International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 3 17 - 22

Data mining Techniques for Digital Forensic Analysis

Ashwinkumar Malwadkar
Department of Information Technology
K. J. Somaiya College of Engineering
Mumbai, Maharashtra
ashwin.malwadkar1990@gmail.com

Prof. Sonali Patil
Department of Information Technology
K. J. Somaiya College of Engineering
Mumbai, Maharashtra
sonalipatil@somaiya.edu

Abstract: The computer forensic involve the protection, classification, taking out information and documents the evidence stored as data or magnetically encoded information. But the organizations have an increasing amount of data from many sources like computing peripherals, personal digital assistants (PDA), consumer electronic devices, computer systems, networking equipment and various types of media, among other sources. To find similar kinds of evidences, crimes happened previously, the law enforcement officers, police forces and detective agencies is time consuming and headache.

The main motive of this work is by combining a data mining techniques with computer forensic tools to get the data ready for analysis, find crime patterns, understand the mind of the criminal, assist investigation agencies have to be one step ahead of the bad guys, to speed up the process of solving crimes and carry out computer forensics analyses for criminal affairs.

Keywords-*Digital Forensics, NTFS, MFT, PDA, MBR, Data mining, IDS*

*****

I. INTRODUCTION

Digital Forensics is the application of science to identify, collect, examine, and analysis the data, while preserving the integrity of the information and maintaining a strict chain of custody for the data. Data contains the distinct pieces of digital information that have been formatted in a specific way. Organizations have an escalating amount of data from many sources. For example, data can be transferred or stored by standard networking equipment, computer systems, computing peripherals, personal digital assistant (PDA), consumer electronic device and different types of media, enclosed by other sources.

Data is an important tool and weapon for companies, to capture larger marketplace. Due to the importance of Data, its' security has become a major issue in the I.T. industry. So the organization will have difficulty determining what events have occurred within its systems and networks, such as exposures of secured, sensitive data.

The law enforcement officer, detective agencies, police departments having problem to solve this cases because of the large volumes of crime-related data are existed. Due to the crime-related complexity relationships, the widely used methods of crime analysis are out-of-date that consume many time and human resources. Moreover, these methods are not able to obtain all influential parameters because of their high amount of human interference, therefore, using an intelligent and systematic approach for crime analysis more than ever. Whereas, the data mining techniques can be the key solution.

With the use of data mining techniques like clustering, classification used to track, identify crimes, crimes patterns, which have started helping the law enforcement officers and detectives to speed up the process of solving crimes. Here we will take an interdisciplinary approach between computer science and criminal justice to develop a data mining paradigm that can help solve crimes faster.

II. LITERATURE SURVEY

Digital forensics is about finding evidence present in the digital devices that is sufficiently reliable to stand up in court and be convincing. Digital forensics mainly used to preserve, identify, extract, and document the digital evidence stored as data or magnetically encoded information[8].

The process of acquiring, examining, and applying digital evidence is crucial to the success of prosecuting a cyber-criminal. With the continuous evolution of technology, it is difficult for law enforcement and computer professionals to stay one step ahead of technologically savvy criminals. To effectively combat cyber-crime, greater emphasis must be placed in the digital forensic field of study.

A. Steps for Digital Forensic

1) Assessment:

You must be able to distinguish between evidence and junk data. For this, you should know what the data is, where it is located, and how it is stored.

2) Acquisition:

The evidence you find must be preserved as close as possible to its original state. Any changes made during this phase must be documented and justified.

3) Authentication:

At least two copies are taken of the evidential computer. One of these is sealed in the presence of the computer owner and then placed in secure storage. This is the master copy and it will only be opened for examination under instruction from the court in the event of a challenge to the evidence presented after forensic analysis on the second copy.

4) Analysis:

The stored evidence must be analysed to extract the useful information and recreate the chain of events.

5) Articulation:

The manner of presentation is important, and it must be understandable to court effectively. It should remain technically correct and credible. A good presenter can help in this respect.

6) Archival:

After the case is closed seal the original evidence and keeps it in secure storage place because it is a chance to reopen the case after some time or years, then it's required to resubmit in court.

