IMPLEMENTATION OF DOCUMENT CLUSTERING USING MULTI-VIEW POINTS WITH SIMILARITY MEASURES

#1POTURAJU CHEBROLU, #2Dr. JAYARAMAKRISHNAIAH VEMULA

#1Associate Professor, VRS & YRN College, Chirala, India.

#2Associate Professor, ASN Degree College, Tenali, India.

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 multi-viewpoint 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. . Existing clustering algorithms are implemented based on partitioning, hierarchical, density based and grid based. These methods assume some kind of cluster relationship among the clustered objects. Similarity among the pair of objects may be defined as implicitly or explicitly. Our main objective is to cluster web documents. So, in this paper we propose “multi view- point based clustering methods with similarity measure” approach for clustering high dimensional data. This approach makes use of different viewpoints from different objects of multiple clusters and more useful assessment of similarity could be achieved. Analysis and experimental study are conducted in support of this approach.

Keywords: *multi view-point clustering, web document, high dimensional data, similarity measure.*

I.INTRODUCTION

One of the important researches in data mining is text mining which refers to the process of automatically extracting information from a usually large amount of different unstructured textual sources. In text mining, the goal is to discover unknown information, something that no one yet knows are to be extracted from large database. Clustering is extensively used for getting text with highest accuracy. It is used for grouping a set of objects into classes of similar objects and is the most interesting concept of data mining. Purpose of Clustering to group essential structures in data and organize them into meaningful subgroup for further analysis. It also makes search mechanism too easy and reduces the bulk of operations and computational cost. There have been many clustering algorithms in the data mining. The most favorite is K-means and top 10 among all data mining algorithms [1]. Even though it is a top most algorithm, it has a few basic drawbacks such as sensitiveness to initialization and to cluster size [2]. It means one should need to specify the number of clusters in advance. In spite of that, it is still popular due to its simplicity, understandability, and scalability. While offering best outcome, K-means is quick and simple to combine with other methods in larger systems. To meet various requirements k-means has many variants. For instance spherical k-means (uses cosine similarity) is used to cluster text documents while original k-means can be used to clustering using Euclidean distance [3], [4].A hierarchical clustering algorithm [8] creates a hierarchical decomposition of the given set of data objects.

Depending on the decomposition approach, hierarchical clustering algorithms are classified into agglomerative and divisive. An agglomerative clustering is bottom-up approach in such way that each object is assign to a separate cluster and merges the object with the shortest distance to form a large cluster. Generally the problem of clustering can be thought as optimization process, by optimizing similarity measures, the optimal clusters can be formed and its performance is improved. The efficiency of clustering algorithms depends on the accuracy of the similarity measure to the data. Hence the similarity measure plays a very important role in the success or failure of a clustering method. A variety of similarity measures have been proposed so far and widely used measures are cosine similarity, Jaccard coefficient and Pearson correlation coefficient. To improve the accuracy of document clustering, Correlation similarity measure is integrated to Hierarchical Agglomerative Clustering with Multi viewpoint Similarity Measure. The proposed work is motivated by research of similarity measures in document clustering.

Similarity measures play a vital role in clustering the documents.

II.RELATED WORK

This paper to represent documents and related concepts. 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. The principle definition of 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 problem formulation itself implies that some forms of measurement are needed to determine such similarity or dissimilarity. There are many state-of-the-art clustering approaches that do not employ any specific form of measurement, for instance, probabilistic model-based method [9], and nonnegative matrix factorization [10], information theoretic co clustering [11] and so on. In this paper, though, we primarily focus on methods that indeed do utilize a specific measure.

Document clustering is required in the real world applications such as web search engines. It comes under text mining. It is being used for many years. It is meant for grouping documents into various clusters. These clusters are used by various applications in the real world such as search engines. A document is treated as an object a word in the document is referred as a term. A vector is built to represent each document. The total number of terms in the document is represented by m.

