Identification of User Ownership in Digital Forensic using Data Mining Technique

Kailash Kumar
Phd(CSE) Student
PEC University of Technology, Chandigarh

Sanjeev Sofat
Prof. and Head CSE Deptt.
PEC University of Technology, Chandigarh

Naveen Aggarwal
Asst. Professor
UIET, Chandigarh

S.K.Jain
Deputy Director (Ball.)
CFSL, Chandigarh

ABSTRACT
As existing technology used by criminal rapidly changes and growing, digital forensics is also growing and important fields of research for current intelligence, law enforcement and military organizations today. As huge information is stored in digital form, the need and ability to analyze and process this information for relevant evidence has grown in complexity. 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. Data Preparation/ Generation, Data warehousing and Data Mining, are the three essential features involved in the investigation process. The purpose of data mining technique is to find the valuable relationships between data items. This paper proposes an approach for preparation, generation, storing and analyzing of data, retrieved from digital devices which pose as evidence in forensic analysis. 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.

Keywords
Digital Forensic, Hard Drive, Framework, Data Preprocessing.

1. INTRODUCTION
Providing accurate information derived through the use of proven and well-understood methodologies has always been the goal of traditional forensic analysis. Forensic Science applied in courts of law has sought to use commonly applied techniques and tools only after rigorous, repetitive testing and thorough scientific analysis. In 2006 Richard and Roussev discussed the urgent need for new tools and strategies for the rapid turnaround of large forensics targets. They focused on the acquisition and analysis of forensic evidence and argued that current forensics tools are inadequate given the increased complexity of cases, increased size of targets, better awareness of the capabilities of digital forensics, and multi-computing scenarios. Data mining is the application of algorithms for extracting patterns from data [12]. These extracted patterns will provide useful knowledge to decision makers. As such, there has been an increasing demand for data mining tools to help organizations uncover knowledge that can guide in decision making.
For law enforcement agencies it's tedious and time consuming task to find out the user from the large quantity of the acquired digital evidence drive, here we propose a method for finding user ownership information based on attribute analysis of evidence drive. With the growing sizes of databases, law enforcement and intelligence agencies face the challenge of analyzing large volumes of data involved in criminal and terrorist activities [7]. Thus, a suitable scientific method for digital forensics is data mining.

The DFRW has identified media analysis as one of the three main distinct types of digital forensic analysis, the other two being Code Analysis and Network Analysis. This paper introduces a framework for the digital forensic investigation process of physical storage device. It also takes a specific case of accessing the hard drive as a device and analyzing its contents for ownership.

This paper explains the importance of the information that exists in hard disk drive for forensic analysis and investigations. The remainder of this paper is organized as follows. Section 2 gives an overview of digital forensics. Section 3 provides a short background on data mining and its functionalities. Section 4 focuses on forensic analysis process model for extracting sensitive information from hard disk drive. Finally section 5 concludes the paper and our future work.

2. DIGITAL FORENSIC SCIENCE
The growing and relevant dependency on digital forensics is important because of the vast increase of digital source of information in the form of computerized systems, networks, which involved for data storage, processing, and transmission in all aspects of our lives. Existing technologies and those that are evolving, in support of law enforcement, to courts, military operations, business and industry, and critical infrastructure digital forensics will continue to play increasing rolls of importance. As technology rapidly became advanced we will reach unprecedented volumes of data and be faced with the challenges of managing it all [11].
Figure 1 further illustrates where digital forensic research has found its niche, and where it lies amongst the broad battle space given its enormous potential for application.

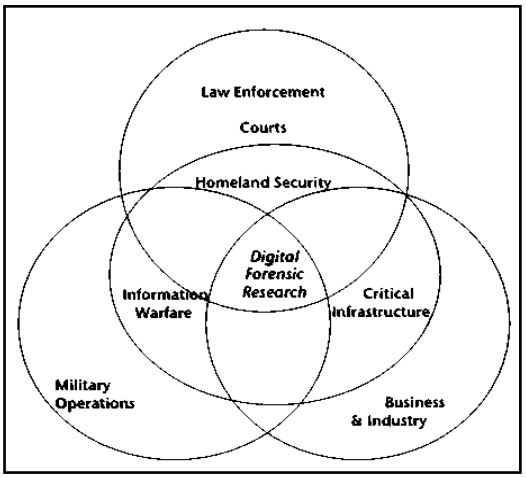

Fig1. The Nucleus of digital forensics

The founders of the first DFRWS characterized the discipline of digital forensic science with the following associated entities:
• Theory: a body of statements and principles that attempts to explain how things work
• Abstractions and models: considerations beyond the obvious, factual, or observed.
• Elements of practice: related technologies, tools, and methods.
• Corpus of literature and professional practice.
• Confidence and trust in results: usefulness, purpose [11]

