An Assessment of Fuzzy Temporal Event Correlation towards Cyber Crime Investigation

Sourabh Jain
M.Tech Student NRI College,Bhopal
Email ID –sourabhjn755@gmail.com

Shatendra Dubey
Asst. Prof. NRI College,Bhopal
Email ID –shatendradubey@gmail.com

Abstract

Event logging and event logs play an important role in modern IT systems criminal investigation which is generated when end user with each other in web environment and stored in various logs like firewall log file at side ,network log file at gateway and web log file at server side. But log file is not to be over emphasized as a source of information in systems and network management. Whereas conduct efficient investigation and gathering of use full information need to correlate different log file. Task of analyzing event log files with the ever-increasing size and complexity of today's event logs has become cumbrous to carry out manually. Nowadays latest spotlighted is automatic analysis of these logs files. . This paper present an bird eye on two basic concepts one is temporal data mining and another is fuzzy association rules. Using log files it is possible to classify the attacker from the normal user.

Keywords- Event Logging, Fuzzy Logic, Temporal Correlation

I INTRODUCTION

Event logging and event logs play an [1] important role in modern IT systems. Today, many applications, operating systems, network devices, and other system components are able to log their events to a local or remote log server. For this reason, event logs are an excellent source for determining the health status of the system, and a number of tools have been developed over the past 10-15 years for monitoring event logs in real-time. However, majority of these tools can accomplish simple tasks only, e.g., raise an alarm immediately after a fault message has been appended to a log file. On the other hand, quite many essential event processing tasks involve event correlation a conceptual interpretation procedure where new meaning is assigned to a set of events that happen within a predefined time interval. Event correlation is one of the most prominent real-time event processing techniques today. It has received a lot of attention in the context of network fault management over the past decade, and is becoming increasingly important in other domains as well, including event log monitoring. A number of approaches have been proposed for event correlation, and a number of event correlation products are available. Unfortunately, existing products are mostly expensive, platform-dependent, and heavyweight solutions that have complicated design, being therefore difficult to deploy and maintain, and requiring extensive user training. For these reasons, they are often unsuitable for employment in smaller IT systems and on network nodes with limited computing resources. So far, the rule-based approach has been frequently used for monitoring event logs – event processing tasks are specified by the human analyst as a set of rules, where each rule has the form IF condition THEN action[2]. For example, the analyst could define a number of message patterns in the regular expression language, and configure the monitoring tool to send an SMS notification when a message that matches one of the patterns is appended to the event log. Despite its popularity, the rule-based approach has nevertheless some weaknesses – since the analyst specifies rules by hand using his/her past experience, it is impossible to develop rules for the cases that are not yet known to the analyst; also, finding an analyst with a solid amount of knowledge about the system is usually a difficult task. In order to overcome these weaknesses, various knowledge discovery techniques have been employed for event logs, with data mining methods being a common choice [3]. It should be noted that while event log monitoring tools conduct on-line (real-time) analysis of event log data, data mining methods are designed for off-line analysis an existing event log data set is processed for discovering new knowledge, with the data set remaining constant throughout the discovery process.

HTTP & TCP METHOD

Both protocols are responsible for the data communication from source node to destination node. In browser the http protocol works as application layer protocol. Computer store all the information regarding access the server store in log file. At that time which protocol has used during the communication will note in the log files.

LOG AND FUZZY LOGIC

Log file that have all the entry related to incoming user and outgoing user. These file are generated by the process of installation. It can maintain by server machine, firewall, web servers, and routers etc. Generally the log files are in the text format can be read by notepad or simple text editor. Due to the plain text the size of log file will also reduces

Group of items having similar sort of properly is known as set. These set items are the elements of set. This is a traditional approach to represent the set theory. There are some problems with the traditional approach of set theory. The set theory is used for the specific data set. It always gives the answer in Yes or No. some time it seems to be that it is not practically suitable. To remove such types of complexity fuzzy set theory comes in existence.

