Webology (ISSN: 1735-188X)
Volume 18, Number 6, 2021

Data Mining Approaches In Digital Forensics

Mohammed Ashfaq Hussain , Ahmed Unnisa Begum , Subuhi Kashif Ansari , Durdana Taranum Khan , Shazia Ali, Manju Sharma

Lecturer, Jazan University, Kingdom of Saudi Arabia.

Abstract:

Data mining is the process of extracting information from massive data sets such as databases and data marts. This could include "statistics, machine learning, data management and databases, pattern recognition, artificial intelligence, and other fields.". Computer forensics using data mining for investigation of crimes. This also is helpful for effectively finding the evidence.

Digital forensics is focusing on gathering evidence for potential computer crimes.
It can also be used to collect the activity from authorized and unauthorized users. Digital forensics is used to help law informants with fraud investigations and with theft to protect knowledgeable property. For both businesses and law enforcement organisations, digital forensics is a crucial discipline. The major focus of this research is on the various data mining methodologies used in digital forensics.

Keywords: Data mining, Cybercrime, Digital forensics

Introduction:

Data mining is defined as a technique that converts a vast quantity of data or a large database into critical information that can be used to establish a pattern, identify fraud, spam email, and anticipate whether you are earning money in marketing or whether you are losing money in marketing.

Usually, the data analysts or sometimes even forensic accountants in general that are in charge of a company that collects data and then with that data that they collect they interpret it to then put it into like a software application that could be beclouded, Linux, Microsoft, excel whichever software you want to use and then all that information will be put into an easy-to-read sort of like eligible graph so other people can read it and understand what the information on that graph says.

The numerous patterns to be looked for are identified using data mining concepts. We will investigate and discuss how data mining is used in the realm of digital forensics in this article. As a result, classic forensic tools with a common thread will be studied. The findings of this evaluation will be utilised to develop a fundamental classification of computer forensic tools. This classification will serve as a framework for identifying basic limits of current computer forensic tools and making recommendations for data mining-based enhancements. Techniques for helping in use of trying to cut digital forensics will be presented.

The analysis phase described the useful continuous integration of datasets with digital forensic science. This will aid in the improvement of study is to evaluate performance and consistency.

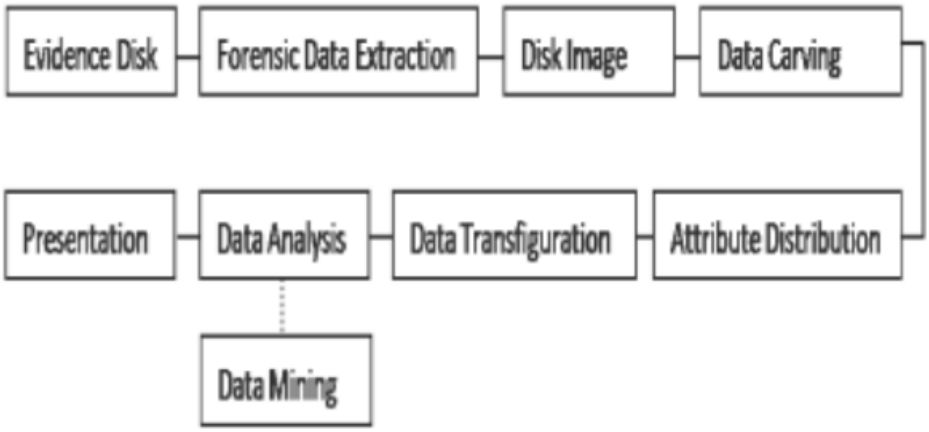
The formal methodology of data mining includes following basic steps [2]:

• 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.

The formal data mining process consists of the phases listed below [5]:

• Control the type and structure of data set representation.
• Determine how to quantify the data; assess how well different representations suit the data.
• Specify an analytical approach to maximise the scoring function; and
• Determine which data management principles are necessary to efficiently apply the algorithms.

