ABHIYANTRIKI
An International Journal of Engineering & Technology
(AIJET)
Vol. 1, No. 2 (December, 2014) http://www.aijet.in/ eISSN: 2394-627X

A Practical Approach for Document Clustering in Computer Forensic Analysis

Tanushri Potphode*
Deptt. of Computer Science & Engineering,
G.H.R.A.E.T, Nagpur University, INDIA

Amit Pimpalkar
Deptt. of Computer Science & Engineering,
G.H.R.A.E.T, Nagpur University, INDIA

Abstract

In recent years, level of crime is becoming severe problems in several countries. In today's world, criminals have greatest use of all new technologies and technical methods in commit crimes like ethical hacking, fraud in different domains, prohibited access etc. For that the law enforcers have to effectively meet out challenges of crime control and maintenance of public order. So, the investigation of such cases is difficult and more important task. That's why need to do the forensic analysis. In digital forensic analysis, thousands of files are generally examined. Much of the data in those files consists of indistinct text, whose investigation by computer examiners is very tough to achieve. Digital forensics deals with study such huge set of documents to collect the evidence from computer devices. So, to do digital forensic analysis time limit is an also major factor. So it's a difficult task for examiner to do such analysis in quick period of time. That's why to do the digital 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. So, clustering algorithms are of great interest. This document clustering analysis is very helpful for crime investigation to analyze the information from seized digital devices like computers, laptops, hard disks, tablets. In this paper we proposed practical approach to achieve efficient document clustering in computer forensic analysis. And show the implementation of proposed system. We also propose enhance text clustering algorithm which will improve accuracy of clustering to finding relevant documents from huge amount of data which will helps to improve the document clustering for forensic analysis.

Keywords: *Document Clustering, Forensic Analysis, Investigation, Crime.*

Author for Correspondence tanushripotphode44@gmail.com

1. Introduction

A. What is Digital Forensic Analysis?

Digital Forensic analysis is the branch of systematic forensic analysis process for investigation of matter found in digital devices interrelated to computer crimes. Digital evidence equivalent to particular incident is any digital data that provides suggestion about incident. The important part

of Digital forensic Process is to examine 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 is proper for presentation 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 enormous 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.

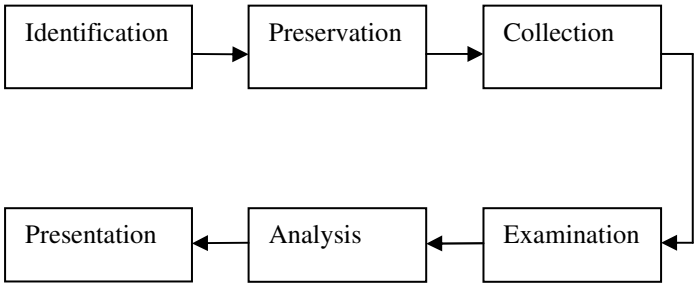

Fig. 1: Process of Digital Forensic Investigation (DFI)

Above figure 1 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) [14]. 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 reviews 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 systematic search for pertinent evidence.

B. Document Clustering

Clustering is the task of grouping data objects into subsets based on their relationship i.e., objects with similar properties are grouped together. 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:

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

Document clustering is the process of grouping similar documents into cluster. The main benefit 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. It provides the well-organized 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 importance of document clustering in computer forensic analysis process.

2. Literature Review

T. Potphode et al. studied various approaches for document clustering in forensic analysis to improve the efficiency of forensic analysis using computer seized investigation for that they studied various text clustering algorithms. L.F.C Nassif et al. proposed 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. B. Vidhya et al. 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 important among swarm intelligent algorithm. K-mean was one of the simplest algorithms for document clustering which was efficient to giving better clusters form huge amount of datasets. T. Thopte et al. proposed system is preprocessing 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 their method much simpler than any other methods. They proposed system neglecting unwanted extension's considering only extensions which were rich in text like .pdf, .doc, .txt. As the final step of clustering, system creates a score matrix of all the documents by comparing with one another to yield a score matrix which was contains aggregate feature score. The grouping of these scored values represents the most accurate clustered documents which was very efficient for improving computer inspection in forensic analysis. A. Maind et al. 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. They proposed hybrid hierarchical algorithm such as Density Based Spatial Clustering of Applications with Noise such as DBSCAN algorithm which had many features such as Discover clusters as random shapes, Handle noise and one scan .good for data sets with large amounts of noise, allows a compact mathematical description of arbitrarily shaped clusters in high-dimensional data sets which was better to achieve fast and efficient analysis. S. Karol et al. 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). C. Jadon et al. studied of various clustering technique like K-means Agglomerative Hierarchical Clustering. In addition to that various clustering techniques they discusses about various document-representing techniques in

graph and particular Vector Space Model and Matrix Representation. After studying all these things, they created a new approach of clustering algorithm and also a new representation technique of documents. They compared their results with K-means Algorithm which was given giving good results.

