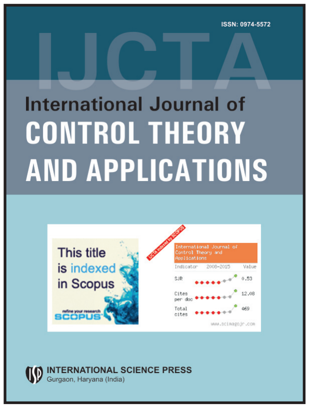

International Journal of Control Theory and Applications

ISSN : 0974-5572

© **International Science Press**

Volume 10 • Number 12 • 2017

Hybrid MVS Clustering with KL Divergence Approach in Data Mining

G. Jagatheeshkumar[a] **and Selva Brunda**[b]

[a]*Research Scholar, R&D Centre, Bharathiar University, Coimbatore-46, Tamilnadu, India*
[b]*Research Guide, R&D Centre, Bharathiar University, Coimbatore-46, Tamilnadu, India*

Abstract: In the past few years, the data sets have become available and the usage of data sets gets increased. Data mining searches large stores of data for discovering patterns and trends for simple analysis. It uses sophisticated mathematical algorithms to partition the data. Identification of similar and dissimilar attribute is a challenging task. There are many clustering algorithms published to identify the similarity between the elements in the given data set. The aim of clustering is to find intrinsic structures in data, and organize them into meaningful subgroups for further study and analysis. The existing system introduced a MultiViewpoint-based Similarity (MVS) measurement and traditional K- Means clustering methods for similarity measurement and document clustering respectively. The MVS utilizes many different viewpoints on similarity measure and it has been successfully applied in data clustering. However, it does not produce a satisfactory clustering result. To solve this problem the proposed system introduced a hybrid MVS clustering with KL divergence mechanism. The dimensionality reduction is performed by Term Variance (TV) which is used to increase both the effectiveness and efficiency of clustering algorithms. In order to compute distances between documents, two measures have been used, namely Kullback-Leibler divergence (KL divergence) and MVS. Incremental MVS based clustering technique is used for document clustering. The simulation result shows the improved accuracy for the Hybrid MVS clustering with KL divergence approach.

Keywords: Clustering, Similarity measurement, KL divergence and Term Variance.

1. INTRODUCTION

Data mining is the exploration and analysis of large data sets, in order to discover meaningful pattern and rules [1] [2]. The key idea is to find effective way to combine the computer's power to process the data with the human eye's ability to detect patterns. The objective of data mining is designed for, and work best with large data sets. Data mining is the component of wider process called knowledge discovery from database [3]. Data mining is a multi-step process, requires accessing and preparing data for a mining the data, data mining algorithm, analyzing results and taking appropriate action [4].

Clustering is one of the most interesting and important topics in data mining. It is "the process of organizing objects into groups whose members are similar in some way". Developing methods to organize large amounts of unstructured text documents into a smaller number of meaningful clusters would be very helpful as document clustering is vital to such tasks as indexing, filtering, automated metadata generation, word sense disambiguation.

An important step in any clustering is to select a distance measure, which will determine how similarity [5] of two elements is calculated. This will influence the shape of the clusters, as some elements may be close to one another according to one distance and farther away according to another. It is expected that distance between objects within a cluster should be minimum and distance between objects within different clusters should be maximum. A clustering using distance function, called distance based clustering, is a very popular technique to cluster the objects and has given good results. A variety of similarity measures have been designed and widely applied in literature, such Euclidean distance, Manhattan distance, Minkowski distance, Cosine similarity, etc. to group objects in clusters.

In order to achieve better similarity measurement novel SMTP is introduced by Yung Shen Lin et. al., [6]. It computes the similarity between two documents by embedding several properties. Kalaivendhan K. et. al., [7] presented HAC and Correlation similarity techniques which are used for any type of text document to display the most relevant document of the clusters. It achieves high efficacy and lower computational complexity. Based on the degree of closeness a similarity/distance measurement has been computed by Anna Huang [8]. It should correspond to the characteristics that are believed to distinguish the clusters embedded in the data.