2.1 Digital Forensic Investigation Process Model

The first framework developed at the DFRWS broke digital forensics down into a seven step process. They are identification, preservation, collection, examination, analysis, presentation, and lastly decision [2]. It was this time also that data mining was classified as a key area of research under the analysis step illustrates in figure2.

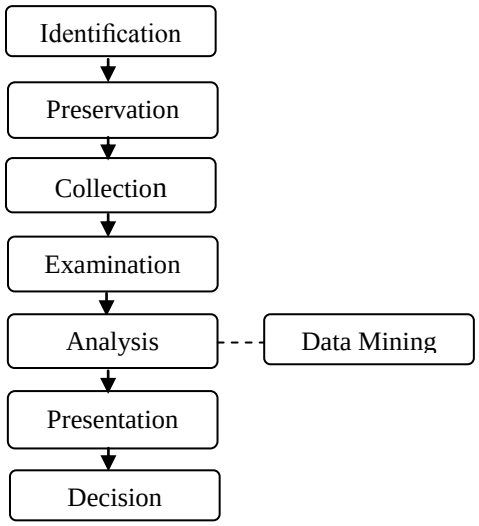

Fig2. Data Mining In Digital Forensic Investigation Process

1. Identification: This recognizes an incident from indicators and determines its type.

2. Preservation: This 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. All potential sources of evidence should be identified and labeled properly before packing.

3. Collection: Evidence collection of the digital or mobile devices is an important step and required a proper procedure or guideline to make them work. We can categorize evidence collection of the digital devices into two categories:
• Volatile Evidence Collection
• Non-Volatile Evidence Collection

4. Examination: This phase involves examining the contents of the collected evidence by forensic expert and extracting information, which is critical for proving the case. Appropriate number of evidence back-ups must be created before proceeding to examination. This phase aims at making the evidence visible, while explaining its originality and significance. Huge volumes of data collected during the volatile and non-volatile collection phases need to be converted into a manageable size and form for future analysis. Data filtering, validation, pattern matching and searching for particular keywords with regard to the nature of the crime or suspicious incident, recovering relevant ASCII as well as non- ASCII data etc. are some of the major steps performed during this phase. Personal organizer information data like address book, appointments, calendar, scheduler etc, text messages, voice messages, documents and emails are some of the common sources of evidence, which are to be examined in detail. Finding evidence for system tampering, data hiding or deleting utilities, unauthorized system modifications etc. should also be performed. Detecting and recovering hidden or obscured information is a major tedious task involved. Data should be searched thoroughly for recovering passwords, finding unusual hidden files or directories, file extension and signature mismatches etc.

5. Analysis: This step is more of a technical review 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.

6. Presentation: This 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. All potential sources of evidence should be identified and labeled properly before packing. Use of ordinary plastic bags may cause static electricity. Hence anti-static packaging of evidence is essential. The device and accessories should be put in an envelope and sealed before placing it in the evidence bag.

7. Decision: The final stage in the model 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. In many cases, much iteration of examination and analysis phases are required to get the total picture of an incident or crime. This information will also help to establish better policies and procedures in place in future.

3. DATA MINING IN DIGITAL FORENSICS

Definition: Data mining can be defined as the analysis of often large observational data sets to find un-suspected relationships and to summarize the data in novel ways that are both understandable and useful to the owner [15].

First and foremost, for data mining to be relevant, the data sets of interest must be large; if they were not then it might be feasible to manually explore the data and make a decision. There are varying scales of data sets that may be considered to be large, but this paper focuses on data and files on a single hard drive which may easily number to millions of files at a time. Once validated on a single hard drive, the goal is scale the data mining effort to much larger data sets across multiple hard drives and possibly networks. Secondly as defined, a goal of data mining is to find unsuspected or unknown relationships within data. Obviously there is no reason to report or to repackage already known relationships through data mining. This paper seeks to find the relationships among files or data, specifically the ascription or ownership of the data. From a forensics perspective such correlations or relationships discovered may be used to tie criminals, terrorists, or people of interest together. Lastly, the results of data mining must be understandable and useful.

