Research and Design of Network Situation Awareness System Based on Big Data

Ran Meng¹, Xiaobing Pei²

¹School of management, Tianjin University of Technology, Tianjin, 300000, China; ²School of management, Tianjin University of Technology, Tianjin, 300000, China; ¹Corresponding author’s e-mail: 296370139@qq.com

Abstract. Network situational awareness has become a new hotspot in the research of network security. Based on the analysis of existing network situational awareness models, this paper establishes a network situational awareness model based on the SimHash algorithm in a big data environment, which provides reference value for the next step of the network situational awareness system.

1. Introduction
The Internet has penetrated into all aspects of social life, and malicious activities against the network have also been increasing. Ensuring the safe operation of the network has become the core content of the national information strategy. In order to cope with the information explosion in the field of network security, network situational awareness technology has been proposed. Its main purpose is to extract, refine, integrate, deepen, and manage various related information provided by the network, to help network managers understand the status of the network and the next development trend provides the basis for network deployment and emergency decision-making.

2. Related research
In the traditional network security situation assessment scheme, some representative methods have emerged. Chen Xiuzhen et al. [¹] proposed a hierarchical network security situation assessment model from the bottom up, first to the whole. Wei Yong et al. [²] introduced the Dempster / Shafer evidence theory and calculated the network security situation by fusing the elements of the network security situation. Noel et al. [³] used Attack Graph to characterize the attacker's attack implementation steps, and then quantified the network security status by simulating network penetration attacks and simulating attack propagation paths. These methods work well in some smaller networks, but in today's big data environment, there are some shortcomings. For example, the traditional network security situation assessment model will encounter bottlenecks such as large data storage, fast acquisition frequency, and complex structure, which will directly lead to deviations in network security situation assessment results and insufficient feedback in time.

3. SimHash algorithm
The SimHash algorithm is an efficient algorithm for finding similar text. The SimHash algorithm is essentially a dimensionality reduction method. It maps a high-dimensional vector to a smaller number of hash values and is used to represent the characteristics of the original vector. Its characteristic is that Bright distances have a positive correlation with the cosine similarity between their feature vectors.
Liu et al.\textsuperscript{[4]} improved the SimHash algorithm and applied the improved algorithm to different fields, which has greatly inspired our research.

Traditionally, the SimHash algorithm is generally used for web page deduplication, document similarity detection, and so on. Because in a large-scale network, various security devices generate network security posture elements quickly, in large numbers, and with high dimensions, the traditional network security posture assessment model is difficult to effectively deal with this, both in terms of execution efficiency and storage. Therefore, network security managers urgently need an efficient evaluation algorithm so that they can quickly understand the current security status of the network. In response to the above problems, we introduced the SimHash algorithm into network security situational awareness. Based on the SimHash algorithm, we proposed our node security situation assessment algorithm to quickly quantify the severity of an attack on a node over a period of time.

4. Network Security Situation Assessment Model

4.1 Research on network security situation assessment based on SimHash algorithm

Traditional network security situation assessment schemes generally only detect based on a single alert or a single log, and a single data source will directly lead to deviations between the assessment results of the network security situation and the actual situation, and the assessment methods often use relatively complex algorithms directly affect the timeliness of the assessment and delays the optimal time for the network administrator to take measures. To address these issues, this chapter presents a network security situation assessment model based on the SimHash algorithm in a big data environment.

![Network Security Situation](image)

The model is first based on a complex network community structure discovery algorithm, which divides a large-scale network into several relatively small modules; then, a big data platform is built to collect network security on each node within each module network Situational factors, and based on the SimHash algorithm to design an efficient node security situation assessment algorithm; finally, combining the weights of nodes and modules, respectively, to gradually and efficiently quantify the node security situation, module security situation, and network security situation, network managers can understand the entire network at any time Safe state.

4.2 Experimental results

In order to verify the validity of the model proposed in this chapter, we used the 2000 DARPA dataset provided by the MIT Lincoln Laboratory as the initial data for the experiment. This data set consists of two attack scenarios, LLDOS1.0 and LLDOS2.0.2. The data mainly includes network traffic and host
security logs, which can be used as a data source for the evaluation model. In this article, we will evaluate the cyber security posture for these attack scenarios.

First, write rules for intrusion detection systems to analyse network traffic and get alerts; then upload these alerts and system security log information to a distributed file system (HDFS) for storage. Because part of the alarm information in the intrusion detection system can also be reflected in the system security log, you need to design a MapReduce program to analyse the data on HDFS to exclude redundant data. Finally, the attack information A is calculated and the vulnerability information V is constructed. Using A and V as the input data of the algorithm, the security situation value of each node are generated.

![Figure 2. New intrusion detection system model](image)

During the first, second and sixth time periods, the attacker scanned a small number of data packets from key nodes in the network to determine which nodes were active. Because scanning nodes is a relatively normal behaviour, the impact is small, so the node security posture values calculated during these three time periods are low. In the third period, because the attacker determined that the Mill and Pascal nodes were active and had obtained the vulnerabilities of these two nodes, a buffer overflow attack was launched on these two nodes. The buffer overflow attack is a very dangerous attack, which may damage the running of the program, gain control of the program, and even the system. Therefore, the node security posture value calculated during this period is high. During the third and tenth time periods, the attacker controlled the Mill and Pascal nodes to launch DoS attacks on the Mil node. The purpose of a DoS attack is to deplete the network or system resources of the target computer, temporarily interrupt or stop services, and make normal users inaccessible.

It can be seen that the node security situation assessment based on the SimHash algorithm is accurate and can reflect the severity of each node's attack.

5 Conclusions
This article first analyses the current situation of network situation awareness systems, proposes a network security situation assessment model based on SimHash algorithm in a big data environment, and analyses and validates the effectiveness of the assessment model through experiments. The next step will continue to improve models and methods for security situation assessment in large-scale network environments, predict the security situation in large-scale network environments.

References
[1] Chen Xiuzhen, Zheng Qinghua, Guan Xiaohong. Quantitative Evaluation Method of Hierarchical Cyber Security Threat Situation [J]. Journal of Software, 2006, 17(4):885–897.
[2] Wei Yong, Lian Yifeng, Feng Dengguo. Network Security Situation Assessment Model Based on Information Fusion [J]. Journal of Computer Research and Development, 2009, 46(3):353~362.

[3] Noel S, Jajodia S, Wang L, Singhal A. Measuring security risk of networks using attack graphs [J]. International Journal of Next-Generation Computing, 2010, 1(1):135~147.

[4] Liu J, Jin T, Pan K, Yang Y, Wu Y, Wang X. An improved KNN text classification algorithm based on Simhash [C]. 2017 IEEE 16th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC). IEEE, 2017:92~95.