2. LITERATURE SURVEY

Yung-Shen Lin, Jung-Yi Jiang, and Shie-Jue Lee [9] introduced a new measure for computing the similarity between two documents. It is a symmetric measure. The difference between presence and absence of a feature is considered more essential than the difference between the values associated with a present feature. The similarity decreases when the number of presence-absence features increases. An absent feature has no contribution to the similarity. The similarity increases as the difference between the two values associated with a present feature decreases. This work mainly focuses on textural features. Furthermore, the contribution of the difference is normally scaled.

Venkata Gopala Rao S. et. al.[10] found that except for the Euclidean distance measure, the other measures have comparable effectiveness for the partitioned text document clustering task. Pearson correlation coefficient and the averaged measures are slightly better in that their resulting clustering solutions are more balanced and have a closer match with the manually created category structure.

In [11] distance based similarity measures of Fuzzy sets are considered which have a high importance in reasoning methods handling sparse fuzzy rule bases. The rule antecedents of the sparse fuzzy rule bases are not fully covering the input universe. Therefore the applied similarity measure has to be able to distinguish the similarity of non overlapping fuzzy sets, too.

Hung Chim and Xiaotie Deng [12] found that the phrase has been considered as a more informative feature term for improving the effectiveness of document clustering. They proposed a phrase-based document similarity to compute the pairwise similarities of documents based on the Suffix Tree Document (STD) model. By mapping each node in the suffix tree of STD model into a unique feature term in the Vector Space Document (VSD) model, the phrase-based document similarity naturally inherits the term TF-IDF weighting scheme in computing the document similarity with phrases. They applied the phrase-based document similarity to the group-average Hierarchical Agglomerative Clustering (HAC) algorithm and developed a new document clustering approach. Their evaluation experiments indicate that the new clustering approach is very effective on clustering the documents of two standard document benchmark corpora OHSUMED and RCV1.

3. PROPOSED METHODOLOGY

From a more technical viewpoint, our datasets consist of unlabeled objects—the classes or categories of documents that can be found are a priori unknown. The rationale behind clustering algorithms is that objects within a valid cluster are more similar to each other than they are to objects belonging to a different cluster. Thus, once a data partition has been induced from data, the expert examiner might initially focus on reviewing representative documents from the obtained set of clusters. This proposed system consists of many functional sub-modules such as Pre-Processing Module, Term Frequency, Calculating the number of clusters, incremental MVS Clustering techniques and Query Processing**.**

Pre-Processing Module

Before running clustering algorithms on text datasets, it need to perform some pre-processing steps. In particular, stop-words (prepositions, pronouns, articles, and irrelevant document metadata) have been removed. Then, it adopted a traditional statistical approach for text mining, in which documents are represented in a vector space model. In this model, each document is represented by a vector containing the frequencies of occurrences of words, which are defined as delimited alphabetic strings, whose number of characters is between 4 and 25. It also used a dimensionality reduction technique known as Term Variance (TV) that can increase both the effectiveness and efficiency of clustering algorithms. TV selects a number of attributes (in our case 100 words) that have the greatest variances over the documents. In order to compute distances between documents, two measures have been used, namely: cosine (KL divergence) and MVS. The later has been used to calculate distances between file (document) names only.

Similarity Measurements

In similarity based clustering, an object is considered as a probability distribution of terms. The similarity of two objects is measured as the distance between the two corresponding probability distributions. The Kullback-Leibler divergence (KL divergence), also called the relative entropy, is a widely applied measure for evaluating the differences between two probability distributions. Given two uncertain objects P and Q and their corresponding probability distributions, D(P Q) evaluates the relative uncertainty of Q given the distribution of P.

$$D_{KL}(P \| Q) = P\log\left(\frac{P}{Q}\right) \tag{1}$$

In the document scenario, the divergence between two distributions of words is:

$$D_{KL} = (\vec{t_a} \| \vec{t_b}) = \sum_{t=1}^{m} \omega_{t,a} \times \log\left(\frac{\omega_{t,a}}{\omega_{t,b}}\right) \tag{2}$$

