International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622
International Conference on Industrial Automation and Computing (ICIAC- 12-13th April 2014)

RESEARCH ARTICLE OPEN ACCESS

Role of Text Clustering and Document Clustering Techniques in Computer Forensic Analysis: A Review

Mr. Nitin S. Kharat *, Prof. Harmeet Khanuja **
*(Department of Computer Engineering, MMCOE, University of Pune, India
Email: ni3kharat@gmail.com)
** (Department of Computer Engineering, MMCOE, University of Pune, India
Email: harmeetkaurkhanuja@mmcoe.edu.in)

ABSTRACT
Recently, in the world of digital technologies especially in computer world, there is a tremendous increase in crimes like ethical hacking, fraud in different domains, money laundering, unauthorized access etc. So investigation of such cases deserves a much more importance. In such investigation computer seized devices plays a significant role. But computer seized devices contains much of data so here computer forensics comes into picture. Computer forensics deals with analyzing such huge set of documents to collect the evidence from computer devices. So, to do computer forensic analysis time limit is an also significant factor. So it's a challenging 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 Text Clustering and Document Clustering. This paper reviews different existing Text clustering and Document clustering methods in accordance with computer forensic analysis.

Keywords - Computer forensic, document clustering, forensic Analysis, money laundering, text clustering

I. INTRODUCTION

Basically forensic analysis is the application of broad spectrum of sciences technologies to investigate situation after fact and to establish what occurred based on the collected evidences [1]. Computer forensics deals with the preservation, identification, extraction as well as documentation of digital evidences [2].Computer forensic is analyzing huge number of files from computer seized devices. 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 seized devices is also key task of forensic analyst. Digital evidence is defined as the information and data of investigative value that are stored on, received or transmitted by digital device [3]. 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. Once the forensic examiner collects such evidences then his job is to establish what occurred based on the collected evidences. But such computer seized devices contains huge set of documents and files, so it is not easy to do the analysis of each and every files individually. At the same time court of law requires quick result of such cases, so to improve speed of such forensic analysis process has a significant wattage. So the key factor to improve such forensic analysis process requires text clustering and document clustering techniques. The text clustering and document clustering simplifies the job of forensic examiner in forensic investigation. The paper outlines the significance of text clustering and document clustering in computer forensic analysis process. The remainder of the paper is organized as section 2 explores role of text clustering and document clustering, section 3 outlines the text and document clustering techniques in forensic analysis, section 4 specifies comparative study and section 5 highlights the conclusion and future work.

II. ROLE OF TEXT CLUSTERING AND DOCUMENT CLUSTERING

2.1 Role of Text Clustering in Forensic Analysis

Generally, most of computer seized devices consist of textual data as an input to digital forensic analysis process. The textual information is the key point for forensic analysis process. These computer seized devices composed of huge amount of information which needs to be analyze by forensic examiner. But this textual information resides in unstructured format which makes difficult job for forensic examiner. That's why forensic examiner finds more difficulties during forensic analysis and to

search the required patterns for further investigations. So we need improved process for the forensic analysis of such computer seized devices which can be achieved by text clustering techniques. Text clustering consists [5] of two phases:

[1] Extraction of Textual Information

[2] Analysis of textual data using clustering algorithm.

As shown in Fig. 1, text clustering involves following steps:

1. Collection of documents*:* It includes different processes like crawling, indexing, filtering in order to collect the document.

