ISSN:0975-9646

Survey on Document Clustering Approach for Forensics Analysis

Bhagyashree Umale
Research Scholar,
Dr. D. Y. Patil School of Engineering and Technology,
Lohgaon, Pune

Prof.Nilav M
Assistant Professor,
Dr. D. Y. Patil School of Engineering and Technology,
Lohgaon, Pune

Abstract: In recent time it is common that use of digital devices such as computers are used for the analysis of crimes by investigation officers. The crimes include hacking, drug trafficking, pornography, various theft crimes. The method of analyzing the various crimes using the computer based methods is called as digital forensic analysis (DFA). The main input to this system is number of raw and unstructured input text files. In computer there are multiple files are present in order to process them for the investigation of particular crime by investigation officer. Therefore, to automate this process, there are many methods and tools presented for forensic investigations. The key part of these tools or methods is the use of clustering algorithms in which the number of unstructured text files is given as input and the output is generated in structured format. Recently the document clustering methods are used for digital forensic analysis. The basic goal of this paper is to present the various text analysis methods using clustering algorithms.

Keywords: ***Digital forensic analysis, crimes, document clustering, structured data, text analysis.***

I. INTRODUCTION

The method of Digital Forensic Investigation (DFI) is the process in which the digital devices like computers are used to analyze the digital evidence those are facts which are under the investigation. The digital evidence is the digital data which supports or refutes the incident hypothesis [2]. Documents analysis process in computer device is key task of the digital forensic investigation process. But, this process of document analysis is becomes more complex if the number of documents available to process are more in number. Further this process becomes more complex, if the size of particular storage device increases. There are some methods and tools already presented by various researchers for the analysis of multiple documents. These existing methods of DFI present the multiple level searching approach for giving the accurate results and producing digital evidence that is related to the current investigation task. But the limitation of such methods is that they stop allowing the end user means the crime investigator for searching the documents which are belonging to a specific subject in which end user interested, or to group the document set based on a given subject.

Recently the clustering algorithms are used in the process of digital forensic analysis. These methods are basically used to convert unstructured documents to structured documents for further investigation. This is precisely the case in many applications of Computer Forensics From a more technical viewpoint, our datasets consist of unlabeled objects—the classes or categories of documents that can be found are a priori unknown. Moreover, even assuming that labeled datasets could be available from previous analyses, there is almost no hope that the same classes (possibly learned earlier by a classifier in a supervised learning setting) would be still valid for the incoming data, obtained from other computers and associated to different investigation processes. More precisely, the new data sample come from a different population. In this context, the use of clustering algorithms, which are capable of finding latent format from text documents found in seized computers, can enhance the analysis performed by the expert examiner.

This is our first review paper on document clustering approaches those are widely used for digital forensic analysis. The text clustering is having two steps like extraction of textual information and analysis of text data using the clustering algorithms. During this paper, following different sections presented. In section II, we are discussing about Digital Forensic Investigation (DFI) process and its associated problems. In section III, we are discussing the process of text analysis in detail, in this section we are also discussing the problems associated with text mining process. In section IV we will discuss the document/text clustering algorithms, finally in section V conclusion is made.

II. REVIEW OF DIGITAL FORENSIC INVESTIGATION

The process of investigating digital devices for the purpose of generated digital evidence related to an incident under investigation is common reference to as Digital Forensic Investigation (DFI). [1], Digital evidence is a digital data can be supports or refutes a hypothesis about the incident. The task of analyzing persistent documents found on a storage device of a suspect's computer is an essential part of the DFI process to gather credible and convincing evidence. However, this task may be daunting due to the large number of documents usually stored on a hard disk. The continuous increase size of storage devices makes the task even more difficult.

Existing digital forensic tools for analyzing a set of documents provided multiple levels of search techniques to answer questions and generate digital evidence related to the investigation. However, these techniques stop short of allowing the investigator to search for documents that due to a certain subject he is interested in and a group the document set based on a given subject.

There have been several attempts to define a digital forensic model that abstracts the forensic process from any specific technology, such as the Digital Forensics Research Workshop (DFRWS) model for digital forensic analysis [5], Lee's model of scientific crime scene investigation [4], Casey's model for processing and examining digital evidence [2], and Reith's model for digital forensic analysis [3]. DFRWS is a pioneer that developed the forensic process. It defined Digital Forensic Science as a linear process:

The use of scientifically derived and proven methods toward the preservation, collection, validation, identification, analysis, interpretation, documentation and presentation of digital evidence derived from digital sources for the purpose of facilitating or furthering the reconstruction of events found to be criminal and helping to anticipate unauthorized actions shows to be disruptive to planned operations. [5]

Figure 1 illustrates the Digital Forensic Investigation (DFI) process as defined by DFRWS. After determining items, components, and data associated with the incident (Identification phase), the next level step is to preserve the crime scene by stopping or preventing any activities that can damage digital information being collected (Preservation phase). Follows that, the next level step is collecting digital information that might be related to the incident, such as copying files or recording network traffic (Collection phase). Next step, the investigator conducts an in-depth systematic search of evidence related to the incident being investigated, such as filtering, validation and pattern matching techniques (Examination phase) [3]. 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 investigator summarizes and the findings by explaining the reasons for each hypothesis 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-depth systematic search for pertinent evidence.