Therefore it is not a true metric. As a result, use the averaged KL divergence instead, which is defined as

$$|D_{AvgKL}\ (P \| Q) = \pi_1 D_{KL}(P \| M) + \pi_2 D_{KL}(Q \| M) \tag{3}$$

Where

$$\pi_1 = \frac{P}{P+Q},\ \pi_1 = \frac{Q}{P+Q} \text{ and } M = \pi_1 P + \pi_2 Q.$$

The KL divergence similarity can be expressed in the following form without changing its meaning

$$\text{Sim}\ (d_i, d_j) = D_{AvgKL}(d_i - 0, d_j - 0) \tag{4}$$

$$\text{Probability distribution } (\mathrm{P}, \mathrm{Q}) = \mathrm{D}_{\mathrm{AvgKL}}(\mathrm{P} - 0, \mathrm{Q} - 0) = (\mathrm{P} - 0), (\mathrm{Q} - 0) \tag{5}$$

The average weighting between two vectors ensures symmetry, that is, the divergence from document *i* to document *j* is the same as the divergence from document *j* to document *i*.

The final form of MVS depends on particular formulation of the individual similarities within the sum. If the relative similarity is defined by dot-product of the difference vectors, From this point onwards, system will denote the proposed similarity measure between two document vectors d_i and d_j by MVS(d_i, d_j).

The KL divergence is always non-negative, and satisfies Gibbs' inequality. That is, $\mathrm{D}(\mathrm{P} \| \mathrm{Q}) \geq 0$ with equality only if P = Q. Therefore, the smaller the KL divergence, the more similar the two uncertain objects.

The final form of MVS depends on particular formulation of the individual similarities within the sum. If the relative similarity is defined by dot-product of the difference vectors, we have

$$\mathrm{MVS} = (d_i, d_j | d_i, d_j \in \mathrm{Sr}) = \frac{1}{n - n_r} \sum\nolimits_{d_h \in \mathrm{S/S}_r} (d_i - d_h)^t \, (d_j - d_h) \tag{6}$$

$$\mathrm{MVS} = (d_i, d_j | d_i, d_j \in \mathrm{Sr}) = \frac{1}{n - n_r} \sum\nolimits_{d_h} \mathrm{KL}_{div} \, (d_i - d_h, d_i - d_h) \| d_i - d_h \| | d_j - d_h | \tag{7}$$

The similarity between two points d_i and d_j inside cluster Sr, viewed from a point d_h outside this 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 these two points. This definition is based on the assumption that d_h is not in the same cluster with d_i and d_j. The smaller the distances $\| d_i - d_h \|$ and $\| d_j - d_h \|$ are, the higher the chance that d_h is in fact in the same cluster with d_i and d_j, and the similarity based on d_h should also be small to reflect this potential. Therefore, through these distances provides a measure of inter-cluster dissimilarity, given that points d_i and d_j belong to cluster Sr, whereas d_h belongs to another cluster. The overall similarity between d_i and d_j is determined by taking average over all the viewpoints not belonging to cluster S_r. It is possible to argue that while most of these viewpoints are useful, there may be some of them giving misleading information just like it may happen with the origin point.

Algorithm 1: Build similarity matrix

```
1.  for $r \leftarrow 1{:}\, c$ do
2.  $\mathrm{D}_{\mathrm{S/S}_r} \leftarrow \sum_{d_{i \in \mathrm{S}_r}} d_i$
3.  $n_{\mathrm{S/S}_r} \leftarrow |\mathrm{S/S}_r|$
4.  end for
5.  for $i \leftarrow 1{:}\, n$ do
6.  $r \leftarrow$ class of $d_i$
7.  for $j \leftarrow 1{:}\, n$ do
8.  if $d_j \in \mathrm{S}_r$ then
9.  $a_{ij} \leftarrow d_i^t d_j - d_i^t \genfrac{}{}{0pt}{}{\mathrm{D}_{\mathrm{S/S}_r}}{n_{\mathrm{S/S}_r}} - d_{\mathrm{J}}^t \genfrac{}{}{0pt}{}{\mathrm{D}_{\mathrm{S/S}_r}}{n_{\mathrm{S/S}_r}} + 1$
10. else
11. $a_{ij} \leftarrow d_i^t d_j - d_i^t \genfrac{}{}{0pt}{}{\mathrm{D}_{\mathrm{S/S}_r - dj}}{n_{\mathrm{S/S}_r} - 1} - d_{\mathrm{J}}^t \genfrac{}{}{0pt}{}{\mathrm{D}_{\mathrm{S/S}_r - dj}}{n_{\mathrm{S/S}_r} - 1} + r$
```

```
end if
end for
end for
return A = $\{a_{ij}\} n \times n$
end process
```

