Cyber Security Situation Awareness Based on Data Mining

Liu Jie[1], Feng Xuewei[1, 2], Li Jin[1, 2] , Wang Dongxia[1, 2]
1 Beijing Institute of System Engineer, Beijing 100101, China
2 Nation Key Laboratory of Science and Technology on Information System Security, Beijing, 100101, China
e-mail: brafum@yeah.net, zyan1981@gmail.com, dongxiawang@126.com

***Abstract*—Situation awareness is a kind of the third generation of information security technology, which aims to provide the global security views of the cyberspace for administrators. A framework of cyber security situation awareness based on data mining is proposed in this paper. The framework can be viewed from two perspectives, one is data flow, which presents the abstracting of cyber data, and the other one is logic view, which presents the procedure of situation awareness. The framework's core component is correlation state machine, which is an extension of state machine. The correlation state machine is a data structure of achieving situation awareness, which is created based on the technology of data mining. After being created, it can be used to assess and predict the threat situation to achieve cyber knowledge. We conclude with an example of how the framework can be applied to real world to provide cyber security situation for administrators.**

Keywords- cyber security, situation awareness, correlation state machine, threat prediction, threat assessment

I. INTRODUCTION

With the dramatically increasing of cyber attacks during the last several years, cyber security situation awareness has become a major contributor for cyber security. Cyber security situation awareness belongs to the third generation of information security defense technology, and it aims to provide the cyberspace's global security views and states for administrators. Based on the comprehensive knowledge provided by situation awareness, administrators can make decisions timely and accurately, this will reduce the compromising considerably.

When attackers exploit the cyber infrastructure, security devices such as IDS (Intrusion Detection System) etc. which are deployed in cyberspace distributedly will generate hundreds of thousands of security events. It is impossible to recognize attack scenarios and be aware of security state of the cyberspace from these events by the manual way, so how to correlate these events to reconstruct attack scenarios and assess the attack activity's consequence automatically is the main goal of cyber security situation awareness.

In this paper, after analyzing the current researches on situation awareness, we propose a framework of cyber security situation awareness based on state machine. It is more operable than the concept model provided by TimBass[1][2], and can be used to guide the whole process of cyber security situation awareness. The framework can be viewed from two perspectives, one is data flow, which presents the abstracting of cyber data, namely along with the processing of situation awareness, the quantity of cyber data is decreasing while the quality is increasing, data is abstracted to knowledge. The other one is logic view, which presents the procedure of situation awareness, that is methods or approaches can be acquired to processing cyber data to achieve cyber information and cyber knowledge.

The remainder of this paper is organized as follows: section 2 analyzes the exiting researches on cyber security situation awareness. Section 3 describes the structure of the situation awareness framework, and the main contents and methods of the framework are explained detailedly in this section too. In section 4, we use an experiment to validate the technologies in the framework. At last we summarize this paper and suggest the future work in section 5.

II. RELATED WORK

A. Models or Frameworks of Cyber Security Situation Awareness

We presented a framework for network security situation analysis in [3], which specializes in situation analysis of network, and situation prediction is not comprised in the framework. The framework is derived from the data fusion model of TimBass. The data fusion model proposed by TimBass in 1999 is the rudiment of network security situation awareness. The model includes five levels, and the objectives of each level are defined in this model. It mainly focuses on the concept of situation awareness, so in the practical environment, available technologies and methods for analyzing security situation are lacking.

Wang[4][5] proposes a hierarchical implementation model for network security situation awareness, but the situation information presented by this model is only security threat level, the level can not indicates the whole situation of the network system. A security situation evaluation model for inter-domain routing system in the internet is proposed in [6], the model is a two tuples (TREE, EA), TREE is the routing tree of the routing system, and EA is the assessment algorithm. The ultimate results presented by this model are security metrics too. Otherwise, the EA's input is abnormal routing information, this information is always hard to acquire accurately.

