International Journal of Security and Its Applications
Vol.8, No.6 (2014), pp.353-364
http://dx.doi.org/10.14257/ijsia.2014.8.6.30

Research on Intrusion Event Sequence Correlation Method for Privacy Protection

Duan Xueying

Department of Intelligence Engineering, Jilin Police College,
Jilin 132000, China
17222433@qq.com

Abstract

For the shortcomings of traditional network security alarm correlation method, combined with the original alarm information privacy protection needs. On the basis of analyzing the characteristics of multi-step attack, proposed the use of sequential pattern mining techniques associated with rapid multi-step attack methods QSPM. And on this basis, proposed privacy protection security alarm multistep attack sequence pattern mining method PPSPM, Achieve the associated sequence of events following the invasion premise security. Experiments show that the algorithm makes quantitative analysis. And compared with typical sequential pattern mining algorithm. Results show the new methods have a positive accuracy and efficiency.

Keywords: *Network Security, Security Privacy Preserving, Sequential Patterns, Security Assessment; Data Mining*

1. Introduction

Detection of complex multi-step attack has always been the research focus and difficulty of safety alarm correlation. This attack is kind of a series of related attacks steps, and in a sequence of the form to complete, such as intrusion detection, vulnerability analysis, network penetration, privilege escalation attack, the back door opened, and log removal etc. In one attack step, some specific operation can be interchanged; the attacker will usually perform some action to conceal its intrusion intention and try to avoid intrusion detection equipment. Therefore, mastering the sequence relation intrusion events which found that complex attack logic connection between multiple attack steps. For the prediction and recognition of the attacker's next move, and then take the initiative defense measures to reduce the losses [1-2].

However, Multi step attack correlation method is most traditional reliance on expert knowledge, and the need for complex predefined rules. Less consider the privacy issues of original alerts. Some privacy protection method has been proposed in the field, the computational complexity is high. Therefore, the sequence association method to study intrusion events, and it has important and positive significance to extend this approach to privacy and security environment.

2. Related Theory

2.1. Sequential Pattern Mining

Many scholars in the study of the process of mining association rules discovered temporal association rules than the basic association rules Contains more extensive information, and it is more practical. They are engaged in mining cycle mode and event sequential pattern mining, and other related research work in the field of data mining and gradually formed a

ISSN: 1738-9976 IJSIA

new research direction-sequential pattern mining. Agrawal and Srikant in [3] proposed the problem of sequential pattern mining. Sequential pattern mining is the sequence of events associated with the relationship, to reveal the specific sequence of events that can be used to predict the occurrence of future events. And frequent pattern mining, a sequence pattern mining concern is "sequence". The mining results point out what frequent event (a) occurred in a certain order. And the sequence events must be different from the related transaction, and does not require these events simultaneously or successively. You can find the sequence relationship between target and characteristics of sequential pattern mining and complex multi-step attack the various steps of the mining has great similarity. This paper analyzes and summarizes the characteristics of multi-step attack, to study sequential pattern mining ideological intrusion alarm events associated methods

2.2. Analysis of Multi-step Attack Characteristics

It is different operating system platform [4], adopting a variety of vulnerabilities implement multi-step attack has been a lot of research, which, buffer overflow vulnerability using Solaris platform Sadmind procedures with the implementation of the remote DDos attacks on the target server. Buffer overflow vulnerability worm Sasser A and Sasser C using the LSASS service in Windows system to obtain full control permissions infected hosts. Buffer overflows attacks and FTP attacks against Microsoft SQL Server loophole. It summarized some characteristics of multi-step attack: The attacker can take a variety of ways to achieve the purpose of the attack. Multi-step attack is linked together multiple attack step. Each attack step the same multi-step attack has the characteristic of time series [5-6]

2.3. Sequential Pattern Mining Method

Due to the widespread application, the sequential pattern mining problem was put forward on the concern. This section describes some of the classic sequential pattern mining algorithm, and the comparison of these algorithms. Most of the classical algorithm for mining sequential patterns in association rules mining based on APRIORI theory. Any frequent pattern of sub patterns is frequent pattern. Based on this feature, researchers have proposed a series of APRIORI algorithm, Such as APRIORIAll, APRIORISome and GSP algorithm [7].