Calculating the Number of Clusters: In order to estimate the number of clusters, a widely used approach consists of getting a set of data partitions with different numbers of clusters and then selecting that particular partition that provides the best result according to a specific quality criterion (e.g., a relative validity index). Such a set of partitions may result directly from a hierarchical clustering dendrogram or, alternatively, from multiple runs of a partitional algorithm (e.g., K-means) starting from different numbers and initial positions of the cluster prototypes.

Incremental MVS based Clustering Technique: The proposed system used a simple approach to remove outliers. This approach makes recursive use of the silhouette. Fundamentally, if the best partition chosen by the silhouette has singletons (i.e., clusters formed by a single object only), these are removed. Then, the clustering process is repeated over and over again until a partition without singletons is found. At the end of the process, all singletons are incorporated into the resulting data partition (for evaluation purposes) as single clusters.

This system defined a clustering framework by MVSC, meaning Clustering with Multiviewpoint-based Similarity. Subsequently, propose an incremental MVS Clustering, which is MVSC with criterion of incremental function. The main goal is to perform document clustering by optimizing incremental function (I_R or I_V). The incremental *k*-way algorithm, a sequential version of *k*-means is employed. Considering that the expression which depends only on n_r and D_r, can be written in a general form

$$I_v = \sum_{r=1}^{k} I_r(n_r, D_r) \tag{8}$$

Where, $I_r(n_r, D_r)$ corresponds to the objective value of cluster *r*. With this general form, the incremental optimization algorithm, which has two major steps Initialization and Refinement.

At Initialization, *k* arbitrary documents are selected to be the seeds from which initial partitions are formed. Refinement is a procedure that consists of a number of iterations. During each iteration, the *n* documents are visited one by one in a totally random order. Each document is checked if its move to another cluster results in improvement of the objective function. If yes, then document is moved to the cluster that leads to the highest improvement. If no clusters are better than the current cluster, the document is not moved. The clustering process terminates when iteration completes without any documents being moved to new clusters. Unlike the traditional *k*-means, this algorithm is a stepwise optimal procedure. While *k*-means only updates after all *n* documents have been reassigned, the incremental clustering algorithm updates immediately whenever each document is moved to new cluster. Since every move when happens increases the objective function value, convergence to local optimum is guaranteed.

During the optimization procedure, in each iteration, the main sources of computational cost are

Searching for optimum clusters to move individual documents to: $O(n_z, k)$

Updating composite vectors as a result of such moves: $O(m, k)$.

where, n_z is the total number of nonzero entries in all document vectors. Our clustering approach is partitional and incremental; therefore, computing similarity matrix is absolutely not needed. If τ denotes the number of iterations the algorithm takes, since n_z is often several times larger than m for document domain, the computational complexity required for clustering with I_R and $O(n_z \cdot \tau \cdot k)$.

Algorithm 2: Incremental clustering algorithm

```
Initialzation
    select $k$ seeds randomly
    clustered $[d_i] \leftarrow p$
    $D_r \leftarrow \sum_{d_i \in s_r} n_r \leftarrow |s_r|$
end process
procedure REFINEMENT
    repeat
        $\{v[1{:}n]\} \leftarrow$ random permutation of $\{1, ..., n\}$
        for $j \leftarrow 1{:}n$ do
            $i \leftarrow v[j]$
            $p \leftarrow$ cluster$[d_i]$
            $\Delta I_p \leftarrow I(n_p - 1, D_p - d_i) - I(n_p, D_p)$
            $q \leftarrow \arg \max_{r \neq p} r \{I(n_r + 1, D_r + d_i) - I(n_r, D_r)\}$
            $\Delta I_q \leftarrow I(n_q + 1, D_q + d_i) - I(n_q, D_q)$
            if $\Delta I_p + \Delta I_q > 0$ then
                move $d_i$ to cluster $q$: cluster $[d_i] \leftarrow q$
                update $D_p, n_p, D_q, n_q$
            end if
        end for
    until No move for all n documents
end procedure
```