B. Methods or Technologies in Situation Awareness

In order to acquire the overall information of attack activities, such as time, place, multistage etc., we

Published by Atlantis Press, Paris, France.

reconstruct attack scenarios from raw security events in [7] using the rudiment of correlation state machine, where the correlation state machine can not contribute to predict and assess cyber threat situation. Liu[8] presents a method of network security situation awareness based on artificial immunity system. The method uses network intrusion detection based on the theory of immunity as the base of situational awareness, to detect known and unknown intrusions with the help of biological technology. The goal of this method is to quantify the security situation of the network system. However, sometimes network security situation can not be boiled down to quantifiable information, such as attack scenarios. Gorodetsky[9] uses agents to detect and analyze network traffic, and then gets the security situation. The situation assessment method is based on asynchronous data flow, usually data flow can not represent all the basic security information in networks. Chen [10] gives a hierarchic assessment method to analyze and quantify the network security threats, but he does not give a practical algorithm when assessing the system level threat, and this method only focuses on threat assessment. Yegneswaran[11] uses IDS Bro to analyze network's activity information provided by honeynets, and then depict the curve of the security situation. However, the result will be intelligible only in the case of large scale attack happening.

Predicting belongs to the third level of situation awareness according to TimBass' concept model, which aims to deduce the security state of cyberspace. GENSHE CHEN[12] analyzed the insufficiency of Bayesian Network in attack intent predicting and proposed a data fusion method to predict security situation of cyberspace based on high-level knowledge and game theory. Firstly, threats are detected and classified based on intelligent agents, and then a game theory model is used to trap attacks and predict intents of the corresponding attacks. There are some other researches on situation predicting, such as Zhang[13] proposed a time serial method for achieving cyber security situation awareness based on Box-Jenkins model and Holt-Winter model.

III. A FRAMEWORK OF CYBER SECURITY SITUATION AWARENESS BASED ON STATE MACHINE

In order to achieve cyber security situation accurately and timely especially the threat situation, we present a framework based on correlation state machine. The framework is a successor of what was proposed in [3] by us, and more accurate architecture is proposed and the core technology is implemented. As shown in Fig.1, the framework comprises four parts mainly, database module, process module, correlation state machine and GUI. The database module is used to store sample/history data and attack scenarios which are discovered after mining. The process module comprises four processing components, namely Knowledge Mining, Correlation Analyzing, Threat Predicting and Threat Assessing. The correlation state machine is the core data structure, which is used to predict and assess cyber threat. GUI is an interactive interface between administrators and the framework. The green broken line in Fig.1 represents the data flow view while the black line represents the logic flow view, the two views are coordinated.

Firstly, sample/history data is collected from experimental cyberspace or known data sets, and then stored in databases, it is data source of Knowledge Mining and could be updated every now and then. After analyzing and discovering, Knowledge Mining generates various patterns of attack scenarios and stores them in databases. Secondly, the Correlation Analyzing component analyzes cyber security events at real-time based on the known attack patterns to reconstruct the occurent attack scenarios. The occurent attack scenarios are represented by the correlation state machines, namely a correlation state machine is an attack scenario which is taking place in cyberspace at present. After being created, the correlation state machines can be input of the Threat Predicting and Threat Assessing components. The threat predicting component can deduce possible attack paths of an attack scenario based on the corresponding correlation state machine, and the Threat Assessing component assesses the consequence resulting from the attack scenario based on the corresponding correlation state machine too. These two components support administrators to make decisions directly. Finally, cyber security situation knowledge including the results of predicting and assessing is presented in a friendly way by the GUI, and the knowledge can be understood easily by administrators after being visualized. This can decrease administrators' cognitive burden considerably.

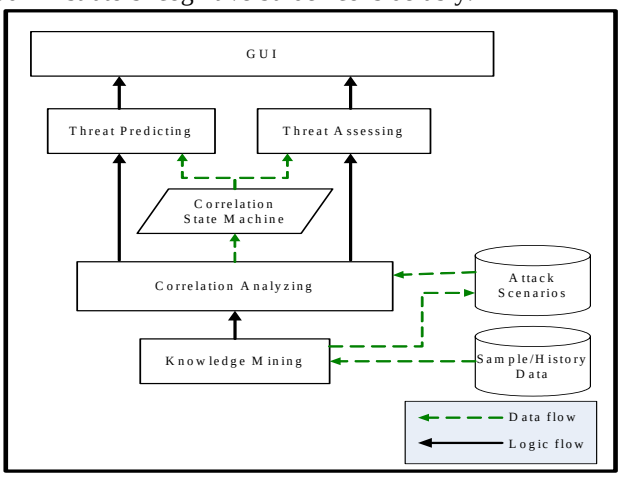

Figure 1. The framework of cyber security situation awareness

A. *Discovering Patterns of Attack Scenarios Based on Data Mining*

Correlation state machine is the core of the framework, which aims to model and abstract the occurent attack activities. It is a dynamic data structure, and created based on the known patterns of attack scenarios. So, in order to

achieve cyber security situation awareness, various patterns of attack scenarios must be discovered. We implemented this by data mining as shown in Fig.2.

