ISSN 2047-3338

Data Classification for Recognizing the Web Application

Tran Thi Dung[1], Trinh Ngoc Minh[2] and Tran Van Lang[3]
[1,2]Information Security Lab, Vietnam National University-HCM City
[3]Institute of Applied Mechanics and Informatics, Vietnam Academy of Science and Technology
[1]dungtt@vnu-itp.edu.vn, [2]minhtn@isepro.vn, [3]tvlang@vast-hcm.ac.vn

***Abstract*—The paper presents a solution to distinguish the different applications running on port 80 with HTTP protocol; from that helping firewall to detect the attack and to try to prevent it. This is relatively a typical problem, by there are many different applications such as game, and download, video, etc. are running on HTTP port 80. Filtering based only on port 80 would not realize which application is unallowed and which application is allowed. The solution of this paper is recording all network transactions when the application is running; building a module which analyzes the information, and then organize them into vector information using some of related parameters; using classification C5.0 algorithm (upgrade of C4.5 algorithm) to produce a decision tree; firewall uses this decision trees to identify/distinguish the applications. The testing results on the computer network system of the Information Security Lab (ISeLAB) using 230 transactions; the accuracy reached 99.5%.**

***Index Terms*—Firewall, Behavior IDS, Data Mining and HTTP**

I. Introduction

NOWADAYS, the firewall is an irreplaceable component of any computer network. Two core functions of the firewall are recognize the information flow passing through the firewall and to deal (permit/deny for example) with the packet basing on collected information [4]. Identifying the packet is usually done through their parameters (such as IP address, service port address and so on), or others characterized by the application data. Data identification features could be deployed on separate and specialized equipment named IDS (Intrusion Detection System). The combination of identification features of IDS with prevention features of the firewall for stream/packet formed the Intrusion Prevention System - IPS. This device has the ability to detect and prevent attacks/unallowed service. Ability to identify the application's information (application awarded) plays a very important, even critical role for the effectiveness of the firewall. It is clear that if firewall/IPS does not know the information traffic well, the next everything is pointless, even harmful.

Ability to identify information of IDS/firewall has been developed in both "height" and "width" directions. “Height” development is to develop the ability to recognize the application by using the information at higher layer in the OSI model. Development in “width” means the extension of ability to recognize the application using the information from multiple packets, over a longer period. Comparing the former firewall that only works with the information of 3rd layer and each individual packet with the nowadays firewall, it is clear that there is an outstanding development of current information screening technology, characterized by the ability to identify information at the application layer, as well as along the data stream. The purpose of this paper is also in this research and development trends, and further in the web-based application recognizing by analyzing the header of the session of the HTTP protocol.

The difficulty that firewall facing to now be there are too many different applications using the HTTP protocol of the application layer. For example, browsing web, watching videos, listening to music, downloading file, playing games, etc. All these applications have in common characteristic is using port 80 and the HTTP headers of data flow are similar.

The convenience of these applications is obvious because of the easy to use features of the web; popularization of the web browser on all computers, an on the smartphone, as well as the ability to almost certainly is not blocked by the firewall. In contrast, for the firewall, this is a big issue because firewall faced to lots of difficulties in identifying which applications are active in the network that needs to be in protection. This is a large problem attracting the interest of many researchers. Some fundamental research directions as follows:

- The papers [2] [3] of Tomasz Bujlow, Tahir Riaz, Jens Myrup Pedersen mentioned on the classification of application that uses the C5.0 algorithm. In these papers there are also the classification method on the HTTP data stream. However, these studies only classified three types of applications so watching movies, listening to music, downloading files and accessing the normal web. Moreover, the classification of data stream mainly based on the content-type field in the HTTP header and on the type of workstation applications that access by the HTTP protocol. This is very difficult to apply to the firewall,

because the firewall does not know on which workstation application uses to access HTTP.

- The iptables command of Linux allows integrating the filtering module of application layer [5]. With this module, iptables can identify the applications if there is string signature in this application that is characteristic in the communication process.
- With existing methods, identifying some of applications active on the HTTP protocol without keywords, as single signature, remains a major challenge. The paper suggested approach is to analyze the combination of several characteristics of the application information exchange in order to extract rules that help to distinguish, identify applications. The analysis of the characteristics of the application session based on a developed tool. Extracting characteristic patterns and providing decision trees for classification is done by using the program See5/C5.0. The assessment of the ability to identify, error rate, demand for system resources such as memory, IDS latency is also considered in this paper.

The rest of paper was organized as follows. The section 2 describes method to solve problem. In the section 3, some experimental results were presented, and the section 4 concludes and describes some future work.

