|| Volume 2 ||Issue 11 ||JUNE 2017||ISSN (Online) 2456-0774

INTERNATIONAL JOURNAL OF ADVANCE SCIENTIFIC RESEARCH

AND ENGINEERING TRENDS

A Comparative Study of K-means and K-medoid Clustering for Social Media Text Mining

Shivani Vishwakarma[1], Dr. Pramod S Nair[2], D. Srinivasa Rao[3]

Research Scholar, Department of CSE, MITM, Indore, M.P, India[1]
Professor, Department of CSE, MITM, Indore, M.P, India[2]
Associate Professor, Department of CSE, MITM, Indore, M.P, India[3]

Abstract— **The text mining is a technique for analyzing text data or unstructured data using data mining algorithms. Data mining algorithms works on the basis of two different manners namely supervised and unsupervised. In this context, when the data is available in pre-defined patterns and need to find the similar kinds of pattern data, then the supervised approaches such as classification techniques are used. On the other hand when the data pattern is not available on a specific manner then the unsupervised approaches are used such as clustering. The unsupervised learning techniques works on the basis of internal similarity of data objects and on the basis of this data clustered. Nowadays, in number of places the unsupervised learning approaches are used for recovering the data text patterns, trends, similarities and other kinds of patterns. In this proposed work, the social media text analysis is the main motive of the work using the unsupervised learning techniques. Therefore, two popular algorithms which are slightly different from each other are used for finding the best performing algorithm for text mining. Thus the k-means clustering and k-medoid clustering algorithms are used for analyzing text data and performing clustering. The proposed work of analyzing text data using clustering approach includes data pre-processing, feature selection and clustering. During pre-processing the data is refined and filtered for finding the required data, in next phase during the feature selection the word frequency is computed and data is transformed in 2D vector. Finally using the implemented algorithms the data is processed. The implementation of proposed technique is provided using JAVA technology. Additionally, the performance in terms of inter cluster similarity is measured. According to experimental evaluation K-medoid clustering is efficient in terms of time space complexity and able to prepare the uniformly distributed clusters of data.**

Keywords- *social media text mining, unsupervised learning, clustering, k-means, k-medoid*

I INTRODUCTION

Social media such as blogs, message boards, micro-blogging services and social networking sites have grown significantly in popularity in recent years. By lowering the barriers to online communication, social media enables users to easily access and share content, news, opinions, and information in general [1 However navigating this information can be problematic due to the fact that contributions are often much shorter than a typical Web document, and the quality of content in social media is highly variable [2]. As the data present on social media is huge so millions of people are turning to micro-blogging services like Twitter to gather real-time news or express their views on various topics. Such services are used for social networking to stay in touch with friends and colleagues. In addition, micro-blogging sites are used as publishing platforms to create and consume content from sets of users with overlapping and disparate interests. As micro-blogging are growing in popularity so services like Twitter are coming to support information gathering needs above and beyond their traditional roles as social networks [3].

In this proposed work, the text data for Twitter analysis is used for preparing the clustering algorithm. Basically, Twitter is frequently used now in these days with the small amount of communication data. Hence, applications using social media data, such as reviews, discussion posts, and (micro) blogs are becoming increasingly popular. The developed system is a Social Media based topic categorization technique that works on the given label data and provides the outcomes for clustered centroid. First, the data is processed in order to clustering basis and then the learning on evaluated data is performed.

II PROPOSED WORK

The aim of the initiated research work is to find the optimal technique of text clustering for social media text analysis. Therefore this section includes the description of the proposed comparative performance study and understanding of the implemented algorithms.

A. Proposed System

Data mining techniques are applied to text data to perform text mining. Therefore, text mining is an approach of obtaining essential patterns from the text data or unstructured data. The mining approaches can be divided into two major classes namely supervised and unsupervised. The supervised approaches first take training on the predefined data patterns and then are utilized with the applications. On the other hand, the unsupervised learning approaches directly work on the input data to recover different hidden patterns. In this context, two unsupervised learning approaches are applied on text data to perform clustering and obtaining the performance study among both the algorithms. Basically now in these days, the use of Web-based applications is increasing significantly. In addition of that, social media applications are in higher demand. A number of new and young people are utilizing their time on various social media sites. During the social media site usage, the data and text are shared on this platform. The analysis of text data on this platform provides the important information about the trending topics and subjects on which the traffic is attracted. These are always the subject of interest for various webmasters and web managers.

