An Enhanced Approach for Digital Forensics using Innovative GSP Algorithm

Priyanka Kayarkar
NRI Institute of Research and Technology,
Bhopal, India

Prashant Richhariya
Prof., NRI Institute of Research and Technology,
Bhopal, India

ABSTRACT
The advent of world-wide web not only changes our life view but also gives rise to advanced forms of digital crimes. Today's era is the digital age, over the internet most of the facts are transferred through the digital devices. Cyber criminals always use Digital devices to conduct digital crime. The applicability of various forensics techniques in digital forensics helps the forensics investigators to adopt practical frameworks and methods to recover data for analysis which can comprise as evidence. In the field of Digital Forensics Data Mining has huge prospective. Computer forensics is a rising discipline investigating the computer crime. The goal of data mining technique is to find the valuable relationships between data items.

This paper proposes a data mining approach for digital forensics investigations which is very important in today's information age. Frequent Sequence Mining in data mining is one of the most important concepts used in Digital forensics Science. This thesis is an imperative work for Digital forensics investigations with maximum accuracy by using GSP algorithm.

Keywords: Digital Evidence, Cyber Forensics, Sequence mining Data SOM classification, BARTLETT'S test

1. INTRODUCTION
The amalgamation of growth in the number of cyber business processes and internet has created opportunity for the criminals to conduct crime. With the rapid and advanced growth in internet gives rise to advanced forms of digital crime. For conducting illicit activities criminal use digital devices, So There is a need of Digital Forensics. Digital forensics is a emergent and imperative field of research for today's world wide web in order to trap digital criminals. The goal of digital forensics is to find out the digital evidence for the forensics investigation. A digital forensic investigation is an inquiry into the unfamiliar or questionable activities in the computer-generated space or electronic world and Data Mining is the essential feature involved in the investigation process [10].Digital Evidence can be defined as the clues which can be recovered from digital sources and helps in digital forensics investigations. Evidences are very fragile to deal because if it is handled indecently it can be disfigured. The evidence is considered as accurate and reliable if the core of the story the material tells is believed and is consistent, and there are no reasons for doubt. Digital evidence can be classified, compared, and individualized in several ways. One of those ways is by the contents of the evidence. For example, investigators use the contents of an e-mail message to classify it and to determine which computer it came from. Sequences are elementary to modelling the three key medium of human communication like speech, handwriting and language and are considered as crucial data types in several sensor and monitoring applications. Mining models for network intrusion detection view data as sequences of IP packets. Text information extraction systems model the input text as a sequence of words and delimiters. GSP Algorithm (Generalized Sequential Pattern algorithm) is an application of sequence mining [9]. A priori_algorithms are mostly used for solving sequence mining problems. Firstly all the frequent items are discovered in a level-wise fashion and the occurrences of all singleton elements in the database are counted. The non-frequent items are removed by filtering the transactions. At the end of this step, each transaction consists of only the frequent elements it originally contained. This customized database becomes an input to the GSP algorithm. This process requires one pass over the complete database.

2. LITERATURE SURVEY
SPADE algorithm is used for Detection of malicious executables that are known beforehand is usually performed using signature-based techniques [1]. The techniques which are used here are usually rely on the former precise knowledge of the malicious executable code, and are represented by one or more rules that are stored in a database. The new signatures, based on new observations are frequently added to the database in order to get updated data. The weakness of these techniques is to detect totally new and un-encountered malicious executables. The behavior of malicious programs is determined and assigned as a sequence of system calls during training phase. A detection phase, identifies malicious executables by comparing their own run time sequences of system calls in the database , that are characteristic to only malicious executables. These techniques developed system for the MS-Windows operating system but the same ideas can work for other OS's.

Discriminant analysis has been employed in digital forensic to determine whether contraband images, such as child pornography, were intentionally downloaded or downloaded without the consent of the user [2]. Often, individuals prosecuted for crimes based on digital evidence claim that a Trojan horse or virus installed on their computer system was responsible. In this instance, Discriminant analysis provided a mechanism for event reconstruction and enabled digital investigators to counter the Trojan defence by examining the characteristics of the data.