2.3.1. APRIORI Algorithm

As mentioned above, Agrawal and Srikant proposed the problem of mining sequential patterns in [1], and gives a set of mining algorithms. These algorithms process of sequential pattern mining is divided into the following five stages:

Sort stage. According to the mining goal, selecting relevant attributes from the databases to be mined, and according to the property of the selected to sort, the original database is converted to a sequence database.

L-item sets stage. Predefined support threshold, scan sequence databases to get all meet the threshold of frequent 1-itemsets. In practical applications, the frequent item sets are mapped to consecutive integers, in order to be efficient processing.

Conversion stage. Each sequence in the database with the maximal frequent item set which includes alternative.

Sequence of stages. The maximum 1- sequence as seed set, using APRIORI algorithm. Through an iterative algorithm, from the converted sequence databases to dig out all the frequent sequential patterns.

Maximal Sequential stages. By cropping, select maximal frequent sequence pattern from Stage 4 results are not included in other sequence mode [8-9].

2.3.2. GSP Algorithm

Srikant and others on the basis of previous work, we proposed a more efficient sequential pattern mining algorithm GSP-Generalized Sequential Pattern. The algorithm is still APRIORI algorithm. Compared APRIORI algorithm, GSP algorithm adds the following restrictions: Time constraints define the maximum and minimum spacing interval between two adjacent sequence parameters affairs constraints. If the time interval between adjacent transactions is not between the minimum and maximum interval, then that is not a sequence of two successive transactions affairs. The definition of a sliding window to extend business, allowing sets from different transaction. As long as the sliding window transaction time range of these items are located within a specified. Conceptual level classification of a sequence using the concept of hierarchical classification items. In calculating the support, the underlying adjacent elements can be used to support the higher level elements. These restrictions greatly reduces the number of candidate sequences, so that the efficiency of GSP algorithm than APRIORI algorithm [10].

3 Design of Intrusion Event Sequence Correlation Method

According to characteristics of multi-step attack, the attack behavior of multi-step attack sequence with features, but the attack steps specific multi step attack has a specific pattern. Based on the mining method in the sequence mode, this section uses the sequential pattern mining idea of multi-step attack related safety alarm. In the privacy of the original alarm demand, proposed method for intrusion events correlation by a sequential pattern recognition intrusion events. First, give the relevant definitions.

3.1. Definition

Definition 1 Attack Set

The safety alarm attack-type property values in a vector set, Represented by $S = \{s_1, s_2, \dots, s_g\}$, Where s refers to the different types of attacks, g represents the total number of attack types.

Definition 2 Attack Sequence

The sequence consists of aggressive behavior called the attack sequence, Represented by $< a_1, a_2, \dots, a_n >$

Definition 3 Sequence Support

Two attack sequence $A =< a_1, a_2, \dots, a_n >$ and $B =< b_1, b_2, \dots, b_m >$, If $a_1 = b_1$, and when the sequence A sequence of sub-sequence B, Said A sequence B contains sequence. At the same time, the definition of the conditions for the sequence, sequence B supports sequence A.

Definition 4 Maximal Attack Behavior

Meet the preset attacks a minimum support threshold is called the maximum attack behavior.

Definition 5 Maximal Attack Sequence

If an attack sequence is not any other sequence contains, say the attack sequence for the biggest attack sequence.

Definition 6 Support of Attack Behavior- SUP_{AB}

Expressed as AB: TB, where AB is the number of a specific attack behaviors global attack sequence a_i, TB is the total number of all attacks the global sequence. As shown in the formula (1)

$$SUP_{AB}(a_i) = \frac{AB}{TB} \times 100\% \quad (1)$$

Definition7 Support of Attack Sequence- SUP_{AS}

Expressed as CS: TS, where CS is the number of support attack sequence candidate attack sequence in A. TS is the total number of all candidate attack sequence. As shown in the formula (2).

$$SUP_{AS}(A) = \frac{CS}{TS} \times 100\% \quad (2)$$

3.2. Design of Quick Sequential Pattern Mining Algorithm

In order to improve the efficiency of intrusion event sequence pattern mining, based on the idea of Mining Attack Sequence Patterns (MASP), this paper presents Quick Sequential Pattern Mining (QSPM) algorithm. The algorithm does not need complex predefined rules, it can effectively find the security event between the sequence and causal relationship. QSPM mining process consists of the following five main stages.

