ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 2, Issue 6, June 2013

A Modified Hierarchical Clustering Algorithm for Document Clustering

Merin Paul, P Thangam

Abstract— **Clustering is the division of data into groups called as clusters. Document clustering is done to analyse the large number of documents distributed over various sites. The similar documents are grouped together to form a cluster. The success or failure of a clustering method depends on the nature of similarity measure used. The multiviewpoint-based similarity measure or MVS uses different viewpoints unlike the traditional similarity measures that use only a single viewpoint. This increases the accuracy of clustering. A hierarchical clustering algorithm creates a hierarchical tree of the given set of data objects. Depending on the decomposition approach, hierarchical algorithms are classified as agglomerative (merging) or divisive (splitting). This paper focuses on applying multiviewpoint-based similarity measure on hierarchical clustering.**

Index Terms—**Document Clustering, Hierarchical Clustering, Similarity Measure.**

I. INTRODUCTION

Clustering is the process of organizing objects into groups whose members are similar in some way. Thus a cluster is a collection of objects which are “similar” to each other and are “dissimilar” to the objects that are in other clusters. Clustering [1] is ultimately a process of reducing a mountain of data to manageable piles. For cognitive and computational simplification, these piles may consist of "similar" items.

A. Document Clustering

Document clustering has become an increasingly important task in analysing huge numbers of documents distributed among various sites. The important feature is to organize the documents in such a way that it results in better search without having much cost and complexity. The Cluster Hypothesis [2] is fundamental to the issue of improved effectiveness which states that relevant documents are more similar to each other than non-relevant documents and thus tend to appear in the same clusters. In a clustered collection, this relevant document may be clustered together with other relevant items that have the required query terms and could therefore be retrieved through a clustered search. Document clustering offers an alternative file organization to that of best-match retrieval and it has the potential to address this issue, thereby increase the effectiveness of an IR system.

There are two approaches to document clustering, particularly in information retrieval, they are known as term and item clustering. Term clustering is a method, which groups redundant terms. The grouping reduces noise and increase frequency of assignment. The dimension is also reduced if there are fewer clusters the original terms. But the semantic properties will be affected.

There are many different algorithms available for term clustering. These are cliques, stars, single link and connected components. Cliques need all items in a cluster to be within the threshold of all other items. In single link clustering the strong constraint that every term in a class is similar to every other term is relaxed. The star technique selects a term and then places in the class all terms that is related to that term. Terms not yet in classes are selected as new seeds until all terms are assigned to a class. There are many different classes that can be created using the star technique. Item clustering helps the user in identifying relevant items**.**

When items in the database have been clustered, it is possible to retrieve all of the items in a cluster, even though the search statement does not identify them. When the user retrieves a strongly significant item, the user can look at other items like it without issuing another search. When significant items are used to create a new query, the retrieved hits are similar to what might be produced by a clustering algorithm. However, term clustering and item clustering in a sense achieve the same objective even though they are the inverse of each other. For all of the terms within the same cluster, there will be significant overlap of the set of items they are found in. Item clustering is based upon the same terms being found in the other items in the cluster. Thus the set of items that caused a term clustering has a strong possibility of being in the same item cluster based upon the terms.

B. Similarity Measures

The set of terms shared between a pair of documents is typically used as an indication of the similarity of the pair. The nature of similarity measure plays a very important role in the success or failure of a clustering method.

Text document clustering groups similar documents to form a cluster, while documents that are different have separated apart into different clusters. Accurate clustering requires a precise definition of the closeness between a pair

Manuscript received June, 2013.

Merin Paul*, PG Scholar, Computer Science and Engineering, Coimbatore Institute of Engineering and Technology. Narasipuram, Coimbatore, Tamil Nadu,,India, 8129898069*

P Thangam*, Assistant Professor, Computer Science and Engineering, Coimbatore Institute of Engineering and Technology, Narasipuram, Coimbatore, Tamil Nadu,,India, 8098099829.,*

of objects, in terms of either the pair wise similarity or distance. Five measures are discussed and tested in [3].