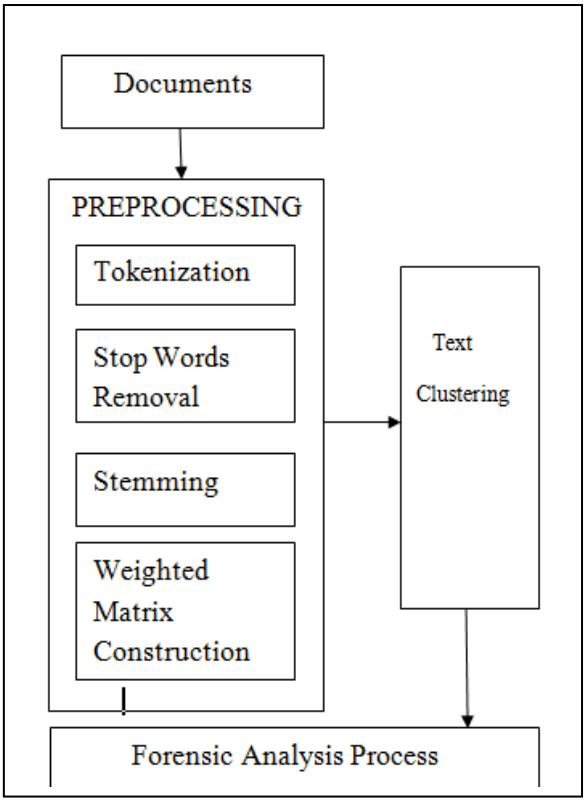

Fig. 1: Relation of Text Clustering and Forensic Analysis

2. Preprocessing:

A. Tokenization*:* It takes text as input and outputs the number of tokens.

B. Stop Words Removal: It includes removal of stop words i.e. the, and etc.

C. Stemming: Stemming is defined as reducing the words to their base form like 'extracting',' extracts',' extracted', extraction' all are converted to stem 'extract'.

D: Weighted Matrix Construction: It involves the construction of weighted matrix based on frequency of occurrence of word.

3. Text Clustering:

Based on preprocessing of data, text clustering is performed on preprocessed data. Text clustering produces sets of clusters as an output.

4. Forensic Analysis Process:

Forensic analysis process utilizes the result of text clustering for collection of relevant files and documents according to reported case.

2.2 Role of Document Clustering in Forensic Analysis

Computer forensic analysis involves the examining the huge set of files. Among all of that files are not relevant to the forensic examiner interest. So analyzing such files and documents which are out of interest tends to more time consuming task. So the key approach is to apply document clustering [7] 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 as shown in Fig. 2:

1. **Collection of Data:**

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

2. **Preprocessing:**

As discussed earlier preprocessing involves the tokenization, removing of stop words, stemming process and weighted matrix construction phases.

3. **Document Clustering:**

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

4. **Post Preprocessing:**

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

5. **Forensic Analysis:**

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.

3. Text Clustering and Document Clustering Techniques

3.1 Text Clustering Based Techniques for Forensic Analysis

The author [5] of 'Text clustering for digital forensic analysis' proposes effective digital text analysis strategy based on clustering based text

mining techniques. He uses two step procedures. In first phase he performs the textual information extraction and second step involves textual data analysis via clustering methods. This text clustering technique helps the forensic examiner at greater extent because forensic examiner need not to analyze each and every textual information rather than he only analyze the interested information that found in resultant cluster. So this approach helps to enhance the performance of forensic analysis process. The resultant data clusters [5] are as shown in following Fig. 3. The forensic examiner only analyzes the clusters which are relevant to the admitted case, hence this speed up the forensic analysis process.

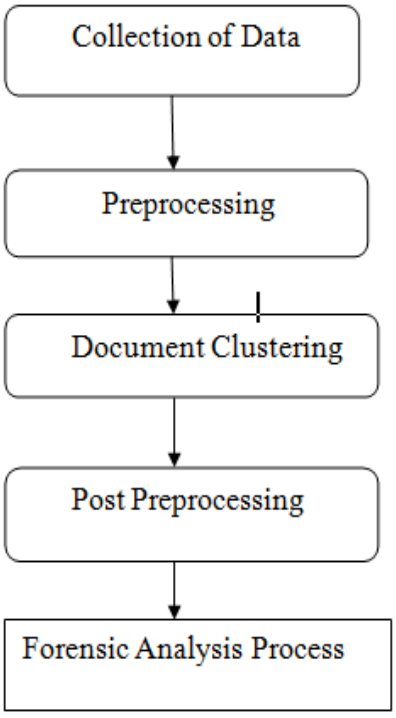

Fig. 2: Relation of Document Clustering and Forensic Analysis

Cluster	Most Frequent and Relevant Words
1	employee, business, hotel, Houston, company
2	pipeline, social, database, report, link, data
3	ECT, EnronXg
4	coal, oil, gas, nuke, west, test, happy, business
5	Yahoo, compubank, NGCorp, Dynegi, night, plan
6	shank, trade
7	travel, hotel, continent, airport, flight, Sheraton
8	Questar, Paso, price, gas
9	schedule, London, server, sun, contact, report
10	trip, weekend, plan, ski

Fig. 3: Text Clustering For Forensic Analysis

In [6], they propose Topic Modeling in Financial Documents which introduces method for clustering of financial documents.

