Research on The Network Security Management Based on Data Mining

Lin Li
Department of Computer Science
HuaZhong Normal University
Wuhan, P. R. China
liling1221@126.com

De-bao Xiao
Department of Computer Science
HuaZhong Normal University
Wuhan, P. R. China
dbxiao0227@vip.sina.com

***Abstract*—Against the defect that the traditional computer network security management system processes a mass of data with low efficiency and accuracy, a novel network security management model based on data mining is proposed in this paper. It makes use of multi-source data collection strategy to acquisition and integrate the relevant data of different security products, and takes advantage of data mining technology for massive data to analyze them comprehensively and intelligently and response automatically. The experimental result reveals that the model performs well. It can identify the real attacks from a large number of security events with a good performance, so that the alarm information will be refined with low false or wrong alarms, the accuracy, intelligence and adaptability of network security management will be enhanced and the requirements of new security situation will be met.**

Keywords-Network Security Management; data mining; intelligent analysis

I. INTRODUCTION

As the attacks to computers and networks are becoming multi-type, complex and intellective, it is difficult to go along with the needs of network security only depending on the static defense technologies, such as data information encryption and firewall. In addition, there are a lot of redundant or unreliable information in increasing security events. Only mining the real attacks from these abundant and complicated security events can make a reasonable assessment and correct responsibilities on the network security. It therefore turns into an active subject for the research on network management and security technology to design and realize a unified and intellective security management system in different environment to integrate and mine large amounts of data. Based on this, a security management network model based on data mining is designed in this study. It adopts a variety of information collection strategies for the collection and integration of the relevant data of all kinds of security products and data mining technologies for multidimensional analysis which implements the automatic comprehensive analysis on the alarm events from various security products, and then submits the mining results to the analyst in a visual way. The experiment results show that the model is superior to the existing network security management system in compression rate, false alarm rate, construction rate and scene detection rate, and significantly improves the accuracy, intelligence and adaptability of the network security management.

The rest of this paper is organized as follows. Section 2 introduces the related works of network security management by leveraging data mining. A novel network security management model based on data mining is formulated in detail in section 3. Experimental results are reported in section 4. The final section summarizes the conclusions.

II. RELATED WORKS

Exploiting the technologies of different fields such as database, statistics, artificial intelligence and machine learning, data mining can intelligently analyze the sea of data effectively to mine unknown, useful patterns or knowledge. Network security management can get several times data as much as common security products because of its comprehensive security services. Some researchers have put forward an idea that apply data mining technologies into network security management, in view of its foregoing advantages, to make full use of its superiority on processing an enormous amount of data to discover unknown knowledge and regular patterns, improving the efficiency and accuracy of security detection. In 1998, Lee and Stolfo [1] first adopted data mining technologies in Intrusion Detection and proved its feasibility in theory and technology. Later, they constructed the first misuse detection system based on data mining—JAM system, in 2000 [2], [3]. Shin et al [4] utilized it in strategy-based network security management, offered an alarm analyzer using data mining mechanism which consists of association rule digger, frequent item digger and clustering digger, and finally realized the effective support of alarm analysis and high-level analysis mining systems to security strategy management. Cabrera et al [5] used it to detect the DDOS attacks in MIB database and gained a good detecting result. Wang et al [6] suggested a deep system intrusion avoid model based on data mining to manage a great many unreliable and unmanageable security events. It made full use of the online detection and the off-line mining to manage network system. Tavallaee et al [7] put the classification technology of data mining in the online classification of network data stream to find out the data stream in danger. All these researches introduced data mining into network security management and improved the

978-1-4244-6542-2/$26.00

efficiency and accuracy of network management extremely. They, however, didn't analyze the alarm events of various security products comprehensively and didn't analyze the comprehensive decision with multiple methods as well, which resulted in the security problems of network management solved incompletely.

III. NEKWORK SECURITY MANAGEMENT MODEL BASED ON DATA MINING

This paper proposes a network security model based on data mining as shown in Figure 1. Its basic idea is to integrate the original data of the all network security products, then analyze and response them automatically and intelligently with the advantages of data mining for a mass of data — it releases the manager from massive alarm data and arduous tasks of security management. Then the corresponding security strategy is made and carried out according to the mining results to enhance the security protection capability of the whole network.