There exists a formal methodology for data mining which includes these basic steps:
• Determine the nature and structure of the representation of the data sets to be used.
• Decide how to quantify the data; compare how well different representations fit the data
• Choose an algorithmic process to optimize the scoring function
•Decide what principles of data management are required to implement the algorithms efficiently [15]

3.1 Literature Survey of Data Mining Applied In Digital Forensics

Data mining techniques are designed for large volumes of data. Hence, they are able to support digital investigations. While such techniques have been employed in other fields, their application in digital forensics is still relatively unexplored.

Some of the data mining techniques applied in digital forensics are as follows:

3.1.1 Association rules

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].

In digital forensic, association rule mining can play an important role as to extract the login information of a user from log files of a computer system. By generating the rule sets, with the help of behavioral profiles of the user, forensic expert can detect the behavioral anomalies of users.

3.1.2 Outlier analysis

It 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 directories where the characteristics of directories at the same level are compared.

3.1.3 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 [16]. 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 [8]. In another instance, a support vector machine was applied to determine the authorship of an e-mail [10]. Based on the content of the e-mail, each e-mail was classified according to its likely author.

Image mining is one of the many activities undertaken during a digital investigation. A support vector machine can also be used to recognize certain patches or areas of an image [5]. Other instances where support vector machines have been utilized include image retrieval [5] and in executing search queries on images containing suspicious objects [4].

3.1.4 Discriminant analysis

Discriminant analysis has been employed in digital forensic to determine whether contraband images, such as child pornography, were intentionally downloaded or downloaded without the consent of the user [6]. Often, individuals prosecuted for crimes based on digital evidence claim that a Trojan horse or virus installed on their computer system was responsible. In this instance, discriminant analysis provided a mechanism for event reconstruction and enabled digital investigators to counter the Trojan defence by examining the characteristics of the data.

3.1.5 Bayesian networks

It has been used to automate digital investigations. Bayesian networks are based on Baye's theorem of posterior probability [9]. A Bayesian network is a directed acyclic graph which models probabilistic relationships among a set of random variables [13]. The aim was to gather information about likely attacks, actions performed by attackers, the most vulnerable software systems and the investigation techniques that should be used.

4. FRAME WORK FOR DIGITAL FORENSIC INVESTIGATION

A framework, for seamless communication, between the technical members of the digital forensic investigation team and the non-technical members of the judicial team, is very necessary. Defining a generic model for digital forensic investigation, sometimes pose a problem taking into account the varied devices available today. This framework is logical in its outline, scientific in its approach though it is to be adapted to comply with all the legal requirements of the country where the incident has occurred. We are proposing an efficient model for both as economical and time factor. The architecture of the model is illustrates in figure 3.

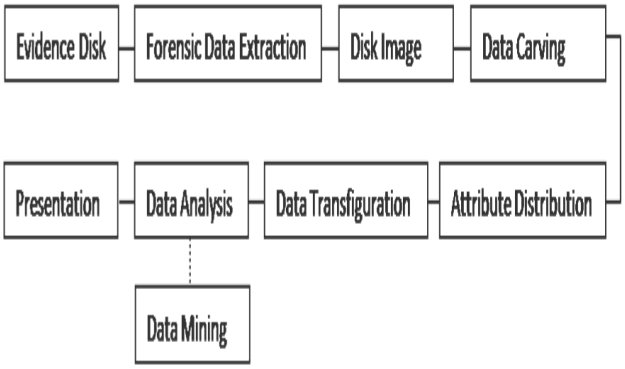

Fig3. Architecture of the Model

4.1 Forensic Data Extraction

In this section, we describe the forensic data extraction process. We conducted an empirical study using selected digital forensic tools that are predominantly used in practice. Since each utility does some specific functionality, collection of such tools were necessary to perform a comprehensive set of functionalities. Hence, the following forensic utilities / tools were adopted to conduct the experimental investigation in this research work:

In an investigation, the analysis phase is the one that most relies on the investigator's skills and experience. To analyze the data collected about a case, an investigator wants to understand and know where the suspect might have hidden the data and in what formats and what application he might have used. Some patterns in the data are important; once found and fully examined; they can lead to more evidence. In order to achieve a successful analysis, many tools are adopted to aid investigators analyze the collected data. In this research we are using **Encase forensics** Guidance Software Inc. [14] for extracting disk image from the suspect hard drive. This tool displays the files of a storage media and allows the user to navigate through the files similar to traditional file explorers. However, they provide additional features that are useful in forensics context such as displaying file headers and opening compressed files. Some of these tools provide more contextual analysis features such as queries and a time-line view of the files. However, the investigator is responsible for manually performing the analysis and gathering knowledge from the extracted data.

