ATLANTIS PRESS
Advances in Social Science, Education and Humanities Research (ASSEHR), volume 130
2nd International Forum on Management, Education and Information Technology Application (IFMEITA 2017)

A Network Security Event Correlation Analysis Method Based on Attribute Similarity

Yue Gao[a], Shuying Zhang[*, b]

College of Computer Science and Technology,Beihua University

Jilin, Jilin, 132021, China

*Corresponding author

[a]lunagao@126.com, [b]jlzhangsy@126.com

Keywords: Attribute Similarity; Network Security; Security Events; Correlation Analysis
Abstract: On the basis of studying the characteristics of network security incidents and the methods of correlation analysis. The paper defines the network security event attribute similarity algorithm. A network event correlation analysis method based on attribute similarity is proposed, and the detailed description of the algorithm is given. The experiment proves that this method can effectively reduce the number of network security incidents and solve the problem of massive alarm.

Introduction

Due to the heterogeneity and complexity of the network, network security events are various. The resulting network security data is characterized by uncertainty, incompleteness, variability and ambiguity due to different sources and different collection methods. Because a lot of network events are actually the same kind of events, or similar events. If you don't deal with them and analyze them, you'll have a huge amount of data, and it is not conducive to network security status analysis and operation management. In order to better analyze and deal with network security events, we need to preprocess network security data. Considering the correlation of network security incidents, combining the same and similar network events, eliminating redundancy, reducing false alarm and alarm probability, and can improve the accuracy and efficiency of network security status evaluation and prediction.

At present, scholars at home and abroad have conducted a lot of researches on the correlation analysis method of network events. A number of methods are proposed. Z. J. Liu et al. proposed an alarm correlation algorithm based on compound attack path graph[1]. F. Chen et al. proposed a hierarchical network security assessment method based on threat propagation model[2]. G. Liu et al. proposed a method of network security assessment for orthogonal projection decomposition of reliability vectors [3]. H. H. Ge proposed a network security risk assessment method based on dynamic correlation analysis[4].Based on the study of the correlation theory of network correlation method and network traffic characteristics[5], this paper puts forward a security event correlation analysis method based on attribute similarity.

Attribute Similarity Definition

We know that every alarm record is made up of a series of attributes, Among them, the more important attributes are source IP, destination IP, source port, destination port, protocol type, time attribute, etc. To compare the degree of similarity between two alarms, you only need to compare these important attributes to calculate their respective degree of similarity. The value range of the similar degree is [0, 1]. The larger the value of the similarity, the closer the two attributes are, 1 represents exactly the same, 0 represents completely different. And then after each of the attributes are compared, the approximate degree of the two alarms can be calculated by using the formula of alarm similarity function. Finally, calculate its similarity with the whole clustering, and decide to add a new alarm to the alarm collection, or merge with an existing alarm.

Here is the definition of each of the similar functions:

1)Definition of the Similarity function of IP address

The composition of IP addresses is generally made up of 4 parts (IPV4) or 8 parts (IPV6). Each component can be represented as the corresponding binary number. IPV4 is a byte 8 bit and IPV6 is 2 bytes 16 bits. The Similarity function of the IP address is defined as formula (1) :

$$S(IP) = \frac{\sum_{i=1}^{m}(\sigma_i * S(a_i))}{\sum_{i=1}^{m}\sigma_i} \qquad (\sum_{i=1}^{m}\sigma_i = 1) \tag{1}$$

In the formula, $S(a_i)$ is the Similarity degree of each part of the IP address, $S(IP)$ is the total similarity of IP addresses, a_i can take the value of 1 to m, It represents each part of an IP address. m could be 4 or 8, σ_i represents the weight of the corresponding part from 1 to m.

2) definition of port Similarity function

Because the port has the characteristics of continuous distribution, the distribution of the attack on the port should have certain regularity, the Similarity of our definition ports is achieved by defining the distance between the two ports. The smaller the two ports are, they are adjacent ports, and the more likely it is to be an attack, the more similar it is. The greater the distance, the less likely they are to come from the same attack. We define a function that is similar to a port, as shown in formula (2) :

$$S(port) = \begin{cases} \frac{\lambda - |X_p - Y_p|}{\lambda} & |X_p - Y_p| \leq \lambda \\ 0 & |X_p - Y_p| > \lambda \end{cases} \tag{2}$$