II TEMPORAL DATA MINING

Time and space are the two basic category of data mining. The temporal [4] data mining is a newly data mining approach with respect to time. Some time it used the temporal data base. There are basically three types of time used in this approach these are valid time, transaction time and Bi-temporal. The valid time shows the duration in which the event was performed in province of real world. The transaction time give the detail about the duration in which the event was saved in the records. The combination of both time periods is known as a Bi-Temporal.

Extraction of Temporal Association Patterns for Temporal data mining has been productively applied in number of fields including trading, marketing, social analysis, medical, fraud detection, robotics and assisted design Because of that explorer’s number of efficient algorithms for temporal data model like symbolic time series, symbolic time sequences, symbolic interval series, numeric time series, item set sequences, etc have been proposed.

III FUZZY ASSOCIATION RULES

Classic set theory was unable to solve many types of problem where there is a probability [5] to solve such sorts of problem fuzzy logic comes in existence. Fuzzy set theory has been used more and more frequently in data mining because of its simplicity and similarity to human reasoning. The rules which associate two or more attributes called association rules. These are the set of rules. The most popular use of association rules are the classification of data.

The fuzzy association rules are used for the classification of data. These rules can applicable where there is some probability. This method is efficient to get the results from the boundary cut problems. Fuzzy rule also can apply on the genetic selections.

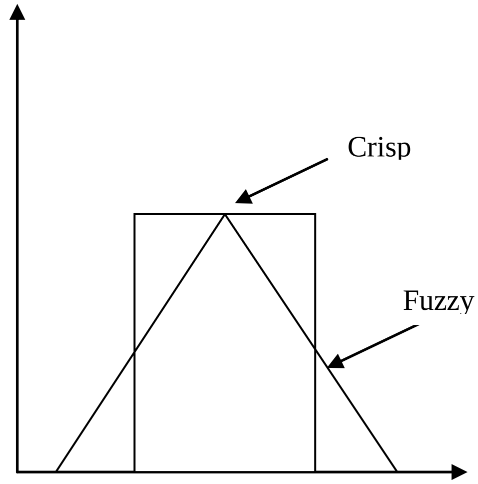

Fig 1. Boundary of fuzzy set

Fuzzy Association Rules mining is a novel approach based on classical association rule mining. Whenever we have a data set having a certain range of values then we might face the sharp boundary problem.

Suppose we have three range of age.

F(x) is a function such that

0<x<=30 then f(x) = younger

30<x<=45 then f(x) = middle ager

x>45 then f(x) = older

However in this example, a person aged 44 years would be middle age and a 46 year old would be older where as in reality, the difference between those ages is not that great. Therefore, here is a problem of sharp boundary. As shown in figure.

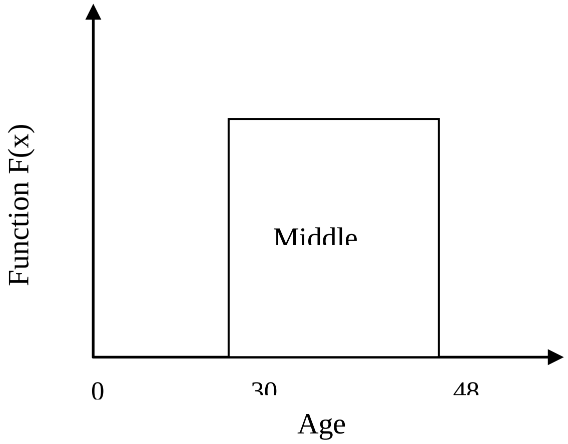

Fig 2. Example of sharp boundary problem

To resolve the sharp boundary problem by using Quantitative approach divide the variable Age into three fuzzy sets. The fuzzy sets and their membership functions will have to be defined by a domain expert. For easy demonstration, we will just define the borders of the sets and split the overlapping part equally between the so generated fuzzy sets. For an example, we will use the following borders for the fuzzy sets of the variable age: Age. Low= {0−33}, Age. Medium= {27−48}, Age. High= {42−∞}. The generated fuzzy sets are shown in Figure 1. For all areas having no overlap of the sets, the support will simply be 1 for the actual item set. If there is an overlap, the membership can be computed by using the borders of the overlapping fuzzy sets. The added support will here always sum up to 1.

