# An Efficient Clustering Process using Optimized C Means Algorithm in Social Media Data

Aratakatla Hari Kusuma[1], P. Mohana Roopa[2]
*Final M.Sc Student[1], Lecturer[2]*
*[1, 2] M. Sc Computer Science, Chaitanya Women's PG College, Old Gajuwaka, Visakhapatnam Andhra Pradesh*

**Abstract:**
*Now a day's social media place an important role for sharing human social behaviour's and participation of multi users in the network. The social media will create opportunity for study human social behaviour to analyse large amount of data streams. In this social media one of the interesting problems is users will introduce some issues and discuss those issues in the social media. So that those discuss will contain positive or negative attitudes of each user in the social network. By taking those problems we can consider formal interpretation social media logs and also take the sharing of information that can spread person to person in the social media. Once the social media of user information is parsed in the network and identified relationship of network can be applied group of different types of data mining techniques. However, the appropriate granularity of user communities and their behaviour is hardly captured by existing methods. In this paper we are proposed optimized fuzzy means cluster distance algorithm for grouping related information. By implementing this algorithm we can get best group result and also reduce time complexity for generating cluster groups. The main goal of our proposed framework is twofold for overcome existing problems. By implementing our approach will be very scalable and optimized for real time clustering of social media.*

**Keywords:** *Clustering, social media, k means algorithm, Manhattan distance, tweeter server, data mining.*

## I. INTRODUCTION

Clustering is the process of partitioning or grouping a given set of patterns into disjoint clusters. This is done such that patterns in the same cluster are alike and patterns belonging to two different clusters are different. Clustering has been a widely studied problem in a variety of application domains including neural networks, AI, and statistics. Data clustering is considered an interesting approach for finding similarities in data and putting similar data into groups. Clustering partitions a data set into several groups such that the similarity within a group is larger than that among groups . The idea of data grouping, or clustering, is simple in its nature and is close to the human way of thinking; whenever we are presented with a large amount of data, we usually tend to summarize this huge number of data into a small number of groups or categories in order to further facilitate its analysis. Moreover, most of the data collected in many problems seem to have some inherent properties that lend themselves to natural groupings. Nevertheless, finding these groupings or trying to categorize the data is not a simple task for humans unless the data is of low dimensionality (two or three dimensions at maximum.) This is why some methods in soft computing have been proposed to solve this kind of problem. Those methods are called "Data Clustering Methods" and they are the subject of this paper. Clustering algorithms are used extensively not only to organize and categorize data, but are also useful for data compression and model construction. By finding similarities in data, one can represent similar data with fewer symbols for example. Also if we can find groups of data, we can build a model of the problem based on those groupings.

As mentioned earlier, data clustering is concerned with the partitioning of a data set into several groups such that the similarity within a group is larger than that among groups. This implies that the data set to be partitioned has to have an inherent grouping to some extent; otherwise if the data is uniformly distributed, trying to find clusters of data will fail, or will lead to artificially introduced partitions. Another problem that may arise is the overlapping of data groups. Overlapping groupings sometimes reduce the efficiency of the clustering method, and this reduction is proportional to the amount of overlap between groupings. Usually the techniques presented in this paper are used in conjunction with other sophisticated neural or fuzzy models. In particular, most of these techniques can be used as pre-processors for determining the initial locations for radial basis functions or fuzzy if then rules. The common approach of all the clustering techniques presented here is to find cluster centers that will represent each cluster. A cluster center is a way to tell where the heart of each cluster is located, so that later when presented with an input vector, the system can tell which cluster this vector belongs to

by measuring a similarity metric between the input vector and al the cluster centers, and determining which cluster is the nearest or most similar one. Some of the clustering techniques rely on knowing the number of clusters . In that case the algorithm tries to partition the data into the given number of clusters. K-means and Fuzzy C-means clustering are of that type. In other cases it is not necessary to have the number of clusters known from the beginning; instead the algorithm starts by finding the first large cluster, and then goes to find the second, and so on. how ever if the number of clusters is not known, K-means and Fuzzy C-means clustering cannot be used. Another aspect of clustering algorithms is their ability to be implemented in on-line or offline mode. On-line clustering is a process in which each input vector is used to update the cluster centers according to this vector position. The system in this case learns where the cluster centers are by introducing new input every time. In off-line mode, the system is presented with a training data set, which is used to find the cluster centers by analysing all the input vectors in the training set. Once the cluster centers are found they are fixed, and they are used later to classify new input vectors. The techniques presented here are of the off-line type. A brief overview of the four techniques is presented here. Full detailed discussion will follow in the next section.