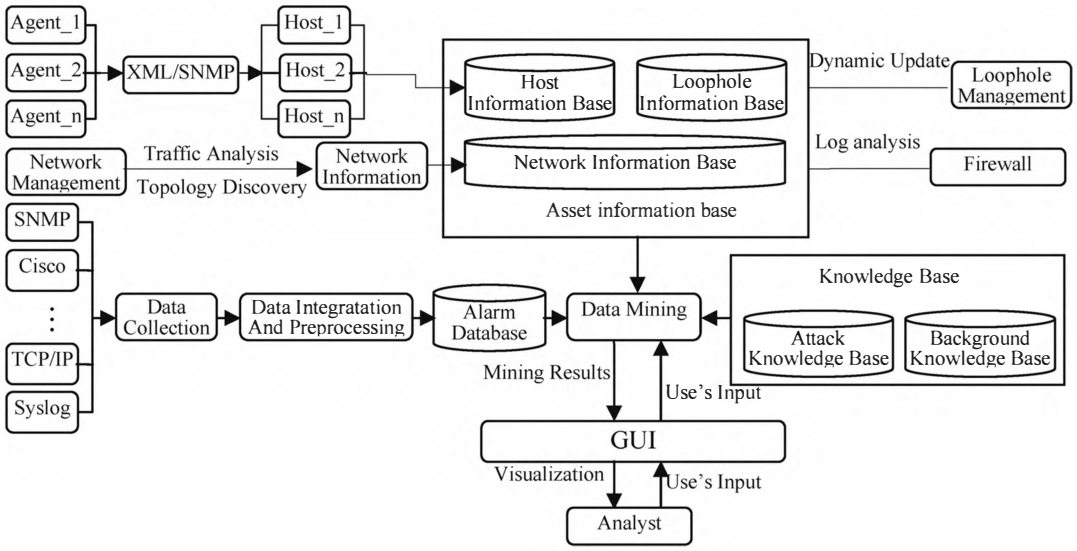

Figure 1 Network security management model based on data mining

This model mainly includes five modules. They are data acquisition, data integration and preprocessing, data mining, user interface and information database. The following will briefly introduce the function of each module.

A. *Data acquisition*

The acquisition of data sources is with the help of Agent automatic program, deliberating on the richer data resource of the security system which is polygenetic and isomeric. There, Agent undertakes a twofold task. One is to acquire data of target system and submit it to Sever for processing. The other is to monitor the changes of target system automatically and update them immediately. Many data acquisition strategies—Log, Web, Syslog, Command line and so on, are applied in Agent automatic program to acquire and standardize the data from the security system which is polygenetic and isomeric, such as security log, alarm, information etc. These security systems are always distributional and have different log formats. Agent can extract information from them uniformly and sent them in the form of event type or IDMEF packaged in XML after being standardized to convenient the mining for Server. Agent uses configuration settings to acquire information on purpose of supporting more security systems. It therefore has a good expansion capability and can add new data sources dynamically.

B. *Data integration and preprocessing*

There are security products of different kinds or from various manufacturers in network security management tools, so most data the Agent collects from them are polygenetic, distributional and isomeric. They must be integrated and preprocessed so that can be analyzed efficiently and intelligently, which is the chief problem of security management needing be solved. It influences not only the species and quantities of the security products supported by management systems, but also the accuracy of analysis results. And it can lay a good foundation for promoting the performance and robustness of data mining engines.

1) Standardize the alarm format. Since formats of security events generated by security systems are different, there may be several alarms to a same intrusion, resulting in redundancy events. So all kinds of security events should be standardized in an uniform format. Then unify the alarm format.

