Suneeta Satpathy *et al*, International Journal of Computer Science and Mobile Applications,
National Conference on "The Things Services and Applications of Internet of Things",
Gandhi Institute for Education and Technology (GIET) Baniatangi, 23-24 March 2018, pg. 129-136
ISSN: 2321-8363
Impact Factor: 5.515

Big Data Computing Application in Digital Forensics Investigation and Cyber Security

Suneeta Satpathy
P.G Department of Comp. Sc. &, Application CEB, Bhubaneswar

Chandrakant Mallick
Department of Comp. Sc. & Engg, CEB, Bhubaneswar.

Sateesh K. Pradhan
P.G Department of Comp. Sc. & Application, Utkal University, Bhubaneswar.

Abstract

*The potential advances and applications of Digital Information Technology (DIT) in several areas of business, engineering, medical and scientific studies are resulting in information/data explosion. The demanding and growing reliance on digital media and devices has increased the volume of data creation and storage exponentially around the world [**3**], with a need of keeping logs about what data is stored and how the data is used. With overwhelming use of Digital Technology, security in Cyberspace has become a prime concern. Knowledge discovery and decision making from such rapidly growing voluminous data is becoming a challenging task for the law Enforcement and Investigative Agencies. As a new research area, Digital Forensics requires to seize the digital evidence to locate who has done it and what has been done maliciously and possible risk/damage assessing what loss it could lead to. The forensic digital analysis is unique among all forensic sciences in that it is inherently mathematical and generally comprises more data from an investigation than is present in other types of forensics. The potential of Big Data for enhanced decision making and analytic process can be seen as a tool to improve operational efficiency in digital forensic investigation with the purpose of constructing potential valuable evidence from it. So, there is a need for Law Enforcement and Investigating Agencies to have a holistic view of the Big Data challenges and opportunities for its application in Digital Forensic Domain with the objective of making robust investigation decisions. With Big Data and Data Science, this paper describes the trends of Digital Forensics served for Big Data, the challenges of evidence acquisition and further suggests the application of machine learning algorithms to process large amounts of data effectively.*

Keywords: Digital Forensics, Digital Evidence, Big Data, Machine Learning, Data Science.

1. Introduction

Digital Forensics is a new branch of forensic science of study and majority of the digital forensic investigation involves the sourcing of digital information stored on computers, mobile devices, game console and other various media storage mediums for the purpose of relevance to civil and criminal investigation [6][7]. The digital forensic investigation involves steps involves the

ISSN: 2321-8363
Impact Factor: 5.515

conventional process like Identification, Acquisition, Preservation, Examination and Presentation of findings to the chief investigator, court of law and other stakeholders by forensic expert investigators where decisions are made on the outcome of an investigation [4][6]. Forensic tools such as Forensics Tool Kit (FTK) and ENCASE [8] are used to carry out this process. The data offered by computer forensic tools can often be misleading due to the dimensionality, complexity and amount of the data presented. Also the scope of investigation and examination of evidence are limited to the examiners and investigators. Hence it can raise challenges with the procreation nature of big data spreading heterogeneously.

The data collected from several computer crime and cyber fraud and digital crime investigations require tools to facilitate efficient data management, analysis, validation, visualization and dissemination, while preserving the intrinsic value of the data and original copy of the data unmodified[6][15]. The magnitude of data generated and shared by various sectors like businesses, public administrations numerous industrial and not-to-profit sectors, and scientific research, social media sites, sensors networks, cyber-physical systems, and Internet of Things, has increased immeasurably [3]. The amount of complex and heterogeneous data including textual to multimedia content pouring from any-where, any-time, and any-device, is an era of Big Data has become an emerging data science paradigm of multi dimensional information mining for scientific discovery and analytics [9] [10].

The existing digital forensics investigation tools are based on pre-determined or signature based analytics over the filtered data that is cleaned and transformed into another form [6]. But, Big Data systems work on non predetermined analytics and organizes and extracts the valued information from the rapidly growing, large volumes, variety forms, and frequently changing data sets collected from multiple, and autonomous sources in minimal possible time, using several statistical, and machine learning techniques [5]. So with the valued nature of the Big Data and Data Science this paper evaluates the current state of digital forensics investigation and its challenges with emphasis on role of Big Data Science in it. The paper gives a closer look at the information's collected from various stages of the research, analyzes it to model further directions for carrying out effective digital investigation with Big Data Science[5][7][8][10].