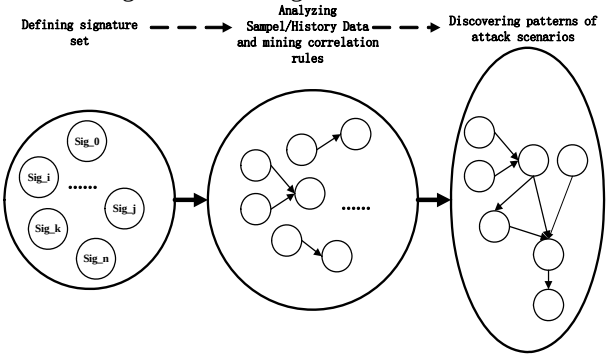

Figure 2. Discovering patterns of attack scenarios based on data mining

Firstly, we define a signature set, which is composed of characteristics of various attack activities, such as SYN-Scan, exploiting of certain vulnerabilities etc., this can be done referring to the existing projects or researches, such as Snort and so on. Every signature word in the set represents certain actions or steps of various attacks, and new words can be added to the set continually, that is the set should be expanded and updated continually in order to discover new attack activities. Secondly, the Knowledge Mining engine analyzes sample/history data to discover frequent itemsets which are composed of signature words. When the frequent itemsets are found out, correlation rules can be generated easily. Finally, if two rules have the same word, or they have the causal relationship, they are combined. After searching all of the rules in this way, various patterns of attack scenarios can be discovered and generated, which will be used when creating correlation state machine.

After generated, patterns of attack scenarios are formalized by XML. Specially, the sample/history data and signature set are critical to discover attack patterns, so administrators should collect useful information at daily work, but if administrators have known some patterns of attack scenarios, they can write the XML files directly, no need to mine the sample data, as what we did in [7].

B. Correlation Analysis of Cyber Security Events

Correlation analysis of cyber security events is the foundation of acquiring high-level knowledge of cyber security situation, it distinguishes and standardises raw security events generated by security devices, and then analyzes the logic relationships between these events based on attack patterns. When events are submitted, correlation engine will match security events with attack patterns discovered by the Knowledge Mining till finding out the corresponding attack pattern or patterns. After the pattern or patterns are found out, one or more correlation state machine will be created according to the patterns, that is one pattern responses to one state machine.

Definition 1 correlation state machine: correlation state machine is a data structure, which maps the XML file of corresponding attack pattern to memory. Each state of the machine is a tuple and derived from the corresponding rule node of the pattern, that is *state_i* =(*plugin_id*, *plugin_sid*, *src_ip*, *dst_ip*, *src_port*, *dst_port*, *protocal*, *timeout*, *occurrence*, *srcIP_record*, *dstIP_record*, *srcPort_record*, *dstPort_record*, *eventCounter*, *startTime*). The first nine attributes of the state indicate the characteristics of the security events which can be processed by this state, and their meanings are the same as the rule presenting in [7]. *srcIP_record*, *dstIP_record*, *srcPort_record* and *dstPort_record* are used to record the characteristics of the events processed by this state, *eventCounter* is used to record the number of the events processed successfully and *startTime* indicates the start time that this state takes effect.

As shown in Fig.3, the correlation state machine is an intermediate process, it is used to track and record the process of matching between security events and attack patterns timely. *timeout* and *occurrence* are the two core concepts of the state machine. *timeout* indicates that how long the engine monitors one state, it corresponds to the certain attack step's lasting time in multi-step attack. *occurrence* presents the number of the security events that the state can process, essentially, this attribute means to classify the similar security events, it reflects the idea of clustering analysis. *timeout* and *occurrence* work cooperatively, if the correlation engine processed "*occurrence*" security events successfully in the time limit of "*timeout*", the current states of the state machine will transfer, this reflects the development of the multi-step attack. The cooperative working of *occurrence* and *timeout* realizes the unification of clustering analysis and causal analysis.