Euclidean distance is the default distance measure used with the K-means [4] algorithm. Cosine Similarity is quantified as the cosine of the angle between vectors. An important property of the cosine similarity is its independence of document length. For text document, Jaccard Coefficient compares the sum weight of shared terms to the sum weight of terms that are present in either of the two documents but are not the shared terms. The Jaccard coefficient and Pearson Correlation Coefficient are other similarity measures.

C. Hierarchical Clustering

A hierarchical clustering algorithm creates a hierarchical decomposition of the given set of data objects. Depending on the decomposition approach, hierarchical algorithms are classified as agglomerative (merging) or divisive (splitting). Agglomerative algorithms are more widely used in practice. Thus the similarities between clusters are more researched.

II. RELATED WORK

Text document clustering groups similar documents to form a cluster, while documents that are different have separated apart into different clusters. Accurate clustering demands an exact definition of the closeness between a pair of objects, in terms of distance or the pair wise similarity. In general, similarity/distance measures map the distance or similarity between the symbolic description of two objects into a single numeric value, that depends on the properties of the two objects and the measure itself. Five measures are discussed and tested in [1].

For high-dimensional data such as text documents (represented as TF-IDF vectors) and market baskets, cosine similarity has been shown to be a superior measure to Euclidean distance. The efficient online spherical k-means clustering [7] focus mainly on achieving non-empty balanced clusters rather than efficiency or quality. Different learning rate schedules are used. The online update of cluster centroids can be viewed as a gradient ascent approach—the cluster centroids (parameters) are updated following the gradient direction. The learning rate used is effectively inversely proportional to the size of a cluster, aiming to balance clusters.

To achieve a more accurate document clustering, a more useful feature term, phrase, has been considered in recent research work and literature [3]. A phrase of a document is an ordered sequence of one or more words. Bigrams and trigrams are commonly used methods to extract and identify meaningful phrases in statistical natural language processing. The quality of clustering achieved based on this model significantly surpassed the traditional VSD model-based approaches in the experiments of clustering Web documents.

The quality of the clustering results is higher than the results of traditional single-word tf-idf similarity measure in the same HAC algorithm, mainly in large document data sets.

The structure of the clusters produced by the spherical *k*-means algorithm when applied to text data sets with the aim of gaining novel insights into the distribution of sparse text data in high-dimensional spaces is studied in [5].

Clustering algorithms, and recently a new wave of excitement has spread across the machine learning community mainly because of the important development of spectral methods [10]. There is also growing interest around fundamental questions regarding the very nature of the clustering problem. Yet, despite the tremendous progress in the field, the clustering problem remains vague and a satisfactory solution even to the most basic problems is still to come.

The partitional approach is attractive as it leads to elegant mathematical and algorithmic treatments and allows us to employ powerful ideas from many sophisticated fields like linear algebra, optimization, graph theory, statistics and information theory. Yet, there are several reasons for feeling uncomfortable with this oversimplified formulation.

The best limitation of the partitional approach is the requirement that the number of clusters be known in advance. The game-theoretic perspective [6] has the following advantages. It makes no assumption on the underlying (individual) data representation like spectral clustering. It does not require that the elements to be clustered be represented as points in a vector space.

III. CLUSTERING WITH MULTIVIEWPOINT-BASED SIMILARITY

The proposed document clustering in this paper uses multiviewpoint-based similarity measure or MVS on hierarchical agglomerative clustering.

A. Overview of Multiviewpoint-based Similarity

The multiviewpoint-based similarity measure or MVS [9] uses different viewpoints unlike the traditional similarity measures that use only a single viewpoint.