In the formula, $S(port)$ represents the similarity of the port, λ is the threshold of the port we set. That's what we think of as the maximum distance between ports. If it's greater than this value, then it's close to 0, A and b represent the two alarm events X, Y value of the comparison, X_p and $Y_p$represent the port values of the two alarm events X,Y that participate in the comparison

3) the definition of the protocol similarity function

The similarity of the agreement is relatively simple, or the same, or different. Therefore, the measurement method is shown in formula (3) :

$$S(protoctol) = \begin{cases} 1 & X_{pt} = Y_{pt} \\ 0 & X_{pt} \neq Y_{pt} \end{cases} \tag{3}$$

In the formula, $S(protoctol)$ represents the approximation of the protocol, X_{pt} and Y_{pt}represent the corresponding port values of the two records that are compared.

4) definition of time similarity function

For attacks like DOS or DDOS. They're launching a lot of packets in a short period of time. these alarm records have a certain continuity in time, the distribution density is very concentrated, So when we're thinking about the approximation, we're going to think about the approximation of time. The time similarity function is defined as the formula (4).

$$S(Time) = \begin{cases} \tau - |X_t - Y_t| & (|X_t - Y_t|) \leq \tau \\ 0 & (|X_t - Y_t|) > \tau \end{cases} \tag{4}$$

In the formula: $S(Time)$ represents the time similarity, τ represents the threshold of time, That is to think,about the approximate degree of time in the τ range. If it's not in this range, we don't think it's similar, the value is 0. X_t, Y_t respectively represent the corresponding occurrence time of event *X* and *Y*.

5) definition of overall similarity function between alarm events

The overall similarity function of alarm events is defined as the formula (5) :

$$S(X,Y) = \sum_{f-1}^{n} \omega^{(f)} S_{xy}^{(f)} \qquad (\sum_{f-1}^{n} \omega^{(f)} = 1) \tag{5}$$

In the formula, $\omega^{(f)}$ represents the weight of the property *f*, *n* is the number of attributes, *f* is the source IP address, destination IP address, source port, destination port, protocol type, time, etc,

which are all properties that are used for similarity comparison. $S_{xy}^{(f)}$ is alarm *X, Y* corresponds to the similarity of attribute *f*.

6) the new alarm and the corresponding clustering overall similarity calculation function are shown in formula (6) :

$$S(n+1) = \alpha * S(n) + (1-\alpha) * S(X,Y) \quad (6)$$

In the formula, *S(n+1) i*s the similarity of the new cluster, $S(n)$ is the clustering approximation that already existed. $S(X,Y)$ is Q is the similar degree between the nearest alarm and the new alarm in the cluster. α is weight parameter.

The Description of Security Event Correlation Analysis Algorithm Based on Attribute Similarity

There are a lot of cyber security incidents that are not isolated, but there is a lot of connection, even many of them belong to the same alarm. In this way, there is a certain amount of similar information between the alarm information of this kind. We should combine this information with the same type of alarm into an alarm recording information. There is only one warning message, We know that a warning message is supposed to be a virtual alarm, and it's almost impossible to attack a single attack, and we should get rid of that information. In this way, we can associate and merge the alarm information, greatly reducing the number of alerts. The proximity of the alarm is measured by the similarity of the characteristic attributes of the alarm record. we need to define an algorithm here, The description of the algorithm is shown below:

Input: Alarm record Alertj which is preprocessed
Output: the result of the final alarm
Begin
T // Initializes, sets the threshold value of the similar degree
Time // Set the time initialization value
D=0 // set the initial similarity to 0
Alert-database={Alert1 , Alert2…Alertn} // Initializes alarm records in time
K =1 // counter
While (k<=n)
Begin
Get(Alert-database(k)) // Take the k alarm information
d= $a * S(old) + (1-a) * S(\mathrm{Alert}_j, \mathrm{Alert}_k)$ *// Calculate Alerj similarity*
If (d>D) then
D=d // Update similarity
Else
k=k+1 // counter
End
If (D<T) then
Add(new_ class) // Add new alarm class
Else
Add(new_alert) // Add new alarm
End

Simulation Experiment and Result Analysis

The effectiveness of the security event correlation analysis algorithm is needed to obtain a valid security event data source. In this experiment, we adopted the Snort network intrusion detection system in the campus network to detect the security incidents. Thus provides an effective security event data source. Three months of network security incidents were collected and analyzed.