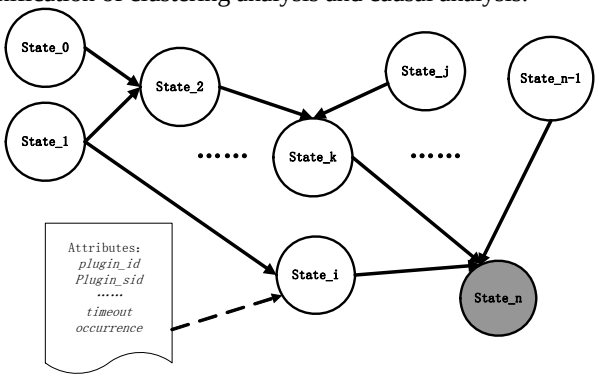

Figure 3. Example of correlation state machine

Because one correlation state machine represents one attack scenario which is happening in cyberspace, so the number of the machines equals the number of attack activities which are detected. The state of the attack activities is represented by the current states of the machine, so administrators could be aware of the situation of cyber threats through observing the current states. The correlation state machine implements modeling of attack threats.

C. Predicting the Threat Situation Using State Machine

We research on attack intent reasoning mainly in this paper, so in our framework, predicting is used to deduce the trend of attack scenarios, that is estimating the next step of attack scenarios, this is helpful to decrease the damage caused by attack activities.

When attackers launched an exploitation, they always want to find out the optimized path to exploit a target system, so in order to simulate this procedure, we use the algorithm of trace-back to find out the optimized path in a correlation state machine. This is similar to an attacker's exploiting in the real world. If we can find out the optimized path in each correlation state machine, we can predict the corresponding attack scenarios' trend easily, that is we can predict the threat situation in the whole cyberspace.

For example, in Fig.4, there are more than one possible attack paths to State_n from current state State_1, such as attack path_i and attack path_j. To find out the optimized path in these different paths using the algorithm of trace-back, that is to find out the path whose costs is the lowest, then the most possible attack path that an attack will adopt is presented. But, there are many factors that should be considered when predicting cyber threat situation, for example defense strategies on target system, vulnerabilities, attackers' capacity and so on. So it is impossible that the predicting path is always the same as the real path. In order to improve the accuracy of predicting, we consider some weight factors that could affect attackers' decision when exploiting, such as defense strategies, vulnerabilities etc.. Using these factors to modify the optimized path timely, and presenting the probability of each path in various correlation state machines, trends of various attack scenarios and the threat situation of cyberspace can be predicted.

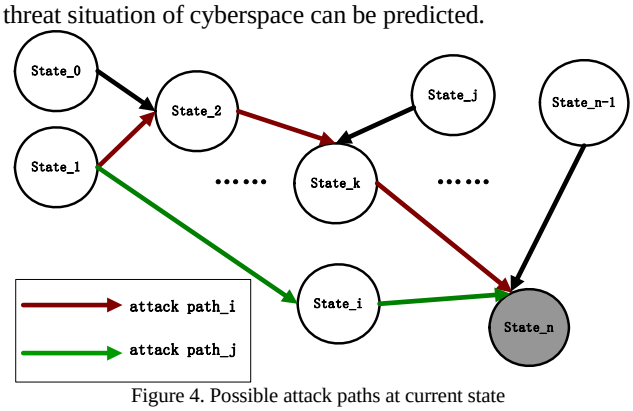

Figure 4. Possible attack paths at current state

D. *Assessment of the Threat Situation*

Administrators can be aware of the global security state and security level of cyberspace through assessment. After creating the correlation state machine and predicting the trends of various attack scenarios, we implements threat situation assessment through calculating the threat metric of cyberspace based on each correlation state machine. The whole algorithm of computing the cyberspace's threat metric can be illuminated by (1) and (2).

$$f_{machine_i}(opt\_path_i) = \sum_{1}^{n} state_j \quad : state_j \in opt\_path_i \quad (1)$$

$$F_{Threat\_Situation}(machine) = \sum_{1}^{n} f_{machine_i}(opt\_path_i) \quad (2)$$

