# Hybrid-Fast Multi-View Clustering Algorithm for Mining Web documents

Mahabbuni Shaik [1], Mr. T Seshu Chakravarthy[2], T.Surekha[3]
[1] Student(Mtech),
[2]Assistant Prof, CSE department, CSE Department, Narsaraopet Engineering College, JNTU Kakinada.
[3]Associate Prof, CSE department, CSE Department, Narsaraopet Engineering College, JNTU Kakinada.

**Abstract:** Clustering is one of the raising techniques in data mining. From the past more and more data is collected by the companies for various purposes from multiple databases. Finding the patterns is one of the challenging tasks from various sources. Clustering is the method for finding similar groups in same group or another and from different groups. Number of multi-view clustering has been introduced to integrate different views of data. The challenging problem is to join the large-scale data clustering's with heterogeneous features. In this paper, proposed system focuses on hybrid-fast multi-view clustering algorithm to integrate heterogeneous representations of large scale data. Results will show the performance of our proposed system.

**Keywords: Clustering, multi-view, large scale data.**

## Introduction:

Clustering is an effort to characterize comparable articles in the same gatherings. Cluster investigation develops good cluster when the individuals from a cluster have a high level of closeness to one another and are not care for individuals from different groups. There numerous uses of clustering differing in numerous fields, for example, information mining, example acknowledgment, picture arrangement, organic sciences, advertising, city-arranging, report recoveries, and so forth. The World Wide Web is an incomprehensible asset of data and services that keeps on grow quickly. Effective internet searchers have been created to help in finding new reports by classification, substance, or subjects. Notwithstanding, questions regularly return conflicting results, with archive referrals that meet the pursuit criteria yet are of no enthusiasm to the client. Web records have very much characterized structures, for example, letters, words, sentences, passages, segments, accentuation marks, HTML labels et cetera. Henceforth, creating enhanced systems for performing machine learning strategies in this incomprehensible measure of non even, semi organized web information is highly desirable.

The effortlessness of K-means made this algorithm utilized as a part of different fields. K-means is a clustering bunching strategy that isolates information into k commonly over the top gatherings. By iterative such apportioning, K-means minimizes the entirety of separation from every information to its bunches. K-implies technique is extremely famous in view of its capacity to group colossal information, furthermore exceptions, rapidly and effectively.

Dimension reduction is the procedure of decreasing the quantity of arbitrary variables under thought, and can be partitioned into highlight choice and highlight extraction. As dimensionality expands, question execution in the record structures debases. Some well known and generally utilized information mining clustering strategies, for example, various leveled and k-means clustering systems are measurable methods and can be connected on high dimensional datasets. High dimensional information are frequently changed into lower dimensional information by means of the central segment investigation. The principle premise of PCA-based dimension reduction is that PCA gets the dimensions with the largest variances.

Principal component analysis (PCA) is a broadly utilized measurable method for unsupervised dimension reduction.K-means clustering is a regularly utilized information clustering for unsupervised learning tasks. Here we demonstrate that key segments are the nonstop answers for the discrete bunch participation pointers for K-means grouping. Bunching is gathering specimens base on their comparability as tests in distinctive gatherings ought to be unique. Both likeness and disparity should be clarified in clear way. High dimensionality is one of the real causes in information many-sided quality. Innovation makes it conceivable to consequently get an immense measure of estimations.

In this paper, we propose to joining relational meaning of clustering with measurement decrease system to overcome previously stated troubles and enhancing proficiency and exactness in K-Means algorithm to apply in high dimensional datasets which give answer for issues, for instance, high dimensionality and flexibility associated with existing methodologies of mining web gives an account of the web. K-Means clustering figuring is joined with reduced datasets which is done by fundamental portion measurement decrease strategy.

## II RELATED WORK

Ruma et al proposed a partitioning approach to cluster the Web-page based on information provided by the hyperlink structure of Web-pages and also by the content of the Web-pages. The proposed approach of Web-page clustering exhibits better result than K-medoid partitioning clustering approach as the centroids are chosen by HITS Algorithm. The partitioning approach like K-mediod, K-means require number of clusters. It has been observed that the performance of these approaches depend on the initial selection cluster centroids.

