ISSN:2249-5789
Amit Pratap Singh et al , International Journal of Computer Science & Communication Networks,Vol 5(2),62-66

Common XML Creator for Various Web Log Formats used in Digital Forensic Investigation (DFI)

Amit Pratap Singh[1]
[1]*Research Scholar*
Department of Computer Applications,
Samrat Ashok Technical Institute,
Vidisha, India
[1]*amitjimail@gmail.com*

Dr. R. C. Jain[2]
[2]*Ex. Director*
Samrat Ashok Technical Institute,
Vidisha, India

Abstract

The Digital forensic emerges as vital tool for facilitating the preservation of digital information, significantly helpful in acquiring evidence from crime scene. With the advancement in digital technologies the requirement for forensic tool is gaining importance. The acquired digital information with the help of forensic tool handles with extensive care so as to retain its evidence value with it. Forensic science enables to retrieve metadata and information; ensure effectual searching over system by curious; and fine grain access rights. This proves as an effective tool for handling numerous sources of data by assigning priority to individual source and validates the integrity of acquired data. The judicious utilization of digital forensic is helpful in providing the theoretical perspective, practical solution to problem, and gaining insight to different circumstances. The aim of introducing this paper is to present a brief overview of digital forensic science and data aggregation from different potential sources which proved as evidence under crime commitment.

Keywords

Digital Forensic, XML, Common format, weblog.

1. Introduction

In today era a new class of crime has been confronting with exponential rate, significantly the crime perpetrate within the electronic or digital domain. A large number of judicial agencies in different corner of the world are setup with an enhanced requirement to investigate the crime committed over the digital or electronic media i.e Internet. Here we introduced the term "Digital Forensic" which is the science dealing with the digital information (information generated, saved, transmitted via electronic device) which becomes a significant evidence for criminal processing. The Digital Forensic define by Forensic Research Workshop as "The use of scientifically derived and proven methods toward the preservation, 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." [1] This Digital Forensic word is stimulated from the Latin word 'Forensis', allude to forum, the dictionary representation of 'forensics' represent the use of information in a court of law. However the itemize study of forensic science is a time consuming process. In commitment of crime the law bodies are under a moral obligation to go with the evidence for resolving the mystery of crime, thus paying less vigilance to the cost associated with the resources. In considering the archives of crime, it is essential for the Digital Capture Exigent that the information be protected and stored for further utilizing this in future context, thus helpful in trailing case under different circumstances.

2. Data Aggregation in Forensic

In digital forensic science the procedure of investigation is segmented into different phases. The commonly fragmented phases of investigation are: preservation, collection, examination, and analysis. In this paper we laid an emphasis on the data aggregation phase of the investigation process. Data aggregation in forensic science is a significant phase in which the potential source of data is recognized following that information is obtained from recognized sources. As the digital information is stored on the electronic devices, so compile of the digital information is either collection of devices which contain the information or stored the information in some other media. Thus aggregation is simply taking under control the computer, mobile, and other electronic device present in the crime scene, copying the information present on server, analyzing and recording the network traffic. One important characteristic in data aggregation is " time- bound" i.e the information is compile in timely manner because of the nature of dynamic data which losses with the passage of time and losing data present in battery operated devices.[10]

Identify data source

As the usage of digital technology in personal or professional increases exponentially, as with there is an enhancement in sources of data. The commonly used data sources include such as desktop computers, storage servers, networking device, and laptops. All these commonly used data sources comprises the internal DVD's drives, associated with several other port (such as Universal Serial Bus, Personal Computer Memory Card International Association) through which the external storage device is connected. Thus different external storage devices include blu-ray, flash cards, optical discs, magnetic disks and various others. Along with this nonvolatile data being present the system also comprises a volume of volatile data (loss when power is shut down for example the data present in RAM, or log file data which get overwrite when new event performed). Though with the computer associated data their present different portable device such as PDA, camera, mobile phone which comprises the data. The duty of an analyst carefully

examines the crime area and identifies all possible sources which contain the digital data.