A framework(i) is required for a unified statement between the technical resources of the members of the digital forensic investigative committee and the non-technical representatives of the legal team. Defining a generic model for digital forensic inquiry, define a problem every now and again, taking into consideration the many devices accessible today. This framework is reasonable in its outline and technical in its method, but it must be modified to meet all legal criteria in the nation where the incident happened. We propose a resourceful approach for both economic and temporal considerations.



(i) Model Architecture

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.

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.

I. Data mining and Digital Forensics:

Data mining is the act of utilising a computer programme or an algorithm to uncover patterns or correlations in data. For example, we may wish to seek for what combinations of symptoms would allow us to correctly diagnose a patient [1]. Another aspect of data mining is if we are running a business and want to know what sorts of things our consumers like to buy together [2].

1) Association rule mining defines frequently occurring item sets in a database and assists some patterns as policies used in connectivity delay recognition in developing link rules from users' contact history. This strategy can also be applied to network intruder reports to aid in the detection of potential future network issues. This approach may find attack designs amongst time-stamped data in network intrusion detection. Detecting previously undiscovered patterns aids in criminal probe, but obtaining substantial findings necessitates a huge portion of well-defined data.
2) Clustering is a technique that groups data items into classes based on comparable attributes in order to reduce or maximise similarity, for example, to categorise suspects who have

committed infractions in similar ways or to discriminate between units belonging into various groups. There have been no specified classes for allocating things in these procedures.

3) Classification identifies common traits shared by many Crime units and groups them into predetermined categories that were used to detect the origins of email spamming based on the sender's underlying attributes and semantic patterns. Classification, which is frequently used to foresee crime patterns, can reduce the time required to identify criminal entities. The technique, however, includes a specified classification scheme [4].

4) Variation detection employs certain procedures to investigate data that differs significantly from the rest of the data. This technology, also known as outlier detection, may be used by detectives to discover fraud, network intrusions, and other crimes. However, such actions might appear to be typical at times, making it difficult to recognise outliers.

5) String comparator procedures that show the relation between the documented areas in pairs of database files and analyse the correspondence among the records that can detect fraudulent data in criminal documents for example the name and address. The investigators can develop string comparators to calculate textual data that often need severe computation.

Digital Forensic Techniques	**Data Mining Techniques**	**Tool**
Data Analysis	Clustering – K-means, Hierarchical Clustering, EM	Weka
	Unsupervised learning – PCA, Karnohuen Map	-
	Types - Baiyesian, SVM, Naïve, Decision tree, Neural networks & Supervised learning.	Weka
	Recognition / Named Entity	Ling Pipe
	Regression	-
	Phishing	Invisible Witness
	Anamoly Detection and statistical analysis	EMT/MET
Data Recovery, data generation and pre processing	Analysing Statistical Tests Bartlett's sphericity test Kaiser-Meyer-Olkin (KMO)	Recuva Autopsy Rediscover /FTK Encase Sleuth kit

(ii) Analysis of Digital forensics tools & techniques

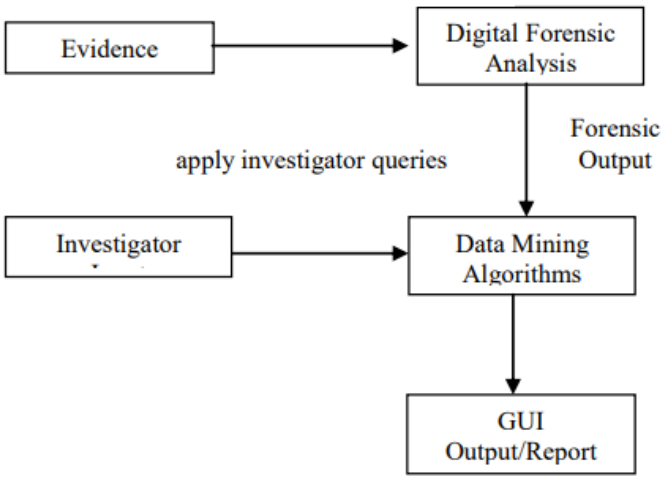
At the heart of the data mining process is some sort of machine learning, in which we apply an algorithm and it learns something about data; these are known as machine learning algorithms. Digital evidence will deal with data that is saved and accessible over the internet, as well as how and where data is kept on our devices, phones, and in the cloud, among many other things.

