# An Efficient Approach for Clustering Using Hybrid Algorithm Analysis

**Ms. Rajnee Kanoje[1], Dr. S. D. Choudhari[2]**
MTech. CSE, SBITM COE, Betul, Professor SBITM COE, Betul
**Email -** rajnee03kanoje@gmail.com, choudhari.sachin1986@gmail.com

**Abstract:** Now days, criminals frequently use all latest technologies to commit serious crimes like cracking sites, fraud in different domains, prohibited access etc. Thus, the inquiry of such cases is very difficult and more significant task. So, we need to do the analysis of crime scene data. In digital forensic analysis time factor play very critical role. So it's a not an easy task for investigator to do such complex analysis in very short period of time. This is the main reason we used the digital forensic analysis of documents technique where complex task is accomplished using a simpler approach. Such type of analysis technique includes document clustering. So, clustering algorithms play very important role for efficient results. In this paper we used proposed novel approach to achieve more efficient document clustering in forensic analysis.

**Key Words:** Document Clustering, Forensic Analysis, Investigation, Data Mining.

## 1. INTRODUCTION:

Recently in the world of digital technology especially in the computer world there is tremendous increase in crime like unauthorized access, money laundering etc. So, investigation of such cases is much more important task for that kind of crime investigation that's why we need to do digital forensic analysis.

Digital Forensic analysis is the branch of systematic forensic analysis process for investigation of matter found in digital devices interrelated to computer crimes [1]. Digital evidence equivalent to particular incident is any digital data that provides suggestion about incident. The vital component of digital forensic process is to inspect the documents that present on suspect's computer. Due to huge count of documents and larger size of storage devices makes very difficult to analyze the documents on computer. Typically, digital forensics is the use of examination as well as analysis technique to collect and protect evidence from exacting computing device in a way that is proper for presentation in a court of proceeds. Digital evidence is generally defined as the information and data of examine value that are stored on, transmitted or received by digital device. Such type of digital evidences needs to be collected from computer seized devices in order to confess the case in court of law [2]. So such digital evidences provide an immense benefit for the forensic examiner .So the key factor to improve such forensic analysis process requires improved document clustering method. The digital forensic analysis process is shown using figure 1.1

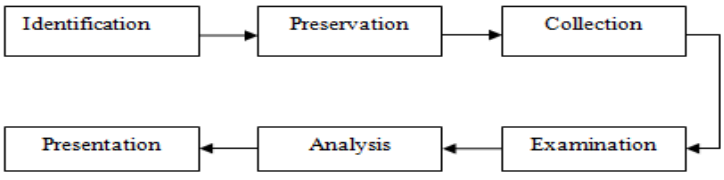



Figure 1.1: Process of Digital Forensic Investigation (DFI)

After identifying items, components, and data related with the unpleasant incident called an Identification phase, the next step is to protect the crime scene by stopping or preventing several actions that may harm digital information being collected can be called preservation phase. Follow that, the next level step is collect digital information that might be related to the incident, for example copying files or recording network traffic (Collection phase). Next step, the investigator conducts an in detail efficient search of evidences related to the incident being analysis such as filter, validation and pattern matching techniques (Examination phase) [1].

Document clustering methods are being studied from many decades but still it is far from a trivial and solved problem. The key challenges are:

1. Finding appropriate features of the documents that should be used for clustering.
2. Applying appropriate similarity measure between documents.
3. Identifying an appropriate clustering method utilizing the above similarity measure.
4. Applying the clustering algorithm in an efficient so that required memory and CPU resources will be optimized.
5. Finding best ways of assessing the quality of the performed clustering.

Additionally, with medium to large document collections (15,000+ documents), the number of term-document relations is fairly high generally millions+, and the computational complexity of the algorithm applied is thus an essential factor to identify feasibility in terms of real-life applications.

## 2. LITERATURE REVIEW:

This section contains the number of scholarly papers that includes the current knowledge as well as the theoretical and methodological contributions to the clustering process and related work to digital forensic analysis.

