Forensic Analysis using Document Clustering: A Survey
Saad S. Ansari¹, Prof. T.M. Pattewar²

¹ Department of Computer Engineering, SES’s RCPIT, Shirpur, ansari.saad91@gmail.com
² Department of Information Technology, SES’s RCPIT, Shirpur, tareekpattewar@gmail.com

Abstract—Now days, Forensic analysis is doing crucial job in crime investigation. In earlier, it’s common that use of digital devices like computers is used for the analysis of crimes by investigation officers. The crimes for investigation are embrace hacking, drug trafficking, erotica, numerous larceny crimes etc. In computer there are multiple files are present in order to process them for the investigation of explicit crime by investigation officer. So it’s a difficult task for forensic examiner to do such analysis in quick period of time. That’s why to do the forensic analysis of documents within short period of time requires special techniques to make such complex task in a simpler approach. Such special technique includes Document Clustering. Because clustering is the unverified organization of designs that is data items, remarks, or feature vectors into groups. Clustering algorithms such as K-Means, K-Medoids, Single Link, Complete Link and Average Link can simplify the detection of new and valuable information from the documents under investigation. This survey paper gives study of different existing Document clustering methods in accordance with computer forensic analysis. We also give comparative study of different computer forensic analysis techniques and enhance clustering algorithm which will improve accuracy of clustering to finding relevant documents from huge amount of data.

Keywords-forensic analysis, crimes, clustering algorithms, text mining

I. INTRODUCTION

The Forensic Investigation is the process in which the digital devices like computers are used to analyze the digital evidence from a particular computing device in a way that is proper for presentation in a court of act. It also deals with the preservation, identification, extraction as well as documentation of digital evidences. It is task of analyzing huge number of files from computer seized devices. But in computer forensic procedure all the information and files are stored in digital form. This digital information stored in computer seized devices has an important factor from an investigative point of view which treated as evidence in the court of law to prove what occurred based on such evidences [2].

Clustering algorithms are generally used for exploratory data analysis, where there is very little or no previous data about the data. From a lot of technical viewpoint, data contains unlabeled objects—the categories or classes of documents which will be found are a priori unknown. Moreover, even assuming that labeled datasets may be available from previous analyses, there is almost no hope that a similar categories would be still valid for the upcoming data, obtained from different computers and associated to different investigation processes. More exactly, it is possible that the new knowledge sample would come from a different population. In this context, the utilization of cluster algorithms, which are capable of finding latent patterns from text documents found in seized computers, will enhance the analysis performed by the professional examiner [1].

The principle of clustering algorithms is that objects within a valid cluster are additional the same as one another than they are to objects belonging to a distinct cluster. Thus, once information knowledge partition has been induced from data, the expert examiner may initially target reviewing representative documents from the obtained set of clusters. Then, after this preliminary analysis, he could eventually commit to scrutinize alternative documents from every cluster. By doing so, one can avoid the hard task of examining all the documents but, even if so desired, it still may be done [1]. The forensic experts have restricted time available for performing examinations. Thus, it's reasonable to assume that, when finding a relevant document, the examiner may grade the analysis of alternative documents belonging to the cluster of interest, because it's seemingly that these are also...
relevant to the investigation. Such an approach, based on document clustering, will so improve the analysis of seized computers.

The principle of clustering algorithms is that objects among a valid cluster are further the same as one another than they are to objects belonging to a distinct cluster. Thus, once information partition has been induced from knowledge, the professional examiner could initially target reviewing representative documents from the obtained set of clusters. Then, once this preliminary analysis, he may eventually decide to scrutinize various documents from each cluster. By doing so, one will avoid the exhausting task of examining all the documents however, even if so desired, it still could also be done [2].

