Association Rules based on NTFS File System Metadata for Computers Forensics

Punam
Assistant Professor, Dept. of CSE,
University College, Miranpur, Patiala, India

Abstract. Computer related crime keep on increasing due to increasing storage capacity, rapid development of information technology and internet. Investigating whole hard disk data takes lots of time and effort. In forensics, File system metadata is good indicator of user's action. Today available forensic tools presents file system metadata but they lacks in providing relationship that exist between file system metadata. This paper focuses on analysis of file system metadata and to project association rules based on metadata collected from system. These rules provide relationship between metadata residing on hard disk. This relationship helps in predicting user behaviour and grasp user's system usage patterns.

Keywords: Association Rules, File system metadata, Forensics.

I. INTRODUCTION

Earlier Computer was not in common man's hand. Rather it was a just commodity of elite class. These people were less involved in crime. Moreover, usage of computer was very less because knowledge related to computers was in few hands.

But, if we talk about present scenario, a trend has changed. Computer is very common now a day as it is in reach of everyone's pocket. Computers have been widely used in every field of life. As a result, computer related crime occurs very often. Most of important evidences are increasingly stored in computer's hard disk. Therefore, analysing data stored in storage devices become very important in collecting important evidences. But with increase of storage capacity, rapid development of information technology and internet, posses challenges to forensic investigators.

Generally in forensics investigation an image of suspect computer is created and it is mounted on forensic investigator's computer. But this takes lot of time due to increase in storage capacity. Also finding case related data from large amount of data residing in suspect storage device takes lot of time and effort. However discovering evidences from such large data, we can take help from data that is created by system. Data that is created by system is generally hidden and is in form of files. There are various system files from which we can discover usage pattern of a suspect from target system. These include file system metadata, prefetch files, registry, web browser files, and specific document file; recycle bin structural analyses and files hidden in slack etc.

File system metadata is a good indicator of user's action. If computer usage can obtain from file system metadata then it is more effective method than to trace whole hard disk data. Today most widely used file system is New Technologies File System (NTFS). It is designed by Microsoft and is default file system in Windows operating systems and even in free UNIX distributions. In NTFS everything is file. The MFT (Master File Table) is the heart of NTFS as it contains information about all files and directories. Every file and directory has at least one entry in table[2]. Metadata can be obtained from MFT such as file name, extension, creation date and time, modification date and time, last accesses date and time etc.

Today there are various forensic tools such as Forensic Toolkit (FTK) [6], Encase [7], and The Sleuth Kit (TSK) [8] etc. These tools present metadata of data which exist on hard disk. However, these tools lacks in providing the information about relationship exist between metadata reside on suspect hard disk that can help in tracing usage patterns. Hence, there must be some rules based on the metadata collected from hard disk that can help in getting some information about user and his/her behavior and tracing the user's usage history from the target system. This helps in uncovering information that can serve as evidences.

The research described in this paper focuses on association rules that can be drawn from file system metadata are discussed. These rules provide relationship exist between metadata of file system collected from suspect's hard disk. They can help in discovering usage patterns and some information about suspect and his/her behaviour.

II. LITERATURE REVIEW

As increasing storage capacity makes investigation very complicated. Analyzing whole disk data takes lots of time. However, there are various ways to begin the analysis of system for digital evidences. This may include file system metadata, prefetch files, registry, web browser files, and specific document file; recycle bin structural analyses and files hidden in slack etc. Lee et. al. proposed methods for selective acquisition of file system metadata, registry & prefetch files, web browser files, specific document files without duplicating or imaging the storage media. Furthermore they suggested a method to analyze the acquired data stepwise and quickly and effectively trace the use of computer in the crime scene [1].

Buchholz et. al. suggested the role of file system metadata in digital forensics. They discussed four benefits of using metadata in forensics. Firstly, the information is automatically collected and stored by the system. Secondly, the information is collected automatically with no extra cost

for setting up logging mechanisms. Thirdly, information is directly stored with the object of interest. It is not necessary to correlate various system logs to obtain the desired information. Lastly, Tampering with the information is not as simple as tampering with a file [3].

Timestamp analysis is not new in forensic investigation. Large amount of work has been done on temporal analysis. Timestamps help in event reconstruction that occurred in past. Chow et. al. focused on temporal analysis on NTFS file system and projected 7 rules that help in catching user's actions based on MAC times of NTFS file system. For example out of 7 rules one rule is "In a folder, if files M times are equal to C times and the files have "very close" C (M) times, the files may have been downloaded in a batch from another system over the network" [4].