It is very important to highlight that getting from a collection of documents to a clustering of the collection, is not just a single operation, but there is more a process in multiple phases. These stages include more traditional information retrieval operations such as indexing, weighting, filtering etc. Some of these other processes are central to the quality and performance of most clustering algorithms, and it is thus necessary to consider these phases together with a given clustering algorithm to join its true potential. We will give a brief overview of the clustering process, before we begin our literature study and analysis.

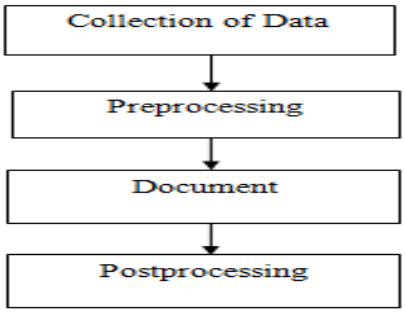



Figure 2.1: The Phases of the Process of Clustering

Y.Zhao etal.[5] focused on document clustering algorithms that build such hierarchical solutions and (i) presents a comprehensive study of partitioned and agglomerative algorithms that use different criterion functions and merging schemes, and (ii) presents a new class of clustering algorithms called constrained agglomerative algorithms, which combine features from both partitioned and agglomerative approaches that allows them to reduce the early-stage errors made by agglomerative methods and hence improve the quality of clustering solutions.C.Charu etal. [6] Studied each object initially represents a cluster of its own. Then clusters are successively merged until the desired cluster structure is obtained. Agglomerative algorithms are usually classified according to the inter-cluster similarity measure they use. The most popular of these are single-link, complete-link and group average.

### 2.1 Problem Statements

- In the existent system the clusters are formed on the basis of exact match in that case examiner have to give exact query for searching relevant information it was time consuming so in our proposed system we overcome this drawback. In this paper we mainly work forming a cluster on the basic of relative match.

- The hybrid approaches for document clustering implemented for getting efficient clustering results and improve speed of forensic analysis process.

## 3. PROPOSED METHODOLOGY AND IMPLEMENTATION:

This section major focus is on the methodologies used and the implementation part of the project. The brief introduction of the methodologies and the algorithms used are given in this paper.

### 3.1. System Architecture

Figure shows system flow of implemented system for digital forensic analysis .In our system we apply most commonly used preprocessing procedures of text mining .The brief discussion of system flow is briefly explain below.

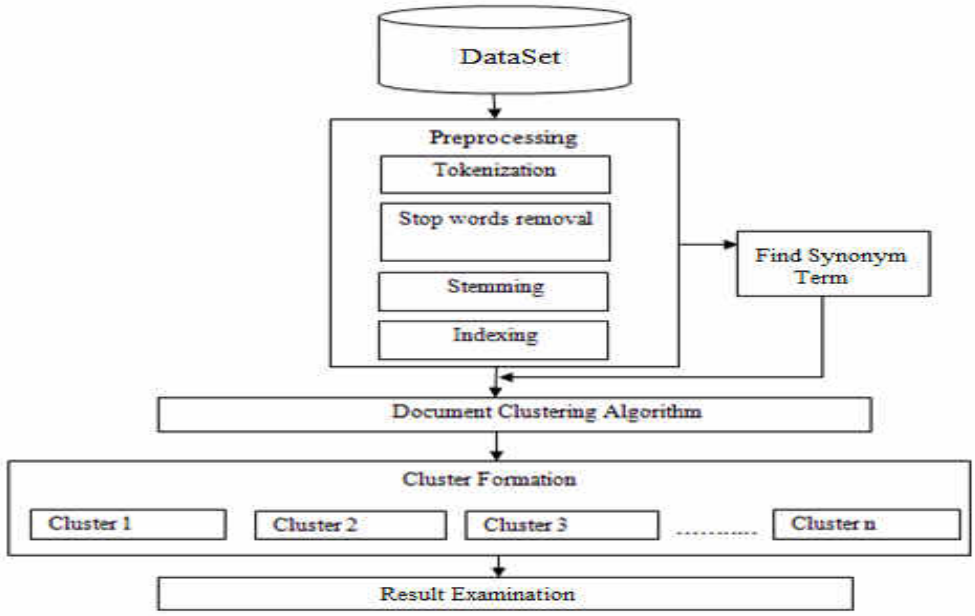



Figure 4.1: System Architecture

### 3.2 Proposed K-representative Algorithm

Let's first observe some special requirements for good document clustering algorithm:

1. The document model should better conserve the relationship between words like synonyms in the documents since there are different words of same meaning.
2. Relate a meaningful label to each final cluster is necessary.
3. The high dimensionality of text documents must be reducing.

