A Forensics Method of Web Browsing Behavior Based on Markov Chain Model

Sanjeev Shukla[1] Nisha Patil[2] Narendra Joshi[3]
[1,2,3]Department of Computer Engineering
[1,2,3]Sandip University, Nashik, Maharashtra, India

Abstract— Issues about privacy-preserving data mining have emerged globally, but still the main problem is that non-sensitive information or unclassified data, one is able to infer sensitive information that is not supposed to be disclosed. Data collection is a necessary step in data mining process. Due to privacy reasons, collecting data from different parries becomes difficult. Privacy concerns may prevent the parties from directly sharing the data and some types of information about the data. How multiple parties collaboratively conduct data mining without breaching data privacy presents a challenge. The objective of this Paper is to provide solutions for privacy-preserving collaborative data mining problems.

Keywords: Markov Chain Model, Forensics Method of Web Browsing Behavior

I. INTRODUCTION

With the rapid development of web technology, more and more people go online for information, transactions and doing other business activities. Web history data contain a lot of user's view information and is helpful for computer forensics analysis. The cyber world provides a convenient platform for criminals to conduct their illegal activities. Web browser is one of the most commonly used tools that is used to getting information to help cybercrimes. The key part of network forensics is to collect the various sources of digital evidence. The browser retains the user's browsing time, URL and other log information. In the web browser forensics, to analyze these logs can find some non-direct found browser behaviors of some computer users. Data of cybercrime cases provide a new and appropriate auxiliary information for the detection. However, due to the data volume, the diversity of the data structure and other characteristics, analysis of the data was difficult in the past. It drew helpful information on the case manually from browsing history information. Forensics officers transformed the data into an organized database, and used criminal network analysis tools to evaluate the organized data. Currently, some companies and organizations have issued their own web forensics tools, such as Web Historian and Forensic Tool Kit, to analyze various browsers history data. These tools support most browsers, like Internet Explorer, Mozilla Firefox, Opera. However, and so on. However, these tools are limited to analyze the storage formats of historical data of web browser, and provide only a URL history list, cookies, cached pages and other basic information. Forensics officers still need to manually analyze electronic evidence in complex information. This operation is usually very time-consuming, and the error rate is quite high. A framework for finding the suspects on the basis of scientific analysis of the related web browsing history sequences and links, thereby improving the convenience and reliability of the investigation process. Installing antiviruses, filters, firewalls, and scanners is insufficient to secure e-mail communication. In this context, cyber forensic investigation (also called digital investigation) is employed to collect credible evidence by analyzing e-mail collections to prosecute criminals in the court of law. The scope of e-mail analysis ranges from simple keyword searching to authorship attribution of anonymous e-mails. For instance, an investigator may want to get an overview of an e-mail collection by computing simple statistics such as the distribution of e-mails per sender/recipient domains. In some situations an investigator may try to narrow down the scope of investigation by selecting (usually few) malicious e-mails from regular ones. For this purpose, usually content-based clustering is applied to divide e-mails into different groups on the basis of the subject matter of e-mails. The conceived subject matter could be the type of crime, such as pornography, hacking, or terrorism, etc., in which e-mails were instrumental in committing those crimes. E-mails can be clustered on the basis of stylometric features to determine the writing styles of different individuals contained in an e-mail collection.

Frequent pattern mining plays a major field in research since it is a part of data mining. Many research papers, articles are published in the field of Frequent Pattern Mining (FPM). This chapter details about frequent pattern mining algorithm, types and extensions of frequent pattern mining, association rule mining algorithm, rule generation, suitable measures for rule generation. This chapter describes about various existing FPM algorithms, data mining algorithm for crime pattern. By applying frequent pattern mining algorithm and suitable measures, the proposed new algorithm in future is applied to crime dataset in order to find out the suspects in the short span of time.