Durga et al proposed an algorithm for clustering unstructured text documents using native Bayesian concept and shape-pattern matching. The Vector Space Model is used to represent their dataset as a term-weight matrix. In any natural language, semantically linked terms tend to co-occur in documents. Hence, the co-occurrences of pairs of terms in the term-weight matrix are observed. This information is used to build a term-cluster matrix where each term may belong to multiple clusters. The native Bayesian concept is used to uniquely assign each term to a single term-cluster. The documents are assigned to clusters using mean computations. They applied shape pattern-matching to group documents within the broad clusters obtained earlier.

Kun Yang et al proposed a hierarchy clustering method to organize web resources in hierarchy by making full use of the linkage relationship between the web resources and their annotations. In this way, users can browse or search a web resource collection in a Google map’s way by going deep into a collection with more details. Some experiments on data set from Deli.icio.us are performed to justify the effectiveness of their method.

Wai-Tat Fu et al developed user models of knowledge exploration in a social tagging system to test the expertise rankings generated by a link-structure method and a semantic-structure method. The link-structure method assumed a referential definition of expertise, in which experts were users who tagged resources that were frequently tagged by other experts; the semantic-structure method assumed a representational definition of expertise, in which experts were users who had better knowledge of a particular domain and were better at assigning distinctive tags associated with certain domain-specific resources.

Sanghyun Ryu et al proposed an agent based recommendation model that can reduce analysis time when new users or new services appear in the system and recommend more user centric services. Proposed model clusters existing users by using decision tree and analyzes new incoming users by traversing the decision tree, which has already been constructed into the structure that reduces the analysis time. To prove the effectiveness of the proposed model, they implemented user clustering and service recommendation scheme using decision tree.

Caimei Lu et al investigated how to enhance web clustering by leveraging the tripartite network of social tagging systems. They proposed a clustering method, called “Tripartite Clustering”, which cluster the three types of nodes (resources, users and tags) simultaneously based on the links in the social tagging network. The proposed method is experimented on a real-world social tagging dataset sampled from del.icio.us. They also compared the proposed clustering approach with K-means. All the clustering results are evaluated against a human-maintained web directory. The experimental results show that Tripartite Clustering significantly outperforms the content-based K-means approach and achieves performance close to that of social annotation-based K-means whereas generating much more useful information.

## III Multi-View K-Means Clustering Algorithm

K-means is a commonly used partitioning based clustering technique that tries to find a user specified number of clusters (k), which are represented by their centroids, by minimizing the square error function. Given a set of numeric objects X and an integer number k, the K-means algorithm searches for a partition of X into k clusters.

The input data points are then allocated to one of the existing clusters according to the square of the Euclidean distance from the clusters, choosing the closest. The mean (centroid) of each cluster is then computed so as to update the cluster center. This update occurs as a result of the change in the membership of each cluster. The processes of re-assigning the input vectors and the update of the cluster centers is repeated until no more change in the value of any of the cluster centers.

The steps for k-mean algorithm are:

**Step 1**: Initialization: choose *K* input vectors (data points) to initialize the clusters.

**Step 2:** Nearest-neighbor search: for each input vector, find the cluster center that is closest, and assign that input vector to the corresponding cluster.

**Step 3:** Re-calculate the mean of the input vector.

**Step 4:** Mean update: update the cluster centers in each cluster using the mean (centroid) of the input vectors assigned to that cluster.

**Step 5:** Stopping rule: repeat steps 3 and 4 until no more change in the value of the means.

### IV. Hybrid-Fast Multi-View Clustering

**A**s one of the most effective clustering algorithms Multi-View clustering algorithm is applied to large-scale data clustering, to cluster multi-view data, we proposed a new hybrid multi-view clustering (HFMVC) method.

