Journal of Theoretical and Applied Information Technology
20th April 2013. Vol. 50 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

THE APPLICATION OF APRIORI ALGORITHM FOR NETWORK FORENSICS ANALYSIS

XIUYU ZHONG
Prof., School of Computer Science, Jiaying University, Meizhou, Guangdong, China
E-mail: wch@jyu.edu.cn

ABSTRACT

With frequently network attack crimes, it causes serious economic loss and bad social influence. Network security products are practically impossible to guard against intrusion methods, network forensics is needed. The massive network data must be captured and analyzed in network forensics, and the data is often related, the application of Apriori algorithm is proposed for network forensics analysis. After capturing and filtering network data package, and the Apriori algorithm is used to mine the association rules according to the evidence relevance to build and update signature database of offense, current user behavior is judged legal or not through pattern match results of user behavior and association rules which are stored in databases. The crime behaviors are saved in evidence database, which can be used as primitive evidence for network forensics. Simulation results show that the application of Apriori algorithm can raise the speed, exactitude and intelligence of data analysis for network forensics, the application can help to resolve the real-time, efficient and adaptable problems in network forensics.

Keywords: *Apriori algorithm, Application, Network forensics, Data analysis*

1. INTRODUCTION

Along with the popularization and development of internet, network security is confronted with more and more severe threat. Network security products, such as firewall and intrusion detection system, which often cause excessive large false alarm linked to wrong correspondence, are practically impossible to guard against the intrusion methods, network forensics is needed. For each kind of crime, there is often association in crime time, crime tool and crime technology. Using data mining technology, it can discover the relations of events and the related data of specific crime.

Many scholars have made research on network forensics and application of frequent sequence mining algorithm. E.J. Palomoa had presented a novel approach to analyse and visualising network traffic data based on growing hierarchical self-organising maps (GHSOM), the GHSOM was used to cluster network traffic data and to represent this in a manner [1]. Emmanuel S. Pilli had presented an overview on network forensics covering tools, process models and framework implementations [2]. Clay Shields had presented a system named Proactive Object Fingerprinting and Storage (PROOFS) that continuously and efficiently creates fingerprints based on the contents of files [3]. Robert Beverly had developed a tool that can be used to automatically extract memory information [4]. Bilal Shebaro had proposed a privacy-preserving method for recording and storing network flow records, network operators can use to enforce a privacy policy [5]. Simson Garfinkel had presented an abstract differencing strategy and applies it to all of these problem domains. Use of an abstract strategy allows the lessons gleaned in one problem domain to be directly applied to others [6]. M.Y. Su had proposed a real-time NIDS with incremental mining for fuzzy association rules [7]. J. Pei had proposed mining sequential patterns by patter-growth [8]. C. Lei had explored how to efficiently maintain closed sequential patterns in a dynamic sequence database environment [9]. F. Wu and S.W. Chang had proposed item-transformation methods to mine frequent patterns [10]. Syed Khairuzzaman Tanbeer had presented a compact pattern tree that captured database information with one scan and provided the same mining performance as the FP growth method [11]. En Tzu Wang and Guan-ling Lee had proposed sanitization algorithm to modify databases for hiding sensitive patterns [12]. Apriori algorithm is one kind of most influential mining Boolean association rule algorithm, the application of Apriori algorithm for network forensics analysis can improve the credibility and efficiency of evidence.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Section 2 presents the model of network forensics based on applying Apriori algorithm. In section 3, Apriori algorithm is introduced. Section 4 presents the application of Apriori algorithm for network forensics analysis. In section 5, the result and analysis of test is given. Section 6 gives a conclusion to the whole paper and the further work.

2. THE MODEL OF NETWORK FORENSICS BASED ON APPLYING APRIORI ALGORITHM

2.1 Logic Design

The model of network forensics based on applying Apriori algorithm is shown in Figure 1. The system mainly consists of four parts: Data capture, intrusion detection system (IDS), data mining and data analysis.