2. Digital Forensics

According to a definition from the National Institute of Standards and Technology (NIST) [8], Digital Forensic is "an applied science to identify an incident, collection, examination, and analysis of evidence data". Digital forensics comprises four main processes [6] [7]:

Identification: The first step of a digital forensics investigation is identification, where an investigator identifies the incidents that are important to prosecute litigation and identifies the evidence related to those incidents.

ISSN: 2321-8363
Impact Factor: 5.515

Collection: After identifying the evidence, an investigator needs to collect evidence from various digital medias, such as cell phone, hard disk, router, etc.

Organization: Organizing the collected evidence efficiently leads towards the facts of a criminal incident. First, an investigator inspects the data and its characteristics. After that, the investigator interprets and correlates the available data to determine the facts.

Presentation: In the final phase, an investigator prepares an organized report to state his or her findings about the case, which should be admissible to the court.

3. Big Data and Big Data Digital Forensics

Big Data [9] is said to be a large volume of data sourced from various mediums such as Social Media, Health Care, Transport, e-commerce, mobile phones, satellite communications, GPS systems, media sharing through handheld devices and other modern day means of communication. Big data is also characterized by the five V's: variety, velocity volume, veracity, and value.

Variety

Big data can come from different sources, such as web pages, network or process logs, posts from social Medias, emails, documents, and data from various sensors [10]. These data can be categorized in three general types: structured (data stored in regular database), semi structured (data may not have fixed fields but can contain tags to expedite the data analysis), and unstructured (data do not have a fixed format). Hence, it is much harder to execute forensics analysis on such data.

Volume

In the age of big data, organizations deal with terabytes and petabytes of data. For example, Walmart handles more than a million customer transactions each hour and imports those into databases estimated to contain more than 2.5 petabytes of data [3]. Google processes 20 Petabytes of data per day [10].

Velocity

Velocity of new data that comes from different systems makes the big data even bigger every day. For example, International Data Corporation (IDC) estimates that by 2020, the number of business transactions on the Internet, which includes Business-to-Business (B2B) and Business-to-Consumer (B2C), will reach 450 billion per day [3].

ISSN: 2321-8363
Impact Factor: 5.515

Veracity

Veracity refers to correctness and accuracy of information. Veracity involves data quality, data governance, and metadata management, along with considerations for privacy and legal concerns [10].

Value

Value refers to the ability of turning the data into a value. The availability of big data can reveal previously unknown insights from data and this understanding leads to the value [10] [11].

4. Big Data Digital Forensics

Big Data Digital Forensics is a special branch of digital forensics where the identification, collection, organization, and presentation processes deal with a very large-scale dataset of possible digital evidence to establish the facts about a digital crime [1][2][5]. It can be described as a set of information recorded automatically by the system in relation to registered values, memory, timers, network events that are useful to the course of digital investigation [6]. This information can be used to recreate all the events that happened on a system within various spaces of times. The evidential digital information may include title, Authors, Size, Date modified, tags, categories, content status, content types, date created, date last accessed, user information etc.

Big Data Digital Forensics can face many challenges as listed below [6][12].

A. Storing and backups of Petabytes of information

The large set of data created as a result data explosion always creates the issues of data storage. Trillions of data sets are created on daily and to analyze this set of data, the investigation team has to store the data in some form or capacity [6].

B. Faster indexing of huge amounts of data

Due to the size and heterogeneity of data that need to be analyzed, faster device and methods are needed to be able to analyze data within a given time frame. So faster Indexing of data will always remain a challenge for digital investigators when it comes to big data compared to small set of data [11].

C. Methods for presenting large amount of data to the court of law/ visualization methods

With the emergence of big data, wider varieties of data are often needed to be presented after investigation or during various stages of the investigation. The large size of data always makes it a difficult task for investigators to be able to present the whole data

ISSN: 2321-8363
Impact Factor: 5.515

without cutting it down to smaller pieces and potentially omitting key information [6][11].So Presentation of digital findings from research is always a debatable aspect of digital forensics which includes technical knowledge of the judge and other courtroom members or stakeholder [4].

