AN APPROACH TOWARDS MULTI VIEW BASED DOCUMENT CLUSTERING

R.Goutham[1], M.Srinivas[2]
[1]MCA Student, Dept of MCA, DRK Institute of Science and Technology, Hyderabad, Andhra Pradesh, India

[2]Assistant Professor, Dept of CSE, DRK Institute of Science and Technology, Hyderabad, Andhra Pradesh, India

ABSTRACT:

Cluster analysis is a learning technique aiming to gather instances into different subsets, or clusters, such that each subset encloses the similar observations with reverence to some predefined measures. Determining the fundamental structures in data and organize them into significant subgroups for research studies and assessment is the main proposal of clustering. Clustering procedure which is exclusively designed for automatic topic extraction or unsupervised document association is known as document clustering. The substances that are functional have all clustering methods to imagine some cluster relationship among the data objects. Similarity of two documents in the related cluster is based on the general resemblance measured comparatively from the explanation of other documents peripheral to the cluster. In recent times, multi-view clustering methods have been proposed that has been shown to improve over conventional single-view clustering. The basic idea is to control every view individuality in order to do enhanced views than simply concatenating views. Multi viewpoint-based Similarity method is probably making use of additional points than one point of reference and can have more accurate estimation of how close or distant pair of points are if they are viewed from many viewpoints. It is possible to make use of more than one point of indication for creating new concept of identity. The closeness and the separation distances of the pair of the points can be perfectly estimated when they are glimpsed from dissimilar view points.

Keywords: Cluster analysis, Multi-View Clustering, Single-View Clustering, Document Clustering, View points, Clustering procedure.

1. INTRODUCTION

The practice of inspecting a collection of points, combining them into clusters based on some designed distance is known as clustering. for various fields such as statistics, machine learning, data mining, image analysis and information retrieval, clustering has been an intensive subject. Most clustering methods fall into three categories such as similarity based clustering methods which, builds k partitions of the information where each cluster maximizes a clustering approach on the basis of a similarity measurelike minimization of the summation of squared distance from the mean contained by each cluster [1] [2]. Projection based approach makes general use of the set of the Eigen values of a similarity matrix of information to perform dimensionality reduction for clustering in fewer dimensions and Density based clustering methods which can estimate the probability distributions of clusters and then allocate each instance to its most probable cluster [3]. Clustering procedure which is exclusively designed for automatic topic mining or unsupervised document organization also called document clustering. In Similarity-based clustering algorithms the separation of the data is based on how similar instances are that is the more similar, the more possible that they belong to the same cluster. This combination function is based on similarities among instances according to a given calculated distance Similarity-based clustering techniques therefore means at identifying clusters that make the most of the intercluster distance and minimize the intracluster distance, so that distinct groups of similar entities are obtained [4] [5] [6]. The most common measures is the Euclidean distance where two data points are understood to be vectors in a given vectorial space and their distance indicates the length of the straight line which is connecting them [9] [10] [11]. Cosine-measure conveys the similarity between two objects with reverence to the angle among the relative directions of their corresponding vectors as a substitute of their distance [12] [14] is shown in fig 1.

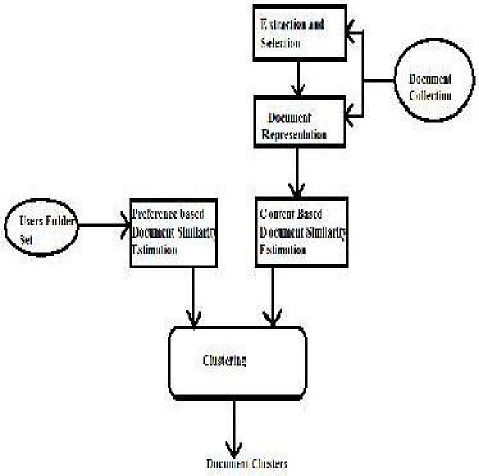

Fig 1. Document Clustering Method

2. PARTITIONAL CLUSTERING ALGORITHM:

A partitioned clustering algorithm achieves a single partition of the data as a substitute of a clustering structure. Partitional method has advantage in functions which involve large data sets for which the structure of a dendrogram is computationally excessive [13]. A difficulty associated with the use of a partitional algorithm is the selection of number of desired output clusters. The partition method normally constructs clusters by maximizing a function which is defined either locally or globally. The algorithm naturally runs multiple times with different initial states, and the best configuration attained from all of the runs is used as the output clustering [14]. The most instinctive and regularly used function in partitioned clustering method is the squared error approach, which have a tendency to work well with isolated and compact clusters. For a clustering, the squared error of a collection contains K clusters. The kmeans is the simplest and most commonly used algorithm which employs squared error criterion [16]. It begins with a random initial partition and reassigns the patterns to clusters on the basis of similarity between the pattern and the cluster centers awaiting a convergence criterion is met. Due to the ease and its good performance in most of the scientific areas, k-means is widely used for clustering. K-means algorithm is the simple, fast and easy to merge with other methods and also it is understandable and scalable. It is mostly applicable for its improved performance in other better systems. For selection of the initial points of k corresponding to the clusters, there are many ways for the selection of points. Each point other than k preferred points are allotted to nearby centre of the cluster and it moves about around during the allocation of points to it. In the examination of the point that only points in close proximity of the cluster are expected to be allocated therefore the centre of the cluster have a tendency not to move to a great amount. Cosine dimensions can be functional in a variant of k-means known as spherical k-means which means for reducing the Euclidean distance and which also intends for making most of the cosine similarity between the documents in a cluster. The most significant discrimination among the Euclidean distance and cosine similarity as a result connecting k-means and spherical k-means is that Euclidean distances mainly focus its attention on vector magnitudes while cosine resemblance highlights on vector directions in addition to
the straight function in spherical k-means cosine function also expansively functional in various additional document clustering methods. A major difficulty regarding this algorithm is that it is susceptible to selection
of the initial partition which initializes dimensions of the cluster and its presentation is of inferior quality in many other areas when compared to other algorithms and probable to reduce them by trial and error methods. It can come together to a local minimum of the criterion function value.