G. Thilagavathi et al. studied computer forensic process is to examine the documents present in suspect's computer. Due to huge amount of amount of documents and larger size of storage space devices makes very difficult to evaluate the documents on computer. To overcome those problems, they had propose 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). The accuracy of clustering of files has been improved by means of this enhance approach. S. Oliver et al. proposed SOM-based algorithms were used for clustering files with aim of making the decision-making process performed by the examiners more resourcefully. The files were clustered by taking into account their creation dates/times and their extension. That kind of algorithm has also been used in order to cluster the results from keyword searches. The underlying assumption was that the clustered results can increase the information retrieval efficiency, because that could not be necessary to assess all the documents found by the user any longer. K. Nagarajan et al. 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. 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. R. Hadjidj et al. developed an integrated approach for mining e-mails for forensic analysis, using classification and clustering algorithms. K. Stoffel et al. provided techniques and an automatic process for infer accurate and simply clear expert-system-like rules from forensic data. That method was based on the fuzzy set theory. C. Charu et al. clustering is an extensively studied finds many applications in customer segmentation, mutual filtering, visualization, document organization and indexing.

3. Proposed Work

Let's examine closely 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 implementing enhance technique to accomplish this proposed approach. We implement 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

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

In our proposed system, investigator firstly defines the subject vectors i.e. the terms belong to particular subject such as crime. Afterward suspect document set is preprocessed into which parsing, indexing and analyzing to find the terms. Preprocessing involves the stop word removal and stemming words procedure. Indexing provides how many times the particular term come out with in document. Weight for each term in each document is provided by term frequency and Inverse document frequency.

Implementation

Forensic analysis involves the analytical huge set of documents. Among 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. Accordingly, these document clustering provides different set of clusters amongst which forensic examiner analyze only associated documents related to investigation of reported case. It helps to improve speed of the forensic analysis process. It will also assist for forensic examiner to analyze the files and documents by only analyze representative of the clusters. The forensic examiner is involved in clustering the collection of documents based on their related subject initially the examiner defines the set of terms that belongs to particular subject. Now, here we show the implementation of proposed system in figure 2.

