Automatic Labelling and Document Clustering for Forensic Analysis

Ms. Raksha K.Mundhe,
IVth SEM, M.Tech – CSE,
Tulasiramji Gaikwad-Patil College of Engineering and Technology, Nagpur, India
raksha.k.mundhe@gmail.com

Prof. Ankush Maind
Department of CSE
Tulasiramji Gaikwad-Patil Collegeof Engineering and TechnologyNagpur, India
ankushmaind@gmail.com

Abstract - In computer forensic analysis, retrieved data is in unstructured text, whose analysis by computer examiners is difficult to be performed. In proposed approach the forensic analysis is done very systematically i.e. retrieved data is in unstructured format get particular structure by using high quality well known algorithm and automatic cluster labelling method. Indexing is performed on txt, doc, and pdf file which automatically estimate the number of clusters with automatic labelling to it. In the proposed approach DBSCAN algorithm and K-mean algorithm are used; which makes it very easy to retrieve most relevant information for forensic analysis also the automated methods of analysis are of great interest. In particular, algorithms for clustering documents can facilitate the discovery of new and useful knowledge from the documents under analysis. Two methods are used for document clustering for forensic analysis; the first method uses an x^2 test of significance to detect different word usage across categories in the hierarchy which is well suited for testing dependencies when count data is available. The second method selects words which both occur frequently in a cluster and effectively discriminate the given cluster from the other clusters. Finally, we also present and discuss several practical results that can be useful for researchers of forensic analysis.

Index terms: *forensic analysis, clustering algorithm, automatic labelling method, indexing.*

____*****____

I. INTRODUCTION

In forensic analysis, hundreds of thousands of files are usually examined. Much of those files consist of unstructured text, the term unstructured refers to data which does not have clear, whose analysis by computer examiners is difficult to be performed. Generally search engines are used to retrieve the data. The search engines commonly build a very large centralized database to index a portion of Internet and help to reduce the information overload problem by allowing a user to do a centralized search. However, they also bring up another problem: too many web pages are returned for a single query. To find out which documents are useful, users have to sift through hundreds of pages to find out that only a few of them are relevant. One way to tackle this problem is to cluster the search result documents based on their extensions so that users can scan a few coherent groups instead of many individual documents. This activity exceeds the expert’s ability of analysis and interpretation of data. Therefore in proposed approach, methods and clustering used keep paramount importance. Explosion of research aimed at facilitating retrieval and organization of clustering documents into meaningful groups, for this Scatter/Gather cluster based approach is used. The use of clustering algorithms, which are capable of finding latent patterns from text documents found in seized computers, can enhance the analysis performed by the expert examiner.

The proposed approach provides the rationale behind clustering algorithms is that objects within a valid cluster are more similar to each other than they are to objects belonging to a different cluster. Such an approach, based on document clustering, can indeed improve the analysis of seized computers. Computer forensics is the application of investigation and analysis technique to gather and preserve evidence from a particular computing device in a way that is suitable for presentation of court of law. The goal of computer forensics is to perform a structured investigation while maintaining a documented chain of evidence to find out exactly what happened on a computing device and who was responsible for it. Therefore, I decided to choose powerful well known representative algorithms in order to show the potential of the proposed approach, namely: the K-means, DBSCAN, Hierarchical Algorithm and Automatic Labelling Cluster Method.

Information Retrieval Using Document Clustering:

Information retrieval (IR) is a technique of finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers). The term “unstructured” refers to data which does not have clear. The field of information retrieval also covers supporting users in browsing or filtering document collections or further processing a set of retrieved documents. Given a set of documents, clustering is the task of coming up with a good grouping of the documents based on their extensions. Clustering is the process of finding group of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other group. (IR) systems have to address the problem of returning relevant information in response to a user's information need.

