International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 12, December - 2013

Review on Email Forensics Using Sequence Mining

Priyanka V. Kayarkar
NIRT, RGPV, Bhopal

Prof. Prashant Ricchariya
NIRT, RGPV, Bhopal

Prof. Anand Motwani
NIRT, RGPV, Bhopal

Abstract

Digital forensics is the use of scientifically derived and proven methods to preserve, validate, identify, analyse, interpret and present the digital evidence stored in digital devices.

With the rapid and advanced growth in internet gives rise to advanced forms of digital crime. During criminal activities crime committed use digital devices, forensic examiners have to adopt practical frameworks and methods to recover data for analysis which can comprise as evidence. Investigation of Digital forensics adopts three essential processes: Data Generation, Data Preparation and Data warehousing. Data Mining has unlimited potential in the field of Digital Forensics. Computer forensics is an emerging discipline investigating the computer crime.

In this paper we are introducing the cyber Forensics using Sequence Mining algorithm, by comparing it with association rule mining algorithm parameters.

1. Introduction

Digital forensics is a growing and important field of research for current intelligence, law enforcement, and military organizations today. The goal of digital forensics to find out the digital evidence for the forensics investigation.

1.1. Digital Evidence

It can be defined as the clues which can be recovered from digital sources and helps in digital forensics investigations. Evidences are very delicate to deal with it, if it is handled improperly it can be spoiled. The evidence is accurate and reliable if the substance of the story the material tells is believed and is consistent, and there are no reasons for doubt. Digital evidence can be classified, compared, and individualized in several ways. One of those ways is by the contents of the evidence. For example, investigators use the contents of an e-mail message to classify it and to determine which computer it came from. Digital Forensics includes various sub-branches relative to investigation of various types of devices, media and artefacts. It is an important part of computer investigation to recover data.

Digital Forensics can be categorised into four areas:

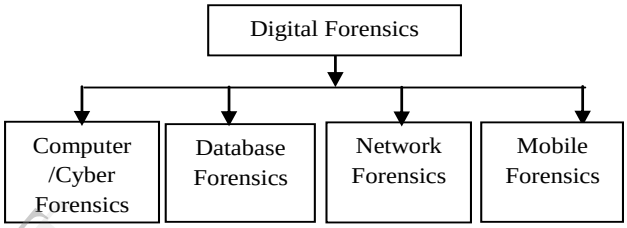

Fig.1: Types of Digital Forensics

Table 1: Information of types of Digital Forensics

Computer /Cyber Forensics	• It deals with broad range of digital information from system logs such as browser history with the help of actual files stored on the drive.
Database Forensics	• It studies databases and their metadata. • It uses database contents, log files in order to retrieve the relevant information.
Network Forensics	• It observes computer network traffic to gather information for the purpose of legal evidence. • It allows us to make forensic determinations based on the observed traffic of the network ((both LAN and MAN/internet).
Mobile Forensics	• Recovers data from mobile devices. • Here investigation usually focuses on simple data such as call details and SMS or Emails Mobile devices are also gives the information about the location.

1.2. Cyber Forensics

Cyber forensics is "The application of computer investigation and analysis techniques in interest of determining potential legal evidence."

1.3. Cyber Crime

It refers to the criminal activity where computer is target for conducting crime. Computer criminals can infiltrate wide variety of platforms and commit wide array of crimes. Computers are everywhere and have virtually penetrated all industries. Investigators uses computer evidences in variety of ways where incriminating documents or files can be found.