II. RELATED WORK

L.F.C Nassif has been proposed an approach that applies document clustering algorithms for the forensic analysis of computer devices. They illustrated an approach by carrying out wide experimentation with six well known clustering algorithms (K-mean, K-medoids, Single Link, Average Link, complete Link and CSPA) applied to five real world datasets obtained from computer seized. They were also studied uses of the comparative validity index criteria for the estimating the number of clusters in an automated manner which overcomes the limitations of previous techniques [1].

A. Maind has been proposed approach the forensic analysis was done very scientifically i.e. retrieved data is in unstructured format get particular structure by using high quality well known algorithm and automatic cluster labeling method. Two relative validity indexes were used to repeatedly approximation the amount of clusters with automatic labeling to it; which makes it very easy to retrieve most relevant information for forensic analysis. They proposed hybrid hierarchical algorithm such as density based clustering such as DBSCAN algorithm which had many features such as Discover clusters as random shapes, Handle noise and one scan. Which was better to achieve fast and efficient analysis [8]?

S.Oliver have proposed SOM-based algorithms were used for clustering files with intend of making the decision-making process performed by the examiners more efficient. The files were clustered by taking into account their creation dates/times and their extension. That kind of algorithm has also been used to cluster the results from keyword searches. The underlying hypothesis was that the clustered results can increase the information retrieval efficiency, because that could not be necessary to review all the documents found by the user any longer [9].

The use of clustering has been reported by only few studies in the computer forensics field. Basically, the use of classic algorithm for clustering data is described by most of the studies such as Expectation-Maximization (EM) for unsupervised learning of Gaussian Mixture Models, K-means, Fuzzy C-means (FCM), and Self-Organizing Maps (SOM). These algorithms have well-known properties and are widely used in practice [10].

The literature on Computer Forensics only reports the use of algorithms that assume that the number of clusters is known and fixed a priori by the user. Aimed at relaxing this assumption, which is often unrealistic in practical applications, a common approach in other domains involves estimating the number of clusters from data. Essentially, one induces different data partitions (with different numbers of clusters) and then assesses them. With a relative validity index in order to estimate the best value for the number of clusters. This work makes use of such methods, thus potentially facilitating the work of the expert examiner—who in practice would hardly know the number of clusters a priori [2], [3], [9].

III. METHODOLOGY

Computer forensic analysis involves the investigative the huge set of files. Between all of that files are not relevant to the forensic examiner interest. So analysis of those files and documents which are
out of interest tends to more time consuming task. So the key approach is to apply document clustering on such huge set of files and documents. As a result, these document clustering provides different set of clusters among which forensic examiner analyze only relevant documents related to investigation of reported case. It helps to improve speed of the forensic analysis process. It will also help for forensic examiner to analyze the files and documents by only analyzing representative of the clusters. The document clustering process involves the following phases such as collection of data, preprocessing, Apply document clustering algorithm.

A. Collection of data
Collection of data involves the processes like obtain the files and documents from the computer seized devices. The collection of such files and documents involves special techniques.

B. Preprocessing
Preprocessing is very important phase in proposed system which will used to reduce the noise, dimensionality computational complexity and loss of information. Preprocessing involves [1]:

1) Tokenization: The procedure of break stream of text into words or phrases into tokens called as “Tokenization”. In a document, tokenization separates the sequence of characters into tokens by using punctuation and white space consider as separators. For example, regard as the string “John, Peter and Jim” produce the tokens such as: “John”, “Peter”, and “Jim”.
2) Stop word removal: Stop word removal is used to save space and to speed up searching procedure; the words which are considered as less significant should be removed. Any group of words can be chosen as stop word for instance “the”, “which”, “what”, “at”, “on”, etc.
3) Stemming: Stemming technique is used to reduce the word to its root or stem. The input terms used in document are expressed by stem rather than original words. For example, consider the words “Inviting”, “invited”, “invitation”, and “invites” can be reduced to the root word, “invite”.
4) Indexing: Indexing is done by using vector space model, Vector containing the frequencies of occurrences of words, which are defined as delimited alphabetic strings, whose number of characters is between 4 and 25. We also used a dimensionality reduction technique known as Term Variance that can increase both the effectiveness and efficiency of clustering algorithms. TV selects a number of attributes that have the greatest variances over the documents.