The identification of various privacy issues in cyber security and digital forensics, issues that use for protecting privacy of data in forensic investigation, whereby how forensics investigators may have infringed user privacy while conducting forensics investigations, and how user privacy is always under threat without proper protection is discussed by Asou Aminnezhad. It has also reviewed the current development trend shift in this industry, why such trend could have happened and its drive. The paper has reviewed various fields and their development in the technicalities and technologies to address this problem. This paper describes

each field in a nutshell that explains how these technologies work, and what are their approaches in solving the problem of preserving privacy. The reviews are split into three sections, each with its corresponding fields of reviews and explanation. The paper then analyses these reviews and view them from the user and forensics investigator's perspectives, whether such development in cyber security and digital forensics actually improve the efforts on preserving privacy [3]. The paper concluded that while every development has its positive approach and finds the solution for the issue of privacy preservation at rest exists, by considering non-technical aspects in professionalism in practice and the ambiguity of scenarios causing some approaches to be counterproductive.

Extracting knowledge and information from e-mail text has become an important step for cybercrime investigation and evidence collection. It is one of the most challenging and time consuming tasks due to special characteristics of e-mail dataset. The problem of mining is the writing styles from a collection of e-mails written by anonymous authors is solved by using Cluster analysis for the anonymous e-mail by the stylometric features and then extract the whiteprint, i.e., the unique writing style, from each cluster. They proposed the solution for the problem of authorship identification by building a classifier in which they assumed training dataset is available. They propose a method and developed a tool for the investigator about the potential suspects of the given anonymous e-mails, to visualize, browse, and explore the writing styles. [4]

The most talented directions of network applications regarding the next generation of Internet technology evolutions are Social Networking Services (SNS). Immeasurable global on-line community members allocate general interests with each other by means of the User Generated Content (UGC) platforms. Facebook provides facility to the social networking participants to distribute the digital contents to authorized clients or precise groups. As cybercrimes flourished in latest years, extra digital crime investigations have strong relations to Facebook. Facebook has been exploited via global perpetrators. As a result, This paper highlight on live data acquisition within the RAM of the desktop PC with prominence on some dissimilar strings that could be found in order to renovate the previous Face book session, which plays an tremendously valuable role for the associate digital forensics investigators to nurture additional thoughtful decisions concerning the discovery of breadcrumb digital evidences in this unmatched cybercrime incidents epoch. [5]

Investigating of possibility of predicting several user and message attributes in text-based, real-time, online messaging services is discussed in the paper Chat Mining. For this purpose, a large collection of chat messages is examined [6]. The applicability of various supervised classification techniques for extracting information from the chat messages is evaluated. Two competing models are used for defining the chat mining problem. A term-based approach is used to investigate the user and message attributes in the context of vocabulary use while a style-based approach is used to examine the chat messages according to the variations in the authors' writing styles. Among 100authors, the identity of a chat message's author is correctly predicted with 99.7% accuracy.

An investigation of authorship gender attribution mining from e-mail text documents is described in paper of Appropriate Gender Identification from the Text [7]. A set of topic, content-free e-mail document features such as style markers, structural characteristics and gender-preferential language features are used. Support Vector Machine learning algorithm is used. Experiments using a corpus of e- mail documents generated by a large number of authors of both genders performed for author gender categorization. In their approach, two popular machine learning algorithms are used: decision tree and SVM.

Potentially numerous sources are examined and events are compiled and are grouped according to some criteria and repeatedly happening event sequences are recognized in paper "Event Sequence Mining to Develop Profiles for Computer Forensic Investigation Purposes",. Here, the methodology and techniques to extract and contrast these sequences are described by using Sequential Pattern Mining algorithm [8].

3. PROPOSED WORK

In this paper the Data Mining Technique is used for the purpose of Digital forensics Investigation. Here forensics investigation is done on text data set, where we are using GSP algorithm of Sequence mining. But in this paper we are making this algorithm more effective by adding concept of statistical test analysis and SOM classification Technique. We demonstrate the efficacy of Self-Organizing Kohonen maps (SOM) as a useful technique for the discovery of statistical insights and models from large data sets, i.e. exploratory data analysis .We show that by using SOM, high-dimensional data can be projected to a lower dimension representation scheme that can be easily visualised and understood. This new algorithm works more intelligently and effectively for getting better digital forensics investigation process and produce considerable and desirable output.