Manage the data and find the things that we are looking for easier so imagine we have a computer at home we might have files from the last 10 years if we want to find something from 10 years ago how do we find it efficiently and quickly so whenever we think about digital forensics it gives us the better idea about where data is located how you can find that data very quickly and efficiently where other people who are not used to manage the data it takes them a much longer time and all of this, of course, translates into our everyday lives we want to find cacao talk message very quickly that contains an address well.

II. Purpose of Digital Forensics:

Data mining has a wide range of applications for digital forensics. These include detecting and arranging forensic data into groupings based on similarity called classification, locating groups of hidden facts called clustering, and determining trends in data that may lead to useful predictions called forecasting [5]. While this method is excellent for association, classification, clusters, and forecasting, it is primarily beneficial for visualisation. [6]

Digital forensics is a technique that is used to conduct investigations into digital crimes or occurrences. Data mining can be more robust as technology advances, with faster computers and better systems to manage the vast volumes of data required for data mining [4]. With the usage of data mining, the promise of digital forensics is certain to grow.



(iii) Block Diagram of Applying digital forensics methods

We usually use criminal data mining to sort of pinpoint where cybercrime is happening and then from that, we investigate so that sort of crime could be anywhere from stealing sensitive information for exploitation or personal gain cybercriminals usually use it for financial gain or profit so they will usually sell someone's information and get money for it which is illegal also it's a lot easier for an online criminal to hijack information rather than rob someone in person at the store.

III. Types of Digital Forensics:

Digital forensic tools are generally focused on digital proof recovery, which is the recovery of balance data from a portion of media. These technologies often have limited capabilities to assist in the examination of better data. Data management given by computer forensic technologies might be deceptive at times [7]. The reason for this is because the dimensionality, complexity, and volume of data are still present since computer forensic technologies just show it to researchers. The only requirement for computer forensics has led in the development of computer forensic tools inside the form of software. These technologies ensure that digital data is gathered and kept properly to sustain the reliability of digital evidence.

1. Disk forensics is the process of retrieving data from storage media by looking for active, updated, or deleted files.

2. Network forensics — This is a sub-branch of digital forensics that is concerned with the monitoring and analysis of computer network traffic in order to obtain critical information and legal evidence.

3. Wireless forensics - A subset of network forensics, the primary goal of wireless forensics is to provide the tools required to gather and analyse data from wireless network traffic.

4. Database forensics - This is a subset of digital forensics that deals with the study and evaluation of databases and their associated metadata.

5. Malware forensics - This branch is concerned with the detection of harmful code in order to study its payload viruses, worms, and so on.

6. Email forensics is concerned with the retrieval and investigation of emails, including deleted emails, calendars, and contacts.

7. Memory forensics - It is concerned with gathering data in raw form from system memory system registers cache ram and afterwards carving the data from raw dump mobile phone forensics. It mostly deals with the investigation and analysis of portable devices; it aids in the retrieval of phone and sim contacts, call logs, arriving and leaving SMS, audio and video, and so on.

IV. How data mining involves in digital forensics:

In digital forensics, data mining performed web scraping allows us to extract data from that site that offered people jobs through digital forensics learn how to extract their LinkedIn accounts like all of their all of a lot of people's information instead of just like one person and then doing it all at once so we can gather a lot of people's LinkedIn accounts and their vital information all into one Microsoft spreadsheet in excel within five minutes so that’s a plus of using web scraping sort of like data mining[8].

Data mining occurs when we use our Linux and Windows panel commands in our labs and then extract them or extract that data from the pcs so that when we use those commands we will be able to identify the position, time, and type of pc so this is wonderful for digital forensics

and data analysts in forensics since it enables them to sort of individuality what law breaker is doing what and who that person is that does own that computer that's hijacked whether it's hijacking or stealing information or theft So, using data mining, we may uncover a pattern and analyse what is inconsistent in that pattern. If we see anything out of the ordinary in the pattern, we know something is wrong but we need to investigate [9][11].

