USE OF DATA MINING TECHNIQUES IN DATA ANALYSIS FOR DIGITAL APPLICATIONS

***Ajay, $Rakesh Kumar #Dr. P.K. Jakhar, @Dr. Anuj Kumar**

**, $Research Scholar, CMJ University*

#, @Research Guide

ABSTRACT

With the rapid advancements in information and communication technology in the world, crimes committed are becoming technically intensive. When crimes committed use digital devices, forensic examiners have to adopt practical frameworks and methods to recover data for analysis which can pose as evidence. Data Generation, Data Warehousing and Data Mining, are the three essential features involved in the investigation process. This paper proposes a unique way of generating, storing and analyzing data, retrieved from digital devices which pose as evidence in forensic analysis. A statistical approach is used in validating the reliability of the pre-processed data. This work proposes a practical framework for digital forensics on flash drives.

Keywords— *Digital Forensic, Flash Drive, Framework, Data Preprocessing.*

INTRODUCTION

The digital world has penetrated every aspect of today's generation, both in the space of human life and mind, not even sparing the criminal sphere of the world. According to Jim Christy, Director of Cyber Crime Institute, forensic science is the application of science to legal process and therefore against crime. It relates the use of science and technology, in the process of investigation and establishment of facts or evidence in the court of law [1]. When crime is aided by or involves the use of digital device(s), the investigation is categorized under digital forensic or cyber forensic. If the digital device involved is only a computer or digital storage medium, we refer to the investigation as computer forensic. Computer forensic (aka digital forensic) is a branch of forensic science, whose goal is to explain the current state of the digital artifact [14].

The pool of digital devices used by individuals – for work or for entertainment, on a day-to-day basis, includes cellular phones, laptops, personal digital assistants (PDAs), personal computers, wireless phones, wired landlines, broadband/satellite internet connection modems, iPods etc. Each individual today maintains more than one email account, is a member of many communities, virtual groups, takes active part in chat rooms and other networking sites with his/her identity or under an alias, juggles multiple flash drives and other digital storage media. Departments of the Government and Armed Forces, insurance organizations, telephone industries and banks are a few of the sectors which are eager to track, identify and defend themselves against any digital criminal activities. Digital Forensic Research Workshop (DFRWS) has defined Digital Forensic Science as "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, or helping to anticipate unauthorized actions shown to be disruptive to planned operations" [2]. Digital Forensic Science covers computer forensics, disk forensics, network forensics, firewall forensics, device forensics, database forensics, mobile device forensics, software forensics, live systems forensics etc. Digital Forensic has been described as incident(s) specific and practitioner driven advances which are developed and then applied [3]. 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 flash drive as a device and analyzing its contents. The paper details the preprocessing steps adopted to bring out information of the data stored on the flash drive.

RELATED WORK

The forensic investigation of digital evidence is predominantly employed as a post-incident response to an activity that cannot be defined definitely as legal or to an incident that does not comply to the organizational norms and policies. While the presence of physical forensic investigation model has matured through the years of its presence, refined as revised globally, the involvement of digital evidences have made its presence felt in the recent years. In the year 1995, M Pollitt, suggested a four step process that mapped admission of documentary evidence in the court of law to admission of digital evidence, giving a the process steps included were acquisition, identification, evaluation and admission. In 2001, DFRW came up with a framework which involves identification, preservation, collection, examination, analysis, presentation and decision. This framework is the basis for all the proposed models that followed till date. In 2002, Reith, Carr and Gunsch [5] proposed a model, called an abstract digital forensic model, based on the DFRW model, where the key components of the model involved nine stages. The disadvantages, as quoted by the authors, is that the model is too general for practical use, there is no easy or obvious method for testing the model and that each subcategory added to the model will make it even more cumbersome. In 2006, Kohn, Eloff and Olivier [6] proposed a model, merging the best and essential features of all the models proposed till date. The framework is so designed that any number of additional phases can be easily accommodated. The common model proposed by Freiling and Schwittay in 2007, both for incident response and computer forensic processes, allowed a management oriented approach in digital investigations, while retaining the possibility of a rigorous forensic investigation [7]. [8] in 2008, identifies the five categories of computer forensic research as framework, trustworthiness, computer forensics in networked environment, data detection and recovery and the last category as acquisition. The goal of detection and recovery is stated as to recognize the digital objects that may contain information about the incident and to document them. Our paper focuses on proposing an alternate

Framework for investigation process of physical storage devices, which builds on the models already proposed. It also proposes and chalks out the implementation process for extraction and preprocessing of data extracted from a flash drive. Though there has been significant work in area of extraction and analysis of digital evidence from physical devices such as hard disk, work in the area of simple, portable, storage devices, which are accessible for easy storage like flash drives, cellular phones [9], compact disks and iPods [10] are relatively hard to find. Reference [11] describes methods for digital forensic characterization of physical devices like digital cameras, printers and RF devices. The concept of mapping physical investigating process with the digital investigation process has been discussed in detail [1] [12] [13] [15], which forms the base for our paper, where in the digital device focused on is the flash drive.

