Forensic Document Analysis using Apriori Algorithm

Mr. Amit N. Borkar, Prof. P. S. Kulkarni
Dept. of C.Tech., RCERT, Chandrapur , Gondwana University, Gadchiroli, Maharashtra,India.

Abstract: Today, we find a tremendous increase in crimes like ethical hacking, rackets on different domain packets, unauthorized entrance, in the digital world. So, we need to track such unauthorized access. Our forensic document analysis using apriori algorithm provides an approach which is used to find the evidence by analyzing such massive set of documents. In forensic analysis frequently we examine hundreds of thousands of file which is huge amount of data. Here, we group the retrieved documents into the meaningful categories list which is the most central process by using the k-representative algorithm. Hence we are specifying an approach for forensic analysis using document clustering algorithm helpful in police investigation.

Keywords*: Document clustering, k-representative, k-means, k-medoids, Forensic analysis, Apriori algorithm.*

*****

I. INTRODUCTION:

Now a day's due to highly increase in crime linked with the computer and internet, forensic analysis of computers become emergent need of the digital world. For computer forensic analysis generally we requires some computer forensic tools that can subsist in the form of computer software. Because of storage media expanding in size of large volume, day by day, it becomes difficult task for investigator to locate their points of interest from large pool of data. Therefore, introducing such tools are helpful for forensic investigation while dealing with computer investigation. On the other hand, it may be difficult for investigators because of the way in which data is present and may result in misinforming. As a result, the method will acquire a very large amount of time for analyzing big volumes of data. Sometimes it is also possible, that computer forensic tools may be meaningless for their generated data due to the fact that current tools of computer forensic is not capable of presenting a clear virtual outline of all the stuff (files) originated on the storage medium, as storage medium can store data in huge amount. So, the clustering of documents in computer forensic plays an important role for identifying the evidence in police investigation of crime. This activity exceeds the capability of expert analysis and understanding of data. In more realistic and practical situation field expert (e.g. forensic examiners) are sparse and have partial time for performing examinations. Thus after finding an appropriate documents, it become sensible to suppose that the examiner might prioritized the analysis of supplementary documents belonging to the significant cluster, because it is likely that these can also be appropriate and applicable to the investigation. Such approaches based o clustering of documents, can certainly advance the progress of analysis of seized computer. Here the number of clusters is a grave bound of various algorithms and it is frequently unknown a priori. Till now the automatic assessment of the cluster number is not evolved in the investigation of the computer forensic literature. Truly, we couldn't specify one work which is reasonably close in its domain application and that intelligence the competent of the algorithm estimating the cluster in specific number, back to the sixties it was suprising that lack of studies over hierarchical clustering algorithm.

II. LITERATURE SURVEY:

The author in this paper[1] Two process of k-representative, instance memberships to cluster and cluster re-estimating are describe where representative is used by replacing centroids as centriods are present only in numerical domain. Representative show the occurring ratio among the possible value of features of the clusters member .Value difference matrix is inserted to compute the distance between the instances, specified. The article[2] demonstrate the proposed approach by targeting extensive test by conducting different experimental test of six well-known clustering algorithms(Single link, Complete link, Average link, k-means, k-medoids, and CSPA) that where applied to five different dataset of real-world obtains from investigation carried over computer seized..

Exploratory of data analysis where done by clustering algorithms, when there is no or little prior knowledge about the data [2]. For mining e-mails for forensic analysis in integrated surrounding via classification and clustering algorithm, was present in [3].In an application domain which is related to email were grouped by using structural, domain-specific, syntactic, and lexical features[5]. The problem of e-mails clustering for forensic analysis was also introduced, using three clustering algorithm (k-means, Bisecting k-means and EM), where K-means of kernel-based variant was applied [4]. For forensic investigation in mining prints write from e-mails anonymous, basically they were written by multiple anonymous author throw collecting e-mails and focusing on the problem of mining the styles of writing those e-mails. The basic way for anonymous e-mail is to be first cluster by the Stylometric (the application of the study of linguistic style, mainly written language .i.e. Stylometry, but it actually applied for music and to fine-art

paintings successfully) features and then pull out the write print, i.e., the inimitable writing style, from each cluster.[3]

This paper speak about the various computer forensic tool that are available on the market, For instance forensic Toolkit, Encase and Pro Discover are the list of available tools.The difference about these tools that some are built for single purpose only while other are designed to provide a whole range of functionalities. Looking over the examples of these functionalities are hashing verification, report generation, advanced searching capabilities and etc. Some computer forensic tools have the common functionalities by difference only in there GUI [6].