To measure the reduction efficiency of alarm quantity, define the formula of reduced rate (7) :

$$\mathrm{R} = \left(\frac{Sum - N}{Sum}\right) * 100\%$$

In the formula, the *R* represents the reduction rate, *N* denoth the number of safe events after the associated processing, and *SUM* represents the number of original security events that are generated, According to the different approximate threshold (T), the experiment is carried out, and the change of rate is as shown in Fig.1:

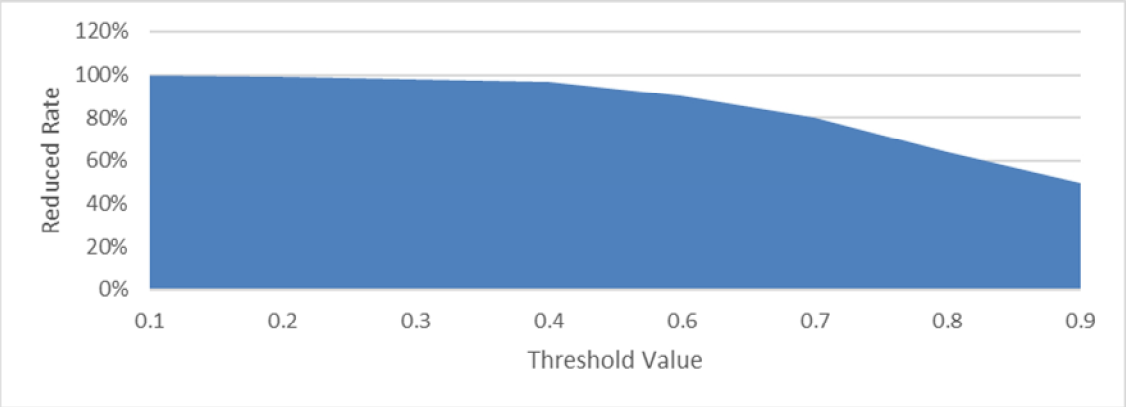

Fig.1 Reduced Rate Variation

As can be seen from Fig.1, with the setting of the similar threshold T, the downsizing rate of network events is very different. The larger the threshold T is, the more strict the criterion of similarity is, and the larger the number of related items is. On the contrary, the smaller the threshold setting, the higher the reduction rate, the less the number of events associated with it. When the threshold value is greater than 0.6, the downsizing rate is accelerated. After repeated experimental verification, the setting range of threshold T should be appropriate between [0.5, 0.7]. In addition, we can adjust other parameters to achieve more reasonable clustering results, such as time threshold value, port threshold value, IP address weight. All these reasonable Settings need to be tested over and over again, Whether the parameter is reasonable or not, the correlation analysis of network security events plays a crucial role, and has a great influence on the final classification result.

Conclusion

The paper analyzes the development of network correlation analysis technology and studies the development status of network security event correlation method. Then, a network security event correlation method based on attribute similarity is proposed and the detailed description of the algorithm is given. The experiment proves that this method can effectively simplify the number of network security incidents and solve the problem of massive alarm. It can generate high quality alarm collection, effectively reduce the number of false alarm, and provide beneficial help to improve the analysis of network security incidents.

Acknowledgment

The research work is supported by the 13th five-year plan project of education science in Jilin province (No.GH16061,No.GH170128).

References

[1] Z. J. Liu, C. J. Wang. An Alert Correlating Algorithm Based on Multi-Step Attack Path Graphs. Journal of NanJing University (Natural Sciences), vol.46 (2010) , p. 56-63

[2] F. Chen, D. H. Liu, Y. Zhang, et al. A hierarchical Evaluation Approach for Network Security Based on Threat Spread Mode, Journal of Computer Research and Development, vol.48(2011), p. 945-954

[3] G. Liu, Q. M. Li, and H. Zhang. Reliability Vector Orthogonal Projection Decomposition Network Security Risk Aassessment Method. Journal of Electronics & Information Technology, vol.34 (2012), p. 1934-1938
[4] H. H. Ge, D. Xiao Da, T. P. Chen, Y. X. Yang et al. Quantitative Evaluation Approach for Real-time Risk Based on Attack Event Correlating, Journal of Electronics & Information Technology, vol.35(2013), p. 2630-2636
[5] B. Li, F. Xie, Z. Chen, A Business-Oriented Risk Assessent Model, Computer Research and Development, vol. 48(2011), p.1634-1642