C. Estimate the number of cluster
In order to estimate the number of clusters, a wide used approach consists of obtaining a collection of data partitions with different numbers of clusters and so selecting that individual partition that provides the simplest result consistent with a selected quality criterion [1].

D. Clustering algorithms
K-means and K-medoids, the hierarchical Single/Complete/Average Link, and the cluster ensemble based algorithm known as CSPA [4], [5], [6], [7].

1) **K-means**: K-means is one of the simplest unsupervised learning algorithms that partition feature vectors into k clusters so that the within group sum of squares is minimized. K-means clustering is a method of vector quantization originally from signal processing that is popular for cluster analysis in data mining. K-Means follows a simple steps [4].
   - **Step 1**: Place randomly initial group centroids into the 2d space.
   - **Step 2**: Assign each object to the group that has the closest centroid.
   - **Step 3**: Recalculate the positions of the centroids.
   - **Step 4**: If the positions of the centroids didn't change go to the next step, else go to Step 2.
   - **Step 5**: End

2) **K-medoids**: K-medoids is same as k-mean but in k-medoids each time select object as medoids and then perform clustering on it [5].

Considering the partitional algorithms, it's widely known that each K-means and K-medoids are sensitive to initialization and typically converge to solutions that represent local minima. Attempting
to attenuate these issues, we have a tendency to use a nonrandom format during which distant objects from one another are chosen as beginning prototypes. Not like the partitional algorithms like K-means/medoids, hierarchical algorithms like Single/Complete/Average Link provide a hierarchical set of nested partitions, typically represented within the form of a dendrogram, from that the simplest variety of clusters may be calculable. In particular, one will assess the quality of each partition represented by the dendrogram [11].

3) Hierarchical: Hierarchical algorithms (Single/Complete/Average Link), assess every partition from the resulting dendrogram by means of the silhouette [6]. Then, the best partition (elected according to the relative validity index) is taken as the result of the clustering process. For each partitional algorithm (K-means/medoids), we execute it repeatedly for an increasing number of clusters. For each value of k, a number of partitions achieved from different initializations are assessed in order to choose the best value of and its corresponding data partition, using the Silhouette, which showed good results in [12] and is more computationally efficient. In our experiments, we assessed all possible values of in the interval [2, N] where is the number of objects to be clustered [6].

4) CSPA: The CSPA algorithm essentially finds a consensus clustering from a cluster ensemble form by a set of different data partitions. More precisely, a similarity matrix [10] is computed. Each element of this matrix represents pair-wise similarities between objects. The similarity between two objects is simply the fraction of the clustering solutions in which those two objects lie in the same cluster. Later, this similarity measure is used by a clustering algorithm that can process a proximity matrix—e.g., K-medoids—to produce the final consensus clustering. The sets of data partitions were generated in two different ways: (a) by running K-means 100 times with different subsets of attributes and (b) by using only two data partitions, namely: one obtained by K-medoids from the dissimilarities between the file [7].

E. Dealing with outlier
A simple approach for removing outliers simply remove singleton. The approach makes recursive use of the silhouette. Fundamentally, if the best partition chosen by the silhouette has singletons (i.e., clusters formed by a single object only), Then, the clustering process is repeated over and over again—until a partition without singletons is found. At the end of the process, all singletons are incorporated into the resulting data partition as single clusters [1].

CONCLUSIONS
By doing the survey on computer forensic analysis it can be conclude that clustering on data is not an easy step. There is huge data to be cluster in compute forensic so to overcome this problem, this survey paper present an approach for document clustering methods to forensic analysis of computers seized in police investigations. Again by using Labeling technique there will be document clustering for forensic data which will be useful for police investigations.

REFERENCES