III. PROPOSED SYSTEM:

Clustering play very important role in data mining, there are some clustering algorithm by which the formation of clusters center is done. So, for obtaining our objective we proposed k-Representatives Algorithm by which the clustering of document is carried out and then the clustered documents are analyzed using Apriori Algorithm.

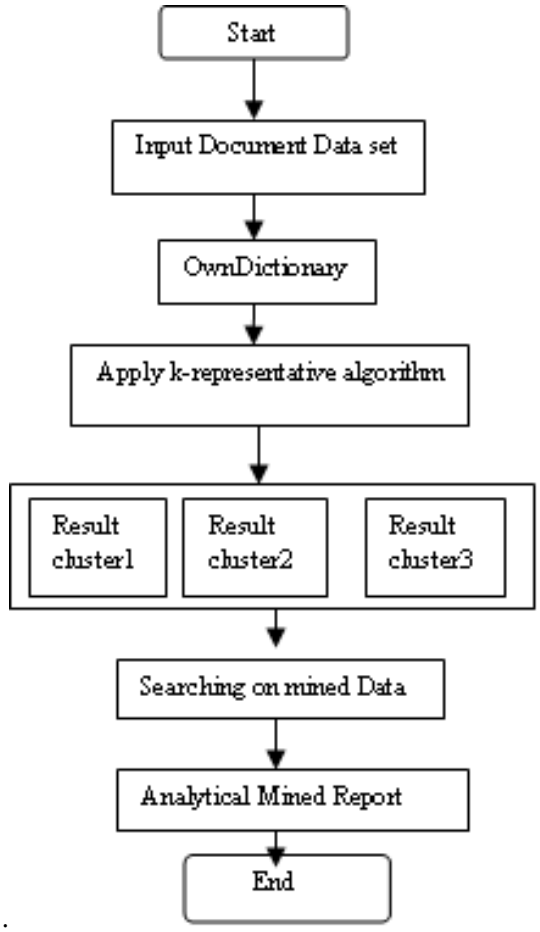

Fig.1.Flow of proposed system

In our previous paper, we explained how to gather the data, preprocessing of data for that we collect the real time data which may contains, FIR report, Empty files, private data, etc. After this above two processes it become easier for the formation of cluster and the analysis on different clustering algorithm. Here the cluster formation is done using k-representative algorithm and also shown the comparative analysis between different clustering algorithm(k-means.k-medoids and k-representative).

IV. CLUSTER FORMATION MODULE:-

The very first think we had done is the gathering of data and then preprocessing, input the data set which is of real time and it's a part of preprocessing where the stemming and stop-word removal process is carried out, then we group the useful data in the form of indexing which is actually going to use for the clustering of documents. After that, we form the cluster based on the input given by the user and that is real time input, the formation of clustering the documents in done using k-representative algorithm, it is the most refinement approach for forming the clusters. Here, we form N number of cluster useful for retrieval of documents depending on the search for mined data. The retrieval of documents is then proceed to analysis the document by using Apriori algorithm. The use of apriori algorithm is to get the output is more optimized form.