The first technique is K-means clustering (or Hard C-means clustering, as compared to Fuzzy C-means clustering.) The k-means method has been shown to be effective in producing good clustering results for many practical applications. However, a direct algorithm of k-means method requires time proportional to the product of number of patterns and number of clusters per iteration. This is computationally very expensive especially for large datasets. We propose a novel algorithm for implementing the k means method. Our algorithm produces the same or comparable (due to the round-off errors) clustering results to the direct k-means algorithm. It has significantly superior performance than the direct k-means algorithm in most cases. This technique has been applied to a variety of areas, including image and speech data compression data pre-processing for system modelling using radial basis function networks, and task decomposition in heterogeneous neural network architectures. This algorithm relies on finding cluster centers by trying to minimize a cost function of dissimilarity (or distance) measure. The second technique is Fuzzy C-means clustering, which was proposed by Bezdek in 1973 as an improvement over earlier Hard C means clustering. In this technique each data point belongs to a cluster to a degree specified by a membership grade. As in K-means clustering, Fuzzy C-means clustering relies on minimizing a cost function of dissimilarity measure

## II. RELATED WORK

Clustering is a fundamental form of data analysis that is applied in a wide variety of domains, from astronomy to zoology. With the radical increase in the amount of data collected in recent years, the use of clustering has expanded even further, to applications such as personalization and targeted advertising. Clustering is now a core component of interactive systems that collect information on millions of users on a daily basis . The ultimate aim of the clustering is to provide a grouping of similar records. Clustering is a technique in which, the information that is logically similar is physically stored together. In order to increase the efficiency of search and the retrieval in database management, the number of disk accesses is to be minimized. In clustering, since the objects of similar properties are placed in one class of objects, a single access to the disk can retrieve the entire class. However, with the never ending data in today's era it is becoming impractical to store all relevant information in memory at the same time, often necessitating the transition to incremental methods called Incremental Clustering.

Clustering problems arise in many different applications, such as data mining and knowledge discovery, data compression and vector quantization, and pattern recognition and pattern classification. The notion of what constitutes a good cluster depends on the application and there are many methods for finding clusters subject to various criteria, both ad hoc and systematic. These include approaches based on splitting and merging such as ISODATA, randomized approaches such as CLARA, CLARANS, methods based on neural nets , and methods designed to scale to large databases, including DBSCAN BIRCH, and ScaleKM . For further information on clustering and clustering algorithms, see . Among clustering formulations that are based on minimizing a formal objective function, perhaps the most widely used and studied is k-means clustering. Given a set of n data points in real d-dimensional space, Rd, and an integer k, the problem is to determine a set of k points in Rd, called centers, so as to minimize the mean squared distance from each data point to its nearest center. This measure is often called the squared-error distortion and this type of clustering falls into the general category of variance based clustering. Clustering based on k-means is closely related to a number of other clustering and location problems. These include the Euclidean k-medians (or the multisource Weber problem) in which the objective is to minimize the sum of distances to the nearest center and the geometric k-center problem in which the objective is to minimize the maximum distance from every point to its closest center. There are no efficient solutions known to any of these problems and some formulations are NP-hard. An asymptotically

efficient approximation for the k-means clustering problem has been presented by Matousek , but the large constant factors suggest that it is not a good candidate for practical implementation. One of the most popular heuristics for solving the k-means problem is based on a simple iterative scheme for finding a locally minimal solution. This algorithm is often called the k-means algorithm. There are a number of variants to this algorithm, so, to clarify which version we are using, we will refer to it as Lloyd's algorithm. (More accurately, it should be called the generalized Lloyd's algorithm since Lloyd's original result was for scalar data.

