ARTIFICIAL INTELLIGENCE APPLIED TO DIGITAL EMAIL FOR FORENSIC APPLICATION

Mr. Shrimant B. Bandgar*

Mr. Mahesh Sale**

Dr. B. B. Meshram***

ABSTRACT:

The number of computer security incidents is growing exponentially and society's collective ability to respond to this crisis is constrained by the lack of trained professionals. The increased use of the Internet and computer technology to commit crimes indicates an abuse of new developments that requires a response by those involved in law enforcement. In this paper we see new research approach that uses artificial intelligence and data mining techniques to study spam emails with the focus on law enforcement forensic analysis. In this 1st we retrieve useful attributes or features from spam emails, these are use in intelligence toolkit to reduction size to investigation then we use clustering algorithm to form relationships between messages. These first clusters are then refined by using a weighted edges model where membership in the cluster requires the weight to exceed a chosen threshold and data mining to managed database. Herein, we describe the use of Artificial Intelligence in computer forensics through the development of a multiagent system and span email tracking retrieve useful attributes from spam emails.

Index Terms - computer forensics, Electronic Mail, Spam, artificial intelligence, multiagent systems, Data Mining, Cyber Crime, digital investigation.

* Department of Computer Technology, VJTI, Mumbai.(India)

** Department of Computer Technology, VJTI, Mumbai.(India)

*** Prof & Head of Computer Technology, VJTI, Mumbai.(India)

1. INTRODUCTION:

In recent years, Spam email has become a major problem for society not only because the number of spam emails is astonishingly massive and growing but also because more and more spam emails are related to cyber crimes. Even though these kinds of spam emails have violated laws and caused damage, it is difficult for law enforcement personnel to stop them for the following reason. The forensic examination of computer systems consists of several steps to preserve, collect and analyze evidences found in digital storage media, in such a way that they can be presented and used as evidence of unlawful actions involving those resources. At a crime investigation, digital

Evidence can be of importance in a number of serious crimes such as child exploitation, forgery of documents, tax frauds and even terrorism.

2. Related Work:

2.1Artificial Intelligence Applied to Computer Forensics [1]

To be able to examine large amounts of data in a timely manner in search of important evidence during crime investigations is essential to the success of computer forensic examinations. The limitations in time and resources, both computational and human, have a negative impact in the results obtained. They describe the use of Artificial Intelligence in computer forensics through the development of a multiagent system and case-based reasoning. Their goal is to analyze and correlate the data contained in the evidences of an investigation and based on its expertise; present the most interesting evidence to the human examiner, thus reducing the amount of data to be personally analyzed

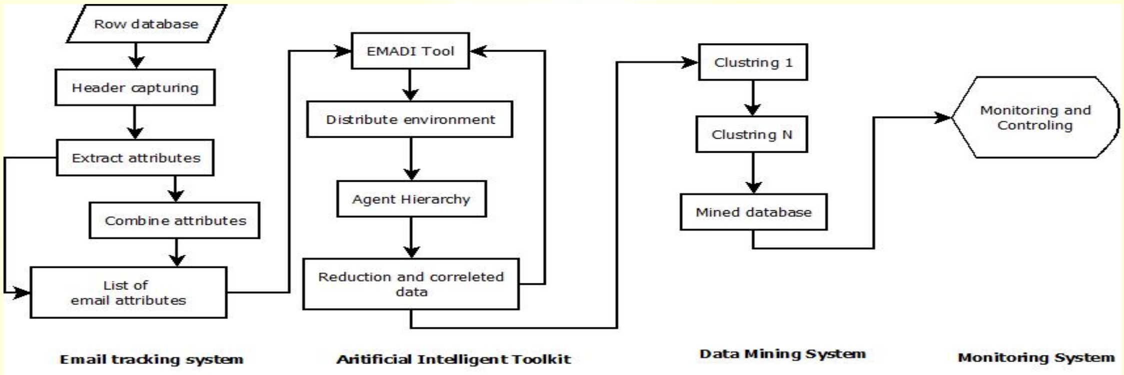
2.2 Mining Spam Email to Identify Common Origins for Forensic Application [2]

The conduct illegal business on the Internet. Therefore, in this paper they describe a new research approach that uses data mining Techniques to study spam emails with the focus on law enforcement forensic analysis. After exact useful attributes from spam emails, use a clustering algorithms to form relationships between messages. This technique has been successful in identifying relationships between spam campaigns that were not identified by human researchers, enabling additional data to be brought into a single investigation.