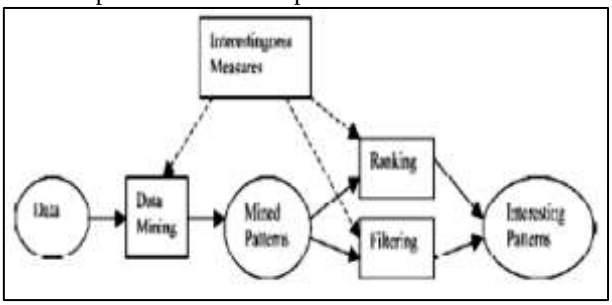

Fig. 1: Roles of interestingness measures in the data mining process

II. ASSOCIATION RULE MINING

Association rule mining concept has been applied to market domain and specific problem has been studied, the management of some aspects of a shopping mall, and an architecture that makes it possible to construct agents capable of adapting the association rules has been used. Data mining refers to extracting knowledge from large quantity of data. Interesting association can be discovered among a large set of data items by association rule mining. The finding of interesting relationship among large amount of business

All rights reserved by *www.ijsrd.com*

transaction records can help in many business decisions making process [24]. Association rules mining is an important task in the field of data mining, and frequent item set mining is a key step of many algorithms for association rules mining. There had been lots of work done for mining of association rules. When the dataset are large, the rules generated may be very large, but some of them are not interesting to the users, so, it is common to set some parameters to reduce the numbers of rules generated, support and confidence are two common parameters. An association rule R is of the form A → B, where A, B are disjoint subsets of the attribute set I. The support for the rule R is the number of database records which contain A ∪ B (often expressed as a proportion of the total number of records). The confidence in the rule R is the ratio:

- Support for R
- Support for A

These two properties, support and confidence, provide the empirical basis for derivation of the inference expressed in the rule, and a measure of the interest in the rule. The support for a rule expresses the number of records within which the association may be observed, while the confidence expresses this as a proportion of the instances of the antecedent of the rule. In practical investigations, it is usual to regard these rules as "interesting" only if the support and confidence exceed some threshold values. Hence the problem may be formulated as a search for all association rules within the database for which the required support and confidence levels are attained. Note that the confidence in a rule can be determined immediately once the relevant support values for the rule and its antecedent are computed. Thus the problem essentially resolves to a search for all subsets of I for which the support exceeds the required threshold. Such subsets are referred to as "large", "frequent" or "interesting" sets [16].

III. ASSOCIATION RULE MINING IN LARGE DATABASE

Association rule mining used to mine the sales transactions between items in large database recognized as a most significant area of database research. Measuring a large database there are different techniques are used. Pruning strategy and interestingness is one of the measuring techniques for measuring large database. Large database consists of many fields. Each field consists of their own process. They different depends on their field of work. Suppose we consider a customer transaction of a large database each transaction consists of items purchased by a customer in a visit items purchased by a customer in a visit , time of purchase, category of payment, net amount etc. so it is a tedious process to maintain for huge amount of customer transaction. An efficient algorithm implemented in association rule mining. Apriori algorithm is best for association rule mining in large database. This algorithm generates all significant association rules between items in the large database. Today, most research related work on data mining in association rules are encouraged by an wide range of application areas, such as financial transactions, engineering, health care, GIS, and broadcastings. Association rule mining used to originate interesting association or correlation relationships among a large set of items in the large database. In large database Application of association rule mining in market basket analysis are:-

- To analyses the point of sales transaction.
- From uses information on what customers buy to provide insights into who they are and why they make certain purchases.
- From which products are purchased together and which are most willing to support.

