

# Data Mining Approach in Security Information and Event Management

Anita Rajendra Zope, Amarsinh Vidhate, and Naresh Harale

**Abstract**—This paper gives an overview of data mining field & security information event management system. We will see how various data mining techniques can be used in security information and event management system to enhance the capabilities of the system.

**Index Terms**—Data mining, security information event management system.

## I. INTRODUCTION

Security information and event management system is the industry-specific term in computer security referring to the collection of data typically log files or event logs from various sources into a central repository for analysis. Event logs are generated by various networking devices, Operating Systems and Application Servers. Event logs give raw input of all activity happening in IT infrastructure of any organization. This raw data act like input to SIEM system which provides us security alerts, reports as an output. The processing of all raw data is achieved using data mining technique.

Data mining derives its name from the similarities between searching for gold in mines. In gold mines we search for very small particles of gold in tons of soil. Similarly in data mining we search for valuable information from huge amount of data collected in various ways. Data mining, a synonym to “knowledge discovery in databases” is a process of analyzing data from different perspectives and summarizing it into useful information. It is a process that allows users to understand the substance of relationships between data. It reveals patterns and trends that are hidden among the data. It is often viewed as a process of extracting valid, previously unknown, non-trivial and useful information from large databases [1]. Data mining is becoming increasingly common in both the private and public sectors. Industries such as banking, insurance, medicine, and retailing commonly use data mining to reduce costs, enhance research, and increase sales [2]. If scope of data mining is applied to all events logs generated by various networking devices, system and application servers then efficiency of enterprise security can be drastically increased.

The real problem in today’s enterprise security is amount of logs generated by various systems. Organizations often put too much faith in their new shiny firewalls, IDSs, or antivirus

software. Once one or more of these solutions are implemented then IT staff realizes that interpretation of all logs generated by this solution is big challenge. A network could either perform as a well-tuned orchestra or as several pieces that play wonderfully by themselves but give you a headache when they are all brought into the same room. Each individual security component could be doing its job by protecting its piece of the network, but the security function may be lost when it is time to interrelate or communicate with another security component. SIEM system helps us to take an architectural view, where we can look at the data flow in and out of the environment, how this data is being accessed, modified, and monitored at different points, and how all the security solutions relate to each other in different situations.

## II. DATA MINING BASICS

Our capabilities of both generating and collecting data have been increasing rapidly. The widespread use of bar codes for most commercial products, the computerization of many business and government transactions, and the advances in data collection tools have provided us with huge amounts of data. Millions of databases have been used in business management, government administration, scientific and engineering data management, and many other applications. It is noted that the number of such databases keeps growing rapidly because of the availability of powerful and affordable database systems. This explosive growth in data and databases has generated an urgent need for new techniques and tools that can intelligently and automatically transform the processed data into useful information and knowledge. Consequently, data mining has become a research area with increasing importance [3]. We need information but what we have is a huge amount of data flooding around. Because of the amount of data is so enormous that human cannot process it fast enough to get the information out of it at the right time, the data mining technology has been established to solve this problem potentially. The ultimate goal of knowledge discovery and data mining process is to find the patterns that are hidden among the huge sets of data and interpret them to useful knowledge and information

### A. Typical Data Mining Architecture

Based on storage & retrieval of data following architecture is possible

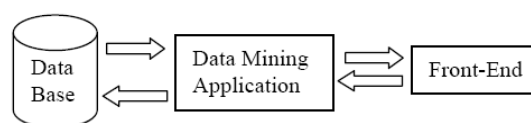


Fig. 1. Data mining architecture.

There are three tiers of data mining Architecture

Manuscript received September 16, 2012; revised October 27, 2012.  
Anita Rajendra Zope and Naresh Harale are with the M. G. M College Of Engineering and Technology, Navi Mumbai, India (e-mail: anitazope@yahoo.com, nareshkumar.d.harale@gmail.com).

Amarsinh Vidhate is with Ramrao Adik Institute of Technology, Navi Mumbai, India (vidhate.amarsinh@gmail.com).

- 1) *Data layer* As mentioned above, data layer can be database and/or data warehouse systems. This layer is an interface for all data sources. Data mining results are stored in data layer so it can be presented to end-user in form of reports or other kind of visualization.
- 2) *Data mining application layer* This layer is used to retrieve data from database. Some transformation routine can be performed here to transform data into desired format. Then data is processed using various data mining algorithms.
- 3) *Front-end layer* This layer provides intuitive and friendly user interface for end-user to interact with data mining system. Data mining result presented in visualization form to the user in the front-end layer

### B. Data Mining Techniques

There are several major data mining techniques have been developed and used in data mining projects recently including association, classification, clustering, prediction and sequential patterns.

