DATA MINING APPROACH IN DIGITAL FORENSICS

***Rajiv Sharma**
***Deptt. of Computer Applications, A.S. College, Khanna.**

Abstract: Data mining is a part of the interdisciplinary field of knowledge discovery in databases. Research on data mining started in the 1980s and grew rapidly in 1990s.Specific techniques that have been developed within disciplines such as artificial intelligence, machine learning have been successfully employed in data mining. Data mining has been successfully introduced in many fields. An important application area for data mining techniques is the World Wide Web Recently, data mining techniques have also being applied to the field of criminal forensics nothing but Digital forensics. Examples include detecting deceptive criminal identities, identifying groups of criminals who are engaging in various illegal activities and many more. Data mining techniques typically aim to produce insight from large volumes of data.

Digital forensics is a sophisticated and cutting edge area of breakthrough research. Canvass of digital forensic investigation and application is growing at a rapid rate with mammoth digitization of an information economy. Law enforcement and military organizations have heavy reliance on digital forensic today. As information age is revolutionizing at a speed inconceivable and information being stored in digital form, the need for accurate intellectual interception, timely retrieval, and nearly zero fault processing of digital data is crux of the issue. This research paper will focus on role of data mining techniques for digital forensics. It also identifies how Data mining techniques can be applicable in the field of digital forensics that will enable forensic investigator to reach the first step in effective prosecution, namely charge-sheeting of digital crime cases.

Keywords: Data Mining techniques, Digital forensics, digital Investigation, Data Recovery

A. INTRODUCTION

Digital forensics is the use of scientific methods for the identification, preservation, extraction and documentation of digital evidence derived from digital sources to enable successful prosecution [7].

Digital forensics, in real meaning, answers the when, what, who, where, how and why concerning a digital crime [29]. When conducting an investigation on a computer system, for example, the 'when' refers to the time interval the activities took place during. The 'what' concerns the activities performed on the computer system. The 'who' concerns the person responsible, the 'where' refers to where the evidence is located, the 'how' addresses the manner in which the activities were performed, and the 'why' seeks to ascertain the motives behind the crime.

Digital forensics is a technology which is exploited to conduct investigations into digital crimes or incidents. The aim of such investigations is to expose and present the truth, which often leads to prosecution and conviction. Dramatic rise in the numbers of digital crimes committed have led to the development of a whole slew of computer forensic tools. These tools ensure that digital evidence is acquired and preserved properly and that accuracy of results regarding the processing of digital evidence is maintained [8]. Such tools exist in the form of computer software and have been developed to assist digital investigators conduct a digital investigation.

The useful seamless integration of data mining techniques with digital forensic science has been depicted at analysis phase. This will help in boosting up the performance and the reliability of investigations of the subjects.

The formal methodology of data mining includes following basic steps [2]:

- Determine the nature and structure of the representation of the data sets.
- 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.

Data mining functionalities are used to specify the various types of patterns to be looked for. The first part of this paper will study and review the current know-how in the field of digital forensics. Subsequently, common thread of conventional forensic tools will be investigated. The results of this analysis will be used to create a basic classification of computer forensic tools. This classification shall serve as a foundation for the identification of inherent limitations of current computer forensic tools and recommendations for the breakthrough improvements through Data mining Techniques for ushering in state of the art digital forensic tools will be suggested.

B. EXISTING STATUS OF DIGITAL FORENSICS TOOLS

In the Literature, Computer Forensics Tools are basically classified as

- Hardware forensic tools
- Software forensic tools

Hardware forensic tools can be used for Single-purpose components or complete computer systems and servers. Software forensic tools are used for Command-line applications and GUI applications. The development of a variety of specialist commercial and freeware tools began in 1980s and 90s.These can generally be broken down into three categories as follows.

General forensic tools: Tools allowing a wide variety of investigation, particularly keyword searching, on digital media.

Specialist forensic tools: Which focus on a specific piece of forensic material for investigation perhaps images, or internet artefacts. Often relying on output from one of the general tools.

Case Management tools: These are used to track, audit and report on cases.

The unique need of computer forensics have resulted in the creation of computer forensic tools in the form of computer software. These tools ensure that digital evidence is acquired and preserved properly to maintain the integrity of digital evidence. For example, copying and pasting data onto another storage medium may not be admitted in a court of law as forensically sound evidence [31]. This is because the process of copying and pasting data can modify it, for example, altering the timestamps of the data. As a result, a typical digital investigation requires the making of an exact bit by bit (or bit stream) copy of all the data on a storage medium. This exact bit by bit copy is called an image and the process of making an image is frequently referred to as imaging [30].