FRAMEWORK FOR DIGITAL FORENSIC INVESTIGATION PROCESS OF PHYSICAL STORAGE DEVICES

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. It charts to add value in the specific case of portable storage digital devices. Made up of six stages, it is practical in approach, easy to implement when the digital device involved is any portable, storage device.

Stage 1: Preparation: The main focus is acknowledging the role of digital storage device(s) in the identified or untoward incident. This step recognizes the presence or absence of the digital forensic investigation. All suspected physical storage devices are to be physically secured to prevent tampering. The concerned authorities are to be notified about the presence of possible evidence(s) and the need for examination of the same, and hence permission to access the device. In case the evidence needs to be removed from the premises or site of the activity, steps for obtaining the necessary permissions for the removal are to be identified and executed. On the whole, based on the nature of the incident or crime, the investigation steps are to be chalked out.

Stage 2: Collection and preservation of digital device: The device collection phase opens with the identification of the ownership of the device along with the identification of supposed users of the device. All the digital devices and any other supporting evidences about the usage of these devices that are present at the scene of crime are to be confiscated for data collection. In case the physical device is password protected, the software necessary for accessing the device contents is identified and verifying that it does maintain the integrity of the data as it works on accessing the device. The device contents should be duplicated or imaged maintaining the integrity of the data in the device. Each step of the activity should be documented.

Stage 3: Data extraction and preprocessing: The device/disk that has been imaged or duplicated is to be accessed and examined for the presence of any hidden or encrypted data and system related data. Required software tools are to be used to decrypt or access the data. These tools

should not tamper the original data. Ensure that nothing will/shall be written on to the device that is under scrutiny. Based on the nature of the incident, the investigation is to be categorized as goal based or non-goal based. The data should be extracted from the digital device and the steps for the preprocessing the data is to be outlined, justifying the reason for the same. The software required for the process is to be identified. All through the stages, concern about maintaining the integrity of data should be the key focus and each step is to be validated before executing. Documentation of the activities carried out should be precise and justified as this would act as the base document for justifying the integrity of the presence or absence of evidence leading to the crime.

Stage 4: Data examination and analyses: Before the data is subjected to examination and analyses, the data is to be cross checked for authentication and integrity. The analyses that can be carried out on the extracted data, based on the nature of the data, are to be considered along with the required tools to perform the same. On justifying the analysis methodology, the actual analysis is to be carried out until stable results are achieved. Interpretation of data is the most difficult step, while at the same time the most important step in this flow.

Stage 5: Reporting and documentation: Though this has been cited as the stage 5, it is a continuous process, which needs to be reviewed and updated finally, before presentation in the court of law, for completeness and accuracy. validity and the acceptance of the process or methodology in the scientific community should also be explored. Documentation of the analyses, conclusions and assumptions if any, are also of importance. The limitations of the procedures/analysis carried out are to be outlined clearly.

Stage 6: Presentation in the court of law: The main focus of this step is to prove the presence or absence of digital evidence, from the digital devices collected from the scene of the incident under examination, in the court of law. While computer forensics is highly technology specific, people handling law in the court of justice are not technology specialists. Hence it is very important for technology specialists to understand the ramifications of the legal world and at the same time, communicate effectively and clearly the complete digital investigation process, emphasizing on the analysis of the findings. The documentation of the entire process may also be submitted in the court of law to cross-examine the steps adopted during the investigation process. While this may suffice the needs of the court to arrive at a decision, it may sometimes be required to complete further analysis or redo a phase, as required by the court, to support any issues. In the United States of America, a pre-trial "Daubert Hearing', conducted in the presence of a judge, to verify the underlying methodology and techniques used in the identification of the evidence and hence authenticating validity of the evidence, is mandatory. The validity of the procedures used, the error rate of the procedures, the cross check of the process by peer reviews and the scientific community, is scientifically and systematically checked. In India, a similar process is to be framed and adopted involving the legal specialists and technology specialists. To validate this framework, it needs to be tested out in its entirety in the real world. The data extraction and preprocessing stage has been tested out for effectiveness, as outlined in the following sections of the paper. The digital device selected for this phase of the framework is the flash drive, which is identified as the most frequently used portable storage device of the generation today.