Following tables gives an idea about various data mining techniques.

TABLE I: VARIOUS DATA MINING TECHNIQUE

Techniques Name	Function
Association	a pattern is discovered based on a relationship of a particular item on other items in the same transaction
Classification	Classify each item in a set of data into one of predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics
Clustering	Makes meaningful or useful cluster of objects that have similar characteristic using automatic technique.
Prediction	Discovers relationship between dependent and independent variables
Sequential Patterns	Discover similar patterns in data transaction over a period

### III. DATA MINING FOR SECURITY APPLICATIONS

In this section we will understand what is role of data mining in security information & event management system.

TABLE II: VARIOUS NETWORK SECURITY DEVICES & THEIR FUNCTION

Device Name	Function
Firewall	permit or deny network transmissions based upon a set of rules and is frequently used to protect networks from unauthorized access
Network & Host Intrusion prevention system (NIPS,HIPS)	monitors network and/or system activities for malicious activities
Mail Gateway	Used to detect & prevent spam mails & unwanted software attached in emails
Web Gateway	Perform URL filtering & block malicious sites , provides proxy function
AAA system	handles user requests for access to computer resources and, for an enterprise, & provides authentication, authorization, and accounting
Data Leakage Prevention system	Systems that enable organizations to reduce the corporate risk of the unintentional disclosure of confidential information.
Vulnerability Assessment Tools	Find vulnerability in operating system, application, Data base server

In today's world every IT administrator staff has to deal with millions of events. These events are generated by various devices for example the staff in Network Operation Center (NOC) has to analyze events generated by networking devices like routers, Switches, load balancer similarly staff in Security Operation Center (SOC) has to analyze events generated by security devices like firewall, IPS, AAA server. The management body of every organization wants administrator to analyze each & every events. This gives burden to staff & likely chances to miss critical events which increase threats to entire organization.

A defense in depth strategy (industry best practice) utilizes multiple security devices. Each device has specific function for which they deploy in IT infrastructure. As a part of risk management many organization generally deploy following devices in their IT infrastructure.

Following tables list various security devices and their functions.

Any organization dealing with all these security devices face problem in monitoring all such events. This force the origination to increase security analysts posts. This huge amount of events creates following problem in any organization

- 1) Security administrator has to manage all devices & analyze the events generated by these devices which increase the work load
- 2) Efficiency decrease by spending long time in finding false alarms.

In order to reduce the number of security events on any given day to a manageable, actionable list and to automate analysis such that real attacks and intruders can be discerned we should apply data mining technique to all such events. If we want see holistic view for enterprise security then we do mining on all security & network events.

### IV. OVERVIEW OF SIEM

SIEM system uses most aspects of data mining. It basically collects logs from various devices normalized the logs & store in data base. On all such data correlation rules are applied to get meaningful information. [4]

#### A. What SIEM can Provide?

SIEM system when properly configured has capacity to become central nervous system of network. SIEM can do real time monitoring & incident management for security related events which are collected from network, security devices, system, applications. It can also used as log management & compliance reporting.

#### B. Functions of SIEM

With some subtle differences, there are four major functions of SIEM solutions system

- 1) Log Consolidation – centralized logging to a server
- 2) Threat Correlation – the artificial intelligence used to sort through multiple logs and log entries to identify attackers
- 3) Workflow – Helps to track and escalate the incident
- 4) Reporting – Gives enterprise reporting for compliance purpose.

### C. SIEM Architecture

SIEM system uses data feeds from various devices.