Rowe et. al. found time patterns associated with disks and files. They used Real Disk Corpus of over 1000 drive images from eight countries as a corpus. They found 14 kinds of drive usages based on three parameters (Weekday Ratio, Day Ratio, and Evening Ratio)[5]. However Rowe et. al. use only timestamps of file system metadata to determine behaviour and usage patterns of user. But using other kind of metadata along with timestamps can be very useful for investigation. For instance Brian Carrier said that Files of size smaller than 700 bytes can also contain evidence [2]. So discovering rules based on file system metadata can be very helpful in predicting user behaviour.

III. PROPOSED FRAMEWORK

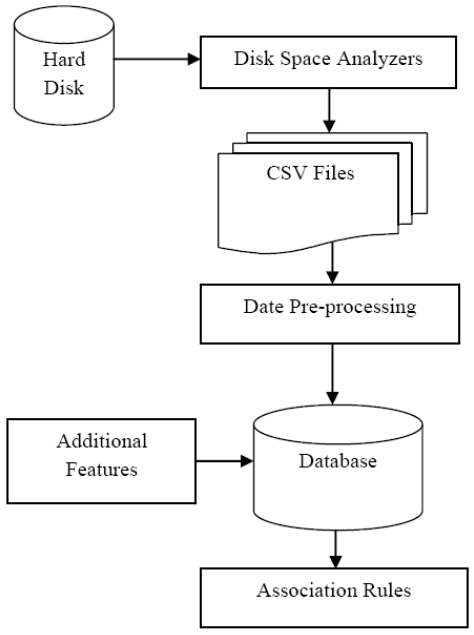

Fig. 1. Proposed Framework

Disk space analyzers (analysis tools) are useful for collecting metadata of file system. These tools have disk space reporting functionality. Out of these tools some have export facility to make a separate CSV file for each reporting. Different tools provide different kind of metadata information (features). So some kind of data pre- processing is required in order to combine features obtained from different tools. After pre-processing, there is one complete CSV file containing all features obtained from different analysis tools. This newly created CSV file can be imported into a database. In order to make rules, some additional features are required. These features can be derived from existing features of database. After collecting all features which are necessary for making rules into database, association rule mining can be performed.

IV. METHODOLOGY

The existing forensic tools provide metadata related to files exist on hard disk. However, these tools lacks in providing relationship that exist between metadata present on disk. This relationship can help in discovering behaviour of user and his/her system usage. Methodology which is opted by this paper for discovering relationship between file system metadata residing on hard disk consists of following steps:

4.1 Disk Space Analyzers

Disk space analyzers are basically software utilities. These are graphical, menu-driven applications. They reveal those files and folders that occupy most of hard drive space. The basic functionality of these analyzers is to display disk space usage by getting size of each folder, its sub-folders and files in a folder or drive. Most of these applications analyze this information to generate a graphical representation showing disk usage distribution according to folders or other user defined criteria. There are a number of excellent free contenders. Different analyzers use different graphical representations to show space utilization. This includes pie charts, bar charts, radial, treemap etc.

Rather than displaying only disk space usage visually these analyzers also provide metadata related to files and folders in disk such as file name, file path, owner, and creation, modification and last accessed data and time etc. But some disk space analyzers have this disk space reporting facility. Some analyzers give an opportunity to get **disk space report** as a separate file in the form of text, CSV, HTML, XML etc.

This work has used GetFoldersize and Xinorbis6 analysis tools. Both the tools are freeware and have export functionality to save disk space reporting. These tools are used to collect metadata related to files and save them as separate CSV files. The term feature is used for each kind of metadata information provided by these CSV files.

4.2 Data Pre-processing

Not CSV files generated by both these analysis tools contain same information. Some of features present in CSV file of one tool may be absent in CSV file of another tool. Similarly CSV files of both tools may contain some features which are present in both the files. So some kind of pre-processing is required, in order to combine features of both CSV files

obtained from two different tools. A new CSV file is created that contains the features of both files.

But combining these features is not easy. As these CSVs contain one entry for each file resides on drive, position of entry for corresponding file may be different in these two CSVs. In order to combine features of different CSVs, features value related to drive file must be in one row. In CSV obtained from GetFoldersize there are feature called File Name and File Path. In CSV obtained from Xinorbis6 there is a feature called File Path. This File Path includes in itself File Name feature. For example, GetFoldersize CSV contains information in form:

File Name: FORENSIC ARTICLES.docx

File Path: E:\Articles\

Xinorbis6 CSV contains information in form:

File Path: E:\Articles\FORENSIC ARTICLES.docx

So in order to combine these CSVs, use of these features are helpful in making a new CSV that contain features of both CSVs.