3.2.1. Alarm Data Preprocessing

Before the alarm data sets for sequential pattern mining, For the purpose of preprocessing is to reduce the number of false alarm data set or redundant alarm. We adopt the following two methods of pretreatment.

Alert aggregation. An attacker will often perform their intrusion within a certain time interval.

[3] proposed a method. This paper presents a time constraint alarm aggregation. For a given time window W_p. The time constraint description is as shown in formula (3):

$$Max\{al.start-time \mid \forall al \in AL\} - Min\{al.start-time \mid \forall al \in AL\} < W_p \quad (3)$$

Alarm filtering. Intrusion Detection System (IDS) generated alarms usually contain a lot of false alarm, referred to as the false alarm. Characteristics of false alarm are continuously or periodically. But the real attack alarm triggered by the characteristics of the general is sudden or random. But does not rule out the existence of false alarm. Some of the normal system configuration information or service request information will also be deemed a security threat and trigger the alarm.

3.2.2. Constructing the Global Attack Sequence

Mapping. For efficiency factor, build a relationship based on mapping data attack-type character and integer data. Using a unique integer to represent a type of attack, as the subsequent mining phase of the key. In order to replace time-consuming string matching operation. Table1 shows the mapping method of sample.

Table 1. Example of Mapping

start-time	The remaining properties	attack-type	The results of mapping
05:35:48	…….	Community SIP TCP/IP Message Flooding Directed To SIP Proxy	2
05:37:42	…….	BAD-TRAFFIC SYN to Multicast Address	5
05:37:42	…….	Port Sweep	7
05:47:31	…….	Fragmentation Overlap	8

Global attack sequence generation. After the mapping process, to generate global attack sequence by the global attack sequence generator. According start-time property values after the mapping process, in ascending order to form a global attack sequence.

3.2.3. Generating Candidate Attack Sequence Database

The attacker usually perform aggressive behavior in a specific time period based on the viewpoint, scene definition will attack Set in the specified time window within the W intrusion events. On this basis, the time series properties for aggressive behavior design an iterative process. By traversing the global attack sequences to generate candidate attack sequence database. The specific process is as follows, Figure 1 describes the generation process of candidate attack sequence

a. Initialization time window, start time window for the first alarm start-time global attack sequence value.

b. Selects all start-time within the time window of alarm record

c. Moving time window, the beginning time alarm record for a start-time value;

d. Repeat steps b) and c), until the end of time window to the entire global attack sequence ends.