DATA EXTRACTION AND PREPROCESSING FROM THE FLASH DRIVE

The classic Extract-Transform-Load steps are applied right from the identification of the data to loading the data for the final analysis. The main highlight of our preprocessing step is that it does not depend on expensive, specific proprietary software for extraction as well as the transformation of the accessed data, instead uses software that are either freeware or versions available for free download for personal usage or those that are readily available on any personal computers or laptops. The economics of a digital forensic investigation is very necessary today along with the time frame required to complete the same. Hence, taking this into consideration, the path taken by our research activity, the usage of free source and existing operating system commands, makes it one of the most economical ways of preprocessing data. Fig. 1 shows the architecture.

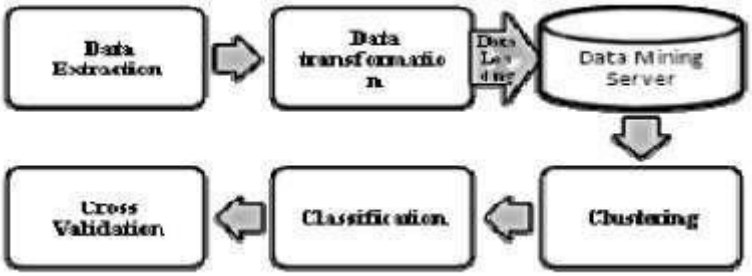

Figure 1. Architecture of the model.

A. Extraction

The digital device focused for investigation is the flash drive. The device is checked for password protection. In case of protected device, the identification of various software required to break the protection code is to be identified. The drive is also checked for encrypted data. The path adopted is non-goal based as at this stage we are not sure of the role of the device in the alleged activity. The data is extracted from the original device, taking care that there is no process that writes on to the digital device under investigation. *Recuva(version 1.23.389)*, a freeware that recovers lost data so long it has not been overwritten by the system, has been used to list the entire contents of the digital device, the flash drive. We recovered the files on to the hard disk of the computer so that the integrity of data on the evidential flash drive are preserved. The time frame for the actual data recovery depends on the duration and frequency of usage of the flash drive.

DOS*:* Disk Operating System has been used extensively to gather the data details and software associations for preprocessing. DOS maintains the time stamps of the recovered file, thus ensuring integrity of data.

B. Transformation

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 transformation process at the database level too. The major data transformations are conversion of the data into any standard format (comma separated format used here), generation of the parent directories and extraction of the file extensions. The *Oracle Express Edition(10g)*, has been used here as a data warehouse. The data extracted and partially preprocessed was loaded into Oracle, using the SQL loader. Further data preprocessing, transformations and basic computations were performed in this environment to get the complete dataset.

C. Loading

Before the dataset is actually loaded for data mining analysis, the data is validated statistically. The statistical tests Bartlett's test of sphericity and Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy, are conducted for exploratory factor analysis. In KMO measure of sampling adequacy, if two attributes, share a common factor with other variables, their partial correlation (aij) will be small, indicating the unique variance they share, and is given by

$$a_{ij} = r_{ij} \qquad (1)$$

where r_{ij} is the Pearson correlation between items i, j where I, $j = 1,2,3 \ldots k$.

KMO is calculated as follows:

$$KMO = \left(\sum_{i=1}^{k} \sum_{j=1}^{k} r_{ij}^2\right) \Big/ \left(\sum_{i=1}^{k} \sum_{j=1}^{k} r_{ij}^2 + \sum_{i=1}^{k} \sum_{j=1}^{k} a_{ij}^2\right) \qquad (2)$$

for all i,j where k is the number of components. If aij ≈ 0.0, then the variables measure a common factor and KMO ≈ 1.0. If aij ≈ 1.0, then the variables do not measure a common factor and KMO ≈ 0.0. Using Bartlett's test of sphericity, we calculate the determinate of the matrix of the sums of products and cross-products (S) from which the inter-correlation matrix is derived. The determinant of the matrix S is converted to a Chi-square statistic and tested for significance. The null hypothesis is that the inter-correlation matrix comes from a population in which the variables are non-collinear (i.e. an identity matrix), and the non-zero correlations in the sample matrix are due to sampling error. Chi-square is calculated as

$$\psi^2 = -\left[(n-1) - {}^1\!/_6\,(2p+1+{}^2\!/_p)\right]\left[\ln|S| + \rho \ln({}^1\!/_p)\sum I_j\right] \quad (3)$$

where n = number of instances, p = number of variables, I_j = j^{th} Eigen value of S.