Some kind of weighting schemes like Term Frequency – Inverse Document Frequency (TF-IDF) is used to represent document vectors. There are many approaches for document clustering. They include probabilistic based methods [8], non- negative matrix factorization [7] and information theoretic co-clustering [6]. These approaches are not using a particular measure for finding similarity among documents. In this paper, we make use of multi-viewpoint similarity measure for finding similarity. As found it literature, a measure widely used in document clustering is ED (Euclidian Distance).

$$\text{Dist}(di,dj) = ||di - dj|| \qquad (1)$$

K-Means is most widely used clustering algorithm due to its ease of use and simplicity. ED is the measure used in K-Means algorithm to measure the distance between objects to make them into clusters. In this case the cluster centered is computed as:

$$\text{Min} \sum_{r=1}^{k} \sum_{di \in Sr} ||di - Cr||^2 \qquad (2)$$

III. PROPOSED ALGORITHM

A.QT clustering algorithm

QT (quality threshold) clustering is an alternative method of partitioning data, invented for gene clustering. It requires more computing power than k-means, but does not require specifying the number of clusters a priori, and always returns the same result when run several times.The user chooses a maximum diameter for clusters. Build a candidate cluster for each point by including the closest point, the next closest, and so on, until the diameter of the cluster surpasses the threshold. Save the candidate cluster with the most points as the first true cluster, and remove all points in the cluster from further consideration. Must clarify what happens if more than 1 cluster has the maximum number of points? Recurse with the reduced set of points.

B. Comparisons between data clustering's

There have been several suggestions for a measure of similarity between two clustering's. Such a measure can be used to compare how well different data clustering algorithms perform on a set of data. Many of these measures are derived from the matching matrix (aka confusion matrix), e.g., the Rand measure and the Fowlkes-Mallows Bk measures. Several different clustering systems based on mutual information have been proposed. One is Marina Meila's 'Variation of Information' metric and another provides hierarchical clustering.

C. Hierarchical Document Clustering Using Frequent Item sets

Document clustering has been studied intensively because of its wide applicability in areas such as web mining, search engines, information retrieval, and topological analysis. Unlike in document classification, in document clustering no labeled documents are provided. Although standard clustering techniques such as k-means can be applied to document clustering, they usually do not satisfy the special requirements for clustering documents: high dimensionality, high volume of data, ease for browsing, and meaningful cluster labels. In addition, many existing document clustering algorithms require the user to specify the number of clusters as an input parameter and are not robust enough to handle different types of document sets in a real-world environment. For example, in some document sets the cluster size varies from few to thousands of documents. This variation tremendously reduces the clustering accuracy for some of the state-of-the art algorithms. Frequent Itemset-based Hierarchical Clustering (FIHC), for document clustering based on the idea of frequent item sets proposed by Agrawalet. al. The intuition of our clustering criterion is that there are some frequent item sets for each cluster (topic) in the document set, and different clusters share few frequent item sets. A frequent item set is a set of words that occur together in some minimum fraction of documents in a cluster. Therefore, a frequent item set describes something common to many documents in a cluster. In this technique use frequent item sets to construct clusters and to organize clusters into a topic hierarchy. Here are the features of this approach.

Reduced dimensionality. This approach uses only the frequent items that occur in some minimum fraction of documents in document vectors, which drastically reduces the dimensionality of the document set. Experiments show that clustering with reduced dimensionality is significantly more efficient and scalable. This decision is consistent with the study from linguistics (Longman Lancaster Corpus) that only 3000 words are required to cover 80% of the written text in English and the result is coherent with the Zipf's law and the findings in Mladenic et al. and Yang et al. High clustering accuracy. Experimental results show that the proposed approach FIHC outperforms best documents clustering algorithms in terms of accuracy. It is robust even when applied to large and complicated document sets. Number of clusters as an optional input parameter. Many existing clustering algorithms require the user to specify the desired number of clusters as an input parameter. FIHC treats it only as an optional input parameter. Close to optimal clustering quality can be achieved even when this value is unknown.