IV. FORENSICS MODEL OF WEB BROWSING BEHAVIOR

Here we can say easily get the user's browsing history log files on any browser. Convenient source of the data and more favorable file structure, combined with increasingly sophisticated data mining techniques, make it possible to process the huge data files. In this paper, computer forensics in web browsing behavior pattern is discussed. The behavior pattern is important for the analysis of anonymous users and suspicious behavior. A behavior pattern library can be build by web browsing behavior pattern mining to help forensics. Browser behavior analysis techniques have received some attention in recent cyber forensic investigation cases. Statistics show that there is a pattern on the site activities of most users. It contains many repetitive movements in the long run, and it shows some unique patterns and trends. Searching for a few weeks or even months of access history of a single client, web log mining can discover some information about these unique patterns and trends. There is a collection of suspicious web browser history in an investigation. The main objective is to extract the user's web browsing behavior patterns. In order to get the user's browsing patterns, a forensics method of web browsing behavior based on association rule mining is described here such as depicted in below figure

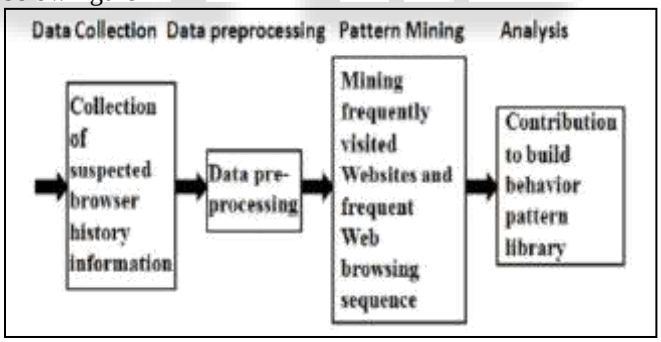

Fig. 2: Forensics steps of web browser behavior.

In above figure the method is divided into 4 steps:-

A. Collection of suspected browser history information:

Collect the browser history data of suspected browser history, mainly including the user's accessed website, accessed time, accessed method and the page classification. The original data implies the user's web browsing information.

B. Data preprocessing:

Convert the data form into the suitable data form for mining association rules. This step provides the interface to the data pattern mining layer. The purpose of this method is to get user's browsing behavior patterns by data mining. Therefore, the data format should be suitable for mining these results.

All rights reserved by www.ijsrd.com

C. Mining frequently visited websites and frequent web browsing sequence:

Use association rule mining algorithm to process the preprocessed data and to get the user's frequently visited websites and frequently browsing URLs sequences. Obtaining the frequent item sets is the main objective in this step. The frequently visited websites reflect the user's point of interest, and the frequent web browsing sequence could be seen as one of the user's browsing patterns.

D. Contribution to build behavior pattern library:

Case analysis usually needs to know the suspects' behavior patterns. It is important to build behavior pattern library. A variety of computer behavior patterns make contribute to the behavior patterns library. Browsing behavior pattern including frequently visited websites and frequently browsing URLs sequences will become an important part of the behavior pattern library.

V. WORK ACCOMPLISHED SO FAR

A. "A Forensics Method of Web Browsing Behavior Based On Association Rule Mining"

In this paper, author discuss about the security issues for a web browser a forensics method of web browsing behavior based on association rule mining is presented. The method aims at providing the necessary data support to build the behavior pattern library for investigation. The records of the user's browsing history are collected to be analyzed. The obtained original data are pretreated to transactional data which are suitable for association rule mining. Frequent browsing time and frequent web browsing sequences are obtained from the transactional data by Apriori algorithm. The mining results are helpful for identification and recognition of anonymous or suspicious web browsing behavior patterns.

In conclusion, this method has a good effect on helping web forensics. The experiments further validate the feasibility and rationality of this method. The setting of minimum support will affect the number of web browsing patterns, which has a great impact on the forensic investigation. However, the experiments are not perfect. For example, the experimental results were not intuitive, and the meaning of parameters is not clear enough.

B. "Towards An Integrated E-Mail Forensic Analysis Framework"

In this article, author present their e-mail forensic analysis software tool, developed by integrating existing state-of-the-art statistical and machine-learning techniques complemented with social networking techniques. As a result of growing e-mail misuse, investigators need efficient automated methods and tools for analyzing e-mails. They developed an e-mail analysis framework to assist investigators gather clues and evidence in an investigation in which e-mail communication is relevant. The framework offers different functionalities ranging from e-mail storing, editing, searching, and querying to more advanced functionalities such as authorship attribution and e-mail account localization. Extending traditional authorship identification techniques, they proposed a new technique of mining style variation. This will help to capture the change that occurs in the style of person with respect to different contexts/recipients.