B. Types of Digital Forensic

1) Computer Forensic:

The core underlying principle within computer forensics is preservation of data. Therefore, during all stages of examination and analysis a forensic examiner will work on duplicates of the original evidence rather than the original.

Computer forensic used to preserve, identify, extract, and document the evidence from the storage media. File management systems or file systems is a part of operating system which organize and locate sectors for file storage.

A computer system fundamentally has two sources of data that are of interest to a forensic examiner: volatile and non-volatile memory. Volatile memory primarily relates to the main RAM of a computer, but also includes cache memory and even register memory and the non-volatile memory does not lost data when the system is switched off i.e. hard disk [8].

2) File System Analysis:

File system analysis examines data in a volume (i.e., a partition or disk) and interprets them as a file system. There are many end results from this process, but examples include listing the files in a directory, recovering deleted content, and viewing the contents of a sector.

File systems provide a mechanism for users to store data in a hierarchy of files and directories. A file system consists of structural and user data that are organized such that the computer knows where to find them [7].

*a) Hidden Evidence Analysis in the File System:*Suspects can hide their sensitive data in various areas of the file system such as volume slack; file slack, bad clusters, deleted file spaces.

i. *Hard Disk*: The maintenance track/Protected Area on ATA disks are used to hide information.

ii. *File System Tables*: A file allocation table in FAT and Master File Table (MFT) in NTFS are used to keep track of files. MFT entries are manipulated to hide vital and sensitive information.

b) File Deletion: file is removed from the table, by that making it appear that it does not exist anymore. The clusters used by the deleted file are marked as being free and can now be used to store other data. However, even if the record is gone, the data may still reside in the clusters of the hard disk. That data can be recovered by calculating start and end of the file in hex format and copy it into a text file and save with corresponding extension.

Restore a JPEG image:

a) Open file in the hex pattern.

b) Analyze the file signature.

c) Replicate from starting signature up to ending signature.

d) For example (JPEG/JPG/JPE/JFIF file starting sig-nature is FF D8 FF E1 XX XX 45 78 69 66 00 (EXIF in ascii Exchangeable image file format trailer is FF D9).

e) Open the file with corresponding application.

c) Partition Tables: Information about how partitions are setup on a machine is stored in a partition table, which is a part of the Master Boot Record (MBR). When the computer is booted, the partition table allows the computer to understand how the hard disk is organized and then passes this information to the operating system. When a partition is deleted, the entry in the partition table is removed, making the data inaccessible. However, even though the partition entry has been removed, the data still resides on the hard disk [7].

d) Slack Space: A file system may not use an entire partition. The space after the end of the volume called volume *slack* that can be used to hide data. The space between Partitions is also vulnerable for hiding data, *file slack* space is another hidden storage. **Figure 1**shows slack spaces in a Disk. When a file does not end on a sector boundary, operating systems prior to Windows 95 a fill the rest of the sector with data from RAM, giving it the name RAM slack [7].

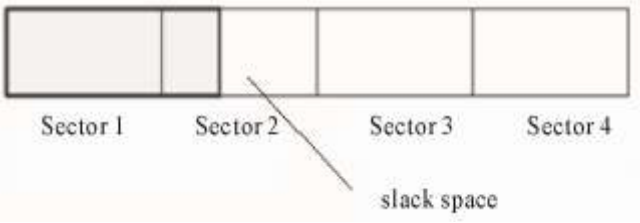

Figure 1: Slack Space Structure

When a file is deleted, its entry in the file system is updated to indicate its deleted status and the clusters that were previously allocated to storing are unallocated and can be reused to store a new file. However, the data are remains on the disk and it is often possible to retrieve a file immediately after it has been deleted. The data will re-main on the disk until a new file overwrites them. Whenever, if the new file does not take up the entire cluster, a some part of the old file might remain in the slack space. In this case, a portion of a file can be retrieved long after it has been deleted and partially overwritten.

e) Free Space:However, when a file is moved from one hard disk or partition to another, it is actually a multistep process of replicating and deleting the file. First, a new copy

of the file is created on the target partition. After the file has been copied, the original file is then deleted. This process also requires some housekeeping in the FAT or MFT tables. A new entry is created in the table on the partition where it has been copied, whereas the record for the deleted file is removed from the table on its partition. When a file get deleted, that space considered as free space, there also criminal can hide sensitive information [6].

f) Faked Bad Clusters: Clusters marked as bad may be

used to hide data. In NFTS, bad clusters are marked in metadata file called $BadClus, which is in MFT entry 8.Originally, $BadClus is a sparse file which file size is set to the size of entire file system. When bad clusters are detected, they will be allocated to this file. The size of data that can be hidden with this technique is unlimited. Suspects can simply allocate more clusters [6].