IV. MULTI-VIEWPOINTBASED SIMILARITY

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

Sim(di, dj) = cos(di 0, dj 0) = (di 0)t (dj 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 di and dj 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 just one point of reference. We may have a more accurate assessment of how close or distant a pair of points are, if we look at them from many different viewpoints [2][3]. From a third point dh, the directions and distances to di and dj are indicated respectively by the difference vectors (di dh) and (dj dh). By standing at various reference points dh to view di, dj and working on their difference vectors, we define similarity between the two documents as:

A. Analysis and practical examples of MVS

In this section, we present analytical study to show that the proposed MVS could be a very effective similarity measure for data clustering. In order to demonstrate its advantages, MVS is compared with cosine similarity(CS) on how well they reflect the true group structure in document collections Clustering is the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters)[6], so that the data in each subset (ideally) share some common trait - often proximity according to some defined distance measure. Data clustering is a common technique for statistical data analysis, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. The computational task of classifying the data set into k clusters is often referred to as k-clustering. Besides the term data clustering (or just clustering), there are a number of terms with similar meanings, including cluster analysis, automatic classification, numerical taxonomy, botryology and typological analysis [6].Document clustering aims to group, in an unsupervised way, a given document set into clusters such that documents within each cluster are more similar between each other than those in different clusters.

It is an enabling technique for a wide range of information retrieval tasks such as efficient organization, browsing and summarization of large volumes of text documents. Cluster analysis aims to organize a collection of patterns into clusters based on similarity. Clustering has its root in many fields, such as mathematics, computer science, statistics, biology, and economics. In different application domains, a variety of clustering techniques have been developed, depending on the methods used to represent data, the measures of similarity between data objects, and the techniques for grouping data objects into clusters.

V. HIERARCHICAL ANALYSIS MODEL

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). The agglomerative approach starts with each data point in a separate cluster or with a certain large number of clusters. Each step of this approach merges the two clusters that are the most similar. Thus after each step, the total number of clusters decreases. This is repeated until the desired number of clusters is obtained or only one cluster remains. By contrast, the divisive approach starts with all data objects in the same cluster. In each step, one cluster is split into smaller clusters, until a termination condition holds. Agglomerative algorithms are more widely used in practice. Thus the similarities between clusters are more researched [9].

VI.ALGORITHMS PROPOSED

Many algorithms have been proposed to work on multi-viewpoint similarity measure. The procedure for similarity matrix is as shown in Listing 1.

```
Procedure BUILDMVSMATRIX(A)
For r ← 1 : c do
    Ds/sr ← Σ di¢Sr di
    Ns/sr ← |S\Sr|
End for
For r ← 1 : n do
```

```
 7. R ← class of di
 8. For j ← 1 : n do
 9. If dj ϵ Sr then
 10. aij ← dtjdj - dtiDs/SrnS/Sr - dtjDs/sr nS/Sr + 1
 11. else
 12. aij ← dtjdj - dti Ds/SrnS/Sr- dtjDs/sr - Dj nS/Sr -1+1
 end if
end for
end for
return A={aij} mxn
end procedure
```

Algorithm 1 –Procedure for making similarity matrix

As per the procedure in Algorithm 1, it is known that dl and di are closer and the dl is also considered closer to di as per the multi-viewpoint simairlity measure. The Algorithm 2 shown the validation procedure.

Algorithm 2 –Validation Procedure

By averaging overall rows, the final validity is calculated. It is as given in line 14. It is known that when validation score is higher, it reflects that the similarity is higher and thus eligible for clustering. Fig. 1 shows the validity scores of multi-viewpoint similarity and cosine similarity.

As can be seen in fig. 1, Series 4 is related to klb-MVS, series 3 corresponds to klb-CS, series 2 corresponds to reutors-7

while series 1 corresponds to reutors -7 CS. As shown in fig. 1 performance of MVS is higher when compared to that of CS.

```
1. Select k seeds S1………….,Sk randomly
2. Cluster[di] ← p=argmaxr{strdi}, ʉi=1,…..,n
3. Dr ← Σ di¢Sr di, nr ← |Sr|, ʉr=1,….,k
4. End procedure
5. Procedure REFEINEMENT
6. Repeat
7. {v[1 : n]} ← random permutation of {1,….,n}
8. For j ← 1: n do
9. I ←v[j]
10. P ← cluster[di]
11. ΔIp ← I(np-1,Dp-di) – I(np,Dp)
12. q ← arg max r,r=p {I(nr+1, Dr+di)-I(nr,Dr)}
13. ΔIp ← I(nq+1, Dq+di) – I(nq,Dq)
14. If ΔIp + ΔIq > 0 then
15. Move di to cluster q: cluster[di] ← q
16. Update Dp,np,Dq,nq
17. End if
18. End for
19. until No move for all n documents
20. end procedure
```

VII.CONCLUSION

This paper presents a novel similarity approach known as multi-viewpoint based similarity measure. The similarity measure is capable of providing informative assessment and bestows high quality clusters. The proposed approach achieves highest similarity between objects of same cluster and lowest similarity between the objects of different clusters. Two criterion functions were implemented with MVS. The proposed similarity measure is tested with bench mark datasets. The proposed clustering algorithms in this paper are compared with five other clustering algorithms used for document clustering. The results revealed that the multi-viewpoint based similarity measure outperforms them.