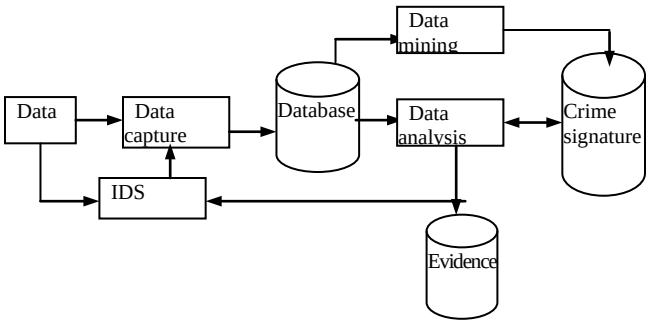

Figure 1. The Model of Network Forensics Based on Applying Apriori Algorithm

The main function of the system is recording the data package of network truly and completely, analyzing to each data package according to the protocol, and analyzing TCP, UDP, IP, ARP, ICMP as well as the partial application layer protocol. The system defines filtration rules in order to filter the first floor data package, and analyzes the filtering data according to the legal science and the information security analysis technology. The system finally matches the pattern, detects attack and extracts evidence. In order to build crime signature, the system analyzes program execution and user's behavior sequence associations with data mining technology, and analyzes the signature association of crime time, crime tool technology in the common each kind of crime. In order to discover the new crime behavior signature, the system extracts association signature of different crime, and excavates the signature of different attack form and associations of the different evidence in identical event.

2.2 Physical Design

The system skeleton is mainly composed of the attack computer, monitor and analyzer. The monitor has two network cards, one network card connects monitoring network, and the other network card connects the institute net which the analyzer is in. Monitor records primitive network data, captures data timely and processes preliminary to the data. Then it analyzes network data package, deposits and withdraws each kind of package section information according to the definition data model. After reading the network data and the diary which is recorded by the attack computer, the analyzer mines data and analyses data, and displays some unusual data. Meanwhile it promulgates the method of new crime and produces the new crime signature database for next data package analysis.

3. APRIORI ALGORITHM

The Apriori algorithm is one kind of most influential mining Boolean association rule algorithm, and the rule is expressed by frequent item collection. The association rule has two important attributes: Support level P(A UB), namely the probability of the two items of collections A and B which simultaneously appear in business collection D. Confidence level P(B|A), namely probability that collection A appears in items of in business collection D, items of collection B also simultaneously display. The rules which simultaneously satisfy the smallest support threshold value and the smallest confidence level threshold value are called the strong rule. Giving business collection D, the association rule mining creates the rules whose support and confidence level is bigger separately than the smallest support and confidence level which the user assigns. The Apriori core algorithm has used the recursion method in order to produce all frequency collections.

4. THE APPLICATION OF APRIORI ALGORITHM FOR NETWORK FORENSICS ANALYSIS

4.1 Network Forensics Analysis

The module analyzes the captured data from data capture module. The analyzer mainly carries on the analysis from two aspects. On one hand, the analyzer captures network integrated data packet in high webpage and matches it with signature information table item of network unusual data packet in the database. If match, the system sends out the warning, notices manager, and stores connected information in the result database. On the other hand, it analyzes logs and compares with the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

new record item and diary signature information table in unusual databases, and then reports to the manager and stores connected information to the result database if matched. The concrete processes mainly are shown as following:

(1) The key and the certificate in the monitor and the analyzer are leaded in, and the SSL security connection is established.

(2) The entity class of data packet in the Vector set is send to the analyzer through the SSL transmission.

(3) The entity class is deposited to Vector set and simultaneously saved into MySql database for the next analysis step.

(4) The analyzer backups and analyzes entity class of data packet that comes from the monitor, and displays the abnormal data, which takes as the crime evidence.

4.2 Application of Apriori Algorithm

The first step of the algorithm is to discover all frequency collection. These appearing frequency of items collections are not smaller than the minimum pre-definition support level. The second step is to generate the strong association rule by the frequency collection. The algorithm mainly contains the four functions: function creatC1 () obtains the first candidate collection, function getL () obtains the frequent collection, function getC () obtains the candidate collection, function count () calculates the record number of the candidate collection. The function getL () obtains frequent collection as follows:

(1) Begin

(2) if (i>=list.size) then goto (8);

(3) q=list.get(i+k-1);temp=q.count/ length;

(4) if (temp>=0.3) then goto (6);

(5) i=i+k, goto (2);

(6) if (n>=k) then goto (2);

(7) list.remove(i), goto (6);

(8) End.

The algorithm final outcomes the frequent record compendium, which corresponds to an instance of class Node. The class attribute significance is shown as follows:

(1) Count means the quantity of the record appearance.

(2) Column is a Vector set, which indicates the set of dimension and stores the record field.