The Text data set is fed as an input and initially for Data recovery and Data generation BARTLETTS test of sphericity is performed on that text data set for verifying the assumption of equal variances. In particular, this assumption is made for the frequently used one-way analysis of variance. This test will check the equality between variable and number and defines the percentage of equality. After verifying the assumption the SOM classification technique is employed on the text data set for Data Analysis. SOM classification is done to obtain trained data Set.

On that Trained data set GSP (Generalized Sequential Pattern minining) algorithm Sequence mining is applied to get sequence data as output.

The following Flowchart elaborate work flow of our Project briefly.

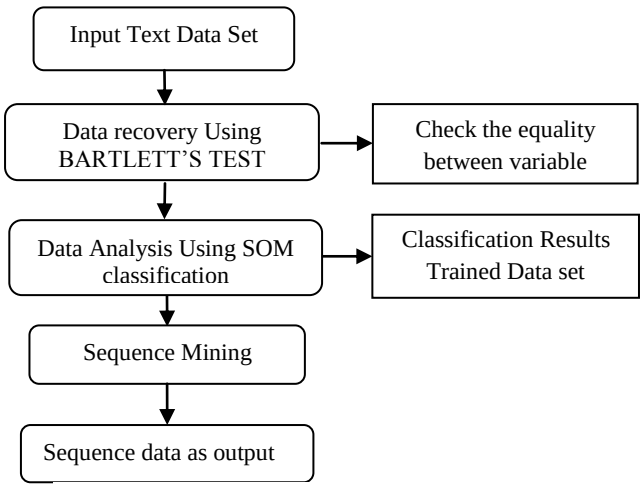

Fig 1: Work Flow of the proposed Digital Forensics System

3.1 Existing GSP Algorithm:

Algorithm makes multiple database passes. In the first pass, all single items (1-sequences) are counted. From the frequent items, a set of candidate 2-sequences are formed, and another pass is made to identify their frequency. The frequent 2-sequences are used to generate the candidate 3-sequences, and this process is repeated until no more frequent sequences are found. There are two main steps in the algorithm.

Algorithm:

F1 = the set of frequent 1-sequence k=2, do while F(k-1)!= Null;

Generate candidate sets Ck (set of candidate k-sequences);

For all input sequences s in the database D

Do

Increment count of all a in Ck if s supports a

Fk = {a Є Ck such that its frequency exceeds the threshold}

k= k+1;

Result = Set of all frequent sequences is the union of all Fks

End do

End do

3.2 Proposed Algorithm:

1. Input Text data set
2. Apply Statistical analysis on input data set
4. Apply BARTLETT's Test of sphericity to check equality of covariance matrices of various class using:

$V=(n-k)\log(|\sum i|)-\sum(ni-1)\log(|\sum i|)$ where,

V= statistical test

n=total no of observations

Ni=Total No. of Observations of class

$|\sum|$=Means of Determinant of Matrix

5. Apply SOM classification for getting the trained data set
6. Apply GSP algorithm For Finding the Sequence
7. Check the accuracy of GSP using

Accuracy=TP+TN/TP+TN+FP+FN

4. EXPERIMENTAL RESULTS

The data set can be obtained from any of the known source and system has the flexibility of importing data set from various sources.

Step 1: Statistical Analysis

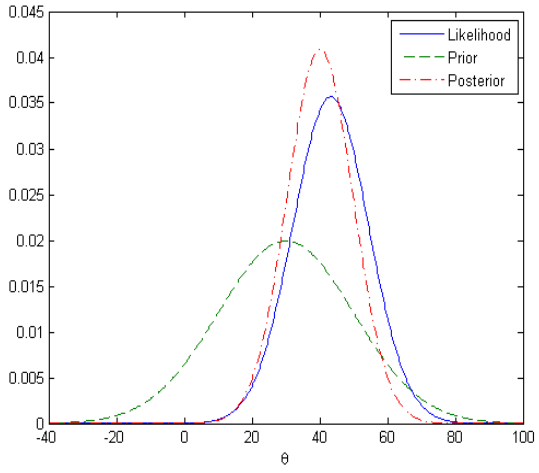

Fig 2: Output of Statistical Analysis

Maximum Likelihood Estimation (MLE) is mostly used in Statistical inferences. MLE selects the parameters that exploit the likelihood of the data, and is spontaneously appealing. In MLE, parameters are supposed to be anonymous but fixed, and are computed with some confidence. In Bayesian statistics, the ambiguity about the unfamiliar parameters is quantified using probability so that the unknown parameters are considered as random variables. Bayesian inference is the process of analyzing statistical models with the inclusion of prior facts about the model or model parameters.