1.4. Emerging Cyber Frauds

There are many types of cyber frauds:

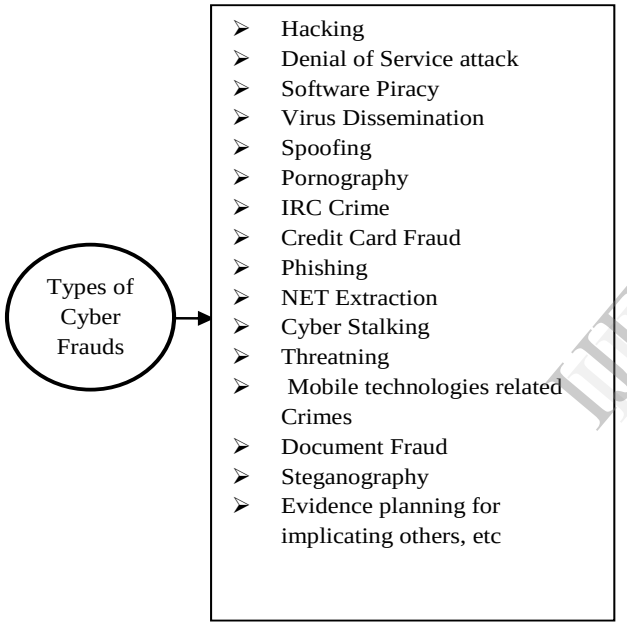

Fig 2: Types of Cyber Frauds

1.5. Computer Forensics Investigation Process

Computer forensics involves the preservation, identification, extraction, interpretation, and documentation of computer evidence. At a very basic level, computer forensics is the analysis of information contained within and created with computer systems, typically in the interest of figuring out what happened, when it happened, how it happened, and who was involved.

The investigation process is as follows.

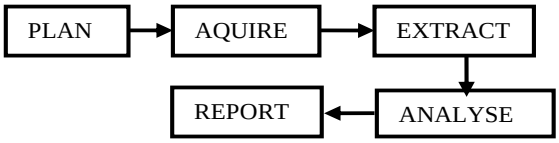

Fig 3: Cyber Forensics Investigation Process Plan

Computer Forensics Process begins with the plan. The ability to build and fallow the targeted work flow guidelines helps to reduce time and thereby cost and increases the amount of relevant data retrieved.

1.5.1. Acquire. This process is similar to taking photographs, blood samples and fingerprints from a crime scene. Allocated and unallocated area of the hard disk are copied that referred as image of an investigation. Forensic tools used in this phase to copy all information from the suspect storage device to a trusted device. These tools modify the suspect digital device as little as possible and copy all data from digital device. The aguisition process ranges from complete forensic disk imaging to gathering information from other devices and sources (like servers and phones).

1.5.2. Extract. Once data has been collected, the next phase is to extract the relevant pieces of information from the collected data.

1.5.3. Analysis. Extracted and relevant data has been analyzed to draw conclusions. If additional data is sought for detail investigation will call for in depth data collection.

1.5.4. Report. Once computer forensic analysis is complete, presenting an understandable, defendable and complete report is key.

In computer forensics, there are three types of data:

- Active Data is the information that we can actually see. This includes data files, programs, and files used by the operating system. This is the easiest type of data to obtain.
- Archival Data is data that has been backed up and stored. This could mean backup tapes, CDs, floppies, or entire hard drives.
- Latent Data is the data that requires specialized equipment to access such as information that has been deleted or partially overwritten. Latent data is the most difficult and time consuming type of information to collect.

A computer investigation deals with all kinds of data. Computer forensics is all about obtaining the proof of a crime or breech of policy. It focuses on obtaining proof of an illegal misuse of computers in a way that could lead to the prosecution of the culprit.

1.6. Data Mining and Digital Forensics

Data Mining is the application of algorithms for extracting the patterns from data and to provide useful knowledge for decision making. Data Mining has several applications in Digital Forensics. Data Mining involves identifying correlations in forensic data (association), discovering and sorting forensic data into groups based on similarity (classification), locating groups of latent facts (clustering), and discovering patterns in data that may lead to useful predictions (forecasting). Data mining have unlimited potential in field of Digital Forensics where tools and models are developed to help investigators to find data or clues which they searching for, much more efficiently and faster.