(3)Data is a Vector set which stored corresponding value set of field. The frequent record set produces the strong association rules, which are stored in the signature database.

On the stage of data analysis, current user behavior is judged legal or not through match results of user behavior and association rules which are stored in databases. From judging that the behavior has the crime signature or is related to some crime, crime evidence which may be crime is withdrawn. The application of Apriori algorithm in data analysis for network forensics is shown in Figure 2.

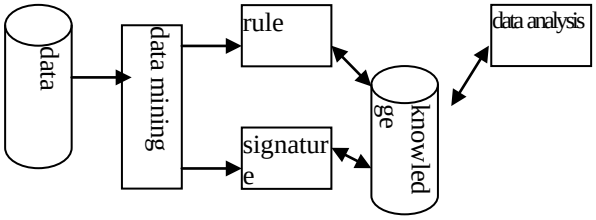

Figure 2. Application Flow of Apriori Algorithm

4.3 The Case

In the SYNFlood attack forensics, an example of Apriori application is given. Criminal sends massive SYN connection requests to the destination host in short time, but each connection is only half connection in direction. That will cause the connection request buffer of the destination host occupied completely, thus make the destination host unable to handle new connection requests, and the criminal achieves the purpose that the host will refuse to serve. This kind of situation will display in connecting records, in short period of time, some ports on a host can receive a number of “50” connecting state, which means half connection. The server is taken as the axis attribute and the destination host is taken as the quotation attribute to discover the frequent sequence service pattern of the same destination host in two seconds window time. Using Apriori algorithm obtains patterns of implicit crimes, which indicates that the HTTP service of the host has encountered the SynFlood attack. There are not “50” patterns in normal data because half connection cannot frequently occurred in a short time.

In the port scanning attack forensics, before the aggressor carries on attack to destination host, he must scan each port of destination host to find loophole. Aggressor must attempt to connect each port of destination host to observe whether this kind

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

of service is started or not. Therefore there are many records that a host receives multiple connections to different ports in short period of time. In the normal condition, many accessing services of port are prohibition and the connections with the port are rejected, which would have much "REJ" marks in connecting records.

The destination host takes as the axis attribute, simultaneously also takes the quotation attribute in order to discover the frequent sequence pattern of the same destination host in two seconds window time. The patterns of implicit attacks are obtained using the Apriori algorithm.

5. TEST RESULT AND ANALYSIS

5.1 Test Step

The crime takes as the client, the monitor and the analyzer take as the server. Crime tool is used for simulation tests in the campus network; the test process is as following:

(1) Server terminal environment is set before catching data package, the database is built, the monitor captures data packet and saves the original data into the database.

(2) SSL is set up, keys and certificates are loaded, and data is transferred from the monitor to the analyzer.

(3) Simulated intrusion is carried on under the unimpeded monitor network.

(4)The data from the crime is saved and the backup data is read.

(5) In server terminal, the system emergency alarm is watched and recorded, the emergency alarm is analyzed and the invasion data is recorded.

(6) Invasion data can be inquired and the invasion process can be analyzed. The analysis time slice may be set and the data analysis is carried on. It also may carry on the data analysis after increasing your own signature description.

The smallest support can be also set, new signature of crime may be joined into the signature database at any time to improve the extensibility of system and the identification ratio of crime.

5.2 Test Result and Analysis

The analogous system has realized functions: crime detection, evidence saved, reappearing criminality and evidence inquiry. The packet of simulated test comes from Ping Flood, SYN Flood and TFN2K tools. The function of evidence saved confirms and filters and saves and backups to the related record. The function of reappearing criminality can reduce network crime data, such as WEB, E-mail, FTP and so on, reproduce the network invasion, analyze new methods and tools of crime, and take them as the basis for lawsuit. The function of evidence inquiry query by the IP address, MAC address, port number, protocol type, analyze unusual phenomena in data flow quickly.

In performance test, simulated attack crimes such as Ping Flood, SYN Flood, ARP deception and TFN2K have been tested many times, especially to the SYN Flood attack. When the minimal support of mining algorithm is set as 0.3 and minimum confidence is set as 0.8, the detection ratio and false alarm ratio to different attacks type and continuous seconds are shown in table 1.