Network data source

Organizations typically have several types of information sources concerning network traffic that might be useful for network forensics. These sources collectively capture important data from all four TCP/IP layers. The following subsections highlight the major categories of network traffic data sources firewalls and routers, packet sniffers and protocol analyzers, IDSs, remote access, security event management software, and network forensic analysis tools as well as several other types of data sources.

Many works have been devoted to preprocess data in log file for web usage mining. But few researches have been developed for preprocessing of log files for detecting malicious action. A web server log file [8, 9] is a simple plain text file which records information each time a user requests a resource from a web site responds to user requests. This file is opened when the web services of a server starts and remain open as the server. Generally there are four types of server logs:

- Access log file
- Error log file
- Agent log file
- Referrer log file

The first two types are the most commonly used. The agent and referrer logs may or may not be enabled at the server. Access log file contains data of all incoming requests and lets you track and get information about clients of the server. Error log file lists internal server errors. This information enables server administrators to correct site content or to detect anomalous activities. Agent log file provides information about user's browsers, operating system and browser version. Referrer log provides information about the link that redirects visitors to my site.

Acquiring the Data:

Once the potential source of data are identified then next is to obtain data from all potential sources. The data acquisition is processed by following the 3 step given below: discover a sketch to acquire data, acquiring data, and integrity verification of acquired data.

Discover a sketch to acquire data**:** This is an important step in acquiring data because there are large numbers of important sources. Thus, it is important for an analyst to assign priority for each source. Here we introduced some parameter for assigning priority:

- Likely Value: The computation of likely value of all potential sources is possible by taking into consideration analyst past experience of equivalent criminal scene and the developed understanding of crime scene.

- Volatility. The basic nature of volatile or live data is that they lost with the expansion of time and power off computer. However this volatile data also lost due to other course of action. The volatile data is dynamic in nature therefore it lost when other event taken place for example log file overwrite with by execution of new event. During large number of investigating cases the acquisition of volatile data is prevail over the nonvolatile data.

- Amount of Effort Required: In identifying different data sources the estimated effort vary with different sources. For example the effort in obtaining data via network router is much less as compared to the effort in obtaining from Internet Service Provider. Thus while estimating the effort not only taken into consideration the time spend by analysts and other adviser of investigating department but also the cost of apparatus and services (for example different investigating department experts).

Forensic tool has proved a useful tool for obtaining the volatile data in a situation where the analysis, monitoring , security tool failed in acquiring volatile data, copying the non volatile source of data, and preserving the actual source of non volatile data. The process of data acquiring can be performed either locally or over the network. The process of acquiring data locally is prevailing over the remote acquisition because one has entire control over the system and the data present in the system. However a certain condition may arise where a local acquisition is not preferable such as computer in locked room, location of computer is far beyond. If the data acquisition is performed over the network, then it is essential to taken into consideration the type of data aggregated and estimate effort in accomplishing aggregation. For example decide whether it is essential to collect the data from the various systems connected through different network or it is enough to duplicate the disk volume of single system over a network. The two well known techniques for file duplication are [2]

- **Logical Backup**: This backup method duplicates the file or directories present on disk volume. However this method does not copies the files which are deleted, data present in slack part of the disk.

- **Bit Stream Imaging:** Another well known method is Bit steam imaging also known as disk image because it duplicate bit- by bit entire disk including the deleted files and slacked part of disk. However this operation is time consuming and requires more space then the logical backup methods. Bit stream imaging processed using two duplication methods such as disk to disk and disk to file imaging. The first disk to disk imaging duplicate the entire content exist on one media to another similar media. The second disks to file method duplicate the disk to one logical file. The prior disk to disk method is useful because duplicate disk is directly connected to any system and content present on it can be viewed, but on other hand this copying method require the storage media similar to the original data source storage. The second disk to file imaging method permit simple backup and transfer of data file

Verify the integrity of the data

The last and important part of data acquisition is to verify the integrity of collected data. This is duty bound of an analyst to prove that the data which

acquired is not tampered data or modifiable data. Thus for verifying the integrity of data aggregated from different sources, message digest of duplicate and original data is estimated then compared this estimated digest to ensure correctness of acquired data .[3]

3. Digital forensic & its classification

The science of digital forensic deal with analysing digital data as important evidence of crime . Digital forensics [4] is in simple term stated as the investigation tool that helpful in approaching toward accused more often this deal with the crime committed by sort of electronic media .This crime is categorized into two classes:
computer based crime and computer facilitated crime.