Computer forensic tools focus primarily on digital evidence recovery, in other words, on recovering residual data from a piece of media. These tools usually have limited abilities to assist in the analysis of the recovered data. The presentation of data offered by computer forensic tools is deceptive at times. The reason is that the dimensionality, complexity and volume of data still exist because the computer forensic tools merely present it to investigators. The digital investigators still have to examine the presented data and draw conclusions.

At present, computer forensic tools are not ideal for the following tasks:

- Association: identifying correlations among data.
- Classification: discovering and sorting data into groups based on similarities of data.
- Clustering: finding and visually presenting groups of facts previously unknown or left unnoticed;
- Forecasting: discovering patterns and data that may lead to reasonable predictions.

C. IMPORTANT FORENSICS TECHNIQUES

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. [12]

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. The uniqueness of the hash depends on the cryptographic function used. MD5 hashing was developed in 1991 by Ron Rivest and was rapidly adopted by the forensics community. NIST soon decided upon SHA-1 as the federal standard[27].

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 [23]. Carving can recover deleted but not overwritten files as well as temporarily cached files on media. An analysis of carving techniques was performed by Mikus in 2005 [25]. Recent advances in carving allowing fragmented files to be recovered with more accuracy. Garfunkel demonstrated file carving with object validation [18].

D. ROLE OF DATA MINING IN DIGITAL FORENSIC

Data mining & soft Computing has several applications in digital forensics. These include 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)[10].While this technique is ideal for association, classification, clustering and forecasting, it is also particularly useful for visualization. [11]

Visualization enables digital investigators to locate vital information that is of interest rapidly and efficiently. In addition, it can guides digital investigators towards the best next step in their search so that digital evidence recovery is carried out in a more efficient and effective manner.[12]

In 2003, the Artificial Intelligence Lab at the University of Arizona, presented an overview of case studies done with relation to their COPLINK project. The project's specific interest was how information overload hindered the effective analysis of criminal and terrorist activities by Law enforcement and national security personnel.

Their work proposed the use of data mining to aid in solving these issues. In their report they define data mining in the context of crime and intelligence analysis to include entity extraction, clustering techniques, deviation detection, classification, and lastly string comparators.

Four case studies in the report showed how data mining was useful in extracting entity information from police narrative reports, detecting criminal identity deceptions, authorship analysis in cyber crime, and lastly criminal network analysis. Today, COPLINK is software that has been successfully deployed in the field, and works by consolidating, sharing, and identifying the information from online databases and criminal records [6].Work done by Hewlett Packard in 2005 applied data mining to solve their problem of finding similar files in large document repositories [15].

The end analysis yielded clusters of related files and was further enhanced by applying a graph bipartite

partitioning algorithm [12].

In 2006, Galloway and Simoff experimented with a case study redefining an approach to network data mining. In their work, they defined network data mining as identifying emergent Networks between large sets of individual data items. [13].

Shatz, Mohay, and Clark in 2006 explored a correlation method for establishing provenance of timestamped data for use as digital evidence. This work has a deep and relevant impact on digital forensics research as it reiterated the complexity issues of dealing with timestamps because of clockskew, drift, offsets, and possible human tampering. [4].

In 2006 as well, research done by Abraham explored event data mining to develop proles for computer forensic investigation purposes. Abraham analyzed computer data in search of discovering owner or usage profiles based upon sequences of events which may occur on a system. Abraham categorized an owner profile with four different attributes: subject, object, action, and time stamp [1].

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 [26].

E. EXISTING TOOLS AND DATA MINING TECHNIQUES FOR DIGITAL FORENSIC

The following table summarizes the existing tools and techniques used to solve some of the issues of digital forensics.

Table1: Analysis of existing tools & techniques

Digital Forensic Techniques	**Data Mining Techniques**	**Tool**
Data Recovery, data generation and preprocessing	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	LingPipe
	Visualization	CyberForensicTimeLab
	Statistical Analysis and Anamoly Detection	EMT/MET
	Recursive data mining	-
	Phishing	Invisible Witness
	Regression	-

F. CONCLUSION

With the increasing number of computer forensic tools available on the market, it is important to be aware of the different features that exist within the domain. The aim of this classification was to provide an overview of the current capabilities of computer forensic tools. It is also taken as the baseline from which limitations and recommendations were identified. These tools usually have limited abilities to assist in the analysis of the recovered data. That is why currently data mining techniques are being used to develop various tools. Thus this paper focuses on various issues of digital forensics that can be solved using existing data mining techniques. Also this paper summarizes the existing tools and techniques for digital forensics.

