International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

A Survey of Forensic Analysis on Document Clustering

Nikhil Nagnath Boriwale[1,*], Satish .R.Todmal[1]

[1] Department of Computer Engineering, JSPM ICOER, Pune, Maharashtra, India

[2] Department of Information Technology, JSPM ICOER, Pune, Maharashtra, India

Abstract: ***Substantial development in advanced world increases quickly. The method of analysing different criminal acts utilizing machine framework is known as Digital Forensic Analysis (DFA). In forensic analysis, large number of reports is overviewed. Information comprise of in unstructured format after processing we get structured framework. In the system, there are such a variety of records which will be utilized for research specific wrong doing by investigation officer. Document clustering is useful to grouped similar records together. There are numerous techniques and tools which are accessible for investigation to mechanize this sort of procedure. Clustering algorithm is currently generally utilized within forensic investigation. In this paper we give review on literature of forensic analysis. This paper gives organized perspective of different clustering methodologies, for example, K-means, K-medoids, Single Link, Complete Link and Average Link.***

Keywords: Document clustering, forensic analysis, survey, structured, unstructured, investigation.

1. Introduction

The large growth of the digital universe means individual can faces the problem of handling large amount of data [1].A very large amount of crime increases to Internet andcomputers has caused rising need for computer forensics. Computer forensics searches evidencewhen computers are used by the police investigations of crimes. It is possible by using the document clustering.Daily, thousands of files can be investigated per computer. This activityinvolves the analysis and understanding of such a large data by expert [2][3][4].

The analysis team uses separate database system to perform on a data and also analyse that data. The basic aim is to search and analyse patterns of fraudulent activity. Generally, computerforensic tools have been used for computer forensic analysis that is available in the form of computer software.So, to make computer investigation easy the tools have been developed to help computer forensicinvestigators. However, due to incredible large scale data investigators mayhave facing difficulty to locateuseful points from hugedata[1][2].Data increases day by day so, the process of analyzing largevolumes of data may consume a lot of time. Many a times itmay happen that data generated by computer forensic tools may be meaningless. It is just because huge data that can bestored on a storage medium and the fact that existing computerforensic tools are not capable of to present a visual overview of all theobjects found on the storage medium [1].

From the Fig.1, it shows forensic analysis to collect evidence. By looking into the repairing systems, tracking illegal use of computers, spam etc. to tackle such problem solution is to authentic accurate thing so that such things could not happen.

Figure 1: Forensic Analysis [5]

Digital evidence has to be Authentic and reliability. Authenticate is most explicit link data to physical person. Must be self-sustained, strong access controls in place, logs and audit in good shape. Accurate means data process reliability determines content reliability and timings issues [5].The remainder of this paper is organized as follows. Section 2 presents literature survey. Section 3 concludes the paper. Section 4 gives references.

2. Literature Survey

Clustering algorithm helps to identify the accurate data from the analysis without less knowledge or no prior knowledge of data. Initially, computer forensics hasobjects which are unlabelled [4]. Previous analysis defines data partition from the data and expert examiner only focus on reviewing representative documents from the obtained set of cluster. Firstly, examiner check for investigation, after finding relevant document then the examiner can pass the analysis of the other document for investigation. Text clustering in digital evidence is defined as the data of the investigate value.Digital investigation is much necessary for textual evidence[6]. Examples of investigations are like e-mails, internet browsing history, instant messaging, and word processing documents.Forensic intelligence is defined as the accurate, timely and meaningful products of logically processed forensic data. The further results of forensic

intelligence have discipline specific process.The clustering has been used by computer forensics field very rarely[1][6].

2.1 Digital Forensic Investigation (DFI)

The present digital forensic tools are used for analyzing manydocuments, which provided the multiple levels of searching techniques toanswer the questions and generate digital evidence related tothe specific investigation. However, these technique works improperly whichallows the investigator to search for specific documents of certain subject of specified searchand grouped thedocument set based on a given subject.There have been many different things to define a digitalforensic model that inherits the forensic process from anyparticular technology, such as the Digital Forensics ResearchWorkshop (DFRWS) model for digital forensic analysis [8], Lee's model of scientific crime scene investigation [7],Casey's model for processing and examining digitalevidence [9], and Reith's model for digital forensic analysis[10]. DFRWS is a pioneer of the forensicprocess.

There are some problems which are occurred during forensic analysis which are mentioned bellow [11].

1) High-level Search: To identify evidence, either tool is used to search documents stored on suspect's computer or by the operating system. Since, manual searching takes much of time investigator do in above mentioned way. The automatically searching techniques which are provided bycurrent DFI tools includes, keyword search, regularexpression search, approximate matching search, andlast modification date search. Accordingly, such techniquesare applied directly on all of the stored fileswithout any high level knowledge about the topicmentioned in each document. So, the results ofthe search techniques generally suffer from the largeamount of false positives and false negatives results [11].
2) Evidence-oriented Design: At present, DFI tools aredesigned for solving problem against crime.The evidence is stored in file format on computer which is not available for addressing particular offenses. Generally,DFI techniques are designed to find evidences to reduce crime scenes [12].
3) Limited Level of Integration: Moreover, existing forensictools are designed as stand-alone applicationsand which provided limited capability for integrating witheach other or other custom tools[11].

Following fig. 2 shows process of digital forensic investigation [11]. The first phase is Identification phase in which data is associated with particular event. Next is Preservation in this step, precautions has to be taken that data could not be lost from any malicious activities. In the collection step digital data is stored. In Examination step, investigator uses certain tool to examine in depth search for particular case. Next is Analysis step, investigator collect all the evidence together and analyse it with the suspect computer. After that, investigator summarizes the entire thesis, and with that event he presents it which he found out during investigation [11].