4.2. Methodology

The following six steps involved for experimental result that fullfill our goal.

Step 1: *Setup of a test computer:* For the experimental investigation of the effectiveness of the above tools, we created test data on a Pentium (R) Core (TM) 2 Due CPU, 1.89 GHz, 0.99 of RAM, 160 GB hard drive, write blocker device with Windows XP professional.

Step 2: *Forensically Image the Drive using Encase forensics version 6.2*: Here we create an image and analyze the registry of the suspect hard drive using the Encase program. The Encase forensics is a simple but concise tool. It saves an image of a hard disk in one file or in segments that may be later on reconstructed. The data is extracted from the original device, taking care that there is no process that writes on to the digital device under investigation. It calculates MD5 hash values and confirms the integrity of the data before closing the files. The result is an Encase raw image that we have to save in system or another fresh

hard drive of at least equal capacity. Now the raw image created by Encase program is used for analysis and examination purpose.

Step 3: *Data extraction/carving:* Given the working Encase image we will now extract whole imaged data in readable form with the Encase copy folder program. We recovered all files on to system and/or fresh hard drive of the forensic computer. The time frame for the actual data recovery depends on the duration and frequency of usage of the hard drive.

Step 4: *Attribute Distribution:* In figure 4 the attributes or the metadata extracted from each file off the hard drive is also listed. For our experiment, the attribute id is purely a sequential generated number starting with the number one, associated with each instance of data. The partition attribute details which partition of the hard drive the file belongs to; typically there are 1 primary and two secondary partitions and its data type is numeric. The file size details the size of the file as a numeric data type. The *mtime, ctime, atime,* and *dtime* attributes represent the modified time, created time, access time, and deleted time of the file in year, month, day, hour, minutes, and second's format.

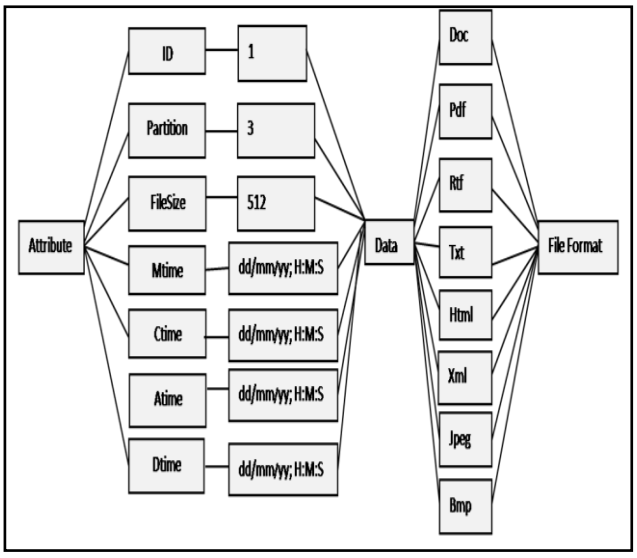

Fig4. Attribute distribution format

Step 5: *Data transfiguration:* Use any spreadsheet which is available on your laptop/personal computer. The role played by the spreadsheet can also be achieved by running the data conversion process at the database level too. The major data transformations are conversion of the data into any standard format (comma separated value used here), generation of the parent directories and extraction of the file extensions [17].

Step 6: *Data Analysis:* With the creation of our attribute distribution file, we now have the file metadata that has been extracted from our test drive. As is common in data mining, before running tests on data instances, it was necessary to clean and prepare our data for use into the WEKA workbench. We can also test our metadata statistically using Bartlett's test of sphericity and Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy [17].