In this proposed work for performing this task, two different unsupervised clustering algorithms are selected namely k-means and k-medoid. In addition of that, by applying the various data mining techniques and pre-processing methods the data is converted to such format that helps to perform the efficient and accurate text mining. In addition of that, it is also required to investigate which algorithm is providing the efficient and accurate results when these are applied on text clustering applications. This section provides the core concept involved in the proposed work. Additionally, the further section helps to understand the implemented system by which the proposed comparative study is performed.

B. Algorithm Study

This section provides the study of the algorithm which is applied for performing the proposed comparative study.

a. K-means Algorithm

The K-Means clustering algorithm is a partition-based cluster analysis method [4]. According to the algorithm, we first select k objects as initial cluster centers, then calculate the distance between each object and each cluster center and assign it to the nearest center, take the averages of all clusters and repeat this process till the criterion function is converged. Square error criterion for clustering:

$$E = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \| x_{ij} - m_i \|^2$$

$x_{ij}$is the sample j of i-class,$m_i$is the center of i-class,n_i i is the number of samples of i-class. Table 1 shows the steps of K-mean clustering.

Table 1 K-mean algorithm steps

Input: N objects to be cluster x_j, x_z, x_n, the number of clusters k;
Output: k clusters and the sum of dissimilarity between each object and its nearest cluster center is the smallest;
Process: 1. Arbitrarily select k objects as initial cluster centers$(m_1, m_2, \dots, m_k)$; 2. Calculate the distance between each object Xi and each cluster center, then assign each object to the nearest cluster, formula for calculating distance as: $d(x_i, m_i) = \sqrt{\sum_{j=1}^{d} (x_i - m_{j1})^2}, i = 1 \dots N, j = 1 \dots k$ $d(x_i, m_i)$is the distance between data i and cluster j. 3. Calculate the mean of objects in each cluster as the new cluster centers, $m_i = \frac{1}{N} \sum_{j-1}^{n_i} x_{ij}, i = 1,2, \dots, K$ N_i is the number of samples of current cluster i; 4. Repeat 2) 3) until the criterion function E converged, return$(m_1, m_2, \dots, m_k)$Algorithm terminates.

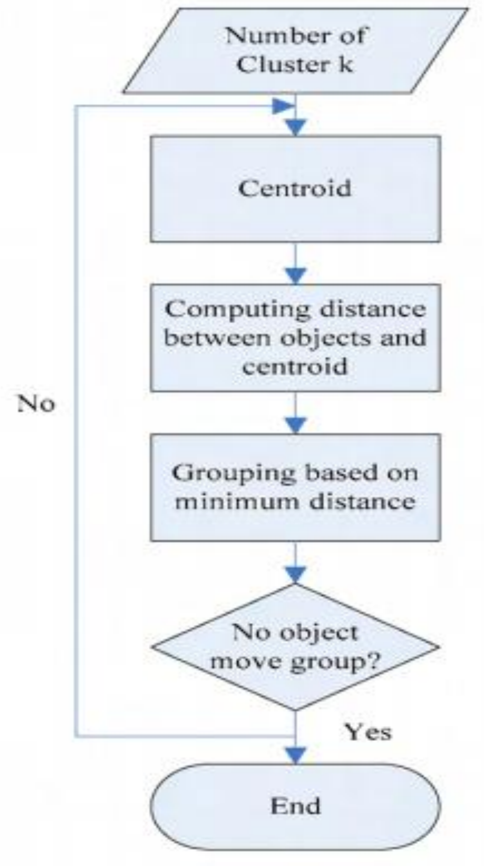
Number of Cluster k
Centroid
Computing distance between objects and centroid
Grouping based on minimum distance
No object move group?
No
Yes
End

Figure 1 K-mean clustering algorithm

b. K-Medoid Algorithm

K-Medoid clustering is also a partition-based clustering algorithm. It uses medoids to represent the clusters. A medoid represents the most centrally located data item of the data set. In medoid, the data member of a data set whose average dissimilarity to all the other members of the set is minimal.

