IJCSE International Journal of Computer Sciences and Engineering Open Access

Review Paper Volume-3, Issue-3 E-ISSN: 2347-2693

Role of Document Clustering For Forensic Analysis Investigation System

Vilas V Pichad[1*] and Sachin N Deshmukh[2]

[1*2]*Department of Computer Science and Information Technology, Dr.BAMU, Aurangabad, India*

www.ijcseonline.org

Received: Feb/21/2015 Revised: Mar/02/2015 Accepted: Mar/20/2015 Published: Mar/31/2015

Abstract— Today's digital word technologies information in computer world, there is extremely large increase in crime like money laundering, unauthorized access, ethical hacking, fraud detection in different domain etc. So, investigation of such cases deserve a lot more important, in this forensic investigation computer devices plays a major role. In Digital forensic analysis seized digital devices can provide precious information and evidences about facts and/or individuals on which the investigational activity is performed. In this paper we proposed document clustering algorithms to digital forensic analysis of computers seized devices in police investigations. In the Digital forensic analysis of investigation, used total six famous partition (k-means, k-medoids and CSPA) algorithm and hierarchical (Single Link, Complete Link and Average Link) document clustering algorithms are used. This is applied datasets for five real-world investigation cases conducted by the Brazilian Federal Police Department. Also two validity index are used to find out how many clusters are formed. This experiments show that the Average Link and Complete Link algorithms provide the best results for the application domain. This reviews different existing text clustering and Document clustering multithreading methods is used with computer forensic analysis.

Keywords— Digital Forensic Analysis, Document Clustering, Text Clustering,Multithreading**.**

I. Introduction

There is explored and utilize the huge amount of text documents is a major question in the areas of information retrieval and text mining. Document clustering (referred to as text clustering) is one of the most important text mining methods are used, to help the organizing a large amount of documents into a number of clusters. Today there is fast increase in crime relating cases to Internet and Computers device has caused a growing need for computer forensics. Computer forensic is analyzing huge number of files from computer seized devices, which is computer forensic tools, can be exist in the form of computer software. These tools have been developed to help computer forensic investigators in a computer investigation. But in computer forensic process 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 perspective which treated as evidence in the court of law to prove what occurred based on such evidences. So collection of evidences from digital devices is also key task of forensic analyst. When the forensic examiner collects all such evidences then his job is to establish based on the collected evidences. But such computer seized devices contains huge set of files and documents, so it is not easy to do the analysis of each and every files individually. In a more practical and realistic scenario, domain experts are scarce and have limited time available for performing examinations. When finding a particular document, the examiner could prioritize all the documents to the analysis. Such an approach, can indeed improve the forensic analysis of seized computers. At the same time court of law requires quick result of such cases to improve the speed of forensic analysis process has a more important. The clustering algorithms play important role in forensic analysis of digital documents since it contains very important, complex and unstructured data, to improve such forensic analysis process requires fast text clustering and document clustering techniques.

The process of analyzing large volumes of data may consume a very large amount of storage in small time in the storage media. The police investigations forensic data of clustering algorithms are typically used for examining data analysis, where there is datasets consist of unlabeled objects the classes or categories of documents that can be found are a prior unknown, this is exactly the case in several applications of computer forensics [1]. Therefore, we decided to choose set of (six) representative algorithms for clustering the document, namely: the partition cluster ensemble algorithm known as CSPA [5], K-means [2] and K-medoids [3], the hierarchical Single/Complete/Average Link [4], as shown in Table 1[1].

Table 1: Summary of algorithms and their parameters

\	Algorithm	Attributes	Distance	Initialization	K-estimate
Kms	K-means	Cont.(all)	Cosine	Random	Simp.Sil.
Kms 100	K-means	100>TV	Cosine	Random	Simp.Sil.
Kms 100*	K-means	100>TV	Cosine	[18]	Simp.Sil.
KmsT 100*	K-means	100>TV	Cosine	[18]	Silhouette
KmsS	K-means	Cont.(all)	Cosine	Random	Rec. Sil.
Kms	K-means	100>TV	Cosine	Random	Rec.Sil.