After the data preparation was done, WEKA now could be used to run its suite of algorithms on the test data. The complete dataset, consisting of the file tree, the file attributes the timestamps, file size and the deleted flag, is loaded into Weka, open source

software, for analyses. On an average, the hard drive, used in this study, recovered around 10000 instances of data (excluding operating system files).

The CSV files we have created with free tool available reflect all information that we mentioned in the attribute distribution format (Figure 4).

The occurrence of all different file formats (including deleted) are listed below:

i. The occurrences of all Document files are more than 275.
ii. The occurrences of all PDF files are more than 1200.
iii. The occurrences of all TXT and RTF files are more than 3000.
iv. The occurrences of all other files such as BMP, JPEG, XML, HTML etc are more than 3000.

After categorized all data files, the occurrence of document, pdf and text files are more than any other file format.

The decision tree output of the algorithm C4.5, suggests the usage pattern of the hard drive. It is evident from Table I as the disk used to store a lot of text and pdf files.

Table1 File type distribution across hard drive

File type	%	File Type	%	File Type	%
DOC	7%	RTF	5%	JPEG	5%
PDF	20%	XML	2%	BMP	7%
TXT	25%	HTML	4%	Other	20%

5. CONCLUSION & FUTURE WORK

In conclusion, based upon the experimental results, the distribution of different types of files has been demonstrated. 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 behavior and the area of the person which he/she belongs. Within the experiment, ownership was defined by parsing out user profiles from the windows file system directory structure.

Given a different heuristic, it would be interesting to apply this technique in future to other file systems other than that of Windows and to compare results. This research utilized a basic set of metadata from the files found on each hard drive; for example, file-size, partition, and file format.

6. ACKNOWLEDGMENT

The authors thankful to the Central Forensic Science Laboratory, Chandigarh for generating test data using original Encase digital forensic tool version 6.2 for this research paper.

7. 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] Ankit Agarwal, Megha Gupta, Saurabh Gupta, S.C. Gupta

2011. Systematic Digital Forensic Investigation Model. IJCSS, Volume 5, Issue 1.

[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] Brown, Ross A., Pham, Binh L., & De Vel, Olivier Y. 2003. A Grammar for the Specification of Forensic Image Mining Searches. In Lovell, Brian, Campbell, Duncan, & Fookes, Clinton (Eds.) Eighth Australian and New Zealand Intelligent Information Systems Conference, December, Sydney, Australia.

[5] Brown, Ross A. & Pham, Binh L. 2005. Image Mining and Retrieval Using Hierarchical Support Vector Machines. In Chen, Yi-Ping (Ed.) 11th International Conference on Multi-Media Modeling, Jan, Melbourne, Australia 1550-5502, IEEE.

[6] Carney, M. and Rogers, M. 2004. The Trojan Made Me Do It: A First Step in Statistical Based Computer Forensics Event Reconstruction. International Journal of Digital Evidence, 2(4). 1-11.

[7] Chen, Y., J.R. Miller, J.A. Francis, G.L. Russell, and F. Aires 2003. Observed and modeled relationships among Arctic climate variables. *J.* Geophys. Volume. 108.

[8] 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).

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

[10] De Vel, O., Corney, M. and Mohay, G. 2001. Mining E-Mail Content for Author Identification Forensics, SIGM OD Record, ACM Press, Volume 30, Issue 4, 55–64.

[11] DFRWS. 2001. A road map for digital forensic research. DTR - T001-01 FINAL - DFRWS Technical Report, 1(1).

[12] Fayyad, U., G. Piatetsky-Shapiro and P. Smyth, 1996. Advances in Knowledge Discovery and Data Mining, MIT Press, ISBN-10: 0262560976,560.

[13] F. Pernkopf 2004. Detection of Surface Defects on Raw Steel Blocks Using Bayesian Network Classifiers. Pattern Analysis and Applications, Vol. 7, No. 3, 333–342.

[14] Guidance Software Inc. Encase Forensics. http://www.guidancesoftware.com.

[15] Padhraic Smyth, David Hand, Mannila Heikki 2001. Principles of Data Mining. The MIT Press.

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

[17] Veena H Bhat, Member, IAENG, Prasanth G Rao, Abhilash V. R., P. Deepa Shenoy, Venugopal K. R. and L. M. Patnaik 2010. A Data Mining Approach for Data Generation and Analysis for Digital Forensic Application IACSIT, Vol.2, No.3, ISSN: 1793-8236.