V. K-REPRESENTATIVE ALGORITHM:

K-representative algorithm takes iterative refinement approaches, which include EM and k-means and known as the most effective among various approaches to solve the clustering problems. As the other iterative clustering algorithms, k-representatives algorithm also repeats two processes, deciding the memberships of instances to clusters and re-estimating the centroids. In our algorithm, the representatives replace the centroids of clusters because centroids exist only in numerical domains. A representative of a cluster shows the occurring ratios of all possible values of features in the members in the cluster

Algorithm:

Initialize clusters and their representatives
Repeat
Decide the memberships of instances to clusters
For each feature, derive value difference matrix
For each instance, measure distances between instances and clusters and classifyit to the closest cluster
Re-estimate the representatives of clusters
Until no object has changed clusters.

Here is some experimental results,as shown below, initially we have to make choice of file or flder which we want to

 ISSN: 2321-8169
5705 - 5710

investigate, the file can be present at anywhere in pendrive, harddisk,CD, DVD,etc.We have to browse the folder and then select the file or folder which we want to investigate.The result will shown the first wo processes STOPWORDS and PRE-PROCESS DATA.After that type the word or sentence which is to be cluster(for ex:crime), it will take few seconds for the next process called as indexing and then it will shows the result of tota cluster formation at runtime.

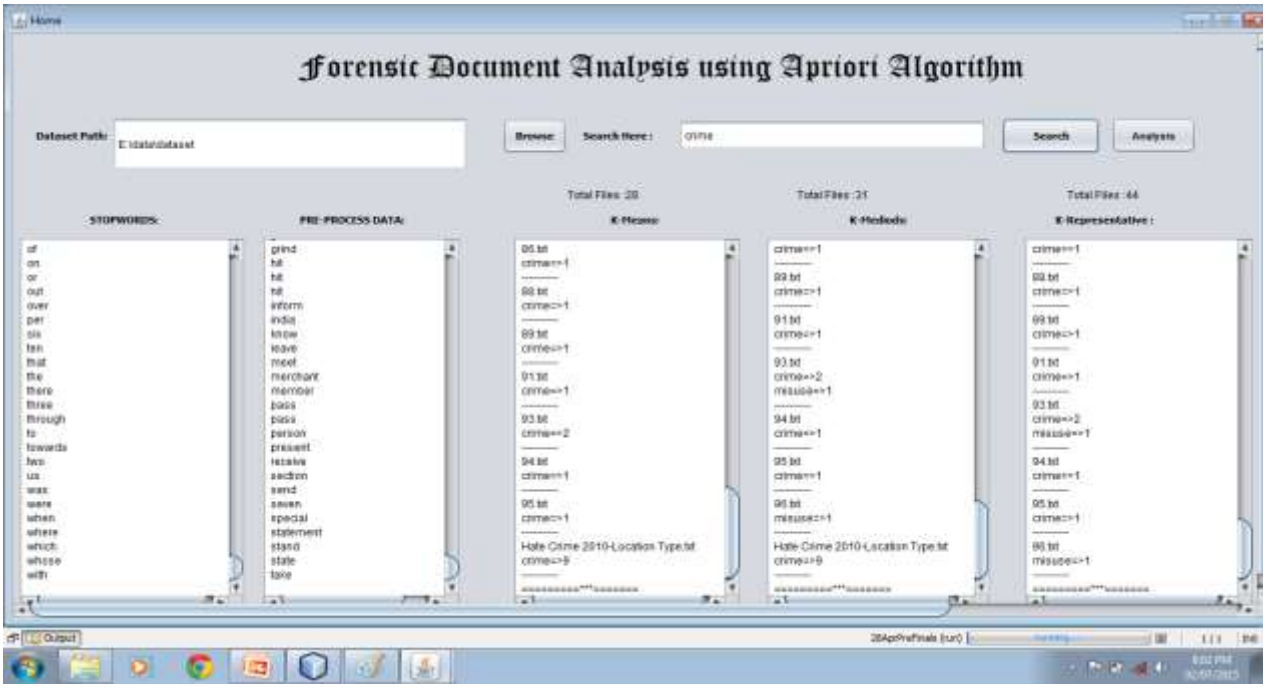

Fig:2 search window fior forensic document analysis using apriori algorithm

The next output window gives the detials of kewords search and their related synonyms that will cluster in different files