Scatter/Gather Cluster Based Approach:

In Scatter/Gather approach most useful algorithm is Hierarchical Algorithm. Scatter/Gather clustering algorithms can be used to compute a hierarchical clustering solution using a repeated cluster bisecting approach. In this approach, all the documents are initially partitioned into two clusters. Then, one of these clusters containing more than one document is selected and is further bisected. This process continues n − 1 times, leading to n leaf clusters, each containing a single document. It is easy to see that this approach builds the hierarchical agglomerative tree from top (i.e., single all-inclusive cluster) to bottom (each document is in its own cluster). In the rest of this section we describe the various aspects of the partitional clustering algorithm that we used in our study. Hierarchical Algorithm divides into two parts first is agglomerative and second is divisive clustering. Figure 1 shows the working principle of hierarchical algorithm also called as scatter/gather approach. In Agglomerative method merging of the most similar pairs of data points is done until one big cluster left. This is called a bottom-up approach. In divisive method splitting of large data is done. It is top- down approach. The concept is explained below diagrammatically.

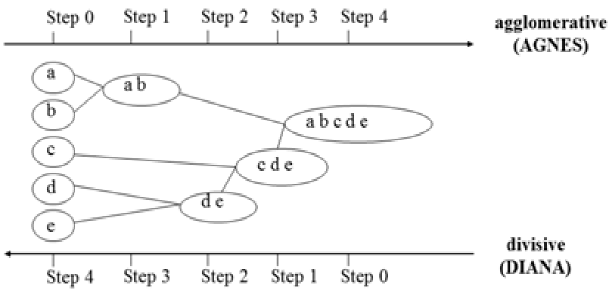

Figure 1: Scatter/Gather Approach

1.3 Hybrid Algorithm:

Hybrid hierarchical clustering algorithm will be the integration of density based clustering and hierarchical cluster.

This paper is organized in the following way. In section II some earlier related work is explained. In section III, System Architecture. In section IV, Automatic Labelling And Document Clustering. Section V, Result and Discussion. In section VI, conclusion and future scope.

II. RELATED WORK

Six well-known algorithms [1] (K-means, K-medoids, Single/Average/Complete Link, and CSPA), Two relative validity indexes were used to automatically estimate the no. of clusters. Dendrograms provide summarized view of the document being inspected also provide very informative descriptions and visualization capabilities of data clustering structures. The k-medoid algorithm is a clustering algorithm related to the k-means algorithm and the medoid shift algorithm.

Both the k-means and k-medoids algorithms are partitional (breaking the dataset up into groups) and both attempt to minimize squared error, the distance between points labelled to be in a cluster and a point designated as the centre of that cluster. In contrast to the k-means algorithm, k-medoids chooses data points as centres. A medoid of a finite dataset is a data point from this set, whose average dissimilarity to all the data points is minimal i.e. it is the most centrally located point in the set. K-medoid could be more robust to noise and outliers as compared to k-means because it minimizes a sum of general pair wise dissimilarities instead of a sum of squared euclidean distances.

In Single-Link Method / Nearest Neighbour (also called the connectedness, or minimum method), the distance between one cluster and another cluster is equal to the shortest distance from any member of one cluster to any member of the other cluster. In Complete-Link / Furthest Neighbour (also called the diameter or maximum method), the distance between one cluster and another is equal to the longest distance from any member of one cluster to any member of the other cluster. In Average-link clustering, the distance between one cluster and another cluster to be equal to the average distance from any member of one cluster to any member of the other cluster.

In Cluster-based Similarity Partitioning Algorithm (CSPA), if two objects are in the same cluster then they are considered to be fully similar, and if not they are dissimilar. This is the simplest heuristic and is used in the Cluster-based Similarity Partitioning Algorithm (CSPA). With this viewpoint, one can simply reverse engineer a single clustering into a binary similarity matrix. Similarity between two objects is 1 if they are in the same cluster and 0 otherwise.