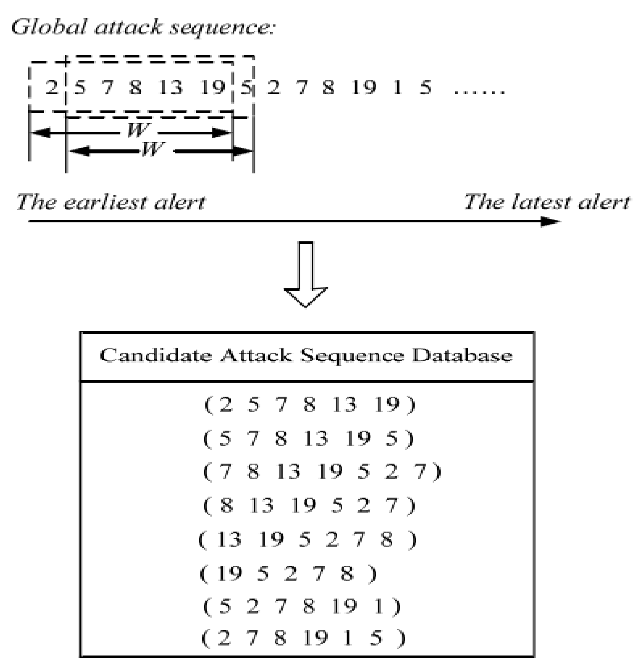

Figure 1. Example of Producing Candidate Attack Sequence Database

3.2.4. To Obtain Maximum Attack Behavior Set

In order to obtain the maximum attack behavior set, maximum attack behavior set processor is according to the logic of global attack sequence traversal as Algorithm 1 shows. Which met the minimum support 1–Sequence. The sequence contains only one attack, expressed as L_1 .QA is the total number of the largest global attack sequence Aggressive Behavior. In Figure 1 global attack sequences for example, Figure 2 shows the obtained maximum attack set L1, and the corresponding support. In this case the sequence of attacks and attacks the minimum support threshold were set at 15%.

```
For the attack type collection of S in each a_i attacks
If SUP_AB(a_i) ≥ Min − sup
    then a_i ∈ L_1
End If
End For
```

Algorithm 1. Pseudo Code for Getting L_1

1-sequence	SUP_{AB}
(2)	2/13
(5)	3/13
(7)	2/13
(8)	2/13
(19)	2/13

Figure 2. Example of L_1

3.2.5. Mining Maximum Attack Sequence

APRIORI characteristics of the known, maximum attack sequence elements must belong to the biggest attack behavior set. Each attack sequence represents maximum attack scenarios. Through the CAS multi round traversal, maximum attack sequence mining is first found small attack scenarios. Then gradually get more attack scenario. In the first round, we use L_1 as the initial seed set, potentially attack sequence new collection C_2 . During the traversal of CAS, count on the support of each sequence in C_2 , Screening satisfied sequence L2 min-sup. As the next generation of candidate maximum attack sequence of seeds. Each round will have a maximum attack sequence candidate maximum attack sequence new collection C_y and corresponds to the set L_y .y represents the maximum attack sequence length, y is the number of maximum attack attacks included the sequence. The process is repeated. Until you cannot get a new sequence from L_y to C_{y+1}. It means that all of the global attack sequence biggest attack sequences have been found.

3.3. Design of Privacy Preserving Sequential Pattern Mining (PPSPM) Algorithm

The security situation is increasingly grim highlights the invisibility and strong counter attack. Based solely on local data is difficult to detect intrusion collusion detection networks, springboard attacks and other types of complex multi-step distributed attacks. Sequential pattern mining is to obtain security intrusion event alarm sequence and causal relationship.

The association has important value for security alarm analysis. However, security alarm data contains difficult to obtain from the external scanning vulnerability information and all kinds of basic network deployment information. For concerns about privacy issues, organizations and institutions are often reluctant to publish and share their direct foreign safety data.

This section QSPM extended sequence pattern mining algorithm, and proposed Privacy Preserving Sequential Pattern Mining algorithm PPSPM. At the same time in mining valuable intrusion event sequence pattern protect alarm information privacy and security. PPSPM algorithm as shown in algorithm 2

Input: The original security alarm data set AL, Alarm sensitive attribute set $A_i(i=1...d)$, Sensitive to the importance of each attribute in the sequence pattern mining W_i, k values. Attack scenario time window W. Minimum support threshold min-sup.
Output: After the data set PPAL privacy intrusion alarm event sequence pattern set
Algorithm:
Adopting alarm aggregation, filtering and other methods to achieve AL data preprocessing, to get AL_n

```
FOR i=1 to d
IF Ai is discrete type
    THEN Entropy as a guide to design the ADH
ELSE IF Ai is continuous type
    THEN Differential entropy as guide to design their ADH
  END IF
END FOR
```

Algorithm 2. Pseudo Code of PPSPM

4. Experimental Analysis and Results

All experiments were performed on the data set DEF CON 9.0, the data set is the industry's attack scenario analysis and validation of authoritative data sets. We build an experimental environment in the network, the use of open-source replay tool TCPReplay to import the network traffic. At the same time in the network deployment of SNORT network traffic monitoring. PPSPM runs at CPU clocked at 2.0G, memory is 1G on the server. Part of the original network intrusion alert information in as shown in Table 2.

4.1. Validity Analysis

This experiment mainly used to evaluate the QSPM and PPSPM algorithms ability to find security attack sequence. Effectiveness that is to build a multi-step attack scenarios. Original experimental data set contains 5290 records intrusion alarm. After the data preprocessing, record reduced to 4833.

