International Engineering Journal For Research & Development
Vol.3
Issue 2

E-ISSN NO:-2349-0721

Impact factor : 2.25

A NOVEL APPROACH FOR DOCUMENT CLUSTERING IN DIGITAL FORENSIC ANALYSIS

[1]Ms. Rajnee Kanoje
MTech. Computer Science & Engineering
SBITM College of Engineering, Betul, India
Email:rajnee03kanoje@gmail.com

[2]Dr. S. D. Choudhari
Professor SBITM College of Engineering
Betul, India
Email: choudhari.sachin1986@gmail.com

ABSTRACT:

Now A days, criminal's uses recent new technologies as well as technical methods to commit crimes like money fraud, unethical hacking, fraud in various domains and prohibited access etc. So, the investigation of such cases is tricky and more important task. That's why need to do the forensic analysis. In recent times digital forensics analysis has become a most important activity in crime investigation since computers are gradually more used as tools to commit various crimes. Throughout forensic investigation the digital devices such as desktops, , smart phones, notebooks, PDAs etc. found at the crime scene are collected for further investigation .In digital forensic analysis, huge amount of files are generally need to examined. Much of the data in those files consists of indistinct text, whose investigation by computer examiners is very tough to accomplish. Digital forensics deals with such huge set of documents to collect the evidence from computer devices. So, to do digital forensic analysis time limit play key role. So it's a not easy task for examiner to do such analysis in short period of time. Thus to do the digital forensic analysis of documents within short period of time requires particular techniques to make such complex task in a simpler approach. Such special technique includes document clustering. So, clustering algorithms are best choice for such operations. This document clustering analysis is very helpful for crime investigation to analyze the information from seized digital devices. In this paper we proposed novel approach to attain more efficient document clustering in forensic analysis. The accuracy of clustering of documents may improve by means of this novel approach.

Keywords: Document Clustering, Forensic Analysis, Investigation, Crime

INTRODUCTION

A. What is Digital forensic analysis?

Digital Forensic analysis is the branch of systematic forensic analysis process for investigation of vital matter found in digital devices interrelated to computer crimes. Digital evidence equivalent to particular incident is

any digital data that provides idea about incident. The important part of Digital forensic process is to scrutinize the documents that present on suspect's computer. Due to increasing count of documents and larger size of storage devices makes very difficult to analyze the documents on computer. usually, digital forensics is the use of investigation and analysis technique to collect and protect evidence from a exacting computing device in a way that it acts as a proper evidence in a court of proceed .It also deals with the preservation, identification, extraction as well as documentation of digital evidences .This is task of analyze massive number of files from computer seized devices. But in computer forensic procedure all the essential information and files are stored in digital form. This digital information stored in computer seized devices has an key factor from an investigative point of view which treated as evidence in the court of law to prove what occurred based on such evidences. Therefore collection of evidence from seized devices is also task of forensic examiner. Digital evidence is defined as the information and data of investigative value that are stored on, received or transmitted by digital device. Such digital evidences needs to be collected from computer seized devices in order to confess the case in court of law. So such digital evidences have a great asset for the forensic examiner .So the key factor to improve such forensic analysis process requires document clustering technique The process of digital forensic analysis is shown in below figure 1

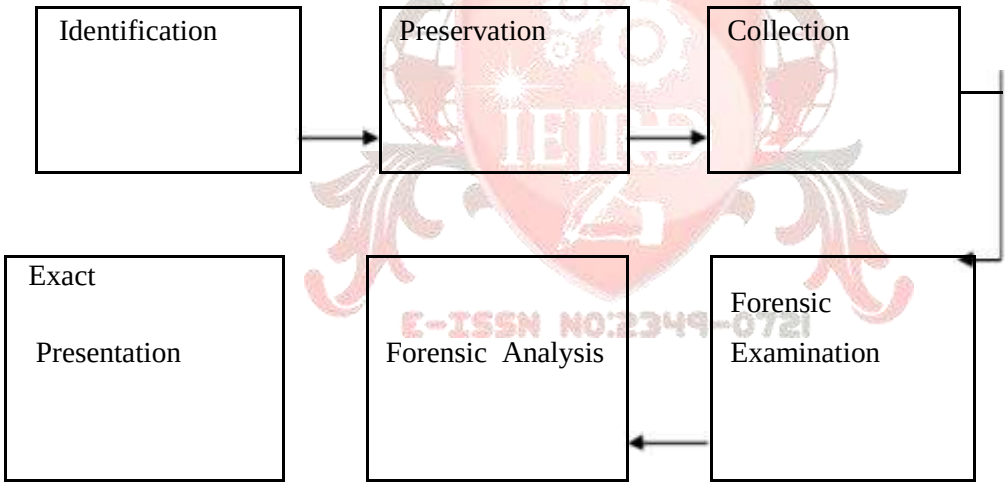


Figure1 : Forensics Analysis Phases

(Identification phase), the next step is to preserve the crime scene by stop or prevent several actions that can harm digital information being collected (Preservation phase). Follow that, the next level step is collect digital information that might be related to particular 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) [16]. The investigator can put the evidence together and tries to develop theories concerning events that occurred on the suspect's computer (Analysis phase). Finally the examiner review and the findings by explaining the reasons for each hypothesis that was formulated during the investigation

(Presentation phase). In the examination phases investigators often utilize certain forensic tools to help examine the collection files and perform an in detail organized search for pertinent evidence

B. Document clustering

Document clustering is the task of grouping similar data objects into subsets based on their relationship, i.e., objects with similar type of properties is grouped together where each subset is called a cluster. Each cluster contains a collection of objects that are similar between them. Objects in different clusters are “dissimilar" to each are requirements for clustering

- Scalability
- Different types of attributes/high dimensionality
- Minimal domain knowledge for determining features

The main advantage of document clustering is to retrieve the information effectively, reduce the search time and space, to identify the outliers, to handle the high dimensionality of data and to provide the review for related documents.

This paper is organized in the following way. In section 2 some earlier work is explained, and in section 3, explain the work of proposed system, section 4 conclude the proposed work.

I. LITERATURE REVIEW

Nassif etal [3], They illustrate 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 restrictions of previous techniques.

S. Mascarnes et al[4] major focused on a novel document clustering model that allows an examiner to semantically clusters the documents stored on a suspect's digital devices with the help of subject suggestions initially provided to him. They provided subject suggestion improves the accuracy and speeds up the process of finding the evidences in forensic analysis.

B.Vidhya etal [5],studied various text clustering and document clustering technique for digital forensic analysis .To improve digital forensic analysis they proposed K-mean algorithm and ant colony optimization algorithm. This was very significant among swarm intelligent algorithm. K-mean was one of the simplest algorithms for document clustering which was efficient to giving better clusters for huge amount of datasets.

T.Thopte etal [6],they discuss about preprocess formless document to structured data. For that they have idea to extract four features of each document like title sentences, numeric words, proper nouns and term weights. That makes There method much simpler than any other methods. Their proposed system neglecting unwanted extension’s considering only extensions which was rich in text like .pdf, .doc, .txt.. The grouping

of these scored values represents the most accurate clustered documents. Which was very competent for improving computer inspection in forensic analysis?

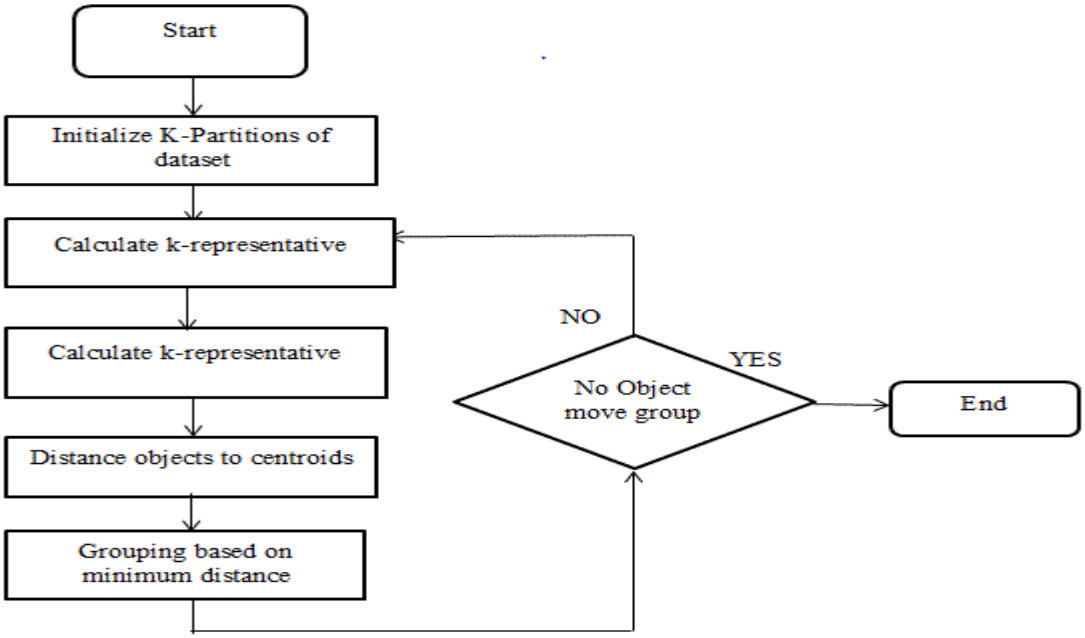
A.Maind etal [7], proposed approach on forensic analysis was done very scientifically i.e. retrieved data is in unstructured format get particular structure by using high quality well known algorithm and automatic cluster labeling method. They proposed hybrid hierarchical algorithm such as Density Based Spatial Clustering of Applications with Noise like DBSCAN algorithm which had many features such as Discover clusters as random shapes, Handle noise and one scan etc.

PROPOSED WORK:

By doing this literature survey we studied that the presented system have some problem such as accuracy, required more time for finding relevant document from huge amount of clusters that's why to overcome this problem we proposed new text clustering algorithm such as K-representative algorithm which will gives us the better computer forensic analysis. 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. It has been shown that K-representative algorithm is very efficient. Due to the modification proposed in forming representatives for clusters of categorical objects, the dissimilarity between a categorical object and the representative of a cluster is defined based on simple matching.

A. Process of document clustering in forensic analysis:

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, result clusters and forensic analysis shown in below figure .



Let's observe the 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 [21].

It has been shown that K-representative algorithm is very efficient. Due to the modification proposed in forming representatives for 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

$$\ldots \quad d(X,Q) = \sum_{j=1}^{m} \sum_{c_j \in D_j} f_{c_j,\delta(x_j}$$

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 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.

CONCLUSION

In this paper our survey shows how different document clustering techniques are used for digital forensic analysis with different phases involve in it. In addition to this, we present an approach for implementation of enhance novel text clustering algorithm which will forming clusters on the basis of relative match. It gives better results and improves the accuracy of clustering technique. By using this approach searching time for finding relevant document from massive amount of datasets will be reduce and improve the 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. S. Mascarnes and J.Gomes “Subject based Clustering for Digital Forensic Investigation with Subject Suggestion”, International Journal of Computer Applications (0975 – 8887) Volume 102– No.11, September 2014.
3. 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, Special Issue 1, March 2014
4. T.Thopte, Y. Indani, M.Jangale and S.Gaikwad,“Heuristic Approach for Document Clustering in Forensic Analysis”, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (1) , 2015, 182-185
5. 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.

6. 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.

7. C. Jadon and A. Khunteta, " A New Approach of Document Clustering", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 4, April 2013.

8. G. Thilagavathi and J. Anitha, "Document Clustering in Forensic Investigation by Hybrid Approach", International Journal of Computer Applications Vol. 91, April 2014.

9. B. K. L. Fei, J. H. P. Eloff, H. S. Venter, and M. S. Oliver, "Exploring forensic data with self-organizing maps", Internatinal Conference Digital Forensics, 2005.

10. K. Nagarajan and Dr. M. Prabakaran, "A Relational Graph Based Approach using MultiAttribute Closure Measure for Categorical Data Clustering", The International Journal Of Engineering And Science (IJES) ,Vol. 3, 2014.

11. 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, 2009.

12K. Stoffel, P. Cotofrei, and D. Han, "Fuzzy methods for forensic data analysis", IEEE International Conference Soft Computing and Pattern Recognition, 2010.

13C. C. Charu, and C. X. Zhai, Eds., "Chapter 4: A Survey of Text Clustering Algorithms", Mining Text Data. NewYork: Springer, 2012.