There are so many tools used in digital forensic investigation. These tools ensure that digital evidence is acquired and preserved properly and that accuracy of results regarding the processing of digital evidence is maintained. . With the help of forensic tools we can determine the security flaws in the computer system in against to the person who destroyed our computer based security.

Digital Forensics Tools are basically classified as:

Table 2: Tools in Digital Forensics

Hardware Forensics Tool	Used for Single-purpose components or complete computer systems and servers.
Software Forensic Tool	Used for Command-line applications and GUI applications.

Table3: Digital Forensics and Data Mining Techniques

Digital Forensic Techniques	Data Mining Techniques	Tools
Data Recovery, Data generation and Pre-processing	Statistical test Analysis Bartlett's test of sphericity Kaiser-Meyer-Olkin (KMO)	Recuva FTK Encase Sleuth kit/Autopsy ProDiscover
Data Analysis	Clustering – K-means, EM, Hierarchical Clustering	weka
	Classification – Supervised learning - Decision Tree, Neural Networks, SVM, Naïve Baiyesian	weka
	Unsupervised learning – PCA, Karnohuen Map	-
	Frequent Pattern Mining/Association rule Mining - Apriori, Eclat	weka
	Named Entity recognition	Lingepipe
	Visualization	Cyber Forensics Time Lab
	Statistical Analysis and Anamoly Detection	EMT/MET
	Recursive data mining	-
	Phishing	Invisible Witness

1.7. Important Forensics Techniques

1.7.1. Imaging. One of the first techniques used in a digital forensics investigation is to image, or copy, the media to be examined. Though this seems to be a straightforward step at first, modern Operating Systems (OSs) perform many operations on file systems when connected, such as indexing or journal resolution. Without care, media can be modified, however slightly, and the integrity of the evidence can be compromised.

1.7.2. Hashing. To quickly identify a file and to provide authenticity that an image or file was not modified, the forensic community adopted cryptographic hashing. Modern hashing functions use one way Cryptographic functions to obtain a hash.

1.7.3. Carving. One category of tools in the digital forensic toolkit is called file carvers. These tools allow the Scanning of disk blocks that don't belong to current files to find deleted data. Carvers use known header and footer signatures to combine these 'unused' nodes into the original files that were deleted. Carving can recover deleted but not overwritten files as well as temporarily cached files on media.

1.8. Role of Data Mining in Digital Forensics Investigation

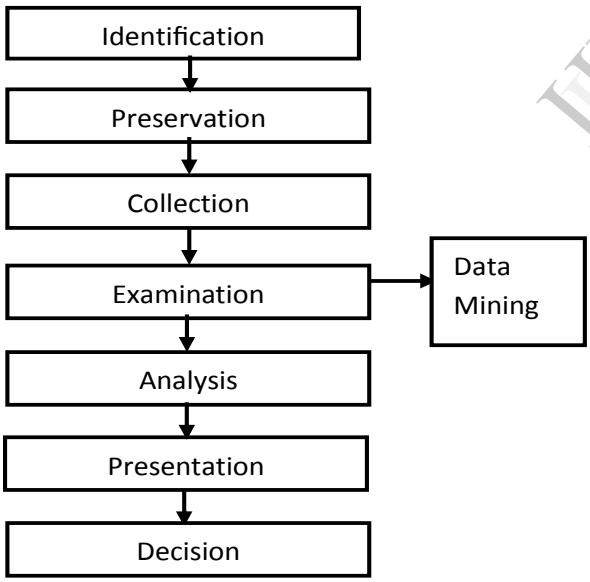

Fig 4: Role of data mining in digital forensics Investigation

Data Mining in Digital Forensic Investigation Process includes processes like Identification, Preservation, Collection, Examination, Analysis, Presentation and Decision. Identification process recognizes an incident from indicators and determines its type. Preservation process includes packaging, transportation and storage. Appropriate procedures should be followed and documented to ensure that the electronic evidence collected is not altered or destroyed. Collection process collects the Evidence of the digital or mobile devices are an important step and required a proper procedure or guideline to make them work. Examination phase involves examining the contents of the collected evidence by forensic expert and extracting information, which is critical for proving the case.

