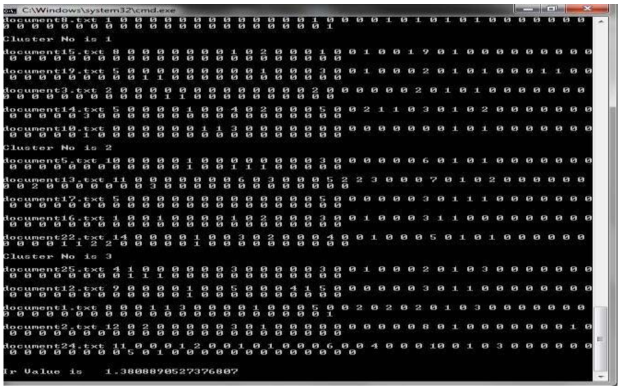
In the formulation of IR, a cluster quality is measured by the average (5) pair wise similarity between documents within that cluster.
IV calculates the weighted difference between the two terms: $\|D_r\|$ and $\frac{D_r^t D}{\|D_r\|}$ which again represent an intra cluster similarity measure and an inter cluster similarity measure, respectively. The first term is actually equivalent to an element of the sum in spherical k-means objective function in

(4) The second one is similar to an element of the sum in min-max cut criterion in (6).By using greedy algorithm optimization of functions are performed and clustering is performed.

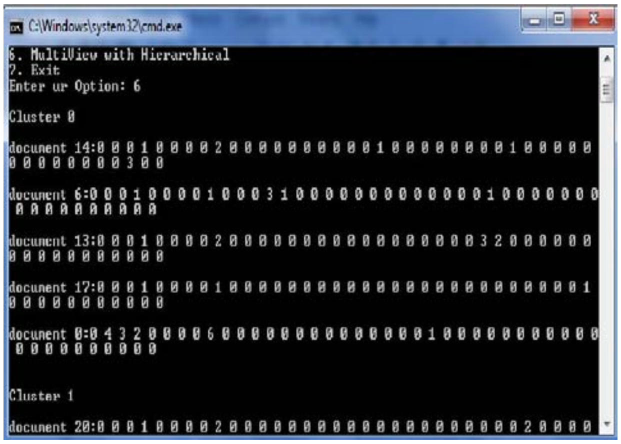
4. Result and Discussion



```
Cluster No is  1
document11.txt  5 0 0 0 0 0 0 0 0 1 2 1 0 0 2 0 0 1 0 0 0 6 0 1 0 0 0 0 0 0 0 1
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
document14.txt  5 0 0 0 0 1 0 0 4 0 2 0 0 0 5 0 0 2 1 1 0 3 0 1 0 2 0 0 0 0 0 0
0 0 0 0 0 0 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
document23.txt  7 0 0 0 0 0 0 1 3 0 2 0 0 0 2 0 0 5 0 0 0 3 0 1 0 3 0 0 0 0 0 0
0 0 0 0 0 0 2 1 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0
document17.txt  5 0 0 0 0 0 0 0 0 0 0 0 0 0 5 0 0 0 0 0 0 3 0 1 1 1 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
document25.txt  4 1 0 0 0 0 0 0 3 0 0 0 0 0 3 0 0 1 0 0 0 2 0 1 0 3 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0
Cluster No is  2
document12.txt  9 0 0 0 0 1 0 0 5 0 0 0 4 1 5 0 0 0 0 0 0 3 0 1 1 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0
document6.txt  5 0 0 0 0 0 0 0 1 0 3 0 0 0 1 0 0 1 0 0 0 4 0 1 0 2 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 1 1 0 0
document5.txt  10 0 0 0 0 1 0 0 0 0 0 0 0 0 3 0 0 0 0 0 0 6 0 1 0 1 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 1 1 0 0 0 0 0
document1.txt  8 0 0 1 1 3 0 0 0 0 1 0 0 0 5 0 0 2 0 2 0 2 0 1 0 3 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
document4.txt  7 0 0 0 0 0 0 1 2 0 1 0 0 0 2 0 0 2 0 0 0 7 0 1 0 3 0 0 0 0 0 0
0 0 0 0 0 0 1 0 0 0 0 0 0 0 2 2 0 0 0 0 0 0 0 0
Cluster No is  3
document22.txt  14 0 0 0 0 1 0 0 3 0 2 0 0 0 4 0 0 1 0 0 0 5 0 1 0 1 0 0 0 0 0 0
 0 0 0 0 1 1 2 2 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0
document13.txt  11 0 0 0 0 0 0 0 6 0 3 0 0 0 5 2 2 3 0 0 0 7 0 1 0 2 0 0 0 0 0 0
 0 0 2 0 0 0 0 0 0 0 3 0 0 0 0 0 0 0 0 0 0 0 0 0
document15.txt  8 0 0 0 0 0 0 0 1 0 2 0 0 0 1 0 0 1 0 0 1 9 0 1 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
document2.txt  12 0 2 0 0 0 0 0 3 0 1 0 0 0 0 0 0 0 0 0 0 8 0 1 0 0 0 0 0 0 0 1
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
document24.txt  11 0 0 0 1 2 0 0 1 0 1 0 0 0 6 0 0 4 0 0 0 10 0 1 0 3 0 0 0 0 0
0 0 0 0 0 0 0 0 0 5 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0

Iv Value is   0.13684443555679296
_
```

