Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.12785/amis/081L06

An Alert Correlation Analysis Oriented Incremental Mining Algorithm of Closed Sequential Patterns with Gap Constraints

*Hui He, Dong Wang, Gui Chen and Weizhe Zhang**

School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

Received: 12 Apr. 2013, Revised: 7 Aug. 2013, Accepted: 8 Aug. 2013
Published online: 1 Apr. 2014

Abstract: Large-scale network attacks will bring great damage to the network. Although the existing detection systems are able to detect a large number of known attacks, when facing large-scale network attacks, log data generated by these systems usually increases rapidly, which forms vast amount of alert information in a short period of time. This paper researches on picking up alert information efficiently and timely, which is an important need. According to the characteristics of intrusion detection log, we put forward the method of using incremental mining algorithm of closed sequential patterns with gap constraints - cispan algorithm to analyze the growing log database, we also compare the performance of cispan algorithm, prefixspan algorithm and clospan algorithm in analyzing intrusion detection log, and proves that cispan algorithm has higher efficiency in analyzing alert log.

Keywords: alert analysis, sequential pattern, gap constraint, closed sequential pattern, incremental mining algorithm of closed sequential patterns.

1 Introduction

With a rapid development of network, network security is becoming more and more important. The attack techniques and tools of hackers become increasingly complicated and varied. Some hackers have been able to carry out some large-scale network attack. The well-known BT site - Mininova has suffered a large-scale botnet attack across three continents recently. Although the existing detection systems are able to detect a large number of known attacks, when facing large-scale network attacks, log data generated by these systems usually increases rapidly, which forms vast amount of alert information in a short period of time. How effectively and timely to pick up useful alert information from the large amount of alert log data is a complicated and meaningful work.

Many complicated intrusions have a fixed time sequences, such as when a hacker attacks, at first he often scans port, executes some specific codes to get special permission, carries out an attack [1] and so on, these acts will leave the same alert sequences in alert log. Picking up alert sequences above has an important guidingsignificance to analyze the true purpose of intruder. In order to accurately pick up the alert sequence, incremental mining algorithm of closed sequential patterns with gap constraints - cispan algorithm will be used in this paper to the analysis of alert log, it find high frequency of frequent sequential patterns, then analyzes the contract between the alert information. This paper also compares the performance of cispan algorithm, prefixspan algorithm and clospan algorithm in analyzing alert log, and proves that cispan algorithm has higher efficiency in analyzing the alert log.

In section 2, the related work about alert log analysis and sequence mining are introduced, in section 3, the related definitions and content of cispan algorithm are given, in section 4, how to use cispan algorithm to analyze alert log, and the performance comparison of cispan algorithm, clospan algorithm and prefixspan algorithm are also presented, finally, our conclusions of this paper are summarized in section 5.

2 Related Work

Some scholars at home and abroad have done some research work on analyzing intrusion detection log, and

* Corresponding author e-mail: wzzhang@hit.edu.cn

have achieved some results. Peng Ning, et al. put forward alert correlation analysis methods based on the precondition and subsequent results [2], and did research on a series of methods to analyze the alert correlation of large sets data. But this method requires a lot of expert knowledge, it's scalability is not strong and can't prevent new attacks. The adaptive alert correlation model A3PC based on pattern mining and cluster analysis was proposed by Zhihong Tian in his paper [1], he introduced the idea of anomaly detection into the problem of alert correlation, but this model needs to build a "normal" alert model and is easy to cause erroneous judgment. Yijun Sun, et al. put forward methods based on pattern mining, sequential patterns mining and so on to analyze alert log [3], but the method is too generalization, lacks pertinence, and will generate the number of redundant even useless information.