C. "A Complete Survey on Application of Frequent Pattern Mining and Association Rule Mining on Crime Pattern Mining"

In this paper author presents the survey of various Frequent Pattern Mining and Rule Mining algorithm which can be applied to crime pattern mining. The analysis of literature survey would give the information about what has been done previously in the same area, what is the current trend and what are the other related areas. This paper explains the concepts of Frequent Pattern Mining and three important approaches that is candidate generation approach, without candidate generation and vertical layout approach. It also explains various frequent pattern algorithms and how it can be applied to different areas particularly in crime pattern detection.

D. "A Novel Approach of Mining Write-Prints for Authorship Attribution in E-Mail Forensics"

In this paper, author introduces an innovative data mining method to capture the write-print of every suspect and model it as combinations of features that occurred frequently in the suspect's emails. This notion is called frequent pattern, which has proven to be effective in many data mining applications, but it is the first time to be applied to the problem of authorship attribution. Unlike the traditional approach, the extracted write-print by our method is unique among the suspects and, therefore, provides convincing and credible evidence for presenting it in a court of law. Experiments on real-life e-mails suggest that the proposed method can effectively identify the author and the results are supported by a strong evidence.

E. "Anomaly Extraction in Backbone Networks Using Association Rules"

In this paper, author use meta-data provided by several histogram-based detectors to identify suspicious flows, and then apply association rule mining to find and summarize anomalous flows. Using rich traffic data from a backbone network, we show that our technique effectively finds the flows associated with the anomalous event(s) in all studied cases. In addition, it triggers a very small number of false positives, on average between 2 and 8.5, which exhibit specific patterns and can be trivially sorted out by an administrator. Our anomaly extraction method significantly reduces the work-hours needed for analyzing alarms, making anomaly detection systems more practical.

F. "An Approach for Web Log Pre-Processing and Evidence Preservation for Web Mining"

In this paper, author examines information preprocessing systems and different steps included in getting the obliged substance adequately. A powerful web log preprocessing technique is constantly proposed for web log preprocessing to concentrate the client designs. The information cleaning method uproots the unessential passages from web log and sifting calculation disposes of the uninterested characteristics from log record. During the past few years the World Wide Web has become the biggest and hottest means of

All rights reserved by www.ijsrd.com

communication and information proliferation and promulgation. It provides a platform for exchanging varied information. The quantity of information accessible on the net is increasing chop-chop with the explosive growth of the World Wide Web and the advent of E-Commerce. While users area unit given additional service options and information, it's become tougher for them to find the relevant information of their interest, the problem unremarkably known as information overload.

Authors have described a fully reversible log file repository scheme capable of significantly reducing the amount of space required to store the compressed and preprocessed log, the obtained test results show it manages to improve compression of different types of log files. It is lossless, fully automatic (it requires no human assistance before or during the compression process), and it does not impose any constraints on the log file size.

VI. NOTEWORTHY CONTRIBUTION IN PROPOSED WORK

The aim of this Paper is to propose an effective and efficient scheme for cyber forensics to resolve following problems that arise during log file analysis for forensic investigation.

- Digital investigations are becoming more time consuming and complex as the volumes of data required to analysis is large in size therefore lossless compression log file in lossless manner.[12]
- There is possibility of manipulation or deletion of log information. Because log files are incriminating evidence against attackers, these files are at risk of attacks. Therefore, a mechanism is needed to prevent the manipulation and deletion of log info and log files by attackers and maintain the contents of log files that are created at the time of outbreak.[15]

VII. PROPOSED FRAMEWORK

The Details of Proposed framework are as follows:

A. Log Fetching

The first tier contains the client that produces the log data. Some hosts run logging client applications or services that make their log data available through networks to log servers in the second tier. Other hosts make their logs available through other means, such as allowing the servers to authenticate to them and retrieve copies of the log files.