Fast and robust clustering algorithms play an important role in extracting useful information in large databases. The aim of cluster analysis is to partition a set of N object into C clusters such that objects within cluster should be similar to each other and objects in different clusters are should be dissimilar with each other. Clustering can be used to quantize the available data, to extract a set of cluster prototypes for the compact representation of the dataset, into homogeneous subsets. Clustering is a mathematical tool that attempts to discover structures or certain patterns in a dataset, where the objects inside each cluster show a certain degree of similarity. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Cluster analysis is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization. It will often necessary to modify pre-processing and parameter until the result achieves the desired properties. In Clustering, one of the most widely used algorithms is fuzzy clustering algorithms. Fuzzy set theory was first proposed by Zadeh in 1965 & it gave an idea of uncertainty of belonging which was described by a membership function. The use of fuzzy set provides imprecise class membership function. Applications of fuzzy set theory in cluster analysis were early proposed in the work of Bellman, Zadeh, and Ruspini This paper opens door step of fuzzy clustering . Integration of fuzzy logic with data mining techniques has become one of the key constituents of soft computing in handling challenges posed by massive collections of natural data. The central idea in fuzzy clustering is the non-unique partitioning of the data into a collection of clusters. The data points are assigned membership values for each of the clusters and fuzzy clustering algorithm allow the clusters to grow into their natural shapes.

## III. PROPOSED SYSTEM

The amount of information shared on online social media has been growing during recent years. Much can be learned about the retail and finance behaviours of users by studying social media analysis. It is nothing new that retail companies market via social networks to discover what consumers think about branding, customer relationship management, and other strategies including risk prevention. A good example is the found correlation of data on Twitter with industry market behaviour and sentiment posted by users. Social network analysis has a well-defined relation and background in sociology. With the rapid growth of the web forums and blogs, the user's participation on content creation led to a huge amount of dataset. Hence the advancement of data mining techniques is required. An overall discussion of one news forum called Slashdot, can be found in Social networks, it focus work like face pager. It is used to access data from social media like Facebook by using this data to develop a clustering framework using optimized fuzzy means cluster distance algorithm that is more accurate than existing methods. Clustering is used as an exploratory analysis tool that aims at categorizing objects into categories, so the association degree between the objects is maximal when belonging to the same categories. Clustering structures the data into a collection of objects that are similar or dissimilar and is considered an unsupervised learning. The application of our method is mainly on finding user groups based on activities and attitude features as suggested in the authority model.

The standard k-means algorithm takes extra time in calculating distance from each cluster's center in each iteration. The implementation process of k means algorithm is as follows.

1. Read the twitter data set from the twitter server.

2. Enter number of clusters to be performing and randomly choose the centroids from twitter dataset.

3. Take each data point ($d_i$) from dataset and calculate the Manhattan distance from data point to centroids' ($c_i$).

$$\text{Distance} = (c_i - d_i)$$

4. If check the closet distance of each centroid from the data point and that data points will be put into those clusters.

5. The step 3 and 4 will be repeated until there is no change in the centroids.

6. After completion of step 6 we can get group of clustered data.

7. The calculation of Manhattan distance we can also calculate each cluster sum squared error by using following equation.

$$SSE = \sum_{i=1}^{n} dis\ (c_i, di)$$

By implementing this algorithm will take time complexity and space complexity. This extra time can be saved by adapting this method. The implementation process of optimized fuzzy means cluster distance algorithm is as follows:

**Optimized Fuzzy Means Cluster Distance Algorithm:**

**Input:**

The number of desired clusters, k, and a dataset D= (d1, d2,……., dn) containing n data objects.

**Output:**

A set of k clusters.

Steps:

1) Randomly select k data objects from dataset D as initial clusters.

2) Calculate the matched words between each data object $d_i$ (1<=i<=n) and each cluster center $c_j$ (1<=j<=k) .