So to achieve this feature in our proposed system we enhance approach to improve document clustering in forensic analysis. For that we were implementing hybrid approach to accomplish this proposed approach. We implementing new text clustering algorithm such as K-representative algorithm which will gives us the better clustering result .The main idea of K-representative algorithm is to use the relative attribute frequencies of the clusters mode in the dissimilarity measures in the K-mode objective function [7].

It has been shown that K-representative algorithm is very efficient. Thus the modification implemented in forming representatives for different clusters of categorical objects, the dissimilarity between a categorical object and the representative of a cluster is defined based on simple matching as follows.Let C be a cluster of categorical Objects, with Xi = (xi . . . xi m), 1≤ i ≤p, and X = (x1…. xm) be a categorical object. Assume that Q = (q1….qm), with qi= {(cj, fcj) | cj .Dj}, is a representative of cluster C. Now we define the dissimilarity between object X and representative Q by

$$d(X,Q) = \sum_{j=1}^{m} \sum_{c_j \in Dj} f_{c_j,\delta(x_j,c_j)} \qquad \text{.....(4.1)}$$

### 3.3 Steps of K-representative Algorithm

- Initialization a k-partition of D randomly.
- Calculate k representatives, one for each cluster.
- For each Xi, calculate the dissimilarities d (Xi, Ql), l = 1, .Reassign Xi to cluster Cl (from cluster Cl0, say) such that the dissimilarity between Xi and Ql is least. Update both Ql and Ql0.
- Repeat Step 3 until no object has changed clusters .After a full cycle test of the whole data set

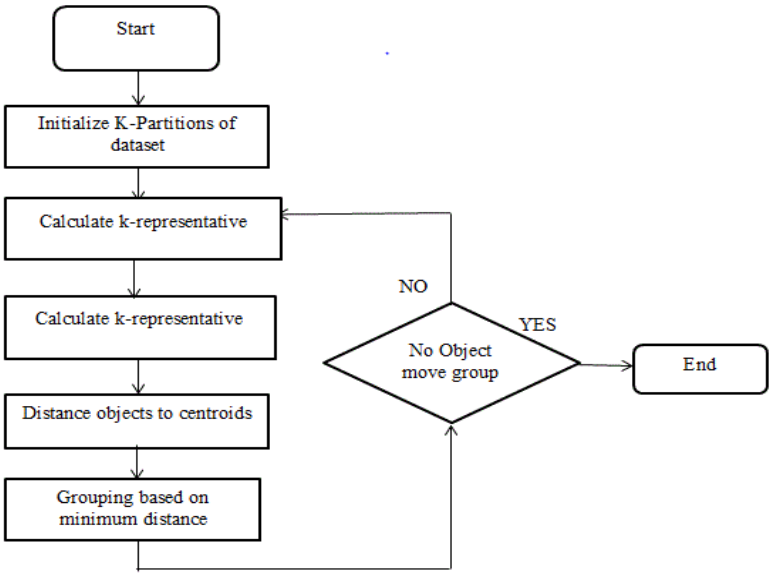
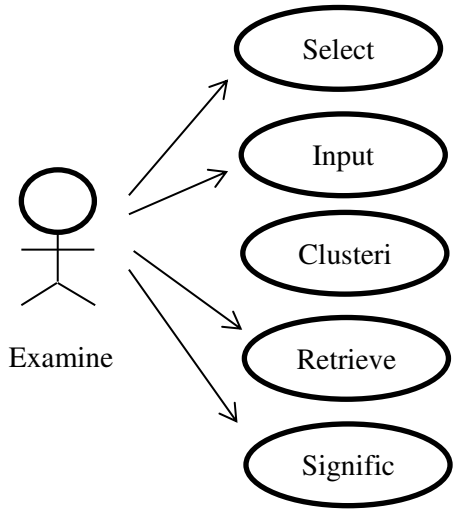



Figure 4.2: Flow chart of K-representative algorithm

Figure shows flow chart of k-representative algorithm in which demonstrates the execution steps of k – representative algorithm. In k-representative algorithm first step is initialization of k-number of clusters randomly after that calculate the K-representative that means centroid of clusters using dissimilarity measures and find distance of object to centroid after that grouping of object to minimum of distance from centroid if no object move to group then repeat the step of find of cluster centers.