B. Log compression

Log compression is storing a log file in a way that reduces the amount of storage space needed for the file without altering the meaning of its contents. Log compression is often performed when logs are rotated or archived. .A multi-tiered log file compression solution shall be proposed. Every of the three tiers address one notion of redundancy. The first tier handles the resemblance between neighbouring lines. The second tier handles

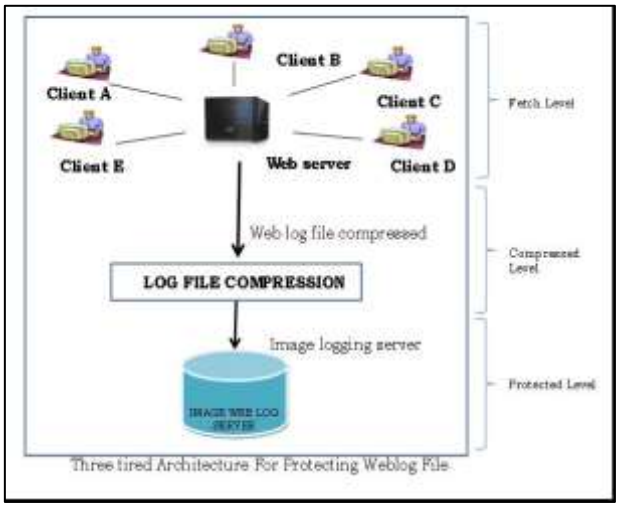

Fig. 3: Three tired Architecture

Proposed architecture for Compact & Secure Logging System global repetitiveness of tokens and token formats. The third tier is general-purpose compressor which handles all the redundancy left after the previous stages. The tiers are not only optional, but each of them is designed in several variants differing in required processing time and obtained compression ratio. This way user with different requirements can find combinations which suit them best. We propose five processing schemes for reasonable ratios of compression time to log file size reduction. A collection of scripts to determine the compression ratios, compression times and decompression times when using data compression was compiled. These scripts were used to run tests on a collection of log files and the obtained statistics recorded.

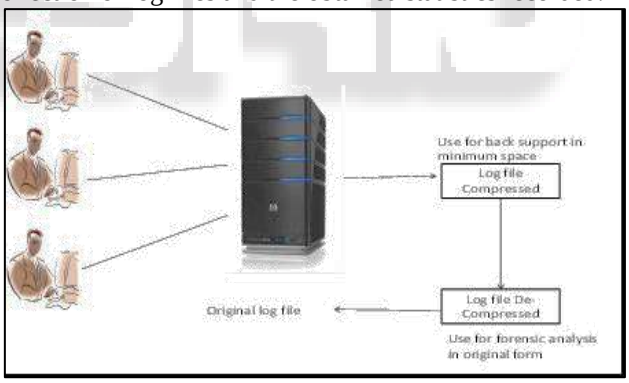

Fig. 4: Log file compression

Maillog files were investigated to determine their structure and evaluate what semantic knowledge can be exploited to improve the performance of standard compression programs.A number of dictionaries were constructed for each maillog file which can be used to perform word-replacement and improve the performance of standard compression programs.

These dictionaries were evaluated and a single dictionary constructed which could be used by a pre-processor for all the log files and achieves a greater amount of improvement than the individual dictionaries. The results were analyzed and the performance of the different techniques was evaluated using the information obtained from the analysis. Conclusions were drawn as to which combinations yielded the best results for different monitoring scenarios

All rights reserved by www.ijsrd.com

C. Log Protection

Log files include log entries associated with system and network activity, they need to be protected from breaches of their confidentiality and integrity. For example, logs might intentionally or accidentally record susceptible information such as users' passwords and the content of e-mails. This raise security and isolation concerns containing both the individuals that assessment the logs and others that might be able to access the logs through authorized or unauthorized means. Logs that are secured improperly in storage or in transit might also be susceptible to intentional and unintentional alteration and destruction. This could cause a variety of impacts, including allowing malicious activities to go unnoticed and manipulating evidence to conceal the identity of a malicious party. For example, many root kits are specifically designed to alter logs to remove any evidence of the root kits' installation or execution.