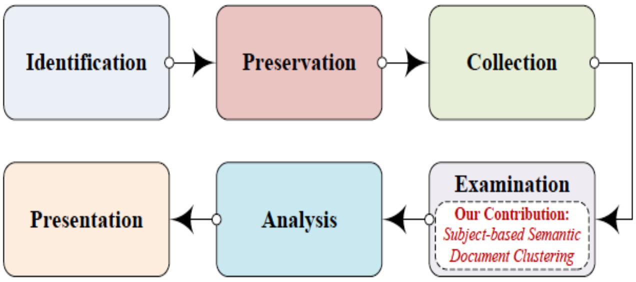

Figure 1: Process of Digital Forensic Investigation (DFI)

However, there are three problems with today's computer forensic tools:

2.1 High-level Search: Since manual browsing is time consuming, investigators often rely on the automatically searching capability provided by either the operating system or existing DFI tools to conduct a search on the documents stored on the suspect's computer in order to identify related evidence. The main automatically searching techniques provided by current DFI tools include keyword search, regular expression search, approximate matching search, and last modification date search. Luckily, such techniques are applied directly against all of the stored documents without any advance knowledge about the topics discussed in each document. Hence, the results on these search techniques generally suffer from a large number of false positives and false negatives.

2.2 Evidence-oriented Design: Existing DFI tools are designed for solving crimes committed against people, in which the existence evidence exists on a computer; they were not created to address cases where crimes took place on computers or against computers. In general, DFI techniques are designed to find evidence where the possession of evidence is the crime itself; it is easy to solve child pornography cases than computer hacking cases. [6]

2.3 Limited Level of Integration: Most existing forensic tools are designed to work as stand-alone applications and provided limited capability for integration with each other or another custom tools or resources the digital forensic team might have already developed.

III. Text Mining Methodology: Review

Text mining involves the application of techniques from areas such as information retrieval, natural language processing, information extraction and data mining. These are various stages of a text mining process can be combined together into a single workflow.

3.1. Information Retrieval (IR):

Systems identify the documents in a collection which match a user's query. The most well known IR systems are search engines such as Google, which identify this document on the World Wide Web that are relevant to a set of given words. IR systems are obtained used in libraries, where the documents are typically not the books themselves but digital records containing information about the books. IR systems allow us to narrow down the set of documents that are relevant to a particular problem. As text mining involves very computationally-intensive algorithms to large document collections, IR can speed up the analysis considerably by reducing the number of documents for analysis. For example, if we are interested in mining information only about protein 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 its synonyms.

3.2. Natural Language Processing (NLP)

This is one of the oldest and most difficult problems in the field of artificial intelligence. It is process for analysis of human language so that computers can understand natural languages as humans do. Although this goal is still away off, NLP can perform some types of analysis with a high degree of success. Shallow parsers identify only the main grammatical elements in a sentence, such as noun phrases 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 provide the systems in the information extraction phase (see below) with linguistic data that they need to perform their task. This is done by annotating documents with information like sentence boundaries, part-of-speech tags, parsing results, which can then be read by the information extraction tools.

3.3. Data Mining (DM): This is the process of identifying patterns in large sets of data. The aim is to uncover past unknown, useful knowledge. DM is applied to the facts generated by the information extraction phase. We have put the results of our DM process into another database that can be queries by the end-user via a graphical interface. The data to be generated by the queries can also be represented visually.

3.4. Information Extraction (IE): This is the process of automatically obtaining structured data from an unstructured natural language document. This involves defining the general form of the information that we are interested in as one or more than templates, which are then used to guide the extraction process.

3.5 Text Mining Problems: One main reason for applying data mining methods to text document collections is to structure them.

A data can be significantly simple collection of document which is accessible for a user and their structure are like library catalogues or book indexes. This is the problem of manual designed indexes that it required time to maintain them. Therefore, they are not obtaining up-to-date or thus not usable for recent publications or frequently changing information sources like the World Wide Web. The existing methods for structuring collections either try to assign keywords to documents based on a given keyword set (classification and categorization methods) and automatically structure document collections to find groups of similar documents (clustering methods). The problem of Text Mining is Classification of data set and Discovery of Associations among data.

3.6 Text Mining Tasks

- Text categorization - assigning the documents with pre-defined categories (e.g. decision trees induction).
- Text clustering - descriptive activity, which groups similar documents together (e.g. self organizing maps).
- Concept mining - modeling and discovered of concepts, sometime combines categorization and clustering approaches with concept logical base ideas in order to find concepts and their relations from text collections (e.g. formal concept analysis approach for building of concept hierarchy).
- Information retrieval - retrieving the documents relevant to the user's query.
- Information extraction - question answering.