### 3.4 Use Case Diagram




The use case diagram for the system is shown in the figure 4.4. The examiner behaves as an actor for this use case diagram. Firstly the examiner selects the dataset for testing system. After the selection of dataset, the examiner gives the files in the system for searching relevant documents according to their requirements Once the query are entered then the clustering process is performed and the clusters are formed. The examiner gets relevant documents which are most significant regarding their need. Then the user logs out from the system.

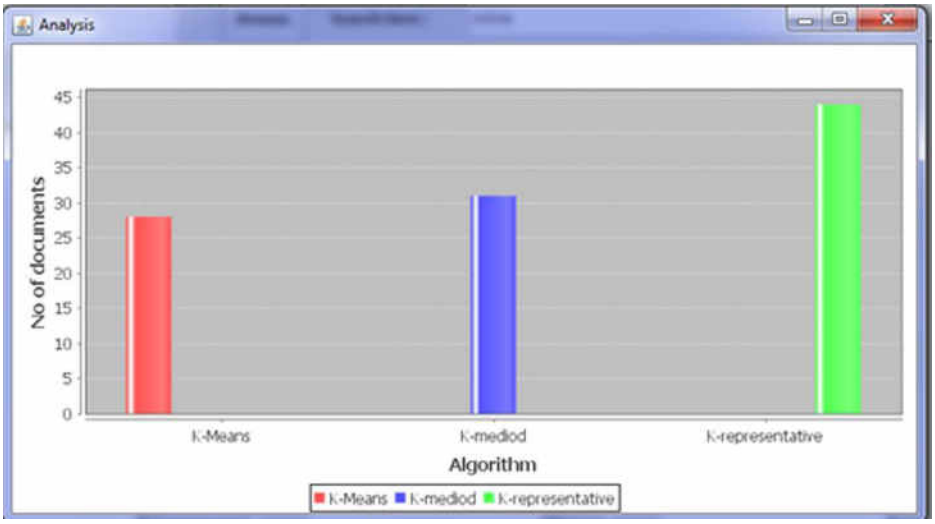



Figure 5.5: Number of Retrieved Documents

Figure 5.5 shows bar chart of no of retrieved documents after applying three clustering algorithms and from that analysis we depicted that which clustering algorithm is efficient. In this analysis red bar shows K-mean algorithm result it retrieved 27 documents, blue bar shows K-medoid algorithm result it retrieved 33 documents and green bar shows K-representative algorithm result it retrieved 44 documents which are related to crime in whole dataset which contain 100 files if we overlook result the k-representative we gives best result than K-mean and K-medoid algorithms.