An automatic procedure and methodology are described [2] for inferring accurate and easily understandable expert-system-like rules from forensic data. For this FCM-based mining association rule based on the fuzzy set theory is used to generate most appropriate result. FCM minimizes intra-cluster variance and also gives best result for overlapped data set and computationally better but Euclidean distance can unequally weight underlying factor. The Fuzzy C-Means Clustering (FCM) is an unsupervised goal oriented clustering algorithm, introduced by Dunn and generalized by Bezdek. In fuzzy set theory the fuzzy methods improve the effectiveness and the quality of the data analysis phase for crime investigation. The paper proposes a framework for applying fuzzy tools in digital investigation. The main goal is the extraction of expert-system-like rule sets based on fuzzy sets that can be presented to the experts in order to support them in their daily activities. This framework is conceived to be a potential starting point to a future standard framework for guiding the use of computational intelligence techniques in gathering digital evidence admissible in a court of law. Fuzzy clustering is used to detect the explanation of criminal activities for crime hot-spot areas and their spatial trends. Compared with two hard-clustering approaches (median and k-means clustering problem), the empirical results suggest that a fuzzy clustering approach is better equipped to handle crime spatial outliers. A two stage fuzzy decision classifier, using reference fuzzy set information, is used to create a text-independent Automatic Speaker Identification. G. Salton choosed effective term-weighting system and provides baseline single-term indexing models with which other more elaborate content analysis procedure can be compared.

SOM-based algorithm [3] clustered the files, based on their creation dates/time and extension which performs analysis process more efficient during computer investigation. A self-organizing map (SOM) assists to computer forensic investigators conducting data analysis in a more efficient manner. A SOM is used to search for patterns in data sets and produce visual displays of the similarities in the data. SOM provide greater abilities to interpret and explore data generated by computer forensic tools.The data set obtained directly from hard drive can be critical to an investigation. Patterns in the dataset could help forensic investigators to locate information and guide them to the next step in their search. The technique is used to create graphical representations of large datasets that offer investigators a fresh perspective from which to study the data.

SOM-based algorithm used to cluster the files, based on their creation dates/times and extensions, which increases the information retrieval efficiency but it is computationally less efficient. Text clustering [4] is one of the central problems in text mining and information retrieval area. So two techniques: Feacher Extraction and Feacher Selection are used to improve the efficiency as well as accuracy of text clustering.

Three effective and efficient combiners are used [5] to solve cluster ensemble problem as an optimization problem based on a hyper-graph model which results cluster ensembles can improve quality and robustness, and Enable distributed clustering. The idea of combining cluster labelling without accessing the original features leads us to a general knowledge reuse framework that we call cluster ensembles.

The consistency of clustering solutions [6] at different levels of granularity allows flat partitions of different granularity to be extracted during data analysis, making them ideal for interactive exploration and visualization.Y. Zhao & G. Karypis uses fast and high quality document clustering algorithm i.e. partitional clustering algorithm to provide intuitive navigation and browsing mechanism by organizing large amount of information into small number of meaningful clusters for better clustering performance. For partitional clustering algorithms, six functions are used [1] that have been shown to produce high-quality partitional clustering solutions. There is the common belief that in terms of clustering quality, partitional algorithms are actually inferior and less effective than their agglomerative counterparts.

The vector-space model is to represent each document. In this model, each document d is considered to be a vector in the term-space.In this model, each document is represented by a vector containing the frequencies of occurrences of words, also uses the technique known as Term Variance (TV) that is o(N log N) versus the overall time complexity of o(N^2 log N) for agglomerative methods. Clustering algorithm indeed tend to induce clusters form by either relevant or irrelevant documents, thus contributing to enhance the examiners job. Limitations-Scalability may be an issue (in order to deal with this issue, a number of sampling and other techniques can be used.In particular, we employed the t f − id f term weighting model, in which each document can be represented as

(tf1 log(n/df1), tf2 log(n/df2), . . . , tfm log(n/dfm)).

Where, tfi is the frequency of the ith term in the document and dfi is the number of documents that contain the ith term.

Because of the problem of mining the writing styles [7] from a collection of e-mails written by multiple anonymous authors, the e-mails are grouped by using lexical, syntactic, and structural and domain specific features. Three algorithms i.e. K-mean, Bisecting K-mean and EM are used to resolve the problem.

Integrated environment used [8] for mining e-mails for forensic analysis, using classification and clustering algorithm that performs a multistage analysis of e-mail ensembles with a high degree of accuracy, and in a timely fashion.

Problem of clustering e-mails for forensic analysis [9] was addressed so Kernal-based variant of K-means was applied. Proposed seized digital device can provide precious information and evidence about facts and law text analysis strategy, relying on clustering based text mining techniques, is introduced for investigational purposes.