Formula (1) presents how to quantify the threats caused by one correlation state machine, namely to quantify the harm degree derived from an attack scenario. (1) indicates that every state machine's threats $f_{machine_i}(\ )$ is only related to its own optimized attack path opt_path_i. When predicting component presents the optimized attack path of a correlation state machine, every state which is in this path will be set a quantitive metric, the metric is a contributor to the machine's threats. Then, all the metrics in the path are added together to compute the machine's threats. In this way, administrators will acquire a number of numerical values, which indicate the harm degree caused by various attack scenarios to cyber infrastructure. Especially, the quantitive metric of each state can be defined depending on the concrete needs. In our experiment, we use the variable of "*occurrence*" to represent the quantitive metric simply.

After acquiring the numerical threats of all the correlation state machines, we can use (2) to compute the global threats that the cyberspace suffers. The global threats $F_{Threat\_Situation}(\ )$ are determined by every state machine's threats $f_{machine_i}(\ )$. The assessment component adds up all the numerical threats to computer $F_{Threat\_Situation}(\ )$, which is an indicator of the whole cyberspace's security level.

Assessment of the threat situation should be executed periodically. In an assessment cycle, the component scans all the correlation state machines which is living in the memory, and computes the threats of the machines based on their optimized attack path. After adding up, the global threats in this cycle could be a point at y axis while time is the other axis. To link these points together in different cycles, a threat situation assessment curve will be presented to administrators.

IV. EXPERIMENT ANALYSIS

We implemented a prototype to validate the framework, and Fig.5 is the topology of the prototype.

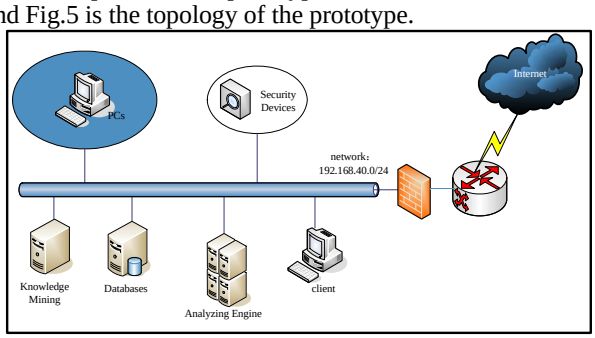

Figure 5. Topology of the prototype

Firstly, Knowledge Mining analyzes the sample data of DARPA 2000 intrusion scenario specific data sets to discover the pattern of Dos attack scenario, and after being generated, the pattern is stored in databases and can be illustrated by Fig,6. Meanings of each step of the Dos attack scenario were explained in our previous work[7]. Secondly,

we replay the data sets in network 192.168.40.0/24. Security devices deployed in this network will generate a large number of events, and these events are sent to Analyzing Engine. Thirdly, after processing these events timely, Analyzing Engine generates a correlation state machine based on the pattern of Dos attack scenario, which can be illustrated by Fig.7. The correlation state machine indicates that it is a distributed dos (DDos) attack activity, and the current state of the activity is "Mstream_Zombie", namely interacting between the controller node and zombie nodes. Attack path from State_1 to the current state State_5 presents the evolving procedure of the DDos activity and the possible attack path later. Especially, the predicting and assessing components are implemented in Analyzing Engine, and after being created, the correlation state machines can be used to assess the threat situation of network 192.168.40.0/24. Finally, administrators can view all the situation information through client, which provides a GUI.