The degrees of freedom (df) is calculated as

$$df = {}^{[(p-1)(p-2)]}\!/_2 \quad (4)$$

The complete dataset, consisting of the file tree, the file attributes, the timestamps, file size and the deleted flag, is loaded into Weka, an open source software, for analyses. On an average, each flash drive, used in this study, recovered around 4000 instances of data.

DATA MINING SERVER

Data mining server is essential to a data mining system and ideally consists of a set of functional modules for tasks such as characterization, association, cluster analysis, classification, evolution and deviation analysis. We have used a two step process to analyze the dataset. The first step is to run unsupervised clustering algorithm. We then used the classification algorithm to verify the visualization pattern of the data instances once loaded into Weka.

A. Clustering

Data Clustering, builds unsupervised data models from the data. Data instances are grouped together, based on similarity schemes, defined by the clustering system, in large, multi-dimensional data set. As clustering attempts to group data instances into clusters of significant interest, evaluate the performance of the model and detect outliers. We have selected clustering as a step in the analysis of the data generated as part of the digital forensic exploration. Our interest is to examine those data instances that do not group naturally into cluster groups, for forensic evidence. We have used simple k-means algorithm for the basic clustering. Simple k-means algorithm takes *k*, the number of clusters to be determined, as an input parameter and partitions the given set of *n* objects into *k* clusters so that the resulting intra-cluster similarity is high while the inter-cluster similarity is low. Euclidean distance measure is used to assign instances to clusters. Cluster similarity is measured as the mean value of the objects in a cluster.

B. Classification

The data instances are fed as input to the classifier. We have selected C4.5 decision tree model, with binary split to visualize the patterns found in the dataset. The C4.5 produces trees with variable branches. The standard way of predicting the error rate of a classifier given a single, fixed stratified data is to use 10-fold cross-validations. When a discrete variable is chosen as the

splitting attribute, one branch is generated for each value of the attribute. C4.5 uses gain-ratio as criterion for splitting, which ensures the largest information gain.

C. Cross Validation

The standard way of predicting the error rate of a classifier given a single, fixed stratified data is to use 10-fold cross-validation. Cross validation technique is adopted, in cases when the amount of data for training and testing is limited.

PERFORMANCE ANALYSIS

The initial statistical analysis for KMO measure of sampling adequacy, and Bartlett's tests of sphericity is summarized in the Table I. The results suggest that the attributes have minimum correlations among themselves. Hence factor analysis and reduction need not be performed on the dataset.

TABLE I. KMO AND BARTLETT'S TEST OF SPHERICITY

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		**0.514**
Bartlett's Test of Sphericity	*Approx. Chi-square*	**6923.777**
	Degrees of Freedom	**21**
	Significance	**0.000**

The decision tree output of the algorithm C4.5, suggests the usage pattern of the flash drive. It is evident from Table III and Table IV that flash drives FD1, FD2 and FD5 are well organized, as the decision tree splits on directories are many. FD1, FD3, FD4 and FD5 are used to store a lot of multimedia files. The generated decision trees for the flash drives under study are shown in Table IV and the file types distribution is tabulated in Table III. Another interesting pattern seen is that files are deleted to accommodate the newer ones, without following priorities in selecting files for deletion, although the preference for retaining folders, is strongly seen here. We have based our clustering on the assumption that there are ideally two distinct clusters, one constituting of the deleted files and the other the not-deleted ones.

TABLE II. RESULTS OF THE CROSS VALIDATION PERFORMED

Flash Drive No.	No. of users	Classifier accuracy (%)	Kappa statistics	Root mean squared error
FD 1	Single	99.6757	0.9876	0.0553
FD 2	Multiple	95.1724	0.9034	0.1887
FD 3	Single	98.4450	0.9397	0.1134
FD 4	Multiple	97.9320	0.9345	0.1444
FD 5	Single	98.8417	0.9764	0.0993

This is validated by the visualization of the clustering output, tabulated in Table IV. The wrongly clustered files are considered to be the files of interest. As a measure for pattern accuracy a 10-fold cross validation was performed and the results for the flash drives considered are tabulated, in Table II.

LIMITATIONS OF THE STUDY

The flash drives accessed for this study were mainly from the students of our university however this needs to be broad based across users in future extensions of this study. The Recuva software cannot restore files if the Windows operating system has overwritten the area where the file used to reside. The files that have been securely deleted, using special tools, cannot be recovered by Recuva. The last accessed date is the date the files are recovered from the disk, hence cannot be considered for forensic analysis. We also need to identify or develop suitable software for analyzing data stored in Devanagari and other Indian / non-English scripts. Since these are widely used by the Indian business community, the current internationally available tools may not address the same.