3. REPRESENTATION OF DATA IN MULTI-VIEW CLUSTERING:

Most of the data is at present available in multiple representations or views that are for an instance multimedia content or web pages which are translated into several languages. The approach to handle these kinds of documents using the relations between the multiple views is known as multiple

view learning The basic idea is to control every view individuality in order to do enhanced views than simply concatenating views. In recent times, multi-view clustering methods have been proposed that has been shown to improve the performance over conventional single-view clustering. Similarity of two documents in the same cluster on the basis of common resemblance considered from the interpretations of other documents outside to the cluster. The objects which are to be estimated should be in the same cluster at the time where the location of the points from some place to begin this dimension should be outer surface of the cluster and this application is known as Multi viewpoint-based Similarity. The similarity between the two documents that appear from the starting point can be estimated based on the angle connecting the two points. Multi viewpoint-based Similarity method is probably making use of additional points than one point of reference and can have more accurate estimation of how close or distant pair of points are if they are viewed from many viewpoints. The proximity and the parting distances of the points can be perfectly predictable when they are glanced
from different viewpoints. A good quality of
viewpoints are functional because of their useful information and few of them may give the false information based on the starting point and for this reason the consequences of viewpoints which are false can be inhibited and minimized.

4. RESULTS:

By applying Clustering with Multi
viewpoint-based Similarity (MVSC) to refine the output of spherical k-means, clustering solutions are improved considerably. Interestingly, there are many situations where spherical k-means result is worse than that of other clustering methods, but after refined by Clustering with Multi viewpoint-based Similarity (MVSCs), it becomes better. Many of the improvements are with a considerably large margin, particularly when the innovative accurateness attained by spherical k-means is small. There are only some exceptions subsequent to refinement that is accurateness becomes bad. However, the diminish in such cases are minute. MVSC preceded by spherical k-means does not necessarily yields better clustering results than MVSC with random initialization. There are only a small number of cases that refinement of spherical k-means by MVSC can be found better than MVSC. Given a local optimal solution returned by spherical k- means, that refinement of spherical k –means by MVSC algorithms as a refinement
method would be inhibited by this local optimum itself and, hence, their search space may be restricted. The innovative MVSC algorithms, are not subjected to this restriction, and are capable to follow the search path of their objective utility from the
foundation. therefore, while performance development after refining spherical kmeans' result by MVSC confirm the appropriateness of MVS (Multi viewpointbased Similarity).

5. CONCLUSION:

In recent times, multi-view clustering methods have been proposed that has been shown to improved performance over conventional single-view clustering. The idea of using this technique is to control every view individuality in order to do enhanced views than simply concatenating views. It makes additional criteria for working out the similarity between the documents when they are computed with cosine similarity and additionally this approach can be computed as an alteration for cosine similarity. Hence this algorithm is used as an enhancement for spherical kmeans which is an appropriate to cosine resemblance. It is more suitable for text documents than the popular cosine similarity is shown by hypothetical analysis and it is to be sure best method for clustering problems. Supplementary informative assessment of similarity than the single initial point-based similarity measure can be obtained by multiple view points.

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.

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

[3] 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.

[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.

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

[8] D. Lee and J. Lee, "Dynamic Dissimilarity Measure for Support Based Clustering," IEEE Trans. Knowledge and Data Eng., vol. 22, no. 6, pp. 900-905, June 2010.

[9] A. Banerjee, I. Dhillon, J. Ghosh, and S. Sra, "Clustering on the Unit Hypersphere Using Von Mises-Fisher Distributions," J. Machine Learning Research, vol. 6, pp. 1345-1382, Sept. 2005.

[10] W. Xu, X. Liu, and Y. Gong, "Document Clustering Based on Non- Negative Matrix Factorization," Proc. 26th Ann. Int'l ACM SIGIR Conf. Research and Development in Informaion Retrieval, pp. 267-273, 2003.

[11] I.S. Dhillon, S. Mallela, and D.S. Modha, "Information-Theoretic Co-Clustering," Proc. Ninth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 89-98, 2003.

[12] C.D. Manning, P. Raghavan, and H. Schu¨ tze, An Introduction to Information Retrieval. Cambridge Univ. Press, 2009.

[13] C. Ding, X. He, H. Zha, 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.

[14] 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.

[15] 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.

[16] 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.

BIOGRAPHY:

R Goutham has completed B.SC (Computers) from St Patrick's Degree College, Hyderabad, Andhra Pradesh, India, and pursuing MCA in DRK Institute of Science and Technology, JNTUH, Hyderabad. His main research interest includes Data Mining, Software Engineering and Computer Networks.

M. Srinivas M.Tech(CSE) ,M.C.A .M.Tech did from JNTU, Hyderabad. M.C.A from Alagappa University, Karaikudi, Tamilnadu, India. He is working as Asst.Prof in DRK Institute of Science and Technology, Hyderabad. His main research interests include Data Base Management System, Security systems, Sensors, Intelligent Systems, Computer networks, Data mining, network protection and security control. He has various publications and presentations in various national and international journals.