**Multi-View Clustering Based Reformulation:**

**In** previous work shows that A-Orthogonal Non-negative matrix factorization (ANMF) it is equivalent to related multi-view clustering indicated as

$$\text{Min } F,A \sum_{F} (M^T-AF^T)^2$$

$$Aik \text{ €}\{0,1\},\sum_{k=1}^{\mathbf{k}} Aik=1,\text{¥}i=1,2,3,----n$$

Where M€R ^DXN is the input data matrix,with n images and D dimensional visual features, A€R^D*K is the cluster centroid matrix, A€R^n*k is cluster assignment matrix each row of A Satisfies I-of-K coding in this paper a matrix M having its X rows and Y columns denoted .Ri, Rj Respectually.

<u>**Algorithm-1:**</u> Applying **PCA** to minimize dimension of data set

**Step 1**: Maintain the dataset in a matrix **M.**

**Step 2:** Normalize the given data set using **Z-**score.

**Step 3:** Now, calculate the singular value decomposition of the matrix**. M =EV T**

**Step 4:** Calculate the variance using the diagonal elements of **E(**data set **Elements).**

**Step 5:** Sort all variances in decreasing order**.**

**Step 6:** Choose the **n** principal components from V which has largest variances.

**Step 7:** Form the transformation matrix **R** consisting of those **nPCs.**

**Step 8:** Find the reduced projected dataset **N** in a new coordinate axis by applying **R to M.**

**<u>Algorithm-2</u>**: Find the initial centroids

**Step 1:** For a data set with dimensionality, *D*, compute the variance of data in each dimension(column).

**Step 2:** Find the column with maximum variance and call it as max and sort it in any order.

**Step 3:** Divide the data points into *K* subsets, where *K* is the desired number of clusters.

**Step 4:** Find the median of each subset.

**Step 5:** Use the corresponding data points (vectors) for each median to initialize the cluster centers.

## Performance

In this paper, the proposed system works on several datasets to show the better performance compare with the other clustering algorithms.

The hybrid algorithm works for huge multi view data. This calculates a set of medians extracted from the dimension with maximum variance to initialize clusters of the k means . The method can give better results when applied to k-means.

| Data Sets | K-Mean (in sec) | Proposed System (in sec) |
|---|---|---|
| Iris | 0.4532 | 0.4121 |
| Cancer | 0.2543 | 0.986 |
| Artificial | 0.6789 | 0.5432 |

**Figure 1: Comparison table**

Here in above table Prove the performance of our algorithm we used data sets like Iris, cancer and artificial. Artificial datasets were generated from a multivariate normal distribution.

**Conclusion:**

In this paper, dimensionality reduction through PCA, is applied to *kmeans* algorithm. We propose a new hybrid algorithm to initialize the clusters which is then applied to k-means algorithm. We propose to combine relational definition of clustering with dimension reduction method to overcome aforesaid difficulties and improving efficiency and accuracy in K-Means algorithm to apply in high dimensional datasets which provide solution to problems such as high dimensionality and scalability associated with existing techniques of mining web documents on the web.

**References**

[1]A. Moth'd Belal, "A New Algorithm for Cluster Initialization". Proceedings of World Academy of Science, Engineering and Technology. Vol. 4 , pp. 74-76. 2005.

[2] E. Diday," The Dynamic Cluster Method in Non-Hierarchical Clustering". Journal of Computer Information Science. Vol. 2, pp. 61- 88. 1973.

**[3]** Yan Jun, Zhang Benyu, Liu Ning, Yan Shuicheng, Cheng Qiansheng, Fan Weiguo, Yang Qiang, Xi Wensi, and Chen Zheng,2006. Effective and efficient dimensionality reduction for large-scale and streaming data preprocessing, *IEEE transactions on Knowledge and Data Engineering*, Vol. 18, No. 3, pp. 320- 333

[4] O. Zamir, O. Etzioni, " Web Document Clustering" Department of Computer Science and

Engineering, University of Washington, Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 46-54.1998.