Analysis phase is implemented by using Data Mining techniques. This step is conducted by the investigative team on the basis of the results of the examination of the evidence. Identifying relationships between fragments of data, analyzing hidden data, determining the significance of the information obtained from the examination phase, reconstructing the event data, based on the extracted data and arriving at proper conclusions etc. are some of the activities to be performed at this stage. Presentation phase includes packaging, transportation and storage. Appropriate procedures should be followed and documented to ensure that the electronic evidence collected is not altered or destroyed. The phase is the decision phase. This involves reviewing all the steps in the investigation and identifying areas of improvement. As part of the decision phase, the results and their subsequent interpretation can be used for further refining the gathering, examination and analysis of evidence in future investigations.

2. Literature Survey

Association Rule Mining is one of the important areas of research, receiving increasing attention. It is an essential part of Knowledge Discovery in Databases (KDD). The scope of Association Rule Mining and KDD is very broad. It has been employed to profile user behavior and identify irregularities in log files such irregularities can assist in locating evidence that might be crucial to a digital investigation [1].

An effective digital text analysis strategy, relying on clustering- based text mining techniques, is introduced for investigational purposes[2]. This methodology is experimentally applied to the publicly available Enron dataset that well fits a plausible forensics analysis context. They proposed the tool by which an analyst can exploit the obtained clusters in order to get useful investigative information, in particular the tool proves effective when one has to cope with a notable amount

of data, when a human operator cannot manually proceed to inspection.

Outlier analysis has been utilized to locate potential evidence in files and directories that have been hidden or that are different from their surrounding files and directories[3]. Outlier analysis applied in digital forensic is to locate hidden files, directory structure of the files, and the characteristics of each file within a directory are compared to detect potential outliers. This approach is similar to that used when locating hidden data.

Support vector machines (SVM) The SVM have been utilized in several research areas in the field of digital forensics. A support vector machine (SVM) is an algorithm for classification that seeks categorized data based on certain fundamental features of the data.[12]

In one instance, a support vector machine was applied to determine the gender of the author of an e-mail based on the gender-preferential language used by the author[4]. In another instance, a support vector machine was applied to determine the authorship of an e- mail. Based on the content of the e-mail, each e-mail was classified according to its likely author[5].

An approach for preparation, generation, storing and analyzing of data, retrieved from digital devices which pose as evidence in forensic analysis[6]. Attribute classification model has been presented to categorized user files. The data mining tools has been used to identify user ownership and validating the reliability of the pre-processed data. This work proposes a practical framework for digital forensics on hard drives. . It is possible to identify a specific user group hard drive with the help of attribute classification of the data retrieve from the digital evidence. The occurrence of text files in the hard drive are more than other files, and the metadata of the document/pdf files are reflecting that the data contains in the hard drive are preferably belongs to any academic person such as research scholar and the contents of the files can be strongly used to identify the behaviour and the area of the person which he/she belongs.

In 2007, Beebe and Clark in their work proposed pre-retrieval and post-retrieval clustering of digital forensics text string search results. Though their work is focused on text mining, the data clustering algorithms used have shown success in efficiency and improving information retrieval efforts[7]

.Digital forensics text string searching: "Improving Information retrieval effectiveness by thematically clustering search results"[11], This paper explains the hidden evidence acquisition from file system. Second section explains investigation on the Network. There are two types of investigation in network, live data acquisition (Packet capturing and analysis) and log file analysis. Third section explains crime data mining. On the basis they propose a new system with Digital forensic tool for decision making in the computer security domain.

In paper "Event Sequence Mining to Develop Profiles for Computer Forensic Investigation Purposes", events are compiled from potentially numerous sources are grouped according to some criteria and frequently occurring event sequences are established[10]. Here, the methodology and techniques to extract and contrast these sequences are described by using Sequential Pattern Mining algorithm.