Here the document is assigned to some number of topics and specifies segment of it which addresses a given topic. It helps the forensic examiner at greater extent because it provides summarized as well as detailed information of financial documents so that forensic examiner will be able to explore the details of financial documents which are of the interest and relevant to the reported case rather than analyzing whole document set .The result[6] is as shown in Table 1. It also plays a vital role in forensically analyzing financial organization reports.

TABLE I
TOPIC MODELING RESULTS

Oil and Natural Gas	Biotech	Real Estate	Media and Networking	Misc
gas	patient	occupants	network	gas
drill	trial	tenant	brand	scrap
rig	clinic	hotel	software	mario
barrel	fda	revpar	wireless	russo
acreage	dose	music	tv	glenrock
haynesville	cancer	noi	video	gasoline

The J.G.Clark [8] also proposed post-retrieval clustering of digital forensic text string search results to improve the information retrieval effectiveness in digital forensic text string searching. It helps the forensic examiner because it provides the result only relevant to the forensic examiner investigative

objectives. This approach proves the effective and significant role in text clustering for the forensic analysis process.

3.2 Document Clustering Based Techniques for Forensic Analysis

S.Oliver [9] proposes Self Organizing Maps (SOM) to support decision making by forensic investigators. Self Organizing Maps (SOM) is basically used to search the pattern in data set. The files are clustered according to date and time of creation and type of files. So it's an easy task for the forensic investigator for the analysis once he got specified pattern. R. Hadjidj [10] proposes email forensic analysis tool based on document clustering technique. It provides automated tool for multi-staged analysis of e-mail for the forensic investigator which helps to gather the evidences related to crime in the court of law. This also notifies the document clustering role in the forensic investigation.

Recently, Nassif and Hruschka[11] proposed an approach that applies document clustering algorithms for the forensic analysis of computer devices. They uses the relative validity index criteria for the estimating the number of clusters in an automated manner which overcomes the limitations previous techniques. Here the forensic examiner can analyze only relevant clusters documents in accordance with reported case. The results [11] are as shown in Fig. 4.

Cluster	Information
C_1	3 blank documents
C_2	4 financial transactions
C_3	2 maternity payments
C_4	2 grocery lists
C_5	1 foreign exchange transaction warning 1 list of documents for registration information
C_6	2 documents from foreign exchange operations
C_7	1 registration form from a brokerage company 1 contract template from the broker
C_8	1 investment club status 1 agreement for joining the club
C_9	2 models for handling cash greater than R$ 100k
C_{10}	8 receipts of foreign exchange insurance transactions
C_{11}	2 warnings about foreign brokerage business hours
C_{12}	3 label designs of a brokerage company
C_{13}	1 notice about working hours 1 check receipt
C_{14}	2 daily reports from buying/selling exchanges
C_{15}	2 sample documents from office application

Fig. 4: Document Clustering for Forensic Analysis

IV. COMPARATIVE STUDY OF FORENSIC ANALYSIS TECHNIQUES

As discussed earlier S. Oliver [9] proposed 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. Whereas in 'Text Clustering for Digital Forensic Analysis'[5], clustering of textual data is performed, but again here we need to specify the number of clusters explicitly before applying clustering techniques. The J.G. Clark [8] 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. Nassif and Hruschka[11] 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.