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 Recognit. Lett., vol. 28, no. 1, pp. 110 – 118, 2007.

[2] A. Banerjee, I. Dhillon, J. Ghosh, and S. Sra, "Clustering on the unit hypersphere using von Mises-Fisher distributions," J. Mach. Learn. Res., vol. 6, pp. 1345–1382, Sep 2005.

[3] I. Dhillon and D. Modha, "Concept decompositions for large sparse text data using clustering," Mach. Learn., vol. 42, no. 1-2, pp. 143–175, Jan 2001.

[4] I. S. Dhillon, S. Mallela, and D. S. Modha, "Information-theoretic co-clustering," in KDD, 2003, pp. 89–98.

[5] S. Flesca, G. Manco, E. Masciari, L. Pontieri, and A. Pugliese, "Fast detection of xml structural similarity," IEEE Trans. on Knowl. And Data Eng., vol. 17, no. 2, pp. 160–175, 2005.

[6] I. Guyon, U. von Luxburg, and R. C. Williamson, "Clustering: Science or Art?" NIPS'09 Workshop on Clustering Theory, 2009.

[7] D. Ienco, R. G. Pensa, and R. Meo, "Context-based distance learning for categorical data clustering," in Proc. of the 8th Int. Symp. IDA, 2009, pp. 83– 94.

[8] Leo Wanner (2004). "Introduction to Clustering Techniques". Available online at: http://www.iula.upf.edu/materials/040701wanner.pdf [viewed: 16 August 2012]

[9] C. D. Manning, P. Raghavan, and H. Sch ¨ utze, An Introduction to Information Retrieval. Press, Cambridge U., 2009.

[10] on web-page clustering," in Proc. of the 17th National Conf. on Artif. Intell.: Workshop of Artif. Intell.for Web Search. AAAI, Jul. 2000, pp. 58–64.

[11] J. Shi and J. Malik, "Normalized cuts and image segmentation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, pp. 888–905, 2000.

[12] A. Strehl, J. Ghosh, and R. Mooney, "Impact of similarity measures.

[13] X. Wu, V. Kumar, J. Ross 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," Knowl.Inf. Syst., vol. 14, no. 1, pp. 1–37, 2007.

[14] W. Xu, X. Liu, and Y. Gong, "Document clustering based on nonnegative matrix factorization," in SIGIR, 2003, pp. 267–273.

[15] S. Zhong, "Efficient online spherical K-means clustering," in IEEE IJCNN, 2005, pp. 3180–3185.

AUTHOR'S PROFILE:

[1].**CH.POTHURAJU**, received Master of Computer Science applications from Osmania University, Hyderabad,I ndia , Master of Philosophy from Periyar University, palkalani nagar,salem,Tamil Nadu. He is currently working as Associate Professor, in the Department of Computer Science, V.R.S &Y.R.N COLLEGE (P.G COURSES),Chirala, which is affiliated to Acharya Nagarjuna University. He has 14 years teaching experience. . He is currently pursuing Ph.D., at Department of Computer Science and Engineering, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India. His research area is Clustering in Databases. He has published several papers in National & International Journals.

[2].V.V.JAYA RAMA KRISHNAIAH, received Doctorate in from Acharya Nagarjuna University and He received Master„s degree in Computer Application from Acharya Nagrajuna University,Guntur, India, Master of Philosophy from Vinayaka University, Salem . He is currently working as Associate Professor, in the Department of Computer Science, A.S.N. Degree College, Tenali, which is affiliated to Acharya Nagarjuna University. He has 14 years teaching experience.. His research area is Clustering in Databases. He has published several papers in National & International Journals.