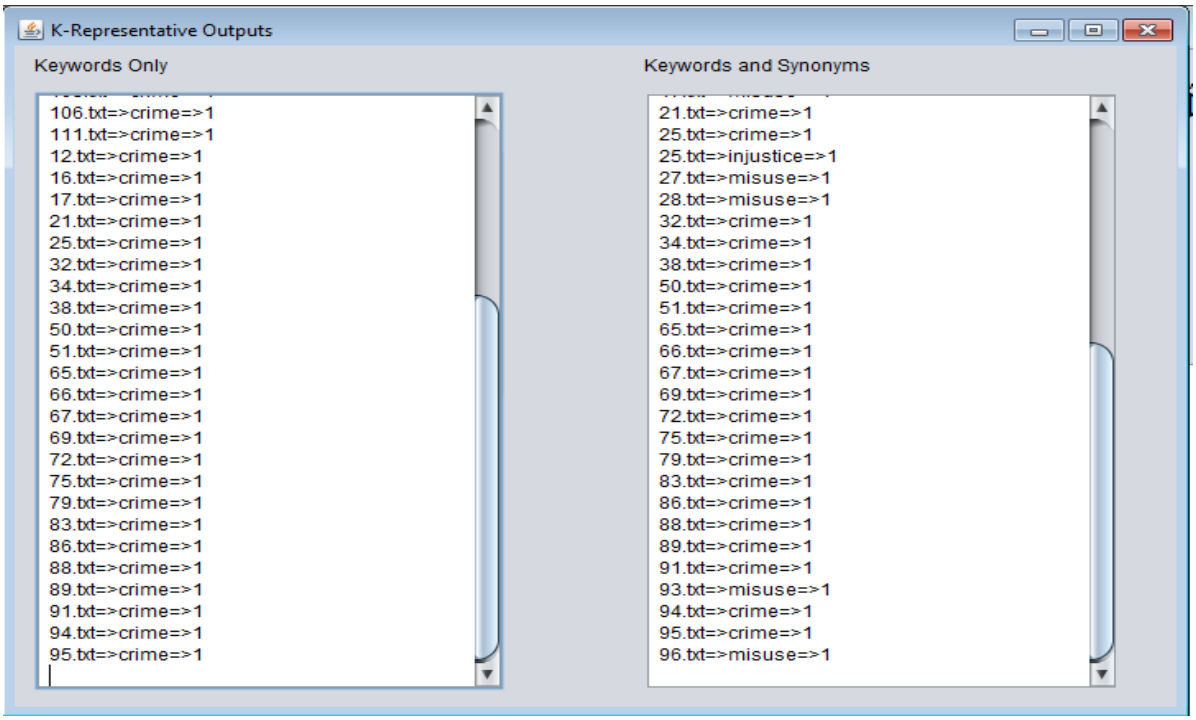

Fig:3 keywords & Synonyms result using k-representative.

 ISSN: 2321-8169
5705 - 5710

VI. COMPARITIVE ANALYSIS:

The table given below will show the experimental analysis on the algorithm regarding different parameters:

Clustering Algorithm / Paramters	k-means	K-medoids	K-representative
Total Cluster	28	31	48
Precission	0.784	0.74	0.84
Recall	0.462	0.6	0.724
Efficiency	0.623	0.672	0.782

Fig:4 Table of comparitive analysis.

Here, we can clearly see the comparitive analysis result between k-means, k-medoids and k-repreentative algorithm by graphical chart where the total clusters formed by them is visible to us using different colours bar chart.It clearly specify that k-representative gives the good number of cluster.