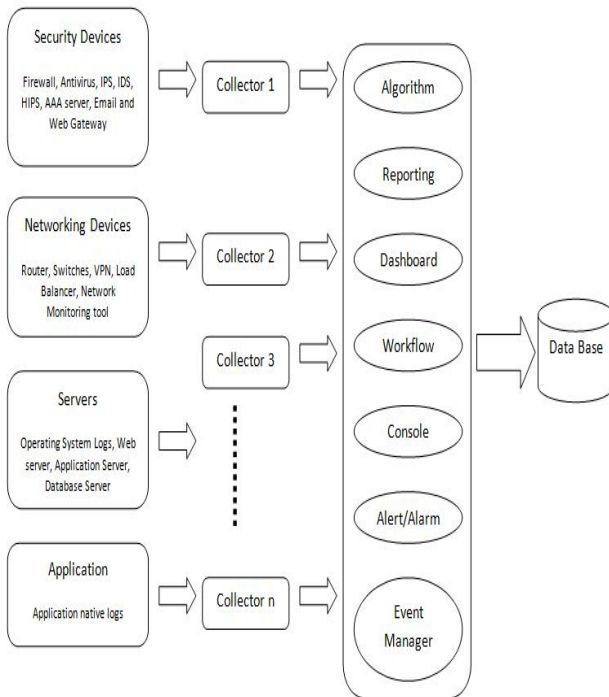


Fig. 2. SIEM architecture

SIEM Architecture has four major parts:

- 1) **Data Sources** : SIEM system gets data feed from various devices which not only include networking devices but also some physical security devices like bio metric devices, card readers.
- 2) **Data Collectors**: primary function of data collector is to do normalization. This normalization happens in two ways it first normalize the values such as time zone, priority, severity in to common format, then they normalize the data structure in to common format. Some time collector do aggregation for example if there are 5 similar events in less than 3 second then collector can send only one such event. This filtering increases efficiency and accuracy and reduce processing time.
- 3) **Central Engine**: This is heart of SIEM system which mainly does applying data mining algorithm. This engine writes events in to database as they stream into the system. It simultaneously processes them through data mining engine where correlation happens. It also has user interface to display result of data mining algorithm. It enables end user to change certain properties of algorithm. Some of other component of this engine is reporting, alerting, and dashboards.
- 4) **Data Base**: As events stream in to central engine they are written in database with normalized schema. This storage helps us to do forensic analysis on historic data. By storing the events we can test new algorithm on historic data.

### V. DATA MINING TECHNIQUES IN SIEM

In this section we will understand various data mining algorithm which can be used in SIEM

#### A. Basic Concept of Association Rules and its use in SIEM

Association rule mining discovers frequent patterns, associations, correlations, or causal structures among huge groups of items or objects in transaction databases, relational databases, and other information repositories [5]. The brief of the basic concept of association rules in transactional and relational databases is presented as follows:

$I = \{i_1, i_2, i_3, \dots, i_n\}$  is a set of items, Let DB be a set of database transactions where every transaction  $T$  is a set of items such that  $T \subseteq I$ . Each transaction is associated with a unique, transaction identifier (TID).

Let  $X, Y$  be a set of items, an association rule is an inference of the form  $X \rightarrow Y$  where  $X \subset I, Y \subset I$ , and  $X \cap Y = \emptyset$ .  $X$  is called the antecedent of the rule, and  $Y$  is called the consequent of the rule. An itemset containing  $i$  itemset called an  $i$ -itemset. The rule  $X \rightarrow Y$  holds in the transaction set  $D$  with Support  $S$ , which is the percentage of transactions that contain both an itemset  $X$  and  $Y$  appearing in the same transaction. For an itemset to be interesting, its support must be higher than a user-specified minimum. Such item sets are said to be frequent. We have the Support of the rule  $X \rightarrow Y$  as

$$\text{Support}(X \rightarrow Y) = P(X \cup Y)$$

There is another measure called Confidence  $C$ , where  $C$  is the ratio of the number of transactions in  $D$  that containing  $X$  and  $Y$  to the number of transactions that contain only  $X$  as Equation

$$\begin{aligned} \text{Confidence}(X \rightarrow Y) &= P(Y|X) \\ &= \text{Support}(X \cup Y) / \text{Support}(X) \end{aligned}$$

Association rule mining is the process of finding all the association rules that pass the condition of min support and min confidence. In order to mine these rules, first the support and confidence values have to be computed for all of the rules and then compare them with the threshold values to prune the rules with low values of either support or confidence. In general association rule can be summarized in two steps [5]

- 1) Find the large item sets, i.e., the sets of items that have transaction support above a predetermined minimum threshold.
- 2) Use the large item sets to produce the association rules for the database that has confidence above a predetermined minimum threshold.