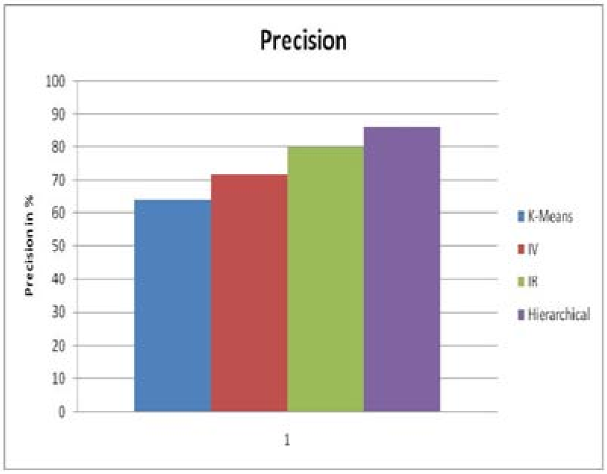


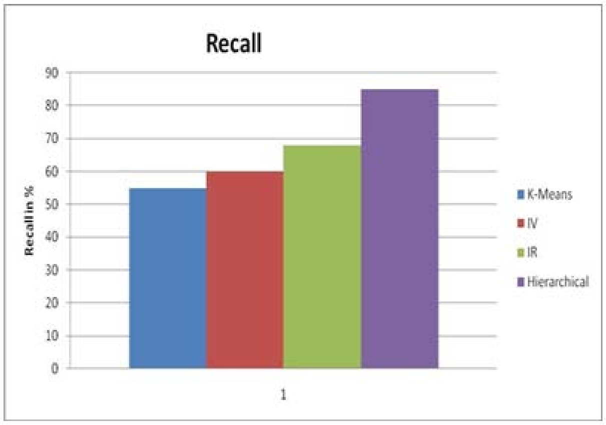
```
document8.txt 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 1 0 1 0 1 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
Cluster No is 1
document15.txt 8 0 0 0 0 0 0 0 1 0 2 0 0 0 1 0 0 1 0 0 1 9 0 1 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
document19.txt 5 0 0 0 0 0 0 0 0 0 1 0 0 0 3 0 0 1 0 0 0 2 0 1 0 1 0 0 0 1 1 0 0
 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0
document3.txt 2 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 2 0 1 0 1 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0
document14.txt 5 0 0 0 0 1 0 0 4 0 2 0 0 0 5 0 0 2 1 1 0 3 0 1 0 2 0 0 0 0 0 0 0
 0 0 0 0 0 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
document10.txt 0 0 0 0 0 0 1 1 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0
 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Cluster No is 2
document5.txt 10 0 0 0 0 1 0 0 0 0 0 0 0 0 3 0 0 0 0 0 0 6 0 1 0 1 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 1 1 0 0 0 0 0
document13.txt 11 0 0 0 0 0 0 0 6 0 3 0 0 0 5 2 2 3 0 0 0 7 0 1 0 2 0 0 0 0 0 0
0 0 2 0 0 0 0 0 0 0 3 0 0 0 0 0 0 0 0 0 0 0 0 0
document17.txt 5 0 0 0 0 0 0 0 0 0 0 0 0 0 5 0 0 0 0 0 0 3 0 1 1 1 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
document16.txt 1 0 0 1 0 0 0 0 1 0 2 0 0 0 3 0 0 1 0 0 0 3 1 1 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
document22.txt 14 0 0 0 0 1 0 0 3 0 2 0 0 0 4 0 0 1 0 0 0 5 0 1 0 1 0 0 0 0 0 0
0 0 0 0 1 1 2 2 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0
Cluster No is 3
document25.txt 4 1 0 0 0 0 0 0 3 0 0 0 0 0 3 0 0 1 0 0 0 2 0 1 0 3 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0
document12.txt 9 0 0 0 0 1 0 0 5 0 0 0 4 1 5 0 0 0 0 0 0 3 0 1 1 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0
document1.txt 8 0 0 1 1 3 0 0 0 0 1 0 0 0 5 0 0 2 0 2 0 2 0 1 0 3 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
document2.txt 12 0 2 0 0 0 0 0 3 0 1 0 0 0 0 0 0 0 0 0 0 8 0 1 0 0 0 0 0 0 0 1 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
document24.txt 11 0 0 0 1 2 0 0 1 0 1 0 0 0 6 0 0 4 0 0 0 10 0 1 0 3 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 5 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0

Ir Value is   1.3808890527376807
```