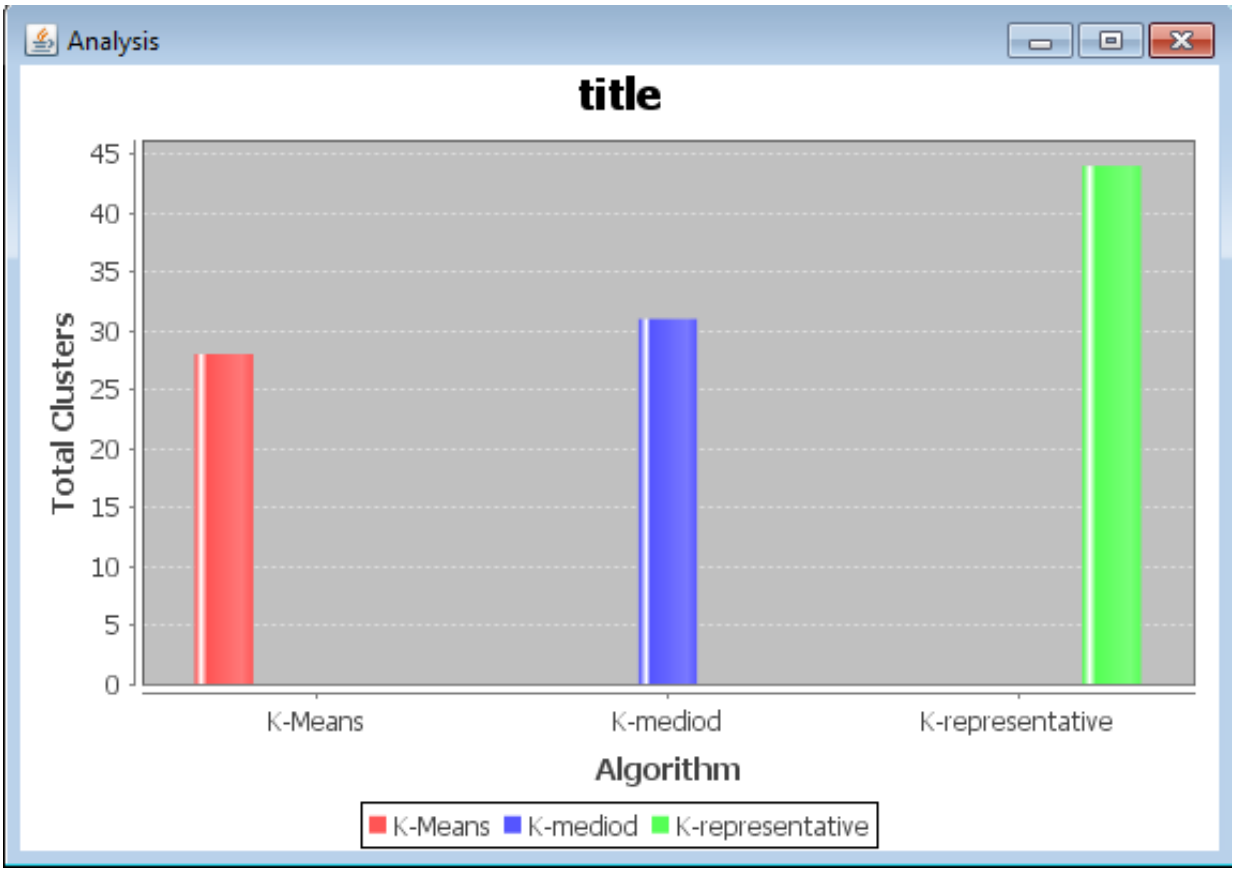

Fig:5 Comparitive analysis graph between k-means,k-medoids and k-representative.

 ISSN: 2321-8169
5705 - 5710

The next window is used to show the precission and recall made by the algorithms, blue colour bar for precossion and red for the recall.