#### B. Association Rules Techniques used in SIEM

In association rules there are many techniques. In this paper we chose the most popular techniques that can be utilized in SIEM

##### 1) Apriori algorithm Introduced by Agrawal [6]

Apriori is the most important algorithm for mining frequent item sets for Boolean association rules in a given database. It uses the property: all nonempty subsets of a frequent itemset must also be frequent [7]. The key idea of Apriori algorithm is to make multiple passes over the database. It employs an iterative approach known as a breadth-first search (level-wise search) through the search space, where  $k$ -item sets are used to explore  $(k+1)$ - item sets. In the beginning, the set of frequent 1-itemsets is found. The set contains one item, satisfying the support threshold and, is

denoted by L1. In each subsequent pass, we begin with a seed set of item sets found to be large in the previous pass. This seed set is used to develop new potentially large item sets, called candidate item sets, and to count the actual support for these candidate item sets during the pass over the data. At the end of the pass, we decide which of the candidate item sets are actually large (frequent), and they become the seed for the next pass. Therefore, L1 is used to find L2, the set of frequent 2- item sets, which is used to find L3, and so on, until no more frequent k-item sets can be found. Then, a very significant property called Apriori property is employed to reduce the search space. Expressly, the Apriori algorithm consists of join and prune steps [7].

2) *Frequent Pattern Growth algorithm*

FP-growth algorithm is an efficient method of mining all frequent item sets without candidate generation. FP-growth utilizes a combination of vertical and horizontal database layouts to store the database in main memory. Instead of storing the cover for every item in the database, it stores the actual transactions from the database in a tree structure and every item has a linked list going through all transactions that contain that item. This new data structure is denoted by FP-tree [7]. Essentially, all transactions are stored in a tree data structure. Every node additionally stores a counter, which keeps track of the number of transactions that share the branch through that node. In addition, a link is stored, pointing to the next occurrence of the respective item in the FP-tree, such that all occurrences of an item in the FP-tree are linked together. Furthermore, a header table is stored containing each separate item together with its support and a link to the first occurrence of the item in the FP-tree. In the FP-tree, all items are ordered in support descending order, as it is hoped that this representation of the database is kept as small as possible since all of the more frequently occurring items are arranged closer to the root of the FP-tree and thus are more likely to be shared.

3) *Sequential association rule mining algorithm [8].*

Given a "sequence database" of past event sequences to learn from, we aim to predict the next event within a current event sequence. The "support" of an itemset is the number of times that the itemset has appeared in the sequence database. Typically in association rule mining, a strict minimum support threshold condition is placed on the support of itemsets within a rule, so that rules falling below the minimum support threshold are simply discarded. There are 2 algorithms written for sequential association rule. Let  $X$  be set of possible itemset and  $S$  be past history of data set which is collection of  $m$  item set.  $S$  is the sequence database  $S = \{Z_i\}_{i=1,2,\dots,m}, Z_i \subseteq X$   $B$  is current sequence itemset. An algorithm uses  $B$  and  $S$  to find rules  $a \rightarrow b$  where  $a$  is in item of current of past history itemset and  $b$  is not in the current item set or this is predicted value

Algorithm 1: Subroutine GenRules.

Input is  $(S, B, X)$ , that is, past itemset  $S = \{Z_i\}_{i=1,2,\dots,m}, Z_i \subseteq X$ , Current item  $B \subseteq X$ , set of item itemset  $X$ . Output is a set of all rules  $\{a_j \rightarrow b_j\}_j$  where  $b_j$  is a single item that is not in the basket  $B$ , and where  $a_j$  either a subset of items in the basket  $B$  is, or else it is the empty set. Also the left-hand side  $a_j$  must be allowed (meaning it is in

A). That is, output rules  $\{a_j \rightarrow b_j\}_j$  such that  $b_j \in X \setminus B$  and  $a_j \subseteq B \subseteq X$  With  $a_j \in A$  or  $a_j = \emptyset$

Algorithm 2: Max Confidence, Min Support Algorithm

Input is  $(\emptyset, X, S, B, GenRules, c)$  that is, minimum threshold parameter  $\emptyset$  set of item  $X$ , past orders  $S = \{Z_i\}_{i=1,2,\dots,m}, Z_i \subseteq X$ , current basket  $B \subseteq X$ , Gen Rule generates candidate rule  $GenRule(S, B, X) = \{a_j \rightarrow b_j\}_j$  number of recommendations  $c \geq 1$

Output: Recommendation List, which is a subset of  $c$  items in  $X$

VI. IMPLEMENTATION OVERVIEW

The process of the association analysis can be divided into three parts usually

- Filtering redundant information and formatting the security information.
- Matching the association rules.
- Generating security events.

