Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 15 (2011) 3511 – 3515

Procedia Engineering

www.elsevier.com/locate/procedia

Advanced in Control Engineeringand Information Science

A Synergetic Pattern Matching Method Based-on DHT Structure for Intrusion Detection in Large-scale Network

Dong Ma[a], Yongjun Wang[b], Zhenlong Fu[c], a*

[a]*Dong Ma, Ph.D, School of Computer, National University of Defense Technology (NUDT), Changsha,Hunan 410073, China*
[b]*Yongjun Wang, Professor, PhD supervisor, School of Computer, National University of Defense Technology, Changsha, China*
[c]*Dong Ma, Ph.D, School of Computer, National University of Defense Technology (NUDT), Changsha,Hunan 410073, China*

Abstract

Research in network security, with the attacks becoming more frequent, increasing complexity means, for the large-scale network intrusion detection, this paper presents a warning by analyzing the behavior of the log, the contents of the relevant association, through the DHT(Distributed Hash Table) distributed architecture, the Collabarative matching, fusion, and ultimately determine the method of attack paths. First, by improving the classical Apriori algorithm, greatly improving the efficiency of the association. At the same time, through the behavior pattern matching algorithms to extract information about the behavior of the alert and the behavior sequence elements to match the template, and through the right path to finally determine the value of the threat of the network path. After the design of a DHT network, the distributed collaborative match the path used to find complex network attacks. Finally, the overall algorithm flow, proposed a complete threat detection system architecture.

Keywords: security; alarm correlation; pattern matching; behavior sequence

1. Introducton

With the popularization of Internet, the network has penetrated into every aspect of people's lives. Large-scale network security issues become increasingly prominent, and substantial intrusion detection for large-scale network requires introducing a variety of key technologies. Firstly, massive data of the large-scale network have to be collected and filtered, in order to get the required core data for our analysis. For the massive data collecting, this paper adopts the idea of data mining methodology, through an improving

* Corresponding author. Tel.:+86-13807318624; fax: +86-0731-85451901.
E-mail address: malencello@tom.com

doi:10.1016/j.proeng.2011.08.657

to the classical Apriori algorithm by applying it to the alert associated handling, to collect the final data. In large-scale networks, the best way to analyze the data with the ability to handle massive amounts of data quickly still relies on the analysis on the behavior information from alert data. For the behavior information presented in the massive alert information, this article proposes new pattern Matching architecture and constructs a behavior sequence model, to accomplish the pattern matching. Based on a peer DHT protocol, a parallel multi-node synergetic pattern matching method is proposed for data processing, so that both accuracy and time efficiency can be achieved. In order to illustrate the efficiency and practicality of the proposed method, an experiment has been conducted based on the collected network data during a large intranet.

2. Related Work

Clifton and Gengo[1] considers that false alarms appear in the alert because normal operation with similar characteristics of the invasion occurs in a particular environment, and the alarms caused by these operations have a certain sequential pattern. Manganaris[2] divides the continuous alarm flow into a lot of alarm bursts, and map every alarm burst into a transaction. Dr. Mei Haibin[3] firstly proposed using the classic Apriori algorithm to implement the association mining for massive alerts. The idea of this method is simple and easy to be implemented and modified, hence, with a high applicability, especially for resolving the false negative and false positives.

Cupid is a pattern matching system based on element level and structure level[4]. It combines the name matching algorithm and structure matching algorithm together, where the similarity of elements can be derived based on the structure algorithm. However, Bernstein et al.[5] pointed out that the existing algorithms are fragile, and usually require manual adjustments, such as setting the threshold, providing a dictionary or strength training. Even the adjusted method will still give incorrect matched patterns, and most methods cannot do large-scale pattern matching. Shanchieh J. Yang[6] proposed a model based on the action elements of network attack behavior. This model has a good generality, clear classification and strong applicability. DHT algorithm uses a distributed hash function to solve the structure of the distributed storage problem. Since the idea of DHT came out, there are many DHT protocols are proposed and applied, typical examples including Chord[7], Tapestry[8] and and so on. Currently, the P2P (Peer-to-Peer) algorithm based on DHT have been well studied.