2.3 Drawback of above system

- In 1st part the system AI is used and then Investigation database reduced but manage the result of the investigation is not handle good way.
- In 2nd system Investigation database is manage good way but reduction of data to investigation is not given.

3. Proposed system:



Proposed System : Artificial Intelligence Applied to digital Email for forensic Application

3.1. Artificial Intelligence in forensic application

We know the forensic examination of computer systems consists of several steps to preserve, collect and analyze evidences (number of spam emails is astonishingly massive and growing but also because more and more spam emails are related to cyber crimes) found in digital storage media, in such a way that they can be presented and used as evidence of unlawful actions involving those resources. At a crime investigation, digital evidence can be of importance in a number of serious crimes such as spam emails are selling pirated software, illegal drugs, or promoting online gambling, tax frauds and even terrorism. The constant growth in the capacity of digital storage media and the widespread presence in everybody's daily life represent also a growth in the demand for those examinations and likewise in the volume of data to examine. It is difficult for law enforcement personnel to stop them for the following reasons [2]:

- The daunting volume of spam emails has made it virtually impossible for human to collect evidence from it;

- Criminals who create and distribute spam emails are using various techniques to disguise their true identities and make it hard to track them down.

- Available set of forensic tools is not robust enough when it comes to analyzing a great number of evidences and correlates the findings. As we know a consequence, computer forensic experts work excessively time consuming. The computational resources required to do such examinations are also a problem, since most of the available forensic tools have no distributed processing capabilities. Our goal is to present a new forensic tool to help experts during specialized forensic examinations in order to obtain significantly better results when compared to those obtained by the currently used tools considering three aspects [1]:

(i) Reduction of routine and repetitive analysis while also reducing the amount of evidence that must be personally reviewed by the expert,

(ii) Correlation of evidences,

(iii) Distribution of processes. With this, human and computational resources can be applied more efficiently.

3.2. EXTRACTING EMAIL ATTRIBUTES

Our research works on spam email usually start with building a word corpus based on the email content or studying email traffic, such as the domain name portion of the sender's email [11]. The email content approach is likely to fail on spam emails with no content, but only an attachment. In fact, our email collection shows that most spam emails with attachment have no body content. Many spam emails contain a fake "From" header, so the sender's email address does not really exist. Therefore, there is no simple solution, and it is necessary to extract as many attributes from the emails as possible. In our study extract attributes have been successfully parsed from the messages **: "message_id", "sender_IP_address", "sender_email", "subject", "body_length", "word_count", "attachment_filename", "attachment_MD5", "attachment_size", "body_URL","body_URL_domain"**. Some attributes are broken down into two sub-attributes, for example, **"body_URL"** into **"machine_name"** and **"path".**

Some attributes are useful for global clustering because most email message have a non-null value in that attribute, such as email subject or sender's IP address. But these attributes may be weak evidence that do not prove two emails are related. Two emails with common subject, such as "Re:" and "Fwd", may actually come from different spammers. Firstly use the email **attribute extraction algorithm** and consider the attribute as artificial agent to criminal investigation data reduction; this process detail sees next sub section. Other attributes are good for clustering a specific subgroup of spam emails, such as "body_URL_domain", which only works with spam email with URLs. But a domain name, especially a spam domain, is very strong evidence showing relationship between two emails if they both point to the same domain. The derived attributes provide further evidence of relationship between spam emails or spammers. For example, if two different domains point to the same IP address, then they are related; and if two IP addresses host the same web pages, then the two IP addresses are related. Derived attributes are very useful in finding non-obvious relationships and validating initial clusters built from inherent attributes.

3.3. ARTIFICIAL INTELLIGENT TOOLKIT

We know many definitions for a multiagent system (MAS) [8,10], but they all refer to a computational system composed by more than one agent. An intelligent software agent (ISA) uses Artificial Intelligence (AI) in the getting of goals. In this work, present the latest results obtained by the use of the ***Email MultiAgent Digital Investigation toolkit*** (EMADIK), a multiagent system to assist the Email forensics expert on its examinations. The system is composed of a set of ISAs that perform different analysis on the digital evidence related to a case on a distributed manner. In EMADIK, each ISA contains a set of rules and a knowledge base, both based on the experience of the expert on a certain kind of investigation. Since the examination of digital evidence in crime investigations share similarities, EMADIK uses CBR to determine which agents are better employed in which kind of investigation. This also allows the agents to reason about the evidences in a way that is more adequate to the specific case in question. At the moment, the EMADIK has six specialized intelligent agents implemented:

HashSetAgent calculates the MD5 hash from a email and compares it with its knowledge base, which contains sets of emails known to be ignorable or important. We might cite that some of these hash sets contain more than 100 million hash values, from different software's, as cited in.