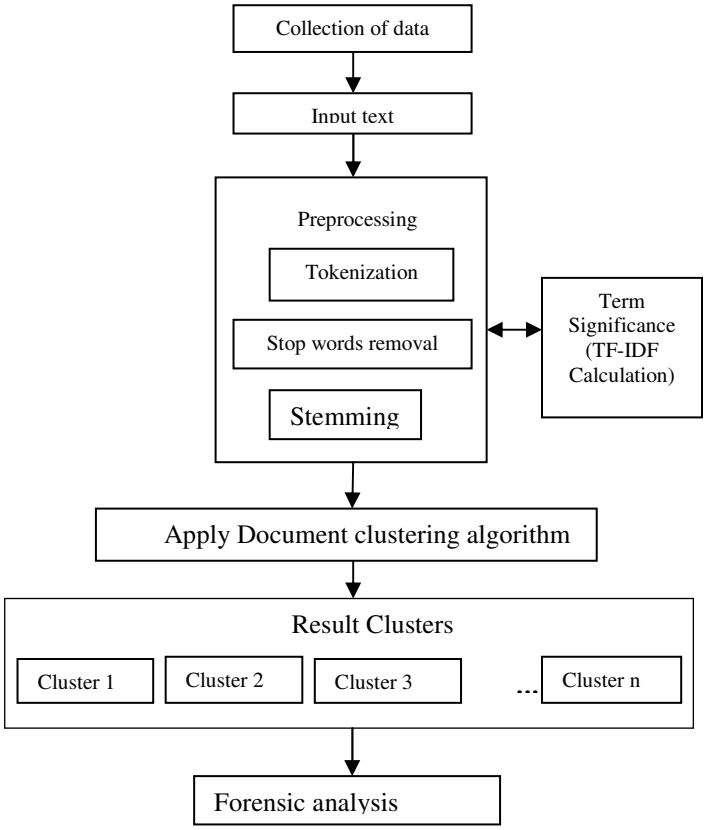

Fig. 2: Implementation of Proposed System

a. Collection of Data

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

Maintain Directory Lookup

We maintain directory because for every input term, expansion vector is generated by looking related words for each term in Extended Synonym List (ESL). Extended Synonym List is an event specific list which contains forensic-related synonyms and acronyms terms which can be used by criminals and it will not present in dictionaries. These directories contain words which are related to the crimes.

b. Preprocessing

Preprocessing is very important phase since it can affect the result of a document clustering algorithm. In the implementation of proposed system which will used to reduce the noise, dimensionality, computational complexity and loss of information. So it is necessary to preprocess the data logically. Preprocessing involves the tokenization, stop words removal, Stemming process and Indexing sub-phases.

Tokenization

The method of breaking stream of text into words or phrases into tokens called as "Tokenization". In a document, tokenization separate the sequence of characters into tokens by using punctuation and white space consider as separators. For example 1, regard as the string "Shama, Vini and Gita" produce the tokens such as: "Shama", "Vini", and "Gita".

For example 2,
Input: Ajinkya Roy will join the board as an executive director.
Mr. Ray is the chairman of Ericson.
Output: Ajinkya Roy will join the board as an executive director. Mr. Ray is the chairman of Ericson.

Stop Word Removal

Stop words are words measured not to convey any meaning. We use a standard list of stop words and remove them from the collection of documents. It is used to save space and to speed up searching procedure; the words which are considered as less significant must be removed. Any group of words can be chosen as stop word for instance 'the', 'which', 'what', 'at', 'on', etc.

Stemming

Stemming method 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 "accomplishing", "accomplished", and "accomplishes" can be reduced to the root word, "accomplish".

Indexing

Indexing is the procedure of how many times the particular term will be appear within document.

c. Term Significance

Term significance defines the control of term with in a document. Weight and meaning of term can be computed by TF-TDF calculation. TF-IDF is used to determine the weight of each term in information retrieval and text mining. It evaluate the importance of word is to a document with

in a collection. Term frequency determines how frequently particular term appear in document. Inverse document frequency evaluates how important the term is.

In particular, each weighted term frequency must be determines as
$\mho(t,d)= tf(t,d)\times idf(t,D)$ (2)

Where tf (t, d) represents frequency of term t in document d and idf(t,D) denotes the inverse document frequency of term t in the document set D

$idf(t,D) = 1+\log(|D|/1 + Freq(t,D))$ (3)

Where, |D| denotes the count of documents in the huge set of documents D, and Freq (t, D) is the count of documents in D that denotes the term t. There will be more number of terms generated for each subject even though it is limited by entry value. To reduce the noise, dimensionality reduction technique such as term variance is used to limit the terms that improves the efficiency and effectiveness of clustering algorithm.

d. Apply Document Clustering Algorithm
After the preprocessing document clustering is applied to form the set of clusters according to specified clustering criteria.

e. 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.

f. 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

4. Conclusion

In this paper our contribution is to shows an practical approach for implementation of proposed system which will uses enhance text clustering algorithm which forming clusters on the basis of relative match. It also gives better results and improves the accuracy of document clustering technique. By using this approach searching time for finding relevant document from huge amount of datasets will be reduce and recover the efficiency of forensic analysis.

References

[1] T. Potphode and A. Pimpalkar, "An Empirical Approach for Document Clustering in Forensic Analysis: A Review" International Journal of Science, Engineering and Technology Research, Vol. 3 (11), November 2014.

[2] 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.

[3] 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.

[4] 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.

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

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

[7] 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.

[8] R. Hadjidj, M. Debbabi, H. Lounis, F. Iqbal, A. Szporer, and D. Benredjem, "Towards an integrated e-mail forensic analysis framework," Digital Investigation, Elsevie*r*, vol. 5, 2009.

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

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

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

[12] 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, Issue 1, March 2014.

[13] T. Thopte, Y. Indani, M. Jangale and S. Gaikwad, "Heuristic Approach for Document Clustering in Forensic Analysis", International Journal of Computer Science and Information Technologies, Vol. 6 (1), 182-185, 2015.

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