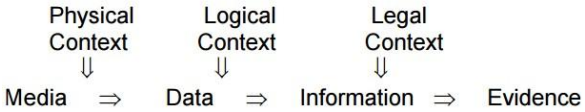

Figure 1 Data analysis and evaluation to evidence [5]

Computer based crime
This class of crime includes the action which are completely executed over system or computer, such cuber-bullying and cyber stalking. Also with this novel committed crime there also a primitive crime that executed on system such as Child pornography.

Computer facilitated crime
This class comprises the crime which executed in real world, but it was facilitated by an electronic media. One popular example of this class crime is “Fraud” some commonly action included in this communication with fraud person, building fraud documents and intentionally plan or record action to harass other person.

The investigation procedure of any committed crime is simply divided in 4 types such as criminal forensic, Intelligence gathering, e-discovery and intrusion investigation. From these the first three types are similar in indulging action, however different in their legal perspective and digital proof.

A. Criminal forensics
A major part of digital forensic is falling under remit of law enforcement including the private contractor under them. This type of investigation comprise broader analysis of crime by law enforcement bodies and different experts thus presenting a report which proved as useful evidence before court and facilitate the investigation procedure. During this investigation process the emphasis is on data obtained by using forensic tool and presenting this evidence in a manner that could be easily understood by each and every person.**B. Intelligence gathering**
In this type of investigation process intelligence are collected so that the criminal action is tracked discovered and stopped. However the evidence here is also a valid proof presented before the court later, but more than the speed is a major requirement for this type of investigation.
C. Electronic discovery (eDiscovery)
By execution nature this investigation is likely to criminal forensic. However it has legal restriction imposed on it. Two common examples which affect the e- discover investigation procedure is privacy law in which employee has legal rights not to disclose his/her personal conversation and human right legislation.

D. Intrusion investigation
The fourth type of investigation which is different from other is Intrusion investigation. This investigation comprises the network investigation in which the unauthorized users steal the information of corporate sector. During execution of this investigation analyst try to exploit the entry point of the hacker, and lessen the actions of hacker. This investigation is performed "live" in real time and inclined to identify intrusion taken place over network.

The most useful piece of evidence that can help is the system logs. One very useful kind of log is a login log, or connection log. These logs tell precisely every connection attempt that is made by recording the precise date, time, the network IP address of the computer that is attempting to log in, and the result of each login. These logs usually show the very first signs of unusual behavior, for example when an unknown address is attempting to connect to an unusual port number or when multiple unsuccessful attempts are made to login to a specific account.

Process accounting logs are very useful for revealing the activities of the intruder by showing exactly which files were executed, when, by whom and for how long. These logs are quite detailed and sometimes very useful. However, reading these logs are difficult because they are sorted in order of when the processes were terminated, so processes that ran longer than others may go unnoticed and those are still running will not be listed.[6,7]

4. Efficient Common Xml Format for Various Web Log Format

There are various ways to perform forensic analysis. But some of those only concentrate on converting various formats into single format to handle or analysis the situation. Here the proposed method concentrates on the merging of these various formats into single format. So that anyone can use this method to perform on verity of datasets.

The proposed work is shown in fig 2. According to proposed work, it will read all datasets than it will find common attributes or field from all datasets. After finding all common fields or attribute proposed algorithm will add all entries of first dataset than second than so on. The proposed algorithm for preparing the common format of all dataset are shown in fig 2.