Assessed by two indicators R_C and R_S corresponding to quantify the accuracy of the algorithm. Such as formula (4) and (5) are shown.

Table 2. Raw Intrusion Alerts (Partial)

Intrusion alarm types	Association of alarm number
BAD-TRAFFIC Loopback IP	48
BAD-TRAFFIC SYN to Multicast Address	146
BAD-TRAFFIC TCP Port 0 Traffic	225
TCP Port Sweep 4	4
SNMP Request UDP	691
Fragmentation Overlap	88
SCAN FIN	232
Community SIP TCP/IP Message Flooding Directed to SIP Proxy	131
ICMP Icmpenum	799
BAD-TRAFFIC Loopback Traffic	48
SNMP Missing Community String Attempt	321
DDOS Mstream Client to Handler	780

$$R_C = \frac{\text{The number of the correct attack scenario}}{\text{The actual number of attack scenarios}} \quad (4)$$

$$R_S = \frac{\text{The number of the correct attack scenario}}{\text{The total number of attack scenarios excavated}} \quad (5)$$

Table 3 gives the accuracy statistical comparison of QSPM, PPSPM and GSP algorithm in multi-step attack correlation analysis. Because the APRIORI algorithm found many invalid attack scenario fragments in the course of the experiment, to cause great performance cost. Here we only give comparative of two algorithms and GSP data.

Table 3. Comparison of Attack Scenario Recognition

Comparison Project	QSPM	PPSPM	GSP
The actual number of attack scenarios	49	49	49
The total number of attack scenarios excavated	47	47	44
The number of the correct attack scenario	43	43	39
The accuracy evaluation index R_C	87.7%	87.7%	79.5%
The accuracy evaluation index R_S	91.4%	91.4%	88.6%

As can be seen, in the assessment of dimension accuracy, QSPM and PPSPM were better than GSP. Using the attack behavior sequential pattern mining, at least 87.7% of the alarm can be accurately Association. QSPM, PPSPM is on the basis of data on protection alarm, its sequential pattern mining capability is without any loss. The main reason is the algorithm using the attack-type attribute as the sequential pattern mining method key. Therefore, QSPM algorithm and PPSPM algorithm for the identification of complex multi-step attack intention has practical guiding significance.

4.2. Performance Analysis

We adopt 1% to 0.25% of the minimum support (min-sup) interval experiment. Figure 3-5 given QSPM, PPSPM, APRIORI and GSP algorithm runs at different time window period comparison.

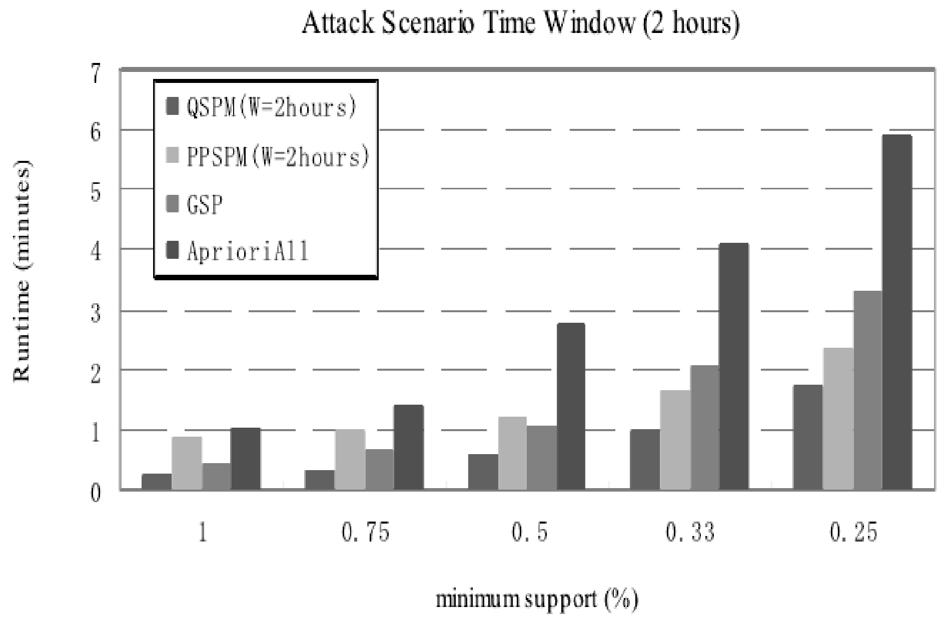

Figure 3. Runtime vs. Min-sup (W=2 Hours)