The test results show that the network forensics system based on Apriori application can detect attack crime accurately, with the increase of attack continuous seconds, the detection ratio would be improved, the system has a high detection ratio and low false alarm ratio. In addition, the system can detect some new attack crimes. The speed and detection ratio in different minimal support of mining algorithm are tested, the test results show that minimal support is smaller, the speed is slower, rules are generated longer, generated rules are more and false alarm ratio will be higher. If the support is set too big, it will be opposite, but the detection ratio may be lower, and it will cause higher miss probability. Many experiments show that minimal support is set as 0.3 and it will be balance on the speed, detection ratio and false alarm ratio.

Table 1. *Test Results of Different Typical Attack for Different Duration*

Attacks type	Continuous seconds	Detection ratio	False alarm ratio
Ping Flood	20	95.63	0.51
	40	97.81	0.64
	60	98.57	0.85
SYN Flood	20	96.33	0.66
	40	97.56	0.68
	60	99.28	0.72
TFN2K Tools	20	94.86	1.01
	40	96.52	1.02
	60	97.68	1.13

6. CONCLUSION

The data analysis is an important step for network forensics. The application of Apriori algorithm for network forensics analysis was

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

studied. After fetching the network data package, the data is pretreated and mined the frequent sequence excavation to obtain the crime signature patterns, the patterns is applied for network forensics data analysis. Apriori algorithm is used to build and update the signature database; it can reduce the number of matching times greatly and improve the efficiency of crime detection. Simulation results show that the application of Apriori algorithm can reconstruct the crime behavior integrally, improve the efficiency of network crime behavior recognition, and make evidence integrity and legal efficiency.

The further work will be:

(1) Deeper research on the network forensic analysis based on mining diary.

(2) Automatic production of signature database to the general attack and the crime.

(3) Research and enhance the evidence law efficiency, including the principle of legality, the process standard, the technical standard, the strategy and so on.

(4) Applying other data mining method, such as grid computing and honey net technology.

ACKNOWLEDGMENT

The authors want to thank the Science and Technology Innovation Project of Guangdong Province and the Natural Science Foundation of Guangdong Province for their general support for the research (with grant NO. 2012KJCX0097 and 9151009001000043 respectively).

REFERENCES:

[1] E.J. Palomoa, Application of growing hierarchical SOM for visualisation of network forensics traffic data, Neural Networks, Vol. 32, no. 16, 2012, pp. 275–284.

[2] Emmanuel S. Pilli, Network forensic frameworks: Survey and research challenges, digital investigation, no.7, 2010, pp. 14–27.

[3] Clay Shields, Ophir Frieder, Mark Maloof, A system for the proactive, continuous, and efficient collection of digital forensic evidence, digital investigation, no. 8, 2011, pp. 3-13.

[4] Robert Beverly, Simson Garfinkel, Greg Cardwell, Forensic carving of network packets and associated data Structures, digital investigation, no. 8, 2011, pp. 78-89.

[5] Bilal Shebaro, Jedidiah R. Crandall ,Privacy-preserving network flow recording, digital investigation, no. 8, 2011, pp. 90-100.

[6] Simson Garfinkel, Alex J. Nelson, Joel Young, A general strategy for differential forensic analysis, Digital Investigation ,no.9,2012, pp.50-59.

[7] M.Y. Sua, G.J. Yub, C.Y. Lin. A real-time network intrusion detection system for large-scale attacks based on an incremental mining approach, Computers & security, no.28,2009, pp. 301–309.

[8] J. Pei, J.W. Han. Mining Sequential Patterns by Patter-growth: The Prefix Span Approach. IEEE Transactions on Knowledge and Data Engineering. Vol. 6, no.10, 2004, pp.1-17.

[9] C. Lei, T.J. Wang. Efficient algorithms for incremental maintenance of closed sequential patterns in large databases. Data & Knowledge Engineering. Vol. 68, no.1, 2009, pp. 68-106.

[10] F. Wu, S.W. Chang. A new approach to mine frequent patterns using item-transformation methods. Information Systems. No.32, 2007, pp. 1056–1072.

[11] Syed Khairuzzaman Tanbeer, Chowdhury Farhan Ahmed. Efficient single-pass frequent pattern mining using a prefix-tree. Information Sciences, no.179, 2008, pp.559–583.

[12] En Tzu Wang, Guanling Lee. An efficient sanitization algorithm for balancing information privacy and knowledge discovery in association patterns mining. Data & Knowledge Engineering, no.65, 2008, pp. 463–484.