MVS uses more than one point of reference. It provides more accurate assessment of how close or distant a pair of points are, if we look at them from different viewpoints. From a third point d_h, the directions and distances to d_i and d_j are indicated by the difference vectors $(d_i - d_h)$ and $(d_j - d_h)$ respectively. By standing at various reference points d_h to view d_i, d_j and finding their difference vectors, similarity between the two documents are defined as

$$Sim(d_i, d_j) = \frac{1}{n - n_r} \sum_{d_h \in S \setminus S'} Sim(d_i - d_h, d_j - d_h), \quad (1)$$

B. Cosine Similarity Identification

The cosine similarity in can be expressed in the following form without changing its meaning:

$$Sim(d_i, d_j) = cos(d_i, d_j) = (d_i - 0)^t (d_j - 0), \quad (2)$$

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 [8].

C. Multiviewpoint-based Similarity Identification

Similarity of two documents d_i and d_j -given that they are in the same cluster-is equal to the product of the cosine of the angle between d_i and d_j looking from d_h and the euclidean distances from d_h to d_i and d_j.

$$MVS(d_i, d_j) = \frac{1}{n - n_r} \sum cos(d_i - d_h, d_j - d_h) \|d_i - d_h\| \|d_j - d_h\|$$

(3)

The two objects to be measured, d_i and d_j , must be in the same cluster, while the points from where to establish this measurement, d_h must be outside of the cluster.

D. Constrained Kmeans Clustering

The traditional kmeans algorithm is used for clustering. I_R and I_V are used to determine the quality of the clusters formed [9].

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

(4)

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

(5)

D denotes the composite vector of all the documents and D_r denotes composite vector of cluster r. The value of α ranges from 0 to 1.

E. Constrained Hierarchical Clustering

Hierarchical agglomerative clustering is used to form clusters. In an agglomerative method, each object forms a cluster. Then the two most similar clusters are merged iteratively until some termination criterion is satisfied. The cluster merging process is repeated until all the objects are merged to form one cluster.

After the clusters are formed, the quality of clusters thus generated is determined using I_R and I_V, which are given in (4) and (5).

Fig 1 shows the system flow diagram of the proposed work.

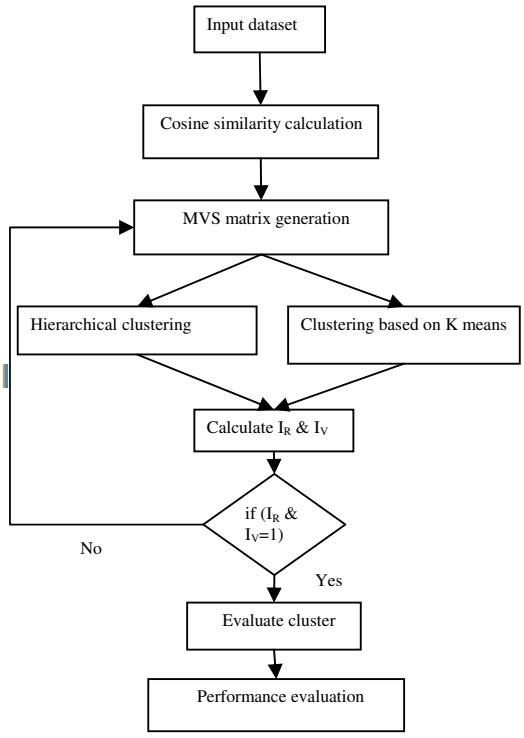

Fig. 1 System Flow Diagram

IV. ANALYSIS OF DOCUMENT CLUSTERING METHODS

To verify the advantages of the proposed work, their performance have to be evaluated. The objective of this section is to compare hierarchical clustering with kmeans clustering after applying MVS on both of the methods.

The analysis of the two document clustering methods are done using two data sets each containing twenty documents. The data sets 'autos' and 'motorcycles' are used to analyse the document clustering methods.

'Autos' and 'motorcycle' are available with the 20 newsgroup. 'Autos' consist of 2344 newspaper articles, among which 20 are taken for this experiment. 'Motorcycle' consist of 2344 newspaper articles, among which 20 are taken for this experiment.

Fig. 2 shows the sample screen shot of the menu page for constrained kmeans clustering.