The above graph shows the prior, likelihood, and posterior for theta.

Step 2: SOM Classification:

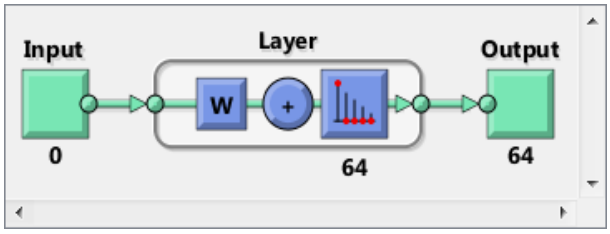

Fig 3: SOM with initial weight 0

The self-organizing map (SOM) is a neural network model that has been fruitfully useful in the clustering and visualization of high-dimensional data. It is usually 2 Dimensional because is used to map high-dimensional data onto a low-dimensional space .The SOM is an unsupervised learning, which means that the learning process is totally without any feedback. There are 2 layers in SOM namely the input layer and the output layer (see Figure 4.2). The neural network starts with random preliminary weights, so the results of this demonstration will diverge to some extent every time it is run. Here, The input size is 0 because the network has not yet been configured to go with our input data. Each neuron in the input layer represents an input signal and this layer is fully connected with nodes at the output layer. A two-dimensional grid of neurons is generated by output layer where each neuron represents a node of the final structure. Weights are representing the connections whose values whose values shows the strength of the connection.

When an input pattern is presented to the input layer, the neurons in the output layer will compete with one another, during learning process. The neurons whose weights are the closest to the input pattern in terms of Euclidian distance will be considered as the winning neuron. Once the winning neuron has been recognized, the weights of the winning neuron and its neighbourhood will be updated. The SOM configures the output neurons into a topological representation of the original data after the learning process, by a process called self-organisation. The SOM is helpful in finding the probable correlations between dimensions in the input data .This can be achieved by process of component visualization of maps. Each component map visualises the spread of values of a particular component (or dimension). Possible correlations are exposed by distinguishing component maps with one another.

Step 3: A neural network generation that will be trained to estimate median house values.

SOM creates self-organizing maps to classify samples with as much in depth as required by selecting the number of neurons in each dimension of the layer.

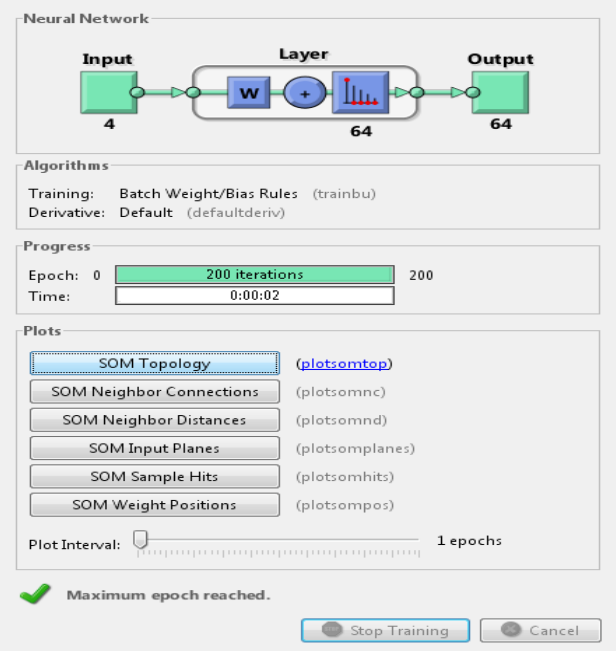

Fig 4: Network Being trained using Algorithm

Here the network being trained by using the algorithms. It also displays the training state during training and the criterion which closed training will be highlighted in green. The Class vectors of each of the training inputs are computed by SOM. The function vec2ind returns the index of the neuron with an output of 1, for each vector. The indices will range between 1 and 64 for the 64 clusters represented by the 64 neurons. To know how good a machine learning algorithm is we need to compare them in a pale way. It's all the time possible to instruct a classifier so that is gives 100% correct answers to the training samples, but it will not applicable for new samples not seen before. The newly generated SOM is then categorized depending on generated system which shows performance in terms of precision and accuracy.