IV. Phase Of Document Clustering Algorithms

Following figure 2 and 3 showing the phase of clustering process and phase of document clustering in details respectively. Based on this figures we are describing details in below points:

4.1 Collection of Data: includes the processes like crawling, indexing, filtering etc which are used to collect the documents that need to be clustered, index them to store and retrieve in a better way, and filter them to remove the extra data, for example, stop words.

4.2 Preprocessing: It consists of steps that take as input a plain text document and output a set of tokens (which can be single terms or n-grams) to be included in the vector model.

4.3Filtering: is the process of removing special characters and punctuation that are not thought to hold any discriminative power under the vector model. This is more critical in the case of formatted documents, such as web pages, where as formatted tags can either be discarded or identified and their constituent terms attributed different weights [7].

4.4 Tokenization: splits sentences into separates tokens, typical words. Mostly sophisticate methods, drawn from the field of NLP, parse the grammatical structure of the text to pick significant terms or chunks, such as noun phrases [8].

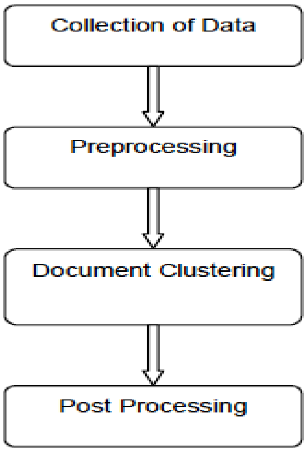

Figure 2: Phase of clustering

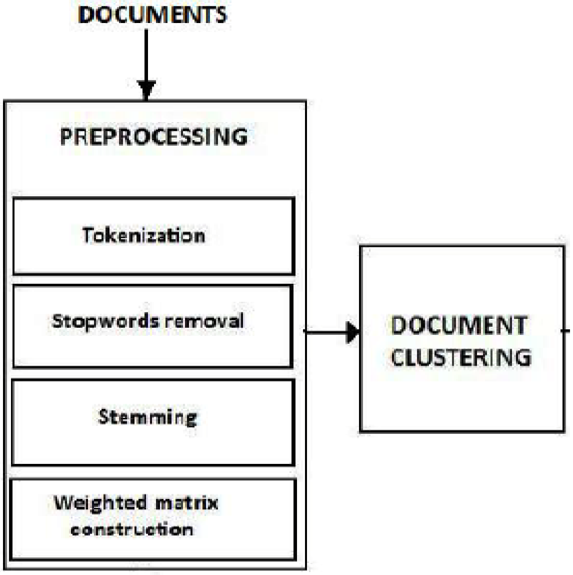

Figure 3: Phase of Document/Text clustering

4.5 Stemming: The process of reducing words into their base form and stem. For example, the words “connected", “connection", “connections" are all reduced to the stem “connect." Porter's algorithm [9] is the de facto standard stemming algorithm.

4.6 Stopword Removal: A stopword is defined as a term, which is not thought to convey any meaning as a dimension in the vector space (i.e. not context). A typical method can remove stopwords by compare each term with a compilation of known stopwords. Another approach is to first apply a part of speech tagger and then rejected all tokens that are not nouns, verbs, or adjectives.

4.7 Pruning: removes words that appear with very low frequency throughout the corpus. The underlined assumption in those words, even if they had any discriminating power, would form too small clusters to be useful. A pre-defined threshold is typically used, e.g. A little or small fraction of the number of words in the corpus. Sometimes words which occur very frequently (e.g. in 45% or more than of the documents) are also removed.

4.8. Post processing: includes the major applications in which the document clustering is used, for example, the application that shows the results of clustering for recommending news articles to the users.

The problem of document clustering is generally defined as follows [10] given a set of document cluster them. An automated derived number of clusters, such documents assigned to each cluster and are more similar to each other than the documents assigned to different clusters. Documents are represented using the vector space model that treats a document as a bag of words [11].

V. Conclusion

During this paper we have introduced the different aspects of text mining and document clustering which are used for the analysis of forensic. The document clustering and text analysis is the key for digital forensic analysis. This review study is presented by considering our future research work over the use of document clustering algorithms for digital forensic analysis. In this paper we have taken the review of process of digital forensic analysis with different phases involved into it. In addition to this, the problems associated with such process are listed in this paper. For the future work, first we like to suggest working on different clustering algorithms for document clustering in digital forensic analysis with practical investigation results. Another future direction to this research field is to investigate Expectation Maximization (EM) algorithm.

References

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

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

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

[4] H. Lee, T. Palmbach, M. Miller, Henry Lee’s Crime Scene Handbook, San Diego: Academic Press, 2001.

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

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

[7] K. M. Hammouda and M. S. Kamel. Efficient phrase based document index for web document clustering. IE3 Transactions on knowledge and data engineering, 16(10):1279{1296, 2004}.

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

[9] 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.

[10] 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.

[11] 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