Clustering algorithms are used [10] for data sets appearing in statistics, computer science, and machine learning; the traveling salesman problem is used to find out the shortest possible distance between clusters.

III. SYSTEM ARCHITECTURE

It is well-known that the number of clusters is a critical parameter of many algorithms and it is usually a priori unknown. As far as we know, however, the automatic estimation of the number of clusters has not been investigated in the Forensics Analysis. Actually, we could not even locate one work that is reasonably close in its application domain and that reports the use of algorithms capable of estimating the number of clusters. This thesis extends our previous work; In addition, we provide more insightful quantitative and qualitative analyses.

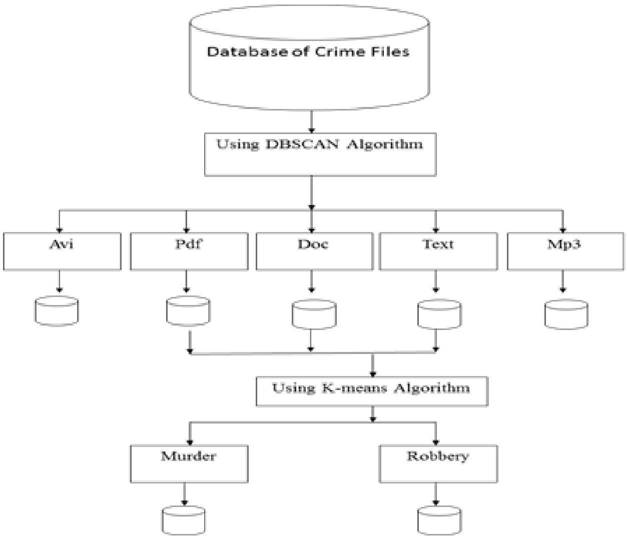

Figure 2: System Architecture

As shown in the Figure 2 the DBSCAN Algorithm is applied on the database of crime files, after applying the DBSCAN Algorithm files gets separated according to their extensions i.e. pdf, txt, doc files; Then K-mean is applied on these files and we get resultant files in particular folder.

System Flow:

As shown in the system flow diagram that is Figure 3 we have to upload the files after signing in the system, Uploading means transmission of a file from one location to another, usually in large computer system. From a network user's point-of-view, to upload a file is to send it to another computer that is set up to receive it.uploading files generates the database of the crimes. Database contains the various text files, document files, and images, audio and video files. We have to apply the DBSCAN Algorithm on generated database. DBSCAN Algorithm helps to separate the files as per their extensions. Once the file moves to the certain folder we need to retrieve the related data to the forensic analysis such as Robbery files, murder files etc. for these reason K-mean is applied to meet the resultant files. For example we are having text files such as murder, rape case, robbery etc. The indexing is performed according to the keyword and we get the perfect result.

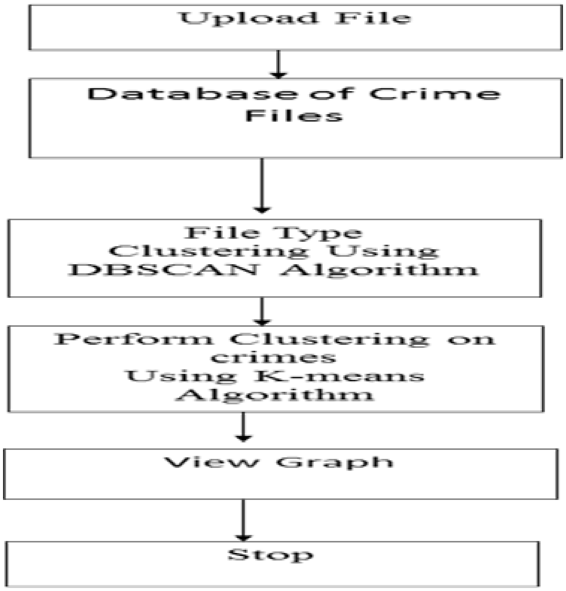
Upload File
Database of Crime Files
File Type Clustering Using DBSCAN Algorithm
Perform Clustering on crimes Using K-means Algorithm
View Graph
Stop

Figure 3: System Flow

DBSCAN Algorithm:

It finds number of clusters starting from estimated density distribution of corresponding nodes. By using two parameters i.e. ε-eps and minPts it perform labelling the cluster on the basis of neighbourhood matching otherwise point is labelled as a noise.