2) Normalize alarm fields. Normalize content representation of each property field, such as the expressions of source and destination addresses like IP, host

name, MAC address and so on, the setting and synchronizing of time's property value and the unifying of attack names, etc. All these is mainly to standardize the expressions of alarm information to help it identify alarms and do further analysis independent of specific system

3) Preprocess data. For the various alarm information from isomeric security devices, it should be examined for errors after removing redundant events according to the set time period, ensuring that the alarms don't include obvious errors in alarm database, loophole information database and asset database. Error detection is to ensure the alarms not containing significant false information, such as illegal time stamp; redundancy elimination is that if the two security events are the same except for the two fields that generate time and system number and the disparity of the generated time doesn't exceed a certain threshold, then we can consider that redundant events appear. All the redundant events are integrated to a single one. The generation time will be set to as the same as the latest one of them and the corresponding security system numbers are added to the integrated one.

C. *Data mining*

Data mining is the power of the whole model. The statistics, association analysis, clustering and series analysis of events can be done for judging events, identifying threatens and generating alarms with appropriate mining algorithms and tools. Its outstanding advantage is to analyze large numbers of security data from various security component elements comprehensively, improving the accuracy of analysis.

1) Association analysis. Association analysis is to find the associations between alarms, alarms and faults, alarms and traffics, i.e., the appearance probability of another alarm, fault, or traffic after one alarm message occurring. All these alarm logs and related backgrounds can be associated with effectively by the multilevel alarm analysis technology based on coordination and time series association to extract and present the real potential dangerous alarms from a large number of logs to the manager. A great many decentral single alarms are associated by way of various styles of alarm analysis to identify the real intrusions effectively. Users are supplied with the assist decision suggestions for special threatens by assist decision system and security expert knowledge base. Association analysis can refine alarm messages, increase the number of available alarm messages, reduce the useless ones and cut down the false alarms from security devices.

2) Clustering analysis. Characteristics aggregation and fuzzy clustering are used in clustering analysis for the compression of repeat alarms and the removal of redundancies in real time. Characteristics aggregation will quickly distinguish and merge the repeat alarms by comparing the property characteristics; fuzzy clustering will construct the fuzzy equivalent matrix for clustering analysis by computing the similarity between alarms to discriminate and lump together the invisible repeat alarms. The reasonable combination and complementary of the two technologies not only shorten the compression time and guarantee the Real-time performance, but also improve the compression efficiency. Clustering analysis lessens the quantity of similar events in isomeric environment and highlights the property feature of similar security events.

3) Series analysis. Series analysis relates the associations to the timeliness both of which are between alarms and finds out the rules of repeat appearance with high probability by time series. It is for mining the relations between data.Series analysis takes alter series as orderly ones with the time as the main line and mines knowledge within a certain time period. It emphasizes the timeliness of alarm messages and usually mines the only two predicates—alter type and alter time, in order to improve the analysis efficiency, to find the trends of alarm messages from them and enhance the user's capability of self-protection and predication. The generated series are mainly to describe the relationship of alarms on time, i.e. if a combination of some alarm messages is occurred in a period, then anther combination will appear in another period.

D. *User interface*

User interface is to present the data mining results in a visual way to the system analyst. Then the system analyst will forecast the develop trend and probable influence of the network action and give some decisions. The mining results are divided into three categories.

1) Information level. Correspond to the lowest alarms and the mining results will be stored for further analysis next time.

2) Alarm level. Correspond to the medium alarms and the mining results will have further analysis as well as being sent to database, and the corresponding decisions will be given.

3) Danger level. Correspond to the high-grade alarms. Take specific action according to the predetermined threshold, such as firewall rules addition, IDS alarms, etc.

E. *Information base*

Information base consists of asset information base and knowledge base — the former one is composed chiefly of the information bases about host, loophole and network; the other one mainly includes the knowledge bases about attacks and backgrounds. The relevant information of host is stored in host information base while the related information from loophole scanning software is in loophole information base, which is an independent database. Network information base includes two main parts. One is the corresponding network information got by means of the traffic analysis and Topology discovery which are implemented by network management tools. The other is the information acquired by the analysis to the log in firewall. The use of information base enhanced the efficiency and intelligence of mining engines extremely.