3) Boot Sector Analysis:

The recent cyber-crime trends are to use different obfuscated techniques such as disguising file names, hiding attributes and deleting files to intrude the computer system. Since the Windows operating system does not zero the slack space, it becomes a vehicle to hide data, especially in $Boot file. Hence, in this study, we have analysed the hidden data in the $Boot file structure. The $Boot entry is stored in a metadata file at the first cluster in sector 0 of the file system, called $Boot, from where the system boots. It is the only metadata file that has a static locationso that it cannot be relocated. Microsoft allocates the first 16 sectors of the file system to $Boot and only half of these sectors contains non-zero values. The $Boot metadata file structure is located in MFT entry 7 and contains the boot sector of the file system. It contains knowledgeabout the size of the volume, clusters and the MFT. The $Boot metadata file structure has four attributes, namely, $STANDARD_INFORMATION, $FILE_NAME, $SECURITY_DESCRIPTION and $DATA. The $STANDARD_INFORMATION attribute contains temporal information such as flags, owner, security ID and the last accessed, written, and created times. $Boot data structure of theNTFS file system could be used to hide data. By analysing the hidden data in the boot sector, onecould provide useful information for digital forensics

4) Network Forensics:

Network forensics deals with the capture, recording or analysis of network events in order to discover evidential information about the source of security attacks in a court of law. With the rapid growth and use of Internet, network forensics has become an integral part of digital forensics [10].

The Network Forensics include

- The analysis of IDS and firewall logs as evidence.
- The back tracking of network packets and TCP connections.
- The analysis of network related artifacts on forensically acquired hard disks.
- The analysis of logs generated by network services and network applications.
- The seizure and analysis of network traffic using sniffers and NFAT1 devices
- Collecting data from remote network services.

1) *Procedure for Network Live Acquisition*:

a) Create a bootable forensic CD.
b) Perform Remote access to the suspected machine or insert bootable CD in suspects' machine directly.
c) Record or keep a log of all the actions of forensic investigator.
d) If need to take out away the evidence then use USB.
e) Next, Take a copy of the physical memory using a forensic tool example memfetch.
f) Create an image of the drive.
g) For Intrusion first check Root kit is installed or not, for that root kit revealers are available.
h) Perform hash value of the created image for integrity checking.

2) *Network Investigation Tools:*

There is a powerful windows tools available at sysinternal

Filemon- shows file system activity.
RegMon- shows all Registry data in real time.
Process Explorer- shows what files, registry keys and dynamic link libraries (DLLs) are loaded at a specific time.
Pstools is a suite created by Sysinternals that includes the following tools.
PsExec-Run processes remotely.
PsGetSid-Displays the security identifier of a computer.
PsKill-Kills processes by name or processes ID.
PsList-Lists detailed information about processes.
PsLoggedOn-Displays who's logged on locally.
PsPassword-Allows user to change account passwords.
PsService-Enables to view and control services.
PsShutDown-Shutdown &optionally restarts a computer.
PsSuspend-Allows to suspend processes.
Tcpdump and Ethereal-Packet sniffers.

5) Email Forensics:

Email is one of the most common ways people communicate, ranging from internal meetings, to distribute the documents and general conversation. Emails are now being used for all sorts of communication including providing authentication, non-repudiation, confidentiality and data integrity.

The tools help to identify the point of origin of the message, trace the path traversed by the message (used to identify the spammers) and also to identify the phishing emails that try to obtain confidential information from the receiver.

EMailTrackerPro analyses the header of an email to detect the IP address of the machine that sent the message so that the sender can be tracked down. It helps to track emails to a country or region of the world, showing information on a global map.

SmartWhoIs is a freeware network utility to look up all the available information about an IP address, hostname or domain, including country, state or province, city, name of the network provider, administrator and technical support contact information [10].