Step 4: Sequence mining

By using the GSP Generalized Sequential Patten algorithm and with the help of Self organizing Map we are getting the aligned sequence.

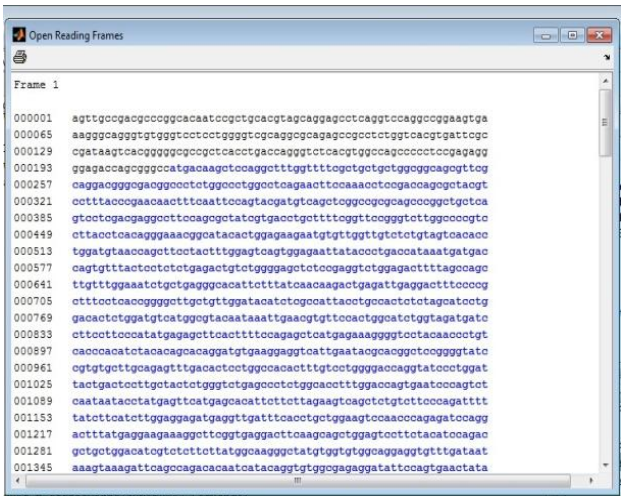

Fig 5: Sequence data as output

We are getting sequence data with the help of frequent sequence patterns and set of candidate sequences.

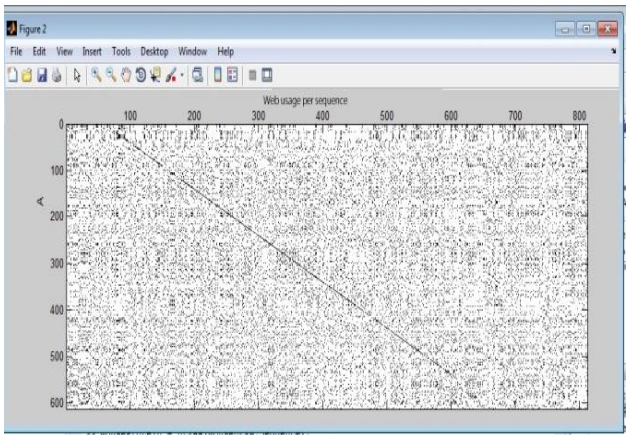

Fig 6: Web usage per sequence

Step 5: Comparing Accuracy

Here accuracy is compared by taking data in USB flash drive FD1,FD2. In Flash drive FD1, result is obtained but on this data BARTLETT'S Test is not applied .BARTLETT'S Test is performed on data which is in flash drive FD2 and the result is obtained. Accuracy of result in Flash drive FD2 is more than in flash drive FD1 as shown below.

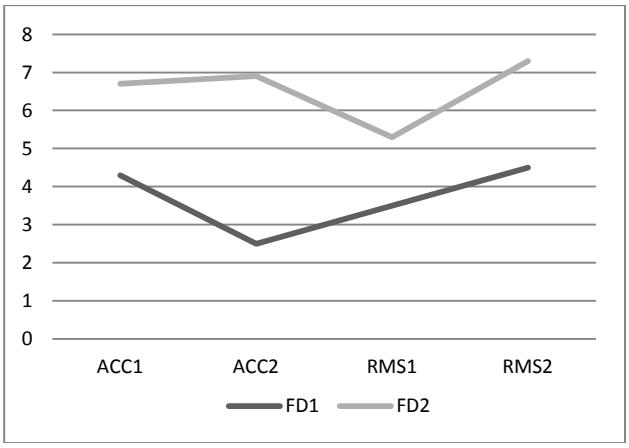

Fig 7: Accuracy graph

Table 1: Comparison of the accuracy of Result

USB Flash Drive	No of users	Classify Accuracy	RMS Value
FD1(Without BARTLETT'S TEST)	SINGLE	98.3	0.021
FD2(Using BARTLETT'S)	SINGLE	99.2	0.05

