International Journal of scientific research and management (IJSRM)
||Volume|| 1 ||Issue|| 6 ||Pages|| 316-322 ||2013||
Website: www.ijsrm.in ISSN (e): 2321-3418

Clustering with Multi view point-Based Similarity Measure using NMF

R.Saranya[1], P.Krishnakumari2

[1]Research Scholar, Department of Computer Science, RVS college of Arts and Science
Coimbatore, Tamil Nadu, India
sarnikovai@gmail.com
[2]Director, Department of Computer Application (MCA), RVS College of Arts and Science
Coimbatore, Tamil Nadu 641402, India

ABSTRACT

Clustering is the one of the major important task in data mining .The task of clustering is to find the fundamental structures in data and categorize them into meaningful subgroups for supplementary study and examination. Existing K-Means clustering with MVS measure it doesn't best position to cluster the data points. This problem will lead to gain less optimal solution for clustering method. This paper presents a solution to the Mulitview point based similarity measure with NMF clustering to predict k value. This paper gives a detailed study on proposing the multiview point clustering approach with the NMF clustering method. Finally experimental result were compared with the parameters in terms of precision and recall to measure the accuracy of the MVS and NMF clustering.

KEYWORDS
Document Clustering, Similarity Measure, NMF Clustering, Clustering Methods.

I.INTRODUCTION

Data mining is that the method of extracting or mining information from great deal of information .It's Associate in analytic method designed to explore giant amounts of information in search of consistent patterns and systematic relationships between variables and to validate the findings by the detected patterns to new subsets of information. It is often viewed as a result of natural evolution in development of

Functionalities like data assortment, information creation, information management, information analysis. It is the process where intelligent methods are applied in order to extract data patterns from databases, data warehouses, or other information repositories Clusters are often thought of the foremost necessary unsupervised learning problem, thus as Each different drawback of this sort, it deals with finding a structure in an exceedingly assortment of unlabelled information. A cluster is so a set of objects that are coherent internally, however clearly dissimilar to the objects to different clusters. A loose definition of cluster could also be "the methodology of organizing objects into groups whose members' are similar in some way".

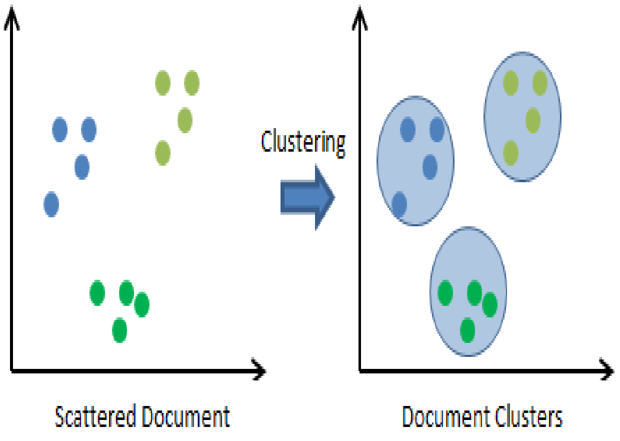

Figure.1Example of cluster formation

Fig1.shows an example of identifies the 3 clusters into which the data can be divided; the similarity criterion is *distance*: two or more objects belong to the same cluster if they are “close” according to a given distance (in this case geometrical distance). This is called *distance-based clustering*, Another type of clustering is *conceptual clustering*: two or more objects belong to the same cluster if this one defines a concept *common* to all that objects .In other words; objects are grouped to their fit to descriptive concepts, not per easy simple similarity measures.

Clustering algorithms could also be classified as listed below

1. **Flat clustering** (Creates a set of clusters without any explicit structure that would relate clusters to each other; It’s also called exclusive clustering)
2. **Hierarchical clustering** (Creates a hierarchy of clusters)
3. **Hard clustering** (Assigns each document/object as a member of exactly one cluster)
4. **Soft clustering** (Distribute the document/object over all clusters)

Algorithms

1. Agglomerative (Hierarchical clustering)
2. K-Means (Flat clustering, Hard clustering)
3. EM Algorithm (Flat clustering, Soft clustering)