IV BACKGROUND STUDIES

This section gives an extensive literature survey on the existing methodology of log correlation. We studied various research paper and journal and find-out some research work carried out by various researchers in the field of log files with various techniques of log correlations and discussed. All methodology and process are not described here. But some related work in the field has discussed by the name of authors and their respective title.

Network administrators are able to correlate log file entries manually [6] Large volume and low quality of log files justify the need for further log processing. The manual log processing is lack of flexibility. It is time consuming, and one doesn't get the general view of the log files in the network. Without this general view it is hard to correlate information between the network components. Events seemingly unessential by themselves can in reality be a piece of a larger threat. In this regard, different log co relation methods are proposed to improve alert quality and to give a comprehensive view of system security. In this paper, the authors show that how different attacks categorized in three categories with different behavior: Denial of service (DoS) attacks, user-to-root (U2R) & remote-to-local (R2L) attacks and probing, are reflected in different logs and argue that some attacks are not evident when a single log is analyzed. The taxonomy of the large ds DNA viruses [7] has been provided in the VIIIth report of ICTV. The phylogenetic tree of large dsDNA viruses has been constructed using CVTree method (Gao and Qi, BMC Evol. Biol.7(2007)41). In this paper, we use the log-correlation distance method analyze the complete genome of the 124 large dsDNA viruses and construct phylogenetic trees based on compositional vectors of DNA sequences or protein sequences. The phylogenetic trees show the large dsDNA virus genomes are separated into nine families. The structures of the trees based on log-correlation distance are mostly consistent with the result of CVTree method and the taxonomy of the VIIIth report of ICTV.Monitoring systems observe [8] important information that could be a valuable resource to malicious users: attackers can use the knowledge of topology information, application logs, or configuration data to target attacks and make them hard to detect. The increasing need for correlating information across distributed systems to better detect potential attacks and to meet regulatory requirements can potentially exacerbate the problem if the monitoring is centralized. A single zero-day vulnerability would permit an attacker to access all information. This paper introduces a novel algorithm for performing policy-based security monitoring. We use policies to distribute information across several hosts, so that any host compromise has limited impact on the confidentiality of the data about the overall system. Experiments show that our solution spreads information uniformly across distributed monitoring hosts and forces attackers to perform multiple actions to acquire important data.Computer forensics [9] searches for evidence to reassemble the actions that led the system from a secure state to the moment an intrusion was detected. The main source of data for a forensic investigation is the information provided by log files. Log files are generated by applications to keep a register of the actions occurred on the system. However, the massive amount of recorded events complicates the forensic investigation. A model composed by a set of agents in order to collect, filter, normalize, and to correlate events coming from diverse log files is proposed in this paper. The purpose of the model is to assist the analyst in the evidence search process of a forensic investigation.The rapidly evolving society, every [10] corporation is trying to improve its competitiveness by refactoring and improving some if not all of its industrial software infrastructure. This goes from mainframe applications that actually handle the company's profit generating material, to the internal desktop applications used to manage these application servers. These applications often have extended activity logging features that notify the administrators of every event encounter at runtime. Unfortunately, the standalone nature of the event logging sources renders the correlation of log event infrastructure prone to continuous queries. This paper described an approach that adapts and employs continues queries for distributed log event correlation with the aim to solve problems that face the present log event management systems. It will present LEC architecture that analyze a set of distributed log events that follow a set of correlation rules; then the main output is a stream of correlated log events. This paper [11] presented a set of innovative algorithms and a system, named Log Master, for mining correlations of events that have multiple attributions, i.e., node ID, application ID, event type, and event severity, in logs of large-scale cloud and HPC systems. Different from traditional transactional data, e.g., supermarket purchases, system logs have their unique characteristics, and hence the authors proposed several innovative approaches to mining their correlations. The authors parsed logs into an n-ary sequence where each event is identified by an informative nine-tuple. The authors proposed a set of enhanced