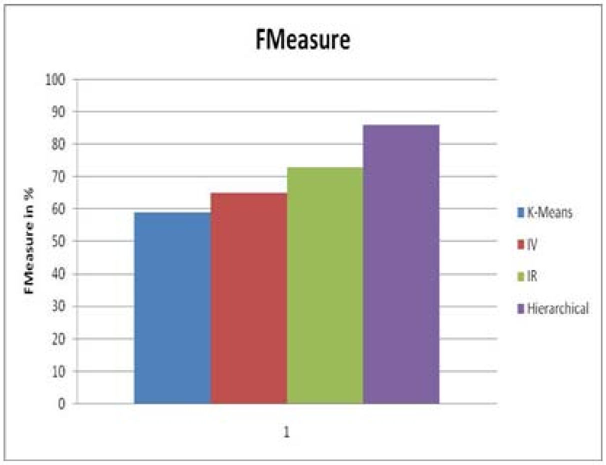


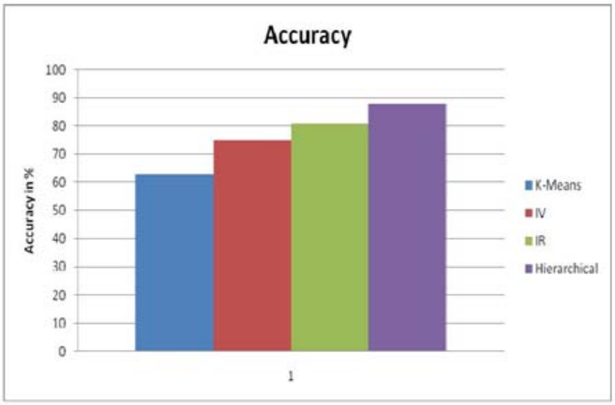
```
6. MultiView with Hierarchical
7. Exit
Enter ur Option: 6

Cluster 0

document 14:0 0 0 1 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0
0 0 0 0 0 0 0 0 3 0 0

document 6:0 0 0 1 0 0 0 0 1 0 0 0 3 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0

document 13:0 0 0 1 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 3 2 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0

document 17:0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
0 0 0 0 0 0 0 0 0 0 0

document 0:0 4 3 2 0 0 0 0 6 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0


Cluster 1

document 20:0 0 0 1 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0
```

5. Performance Analysis









6. Conclusion

In this paper propose a Multi view point-based Similarity measuring method, named MVS. The Theoretical analysis and empirical examples represents that MVS is likely more supportive for documents than the famous cosine similarity. Two criterion functions, IR and IV and the corresponding clustering algorithms MVSC-IR and MVSC-IV have been introduced in this paper. The proposed algorithms MVSC-IR and MVSC-IV shows that they could afford significantly advanced clustering execution ,when compared with other state-of-the-art clustering methods that use distinctive methods of similarity measure on a very large number of document data sets concealed by various assessment metrics. The main aspect of our paper is to introduce the basic concept of similarity measure from multiple viewpoints. Further the proposed criterion functions for hierarchical clustering algorithms would also be achievable for applications .At last we have shown the application of MVS and its clustering algorithms for text data.

References

[1] A. Ahmad and L. Dey, “A Method to Compute Distance Between Two Categorical Values of Same Attribute in Unsupervised Learning for Categorical Data Set,” Pattern Recognition Letters, vol. 28, no. 1, pp. 110-118, 2007

[2] D. Ienco, R.G. Pensa, and R. Meo, "Context-Based Distance Learning for Categorical Data Clustering," Proc. Eighth Int'l Symp. Intelligent Data Analysis (IDA), pp. 83-94, 2009.

[3] H. Chim and X. Deng, "Efficient Phrase-Based Document Similarity for Clustering," IEEE Trans. Knowledge and Data Eng.,vol. 20, no. 9, pp. 1217-1229, Sept. 2008.

[4] H. Zha, X. He, C. Ding, H. Simon, and M. Gu, "Spectral Relaxation for K-Means Clustering," Proc. Neural Info. Processing Systems (NIPS), pp. 1057-1064, 2001

[5] I. 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

[6] I.S. Dhillon, "Co-Clustering Documents and Words Using Bipartite Spectral Graph Partitioning," Proc. Seventh ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 269-274, 2001

[7] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation," IEEE Trans. Pattern Analysis Machine Intelligence, vol. 22, no. 8, pp. 888-905, Aug. 2000

[8] P. Willett. Document clustering using an inverted file approach. Journal of Information Science, 2:223–231, 1990

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

[10] S.C. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, and R.A. Harshman, "Indexing by Latent Semantic Analysis," J. Am. Soc. Information Science, vol. 41, no. 6, pp. 391-407, 1990

[11] S. Flesca, G. Manco, E. Masciari, L. Pontieri, and A. Pugliese, "Fast Detection of xml Structural Similarity," IEEE Trans. Knowledge and Data Eng., vol. 17, no. 2, pp. 160-175, Feb. 2005

[12] Y. Gong and W. Xu, Machine Learning for Multimedia Content Analysis. Springer-Verlag, 2007

[13] Y. Zhao and G. Karypis, "Empirical and Theoretical Comparisons of Selected Criterion Functions for Document Clustering," Machine Learning, vol. 55, no. 3, pp. 311-331, June 2004