EmailSignatureAgent examines the Email headers, to determine if they match the header value. If someone changes the email header value in order to hide the true purpose of the email, this will be detected by this agent.

TimelineAgent examines dates of creation, access and modification to determine events like system and software installation, backups, web browser usage and other activities, some which can be relevant to the investigation.

WindowsRegistryAgent examines Email related to the windows registry and extracts valuable information such as system installation date, time zone configuration, removable media information and others.

EmailPathAgent keeps on its knowledge base a collection of Paths which are commonly used by several application which may be of interest to the investigation like P2P (peer-to-peer), VoIP and instant messaging applications.

KeywordAgent searches for keywords and uses regular expressions to extract information from Email such as credit card numbers, URLs or e-mail addresses. The proposed agents are a reduced set that allows for many rules to be conceived and many examinations to be carried over, as a proof of concept. The case-based approach also provides a way to improve the agent's results over time. As another example of the case-based approach, we can also cite hash set comparisons in order to ignore unimportant emails. If an inexperienced examiner tries to compare every hash set he has available against every single email, the process will take too long and the results will not be much better than those obtained by an experienced examiner who chooses the most likely hash sets so he can have quick but yet effective results. To coordinate and organize the work of these specialists, we propose a four layer hierarchy, similar to human organizations, as used for example in the work of.[12] Figure Presents this hierarchy .

Agents can collaborate by observing and modifying one another's work through the use of a common base named blackboard. This gives the opportunity for agents to cooperate and reach good results. To better understand how the system works we will explain EMADIK's operation processes. Each entry contains the agent's recommendation, an user-friendly description and the time taken to examine the email, for benchmarking purposes. There are three distinct levels of recommendation: (i) **ignore** - the strongest recommendation to ignore a email, indicating its unimportant according to the agent,(ii) **alert** - strongly recommends the selection of a email, and (iii) **inform** - this recommendation is an intermediate value, which contains information to help the human reviewer to decide whether to select that email or not. There can be an additional sign (+ or -) representing an ignore or alert bias, respectively.

3.4. MINING THE EMAIL IN FORNSICS

After the email extraction and AI Agent reduction information which email is important to criminal investigation resource information (we call it as wetness). Then reduced information is applied to next forensics data mining work; with help of that manage such big data. By getting such goal we can use clustering techniques.

3.4.1. CLUSTERING METHODS

Two clustering methods have been used in our experiments thus far. The **agglomerative hierarchical algorithm [11]** is used for the global clustering of the entire dataset. When this clustering method is applied, the largest cluster contained too many emails, indicating the assertion of relationships which were not present. Next, the **connected component with weighted edges algorithm** is used to overcome this false positive situation. If a cluster resulting from the first method is found to be weak, the second clustering algorithm is applied, which is designed to require stronger evidence for clustering.

A] Agglomerative Hierarchical Clustering Based on Common Attributes

An agglomerative clustering method is used for global clustering to group spam emails based on common values of email attributes. In the beginning, each email message by itself is a single cluster. Then clusters that share a common attribute are merged. Each time a new attribute is introduced, clusters from the previous iteration will be merged based on the common values in the new attribute. The old clustering results are backed up in case the process needs to be reversed due to false positives. D(i, j) is defined as the distance between cluster i and j. D(i, j) = 0 if cluster i and j share a common value in an attribute and D(i, j) = 1 if not. Two clusters are merged if distance is 0. A common attribute value means exact string matching. In our experiment, 'subject' is used in the first iteration of global clustering. Therefore, two clusters are merged if they share a common subject. 'Subject' is used because most emails contain a subject and two emails with the same subject are presumed to be more likely to be originated from the same source. Domain name is used as the attribute for the second iteration. A domain name (*.e.g.*, yahoo.com) is the part of a URL that is the human readable representation of an IP address. **Agglomerative Hierarchical Clustering Algorithm :** Two clusters are merged if they contain emails which point to the same domain. The agglomerative clustering method is desirable because only in the first iteration, the runtime of the algorithm is a function of the number of emails, but starting from the second iteration, the runtime is a function of the number of previous clusters, which is constantly reducing. The weakness of the method is that coincidence, common phrases and sheer luck can cause untrustworthy relationships to be introduced since our logic is that two emails are linked as long as they share at least one common attribute. To counter false-positives, a connected component with weighted edge method is introduced in the next section to break the biggest cluster into smaller clusters.