The working of K-Medoids Clustering algorithm is similar to K-Means clustering. It also begins with randomly selecting k data items as initial medoids to represent the k clusters. All the remaining items which are closest to the medoid are included in a cluster. Thereafter, a new medoid is selected which can represent the cluster better. Again the items are assigned to the clusters having closest medoid. In each iteration, the medoids alter their location. The method minimizes the sum of the dissimilarities between each data item and its corresponding medoid. This cycle is repeated till no medoid changes its placement. Here the process ends and we have the resultant final clusters with their medoids defined. K clusters are formed which are centroid on the medoids and all the data members are placed in the appropriate cluster based on nearest medoid [5]. Table 1 shows the steps of K-medoid clustering.

Table 2 K-medoid

Input: number of clusters k, the data set containing n items D

Output: A set of k clusters that minimizes the sum of the dissimilarities of all the objects to their nearest medoids.

$$Z = \sum_{i=1}^{K}\sum |X - m_i|$$

Where, Z is Sum of absolute error for all items in the data set, x is the data point in the space representing a data item and m_i is the medoid of cluster C_i

Process:

1. Arbitrarily choose k data items as the initial medoids.
2. Assign each remaining data item to a cluster with the nearest medoid.
3. Randomly select a non-medoid data item and compute the total cost of swapping old medoid data item with the currently selected non-medoid data item.
4. If the total cost of swapping is less than zero, then perform the swap operation to generate the new set of k-medoids.
5. Repeat steps 2, 3 and 4 till the medoids stabilize their locations.

C. Proposed Methodology

The proposed system for performing the comparative performance study of clustering algorithm to mine social media text data is provided in Figure 1. The involved processes are represented using the blocks of the diagram.

Input Text: the proposed comparative study is performed on the text or unstructured data set. Therefore a set of twitter dataset is prepared for experimentation of both the algorithm. The input data is remained fixed and the performance of the algorithms is computed one by one for both the implemented algorithms.

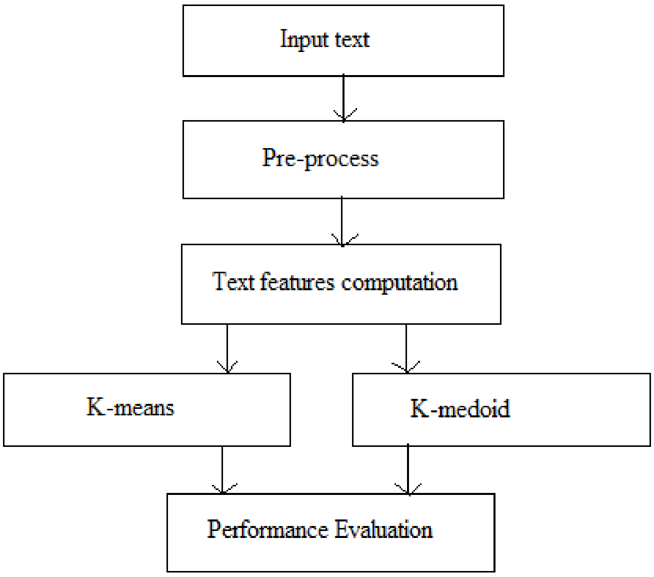

Figure 2 Proposed Methodology

Pre-Process: The pre-processing of data involves three phases of data processing and data recovery which is used with the clustering algorithm for evaluation of performance of both the targeted algorithms.

1. **Recovery of Twits:** The main motive of the pre-processing technique is to reduce or remove the unwanted data from the input set. Therefore the data is refined in this phase for finding the required contents. In addition of that, secondary motive of pre-processing technique is to transform the input data in such manner by which the data becomes acceptable for algorithm processing. Therefore, in this phase the input training data is processed in such manner by which only twits column remain and additional data from the training set is removed.

2. **Removal of Special Character:** After recovering the data from the input training dataset the second phase of pre-processing is applied. In this phase the data is filtered for removing special characters from the input set. Therefore, a function is prepared that accept the list of special characters and replace them using the blank character. After removing the special characters the next step is followed.

3. **Removal of Stop words:** In this phase, the data is processed for removing stop words. The stop words are considered as those words that are frequently occurred in different sentences and much contributing for the domain identification. Therefore, such kinds of words are targeted for removal. In order to perform this task a list of stop words is prepared and then similar function is used for replacing the words.