Data mining in crime is classified as follows:

1) Variables are stored in the system as filesystem tables and network tables. Determine the variables/item sets in a case.

2) Item groupings I = I1, I2, I3, Im.

3) Set of activities D = t1, t2, t3 tn

4) Using the Apriori technique, find similar item sets.

Iterative degree of service to locate a collection of frequent item sets.

For example, if a hacker attacks a database, a login attempt results in data loss/fiddling, and the case report shows activities such as Data removed, Login attempt, attack type = SQL injection, and so on. If certain item sets are widespread, we may establish a rule stating that the motive for the assault is data theft."

5) Use Relationship Rules, which are rules in the form X Y that indicate an association between X and Y that if X occurs, then Y will occur.

If the hacker obtained operating system files, we can conclude that the reason for the assault was a system crash.

If a hacker attacks the database login and password credentials, we may argue that there is a criminal reason for data theft/data alteration.

This maximum common item set displays attack patterns as well.

Identifying further indications Correlations, coincidences, and so on When creating rule sets, keep these values in mind).

6) Configure SQL queries in accordance with the guidelines.

7) Recover the compromised data.

Identity theft, the selling of credit cards, the sale of personal computerised medical information, and cyberbullying are all examples of cybercrime. Crime Network Analysis is utilised in organised crime investigations such as drugs trafficking, fraud, terrorism, and gang offences.

Because information can flow to and from different crime groups in this environment, criminal network analysis must integrate information from multiple crime incidents or even alternative sources in order to determine regular patterns about the structure, operation, and information flow in criminal networks.
Many challenges might arise while mining law administration data, such as erroneous, missing, or inconsistent data. There are three types of criminal network analysis methods [12].

The first category is the manual approach. This method entails an analyst creating an association matrix by categorising criminal associations from original data. When data sets are exceedingly vast, this strategy can become very ineffective and unproductive [10]. The second way is to use a programme to automatically generate graphical descriptions of criminal systems to create a graphics-based style.

Although the instruments in this second way can employ multiple methods to display criminal networks, they are not sophisticated enough to give any analytical description. SNA is the third approach, and it is intended to give more complex analytical descriptions to aid in criminal investigation. To mine vast amounts of data for relevant knowledge about the structure and organisation of criminal networks, complex analysis methods are required.

Conclusion:

To summarise, data mining is the collection of digital data from many sources such as network traffic, databases, or email. This data is used to improve or optimise a company's operations or to collect statistical data for surveys. The goal of this categorization was to provide a high-level overview of the existing capabilities of computer forensic tools. It is depicted as the starting point from which constraints and references were discovered. This article also examines existing digital forensic tools and procedures.

Reference:

[1] Ms. Smita M. Nirkhi, Dr.R.V.Dharaskar, Director, Dr.V.M.Thakre, Data Mining: A Prospective Approach For Digital Forensics. International Journal of Data Mining & Knowledge Management Process (IJDKP) Pg 44, Nov. 2012

[2] S. Garfinkel, "Lessons learned writing digital forensics tools and managing a 30TB digital evidence corpus," Digital Investigation, vol. 9, pp. S80-S89, 2012

[3] 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.

[4] B. Martini and K.-K. R. Choo, "An integrated conceptual digital forensic framework for cloud computing," Digital Investigation, vol. 9, pp. 71-80, 2012.

[5] 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.

[6] 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.

[7] 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

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

[9] prashant K. Khobragade, Latesh G. Ma lik, "A Review on Data Generation for Digital Forensic Investigation using Data Mining," International Journal of Computing and Technology, Volume 1, Issue 3, April 2014.

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

[11] H. Chen, W. Chung, Y. Qin, M. Chau, J. J. Xu, G. Wang, R. Zheng and H. Atabakhsh, "Crime Data Mining: An Overview and Case Studies," Proceeding of ACM International Conference, Vol. 130, 2003, pp. 1-5.

[12] V. Justickis, "Criminal Datamining," Security Handbook of Electronic Security and Digital Forensics, 2010.