3. IACA（Improved Alert Correlation Apriori）Algorithm

The complexity of the algorithm Apriori is O(m^3*n), m is the number of rows of the database, and n is the number of columns of the database. Take the IDS alerts from the network of a large intranet for example, the amount of the alarm alerts in a week is about 15,00000, with this algorithm more than 10 hours is required, the time consumption is enormous. In the algorithm based on Apriori, set a time period t (t> 1 / n), for example t = 3, that means one day will be divided into three segments, each 8 hours. Alerts in a time period will be seen as a set, and the alarms within each set are parallel computed once using Apriori algorithm. Through an experiment, a curve can be obtained, as shown in Figure 1(a).

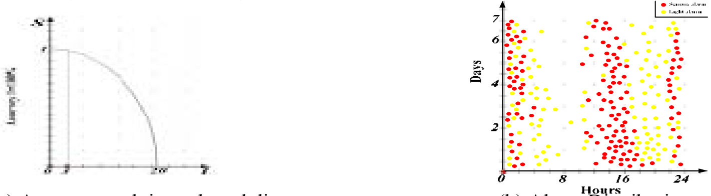


Fig. 1.(a) Accuracy and time-phased diagram (b) Alarm Distribution

Figure 1(b) shows the intensity of the alert appears, the figure shows, the alarm focus on three time periods, so we will take $t = 3$, the algorithm can guarantee the highest accuracy.

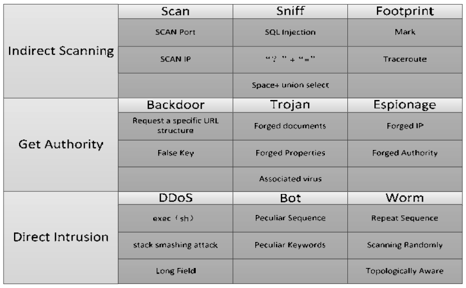
So we divide the algorithm into t segments and calculate them in parallel, then accumulate the results, so the time cost problem is solved and the algorithm becomes scalable. For the follow-up alerts we just need to calculate it according to the alarms in each time segment, and then sum the results up to update data in real time.

4. Behavior Pattern Matching

Behavior pattern matching is the key technology which collects the elements, then classifies and models them to determine whether a behavior sequence matches certain behavior patterns or not. Using a suitable matching algorithm, with necessary improvements, pattern matching can give us the result we want. According to the above idea, the overall architecture of behavior pattern matching system is given.

4.1. Behavior Classification Model

By the difference of network attacks, we divide the attacks into three categories in this paper: indirect scanning, permissions acquisition, and direct invasion. According to the threatening behavior elements in the snort rule library, threatening behavior classification elements of TIAA (a Toolkit for Intrusion Alert Analysis) system and other attack elements found in the long-term network security work , the most common behaviors can be classified into the three types of behavior above, and classification models is shown in Figure 2.



Indirect Scanning	Scan	Sniff	Footprint
	SCAN Port	SQL Injection	Mark
	SCAN IP	"? " + "="	Traceroute
		Space+ union select	
Get Authority	Backdoor	Trojan	Espionage
	Request a specific URL structure	Forged documents	Forged IP
	False Key	Forged Properties	Forged Authority
		Associated virus	
Direct Intrusion	DDoS	Bot	Worm
	exec（sh）	Peculiar Sequence	Repeat Sequence
	stack smashing attack	Peculiar Keywords	Scanning Randomly
	Long Field		Topologically Aware