IV. EXPERIMENT ANALYSIS

A. Experiment Environments and Data Sources

1) Experiment environments

- Internal network. Include the Debian Linux host running OpenNMS, the host installing MySQL, the Windows host running clients like Nessus loophole scanning system and Nmap network topology tool, and the Debian Linux host running system server — agent + server, installing Snort IDS, P0f, Pads, SSH, Iptable and other such plug-in units.
- Hardware security devices like Cisco-PIX firewall and Intrusion Detection System, and network devices like switches and hubs
- The attack hosts from outer net

2) Data sources

The primitive internet data in the experiment include the normal traffic and attack data stream. And the test data is the LLDOS1.0 dataset of DARPA2000. A large number of normal backgrounds and attacks—one of them is DDoS attack, are in this set. The target of test is to exam the efficiency of reducing alarms and the capability of identifying DDoS attacks of the model. In the experiment, we collect the following alarms or data from security device: Snort IDS, Iptables firewall, Apache server log, IIS server log, Nmap network topology structure scanning tool, Ntop network traffic detecting tool, Nessus network loophole scanning system. The data is released by Netpoke a Tcpdump file releasing tool and be detected by Snort 2.0 IDS and Cisco-PIX firewall and other such plug-in components to generate, integrate and mine alarm data.

B. Experiment Result Analysis

The result of the experiment is shown as TABLE 1.

TABLE 1 Experiment result

Alarm	Aggregation	Validation	Attack Graph	Scene Analysis
3865	210	124	12	4
compression rate		false alarm rate	construction rate	scene detection rate
94.56%		2.22%	0.31%	0.103%

From the table we note that the real-time compression rate reach 94.56%, mainly aggregating the alarms from ICMP Ping port scanning and DDoS. There are some wrong reports to a degree owing to that the data set itself doesn't provide the resource allocation information of protected network, so it is necessary to verify the compressed alarms. After that, put causal relationships on the alarms, twelve of which contact with each other to form the DDoS attack scene, to construct the attack scene graph. Then associate them dramatically according to dramatic association rules to match four attack scenes in real time. They are RPC sadmind buffer overflow、Web attack、DDOS attack and so on.

Thus, the network security model based data mining can identify the real attacks from a mass of security events significantly to refine the alarm information, increase its available amount, reduce the useless ones and cut down the false or wrong alarms of security devices.

V. CONCLUSION

The proposed network security management model based on data mining is proved experimentally to enhancing the defensive ability with high efficiency. It can meet the requirement of new security situation with high intelligence, automation, detection performance and adaptability.

REFERENCES

[1] W. K. Lee, S. J.Stolfo, "Data Mining approaches for intrusion detection," Proc. the 7th USENIX Security systems, pp.26-29, 1998.

[2] W. K. Lee, "A Data mining Framework for Constructing Features and Models for Intrusion Detection Systems," New Work: PhD thesis of Columbia University, 1999.

[3] W. K. Lee, S. J.Stolfo, A Framework for Constructing Features and Models for Intrusion Detection Systems," ACM Transactions on Information and System Security, Vol.3, No.4, pp.227-261.

[4] M. Shin, H. Moon, K. H. Ryu, "Applying Data Mining Techniques to Analyze Alarm Data," Proc. the 5th Asia-Pacific Web Conference, 2003.

[5] B. D. Cabrera, P. Lewis, X. Z. Qin, P. Lee, R. K. Mehra, "Proactive Intrusion Detection and Distributed Denial of Service Attacks-A Case Study in Security Management," Journal of Network and Systems Management, Vol.10, No.2, pp.225-254,2002.

[6] J. Wang, X. Zhen, Y. B. Liu, C. H. Shi, "Intrusion Prevention in Depth System Research Based on Data Mining," International Journal of Distributed Sensor Networks, Vol.5, No.1, pp.22-27,2009.

[7] M. Tavallaee, W. Lu, A. A. Ghorbani, "Online Classification of Network Flows," Proc. the 7th Annual Communication Networks and Services Research Conference, pp.78-85,2009.

[8] M. A. Aydin, A. H. Zaim, K. G. Ceylan, "A Hybrid Intrusion Detection System Design for Computer Network Security," Computers and Electrical Engineering, Vol.35, No.3, pp.517-526, 2009.