5. CONCLUSION

In this paper the adoption of new GSP algorithm added the concept of statistical test analysis and SOM classification Technique. In this paper we have demonstrated analysis of textual data by using SOM classification. SOMs provide a more robust learning. Because the SOM adopts unsupervised learning to classify the data, thus, no a priori knowledge about the data distributions are included. Though this paper has only focused in on GSP algorithm of Sequence mining, it would be beneficial to explore other algorithms which are packaged in Data mining.

This work describes one such method based on sequence found within text data. Sequences compiled and are classified according to some criteria. The methodology and techniques to extract sequence data are then described and discussed using Sequence mining concept.The adoption of novel algorithm provides a standardized process for investigators to follow. Thus, this paper will serve the requirement of that system which fulfil the needs of general Digital forensics.

6. FUTURE SCOPE

Digital forensic is very wide concept and covers so many issues related to digital crime .So this system has a lot of future scope in area of digital forensics. The rapid growth in information and communication technology, technically advanced crimes are emerging. When criminals use digital devices, practical frameworks and methods to recover data for analysis which can pretence as evidence are used by investigators.

This proposed system is developed in such a way that it has lot of flexibility to accommodate any small or substantial changes made in its structure to improve the working performance considerably. This system could be easily applied for purpose of email mining. Email is most widely used way of written communication over the internet and with the advent of World Wide Web emails frauds are increasing. In cyber forensics investigation process emails can be considered as powerful evidence. So this paper plays important role in preventing the textual email frauds. This system could be effectively used in Antivirus software systems, on large datasets, so better precise and practical results could be obtained.

7. ACKNOWLEDGMENTS

I sincerely thank to my honourable guide Prof. Prashant Richhariya and others who have contributed towards the preparation of the paper.

8. REFERENCES

[1] Mohammed J. Zakispade: An Efficient Algorithm for Mining Frequent Sequences ,Machine Learning, 42, 31–60, Kluwer Academic Publishers. Manufactured in The Netherlands. 2001

[2] Kailas Kumar, Sanjeev Sofat Naveen Agarwal S.K.Jain, Identification of User Ownership in Digital Forensic using Data Mining Technique International Journal of Computer Applications (0975 – 8887) Volume 50 – No.4, July 2012

[3] Ali Dehghantanha ,Asou Aminnezhad , Mohd Taufik Abdullah , A Survey on Privacy Issues in Digital Forensics , International Journal of Cyber-Security and Digital Forensics (IJCSDF) 1(4): 311-323, 2012 (ISSN: 2305-0012

[4] Farkhund Iqbal,Hamad Binsalleeh, Benjamin C.M. Fung, Mourad Debbabi,Mining Writeprints from anonymous Emails For Forensics Investigation, Science direct journal,dec-2010

[5] Hai Cheng Chu,Der Jiunn Deng,Jhon Hyuk Park. Live Data Mining Concerning Social Networking Forensics Based on a Facebook Session Through Aggregation of Social Data, Selected Areas in Communications, IEEE Journal August 2011 Volume 29 Issue 7

[6] Tyfun Kucukyilmaz, B.barla cambazoglu, Fazli Can ,Chat Mining Predcting User message attributes in Computer Aided Communication ,Science Direct journal, Information Processing and Management 44 1448–1466 march 2008,

[7] Corney, M., de Vel, O., Anderson, A., and Mohay, G. 2002. Gender preferential Text Mining of E-mail Discourse, The 18th Annuals computer security association (ACSAC2002). Press, Volume 30, Issue 4, 55–64.

[8] Tamas Abraham, Event Sequence Mining to Develop Profiles for Computer Forensic Investigation Purposes by, Information Networks Division Defense Science and Technology Organization.nov 2007

[9] Jiawei Han, Kamber M Morgan. Kaufmann Publishers. Data Mining Concepts and Techniques 2005

[10] International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, No.6, November 2012