Fig. 2. Classification of specific acts

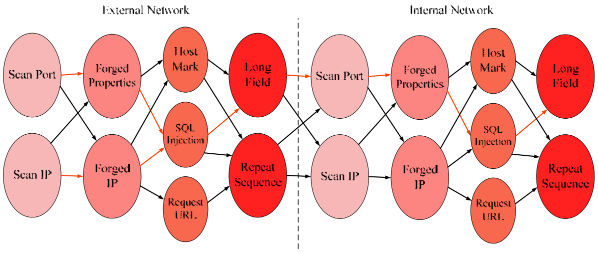


Fig. 3. A behavior sequence template

For example, the behavior of "SCAN" in the alert is scanning, then we can classified it into the indirect scanning category , and "exec (sh)" or "stack smashing attack" and other activities of the class of buffer overflow attacks can be attributed to the direct invasion category, and so on. Here we are not going to list all the kinds of behavior

4.2. Behavior Sequence Template

According to the association between each other, this study created more than 10 general-purpose behavior sequences. Figure 3 is one of the behavior sequences, that is, a behavior sequence template.

For complex attacks may be carried out to an internal node from both internal and external networks, the behavior sequence templates defined in this study can be further divided into internal network part and external network part. In practice, the routers of external network and internal network install IDS, respectively, to get alarm information. In this section, our study is target on the individual pattern matching for the alarm information of any network node, no matter it is internal or external, so the matching method is the same.

4.3. Pattern Matching Algorithm

According to the rules of Cupid pattern matching system, we will divide the studied matching algorithm into three steps:

The first step is the element level matching, and classification by name, data type, and field. The second step, we have to convert the original behavior patterns, i.e. the behavior sequence template into a pattern tree, using tree matching algorithm, making the structure of bottom-up match. The similarity between patterns depends on the similarity of their name and their path similarity. For alarm pattern matching, once the name is matched, then the similarity is set to be 1. Path coefficient is defined in the template, based on the different similarities of behavior elements in different paths, for every behavior sequence template. The weight of every matching path can be calculated by equations:

$$Names(N_1,N_2)=\frac{\sum_{n_1\subset N_1}[\max sim(n_1,n_2)]+\sum_{n_2\subset N2}[\max sim(n_1,n_2)]}{|N_1|+|N_2|} \qquad Paths(P_1,P_2)=\frac{\sum_{p_i\subset P} p_i\times Names(N_{1i},N_{2i})}{|P|\times(|N_{1i}|+|N_{2i}|)}$$

Where n is the element weight, p is the path to the right value, represents a completepath to the value and ownership of, where we can see a name similar to 1, then the pathweight formula can be simplified to:

$$Paths(P_1,P_2)=\frac{\sum_{p_i\subset P} p_i}{|P|}$$

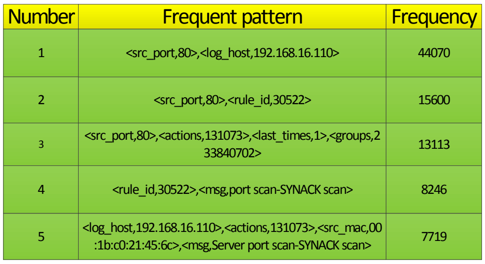
The third step is to select the matched results based on the weighted averages. In this study, the highest value is not the only result; every path is likely to be an attack process. The highest weight can only be considered as the most likely attack path.

5. Network Structure Design Based on DHT

DHT algorithm uses a distributed hash function to solve the structural distributed storage problem. The idea is: every resource is identified by a group of keywords, and the system hashes every keyword to get the keywords identifier “key”; every network node also have a node identifier “ID”, which is obtained by hashed IP address of the node; keyword identifier and node identifier is unique; according to certain mapping function, mapping the keyword identifier onto the node identifier, which indicates the node with the node identifier stores the corresponding resources identified by the keyword identifier. All the <key, value> pairs constitute a huge hash index table of the stored files. Where: “key” is the key hash; value is the address to store the information. And select the chord structure as the underlying P2P communication structure, data transmission between nodes and communication are built based on it. Attack is matched step by step, according to the flow chart.