Corresponding Author: *Assist Prof Sachin N Deshmukh, sndeshmukh@hotmail.com, Department of Computer Science and IT, Dr.BAMU, Aurangabad., India*

100S					
Kmd 100	K-medoids	100>TV	Cosine	Random	Silhouette
Kmd 100*	K-medoids	100>TV	Cosine	[18]	Silhouette
Kmd Lev	K-medoids	Name	Lev.	Random	Silhouette
Kmd LevS	K-medoids	Name	Lev.	Random	Rec.Sil.
AL100	Average Link	100>TV	Cosine	____	Silhouette
CL100	Complete Link	100>TV	Cosine	____	Silhouette
SL100	Single Link	100>TV	Cosine	____	Silhouette
NC	CSPA	Name, Cont.(all)	CSPA	Random	Simp.Sil.
NC100	CSPA	Name 100>TV	CSPA	Random	Simp.Sil.
E100	CSPA	Cont.100 random	CSPA	Random	Simp.Sil.

As shown in table there are six various algorithms with their parameters like distance which has cosine as well as levenshtein distance (Lev), which is nothing but a string metric for measuring the difference between two sequences. The levenshtein distance (Lev) between two words is the minimum number of single character edits (i.e. insertions, deletions or substitutions) required to change one word to the other word. The application for levenshtein distance is to in approximate string matching the objective is to find matches for short and longer text strings, in these situations where a small number of differences are to be expected. Table 1 also gives the initialization of each algorithm. [1]

Attributes (words) consist of 100>TV chosen 100 words randomly from document content, that are the greatest variance occur over the documents. For, the k-estimate used Simp.Sil (Simplified Silhouette) and Rec.Sil (Recursive Silhouette) finding the single object. Where,* denotes the initialization on distant objects only.

II. LITERATURE SURVEY

The use of clustering algorithm has been reported by only few studies in the computer forensic analysis field [1]. 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 Model[12] used to requires an a-priori selection of model order, name is the number of components to be incorporated into the model, and the results is depend on initialization, K-means, Fuzzy C-means (FCM) is a clustering method of data analysis based on the fuzzy membership of each data point to each of the clusters of data formed . The objective of the fuzzy c-means algorithm is to minimize the sum of the weighted squared distances between the data points and the cluster centers. [13]These algorithms have well-known properties and are widely used in e-mail forensic for clustering data. Self-Organizing Maps (SOM) [6] based algorithms were used for clustering files with the aim of making the decision-making process performed by the examiners more quickly. The SOM is used to files were clustered by taking into account their creation dates/times and their file extensions. These types of algorithm have also been used in [7] order to cluster the results from keyword searches. The assumption is that the clustered results can increase the information retrieval efficiency, because it would not be necessary to all the documents found by the user any more.

In Filipe daCruz, Nassif and Eduardo Raul Hruschka [1] use various algorithms and preprocessing technique are used to provide better result in cluster data. At the last in the conclusion they have shown that, the various approach are presented by them to apply document clustering methods and algorithm in forensic analysis of computers seized in police investigations.

Also, [1] has reported and discussed with several practical results that can be very useful for forensic researchers. In their experiments, hierarchical algorithms known as Average Link and Complete Link [4] presented the best results as well as partition algorithms (K-means and K-medoids) can also provide the good results. It is shown these algorithms particularly for the studied application domain because the dendrograms that they provide summarized views of the documents being inspected, this is very helpful tools for forensic examiners to analyze textual documents from seized computers device [1].

The comparative Study on Unsupervised Feature Selection Methods for Text Clustering describe the problems in text mining and information retrieval area is text clustering. The performance of clustering algorithms will considerably reject for the high dimensionality of feature space and the inherent data. There, two techniques are used to deal with this problem i.e. feature extraction method and feature selection method. The Feature selection methods have been successfully applied to text categorization but seldom applied to text clustering due to the unavailability of class label information. There are unsupervised feature selection methods such as DF (document frequency), TC (text clustering) [10] and Reduction technique known as Term Variance (TV) is also used in the base paper to increase efficiency of clustering algorithms. Term variances are used to estimate top n words that the clusters are formed with relevant document. Which have greatest variance occurrences over the documents within clusters? Also it is important to find out distances between two documents used cosine-based distance and levenshtein based distance can also improve efficiency and accuracy of document clustering.[10]

In F. Iqbal, H. Binsalleeh, B. C. M. Fung, and M. Debbabi[8] usually they are suggest collecting e-mails written by multiple anonymous writers or authors and they most highlighting the problem of mining the writing style of those e-mails. The basic idea is to first cluster the anonymous e-mail by the Stylometric (Stylometry is the application of the study of linguistic style, usually to written language) features and then extract the write print, i.e., refer to the unique writing style, from each cluster.