Document clustering has become an increasingly important task in analyzing huge numbers of documents distributed among various sites. The challenging aspect is to organize the documents in a way that results in better search without introducing much extra cost and complexity. The Cluster Hypothesis is fundamental to the issue of improved effectiveness. It states that relevant documents tend to be more similar to each other than to non-relevant documents and therefore tend to appear in the same clusters. If the cluster hypothesis holds for a particular document collection, then relevant documents will be well separated from non-relevant ones. A relevant document may be ranked low in a best-match search because it may lack some of the query terms. In a clustered collection, this relevant document may be clustered together with other relevant items that do have the required query terms and could therefore be retrieved through a clustered search. [8]According to best-match IR systems, if a document does not contain any of the query terms then its similarity to the query will be zero and this document will not be retrieved in response to the query. [7]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.

II.PROBLEM FORMULATION

The clustering problem is expressed as:
The set of N documents $D = \{D_1, D_2, ... D_N\}$ is to be clustered. Every DiεU Nd is an attribute vector consisting of N d real measurements describing the object. The documents are to be classified into non-overlappingclustersC = $\{C_1, C_2, ... C_N\}$ (C is known as a clustering), where, K is the number of clusters, $C_1 \cup C_2 \cup ... \cup C_K$, $C\ i \neq \varphi$ and $C_1 \cap C_2 = \varphi$ for $i \neq j$. Assuming $f: DxD \rightarrow U^+$ is a measure of similarity between document feature vectors. [4]Clustering is the task of finding a partition $\{C_1, C_2, ..., C_K\}$ of D such that $\forall_{i,j} \in \{1,...K\}$, $j \neq i$, $\forall x \in C_i : f(x, O_i) \geq f(x, O_j)$ where, O i is one cluster representative of cluster C_i .

The goal of clustering is explicit as follows:
Given:

- A set of documents $D = \{D_1, D\ 2 ... D_N\}$
- A desired number of clusters K
- An objective function or fitness function that evaluates the quality of a clustering the system should to compute an assignment g: D→ (1, 2... K} and maximizes the objective function**.**

III.MEASURING SIMILARITY BETWEEN TWO DOCUMENTS

Capturing the similarity of two documents using cosine similarity measurement.[3] The cosine similarity is calculated by measuring the cosine of the angle between two document vectors. Using the code:
The main class is TFIDF Measure. This is the testing code:
void Test (string[] docs, int i, int j)

```
// docs is collection of parsed documents
{
StopWordHandler stopWord=new
StopWordsHandler() ;
TFIDFMeasure tf=new TFIDFMeasure(doc) ;
float simScore=tf.GetSimilarity( i, j); // similarity
of two given documents at the // position i,j
respectively }
```

IV.MVS (MULTIVIEW POINT BASED SIMILARITY)

In the existing system, clustering is one among that interesting and very important topic in data mining. The aim of clustering is to hunt out intrinsic structures in data, and organize them into meaningful subgroups for more study and analysis. There are clustering algorithms published every year. They can be proposed for very distinct analysis fields, and developed using totally different techniques and approaches.

This paper proposed a Multiviewpoint-based Similarity measuring method, named MVS. Theoretical analysis and empirical examples show that MVS is potentially more suitable for text documents than the popular cosine similarity. Based on MVS, two criterion functions, I_R and I_V, and their respective clustering algorithms, MVSC-IR and MVSC-I_V, have been introduced [1]. 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 (or distance) 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. Hence, effectiveness of clustering algorithms under this approach depends on the appropriateness of the similarity measure to the data at hand. For instance, the original k-means has sum-of-squared-error objective function that uses Euclidean distance. [8] In a very sparse and high-dimensional domain like text documents, spherical k-means, that uses cosine similarity (CS) rather than Euclidean distance as the measure, is deemed to be more suitable.

DISADVANTAGES

- Sensitiveness to initialization and to cluster size, and its performance can be worse than other state-of-the-art algorithms.
- K-means algorithm’s simplicity, understandability, and scalability are the reasons for its tremendous popularity not based on its Performance.
- Single view point is used for finding similarity of documents.