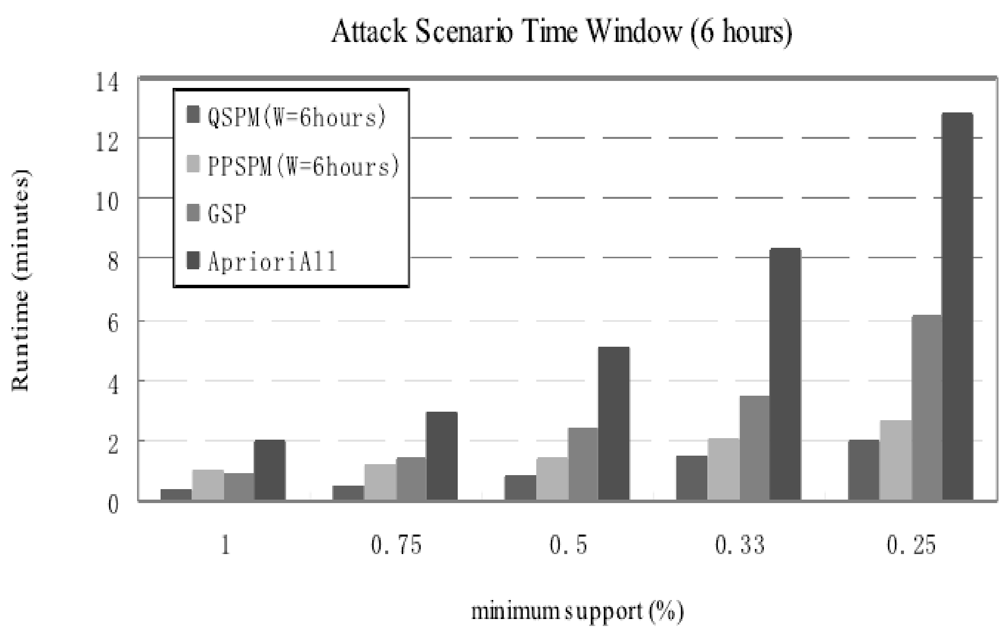

Figure 4. Runtime vs. Min-sup (W=6 Hours)

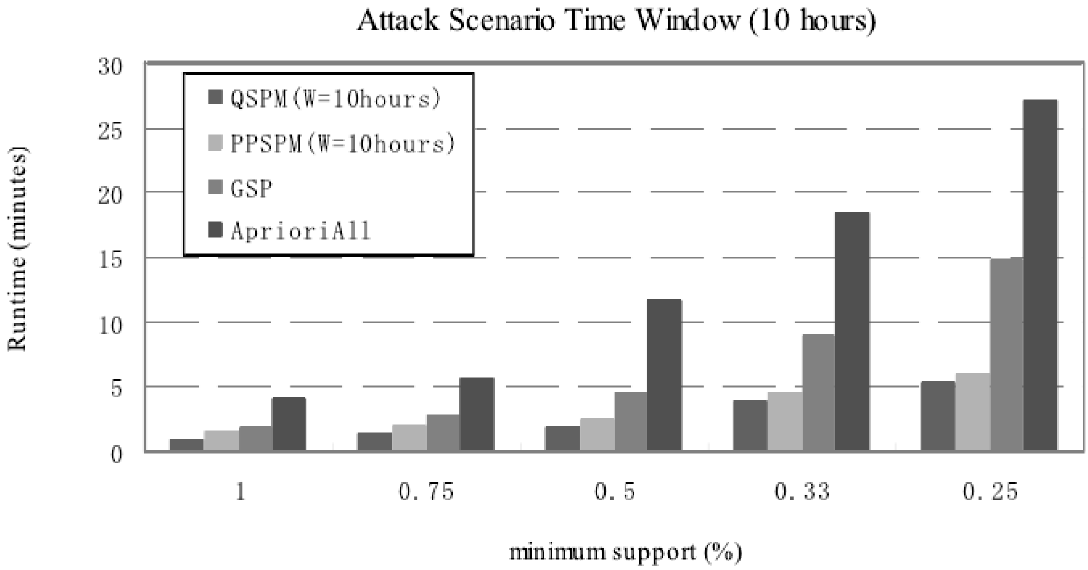

Figure 5. Runtime vs. Min-sup (W=10 Hours)