II. METHOD

A. HTTP Header and parameters for classifying web-based applications

The Hypertext Transfer Protocol (HTTP) is a protocol of application layer, it behinds web to exchange web pages in the network system [1], [8]. A HTTP transaction consists of two types of messages: request and response. Each message consists of three parts: request/status line, HTTP header fields and content. In this three parts, the HTTP header is metadata describing the properties of the data in the body of HTTP packets. This metadata is organized into the fields such as allow, content - encoding, content - length, content - location, with the value of the fields changed depending on the content of the data. The contents of the HTTP header in the form of pairs of attribute - value (also named AV) provide a lots of information to allow classification of applications. Using information of AV pair of HTTP header to identify different application activity over HTTP is the approach of this paper. All of the HTTP header field with selected values are used in the analysis and classification of applications. The different number of value using for the analysis is parameter to assess the ability to distinguish the application of IDS.

B. Building input data system for learning machine

After identifying the information from the fields of the HTTP header that help to distinguish the different activities on the network, another question is how to describe and distinguish the network services using the information of the above mentioned fields? Proposal of the paper is that, although there are many common services on the appearance of the fields, but the rate of occurrence of the field is different and that is the distinguished characteristic of the service. In order to calculate this ratio, first, when the interested service is working in network system, the system records all the information exchanged between the client and server using Wireshark sniffer tool. This information is written to the hard disk as file .pcap. This file contains the various network transactions, which contains the data stream of the services that system is interested in. Based on the information such as the IP address of the server providing services of video, download, game or others the sessions of each service are extracted, labeled corresponding to the service and switch to HTTP analyzer (named ISeHAN). The our software tool ISeHAN runs on the Java platform, using library JNetCap to analyze and calculate the frequency of the AV pairs, given the data vector whose coordinates are the rate of appearance of AV pairs in the HTTP header components on all or part of the session. The the result vectors of ISeHAN were labeled as input for learning machine software using the C5.0 algorithm (Fig. 1).

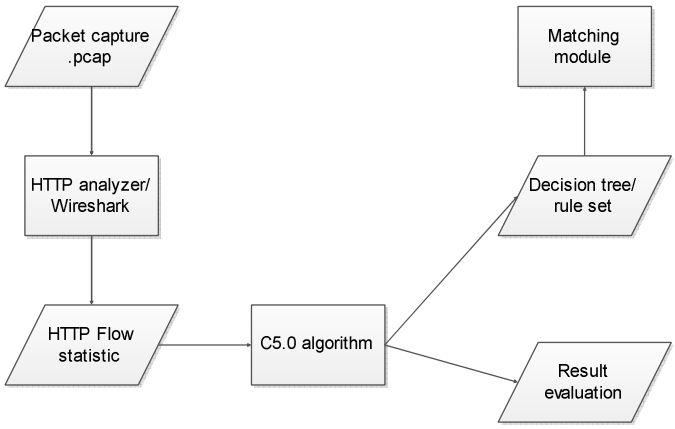

Fig. 1. Flow chart for building the decision tree to apply to the application identification module of IDS

C. Building the decision tree by C5.0 algorithm

C4.5 algorithm and its improved version C5.0 that was proposed by Ross Quinlan [9], [10] allows the construction of a decision tree to classify data based on a set of available classification data (called training data). Once there was a decision tree, the classification of new data is done very quickly based on this decision trees. These are the specific steps to consider on the coordinates of the vector to make the classification decision. After the learning data were completed with a list of classified vector; using the C5.0 algorithm on this data, we obtain not only a decision tree, but also a specific error rate. This error rate characteristics for the fuzzy - the phenomenon not able to classify exactly of the training data. This parameter would be adjusted through changes related to data collection.

D. Using decision tree for IDS application identification system

The decision tree would be "loaded" into the application identification module by IDS. This is the "knowledge" of the IDS for the application.

Then, in the real time, the IDS would record the transaction on the network, filter out the data stream operating under the HTTP protocol and construct the characteristic vector of transactions by the ISeHAN tool.

Applying of data tree for characteristic vector, IDS would make the decision on the class from which this application belongs to. The subsequent action (allow/deny) may depend on the specific policy of the firewall implementation. Thence IDS could just warn about the activity of the applications in the system, or IDS would transfer the information of the application to firewall; from that, the firewall takes action to prevent or allow the activity of application.

E. Quality of IDS Latency

During the period from the moment when the application begins activated until the moment when IDS recognizes it, there would be certain latency. This delay depends on the following factors:

- The number of packets of the application data stream that IDS is using for the construction of characteristic vector of the application. This elapsed time is relatively large and ever changing (not fixed); it depends on essence of the application as well as the network bandwidth.
- Speed of information extraction and construction the characteristic vector from the packets of the data stream.
- The timing for applying of decision tree to classify the application corresponding to the vector data.

Accuracy

The quantity of parameters involves as elements of learning data vector and classification data. Rule of thumb is that if more parameters taken for consideration, less wrong rate of classification; because we have the ability to know better the application from many more different sources. However, the IDS incorrectly identify network activities always happen with a certain rate. This error depends on:

- Quality of data and C5.0 algorithm. With a given learning data set, the generated decision tree also contains a rate of wrong identification with these data. Hence with the data in the future, in principle, it would have greater error.
- The quantity of packets of the application data stream that IDS uses for the construction of vector characteristics of the application. Quantitatively, if there were the more application packets taken, then the identification would be more accurate because we follow the data stream longer.