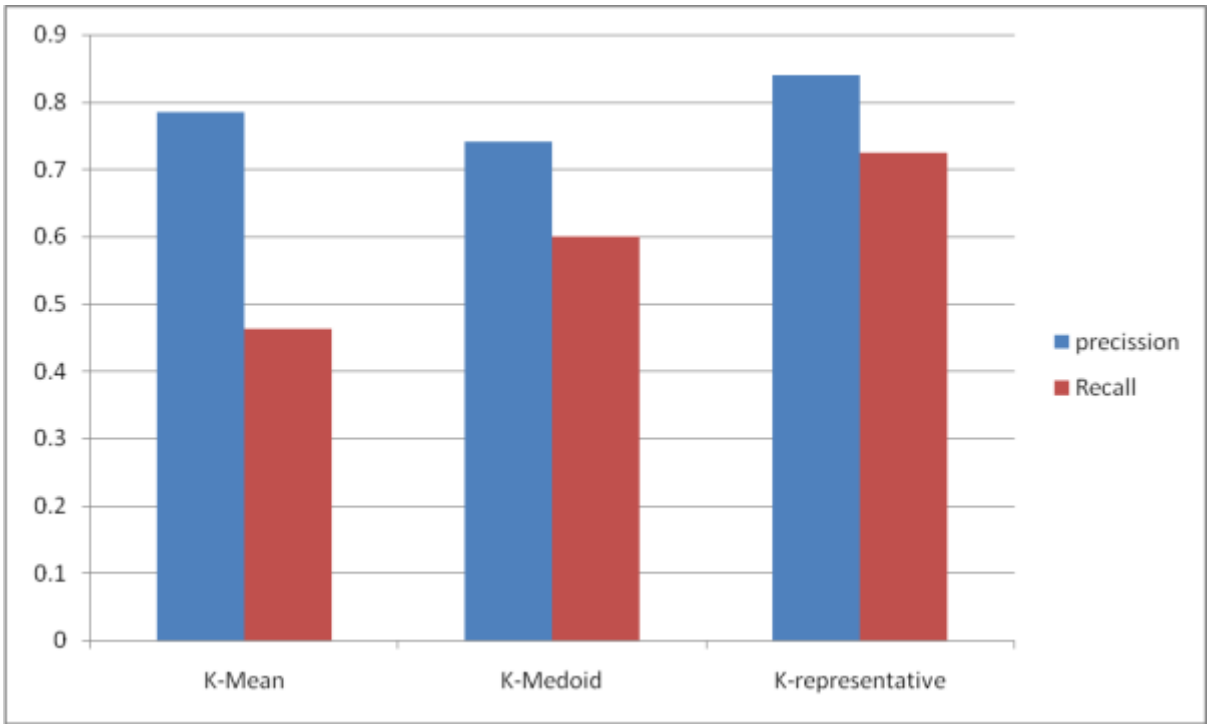

Fig:6 precision and recall by different clustering algorithm

Her;e, is the accuracy rate graph,its shows the accuracy rate of k-means, k-medoids and k-representative.