Step 1: Start
Step 2: Read the first .txt file line by lineStep 3: Split each line to get IP, DateTime and URL from the line (whichever attribute is not found in the file, its place is populated with the string 'none')Step 4: Write these attribute values in a "Log.xml" file with a node labelled as 'info'Step 4.1: the xml file will look like:

```
<LogInfo>
<Info>
        <IP>199.72.81.55</IP>
<DateTime>[01/Jul/1995:00:00:01
0400]</DateTime>
        <URL>none</URL>
</Info>
<Info>
        <IP>none</IP>
```

<DateTime>[01/Jul/1995:00:00:06
0400]</DateTime>
<URL>unicomp6.unicomp.net</URL>
</Info>.....and so on
</LogInfo>

Step 5: Read another .txt file line by line Step 6: Split each line to get IP, Date Time and URL from the line(whichever attribute is not found in the file, its place is populated with the string 'none') Step 7: Append these attribute values in the existing file "Log.xml" file. Step 8: Stop

Fig 2: Proposed Work

5. Simulation and Results

Here, we have two datasets. First of is of comdotzone organization's web logs file and second one is NASA web log file.

Table I: Weblog of first format "comdotzone"

Comdotzone web log contains following fields:

1. IP address
2. Date
3. Time
4. Method
5. Method Responded

IP Address	Date	Time	Method	Method Respond	Browser	URL
66.249.73 .13	30/Sep/ 20 13	06:22: 04 +0000	GET /service/onlinereput ationmanagementin dia/ HTTP/1.1	404	Mozilla/5.0 (compatible; Googlebot/2 .1	http://www.google.com/ bot.ht ml

6. Used Browser

Table I: Weblog of first format "comdotzone"

7. Used Browser

Table II: Weblog of second format "NASA"

At the other end NASA web log files contains followings fields:

1. IP Address/URL
2. Date
3. Time
4. Method

IP Address/URL	Date	Time	Method	Method Respond
199.72.81.55	01/Jul/1995	00:00:01 -0400	GET /history/apollo/ HTTP/1.0	200 6245

5. Method Respond

After performing the proposed work of preparing common format for various web logs through XML, the proposed common format would be like following:

IP Address	URL	Date	Time	Method
199.72.81. 55	NA	01/Jul/19 95	00:00: 01 0400	GET /history/apol lo/ HTTP/1.0

Table III: Common format for all weblogs formats

6. Conclusion

From the above discussion it is cleared digital forensic has its root spread over wide range. Thus, computer forensic plays an important role in governing security. Computer develop an understanding of how different crime committed so it must required to fine tuned forensic tool that effectively helpful in acquiring evidence and mitigate the wave of crime. Forensic science along with useful investigation tool also offer a valuable knowledge from different research community. The different ongoing research in forensic science is imperative, as they confronted different digital technologies with its usefulness in acquiring evidence.

7. References

[1] Jeremy Leighton John, "Digital Forensics and Preservation", DPC Technology Watch Report 12-03 November 2012.

[2] Karen Kent, Suzanne Chevalier, Tim Grance and Hung Dang "Guide to Integrating Forensic Techniques into Incident Response", Recommendations of the National Institute of Standards and Technology August 2006

[3] Chet Hosmer, "Proving the Integrity of Digital Evidence with Time", International Journal of Digital Evidence, 1(1), Spring 2002.

[4] Catherine H. Conly, "Organizing for Computer Crime Investigation and Prosecution", Washington, DC:
National Institute of Justice, 1989

[5] Lecture note on "Computer Forensics: an approach to evidence in cyberspace".

[6] M Reith, C Carr, G Gunsch"An examination ofdigital forensic models". International Journal of Digital Evidence, August 2010.

[7] Pascal Schottle, "Digital Forensics in Computer and Cellular Networks", Chair for Communication Security July 19, 2009

[8] K.R. Suneetha, Dr. R. Krihnamoorthi, "Identifying User Behavior by Analyzing Web Server Access LogFile" , IJCSNS , Vol. 9, No. 4, 2009

[9] L.K. Joshila Grace, V.Maheswari, DhinaharanNagamalai, "Analysis of web logs and web user in web Mining" , IJNSA, Vol.3, 2011.

[10] Lecture note from Sonia Bui, Michelle Enyeart andJenghuei Luong, "Issues in Computer Forensics", May 22, 2003.