6) Web Forensics:

Web forensics deals with collecting critical information related to a crime by exploring the browsing history of a person, the number of times a website has been visited, the duration of each visit, the files that have been uploaded and downloaded from the visited website, the cookies setup as part of the visit and other critical information.

Mandiant Web Historian assists users in reviewing web site URLs that are stored in the history files of the most commonly used web browsers. It allows the forensic examiner to determine what, when, where, and how the intruders looked into the different sites [10].

Index.dat analyser is a forensic tool to view, examine and delete the contents of index.dat files. The tool can be used to simultaneously or individually view the cookies, cache and browsing history. The tool provides support to directly visit the website listed in the output of the analyser and also to open the file uploaded to or downloaded from the website.

7) Packet Sniffers:

A Sniffer is software that collects traffic flowing into and out of a computer attached to a network [11]. Network engineers, system administrators and security professionals use sniffers to monitor and collect information about different communications occurring over a network. Sniffers are used as the main source for data collection in Intrusion Detection Systems (IDS) to match packets against a rule set designed to notify anything malicious or strange. Tools Used Ethereal is an open source software and widely used as a network packet analyser. It captures live network packets. It displays the information in the headers of all the protocols used in the transmission of the packets captured. Depending on user needs it filters the packets.

WinPcap is the tool used for link-layer network access in Windows. WinPcap includes a network statistics engine and provides support for kernel-level packet filtering and remote packet capture.

AirPcap can be used to capture the control frames (ACK, RTS, CTS), management frames (Beacon, Probe Requests and Responses, Authentication) and data frames of the 802.11 traffic.

C. Data Mining Techniques:

Data mining is defined as the identification of interesting structure in data, where structure designates patterns, statistical or predictive picture of the data, and relationships among parts of the data. Data mining in the context of crime and intelligence analysis for national security is still a young field. The following describes our applications of different techniques in crime data mining [3].

1) Clustering technique group data objects into classes by similar characteristics to minimize or maximize interclass similarity for instance, to identify suspects that bearing the crimes in similar ways or discriminate among groups belonging to different gangs. These techniques do not have a set of predefined classes for assigning items.

2) Association rule mining determines frequently occurring item sets in a database and offerings some patterns as rules that been used in network intrusion detection to develop the connection rules from users' interaction history. Investigators also can apply this technique to network intruders' profiles to help detect potential future network attacks. In network intrusion detection, this approach can identify intrusion patterns among time-stamped data. Showing hidden patterns benefits crime analysis, but to obtain meaningful results requires rich and highly structured data.

3) Deviation detection utilizes the particular measures to study data that differs noticeably from the rest of the data. Also called outlier detection, investigators can use this technique to fraud detection, network intrusion detection, and other crime analyses. However, such activities can sometimes appear to be normal, making it difficult to identify outliers.

4) Classification finds mutual properties between various Crime entities and arranges them into predefined classes that have been applied for identifying the source of email spamming according to the sender's structural features and linguistic patterns. Often used to predict crime trends, classification can reduce the time required to identify crime entities. However, the technique requires a predefined classification scheme [5].

5) String comparator techniques that show the relation the textual fields in pairs of database records and calculate the correspondence among the records that can detect deceptive information in criminal records for instance the name and address. The researchers can utilize string comparators to evaluate textual data that often need intensive computation.

D. Data mining algorithm:

1) Identify itemsets/variables from a case report (our proposed system stores these variables as attributes of tables, filesystem table, network table).

2) Item sets I = {I1, I2, I3 ⋯ Im}.

3) Set of actions A = {a1, a2, a3 ⋯ an}.

4) Find frequent item sets by using Apriori algorithm.

5) Make Association Rules

i.e. It is a rule in the form X → Y showing an association between X and Y that if X occurs then Y will occur.

If the attacker accessed operating system files then we can say motive of attack is system Crash.

If the attacker attacked Database login and Password steel then we can say criminal motive for data theft/data change. This maximum frequent item sets also shows attack patterns. Finding other signs of evidence Correlation, contingences (Consider these values while making rule sets).

6) Set SQL queries according to the rules.

7) Retrieve data.

1) Apriori Algorithm:

The Apriori algorithm is the most well-known association rule algorithm and is used in most commercial products. It uses the following property which we call large itemset property.