III. Experimental results

To realize the idea and to assess the ability to recognize the applications of the IDS system, we have recorded network activity when there are the activity of applications for Youtube access, retrieving data through Mediafire, online gaming and browing web by common ways. The objective of this study is to distinguish four types of web-based applications such as watching video, downloading files, playing games online and access to web.

There are 230 sessions over HTTP was recorded and analyzed by ISeHAN. We made a some different senariors when build learning data by changing the number of the first packets taken from session and the number of attribute-value parameters. They are the input values to build the learning data vector and data application to be classified). The identification results and error rates of IDS are summarized in Table 1.

Table 1. The error rate of application identification system in HTTP

	14 parameters	25 parameter
All the stream	1.5%	0.5%
50 packets	2.2%	1.3%
40 packets	3%	1.8%
30 packets	4.6%	0.5%
20 packets	5%	1.8%

Besides, this paper also evaluates the average amount of data that needs to record and the delay to get enough number of packets when the number of the first packet of HTTP session is changed (Table 2).

Table 2. The dependence on quantity of packets to analyze

	Quantity of data (KB)	Latency (s)
Full session	100	4
50 packets	22	0,55
40 packets	10	0,27
30 packets	7	0,22
20 packets	5	0,2

The increase in the number of parameters is generally very little to affect the time to system makes a decision, it only affects mainly to the process of building of training vector and coaching process C5.0 algorithm.

IV. Conclusion

Application Awareness for implementing coordinated policies, priorities, and different governance is a fundamental problem of network equipment system, in general, and information security device, in particular. This problem becomes more complex when there are many similar applications working together, like the context of this paper, where all applications would work on the HTTP protocol. The general approach is that we need to get more and more information from higher levels of the OSI model, following longer the information exchange session and apply smarter artificial intelligence tools, more deeply through which we could distinguish better different applications. The method uses information from the HTTP header with the C5.0

algorithm is a solution that allows identifying different applications at a certain level where the cost of time, storage and capacity of IDS/IPS are acceptable. If we accept the application will be blocked at a little later time, not immediately when the application is started or when the application has the first information exchanges, then solution mentioned in this paper is completely feasible and easily implemented.

ACKNOWLEDGMENT

This paper was funded by the project "Enterprise Non-Standard Firewall Development" of Vietnam National University - Ho Chi Minh Cit . The research team ISeLAB has many remarks to improve this document. We thank all above for these invaluable supports.

REFERENCES

[1] Leon Shklar, Richard Rosen. "Birth of the World Wide Web: HTTP". *Web Application Architecture: Principles, protocols and practices.* pp.32-68. John Wiley & Sons Ltd. 2003.

[2] Tomasz Bujlow, Tahir Riaz, Jens Myrup Pedersen. "A method for classification of network traffic based on C5.0 Machine Learning Algorithm", *2012 International Conference on Computing, Networking and Communications (ICNC)*, pp.237-241,30/01/2012 – 02/02/2012.

[3] Tomasz Bujlow, Tahir Riaz, Jens Myrup Pedersen. "Classification of HTTP traffic based on C5.0 Machine Learning Algorithm", *2012 IEEE Symposium on Computers and Communications (ISCC), 2012 IEEE Symposium on*, pp.882-887, 1- 4 July 2012.

[4] Bert Hubert. Linux Advanced Routing & Traffic Control HOWTO. 29/10/2003

[5] Lucian Gheorghe. "Layer 7 filtering", *Designing and Implementing Linux Firewalls and QoS using netfilter, iproute2, NAT, and L7-filter*. Pp.119-136, PACK publishing, 2006.

[6] Chris Sinclair, Lyn Pierce, Sara Matzner. "An application of machine learning to network intrusion detection", *Computer Security Applications Conference, 1999. (ACSAC '99) Proceedings 15th Annual*, pp.371-377, 1999.

[7] Sebastian Zander, Thuy Nguyen, Grenville Armitage. "Automated traffic classification and application identification using machine learning", *The IEEE Conference on Local Computer Networks,. 30th Anniversary,* Sydney, Australia, pp.250-257, 15-17 Nov. 2005

[8] David Gourley, Brian Totty, Marjorie Sayer, Anshu Aggarwal, Sailu Reddy. *Chapter 15. Entities and Encodings HTTP: The Definitive Guide*. Pp. 317-342. September 2002

[9] Quinlan JR. *C4.5: Programs for machine learning*. Morgan Kaufmann Publishers, San Mateo. 1993

[10] http://www.rulequest.com/see5-info.html, Data Mining Tools See5 and C5.0..