For the DBSCAN algorithm following terms is used in consideration of Database D, Core point (q), Border point (p), Minimum no of points in cluster (Minpts) and Radius (Eps).Two global parameters are explained in Figure 4.

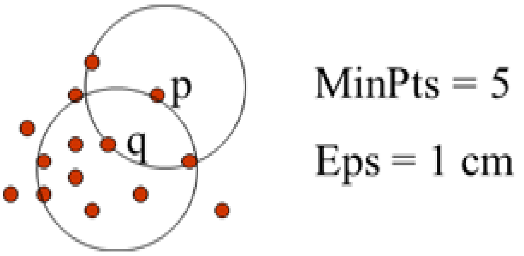

Figure 4: Example of Eps and MinPts

Algorithm steps:

1. Arbitrary select a point p
2. Retrieve all points density-reachable from p wrt Eps and MinPts.
3. If p is a core point, a cluster is formed.
4. If p is a border point, no points are density-reachable from p and DBSCAN visits the next point of the database.
5. Continue the process until all of the points have been processed.

K-mean Algorithm:

K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

Figure 5 shows the execution steps of the K-mean algorithm. K-mean is computationally efficient and does not require the user to specify many parameters.

Followings are the steps of flowchart in Figure 5.7.

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.

2. Assign each object to the group that has the closest centroid.

3. When all objects have been assigned, recalculate the positions of the K centroids.

4. Repeat Steps 2. and 3. until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

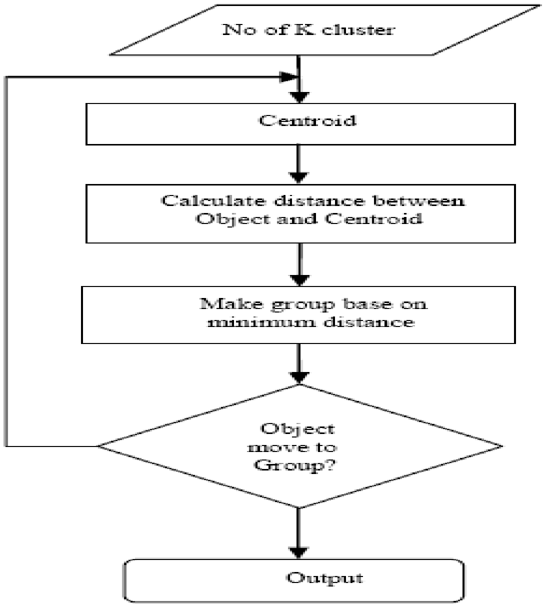
No of K cluster
Centroid
Calculate distance between Object and Centroid
Make group base on minimum distance
Object move to Group?
Output

Figure 5: Flowchart of K-mean

File Extension:

Files get separated according to the extensions i.e. .avi, .txt, .doc, .pdf, .jpeg etc.

IV. AUTOMATIC LABELLING AND DOCUMENT CLUSTERING

The search engines commonly build a very large centralized database to index a portion of Internet and help to reduce the information overload problem by allowing a user to do a centralized search. However, they also bring up another problem: too many web pages are returned for a single query. To find out which documents are useful, users have to sift through hundreds of pages to find out that only a few of them are relevant. In the proposed approach.One way to tackle this problem is to cluster the search result documents based on their extension similarity so that users can scan a few coherent groups instead of many individual documents. The first method uses a x^2 test of significance to detect different word usage across categories in the hierarchy which is well suited for testing dependencies when count data is available. The second method selects words which both occur frequently in a cluster and effectively discriminate the given cluster from the other clusters. Indexing and the technique used for automatic labelling provide the fast and efficient related data retrieval.

X^2 Method:

The X^2test is well suited for testing dependencies when count data is available. The main idea of our method is to use X^2 tests for each word at each node in a hierarchy starting at the root and recursively moving down the hierarchy. If one cannot reject the hypothesis that a word is equally likely to occur in all of the children of a given node, it is marked as general to the current subtree, assigned to the current node's bag of node− specific words and removed from all nodes under the current node.