In the past ten years, a large number of experts and scholars have done research on sequential patterns mining algorithm, and put forward many algorithm, such as GPS, SPADE. Prefixspan algorithm uses the patterns-growth method instead of the previous method of generating candidate frequent patterns, it generates frequent prefix patterns, gets projection database which corresponds to the prefix patterns by projecting, then continues to search for the frequent sequential patterns in the projection database and connects with the prefix patterns to get frequent patterns, but sequential patterns mining algorithm can mine full set of frequent sequential patterns sets in the database of sequential patterns. Closed sequential patterns mining can get the same information as sequential patterns mining, but contains less and more compact sequential patterns. In research of closed sequential patterns algorithm Xifeng Yang, Jianyogn Wang, et al. put forward closed sequential patterns mining algorithm - clospan [5], bide [6]. Clospan algorithm is improved on the basis of prefixspan algorithm, it compares size of the projection database on the basis of the pruning strategy, rather than directly to determine the containment relationship between the sequential patterns to speed up the mining speed. However, clospan algorithm can only analyze the static database, when log database changes, clospan algorithm needs to analyze the entire database again. Ding Yuan, et al. proposed incremental mining algorithm of closed sequential patterns - cispan algorithm [10], which is improved on the basis of clospan algorithm, and only mines the new added sequential patterns, then comprehensives the original mining results of clospan algorithm, finally, gets frequent sequential patterns after increment. In this paper, according to the characteristics of alert information. we put forward the method of using incremental mining algorithm of closed sequential patterns to analyze the vast amount of alert log, and get some good results.

Many sequential patterns database have some characteristics of their own, such as DNA sequential patterns usually requires mining the continuous frequent sequential patterns. In the alert log, the alert information which are reflected by complex invasion behaviors also have limited gap in sequential patterns databases. Cludia Antunes, et al. proposed the prefixspan algorithm [7] with gap constraints, which is based on a new method of generating projection database, and can be used for analyzing such mining of sequential patterns with gap constraints, reduces useless frequent sequential patterns formation. Chun Li, et al. proposed the bide algorithm [8] with gap constraints to analyze closed sequential patterns with gap constraints, it has some advantages of bide algorithm, such as no need to generate candidate sequential patterns. In this paper, on the basis of cispan algorithm and according to the characteristics of gap constraints of alert sequential patterns, we add gap constraints and use this method to analyze alert log and reduce the formation of redundant alert information.

3 Cispan Algorithm with Gap Constraints

3.1 Related Definition

DEFINITION 1. Let $I = \{i1, i2, ..., i_n\}$ be a project set. A subset of I is called an item set. Sequence $s =< t1, t2, ..., tm > (t_j \in I, 1 \leq j \leq m)$ is an ordered item set, without loss of generality, assuming that all items in project set are already sorted in a certain order(for example, in alphabetical order). The size of sequence $|s|$ refers to the number of item sets in the sequence. The length of sequence is described below:

$$l(s) = \sum_{i=1}^{m} l|t_i| \tag{1}$$

It refers to the number of all items in sequence.

DEFINITION 2. Given two sequences $a =< a_1, a_2, ..., a_m >$ and $b =< b_1, b_2, ..., b_n >$, if there is a group of integers $i_1 < i_2 < ... < i_n$that makes $a_1 = b_{i1}$, $a_2 = b_{i2}, ..., a_m = b_{in}$, then a is contained by b, or a is a subsequence of b. It also can be expressed as $a \subseteq b$(If a is not equal to b, it can be expressed as $a \subset b$. If b contains a and their support degrees are equal, then b absorbs a). Givenδ, if $i_k - i_{k-1} \leq \delta$, then a is the $\delta-$ distance subsequence of b, it can be expressed as $a \subseteq b$. If $\delta = 1$, then a is the continuous subsequence of b.

DEFINITION 3. The support degree of sequence a is the number of sequences which are contained by sequential patterns database $-D$:

$$support(a) = |\{s|s \subseteq Danda \subseteq s\}| \tag{2}$$

Given a minimum support degree $-min-sup$, the frequent sequential patterns set (FS) contains all of the sequences that it's support degree is not less than $min-sup$.

DEFINITION 4. If sequence a is a frequent sequential patterns in sequential patterns database $-D$ and there is not the sequence b, which is super-sequence of sequence

a, which have same support degree as sequence a, then a is a closed sequential pattern.

DEFINITION 5. Given a sequence $s =< t_1, t_2, ..., t_m >$ and an item set a, $s \diamond_s a =< t_1, t_2, ..., t_m, \{a\} >$ is called sequence extension, we call S-extension, which means that we can generate a new sequence by adding a new item set a to the end of original sequence s. $s \diamond_I a =< t_1, t_2, ..., t_m$ $\bigcup \{a\} >$ if $\forall k \in t_m, k < a$ is called item set extension, we call I-extension.