Text Feature Computation: The text features are the high priority words that are used for identifying the domain of work. In order to estimate the weighted words in text input various techniques are available such as bag of words, TF-IDF and others. In the proposed technique for preparing the vector of data for providing the cluster analysis word frequency of individual words are computed. In addition of that a fixed size two dimensional vector is prepared. That vector is prepared on the basis of higher word frequency. The word frequency is computed using the following formula:

$$word\ frequency = \frac{total\ times\ a\ word\ appeared}{total\ words}$$

K-Means: The traditional K-means clustering as described in table is implemented and the developed 2D vector is provided as input with the number of clusters required to develop. The clustering algorithm executes the data using this algorithm and generates clusters according to user guidelines.

K-Medoid: In the similar manner the K-medoid clustering is implemented and the prepared 2D vector is produced as input to the system. The system accepts the data and generates the number of clusters of data according to the input amount of clusters.

Performance Evaluation: After applying the required clustering algorithms (k-means and k-medoid) over the similar text data features, the performance of the algorithm is evaluated. The performance criteria of the system are Inter cluster similarity, time and space complexity. The inter cluster distance helps to measure the distribution of data in different clusters. Similarly the time and space complexity is a measurement of efficiency of the clustering algorithm.

III RESULT ANALYSIS

After implementation of two different algorithms namely k-mean clustering and K-medoid clustering algorithm for finding the optimal clustering technique for social media text clustering. The performance evaluation of both the clustering algorithms is performed. The experimental result based on the analysis is demonstrated in this section.

A. Inter Cluster Distance

The inter-cluster distance $dis\,(i,j)$ is the distance between two centroids. Inter-Cluster distance computation shows the similar amount of distance as the input dataset instances. The mean inter-distance of each centroid is computed for both the algorithms. To compute the mean inter-cluster distance among cluster points and final centroid the following formula is used.

$$mean\ inter\ cluster\ distance = \frac{1}{N}\sum_{i=1}^{N} dis(C_{j,i}, P_i)$$

Where N is the number of points in a cluster

$C_{j,i}$ is the jth centroid

P_i is the point which is needs to evaluate with respect to $C_{j,i}$

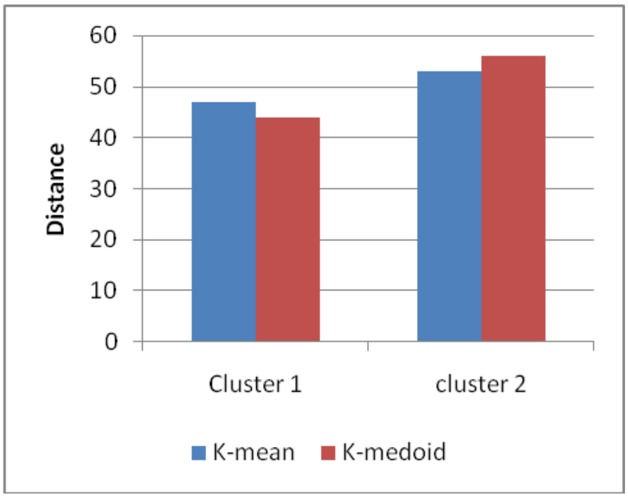

Figure 3 Inter Cluster Distance

And dis is the distance function that computes distance between centroid and points