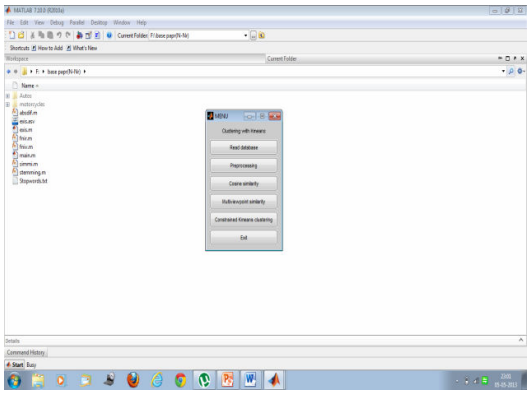

Fig 2 Menu Page for Kmeans

Fig. 3 shows the sample screen shot of the output of multiviewpoint based similarity identification.

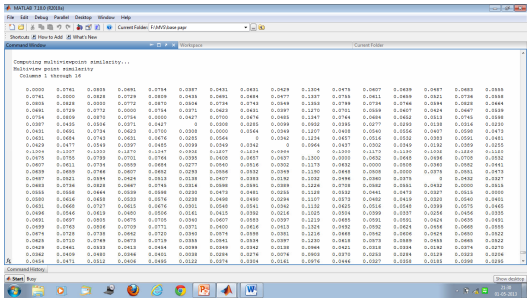

Fig. 3 Multiviewpoint based Similarity Identification

Fig. 4 shows the sample screen shot of the menu page for constrained hierarchical clustering.

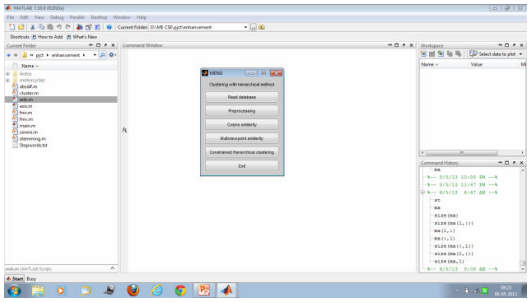

Fig. 4 Menu Page of Hierarchical Clustering

Accuracy measures the fraction of documents that are correctly labels. Table 1 lists accuracy of kmeans clustering and hierarchical clustering on applying MVS.

Table 1 Comparison of Accuracy

Methods	Accuracy %
Kmeans Clustering	85
Hierarchical Clustering	93

Table 2 Comparison of Precision

Method	Data Set Autos	Data Set Motorcycle
Kmeans Clustering	0.83	0.89
Hierarchical Clustering	0.9	0.93

Table 3 Comparison of Recall

Method	DataSet Autos	Data Set Motorcycle
Kmeans Clustering	0.8	0.9
Hierarchical Clustering	0.9	0.95

V. RESULTS AND DISCUSSION

Fig. 5 shows the clustering results based on the parameter accuracy. It is found that accuracy of hierarchical clustering with MVS is higher than kmeans clustering MVS.

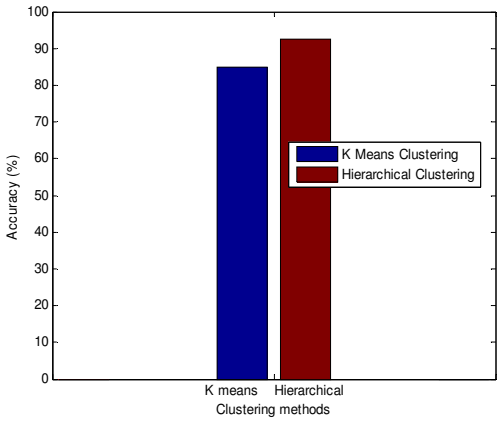

Fig 5 Clustering Results in accuracy

Fig. 6 and Fig. 7 shows the clustering results of kmeans and hierarchical clustering based on the parameter Precision and Recall respectively.