"Any subset of large itemset must be large".

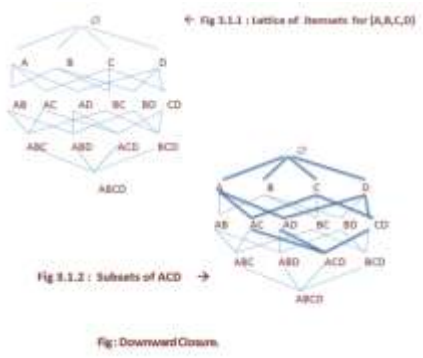

Figure 2. Downword Closure

The large itemset are also said to be downward closed because if an itemset satisfies the minimum support requirements so do all of its subset [9].

The basic idea of the Apriori algorithm is to generate candidate itemsets of a particular size and then scan the database to count these to see if they are large.An itemset is considered as a candidate only if all its subset also are large. An algorithm called Apriori-Gen is used to generate the candidate itemsets for each pass after the first. All singleton itemsets of are used as candidates in the first pass. After the first scan, every large itemset is combined with every other itemset.

Algorithm:

Input:

I // Itemsets

D // Database of transaction

s // Support

Output :

L // Large itemsets

Apriori algorithm :

k = 0 ; // k is used as the scan number.

L = Ø ;

C_1 = I ;

repeat

k = k + 1;

L_k = Ø ;

for each I_i Є C_k do

c_i = 0 ; //Initial counts for each itemset are 0.

for each t_j Є D do

for each I_i Є C_k do

if I_i Є t_j then

$c_i = c_i + 1$;

for each I_i Є C_k do

if c_i>= (s * | D |) do

$L_k = L_k$ U I_i ;

L = L U L_k ;

C_{k+1} = Apriori-Gen(L_k)

Until C_{k+1} = Ø ;

Algorithm Apriori-Gen :

Input :

L_{i-1} // Large itemsets of size i-1

Output :

C_i // Candidates of size i

Apriori-Gen Algorithm :

C_i = Ø ;

for each I Є L_{i-1} do

for each J ≠ I Є L_{i-1} do

if i-2 of the elements in I and J are equal then

$C_k = C_k$ U { I U J };

III. PROPOSED SYSTEM

Our proposed system is the combination of a data mining techniques and computer forensic tools. This helps to organization to get the data ready for analysis, find crime patterns, understand the mind of the criminal, assist investigation agencies have to be one step ahead of the bad guys, to speed up the process of solving crimes and carry out computer forensics analyses for criminal proceedings.

With the use of data mining techniques we can track, identify crimes, crimes patterns that helps to solve crimes fast and digital forensics is the application of science to the identification, collection, examination, and analysis of data while preserving the integrity of the information.

These productive measures can be initiated to alert administrator about similar kinds of attacks happened in future for preventing upcoming cyber attack.

A. Block Diagram of Proposed System:

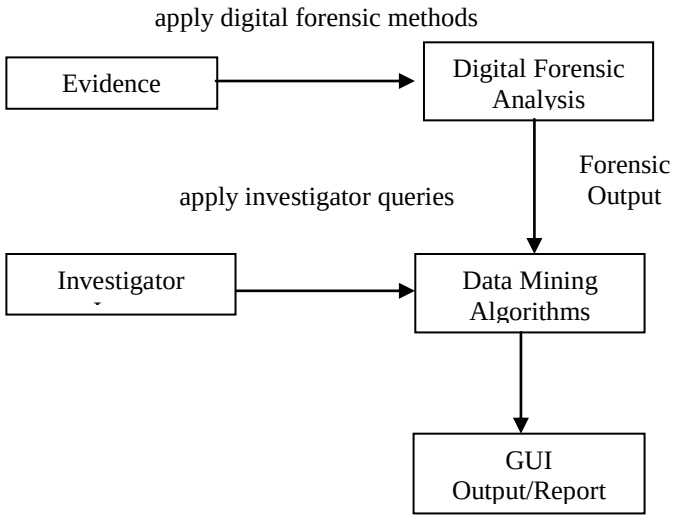

Figure 4: Block DiagramEvidence

Digital evidence or electronic evidence is any probative information stored or transmitted in digital form Digital evidence includes information on computers, audio files, video recordings, digital images, emails, digital photographs, ATM transaction logs, word processing documents, instant message histories, spreadsheets, internet browser histories, databases, contents of computer memory, computer backups, GPS tracks, system logs this evidences must be essential in computer and internet crimes.