In R. Hadjidj, M. Debbabi, H. Lounis, F. Iqbal, A. Szporer, and D. Benredjem[9] use learn about to preferred the individual word and isolate the character, it has targeted on lexical and syntactic features of an mail. The character-based features include the frequency of individual alphabets (26 letters of English), total number of; capital letters used in the beginning of sentences and also calculate the average number of characters per sentence, and average number of characters per word. For writing the e-mails authors are used to indicate the preference of an individual for certain special characters or symbols. For example most of the people prefer to use „&‟ symbol instead of word „and‟, and „$‟ instead of writing the word „dollar”. The syntactic features called as style markers function words such as „although‟, „there‟, punctuation such as „! ‟and „:‟, parts-of-speech tags and hyphenation etc. [9]

Forensic Data Analysis describes the fuzzy methods and an automatic procedure for clustering the document providing accurate and easily understandable system for forensic data. In the data analysis environments choosing the clustering algorithms and the various methodological terms used to be easily implementable. Fuzzy methods are used to improve the effectiveness and the quality of the data analysis phase for crime investigation. [13]

In the literature of the computer forensics analysis the use of algorithms that assume that the number of clusters are known and fixed by the priori user. Common approach in different domains involves estimating the number of clusters from data. It induces different data partitions (with different numbers of clusters) and then assesses them with relative validity index criteria [11], in order to estimate the best value for the number of clusters. This work makes use of methods, facilitating the work of the expert examiner would hardly know the number of clusters a priori. The main motto of this is to focus on self-organizing map (SOM) is an approach that allows computer investigators to visualize all the files on the storage medium and fixed them in locating their points of interest quickly by greatly reducing overall human investigation time and effort [6].

A. Document clustering and text clustering based techniques for forensic analysis

Usually, maximum computer forensic seized devices consist textual data as an input to digital forensic analysis process. The huge textual information is the important point for forensic analysis process. This large textual information exists in unstructured format which makes difficult job for forensic examiner. For further investigations Forensic examiner finds more difficulties during forensic analysis and to search the required patterns. Thus we want improved process for the forensic analysis of such computer seized devices which can be achieved by text clustering techniques.

i. Collection of Documents

Documents collection are used to collect large text document with various source for the purpose of clustering, including the processes indexing, filtering, crawling etc. which are used to store and retrieve in to the database. The basic purpose is to check the file contain content of our input file will be fetched by our software for further processing.

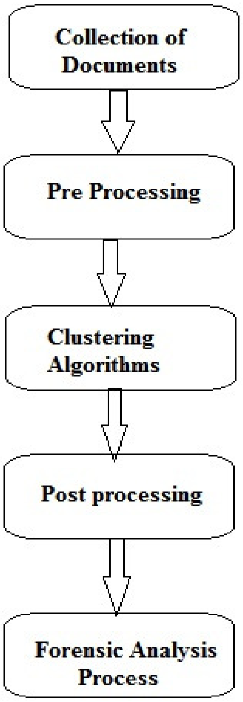

Figure 1: Process of Document Clustering and text clustering of DFI

ii. Preprocessing

It consists of steps that take as input a plain text document and output a set of cluster in the documents (which can be single terms) to be included in the vector model.

Stop word removal:

The document folder contains lot of input contains as stop words such as nouns, pronouns, adjectives, articles etc., These are not affect the real meaning of the document. The removal of stop words is the most common term filtering technique used. There are large numbers of standard stop word lists are available depending on the dataset quality.

Stop word Removal methods are:

• Remove the various terms with low frequencies in the document. This is done to improve the memory consumption and speed of the application.

• The similarities of the documents except date and postal codes numbers do not play much importance. Thus these

can also be removed. Assume our loaded input document contains the sentence as:
“They are going to college”.
In the above sentence, it contains words like ‘They’, ‘are’,’ to’ be stop words which are not needed for further Preprocessing.
Step i.e., stemming. So we will remove those words from our original sentence and we just pass words like college, going to further step.

Stemming:

Stemming is defined as reducing the words to their base form by removing suffix like -ing ,-ed etc. the words like ‘extracting’, ’extracts’ extracted, ‘extraction’ all are converted to stem ‘extract’, By Using the Porter Stemming Algorithm.
These are the necessary step to filtering the data in the text clustering.

iii. Document clustering algorithm

In the document clustering there is used partition algorithms [3], [4], and hierarchical set of algorithms [5] are broadly used to find out correct result. Trying, shrinking the clustering problems choosing distant objects from each other as starting prototype. The hierarchical algorithms frequently represent the best number of clusters showing the dendrogram. In particular, every partition of the best result and their graph represent the dendrogram [11].

iv. Post processing

It is used for application such as forensic analysis in which clustering results are used for further analysis.

v. Forensic analysis process

As discussed in post preprocessing 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. Hence, this clearly specifies the role of document clustering in the process of forensic analysis.