4.3 Database Creation

After pre- processing, a CSV file is obtained that contain metadata for each file which reside on hard disk. Importing CSV into database creates a table. This table contains following attributes:

Table 1. Attributes (Features) of database table

File Name	Last Accessed Date
File Size(B,KB,MB)	Last Accessed Time
File Path (This File Path does not include File Name)	Owner
Creation Date	Extension
Creation Time	Type
Modification Date	Category
Modification Time	Attributes

4.4 Additional Features

For making rules from the metadata, we need some additional features along with features till has been collected. These features can be derived from attributes of database table. These features are added as attributes of existing database table along with other attributes. These features are as follows:

Table 2. Additional attributes (features) added to database table

Feature Name	Derived from Feature (s)	Values
Depth	File Path	Integer
File Name Length	File Name	Integer
Special Character in File Name	File Name	{0, 1}
Size Class	Size	{Very Small, Small, Below Average, Average, Above Average, Large, Very Large}
Working Time	Creation Time Modification Time Last Accessed Time	{NWT,WT} NWT-Non Working Time WT-Working Time
Week Days	Creation Date Modification Date Last Accessed Date	{Weekends, Working Days}
Recently Used Files	Creation Date Modification Date Last Accessed Date	{YES, NO}

4.5 Construction of Association Rules

At this step, metadata and other kind of information about the files is available in form of database table. Now rules can be constructed based on these features. While making the rules, both support and confidence related to rules are calculated. This may lead to formulation of manual expert system. Rules formed like:

1. If Special Character in File Name=1 and File Name Length>20 and Size Class=" VERY SMALL" and Depth>5 ⇒ Working Time="NWT" [14%, 91%]
2. If Special Character in File Name=1 and File Name Length>20 and Size Class="VERY SMALL" and Working Time="NWT" ⇒ Recently Used Files="YES" [15%, 93.4%]
3. If Special Character in File Name=1 and File Name Length>20 and Size Class="VERY SMALL" and Working Time="WT" ⇒ Week Days="Working Day" [12%, 94.6%]
4. If Special Character in File Name=1 and File Name Length>20 and Depth>5 and Week Days="Weekends" ⇒ Category="Uncategorised" [7%, 59.2%]

By using only 8 attributes (Special Character in File Name, File Name Length, Size Class, Depth, Working Time, Recently Used Files, Week Days and Category) it come up with lots of rules by taking each and every combination. From 1st rule it can be concluded that user basically used system at non working hours with the files having special character in their names, more than 20 characters in their names, are of very small size and some are basically located with depth

greater than 5. So such kind of rules can determine user's system usage patterns and his/her behaviour.

V. CONCLUSION AND FUTURE SCOPE

In this work metadata of file system is used in determining user behaviour and user's system usage patterns. For collecting metadata disk space analyzers are used. Some additional features are computed from collected metadata. Rules are constructed based on metadata of files.

However constructing rules are not easy. Taking each and every combination of collected features and calculating their support and confidence takes lots of time and effort. It is not possible to construct these rules manually. So there is a need to automate this process in future in order to grasp user's behaviour and user's system usage patterns.

VI. REFERENCES

[1]. Lee, S.B., Bang, J., Lim, K.S., Kim, J., Lee, S.: A Stepwise Methodology for Tracing Computer Usage. In: Fifth International Joint Conference on INC, IMS and IDC, pp. 8152-8157, IEEE Computer Society (2009)

[2]. Carrier, B.: *File system forensic analysis*. Addison-Wesley (2005)

[3]. Buchholz, F., Spafford, E.: On the role of file system metadata in digital forensics. Journal of Digital Investigation, Vol. 1, No. 4, pp. 297-308 (2004)

[4]. Chow, K.P., Law, F.Y.W., Kwan, M.Y.K., Lai, P.K.Y.: The rules of time on NTFS file systems. In: Proceedings of 2nd International Workshop on Systematic Approaches to Digital Forensic Engineering, pp. 71-85, IEEE Computer Society, Seattle, Washington (2007)

[5]. Rowe, N.C., Garfinkel, S.L.: Global Analysis of Drive File Times. In: Fifth International Workshop on Systematic Approaches to Digital Forensic Engineering, pp.97-108, IEEE Computer Society (2010)

[6]. Access Data, http://www.accessdata.com

[7]. Guidance Software, http://www.guidancesoftware.com

[8]. The Sleuth Kit (TSK) & Autopsy, http://www.sluethkit.org