4. EXPERIMENTAL RESULTS

In this section, the performance of the proposed Hybrid MVS clustering with KL divergence method is evaluated and compared with existing Incremental MVS Clustering scheme. The existing and proposed clustering methods are compared in terms of accuracy, FScore and Normalized Mutual Information (NMI).

1. **Accuracy:** Accuracy measures the fraction of documents that are correctly labels, assuming a one-to-one correspondence between true classes and assigned clusters.

   Figure 1 shows the comparison of accuracy performance for proposed Hybrid MVS clustering with KL divergence and existing Incremental MVS Clustering. In x axis existing and proposed methods are taken and y axis accuracy is taken. Compare to k-means the proposed an increase in the accuracy of the similarity measurement within the clusters.

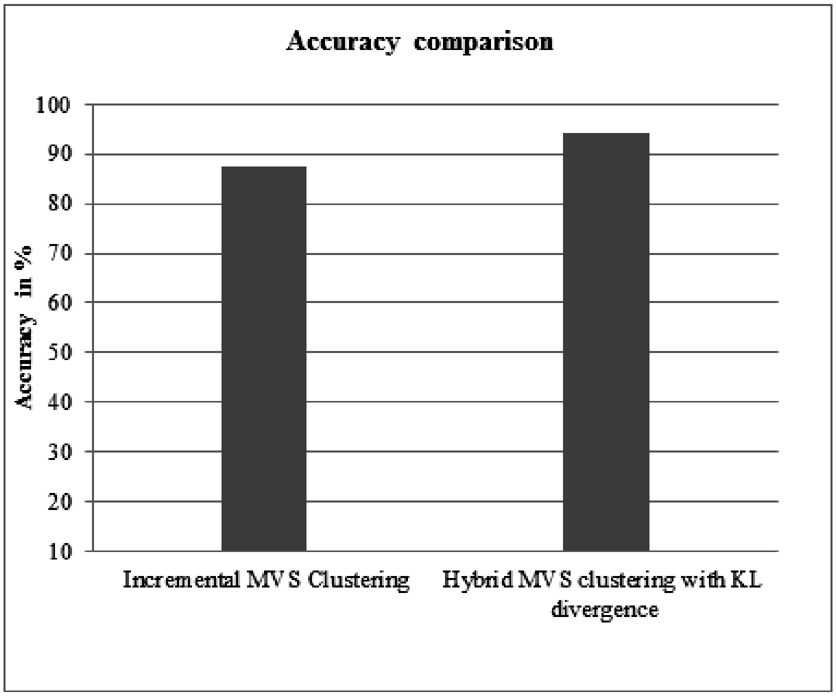

Figure 1: Accuracy comparison

2. **FScore:** FScore is an equally weighted combination of the "precision" (P) and "recall" (R) values used in information retrieval. Given a clustering solution, FScore is determined as