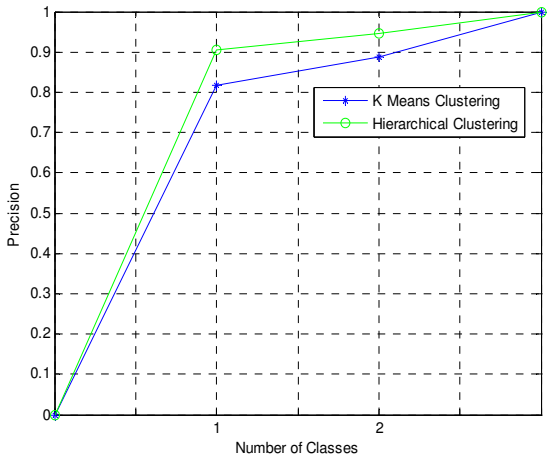

Fig. 6 Clustering Results on Precision

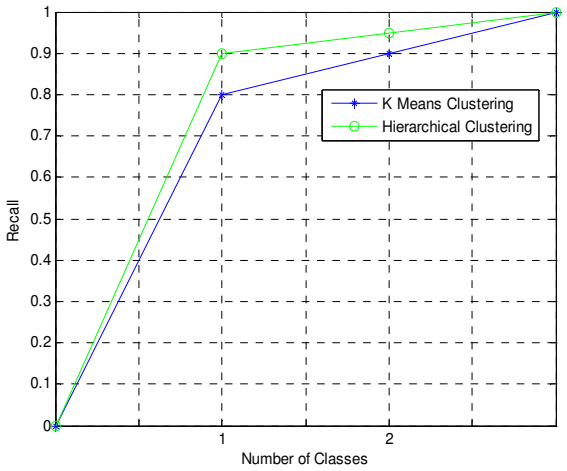

Fig. 7 Clustering Results on Recall

VI. CONCLUSION

In this paper a new hierarchical clustering method using MVS is proposed for document clustering. MVS uses multiple points of reference. The similarity between two documents in a cluster is calculated with respect to another document that is outside the cluster. This increases the accuracy of clustering and thus the quality of the clusters formed increases. Finally the proposed work is compared with kmeans clustering using MVS and is found that performance of hierarchical clustering is higher.

REFERENCES

[1] M. Pelillo, "What Is a Cluster? Perspectives from Game Theory,"*Proc. NIPS Workshop Clustering Theory,* 2009.

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

[3] Anna Huang, Department of Computer Science The University of Waikato, Hamilton, New Zealand "Similarity Measures for Text Document Clustering",2005.

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

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

[6] G. Karypis, "CLUTO a Clustering Toolkit," technical report, Department of Computer Science, Univ.ofMinnesota,http://glaros.dtc.umn.edu/~gkhome/views/cluto, 2003.

[7] Chim H 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.

[8] Zha H, C. Ding, X. He, M. Gu, and H. Simon, "A Min-Max Cut Algorithm for Graph Partitioning and Data Clustering," *Proc.IEEE Int'l Conf. Data Mining (ICDM*), pp. 107-114, 2001.

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

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

Merin Paul is currently pursuing M.E Computer Science and Engineering at Coimbatore Institute of Engineering and Technology, Coimbatore, Tamil Nadu, (Anna University, Chennai). She completed her B.Tech in Information Technology from MES College of Engineering, Kuttipuram, Kerala, (University of Calicut) in 2010. Her research interests include Data Mining and Testing.

Ms.P.Thangam received her B.E Degree in Computer Hardware and software Engineering from Avinashilingam University, Coimbatore in 2001. She has received her M.E degree in Computer Science and Engineering from Government College of Technology, Coimbatore in 2007. She is currently doing her PhD in the area of Medical Image Processing under Anna University, Chennai. Presently she is working as an Assistant Professor in the Department of Computer Science and Engineering at Coimbatore Institute of Engineering and Technology, Coimbatore. Her research interests are in Image Processing, Medical Image Analysis, Data Mining, Classification and Pattern Recognition.