B. Data Set

We are use large textual documents written in various languages (Portuguese and English). These have been corrupted or originally created in different file formats are truly incomplete that they have been (partially) recovered from deleted data. They are assessing five datasets obtained from real-world investigation cases conducted by the Brazilian Federal Police Department. [1]

III. COMPARATIVE STUDY OF FORENSIC ANALYSIS TECHNIQUES

Previously, we have discussed Self Organizing Maps (SOM) to search the pattern in data set which simplifies task of forensic process. SOM is used to cluster the files according to date and time of creation and their extension. So it’s an easy task for the forensic examiner for the analysis of data and organizes the file in particular manner.
As earlier M. S. Oliver [6] proposed this technique for the clustering. In the Digital Forensic Analysis clustering of textual data is performed, but again here we need to specify the number of clusters explicitly before applying clustering techniques.

For Clustering techniques J.G. Clark [7] applied in order to explore only relevant documents to the forensic examiners. But, in real time scenario forensic examiners does not know amount of data resides in computer seized devices therefore, they need some automatic approach for clustering. Email forensic analysis tool. Hadjidj [9] proposes based on document clustering technique. This helps to gather the evidences related to crime in the court of law, also need to provide automated tool for multi-staged analysis of e-mail for the forensic investigator. This also mentions the role of document clustering for forensic analysis.

Newly, Nassif and Hruschka[1] proposed an approaches of document clustering algorithm for forensic analysis in the computer seized devices also overcame the limitation of document clustering algorithm. For the purpose estimating the number of cluster automatically they use the relative validity index criteria [11] which overcomes the limitations of previous techniques. They have tested the result of five dataset, which has real word investigation cases conducted by Brazil police department. For testing they have used six well-known clustering algorithm among these partition (k-means, k-medoids) algorithms does not provide better result. Because when we assign data point randomly starting at centroid, it will change result every time they need to find probabilistic result. Hierarchical (Complete link and Average link) algorithm provide the best result, because provide the furthest distance and average distance between the pair of cluster.

IV. CONCLUSION AND FUTURE WORK

We have introduced algorithms for computer forensic analysis; the scope of clustering data is very difficult step. There is very large data to be clustered in compute forensic to overcome this problem, we presented an approach that applies document or text clustering methods to forensic analysis of computers seized in real world cases conducted by police investigations. In our work, we are trying to implement the k-means algorithms and the algorithm known as Average Link and Complete Link yielded the best results. These algorithms are suitable for our work domain because dendrograms provides a summary view of documents which are being inspected. All the large number of case or documents is collected and providing corresponding output.

Future Work:

Datasets consist of unlabeled objects, Clusters is known and fixed a priori by the user, Scalability may be an issue For the future work we can improve such thing. Also, can be extended to impose these different text clustering and document clustering techniques on real time forensic data sets and comparing the performance of these techniques to speed up forensic analysis process is high.

ACKNOWLEDGMENT

I express my sincere gratitude towards my guide Dr. Sachin N. Deshmukh for his constant help, encouragement and inspiration throughout the project work. Without his invaluable guidance, this work would never have been a successful one. I would also like to thank all the faculty members and staff of Computer Science and IT department for making this journey successful.

REFERENCES

[1] Filipe daCruz, Nassif and Eduardo Raul 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] A. K. Jain, R. C. Dupes, Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[3] L. Kaufman, P. Rousseeuw, Finding Groups in Gata: An Introduction to Cluster Analysis. Hoboken, NJ: Wiley-Interscience, 1990.

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

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

[7] 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, no. 1, pp. 49–54, 2007.

[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] 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.

[10] 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, 2005, pp. 597–601.

[11] 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.

[12] Jensen, J.H., Ellis, D., Christensen, M.G., Jensen, and S.H.: Evaluation distance measures between Gaussian mixture models of mfccs. Proc. Int. Conf. on Music Info. Retrieval ISMIR-07 Vienna, Austria pp. 107–108 (October, 2007)

[13] C. M. Bishop, Pattern Recognition and Machine Learning. New York: Springer-Verlag, 2006.