V.NON-NEGATIVE MATRIX FACTORIZATION CLUSTERING

The proposed is motivated by investigations from the above and similar research findings. It appears to us that the nature of similarity measure plays a very important role in the success or failure of a clustering method. This paper 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, this paper formulate new clustering criterion functions and introduce their respective clustering algorithms, Which are fast and ascendible like k-means, however also are capable of providing high-quality and consistent performance. This paper develop two criterion functions for document clustering and their optimization algorithms .Finally extensive experiments on real-world benchmark data sets are presented. This paper proposing a new way to compute the overlap rate in order to improve time efficiency and "the accuracy" is mainly concentrated with NMF. Experiments in both intra and inter of data and document clustering data show that this approach can improve the efficiency of clustering and save computing time. Given a data set satisfying the distribution of a mixture of Gaussians, the degree of overlap between components affects the number of clusters “perceived” by a human operator or detected by a clustering algorithm [11]. In other words, there may be a significant difference between intuitively defined clusters and the true clusters corresponding to the components in the mixture.

VI.CLUSTERING WITH MULTI-VIEWPOINT USING K-Means

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. The similar documents are grouped together in a cluster, if their cosine similarity measure is less than a specified threshold. It is used in the traditional k-means algorithm. [2]The objective of k-means is to minimize the euclidean distance between objects of a cluster and that cluster's centroid

$$Sim(d_i,dj)=cos(d_i,dj)=d_i^j d_j$$

Cosine measure is used in a variant of k-means called spherical k-means. 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 centroid.

$$\max \sum_{r=1}^{k} \sum_{d_i \in S_r} \frac{d_i^t C_r}{\|C_r\|}.$$

VII.CLUSTERING WITH MULTI-VIEWPOINT USING NMF

NMF uses a multiplicative update algorithm, to factor a non-negative data matrix into two factor matrices referred to as W and H. This represents the W as columns, the H as rows, and the number of clusters to which the samples are to be assigned. Starting with randomly selected matrices and using an iterative approach with a specified cost measurement we can reach a locally optimal solution for these factor matrices. H and W can then be calculated as metagenes and metagenes expression patterns, respectively.

NMF is a matrix factorization algorithm that finds the positive factorization of a given positive matrix. Assume that the given document corpus consists of k document clusters. Here the goal is to factorize X into the non-negative $m \times k$ matrix U and the non-negative $k \times n$ matrix VT that minimize the following objective function:

$$J = \frac{1}{2} \|X - UV^T\|$$

Where $\|.\|$ denotes the squared sum of all the elements in the matrix. The objective function J can be re-written as:

$$J = \frac{1}{2} tr((X - UV^T)(X - UV^T)^T)$$

Non-negative matrix factorization (NMF) has previously been shown to be a useful decomposition for multivariate data. Two different multiplicative algorithms for NMF are analyzed. They differ only slightly in the multiplicative factor used in the update rules. One algorithm can be shown to minimize the conventional least squares error while the other minimizes the generalized Kullback-Leibler divergence. The monotonic convergence of both algorithms can be proven using an auxiliary function analogous to that used for proving convergence of the Expectation- Maximization algorithm. The algorithms can also be interpreted as diagonally rescaled gradient descent, where the rescaling factor is optimally chosen to ensure convergence.

VIII.COMPARSION MVS AND NMF

In this section this paper measure the performance of the existing MVS with IR and IV then we measure the results of the NMF based clustering algorithm. Three types of external evaluation metric are used to assess clustering performance. They are the Measure, precision and Recall.

PRECISION

In the field of data retrieval, **precision** is the fraction of retrieved documents that are relevant to the search.

Precision=|{relevantdocuments}Ω{retrived documents}|/|{retrived documents}|

Precision takes all retrieved documents into account, however it also can be evaluated at a given cut-off measure, considering only the topmost results returned by the system. This measure is called **precision**