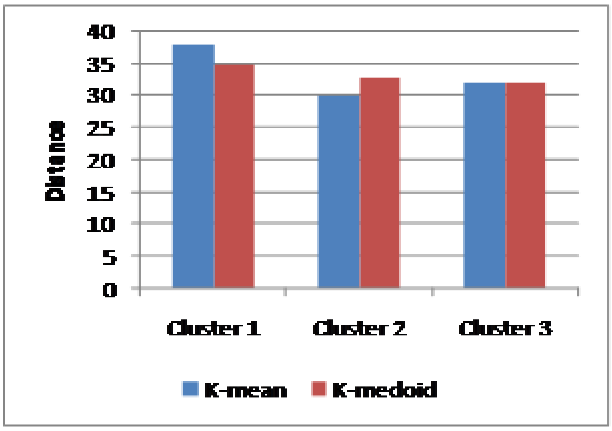

Figure 4 Inter Cluster Distance for three centroids

The experiments are performed on the different amount of data but for the 2, 3 and 4 centroids. Figure 3 shows the inter-cluster distance for the two centroids experimental scenario. In this diagram, the X axis contains the number of clusters and the Y axis contains the distance from the corresponding clusters. Similarly, for three centroids and four centroids, the inter-cluster distance is provided by the figure 4 and figure 5. Figure 4 shows the inter-cluster distance for three cluster points when the k-medoid cluster shows the distributed clusters with respect to the k-means clustering algorithm. According to the obtained results both the clustering approaches produces approximately similar outcomes but sometimes k-medoid outperforms then the k-means clustering. In addition of that, the clustering performance is enhanced as the number of clusters increases. According to the inter-cluster distance, the k-medoid clustering is more acceptable as compared to k-means clustering.

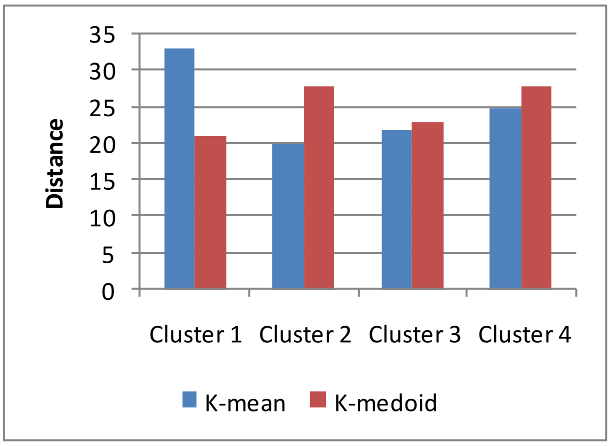

Figure 5 Mean Inter Cluster Distance for four centroids

B. Time Consumption

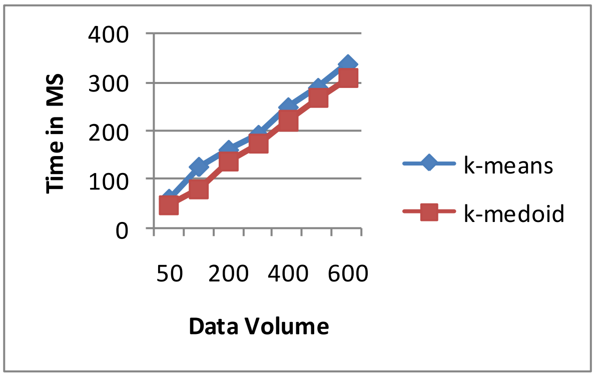

Figure 6 Time Consumption

The time complexity of an algorithm is an additional effective parameter for performance analysis. The time complexity is basically the amount of time which is required to complete the algorithm execution. Different experiments are performed on different size of data; according to the obtained performance on the different volume of data, the performance of the algorithms is demonstrated. Figure 6 shows the time required to perform clustering by both the algorithm. The computed time during these experiments is observed in terms of milliseconds. The graph contains the data volume in the X axis and the corresponding amount of time for clustering is defined in the Y axis. According to the computed time complexity of the system k-means algorithm is taking more time as compared to the proposed k-medoid clustering in terms of millisecond. Meanwhile, the time complexity of both the algorithms increases with the amount of data increase. The performance of k- medoid is much acceptable as compared to k-mean clustering algorithm for time complexity.

C. Space Complexity

In order to perform the computing the data is placed in main memory space. This amount of memory requirement of the algorithms is known as the space complexity or memory consumption of the algorithms.