According to the characteristics of the alert information, there is only one piece of alert message at any time, so in definition 1, $t_j, 1 jm$only have one item set, and it can be simplified into item. In this paper, sequence s is an ordered set which contains m items. In definition 4, if $< abc >$ is a closed sequential pattern, but $< ab >$, $< bc >$ are not closed sequential patterns. Since the size of each item set is 1, we only use S-extension to make frequent sequential patterns grow.

3.2 Cispan Algorithm with Gap Constraints

Yan [5], et al. used a new pruning method on the basis of prefixspan algorithm, he found two sequences s and s', if $s \subseteq s'$and $I(D_s) = l(D'_s)$, then for any of item C in project set D'_s, $support(s \diamond c) = support(s' \diamond c)$. According to the above findings, he proposed two pruning methods - backward sub-pattern pruning method and backward super-pattern pruning method. When extending the sequence s', at first we determine whether there is a sequence s that has extended, which makes $(1) s' \subset s$ or $(2) s \subset s'$, if so ,we can stop extending sequence s. When condition meets (1), we can directly stop extending s'. When condition meets (2), we don't extend s', instead, we directly transplant the offspring of s to the offspring of s'. When combining with gap constraints, the paper [8] put forward a perfixspan algorithm with gap constraints, when extending an element, the author puts forward a method, which records all the positions that the element appears in each of sequence in the database, rather than records the first positions that the elements appears in each of sequence in the database, but each of sequence in the sequential patterns database can only increase support degree count of the element at most once.

Cispan algorithm is future improved on the basis of clspan algorithm to speed up the speed of mining incremental database. Cispan algorithm divides the incremental operation into two steps - remove and insert. When sequence s grows for s', cispan algorithm first removes sequence s, then inserts sequence s'. Let I be the inserted sequence, let R be the removed sequence, let U be the unchanged sequence. Let IS be the frequent sequence that appears in I, let L_i be the prefix case that contains all sequences in IS. Let US be the frequent sequence that appears in U, let L'_o be the sequence that contains all sequences in US.

Cispan algorithm [10] is divided into three steps:

1.For each frequent sequence that appears in I, we call incclospan algorithm to mine L_i

2.Modifying L_o of the original database. When a sequence appears in R, reducing the count. When the count is less than the minimum support degree, removing corresponding node of this sequence, finally we get L'_o

3.Merging L'_o and L_i recursively. During merging, recursively traversing each node in preorder. Since L_i is the new inserted prefix case, when the structure of corresponding node of L'_o is not same as that of L_i, modifying the corresponding node of s. Step 2 and step 3 detailed see the cispan algorithm [10] of Ding, Yuan, et al.

The details of step 1 are described below:

STEP 1 Since the engineering need, we introduce the gap constraints and in the experiment use the method of generating projection database which was proposed by Chun Li, et al. [8] Because of the allowance of memory capacity, projection is pseudo projection, which directly records the positions that each item appears in the database into memory. GCincclospan algorithm is shown in below.

```
GCincclospanMining (D', min_sup,$\delta$,$L_i$)

Input: incremental database D', minimum support degree min_sup, parameter of gap constraints $\delta$

Output: all closed sequential patterns sets $L_i$ in $I$

  1.Find $s$ which meets conditions below

    $s$ appears in $I$

    $s$ is frequent in database $D'$

  2.For each sequence that appears in $I$

  3.Call incclospan (s, $D'_s$, min_sup, $\delta$ , $L_i$)

GCincclospan (s, $D'_s$, min_sup,$\delta$, $L_i$)

Input: sequence $s$, projection database $D_s$, minimum support degree min_sup, parameter of gap constraints $\delta$

Output: incremental prefix searching case $L_i$

  1.Check whether there is sequence $s'$ in $L$, which makes
    (1) $s' \subset s$ or (2) $s \subset s'$, and $I(D_s) = I(D'_s)$.
```

2. If there is s' that meets (2), then directly return. If there is s' that meets (1), then make the pointer of offspring of s' in L point to offspring of s, and return

3. If there isn't s', then inert s into L_i

4. Scan D_s, then find frequent item a that meets conditions below, and for s use a to S-extension

5. Only appear in each sequence's position beginning 1 to in D_s

6. If there are two sequences that contains a in D_s, but they are the same sequence in the original database, then increase the count only once

7. There is sequence in I that contains a

8. If there isn't such a, then return

9. For each a that meets 4

10. Call incclospan ($s \diamond a$, $D_s \diamond a$, min_up,δ, L_i)

Return

In step 1, X. Yan, et al. skillfully use hash table to find all the sequences which have equal $I(D_s)$, then judge whether there is inclusive relationship, this method greatly improves the algorithm's efficiency (complexity is optimized from $O(n^2)$ to $O(n)$). The main difference between GCincClospan algorithm and clospan algorithm is that for GCincClospan the mined frequent sequence must appear in I and it introduces gap constraints.