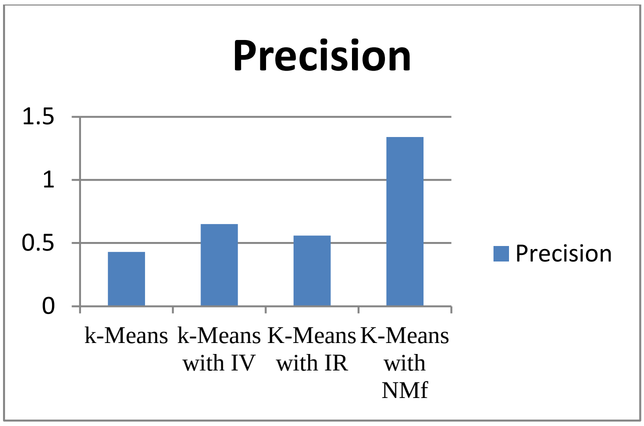

Figure2: Precision comparison

In this figure2 measure precision value of the K-Means, K-Means with IV, K-Means IR and finally measure the NMF clustering .Finally the results shows that the proposed NMF clustering shows the best precision value than the existing K-Means , K-Means with IR , K-Means IV.

Table1.Precision comparison

Measure	k-Means	k-Means with IV	K-Means with IR	K-Means with NMF
Precision	0.43	0.65	0.56	1.34

RECALL

Data retrieval is the fraction of the documents that are relevant to the query that are successfully retrieved is called as recall.

Precision=|{relevantdocuments}Ω{retrived documents}|/|{relevant documents}|

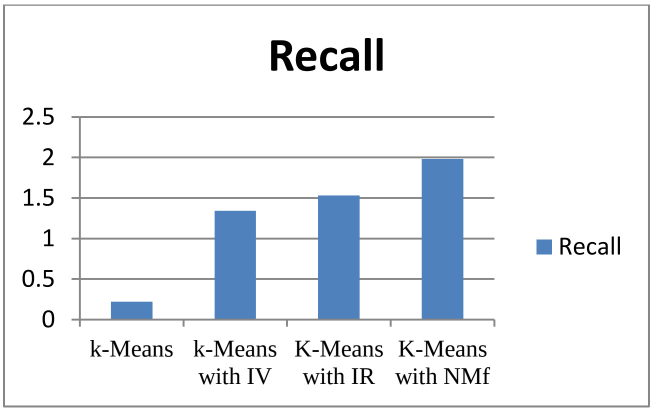

Figure3: Recall comparison

In this figure3 measure recall value of the K-Means, K-Means with IV, K-Means IR and finally measure the NMF clustering .Finally the results shows that the proposed NMF clustering shows the best recall value than the existing K-Means, K-Means with IR , K-Means IV

Table2.Recall comparison

Measure	k-Means	k-Means with IV	K-Means with IR	K-Means with NMF
Recall	0.22	1.34	1.53	1.98

Table3.NMF comparison

Meaure	k-Means	k-Means with IV	K-Means with IR	K-Means with NMf
Precision	0.43	0.65	0.56	1.34
Recall	0.22	1.34	1.53	1.98
FMeasure	0.11	0.72	0.70	1.61
Accuracy	0.21	0.67	0.59	1.21

F-MEASURE

A measure that combines precision and recall is the harmonic mean of precision and recall is called as the traditional F-measure or balanced F-score

F=2. (Precision-recall/precision recall)

This is also known as the F measure, because recall and precision are evenly weighted.

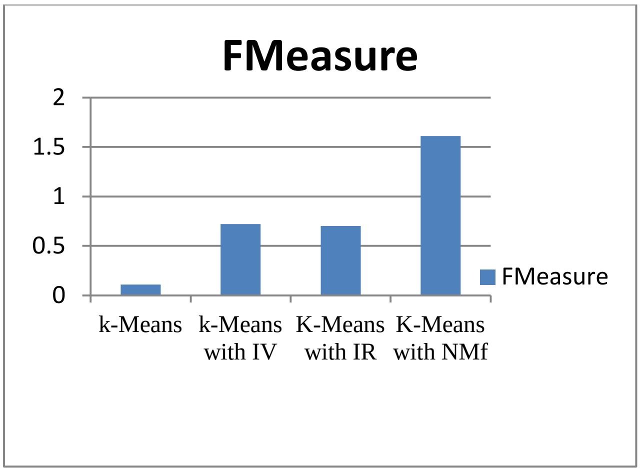

Figure4.FMeasure Comparison

In this figure4 measure Fmeasure value of the K-Means, K-Means with IV, K-Means IR and finally measure the NMF clustering .Finally the results shows that the proposed NMF clustering shows the best FMeasure value than the existing K-Means , K-Means with IR , K-Means IV.