B] Connected Components with Weighted Edges

To eliminate chance conjoining of unrelated spam campaigns into the same cluster, the concept of "connected component of weighted edges" was applied A *connected component* in an (undirected) graph is a set *S* of vertices such that for every vertex *v* of *S*, the set of vertices reachable (by paths) from *v* is precisely *S*. The weight of an edge shows the strength of the connection between the two vertices. The goal is to find connected components of this graph, considering only edges with weight above a threshold. This goal stems from the following reasoning: Suppose a spammer owns 10 domains and has a list of 10 subjects, and he sends out emails by randomly picking a subject and a domain. There are totally 100 possible combinations. If he sends out enough emails and we have enough collection of his emails, we should see examples of all 100 combinations. So if domains are assigned as vertices and subjects as edges, we will evidently find that the ten domains are tightly connected to each other with strong edges. On the other hand, if two domains are owned by two different spammers and they are connected to each other by chance because the two spammers share a common subject, the connection between domains, in this case, will be weak since the probability of two spammers picking the same subject is relatively lower. If a group of domains in the biggest cluster are tightly connected to each other, they are very likely to be owned by the same spammer. Therefore, all domains from the biggest cluster are retrieved and assigned as vertices. The edges connecting them will be any common subject and the weight of the edge is the number of common subjects shared by two domains. A threshold is then selected and all edges with weight below that threshold will be dropped. The remaining connected components should be tightly related. The algorithm is designed to allow the threshold be adjusted to produce a more favorable result. By applying the algorithm to a cluster that has false positives, the cluster is divided into smaller clusters that are more tightly related. If the result still shows too many false positives in our sub-clusters, the threshold will be incremented.

Future Scope:

In future the above system is extended to examine all databases which are going to Forensic Investigation and develop the new data mining technique to easily manage exam data.

4. CONCLUSION:

This paper has proposed a new approach to analyze spam emails with a focus on the needs of law enforcement personnel. Initial extract email attribute and with an application of AI in computer forensics and the latest results obtained with the use of the EMADIK, a MAS to assist the experts during computer forensics examinations with that save investigation time and improve the system approach. Then the data mining technique creates clusters of related emails which can easily be assessed for their validity. The resulting clusters have been primarily related to spam messages which are trying to encourage the purchase of a product or service. Clusters of spam used for spreading viruses through attachments, or spam which sends visitors to hacked websites for purposes of phishing or other fraud were not readily identified using the current method.

5. REFERENCES:

- Artificial Intelligence Applied to Computer Forensics by Bruno W. P. Hoelz, Célia Ghedini Ralha, and Rajiv Geeverghese.
- Mining Spam Email to Identify Common Origins for Forensic Application by Chun Wei, Alan Sprague
- Nicole Beebe and Jan Guynes Clark. A hierarchical, objectives based framework for the digital investigations process.
- Fabio Luigi Bellifemine, Giovanni Caire, and Dominic Greenwood. Developing Multi-Agent Systems with JADE.
- Wiley Series in Agent Technology, Sussex, England. Mining Spam Email to Identify Common Origins for Forensic
- Application by Chun Wei, Alan Sprague, Gary Warner, and Anthony Skjellum.
- Airoldi, E. and Malin, B.ScamSlam: Architecture for Learning the Criminal Relations behind Scam Spam.
- Mark d'Inverno and Michael Luck. *Understanding Agent Systems*. Springer Series in Agent Technology, Berlin, Germany, 2*nd* revised and extended edition,
- George F. Luger. *Artificial Intelligence: Structures and Strategies for Complex Problem Solving*. Addison-Wesley, USA, 4*th* edition, 2002.
- Simson L. Garfinkel. Forensic feature extraction and cross-drive analysis. *Digital Investigation*.
- Han, J. and Kamber, M. Data Mining: Concepts and Techniques. (2nd ed.). Morgan Kaufmann, San Francisco.
- S. Pinson and P. MoraÄ³tis. An intelligent distributed system for strategic decision making.