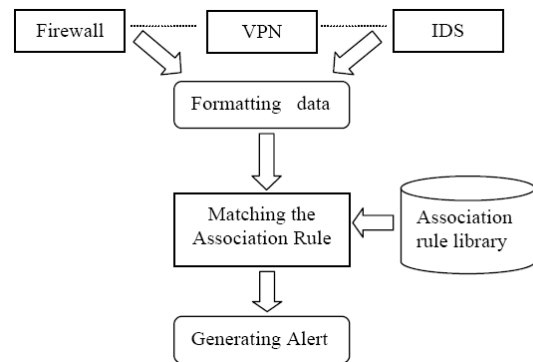


Fig. 3. Framework of the association analysis

First two techniques can be used in SIEM for detection of anomaly. We can call this as *anomaly association rules*. This rule will work only if we define threshold. Anomaly detection is calculated by comparing the rules of normal category dataset with the rules of real traffic category datasets based on similarity measurements. If the similarity result is higher than the user threshold, it means that the dataset has no intrusions, and vice versa. The normal category dataset is reference data, which should not have intrusions. To apply this technique first data has to convert in to datasets for example if we have all packets captures then data sets can be prepared based on SYN, FIN and RST types of TCP flags, number of source IP addresses, number of destination IP addresses, and the total size of packets. We can write some process which will convert all real time continuous data to categorical data. Once the required data set is ready we can apply this algorithm.

Third technique can be used on IDS events the rule will look like  $\{\text{ping sweep, port scan}\} \Rightarrow \{\text{exploit vulnerability}\}$  this rule indicate that if we can see ping scan from single IP address to multiple IP address & port scan event from same IP address then it is likely that same IP address will try to exploit any known vulnerability on open port.

VII. CONCLUSION

This study shows that how data mining can be used in

SIEM system. This paper firstly introduces the related knowledge, architecture of SIEM system and then the rule of algorithm for the correlation analysis. We have seen various association rules to detect abnormal patterns.

One of the areas we are exploring for future research is how we can use other data mining technique like classification, clustering to enhance the system capacity. In addition, we are enhancing the techniques we have mentioned to reduce false positive alerts and to reduce CPU load on system while computing data mining rules. Furthermore we are working to contribute some new modules for open source SIEM project.

#### REFERENCES

- [1] I. K. R. Rao, "Data Mining and Clustering Techniques," *DRTC Workshop on Semantic Web*, DRTC, Bangalore, paperk, pp. 1-1, 8th – 10th December, 2003.
- [2] J. W. Seifert, "Data Mining and Homeland Security: An Overview," *CRS Report*, pp. 1-1, Jan. 2007.
- [3] M. S. Chen and J. H. Philip, "Data Mining: An Overview from a Database Perspective," *IEEE Trans on knowledge and data engineering*, vol. 8, no. 6, pp. 1-1, Dec 1996.
- [4] S. Yuan and C. Zou, "The Security Operations Center Based on Correlation Analysis."
- [5] E. E. Eljadi and Z. A. Othman, "Anomaly Detection for PTM's Network Traffic Using Association Rule," in *Proc. of 2011 3rd Conference on Data Mining and Optimization DMO*, June 2011
- [6] R. Agrawal and Srikant, "Fast Algorithms for Mining Association Rules," in *Proceeding of the 20th VLDB Conference Santiago, 1994*

- [7] J. Han and M. Kamber, "Data Mining Concepts and Techniques. Second Edition," *The Morgan Kaufmann Series in Data Management Systems*.
- [8] C. Rudin, B. Letham, A. S. Aouissi, E. Kogan, and D. Madigan, *Sequential Event Prediction with Association Rules*.



member.

**Mrs. Anita Rajendra Zope** is currently Pursuing Master of Engineering in computer science from Mumbai University in M.G.M college of Engineering and Technology, Navi Mumbai. Received Bachelor of Engineering from Pune University in P.E.S Modern College Of Engineering. Currently working in St.John College of engineering, Palghar , Mumbai University. Published one national and one international paper. Mrs. Anita Rajendra Zope. IEEE



resource management

**Amarsinh Vidhate** received M. Tech in Computer Engg. From Dr. Babasaheb Ambedkar Technological University in 2003. He is currently pursuing PhD from NMIMS University. He is working as a Assistant Professor in Dept. of Computer Engg at Ramrao Adik Institute of Technology, Navi Mumbai. He has almost 18 international and national conference as well as journal papers to his credit. His research interest includes wireless mobile networking, mobility and