Table4.FMeasure comparison

Measure	k-Means	k-Means with IV	K-Means with IR	K-Means with NMf
FMeasure	0.11	0.72	0.70	1.61

ACCURACY

In this paper checked the accuracy of our method. Table 3 and Figure 5 show the result value of the K-Means, K-Means with IV, K-Means IR and finally measure the NMF clustering respectively.

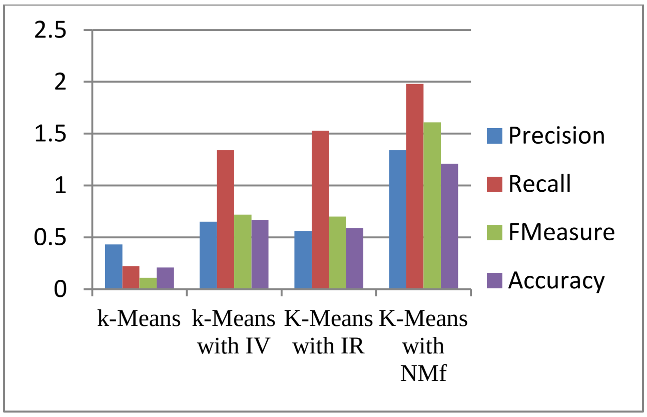

Figure5. NMF comparison

In this figure5 measure the accuracy value of the K-Means, K-Means with IV,K-Means IR and finally measure the NMF clustering .Finally the results shows that the proposed NMF clustering shows the best accuracy value than the existing K-Means ,K-Means with IR ,K-Means IV.

IX.RESULT DISCUSSION

Based on the analysis and comparison, NMF could be a very effective similarity measure for data clustering.NMF is clearly better than MVS for both data sets in the validity test. NMF is more useful for finding the similarity of text document.

The accuracy has illustrated the potential advantage of the NMF compared to the multi viewpoint-based similarity measure.

X.CONCLUSION

Multiview point based Similarity measure method, named MVS. Theoretical analysis and empirical examples show that MVS is potentially more suitable for text documents than the popular cosine similarity. But these MVS based similarity measure doesn't perform efficient in all the dataset, to overcome these problem this paper proposed a NMF based Clustering algorithm to reduce the space complexity and improve the accuracy of clustering. NMF is equivalent to a relaxed form of K-means clustering but the matrix factor W contains cluster centroids and H contains cluster membership indicators, when using the least square as NMF objective to select the best cluster results From document clustering .it leads to improve the clustering accuracy than the MVS.

REFERENCE

[1] Duc Thang Nguyen, Lihui Chen, Senior Member, IEEE, and Chee Keong Chan," Clustering with Multiviewpoint-Based Similarity Measure" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 6, JUNE 2012.

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

[3] **Thanh Da,"Term frequency/Inverse document frequency implementation in C#", 28 Oct 2005 .**

[4] **K. Premalatha** and **A.M. Natarajan**" A Literature Review on Document Clustering", **Aug 26, 2010.**

[5] **I. Dhillon, D. Modha,"Concept decompositions for large sparse text data using clustering", Mach. Learn., Vol. 42, No. 1-2, pp. 143–175, 2001. [4] S. Zhong,"Efficient online spherical K-means clustering", in IEEE IJCNN, 2005, pp. 3180–3185.**

[6] **A. Banerjee, S. Merugu, I. Dhillon, J. Ghosh,"Clustering with Bregman divergences", J. Mach. Learn. Res., Vol. 6, pp. 1705–1749, Oct 2005.**

[7] **E. Pekalska, A. Harol, R. P. W. Duin, B. Spillmann, H. Bunke,"Non-Euclidean or non-metric measures can be informative", in Structural, Syntactic, and Statistical Pattern Recognition, ser. LNCS, Vol. 4109, 2006, pp. 871–880.**

[8] **M. Pelillo,"What is a cluster? Perspectives from game theory", in Proc. of the NIPS Workshop on Clustering Theory,2009.**

[9] **D. Lee, J. Lee,"Dynamic dissimilarity measure for support based clustering", IEEE Trans. on Knowl. and Data Eng., Vol. 22, No. 6, pp. 900–905, 2010.**

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

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