4 Experimental Results and Analysis

Experimental data is the testing sample - Lincoln Laboratory Scenario(DDos) 1.0 [9] which is provided by DARPA 2000, experimental platform environment is 4 cores inter(R) Xeon(TM), CPU frequency is 3.2 GHZ, memory is 4G, OS is Linux, kernel version is 2.6.9, compiler is GCC 3.2.3, programming language is c++.

We put destination IP and source IP in each alert record as an IP pair. In experiment, we use two red-black tree's structures to make IP and s_id in sequential patterns database constitute one-to-one mapping relationship, according to the order of alert time relationship, alert information of the IP pair in log constitute the corresponding sequence of its mapping sequence - s_id.

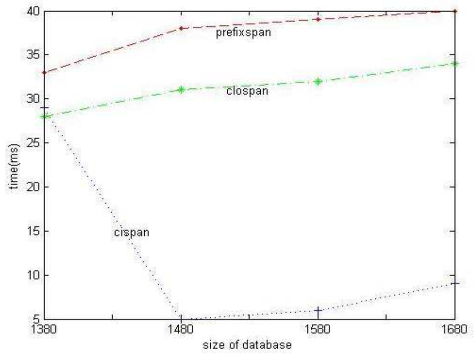

Fig. 1: Performance comparison for 1681 log number

We set minimum support degree - min_sup to 3, gap constraint to 1, according to the help document [9] of MIT Lincoln Laboratory, the attacker first scans the whole network segment to determine which host running the sadmind service, then uses the sadmind vulnerability to carry out buffer overflow attack, and successfully controls three machines - 172.16.112.10, 172.16.112.20, 172.16.112.50. We use prefixspan algorithm, clospan algorithm and cispan algorithm to analyze alert log, and successfully pick up the alert sequence information from three invaded machines.

We compare the spending time of prefixspan algorithm, clospan algorithm and cispan algorithm, and separately analyze the logs that their log number are 1681, 85930 and 688134. For each log we divide the log into four paragraphs to simulate incremental process, at first use program to analyze the first paragraph log, then use the second paragraph log as increment, make program continue to analyze the new generation of log database, and use the third and the fourth paragraphs of log as increment in turn with program analyzing. Each paragraph's increment of three logs are 100, 5000 and 20000. The results are shown in figure 1, figure 2 and figure 3, in first processing time which three algorithms need is similar, in the later log growth, cispan algorithm only needs to analyze the new incremental log, which greatly accelerates the processing speed, however, clospan algorithm and prefixspan algorithm have to deal with the whole sequential patterns database again. All in above confirm that in analyzing the intrusion detection log, cispan algorithm based on increment is better than prefixspan algorithm and cloapan algorithm in efficiency.

We also compare the performance of three algorithms in information compression. Cispan algorithm and clospan algorithm gets 133 items of information separately when analyzing the log of 1681 log number, and prefixspan algorithm gets 211 items of information. When analyzing