V. CONCLUSION AND FUTURE WORK

As the computer forensic have a key point in digital investigation, this paper reviews existing text clustering and document clustering techniques. The paper also explores role of text clustering and document clustering in forensic analysis process. Additionally, it also explores how the existing methods for the text clustering and document clustering are effective and efficient for the forensic analysis process. It also notifies that these techniques improves the performance of computer forensic process within quick period of time as required by court of law as well as speed up forensic analysis process.

The future work 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. It also helps to researchers to employ such techniques in automated forensic tool.

REFERENCES

[1] ForensicsAvailable:https://sites.google.com/site/wnoyolasample/
[2] Forensic Available :http://en.wikipedia.org/wiki/ Computer_forensics
[3] J. Clerk Maxwell, A Treatise on Electricity and Magnetism(3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73).
[4] U.S. Department of Justice, Electronic Crime Scene Investigation: A Guide for First Responders (I Edition,NCJ 219941, 2008, http://www.ncjrs.gov/pdffiles1/nij/219941.pdf)
[5] S. Decherchi, S. Tacconi, J. Redi, A. Leoncini, F. Sangiacomo, and R. Zunino, "Text clustering for digital forensics analysis,"

Computat. Intell. Security Inf. Syst., vol. 63, pp. 29–36, 2009.

[6] Patrick Grafe., "Topic Modeling in Financial Documents", Department of Computer Science Stanford University.

[7] A. K. Jain and R. C. Dubes, Algorithms for Clustering Data (Englewood Cliffs, NJ: Prentice-Hall, 1988.)

[8] 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.

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

[10] 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.

[11] Luis Filipe da Cruz Nassif and Eduarado Raul Hruschka "Document Clustering for Forensic Analysis :An Approach for Improving Computer Inspection"- IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY,VOL.,NO. 1,JANUARY 2013.

[12] B. D. Carrier, E. H. Spafford, An event-based digital forensic investigation framework, in: Proceedings of the 4th Digital Forensic Research Workshop, 2004.

[13] E. Casey, Digital Evidence and Computer Crime: Forensic Science, Computers, and the Internet with Cdrom, 1st ed., Academic Press, Inc., Orlando, FL, USA, 2000.

[14] M. R. Clint, M. Reith, C. Carr, G. Gunsch, an Examination of Digital Forensic Models (2002).

[15] H. Lee, T. Palmbach, M. Miller, Henry Lee's Crime Scene Handbook (San Diego: Academic Press, 2001.)

[16] G. Palmer, M. Corporation, A Road Map for Digital Forensic Research, in: Proceedings of the 1st Digital Forensic Research Workshop, 2001.

[17] L. Garfinkel, Digital forensics research: The next 10 years, Digital Investigation 7 (1) (2010) S64 – S73.

[18] K. M. Hammouda and M. S. Kamel. Efficient phrase-based document indexing for web document clustering. IEEE Transactions on knowledge and data engineering, 16(10):1279{1296, 2004}.

[19] A. Miller. Wordnet: a lexical database for English. Common. ACM, 38(11):39{41, 1995}.

[20] Alessandro Moschitti and Roberto Basili. Complex linguistic features for text classification: A comprehensive study. In ECIR '04: 27th European conference on IR research, pages 181{196, Sunderland, UK, April 2004.

[21] Prof. K. Raja, C. Prakash Narayanan, "Clustering Technique with Feature Selection for Text Documents", Proceedings of the Int.Conf. On Information Science and Applications ICISA 2010 6 February 2010, Chennai, India.

[22] Luiz G. P. Almeida, Ana T. R. Vasconcelos and Marco A. G. Maia," A Simple and Fast Term Selection Procedure for Text Clustering "Seventh International Conference on Intelligent Systems Design and Applications, 0-7695-2976-3/07 © 2007 IEEE, doi:10.1109/ISDA.2007.15

[23] Harmeet Kaur Khanuja and Dr. D. S. Adane. 2012. A Framework For Database Forensic Analysis. Published in Computer Science & *Engineering: An International Journal (CSEIJ), Vol.2, No.3.*

[24] Harmeet Kaur Khanuja and Dr D.S. Adane.'Forensic Analysis of Databases by Combining Multiple Evidences' *International Journal of Computer and Technology , Vol 7,No 3.*