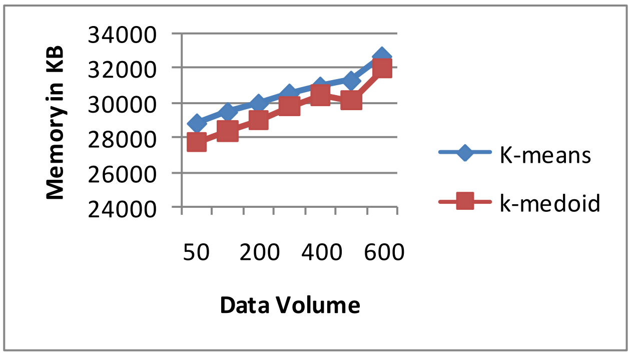

Figure 7 Space Complexity

Figure 7 shows the memory requirements of the both clustering algorithms for text mining. The k-means algorithm's performance is denoted using blue the line and the k-medoid algorithm is represented using red line. In addition of that, for providing the performance, the X axis contains the size of data and Y axis shows the obtained memory consumption of algorithms. The measured performance of both the algorithms is given in terms of kilobytes (KB). According to the experimental results, the memory requirements for both algorithms increases with the amount of data increase in the similar ratio. Additionally, both algorithms are demonstrating similar behaviour for the memory consumption. On the basis of their outcomes, the k-means algorithm consumed more memory space and de-efficient of accurate clustering. Therefore k-medoid is adaptable and performing effective clustering.

IV CONCLUSION AND FUTURE WORK

The proposed work is aimed to find an efficient and accurate clustering technique for social text mining. Therefore the k-means and k-medoid clustering algorithm for text mining are implemented and their performance for twitter text clustering is measured. Based on the observations and experiments the conclusion of the work is made. Additionally, the future extension of the work is also reported in this chapter.

A. Conclusion

Text mining is a branch of data mining where the data mining algorithms are employed for analyzing text data. Basically, the text data is unstructured in nature and there are various complexities to perform the mining such as text file lengths, available data in text files, the weighted text words and the appropriate evaluation of the text data for performing the accurate and efficient clustering. Therefore some additional process is needed to be implemented for obtaining the data in such format by which data mining algorithms becomes implementable. In this proposed work the text mining techniques are applied to the social media text analysis for recovering similar text categories.

The text categorization is also a kind of cluster analysis of data. Therefore clustering techniques of data mining are needed to be implemented. There are the number of clustering techniques available and amongst them, two popular text clustering algorithms k-means and k-medoid algorithms are selected for implementation and performance study. Both the techniques of clustering techniques are a kind of optimization techniques that initially select the random k centroids on the basis of the optimization criteria. During the optimization, the most optimal centroids are tried to find by which the internal distance among the cluster points and centroids are needed to be improved. The key goal of this centroid adjustment is to recover the optimal clustering and uniformly distributed clustering.

The implementation of the proposed comparative study among k-means clustering and k-medoid clustering is performed using JAVA technology. After implementation, the testing of algorithms is performed on twitter based text analysis. The experimental results are computed and compared. Table 3 shows the performance summary.

Table 3 Performance Summary

S. No.	Parameters	k-means	k-medoid
1	Inter cluster distance	Less distributed	Highly distributed
2	Memory space	High	Low
3	Time consumption	High	Low

According to the obtained performance as described in the performance summary table, the comparative study of both the algorithm is performed. Experimental results show the k-medoid algorithm able to perform accurate clustering and also able to provide uniformly distributed clusters. K-medoid algorithm is an efficient algorithm that performs clustering in less amount of time and space complexity.

B. Future Work

The aim of the proposed work is to perform comparative performance study of the clustering algorithm for performing the social media text analysis is accomplished successfully. In near future the following extension is feasible for work:

1. The proposed work considered only two algorithms for performance evaluation. The result can be improved by comparing more clustering algorithms and using the best performed algorithm for the topic categorization.
2. The improvement on the best performed existing algorithm is also another possible future work.

REFERENCES

[1] Agichtein, E., Castillo, C., Donato, D., Gionis, A., Mishne, G.: Finding high-quality content in social media. In: Proceedings of WSDM, 2008.

[2] Ramage, Daniel, Susan T. Dumais, and Daniel J. Liebling, "Characterizing Microblogs with Topic Models", Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media ICWSM 10 (2010).

[3] Johan Bollen, Huina Mao, Xiao-Jun Zeng, “Twitter mood predicts the stock market”, arXiv: 1010.3003v1 [cs.CE] 14 Oct 2010.

[4] Wang, Juntao, and Xiaolong Su. "An improved K-Means clustering algorithm" Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference on IEEE, 2011.

[5] Aruna Bhat, “K-Medoids Clustering Using Partitioning Around Methods For Performaing Face Recognition”, International Journal of Soft Computing, Mathematics and Control (IJSCMC), Vol. 3, No. 3, August 2014.