3) After completion of matched word we can find out sum squared error by using following formula.

$$SSE=1/w^2$$

4) Calculate total number of words in a data point and centroid find out weight of each data points to centroid. The calculation of weight each twit is as follows.

$$\text{Weight } (W_i) = 1/\text{dist } (d_i, ci)^2 / \sum_{q=1}^{k} 1/\text{dist } (C_i, di)$$

5) After completion of weight of each data point to centroids check which data point is near by the centroids.

6) For every cluster center cj (1<=j<=k), it compute the weight of data points d (di, cj) and assign the data object di to the nearest cluster.

Set cluster[i] = j;

Set w[i] = d (di, cj).

7) For each cluster center j (1<=j<=k), recalculate the centers;

9) Until the center is same.

10) Output the clustering result.

The optimized fuzzy means cluster distance algorithm is used to reduce time complexity and also space complexity of data objects.. This paper does not require calculating distance in each iteration. The time complexity of this algorithm is O (nk). If a data point remains in its initial cluster then the time complexity will be O (1) otherwise O (k). If half of the data points move from its initial cluster then the time complexity will be (nk/2). So the proposed algorithm effectively increases the speed of standard k-means algorithm. But this algorithm also requires the value of k in advance. If one wants the optimal solution then he must test for different values of k.

## IV. CONCLUSIONS

This paper we are proposed an efficient clustering algorithm for reduce the time complexity and space complexity. This paper proposes optimized fuzzy means cluster distance algorithm for getting better cluster result in data set. By implementing this process we can easily find out similar data object in data set by calculating weight of each data object to centroids. The calculation of weight of data object will repeat until the no changes occur in the centroids. By applying this process we can reduce number of iteration compared to existing algorithm of k means. So that each data point from each cluster center in each iteration due to which running time of algorithm is saved. By implementing proposed system we can efficiently improve speed of the clustering and accuracy by reducing the computational complexity of standard k-means algorithm.

## REFERENCES


[1] Andreas M Kaplan and Michael Haenlein. Users of the world, unite! the challenges and opportunities of social media. Business horizons, 53(1):59–68, 2010.

[2] Bogdan Batrinca and Philip C Treleaven. Social media analytics: a survey of techniques, tools and platforms. AI & SOCIETY, 30(1):89– 116, 2015.

[3] David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Life in the network: the coming age of computational social science. Science (New York, NY), 323(5915):721, 2009.

[4] Claudio Cioffi-Revilla. Computational social science. Wiley Interdisciplinary Reviews: Computational Statistics, 2(3):259–271, 2010.

[5] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In Proceedings of the 19th international conference on World Wide Web, pages 591–600. ACM, 2010.

[6] Michael D Conover, Clayton Davis, Emilio Ferrara, Karissa McKelvey, Filippo Menczer, and Alessandro Flammini. The geospatial characteristics of a social movemen communication network. PloS one, 8(3):e55957, 2013.

[7] Bruce A. Maxwell, Frederic L. Pryor, Casey Smith, "Cluster analysis in cross-cultural research" World Cultures 13(1): 22-38, 2002.

[8] Kiri Wagstaff and Claire Cardie Department of computer science, Cornell University, USA "Constrained k- means algorithm with background knowledge".

[9] Thomas H. Cormen, Charles E. Leiserson, and Ronald L. Rivest, Introduction to Algorithms, Prentice Hall, 1990.

[10] Anil K. Jain, M. N. Murty, P. J. Flynn, "Data Clustering: A Review," ACM Computing Surveys, 31(3): 264-323 (1999).

[11] Bogdan Batrinca and Philip C Treleaven. Social media analytics: a survey of techniques, tools and platforms. AI & SOCIETY, 30(1):89– 116, 2015.

[12] Emilio Ferrara, Mohsen JafariAsbagh, Onur Varol, Vahed Qazvinian, Filippo Menczer, and Alessandro Flammini. Clustering memes in social media. In Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pages 548–555. IEEE, 2013.