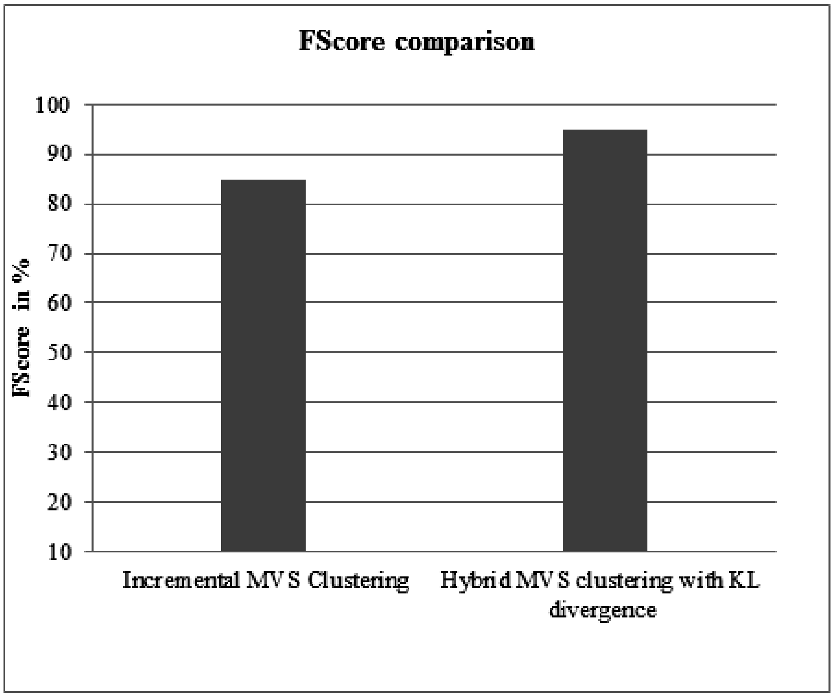

Figure 2: FScore comparison

$$\text{FScore} = \sum_{i=1}^{k} \frac{n_i}{n} \max_j (\text{F}_{ij}) \tag{9}$$

$$(\text{F}_{ij}) = \frac{2 \times \text{P}_{ij} \times \text{R}_{ij}}{\text{P}_{ij} + \text{R}_{ij}}; (\text{P}_{ij}) = \frac{n_{ij}}{n_j}; (\text{R}_{ij}) = \frac{n_{ij}}{n_i}$$

where, n_i denotes the number of documents in class i, n_j the number of documents assigned to cluster j, and n_{ij} the number of documents shared by class i and cluster j.

Figure 2 shows the comparison FScore performance for proposed Hybrid MVS clustering with KL divergence and existing Incremental MVS Clustering. In X axis existing and proposed methods are taken and in y axis FScore is taken. From the graph it is clear that the Hybrid MVS clustering with KL divergence method provides higher FScore than existing method.

3. **NMI Comparison:** NMI measures the information the true class partition and the cluster assignment share. It measures how much knowing about the clusters helps us know about the classes.

$$\text{NMI} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} n_{i,j} \log\left(\frac{nn_{ij}}{nn_j}\right)}{\sqrt{\left(\sum_{i=1}^{k} n_i \log \frac{n_i}{n}\right)\left(\sum_{j=1}^{k} n_j \log \frac{n_j}{n}\right)}} \tag{10}$$