TABLE III. FILE TYPE DISTRIBUTIONS ACROSS FLASH DRIVES.

Flash Drive	File Type	%	File Type	%	File Type	%
FD1	JPG	45%	ASM	7%	EXE	6%
FD2	PDF	31%	GIF	23%	DIR	9%
FD3	JPG	72%	DIR	5%	VBS	6%
FD4	PNG	83%	GIF	5%	DIR	5%
FD5	JPG	68%	CSV	10%	DIR	7%

This paper presents an overall framework that covers the digital forensic analysis process for the diverse range of portable storage devices. The framework is practical and easily adaptable. The tools used here, are open source or those already existing in the present day computing environment, hence easily accessible to the investigation team. It also forms a bridge between the digital forensic investigation team and judicial bodies. It can constitute a guideline for forensic teams in police and other investigation agencies in our country, who do not necessarily work with a common defined process today.

FUTURE WORK

The future work areas identified with respect to the data that has been successfully extracted and preprocessed from the flash drive can be - identification of the files using not so common software(s), identification of the presence of illegal data storage, identification of the hidden and encrypted data, identification of the files with renamed file extensions and identification or development of software's which work seamlessly with non-English scripts and data. Prediction

of the usage pattern of the owner of the flash drive and also a time series analysis for predicting the file type usage along with user profiling including the subjects or topics the user is interested in would be an interesting area to explore. Finally, discussion of the framework with a group of potential users in police / investigation agencies to understand specific areas of development required and also to fine-tune the framework can be conducted to validate the entire process.

REFERENCES

1. Robert Rowlingson, "A Ten Step Approach for Forensic Readiness," International Journal of Digital Evidence, vol. 2, issue 3, 2004.

2. Gary Palmer, "A Road Map for Digital Forensic Research," DFRWS Technical Report, Available:http://www.dfrws.org/2001/dfrwsrmfinal. pdf, 2001.

3. Kara Nance, Brian Hay and Matt Bishop, "Digital Forensics: Defining a Research Agenda," Proceedings of the Forty Second Hawaii International Conference on System Sciences, pp. 1-6, 2009.

4. M. Pollitt, "Computer Forensics: An Approach to Evidence in Cyberspace",Proceedings of the National Information Systems Security Conference, Baltimore, pp. 487-491, 1995.

5. M. Reith, C. Carr and G. Gunsch, "An Examination of Digital Forensic Models," International Journal Digital Evidence, vol. 1, no. 3, 2002.

6. M. Kohn, J. Eloff, and M. Oliver, "Framework for a Digital Forensic Investigation," Proceedings of Information Security South Africa from Insight to Foresight Conference, South Afrika, 2006.

7. F. C. Freiling, and B. Schwittay, "A Common Process Model for Incident Response and Computer Forensics," Proceedings of Conference on IT Incident Management and IT Forensics, Germany, 2007.

8. Mohd Taufik Abdullah, Ramlan Mahmod, Abdul A. A. Ghani, Mohd A Zain and Abu Bakar Md S, "Advances in Computer Forensics," International Journal Of Computer Science and Network Security, vol. 8, no. 2, February 2008.

9. Wayne Jansen and Rick Ayers, "Forensic Software Tools for Cell Phone Subscriber Identity Modules," Conference on Digital Forensics, Security and Law, 2006.

10. Christopher V. Marsico and Marcus K. Rogers, "iPod Forensics," International Journal Of Digital Evidence, vol. 4, issue 2, Fall 2005.

11. Nitin Khanna, K. Aravind, Mikkilineni, Antony F. Martone, Gazi N. Ali, et al, "A Survey ofForensic Characterization Methods for Physical Devices," Digital Forensic Research Workshop, 2006.

12. Brian Carrier and Eugene H. Spafford, "Getting Physical with Digital Investigation Process," International Journal of Digital Evidence, vol. 3, issue 2, Fall 2003.

13. Siti Rahayu Selamat, Robiah Yusof and Shahrin Sahib, "Mapping Process of Digital Forensic Investigation Framework," International Journal of Computer Science and Network Security, vol. 8, no. 10, October 2008.

14. B. D. Carrier, "A Hypothesis-Based Approach to Digital Forensic Investigations," CERIAS Tech Report 2005-06, Purdue University, Center for Education and Research in Information Assurance and Security, West Lafayette, 2006.

15. Mark Rogers, J. Goldman, R. Mislan, T. Wedge, and S. Debrota, "Computer Forensics FieldTriage Process Model," Proc. Of Conference on Digital Forensics, Security and Law, pp. 27-40, 2006.