1) Digital Forensic Analysis:

Digital forensics encompassing the recovery and investigation of material found in digital devices, often in relation to computer crime. The goal of digital forensics is to examine digital media in a forensically sound manner with the aim of identifying, preserving, recovering, analysing and presenting facts and opinions about the digital information.

2) Data mining Algorithm:

Data mining algorithm contains the various data mining algorithm like clustering algorithm, association rule mining algorithm, classification algorithm, which can be used to find patterns, keeps the track of information, identification of interesting structure in data, statistical or predictive models of the data, and relationships among parts of the data etc.

3) GUI Output/Report:

The GUI Output/Report shows output or create report what the investigator wants. It get data ready for analysis, shows the crime patterns, create reports like to find motive behind the crime, pattern of cyber-attacks and counts of attacks types happened during a period.

IV. CONCLUSION

Digital forensics is the science of identifying, extracting, analysing and presenting the digital evidence that has been stored in the digital devices and data mining is the method for the identification of interesting structure in data, where structure designates patterns, statistical or predictive models of the data, and relationships among parts of the data.

Our pre synopsis work defined the new proposed technique with the combination of digital forensic analysis and data mining techniques. The proposed system is designed for to get the data ready for analysis, find crime patterns, finding motive, pattern of cyber-attacks and counts of attacks types happened during a period. Hence the proposed tool helps to enables the system administrators to minimize the system vulnerability, understand the mind of the criminal, assist investigation agencies have to be one step ahead of the bad guys, to speed up the process of solving crimes and carry out computer forensics analyses for criminal proceedings

REFERENCES

[1] Cheong Kai Wee, "Analysis of Hidden Data in NTFS File System,"Edith Cowan University.

[2] Shyam Varan Nath, "Crime Pattern Detection Using Data Mining", Oracle Corporation.

[3] Hsinchun Chen, Wingyan Chung, Yi Qin, Michael Chau, Jennifer Jie Xu, Gang Wang, Rong Zheng, Homa Atabakhsh," Crime Data Mining: An Overview and Case Studies", *Proceeding of ACM Inter-national Conference*, Vol. 130, 2003, pp. 1-5.

[4] Chung-Hsien Yu, Max W. Ward, Melissa Morabito, Wei Ding," Crime Forecasting Using Data Mining Techniques",University of Massachusetts Boston.

[5] Javad Hosseinkhani, Mohammad Koochakzaei, Solmaz Keikhaee, Javid Hosseinkhani Naniz, "Detecting Suspicion Information on the Web Using Crime Data Mining Techniques*", International Journal of Advanced Computer Science and Information Technology (IJACSIT)*, Vol. 3, No. 1, 2014, Page: 32-41, ISSN: 2296-1739

[6] Mamoun Alazab, Sitalakshmi Venkatraman, Paul Watters, "Effective Digital Forensic Analysis of the NTFS Disk Image," *Ubiquitous Computing and Communication Journal*, Vol. 4, No. 3, 2009, pp. 551-558.

[7] Brian Carrier, "File System Forensic Analysis", Addison Wesley Professional, ISBN: 0-32 126817-2.

[8] John R. Vacca, "Computer Forensics: Computer Crime Scene Investigation", Second Edition.ISBN: 1-58450-389-0 ISBN-13: 978-1-58450-389-7

[9] Margaret H Dunham, "Data Mining*:* Introductory and Advanced Topics" Publisher, Pearson Education, 2006.

[10] Natarajan Meghanathan, Sumanth Reddy Allam and Loretta A. Moore, "Tools And Techniques For Network Forensics", *International Journal of Network Security & Its Applications (IJNSA)*, Vol .1, No.1, April 2009.

[11] Bruce J. Nikkel, "Generalizing sources of live network evidence", Whitepaper 2005.

[12] Karen Kent Suzanne Chevalier Tim Grance Hung Dang, "Guide to Integrating Forensic Techniques into Incident Response", National Institute of Standards and Technology Special Publication 800-86.