The detailed description of the algorithm follows:

Input: A hierarchy of documents where the leaves contain bags of words from all of the documents in that leaf unioned together

1. Populate all internal nodes by unioning the bags of words in its children starting from the leaves and moving up to the root;

2. Start at the root and for each word perform a X^2test to discover dependencies:

a) if a test rejects the independence hypothesis, conclude that the word has different probability of occurring in children and thus is specific to one or more categories down the tree;

b) if a test fails to reject the independence hypothesis, conclude that the word is equally likely to occur in all of the children. Retain the word at the current node as being general to the subtree rooted at the current node. Remove all such words from all of the nodes below the node at which the test was performed;

3. Repeat step 2 recursively moving down the tree to the leaves.

Output: A hierarchy of words isomorphic to the initial hierarchy of documents, where each node contains words specific to that node and not occurring in the subtree below the current node.

A label is a list of the most frequent words at the node corresponding to a cluster of documents we want to label. Results are presented below.

Document Clustering:

Clustering based on density (local cluster criterion),
such as density-connected points. Each cluster has a considerable higher density of points than outside of the cluster.DBSCAN requires two parameters: ε (eps: Maximum radius of the neighborhood) and the minimum number of points required to form a cluster (minPts). It starts with an arbitrary starting point p that has not been visited from the group of point D. This point's ε-neighbourhood is retrieved, and if it contains sufficiently many points less than or equal to MinPts it is called a core point and a cluster is started. Otherwise, the point is labelled as noise. Note that p might later be found in a sufficiently sized ε -environment of a different point and hence be made part of a cluster. If in the range of p's ε radius the number of the elements is less than MiniPts, we can call p as the boundary, p is marked as noise node temporarily. Then, DBSCAN will dispose the next document in set D. As the first and the last step is the same as the threshold clustering method, so the two steps are ignored here.

V. RESULT AND DISCUSSION:

Hybrid hierarchical clustering algorithm will be the integration of density based clustering and hierarchical cluster.

The experimental result shows the improved performance of the proposed system after applying the DBSCAN, K-mean and automatic labelling approach. Figure 6 and Figure 7 shows the result in the form of graph also Table1 and Table2 shows the comparative summary. As shown in the Figure 6, the clustering is performed by using the DBSCAN Algorithm. It means that the unstructured data gets converted into the particular structure format. Here the Amps on 120 Volt Line says that the count of files of particular file type. As shown in the mentioned graph the related files gets clustered on the basis of their extensions i.e .txt files into the TXT folder and so on.

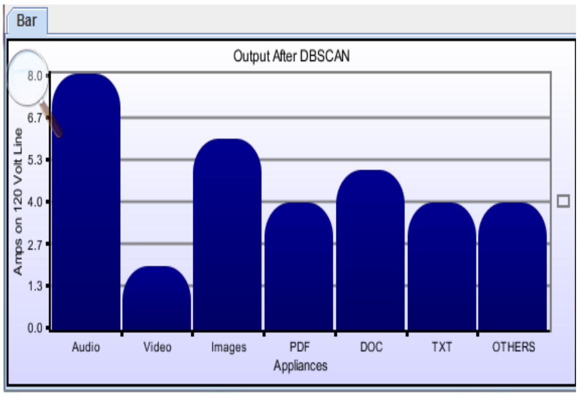

Figure 6: Output after DBSCAN

Table1: Output after DBSCAN

Type of file	Amps on 120 Volt Line
Audio	8
Video	2
Images	6
PDF	4
DOC	5
TXT	4
Others	4

Table1 summarizes the type of files with their count. But these files may have different data of crime. There may be various types of crime cases in a single document. For example PDF may contain the theft file, robbery file. Which can be time consuming process to search the exact file while forensic analysis. To achieve this objective K-mean is applied.

Table3: Comparison between Existing and Proposed System in Term of Time

No. of Files To Search	Time For Existing System	Time For Proposed System
1	10	10
2	23	18
3	15	5
4	21	15
5	10	3
6	60	40
7	55	20
8	26	15

After applying the DBSCAN Algorithm and K-mean algorithm to the input data the graph having name Figure 7 gets generate.Lets explain the graph with the help of example, consider .pdf document contains all type of file related to the crime such as theft, rape case and so on; after applying the DBSCAN and K-mean algorithm the related data i.e. theft related data transfer to the theft file, rape case related data moves to the rape_case file by performing indexing on it. Therefore one can get the relevant data while forensic analysis means it saves the time in investigation process.Table2 summarise the total number of files for particular crime.

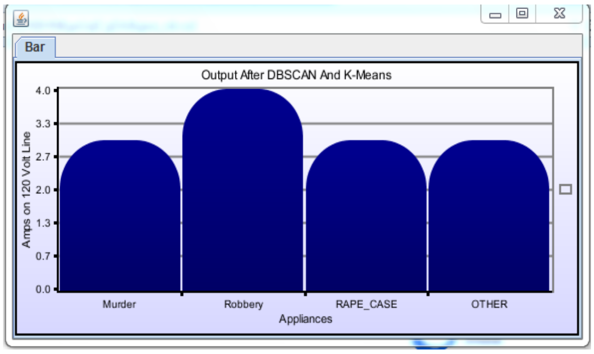

Figure 7: Output after DBSCAN and K-means

Table2: Output after DBSCAN and K-means

Type of Crime	Amps on 120 Volt Line
Murder	3
Robbery	4
Rape_case	3
Other	3

Figure 8 proves that the proposed system performs better and takes less time for searching process comparing the existing system. As shown in the comparison graph the proposed system search the related file within small amount of time comparing the existing system, so the performance gets increase.

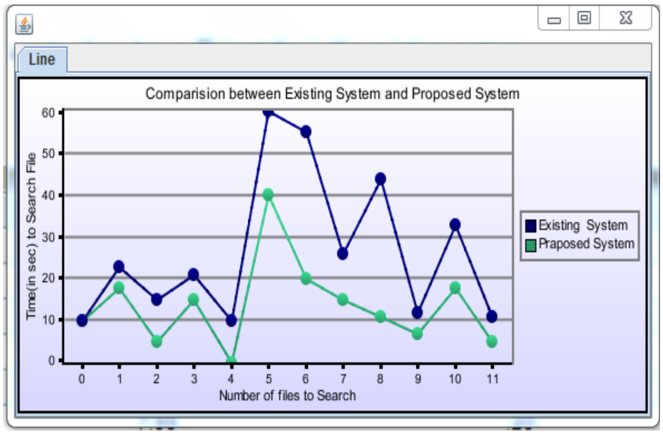

Figure 8: Comparisons between Existing System and Proposed System

The proposed system provides the best result using the document clustering technique that can facilitate the discovery of new and useful knowledge from the document under analysis. In the existing system the file searching process was very time consuming, in proposed system the automated method for cluster labelling is used, so that the proposed system provide the efficient data within small amount of time comparing the existing system.

Table3 gives the comparison between the existing and proposed system which shows that how effectively the relevant data is retrieved. This does the faster operation with greatest accuracy which is the key point of forensic analysis.

VI. CONCLUSION

Many data mining techniques have been proposed in the last decade. The outcome of the

information retrieval using document clustering for forensic analysis in this thesis is the number of labelled cluster, which provides the better visualization in the form of frame which shows the most relevant data present in the particular cluster. The assignment of labels to clusters enables the expert examiner to identify the content of each cluster more quickly—eventually even before examining their contents. K-mean algorithm is effective; it minimizes the squared-error criteria computationally efficient and does not require the user to specify many parameters.

Partitional algorithm that organize large amount of information into small number of meaningful cluster gives better clustering performance.K-means is a classical partitioning technique of clustering that clusters the data set of n objects into k clusters with k known a priori. K-mean algorithm minimizes the distance between the point labelled in the cluster and point designated as a centre of cluster but lead poor performance. Limitations-Inability to identify cluster with arbitrary shapes, ultimately imposing hyper-spherical shapes clusters on the data. The automatic labeling approach provides the fast and efficient analysis, reduce manual work, as system analyses all things the evidences gathered will be more accurate, also increase the performance of forensic analysis with speed up the computer inspection process.