REFERENCES

[1] DFRWS," A road map for digital forensic research", DTR - T001-01 FINAL - DFRWS Technical Report, 1(1), August 2001. http://dfrws.org/2001/dfrws-rm-final. PDF.
[2] Padhraic Smyth David Hand, Heikki Mannila.(2001)," Principles of Data Mining", The MIT Press.
[3] Chidanand Apt´e and Sholom Weiss (1997), "Data mining with decision trees and decision rules",Future Generation Computer System., 13(2-3):197–210,. ISSN 0167-739X.
[4] Andrew Clark Bradley Schatz, George Mohay (2006)," A correlation method for establishing Provenance of timestamps in digital evidence", 6th Annual Digital Forensic Research Workshop, In Digital Investigation, volume 3, supplement 1, pages 98–107.
[5] Brian Carrier. The sleuth kit (tsk). Retrieved 2008-03-10 11:20:22 -0700. http://www.sleuthkit.org/sleuthkit/desc.php.
[6] Hsinchun Chen, Wingyan Chung, Yi Qin, Michael Chau, Jennifer Jie Xu, Gang Wang, Rong Zheng, and Homa Atabakhsh (2003)," Crime data mining: an overview and case studies", Proceedings of the 2003 annual national conference on Digital government research, pages 1–5. Digital Government Research Center.
[7] Kruse II, W.G. and Heiser, J.G. 2002. Computer forensics: incident response essentials.Addison-Wesley.
[8] Marcella, A.J. and Greenfield, R.S. 2002. Cyber forensics: a field manual for collecting, examining and preserving evidence of computer crimes. Auerbach.
[9] O. de Vel, A. Anderson, M. Corney, and G. Mohay (2001),"Mining e-mail content for author identification forensics", SIGMOD Rec., 30(4):55–64, ISSN 0163-5808.
[10] Tamas Abraham,"Event sequence mining to develop profiles for computer forensic investigation purposes" (2006), In ACSW Frontiers '06: Proceedings of the 2006 Australasian workshops on Grid computing and e-research, pages 145–153. Australian Computer Society, Inc., Darlinghurst, Australia, Australia, ISBN 1-920-68236-8.
[11] Usama M. Fayyad, Gregory Piatetsky-Shapiro,Ramasamy Uthurusamy. (2003),"Summary from the kdd-03 panel: data mining: the next 10 years. SIGKDD Explor. Newsl., 5(2): 191–196,. ISSN 1931-0145.
[12] George Forman, Kave Eshghi, and Stephane Chiocchetti,(2005)," Finding similar files in large document repositories.", In KDD '05: Proceeding of the eleventh ACM SIGKDD international Conference on Knowledge discovery in data mining, pages 394–400, ACM, New York, NY, USA, ISBN 1-59593-135-X.
[13] John Galloway, Simeon J. Simoff ,"Network data mining: methods and techniques for discovering deep linkage between attributes", In APCCM '06: Proceedings of the 3rd Asia- Pacific

conference on Conceptual modelling, pages 21–32. Australian Computer Society, Inc., Darlinghurst, Australia, Australia, 2006. ISBN 1-920-68235-X.

[14]Simson L. Garfinkel. Fiwalk program.

[15]Simson L. Garfinkel(2006),"Forensic feature extraction and cross-drive analysis", Digital Investigation, 3(Supplement-1):71–81, http://dx.doi.org/10.1016/j.diin.2006.06.007.

[16]Simson L. Garfinkel, Basis Technology (2008) Aff ,"The advanced forensic format", Retrieved -03-03 10:50:49 -0700. http://www.afflib.org/.

[17]III Golden G. Richard and Vassil Roussev (2006), "Next-generation digital forensics", Commun. ACM, 49(2):76–80, ISSN 0001-0782.

[18]Robert C. Holte(1993),"Very simple classification rules perform well on most commonly used Datasets ", Machine Learning, 11:63–90.

[19]Eibe Frank Ian H. Witten(2005)," Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufman Publishers,.

[20]Knoppix. Shred tool. Retrieved 2007-10-09 10:50:49 -0700. http://www.knopper.net/knoppixmirrors/.

[21]Jan H. Kroeze, Machdel C. Matthee, and Theo J. D. Bothma (2003)," Differentiating data- and Textmining terminology",In SAICSIT '03: Proceedings of the annual research conference of the South African institute of computer scientists and information technologists on Enablement through technology, pages 93–101. South African Institute for Computer Scientists and Information Technologists, Republic of South Africa, 2003.ISBN 1-58113-774-5.

[22]M. Last (2006)," The uncertainty principle of cross-validation. Granular Computing", 2006 IEEE International Conference on, pages 275–280.

[23]Heidi Computers Ltd. Eraser tool. Retrieved 2007-09-10 10:50:49 -0700. http:// www.heidi.ie/node/6.

[24]Piriform Ltd. Ccleaner tool, 2005-2008. Retrieved 2007-09-10 10:50:49 -0700. http://www.ccleaner.com.

[25]Tom M. Mitchell. Instance Based Learning. McGraw Hill, 1997.

[26]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, pages 49–54.