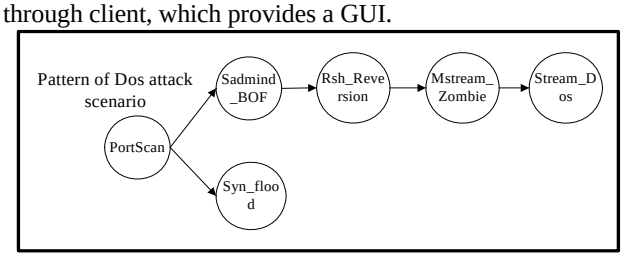

Figure 6. Pattern of dos attack scenario

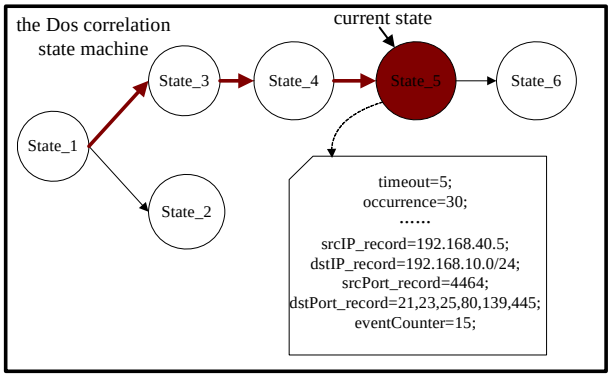

Figure 7. The Dos correlation state machine

V. CONCLUSION

A framework of cyber security situation awareness based on data mining is proposed in this paper, the framework provides guidance and technical support for the whole situation awareness procedure, and it is the foundation of the awareness work. There are four modules in the framework, database module, process module, correlation state machine and GUI. After various attack patterns are discovered, attack scenario will be reconstructed through correlating security events, and the scenarios' abstraction can be used to predict and assess the threat situation in cyberspace, namely the correlation state machine.

In this paper, we don't explain the performance of the analysis methods in the framework detailedly, and just emphasize the structure and profile of the framework. One of the most important works in the future is to analyze and test the performance of the methods in real world.

ACKNOWLEDGMENT

The work of this paper is supported by the National Natural Science Foundation of China Project under grant No.61271252.

REFERENCES

[1] Bass T. Multi-Sensor Data Fusion for next Generation Distributed Intrusion Detection Systems [C]. 1999 IRIS National Symposium on Sensor and Data Fusion, Laurel, USA, 1999(1):24-27.

[2] Bass T. Intrusion Detection Systems and Multi-Sensor Data Fusion: Creating Cyberspace Situation Awareness [J]. Communications of the ACM, 2000,43(4):99-105.

[3] Feng Xuewei, Wang Dongxia. A Framework of Network Security Situation Analysis Based on the Technologies of Event Correlation and Situation Assessment. 2011 International workshop on Frontiers of Secure Networks.

[4] Wang Huiqiang, Lai Jibao, Hu Mingming. Research on the key implement technology of network security situation awareness[J]. Geomatics and Information Science of Wuhan University. 2008, Vol.33 No.10 (in Chinese)

[5] Wang Yanbo, Wang Huiqiang, Wang Xiufeng, Yu Ming. Design of multi-source and heterogeneous log sensor for network situational awareness. Transducer and Microsystem Technologies. 2010. Vol. 29 No.3 (in Chinese)

[6] Liu Xin, Wang Xiaoqiang, Zhu Peidong, Peng Yuxing. Security Evaluation for Inter-Domain Routing System in the Internet[J]. Journal of Computer Research and Development. 2009. 46(10): 1669-1677 (in Chinese)

[7] Feng Xuewei, Wang Dongxia. Analyzing and Correlating Security Events Using State Machine. 2010 International workshop on Frontiers of Secure Networks.

[8] Liu Nian, Liu Sunjun, Liu Yong, Zhao Hui. Method of Network Security Situation Awareness Based on Artificial Immunity System[J]. Computer Science. 2010. Vol.37 No.1 (in Chinese)

[9] Gorodetsky V, Karsaev O, Samoilov V. On-line update of situation assessment based on asynchronous data streams[C] //Knowledge-Based Intelligent Information and Engineering Systems. Berlin/Heidelberg:Springer，2004：1136-1142

[10] Chen Xiuzhen， Zhen Qinghua， Guan Xiaohong， et al. Quantitative hierarchical threat evaluation model for network security[J]. Jouranl of Software， 2006， 17(4)： 885－897 (in Chinese)

[11] Yegneswaran V, Barford P, Paxson V. Using Honeynets for Internet situation awareness [C/OL] //Pro of ACM/ USENIX Hotnets IV. 2005[2008-01-12]. http://www.icir.org/vern/papers/sit-aware-hotnet05.pdf

[12] GENSHE CHEN DAN SHEN CHIMAN KWAN . Game Theoretic Approach to Threat Prediction and Situation Awareness. JOURNAL OF ADVANCES IN INFORMATION FUSION VOL. 2, NO. 1 JUNE 2007

[13] Zhang Yong, Network Security Situation Awareness Model Research and System Implementation. A dissertation for doctor' s degree. University of Science and Technology of China. 2010.5