Future Scope:

Aimed at further leveraging the use of data clustering algorithms in similar applications, Indexing is the critical task on Audio and Video files, a promising venue for future work involves investigating the automatic approach for indexing on the audio and video files which will help in the improvement of the forensic analysis while investigation process. Security and scalability is the issue that can solve in future and can provide the higher security.

VII. REFERENCES

[1] L. Filipe da Cruz Nassif, E. R. Hruschka, “Document Clustering for Forensic Analysis: An Approach for Improving Computer Inspection,” IEEE Transaction on Information Forensics and Security, Vol.8, No. 1, January 2013.

[2] K. Stoffel, P. Cotofrei, and D. Han, “Fuzzy methods for forensic data analysis,” in Proc. IEEE Int. Conf. Soft Computing and Pattern Recognition,, pp. 23–28, 2010.

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

[4] L. Liu, J. Kang, J. Yu, and Z. Wang, “A comparative study on unsupervised feature selection methods for text clustering,” in Proc. IEEE Int. Conf. Natural Language Processing and Knowledge Engineering, pp. 597–601,2005.

[5] A. Strehl and J. Ghosh, “Cluster ensembles: A knowledge reuse framework for combining multiple partitions,” J. Mach. Learning Res., vol. 3, pp. 583–617, 2002.

[6] Y. Zhao and G. Karypis, “Evaluation of hierarchical clustering algorithms for document datasets,” in Proc. CIKM, pp. 515–524, 2002.

[7] 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.

[8] 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.

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

[10] R. Xu and D. C.Wunsch, II, Clustering. Hoboken, NJ: Wiley/IEEE Press, 2009.

[11] A. L. N. Fred and A. K. Jain, “Combining multiple clusterings using evidence accumulation,” IEEE Trans. Pattern Anal. Mach. Intell., vol.27, number 6, pp. 835–850, Jun. 2005.

[12] Aggarwal, C. C. Charu, and C. X. Zhai, Eds., “Chapter 4: A Survey of Text Clustering Algorithms,” in Mining Text Data. New York:Springer, 2012.

[13] Y. Zhao, G. Karypis, and U. M. Fayyad, “Hierarchical clustering algorithms for document datasets,” Data Min. Knowl. Discov., vol. 10, number 2, pp. 141–168, 2005.

[14] K. Kishida, “High-speed rough clustering for very large document collections,”J. Amer. Soc. Inf. Sci., vol. 61, pp. 1092–1104, doi: 10.1002/asi.2131, 2010.

[15] Y. Loewenstein, E. Portugaly, M. Fromer, and M. Linial, “Effcient algorithms for exact hierarchical clustering of huge datasets: Tackling the entire protein space,” Bioinformatics, vol. 24, number 13, pp. i41–i49, 2008.

[16] D.Deshmukh, S. Kamble “Survey on Hierarchical Document Clustering Techniques Fihc & F2 Ihc” ISSN: 2277 128X, Volume 3, Issue 7, July 2013

[17] L. Vendramin, R. J. G. B. Campello, and E. R. Hruschka, “Relative clustering validity criteria: A comparative overview,” Statist. Anal. Data Mining, vol. 3, pp. 209–235, 2010

[18] L. Vendramin, R. J. G. B. Campello, and E. R. Hruschka, “Relative clustering validity criteria: A comparative overview,” Statist. Anal.Data Mining, vol. 3, pp. 209–235, 2010.

[19] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. In KDD Workshop on Text Mining, 2000.

[20] V. Levenshtein, “Binary codes capable of correcting deletions, insertions,and reversals,” Soviet Physics Doklady, vol. 10, pp. 707–710,1966.

[21] Garbriela derban and Grigoreta sofia moldovan, “A comparison of clustering techniques in aspect mining”, Studia University, Vol LI, Number1, pp 69-78, 2006.

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

[23] A. K. Jain and R. C. Dubes, Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[24] R. Xu and D. C.Wunsch, II, Clustering. Hoboken, NJ: Wiley/IEEE Press, 2009.

[25] G. Salton and C. Buckley, “Term weighting approaches in automatic text retrieval,” Inf. Process. Manage., vol. 24, no. 5, pp. 513–523, 1988.*ons*.