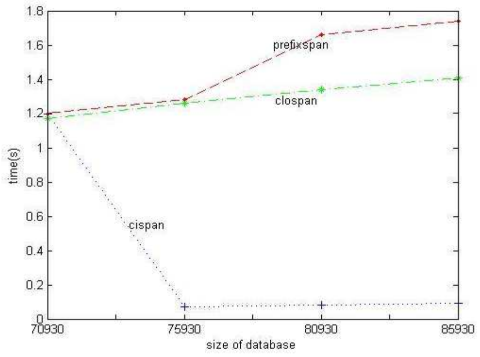

Fig. 2: Performance comparison for 85930 log number

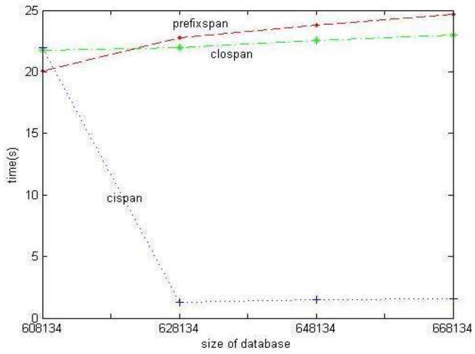

Fig. 3: Performance comparison for 85930 log number

the log of 85930 log number, cispan algorithm gets 168 items of information, and prefixspan algorithm gets 275 items of information, the information that prefixspan algorithm is more than cispan algorithm, which is confirmed to be redundant information. All in above confirm that cispan algorithm and clospan algorithm have a better information compression effect than prefixspan algorithm and reduce about 30% information. The same results of cispan algorithm and clospan algorithm also confirm the correctness of cispan algorithm in incremental mining of closed sequential patterns.

5 Conclusion

This paper use the cispan algorithm with gap constraints to analyze snort log, analyze the growing log database, the performance is better than the primitive clospan algorithm and prefixspan algorithm, at the same time, real-time alert performance and vast log information compression are improved to a certain extent. But cispan algorithm often takes up a lot of memory to save the prefix case information which has been analyzed, when analyzing large amount of log, there may be running out of memory, this problem is worth further researching and improving for us in order to find a more appropriate algorithm to analyze alert log.

Acknowledgement

This work was supported in part by NSFC under Grant Nos. 61173145, National High Technology Research and Development Program of China under Grant No. 2010AA012504 and 2011AA010705, National Basic Research Program of China under Grant No.G2011CB302605. This paper is an extended version from the conference of ISA 2012.
The authors are grateful to the anonymous referee for a careful checking of the details and for helpful comments that improved this paper.

References

[1] Tianhong Zhi. Research on Enhancing the Credibility of Intrusion Detection System. Doctor Thesis of Harbin Institute of Technology, (2006).

[2] Peng Ning, Yun Cui, and Douglas S. Reeves.Analyzing Intensive Intrusion Alerts Via Correlation. In Proceedings of the 5th International Symposium on Recent Advances in Intrusion Detection. Zurich, Switzerland, 74-94 (2002).

[3] Yijun Sun, Hongli Zhang, Hui He. Large-scale Network Security Situation Analysis with the Association Rule Mining. The National Network and Information Security Technology Symposium, **I**, (2007).

[4] J. Pei, J. Han, B. Mortazavi-Asl, Q. Chen, U. Dayal, and M.C. Hsu. PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth. In ICDE'01, Heidelberg, Germany, April (2001).

[5] M. X. Yan, J. Han, and R. Afshar.CloSpan: MiningClosed Sequential Patterns in Large Databases.In SDM'03, San Francisco, CA, May (2003).

[6] Jianyong Wang and Jiawei Han BIDE: Efficient Mining of Frequent Closed Sequences.Int.Conf.on Data Engineering, (2004).

[7] X. Ji, J. Bailey, G. Dong. Mining Minimal Distinguishing Subsequence Patterns with Gap Constraints. ICDM'05.

[8] Chun Li, Jianyong Wang. Efficiently Mining Closed Subsequences with Gap Constraints.SDM, 313-322 (2008).

[9] F. LLS_DDOS_1.0.
http://www.ll.mit.edu/mission/communi-cations/ist/corpora/ideval/data/2000data.html.

[10] LLS_DDOS_1.0.
http://www.ll.mit.edu/mission/communications/ist/corpora/ideval/data/2000data.html.

Hui He received the B.S., M.S. and Ph.D. degree in computer science from Harbin Institute of Technology, Harbin, China. Since September 1999, she has been with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, where she became an Associate Professor in October 2007. Her research interests include network computing, network security

Dong Wang received the B.S. degree from Harbin Institute of Technology, Harbin, China, in 2010. He is currently M.S. student of the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, in 2010. His research interests include parallel computing, cloud computing

Gui Chen was born in Guang Dong in 1990. He is currently B.S. student of the School of Software Engineering, Harbin Institute of Technology, Harbin, China, in 2009. His research interests include parallel computing, cloud computing

WeiZhe Zhang received the B.S.,M.S. and Ph.D. degree in computer science from Harbin Institute of Technology, Harbin, China. Since August 2003, he has been with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, where he became an Associate Professor in October 2007. His research interests include network computing, parallel computing. He is the corresponding author of this paper.