6. Experimental Results

The experimental data source for the a large intranet, China during November 15-21, 2010，a week in four nodes of the IDS alert messages are collected, the data contains a total of 1441440 alert records. Associated by Alert correlation algorithm results in Figure 4, shows the frequency of suspense in the top five of the frequent patterns. So we have these five modes corresponding to 88748 alerts to extract the behavior of the matching elements.



Number	Frequent pattern	Frequency
1	<src_port,80>,<log_host,192.168.16.110>	44070
2	<src_port,80>,<rule_id,30522>	15600
3	<src_port,80>,<actions,131073>,<last_times,1>,<groups,233840702>	13113
4	<rule_id,30522>,<msg,port scan-SYNACK scan>	8246
5	<log_host,192.168.16.110>,<actions,131073>,<src_mac,00:1b:c0:21:45:6c>,<msg,Server port scan-SYNACK scan>	7719

Fig. 4. Results of frequent pattern and frequency

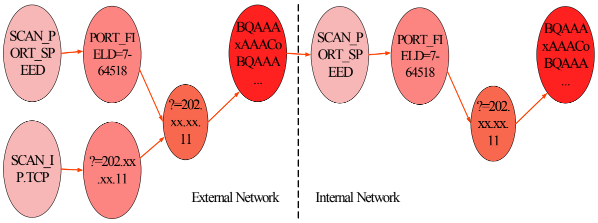


Fig. 5. Results of one attack path

Matching algorithm by matching the element level, 88,748 alerts were matched through the sequence of the template sequence of the 11,030 attacks, including the effective use of the alarm has 78,761. Alert messages will be other manual analysis by creating a new attack sequence template.

The right path by calculating the value of the maximum weight one attack path as shown in Figure 5. Corresponding to a total of 10 alarms, then this is not enumerated.

7. Conclusion and Future Work

In this paper, improved algorithm by association to raise the alarm, the alarm time is only associated with the original 1 / 60, and to ensure the accuracy of the original 90%. Coordination through the DHT structure match, behavior by constructing a sequence of templates and behavior pattern matching algorithms that can accurately identify the attack path, determine the attack.

The next steps, to construct a template for the standardsequence of good behavior, the next step in the path of attack to predict and forecast based on the attack path to do the overall assessment of network security situation.

References

[1] C.Clifton, G.Gengo. Developing custom intrusion detection filters using data mining[C]. In: Military Communications Int'1 Symposium, California; 2000, 440-443.

[2] S.Manganaris, M.Christensen, D.Zerkle, K.Hermiz. A data mining analysism of RTID alarms[J]. Computer Networks; 2000, 34(4):571-577.

[3] Haibin Mei, Large-scale network IDS Alert Correlation and Prediction[D]:[Ph.D Thesis]. Nanjing, China: Southeast University; 2010.

[4] Madhavan J, Bernstein PA, Rahm E. Generic Schema Matching with Cupid[C]. VLDB Conference; 2001:49-58.

[5] Rahm E, Bernstein PA, A Survey of Approaches to Automatic Schema Matching[J]. The VLDB Journal; 2001, 10(4), pp.334~350.

[6] Shanchieh J.Yang, Adam Stotz, Jared Holsopple, Moises Sudit, Michael Kuhl, High level information fusion for tracking and projection of multistage cyber attacks[J]. Information Fusion; 2009, 10, pp. 107~121.

[7] Stoica I, Morris R, Liben D, et al. Chord: a scalable peer-to-peer lookup protocol for Internet applications[J]. IEEE/ACM Trans on Networking; 2003, 11(1):17-32.

[8] Zhao B Y, Ling Huang, Stribling J, et al. Tapestry: a resilient global-scale overlay for service deployment[J]. IEEE Journal on Selected Areas in Commu nications; 2004, 22(1):41-53.