3. Proposed Work

In this paper we are investigating on fraud emails. If any email contents have been changed by attacker and then it is send to recipient. So we are generating our own algorithm named as Fraud Detection algorithm to check whether there is fraud email in inbox or not. When attacker changed the contents of email, then size of file will also be changed, so that it will changes sequence of mails in inbox. So we are introducing the cyber Forensics investigation by using GSP (Generalized Sequential pattern) sequence mining algorithm to rearrange emails and find out which email is fraud and what changes have been made by attacker in that particular email. Sequence mining problems are mostly solved by the algorithms which are based on a priory algorithm of association rules.

3.1. Fraud Detection algorithm

1. Initialize the data set
2. Take content from email
3. Check number of log files through different emails
4. Search the text which is disturbed
5. If text=1

   Result is same

   Else

   Result is changed

3.2. GSP Algorithm

This algorithm makes multiple database passes. In the first pass, all single items (1-sequences) are counted. From the frequent items, a set of candidate 2-sequences are formed, and another pass is made to identify their frequency. The frequent 2-sequences are used to generate the candidate 3-sequences, and this process is repeated until no more frequent sequences are found.

Algorithm. F1 = the set of frequent 1-sequence k=2,

do while F(k-1)!= Null;

Generate candidate sets Ck (set of candidate k-sequences);

For all input sequences s in the database D

Do

Increment count of all a in Ck if s supports a

Fk = {a Є Ck such that its frequency exceeds the threshold}

k= k+1;

Result = Set of all frequent sequences is the union of all Fks

End do

End do

4. Conclusion

This paper introduced new algorithm called Fraud Detection Algorithm and GSP algorithm for cyber forensics investigation. Email is most widely used way of written communication over the internet and with the advent of World Wide Web emails frauds are increasing. In cyber forensics investigation process emails can be considered as powerful evidence. So our paper plays important role in preventing the email frauds.

5. References

[1] Agrawal, R., Imielinski, T. & Swami 1993 A. Mining association rules between sets of items in large databases. Proceedings of the ACM SIGMOD International Conference on Management of Data, 207 – 216.)

[2] Text Clustering for Digital Forensics Analysis by Sergio Decherchi1, Simone Tacconi2, Judith Redi1, Fabio Sangiacomo1, Alessio Leoncini1 and Rodolfo Zunino1.

[3] Brian D. Carrier, Eugene H. Spafford 2005. Automated Digital Evidence Target Definition Using Outlier Analysis and Existing Evidence, Digital Forensic Research Workshop (DFRWS)

[4] Corney, M., de Vel, O., Anderson, A., and Mohay, G. 2002. Gender preferential Text Mining of E-mail Discourse, The 18th annual Computer Security Applications Conference (ACSAC2002). Press, Volume 30, Issue 4, 55–64.

[5] De Vel, O., Corney, M. and Mohay, G. 2001. Mining E- Mail Content for Author Identification Forensics, SIGM OD Record, ACM

[6] International Journal of Computer Applications (0975 – 8887) Volume 50 – No.4, July 2012

[7] Journal of Information Security, 2012, 3, 196-201 doi:10.4236/jis.2012.33024 Published Online July 2012

[8] Data Mining Concepts and Techniques, 2ed by Jiawei Han, Kamber M Morgan 2005. Kaufmann Publishers.

[9] International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, No.6, November 2012

[10] Event Sequence Mining to Develop Profiles for Computer Forensic Investigation Purposes by Tamas Abraham, Information Networks Division Defense Science and Technology Organization.

[11] Jan Guynes, Clark Nicole, Lang Beebe (2007),” Digital forensics text string searching: Improving Information retrieval effectiveness by thematically clustering search results”, In 6th Annual Digital Forensic Research Workshop, volume 4,

[12] Joachims T. 2002. Optimizing search engines using click through data. In Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD).