An Empirical Approach for Document Clustering in Forensic Analysis: A Review

Tanushri Potphode, Prof. Amit Pimpalkar

Abstract:
Now a day, in the world of digital technologies especially in computer world, there is a great increase in crimes like ethical hacking, fraud in different domains, illegal access etc. Thus surveying such information is a critical and more important task. In such investigation computer seized devices plays an important role. Computer forensics deals with analyze such huge set of documents to collect the evidence from computer devices. So, to do computer forensic analysis time limit is an also major factor. 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. This paper reviews different existing Document clustering methods in accordance with computer forensic analysis. In this paper we also give comparative study of different computer forensic analysis techniques and we propose enhance clustering algorithm which will improve accuracy of clustering to finding relevant documents from huge amount of data. Which helps improves the document clustering for forensic analysis.

Keyword: document clustering, forensic analysis, text clustering , digital investigation.

I. INTRODUCTION

A. What is Digital forensic analysis?

In general, Digital forensics is the application of investigation and analysis technique to collect and defend 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. Therefore collection of evidences from seized devices is also key 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 admit 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 techniques. The process of digital forensic analysis is shown in below figure 1

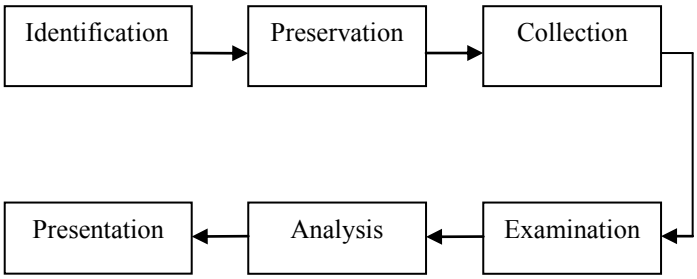

Figure 1: Process of Digital Forensic Investigation (DFI)

Above Figure1 Illustrates the Digital Forensic Investigation (DFI) process as defined by DFRWS. After determining items, components, and data related with the unpleasant incident (Identification phase), the next level 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 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) [15]. The investigator can put the evidence together and tries to develop theories regarding events that occurred on the suspect's computer (Analysis phase). Finally the examiner summarize and the findings by explaining the reasons for each assumption that was formulated during the investigation (Presentation phase). In the examination phases investigators often utilized certain forensic tools to help examine the collection files and perform an in detail systematic search for pertinent evidence.

B. Text Mining: Review

Text mining involves the applications of technique from areas such as information retrieval system, natural language processing, information extraction and data mining. These are a collection of stages of a text mining process can be combined into a single workflow.

- Information Retrieval (IR):

Systems recognize the documents in a collection which match a user's query. The most famous IR systems are search engines such as Google, which identify this document on the World Wide Web that are applicable to a set of given words. Information systems are obtain used in libraries, where the documents are classically not the books themselves but digital records containing information about the books. IR systems allow us to narrow down the set of files that are related to a particular problem. While text mining involves extremely computationally-exhaustive algorithms to large document collections, IR can speed up

the analysis considerably by reducing the number of documents for analysis. for instance if we are concerned in mining information only about interactions, we might restrict our analysis to documents that contain the name of a protein, or some form of the verb 'to interact' or one of the synonyms.

- Natural Language Processing (NLP)

This is the oldest and mainly difficult problems in the field of artificial intelligence. It is method for analysis of human language in order to computers can understand natural languages as humans perform. Although this objective is still away off, NLP can execute some types of analysis with a high degree of success. Shallow parsers recognize only the main grammatical elements in a sentence, for instance noun phrase sentence and verb phrases sentence, whereas deep phrases sentence generate a complete representation of the grammatical structure of a sentence. Role of the NLP in text mining is to supply the systems in the information extraction phase with linguistic data that they need to perform their task. This is done by annotating files with information like sentence boundaries, part-of-speech tags, parsing results, which can then be read by the information extraction tools. The aim is to uncover past unknown, useful knowledge.

- Information Extraction (IE):

This is the procedure of automatically obtain structured data from an unstructured natural language document. This involve defining the general form of the information that we are interested in as one as or more than templates, which are used to guide the extraction process.

C. Document clustering

Document clustering is the process of grouping similar documents into cluster. The main benefit 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 summary for similar documents. It provides the efficient way of representing and visualizes the documents in which it provides better navigation. The Similarity measure used to find similarity between documents, document representation, and algorithm or technique used to cluster the documents plays major role in document clustering. The document clustering simplifies the job of forensic examiner in forensic investigation. The paper outlines the significance document clustering in computer forensic analysis process.

This paper is present in the following way. In section 2 some earlier work is explained, Section 3 present comparative study of document clustering techniques and in section 4, explain the work of proposed system, section 5 conclude the work.

II. LITERATURE REVIEW

L.F.C Nassif etal [1], 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.

A.Maind etal [2], 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.

S. Karol etal [3], suggested fast as well as high-quality document clustering algorithms which plays very important role in document clustering for effective navigation, summarization, and organization of information. They suggests two techniques for efficient document clustering; these suggested techniques relating the application of soft computing concept as an intelligent hybrid PSO based algorithm. The two approaches are partitioning clustering algorithms Fuzzy C-Means (FCM) and K-Means each hybridized with Particle Swarm Optimization (PSO).

G. Thilagavathi etal [4], studied computer forensic process is to examine the documents present in suspect's computer. Due to enhance amount of documents and larger size of storage space devices makes very difficult to evaluate the documents on computer. To overcome those problems, they have been proposed a subject based semantic clustering technique along with bisecting-k means that allows the examiner to examine and cluster the documents based on particular subject and also the terms that does not belong to any subject. For that they proposed Subject vector space model (SVSM), In SVSM, making of vectors and document relationship function will be done in which it finally forms the cluster based on match between the files. Vector creation process is done by compare the terms with absolute Synonym List, Word Net and by Top Frequent terms. Document-subject comparison function determines the similarity of term with in subject and within document. By means of Document-subject clustering, similar documents are clustered based on subjects. There are some terms which do not belong to any subject. The accuracy of clustering of files has been improved by means of this enhance approach.

K. Nagarajan etal [5], studied conventional clustering approaches suffer with the scalability of number of attributes base on which the clustering was performed. There was approaches to cluster data point with a lot of attributes but suffers with overlap and numerous iteration required to perform clustering, also the measure computed for the variation of data points between cluster also will not be efficient when doing with several attributes. To overcome this problem they provided a new graph based approach which represents the relation between the data points and clusters. The relational graph consists of various vertices and edges, each vertex represent a data point. They computed the attribute closure, using all the attributes of the data points. They were also used threshold method to select the data point has closure to other one and the value of threshold was set based on number of attributes the data point has. They proposed method that produces good results which was compare to other approaches discussed in that period and they have been their method with various data sets.

S.Oliver etal, [6] 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 in [13] in order 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.

R. Hadjidj etal [7] developed an integrated approach for mining e-mails for forensic analysis, using classification and clustering algorithms. F. Iqbal etal [8], suggested related application domain, e-mails are group by using lexical, syntactic, structural, and domain-specific features.

S. Tacconi etal [9], provided three clustering algorithms (K-means, bisecting K-means and EM) were used. The problem of clustering e-mails for digital forensic analysis was also addressed in that, based variant of K-means was applied. The obtained results were analyzed subjectively, and the authors concluded that they are interesting and useful from an investigation perspective. K. Stoffel etal [11], provided methodology and an automatic process for inferring accurate and easily clear expert-system-like rules from forensic data. This method was based on the fuzzy set theory.

B.Umale etal [10], studied method of analyzing the various crimes using the computer based methods called as digital forensic analysis (DFA). The main input to that system has number of raw and unstructured input text files. In computer there are multiple files were present in order to process them for the investigation of particular crime by investigation examiner. Therefore, to automate this process there was many methods and tools presented for forensic investigations. The key part of that tools or methods was the use of clustering algorithms in which the number of unstructured text files was given as input and the output is generated in structured format. They provided the document clustering methods were used for digital forensic analysis. They were present the various text analysis methods using clustering algorithms.

B.Vidya etal, [12], studied digital forensics deals with the analysis of artifact on all types of digital devices. The function of digital forensics is to facilitate the study of criminal activities that involve digital devices, to preserve, gather, analyze and provide the scientific and technical evidence, and prepare the documentation for law enforcement authorities. They were also provided enhance technique Ant colony optimization algorithm was a very important among swarm intelligence algorithms. C.Charu etal [14], clustering is a extensively studied data mining problem in the text domain. The problem finds many applications in customer segmentation, mutual filtering, visualization, document organization and indexing.

III. COMPARATIVE STUDY OF FORENSIC ANALYSIS TECHNIQUES

As we discussed earlier many clustering algorithm has been used for document clustering in forensic analysis S. Oliver [6] Self Organizing Maps (SOM) to search the pattern in data set which facilitates task of forensic process. But the number of clusters needs to be specified. The J.G. Clark [13] applied clustering techniques in order to explore only relevant hits to forensic investigators. But, in real time to specify the number of clusters explicitly is not an easy task because forensic investigator does not know amount of data resides in computer seized devices. [1] Applies document clustering algorithm to computer seized devices for forensic analysis as well as overcame the limitation i.e. specifying the number of clusters explicitly using relative index validity criteria.

IV. PROPOSED WORK

By doing this literature survey we studied that the existing 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 as follows.

Let C = {X1, ... Xp} be a cluster of categorical Objects, with Xi = (xi,1, . . . , xi,m), 1≤ i ≤p, and X = (x1, . . . , xm) be a categorical object.

Assume that Q =(q1, . . . , qm), with qj = {(cj , fcj) | cj .Dj},is a representative of cluster C.

Now we define the dissimilarity between object X and representative Q by

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

Below we shown process of document clustering with phases of propose system in figure 2 and figure 3 shown the results [1] after applying clustering for forensic analysis

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 2.

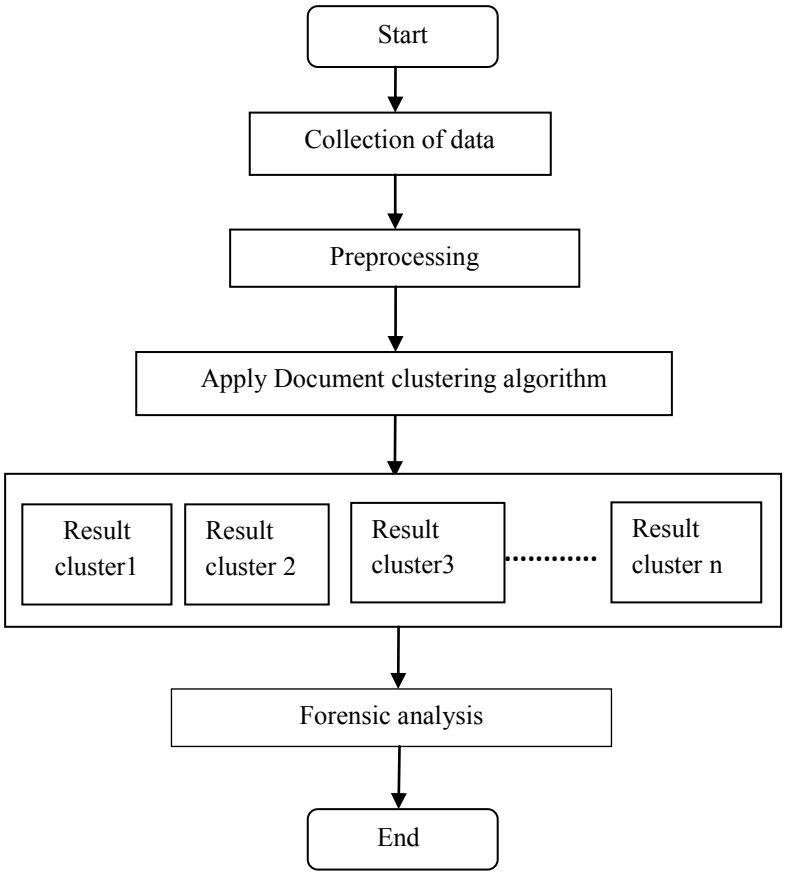

Figure 2: Phases of proposed system

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 the tokenization, stop words removal, Stemming process and Indexing sub-phases.

- 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 "Sonali, Priti and Rita" produce the tokens such as: "Sonali", "Priti", and "Rita".

- 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.

- 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. Porter Stemming technique [18] is used here to remove the stem words. For example, consider the words "Inviting", "invited", "invitation", and "invites" can be reduced to the root word, "invite". One time after preprocessing, the unique terms in a document set are representing as T.

- Indexing:

Indexing is defines as how many times the particular term will be appear within document.

c. Apply Document Clustering Algorithm:

After the preprocessing document clustering is applied to form the set of clusters according to specified clustering criteria.

d. Result clusters:

After applying algorithm we get result clusters It is used for application such as forensic analysis in which clustering results are used for further analysis.

e. Forensic Analysis:

Forensic analysis process uses the result of document clustering for further analysis. The result of document clustering enhances the forensic process within sake of time.

Clusters	Information
C1	3 blank documents
C2	4 Financial Transactions
C3	2 maternity Payments
C4	2 Grocery lists
C5	1 Foreign exchange transaction warning 1 list of documents for information
C6	2 sample documents from office information
C7	1 notice about working hours 1 check receipt
C8	1 investment club status 1 agreement for joining club
C9	8 receipts of foreign exchange insurance transaction
C10	2 warning about foreign brokerage business hours

Figure 3: Document clustering for forensic analysis

V. 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 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 huge amount of datasets will be reduce and improve the efficiency of forensic analysis.

ACKNOWLEDGMENT

I explain my sense of appreciation towards my project guide Prof. Amit Pimpalkar for him valuable guidance at each step of study of this project, also him contribution for the solution of every problem at each stage. I am also thankful to Prof. Amit Pimpalkar for the motivation and inspiration that trigger me for this paper work.

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 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.

[6] 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.

[7] 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.

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

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

[10] B.Umale and M.Nilav, "Survey on Document Clustering Approach for Forensics Analysis", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3), 2014.

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

[12] B.Vidya and P.Vaijayanthi, "Enhancing digital forensic analysis through document clustering", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3, January 2014.

[13] N. L. Beebe and J. G. Clark, "Digital forensic text string searching: Improving information retrieval effectiveness by thematically clustering search results", Digital Investigation, Elsevier, vol. 4, 2007.

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

[15] M. R. Clint, M. Reith, C. Carr, and G. Gunsch, an Examination of Digital Forensic Models (2003).

BIOGRAPHIE

Tanushri Giridhar Potphode, M-Tech student from G.H. Raisoni academy of engineering and technology Nagpur, Maharashtra, India. Her areas of interests is Data Mining and NLP (Natural Language Processing)

Prof. Amit Pimpalkar, Assistant Professor at G.H. Raisoni Academy of Engineering & Technology, Nagpur, Maharashtra, India.. His areas of interests are Natural Language Processing, Data Mining, Data Structure and Image Processing.