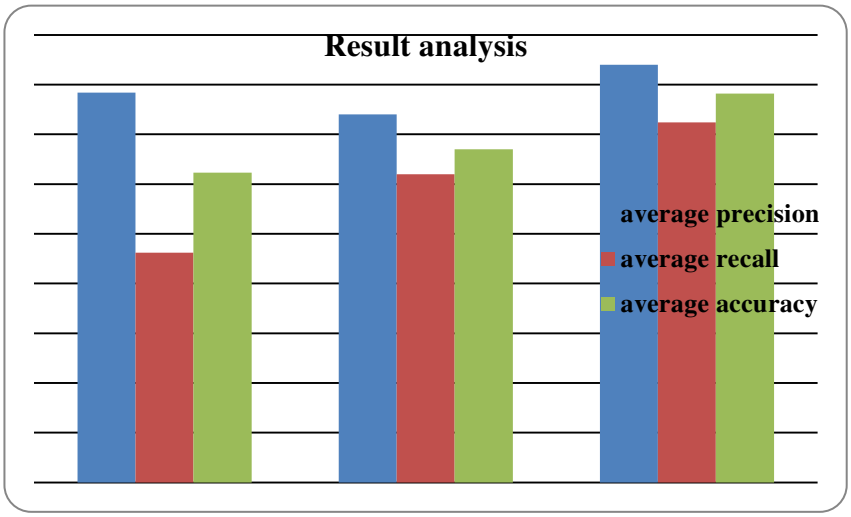



Figure 5.6: Result of Average Accuracy, Precision and Recall of Existing and Proposed system.

Figure 5.6 shows the graphical representation of the average accuracy, average precision and average recall of our proposed system. Here average precision of K-mean algorithm is 0.78 average recall is 0.42 and average accuracy is 0.62 .K-medoid algorithm is little but superior than K-mean their average accuracy is 0.67 next one is hybrid K-representative algorithm which is better than both algorithms .it has average accuracy is 0.78 more than K-mean and K-medoid algorithms.

## 4. CONCLUSION:

In this paper we briefly discuss about our practical approach for implementation of previously proposed system having focus on improved text clustering algorithm which is actually used for forming clusters on the basis of not exact match but on relative match. It also provides better results and improves the accuracy and efficiency of forensic document clustering technique. By using such type of approach searching time for finding relevant document from massive amount of datasets will be significantly reduce and recover the efficiency, effectiveness of forensic analysis.

**REFERENCES:**


1. L.F.D.C Nassif and E.R. Hruschka, "Document Clustering for Forensic Analysis: An Approach for Improving Computer Inspection", IEEE Transactions on Information Forensics and Security,Vol. 8, No. 1, January 2013.
2. R. Mundhe, A. Maind and R. Talmale, "Information Retrieval Using Document Clustering for Forensic Analysis", International Journal of Recent Advances in Engineering & Technology(IJRAET), Vol. 2, 2014.
3. S. Karol and V. Mangat, "Evaluation of a Text Document Clustering Approach based on Particle Swarm Optimization", International Journal of Computer Science and Network Security(IJCSNS), Vol. 13, July 2013.
4. G. Thilagavathi and J. Anitha, "Document Clustering in Forensic Investigation by Hybrid Approach", International Journal of Computer Applications Vol. 91, April 2014.
5. K. Nagarajan and M. Prabakaran, "A Relational Graph Based Approach using Multi Attribute Closure Measure for Categorical Data Clustering", The International Journal Of Engineering And Science (IJES) ,Vol. 3, 2014.
6. B. K. L. Fei, J. H. P. Eloff, H. S. Venter, and M. S. Oliver, "Exploring forensic data with self organizingmaps", Internatinal Conference Digital Forensics, 2005.
7. R. Hadjidj, M. Debbabi, H. Lounis, F. Iqbal, A. Szporer, and D. Benredjem, "Towards anintegrated e-mail forensic analysis framework," Digital Investigation, Elsevier, vol. 5, 2009.
8. K. Stoffel, P. Cotofrei, and D. Han, "Fuzzy methods for forensic data analysis", IEEE International Conference Soft Computing and Pattern Recognition, 2010.
9. C. C. Charu, and C. X. Zhai, Eds., "Chapter 4: A Survey of Text Clustering Algorithms",Mining Text Data. NewYork: Springer, 2012.
10. M. R. Clint, M. Reith, C. Carr, and G. Gunsch, an Examination of Digital Forensic Models (2003).
11. B. Vidhya and R. Priya Vaijayanthi, "Enhancing Digital Forensic Analysis through Document Clustering", International Journal of Innovative Research in Computer and Communication Engineering, Vol.2, Issue 1, March 2014.
12. T. Thopte, Y. Indani, M. Jangale and S. Gaikwad, "Heuristic Approach for Document Clustering in Forensic Analysis", International Journal of Computer Science and Information Technologies, Vol. 6 (1), 182-185, 2015.
13. C. Jadon and A. Khunteta, "A New Approach of Document Clustering", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, Issue 4, April 2013.