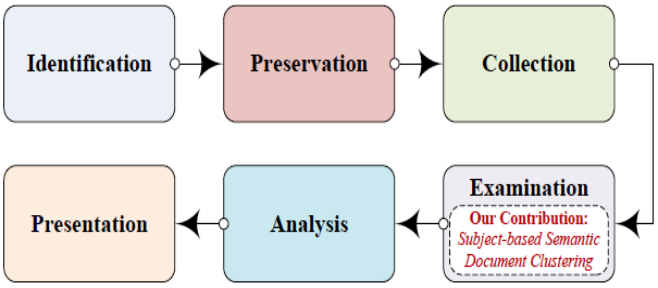

Figure 2: Process of Digital Forensic Investigation (DFI)[11]

2.2 Text mining Methodology

Text mining involves certain methods from different areas like information retrieval, natural language processing, information extraction and data mining. The use of text mining method is to get useful and structured data from large amount of meaningless data.

2.3 Phases of Document Clustering Algorithm:

Following are the phases which are used in document clustering technique[11].

2.3.1 Document Clustering

Clustering meansthe grouping of similar data items together. Clustering methods involves the process of grouping retrieved documents into thesame list of meaningful categories. Document clustering involves descriptors and its extraction process. Descriptors are nothing but the sets of words which describes the contents within the cluster. Document cluster is normally considered as a centralized process. Document clustering can be classify as online and offline.

2.3.2 Pre-processing

Pre-processing is necessary for document clustering. It should be done before clustering. Stemming reduces inflected word to their root word. Stemming algorithm is used to delete suffixes. Stop words are getting removed. Technique searches text by a predefined list so called stop words and delete them from text, stop words like prepositions, pronouns, articles and irrelevant document, Meta data. After that, TDM(Term Document Matrix) is calculated. A distance matrix is used for clustering and multidimensional scaling contains cosine coefficients computed on tf-idf of various words within the documents.

2.3.3 Clustering Algorithm

There are various algorithms which canused in clustering likeK-means, K-medoids, Single Link, Complete Link and Average Link.K-means is one of the simplest unsupervised algorithms.In single link clustering; two groups have been merged such that their closest pair of documents have the highest similarity compared toany other pair of groups. In complete linkage many elements in the clusters are distant to each other. It produces more Compact clusters and most useful hierarchies than any other clustering. In average linkage clustering pairing of clusters takes place with the highest cohesion. The main key point of K-medoids is to determine optimal value from the original set of values.

3. Conclusion

We have surveyed various forensic systems.For the forensic analysis we have introduced the different aspects of text mining and document clustering. The main key point of digital forensic analysis is the document clustering and text analysis. In this paper we have considered the review ofprocess of digital forensic analysis with different phase'salso different phases of document clustering. This survey study is exhibited by considering our future examination work over the utilization of document clustering for digital forensic analysis. In this paper we have taken the survey of methodology of digital forensic analysis with phases included into it. For future work, first we like to recommend dealing with distinctive clustering algorithm for clustering in computerized criminological result with useful examination results. An alternate future bearing to this examination field is to examine Expectation Maximization (EM) algorithm. Future research could request on the possibility of enriching in the cluster technique semantic data. An alternate future direction could comprise in utilizing supervised learning tools to classify information on as of now defined category for investigative purposes.

References

[1] J. F. Gantz, D. Reinsel, C. Chute, W. Schlichting, J. McArthur, S.Minton, I. Xheneti, A. Toncheva, and A. Manfrediz, "*The expandingdigital universe: A forecast of worldwide information growth through2010*," *Inf. Data*, vol. 1, pp. 1–21, 2007.

[2] Luis Filipe da Cruz Nassif and Eduardo Raul Hruschka," *Document Clustering for forensic Analysis: An Approach for Improving Computer Inspection*", IEEE transaction on Information Security, Vol. 8,pp. 46-54,January 2013.

[3] R. Hadjidj, M. Debbabi, H. Lounis, F. Iqbal, A. Szporer,and D. Benredjem, "*Towards an integrated e-mail forensicanalysis framework*," Digital Investigation, Elsevier, vol. 5, no. 3–4, pp. 124–137, 2009.

[4] 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.

[5] Vidhya B, PriyaVaijayanthi R, "*Enhancing Digital Forensic Analysis through Document Clustering*," in International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3, Issue 1, January 2014.

[6] Sergio Decherchi,SimoneTacconi,and Judith Redi,FabioSangiacomo,AlessioLeoncini,and Rodolfo Zunino," *Text Clustering for Digital Forensics Analysis*", Journal of Information Assurance and security", Vol. 5,pp. 384-391,January 2010.

[7] H. Lee, T. Palmbach, M. Miller, Henry Lee's Crime SceneHandbook, San Diego: Academic Press, 2001.

[8] G. Palmer, M. Corporation, *A Road Map for Digital ForensicResearch*, in: Proceedings of the 1st Digital Forensic ResearchWorkshop, 2001.

[9] E. I. Casey, *Digital Evidence and Computer Crime:* ForensicScience, Computers, and the Internet with Cdrom, 1st ed., AcademicPress, Inc., Orlando, FL, USA, 2000.

[10] M. R. Clint, M. Reith, C. Carr, G. Gunsch, An Examination ofDigital Forensic Models (2003).

[11] Bhagyashreeumale, prof. Nilav M, "*Survey on Document Clustering Approach for Forensics Analysis*", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 3335-3338.

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