In spite of the above challenges, it has still tremendous amount of advantages as stated below;

a) Benefits in analyzing almost all media types under one investigation that offers as much resource as needed.

b) Taking advantages of data such as social network feeds and live feeds created through GPS and other devices can help proactive and reactive digital investigations.

c) Access to large set of data that can be analyzed based on the profile of a given suspect of information can be searched for by the digital investigators.

So Big Data with their potential to ascertain valued insights for enhanced decision-making process can be seen as a tool to improve operational efficiency with the purpose of constructing valuable evidential information for Digital Forensic Investigating Agencies [9].

5. Possible applications of Big Data and Machine Learning Techniques in Digital Forensic Investigation

The fast changing landscape in data science and uniqueness of digital forensic investigation doesn't allow the systematic categorization of tools for digital investigations. Though it is hard to systematically categorize tools and techniques due to fast changing landscape in data science and uniqueness of digital evidences some of the tools can be reviewed as follows[9] [10]. Big Data tools may not be treated as a replacement to the existing Forensic tools rather can act as an additional resource to improve the operational efficiency of tools to enhance decision making process of computer crime and cyber fraud investigations.

- Digital Forensic Data sets always lack a lot of internal correlation which leads to misclassification of file fragments obtained from file system image or from unallocated space. Map-Reduce can be a possible framework for all types of parallel tasks. It can be integrated with existing forensic tools for modeling such classification.

- Big data technologies such as the Hadoop ecosystem (e.g. Hive, Pig, Mahout, and R.Hadoop), NoSQL databases, stream mining, and complex-event processing enable to analyze large-scale, heterogeneous datasets at a high speed. They can transform security analytics by improving the maintenance, storage, and analysis of security information [10][11].

ISSN: 2321-8363
Impact Factor: 5.515

- Machine learning classification algorithms –Logistic regression, Support vector machines can be adapted to Map Reduce if the analyst forgoes the possible correlations among single fragments. Since visualization plays an important role a combined approach where a classification algorithm is combined for instance with a decision tree method would yield higher accuracy.

- Decision trees [13][14] and random forests give best results in fraud detection which involves vast data set to find statistical outliers e.g. anomalous transactions, anomalous browsing behavior.

- In audio forensics unsupervised learning techniques gives good results in separating two superimposed speakers or a voice from background noise. They rely on mathematical underpinning to find the least correlated signals.

- In image forensics again classification techniques are useful to automatically review big sets of hundreds or thousands of image files, for instance to separate suspect images from the rest.

- Neural Networks are suited for complex patter recognition tasks in cyber forensics. A supervised learning algorithm can be used to train the network to find a distinction between normal and suspicious behavior. After the event the system can be used to automatically build an execution timeline on a forensic image of a file system.

- Bayesian classifiers and unsupervised algorithms for clustering like k-means, has been successfully employed for authorship verification or classification of large bodies of unstructured texts, emails.

6. Requirements of building Digital Forensic Investigation Model using Big Data Computing

Developing an Investigation model for agencies conducting digital investigations that will utilize the big data technology requires an appropriate methodology for selecting architecture and adopting alternative techniques for cost-effective system requirements [7]. Generally accepted engineering guidelines for big data systems recommend a paradigm in which the design and development flow from an overall system requirements and constraints to a specification of the role for big data within the system. There are several fundamental issues, which should be taken into consideration when building an investigation model [6][7]:

- What architecture should be used?
- What algorithms and techniques are appropriate and optimal for a particular application?
- How should the individual heterogeneous source data be processed to extract the

ISSN: 2321-8363
Impact Factor: 5.515

maximum amount of information?

- How does the data collection environment affect the processing?
- How can the digital investigation process using big data techniques be optimized?
- What accuracy can be achieved by a digital investigation process?

A Fusion based Digital Investigation Model [12] [13] is already developed that fuses data from various heterogeneous sources in order to attain low false alarm rates and high threat detection rates. The model is motivated by JDL data fusion model [12] for internal and external attacks and overall network security in context of high-volume network traffic and heterogeneous data. The model also supports post mortem forensic analysis by preserving the necessary potential legal digital evidence. Big Data analytics can provide help for fraud detection. Big Data can provide security intelligence by shortening the time of correlating long-term historical data for forensic purposes [1][2].