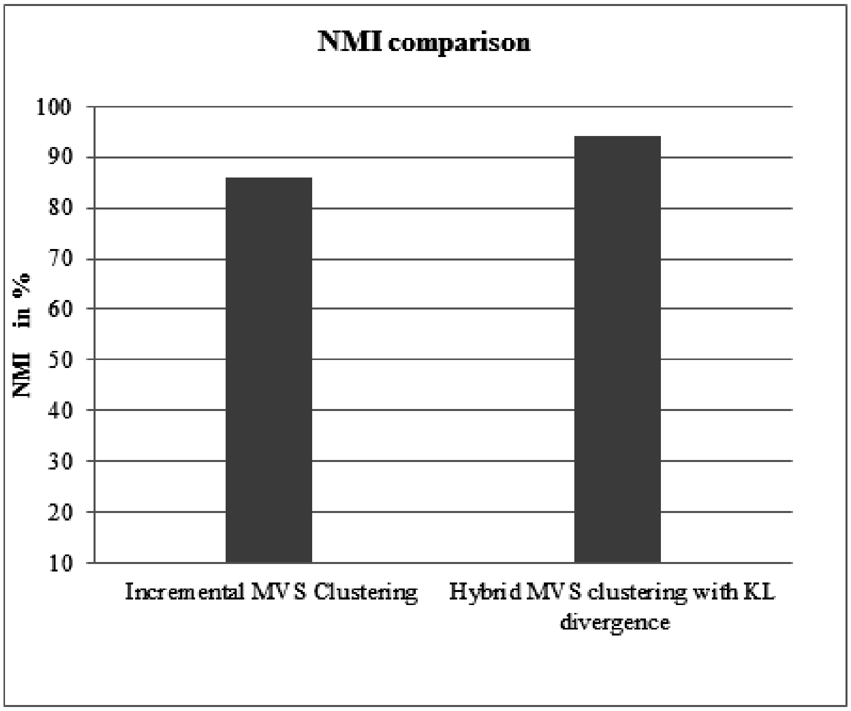

Figure 3: NMI comparison

Figure 3 shows the comparison NMI performance for proposed Hybrid MVS clustering with KL divergence and existing Incremental MVS Clustering. In X axis existing and proposed methods are taken and in y axis NMI is taken. From the graph it is clear that the Hybrid MVS clustering with KL divergence method provides higher NMI than existing method.

5. CONCLUSION

The proposed system introduced a hybrid MVS clustering with the KL divergence approach. In this work two similarity measurement method such Kullback-Leibler divergence (KL divergence) and MVS are used. KL divergence is applied for evaluating the differences between two probability distributions. A novel multi-viewpoint based similarity (MVS) measure utilizes many different viewpoints at same time to assess the similarity between data objects sparse and high-dimensional space, particularly text documents. Based on that similarity value the documents are clustered efficiently. The proposed system provides high accuracy, FScore and NMI than most existing schemes. The results conclude that the proposed algorithm provides better clustering in terms of better performance.

REFERENCES

[1] Pavel Berkhin, “A Survey of Clustering Data Mining Techniques”, pp. 25-71, 2002.

[2] Cheng-Ru Lin, Chen, Ming-Syan Syan , “Combining Partitional and Hierarchical Algorithms for Robust and Efficient Data Clustering with Cohesion Self-Merging” IEEE Transactions On Knowledge And Data Engineering, Vol. 17, No. 2, pp. 145-159, 2005.

[3] Oded Maimon, Lior Rokach, “Data Mining AND Knowlwdge Discovery Handbook”, Springer Science+Business Media. Inc, pp. 321-352, 2005.

[4] Arun K Pujari “Data Mining Techniques” pg. 42-67 and pg. 114-149, 2006.

[5] Gholamreza Esfandani, Mohsen Sayyadi, Amin Namadchian, “GDCLU: a new Grid-Density based CLUstring algorithm”, IEEE 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, pp. 102-107, 2012.

[6] Yung-Shen Lin, Jung-Yi Jiang, and Shie-Jue Lee, Member, IEEE, “A Similarity Measure for Text Classification and Clustering”, in IEEE Transactions On Knowledge and Data Engineering, Vol. 26, No. 7, July 2014, 1575.

[7] Kalaivendhan. K, Sumathi. P, “An Efficient Clustering Method To Find Similarity Between The Documents”, in International Journal of Innovative Research in Computer and Communication Engineering ,Vol. 2, Special Issue 1, March 2014.

[8] Anna Huang, Department of Computer Science, The University of Waikato, Hamilton, New Zealand,” Similarity Measures for Text Document Clustering”, New Zealand Computer Science Research Student Conference (NZCSRSC), Christchurch, New Zealand, April 2008.

[9] “A Study on Existing Protocols and Energy-Balanced Routing Protocol for Data Gathering in Wireless Sensor Networks” published in International Journal of Computing and Technology on Nov 10, 2013. Impact Factor 1.213(Refereed Journal) www.cirworld.com/index.php/ijct/article/view/2780/pdf_293

[10] “Challenges and Authentication in Wireless Sensor Networks by using promising Key Management Protocols” at International Conference in Kristu Jayanthi College, Bangalore on Feb 19th & 20th 2015 and Published in International Journal of Computer Applications. Impact Factor 0.814. www.ijcaonline.org/icctac2015/number1/icctac2005.pdf

[11] Zsolt Csaba Johanyák, Szilveszter Kovács, Distance based similarity measures of fuzzy sets, 2005.

[12] H. Chim and X. Deng, “Efficient phrase-based document similarity for clustering”, IEEE Transactions on Knowledge and Data Engineering, Vol. 20, No. 9, pp. 1217-1229, 2008.