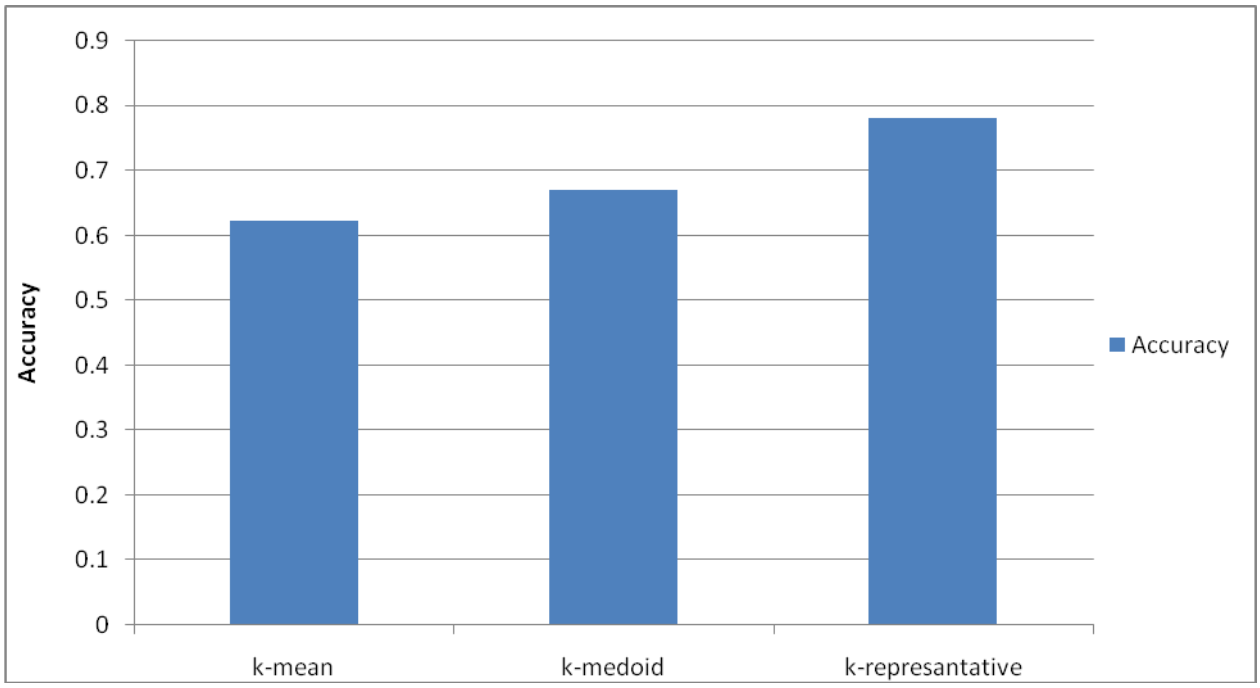

Fig:7 accuracy measure by k-means, k-medoids and k-representative.

VII. Conclusion:

After gathering of data and preprocessing it, the runtime cluster formation is carried out using k-representative where the result shows the good number of cluster formation than any other clustering algorithm. Hence, it is clear that the formation of clustering using k-representative is more effective and accurate. Here, also the comparative analysis between different clustering algorithm is shown. Our next, step is to optimized the search data by using Apriori algorithm which will be the final output for analyzing the data in forensic department.

VIII. Reference:

[1] Jae Heon Park and Sang Chan Park k-representatives Algorithm: a Clustering Algorithm with Learning Distance Measure for Categorical Values" KAIST (Korea Advanced Institute of Science and Technology), Department of Industrial Engineering

[2] B. S. Everitt, S. Landau, and M. Leese, *Cluster Analysis*. London, U.K.: Arnold, 2001.

[3] R. Hadjidj, M. Debbabi, H. Lounis, F. Iqbal, A. Szporer, and D. Benredjem, "Towards an integrated e-mail forensic analysis framework," Digital Investigation, Elsevier, vol. 5, no. 3–4, pp. 124–137, 2009.

[4] S. Decherchi, S. Tacconi, J. Redi, A. Leoncini, F. Sangiacomo, and R. Zunino, "Text clustering for digital forensics analysis," Computation Intell. Security Inf. Syst., vol. 63, pp. 29–36, 2009.

[5] F. Iqbal, H. Binsalleeh, B. C. M. Fung, and M. Debbabi, "Mining writeprints from anonymous e-mails for forensic investigation," Digital Investigation, Elsevier, vol. 7, no. 1–2, pp. 56–64, 2010..

[6] B. K. L. Fei, J. H. P. Eloff, H. S. Venter, and M. S. Oliver, "Exploring forensic data with self-organizing maps," in Proc. IFIP Int. Conf. Digital Forensics, pp. 113–123,2005.

[7] Luis Filipe da Cruz Nassif and Eduardo Raul Hruschka "Document Clustering for Forensic Analysis: An Approach for Improving Computer Inspection" IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, VOL. 8, NO. 1, 1556-6013 IEEE-2012.

AUTHOR BIOGRAPHY:

1) Amit N. Borkar M.TECH(CSE),Final Year, Dept. of C.Tech., RCERT, Chandrapur , Gondwana University, Gadchiroli,Maharashtra,India presenting the paper on forensic document analysis using Apriori algorithm still we have published the two paper on same topic.

2) Prof.P.S.Kulkarni Associate Professor, Dept. of Info.Tech., RCERT, Chandrapur, Gondwana University, Gadchiroli, Maharashtra,India.