Contemporary view on the problem of security is concerned with an idea that particular protective mechanisms and corresponding software must be integrated along with the forensic capabilities into a Fusion based Digital Investigation System[12][13][14] using big data computing tools and techniques interacting via exchange of information and making decisions in a cooperative and coordinated manner. These systems should be adaptive to traffic variations, reconfiguration of the software and hardware components.

7. Conclusion

Profiling, identifying, tracing, and apprehending cyber suspects are the important challenges for the researchers. The amount of digital information is now growing beyond the capacity of current digital forensics tools and procedures. Also the ubiquitous nature of computing devices and anonymity afforded by the criminal encourages destructive behavior while making it extremely difficult to prove the identity of the criminal. So there is a growing need for providing support to digital forensics in big data application domains. Also the challenges of retrieving digital evidence at present highlight the necessity of revising tenets and procedures firmly established in digital forensics. In this paper, we have defined the term big data digital forensics and identified the challenges of executing reliable forensics in the big data paradigm that can be solved with application of data science and machine learning algorithms to open the opportunity of identifying many new insights those were not possible before. Our future work includes extending the investigation model to detect and prevent the various types of computer crime and cyber frauds using big data tools and computing techniques. Application of big data tools and computing techniques along with capabilities data fusion and data mash-up technologies in the existing Fusion based Investigation model can increase the efficiency of dynamic data integration, correlation, transformation, classification and risk analysis of various investigations.

ISSN: 2321-8363
Impact Factor: 5.515

References

1. AICPA, "Big Data" Listed as Top Issue Facing Forensic and Valuation Professionals in Next Two to Five Years: AICPA Survey," http://goo.gl/ 1BgdWB, 2014.
2. Alvaro A. Cárdenas ,Pratyusa K. Manadhata and Sreeranga P. Rajan , "Big Data Analytics for Security", University of Texas at Dallas ,HP Labs, Fujitsu Laboratories of America.
3. B. Davis, "How Much Data We Create Daily," http://goo.gl/a0ImFT, 2013.
4. D. Brezinski and T. Killalea, Guidelines for Evidence Collection and Archiving, RFC3227, Feb 2002.
5. Dolly Das, Urjashee Shaw ,Smriti Priya Medhi " Realizing Digital Forensics as a Big Data Challange" 4th International Conference on "Computing for Sustainable Global Development", New Delhi (INDIA) 01st - 03rd March, 2017.
6. E. Casey (ed.), Handbook of Computer Crime Investigation, Academic Press, 2001.
7. Mohsen Damshenas, Ali Dehghantanha and Ramlan Mahmoud, "A Survey on Digital Forensics Trends", International Journal of Cyber-Security and Digital Forensics (IJCSDF) 3(4): 209-234 209, 2014.
8. National Institute of Standards and Technology (NIST), Computer Forensics Tool Testing (CFTT) project.
9. S. Sagiroglu and D. Sinanc, "Big data: A review," in International Conference on Collaboration Technologies and Systems (CTS). IEEE, pp. 42–47, 2013.
10. SAS, "Big data meets big data analytics," http://www.sas.com/resources/ whitepaper/wp 46345.pdf, Tech. Rep.
11. Shams Zawoad, Ragib Hasan, "Digital Forensics in the Age of Big Data: Challenges, Approaches, and Opportunities", ResearchGate, DOI: 10.1109/hpcc-css-icess" IEEE BigDataSecurity, August 2015.
12. Suneeta Satpathy, Sateesh Pradhan, and B.N.B Ray, A Digital Investigation Tool based on Data Fusion in Management of Cyber Security Systems, International Journal of IT & Knowledge Management, vol. 3, no. 2 pp 561- 565, June2010 .
13. Huda Fatima, Suneeta Satpathy, Satyasundar Mahapatra, G.N.Dash, Sateesh K. Pradhan, " Data fusion & Visualization Application for Network Forensic Investigation –A Case Study" ,IEEE 2nd International Conference on Anti Cyber Crimes, March 26-27,2017, King Khalid University, Abha, Saudi Arabia.
14. Suneeta Satpathy, Sateesh K. Pradhan, B.N.B Ray "A Decision Driven Computer Forensic Classification using ID3 Algorithm", Springer book Series Advances in Intelligent Systems and Computing, International Conference on Intelligent Computing, Communication & Devices, April 18-19, 2014.