To meet data retention requirements, our proposed methodology need capture log entries in image format at server side simultaneously as log entries records by log file in .text format. However size of log file after capturing in image format is very large therefore probability of connection over network traffic gradually increase for resolving this problem, our proposed methodology first compressed log file then change compress log file into image log file in this way traffic over network is decrease and along with that forensic analyser got compact & secure log file in lab for investigation and performed fast and efficient investigation.

Our proposed methodology provide compact & secure log file for investigation and make computer forensic fast & efficient.

VIII. EXPECTED OUTCOME

Log file are mostly captured the behavior of machine not the behavior of end user. Log file provide troubleshooting, security and pro-active system administration that provide significant help in caching suspicious end user and in process of cyber forensic. Moreover, since the logs contain confidential information, they must be protected strictly. Therefore a secure logging scheme that ensures the integrity and confidentiality of the logs is needed. implemented system protect the generated log from illegal tampering by implementing an image logging server that catch .text log file in an image format. Proposed methodology significantly reduce the space requirement at image logging server to hold image of .text log file by applying lossless compression over .Decrease the loss of data

1) Find the user behavior of internet user
2) Compact the security of log protection of sever
3) Find the user chain using markov model
4) Provide secure login for user

REFERENCES

[1] Yiyun Zhang, Guolong Chen "A Forensics Method of Web Browsing Behavior Based on Association Rule Mining" 2nd International Conference on Systems and Informatics, IEEE, 2014. Pp 927-934.
[2] Rachid Hadjidj, Mourad Debbabi, Hakim Lounis, Farkhund Iqbal, Adam Szporer, Djamel Benredjem "Towards an integrated e-mail forensic analysis framework" Elsevier ltd. 2009. Pp 124-137.
[3] Anna Monreale, Dino Pedreschi, Ruggero "Anonymity Preserving Sequential Pattern Mining" IEEE, 2012. Pp 1-31.
[4] Christopher Neasbitt, Roberto Perdisci, Kang Li, and Terry Nelms "ClickMiner: Towards Forensic Reconstruction of User-Browser Interactions from Network Traces" ACM, 2014. Pp 1-12.
[5] Dominik Herrmann, Karl-Peter Fuchs, Hannes Federrath "Fingerprinting Techniques forTarget-oriented Investigations in Network Forensics" 2010. Pp 375-390.
[6] D.Usha, Dr.K.Rameshkumar "A Complete Survey on application of Frequent Pattern Mining and Association Rule Mining on Crime Pattern Mining" IJACST, 2014. Pp 264-275.
[7] Farkhund Iqbal, Rachid Hadjidj, Benjamin C. M. Fung, Mourad Debbabi "A Novel Approach of Mining Write-Prints for Authorship Attribution in E-mail Forensics" IEEE, 2008. Pp 1-10.
[8] Daniela Brauckhoff, Xenofontas Dimitropoulos, Arno Wagner, Kav´e Salamatian "Anomaly Extraction in Backbone Networks Using Association Rules" IEEE, 2011. Pp 1-13.
[9] Richa Chourasia, Preeti Choudhary "An Approach for Web Log Pre-Processing and Evidence Preservation for Web Mining" IJCSE, 2014. Pp 210-215.
[10] Tasawar Hussain, Dr. Sohail Asghar and Nayyer Masood, "Hierarchical Sessionization at Preprocessing Level of WUM Based on Swarm Intelligence", 6th International Conference on Emerging Technologies (ICET) IEEE, 2010. Pp 21-26.

All rights reserved by *www.ijsrd.com*