We can see from the experimental results. QSPM is faster the average 1.7-3 times than GSP. QSPM is faster the average 4-6.5 times than APRIORI. On the other hand, for a fixed time window. The overall operation cycle QSPM min-sup decreases with increasing values. The main reason is when the min-sup value decreases; the initial seeds set increases in the number of L_1 maximum attack. Therefore generate more candidates maximum attack sequence. At the same time, with the decrease of min-sup, Operation period of QSPM and GSP and the value of APRIORI are more. In the time window W smaller, minimum support threshold min-sup is larger circumstances. Due to the small length of the CAS candidate attack sequence, and fewer number of L_1 initial seed. PSPM algorithm requires data sets alarms privacy protection. Therefore, the operation period of PPSPM algorithm is higher than QSPM and GSP, Its performance is better than APRIORI algorithm. With the increase of W , and decreased min-sup. For GSP and APRIORI algorithms, PPSPM reflects the performance advantages are obvious.

5. Conclusion

In this paper, sequential pattern mining methods to achieve security alarms associated with multi-step attack. Combined with the original alarms privacy protection needs. Presents a fast algorithm for mining sequential patterns, mining algorithm QSPM and PPSPM sequence pattern based on privacy protection. The proposed method without expert knowledge need to get attack to scenario and Pre-Design Association Rules. To overcome the defects based on rules, security event causality correlation method. At the same time, by using the support degree evaluation method. Maximum attack sequence generation algorithm is optimized to reduce the time-consuming database scanning operation, it has the features of fast, accurate correlation multi-step attack. Finally, the experiment proved that the use of indicators to assess the accuracy and validity of QSPM PPSPM algorithm performed a quantitative analysis. And with typical sequential pattern mining algorithm APRIORI and GSP are compared. The experimental results show that the proposed algorithm has better performance in finding attacks sequence mode.

References

[1] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules", In Proc. 20th Int. Conf. Very Large Data Bases, VLDB, **(1994),** pp. 487-499.
[2] J. S. Park, M. S. Chen and P. S. Yu, "An effective hash-based algorithm for mining association rules", ACM, **(1995)**.
[3] R. Agrawal and R. Srikant, "Mining sequential patterns", In, Proceedings of the Eleventh International Conference on Data Engineering, **(1995),** pp. 3-14.
[4] W, Li, "Network multi-step research to identify ways to attack", Huazhong University of Science and Technology PhD thesis, **(2012).**
[5] L. Yimin, "Private data access object access control mechanis", Fudan University, **(2012)**.
[6] Daojing, "Wireless network security key technology research", Zhejiang University, **(2012).**
[7] Z. Dongfang, "The research on the key technology of 3G network identity authentication and content security", Beijing University of Posts and Telecommunications, **(2010).**
[8] L. Zhiyuan, "A mobile peer to peer network security key technology", Journal of Nanjing University of Posts and Telecommunications, **(2011).**
[9] C. Juan, "Wireless sensor network node location privacy protection and self healing technology", Journal of Harbin Institute of Technology, **(2013).**
[10] R. Srikant and R. Agrawal, "Mining sequential patterns: Generalizations and performance improvements. Advances in Database Technology—EDBT'96, **(1996),** pp. 1-17.
[11] Z. Rong, "Beijing University of Technology design and implementation", Windows system, network security scanning tool, **(2013).**
[12] S. Yaolin, "Legal protection of personal safety information network", Southwest University of Political Science and Law, **(2012).**

Authors

Duan Xueying, she is a lecturer and graduated from Changchun University of Technology with the major of computer application in April, 2009. Now teaches at the Department of information engineering in Jilin Police College. Research direction for network monitoring.