International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:1, No:12, 2007

Data Mining on the Router Logs for Statistical Application Classification

M. Rahmati, and S.M. Mirzababaei

***Abstract*—**With the advance of information technology in the new era the applications of Internet to access data resources has steadily increased and huge amount of data have become accessible in various forms. Obviously, the network providers and agencies, look after to prevent electronic attacks that may be harmful or may be related to terrorist applications. Thus, these have facilitated the authorities to under take a variety of methods to protect the special regions from harmful data. One of the most important approaches is to use firewall in the network facilities. The main objectives of firewalls are to stop the transfer of suspicious packets in several ways. However because of its blind packet stopping, high process power requirements and expensive prices some of the providers are reluctant to use the firewall. In this paper we proposed a method to find a discriminate function to distinguish between usual packets and harmful ones by the statistical processing on the network router logs. By discriminating these data, an administrator may take an approach action against the user. This method is very fast and can be used simply in adjacent with the Internet routers.

Keywords—Data Mining, Firewall, Optimization, Packet classification, Statistical Pattern Recognition.

I. Introduction

NOWADAYS for internet users there are many new usages of interacts but certainly the main application is accessing data. They easily can access a world of information. Each user can search for many subjects and also every user can present every bulk of information for all others. Beside its usefulness, there is always harmful information available for ill minded people. So some ISP's and authorities are concerned with the harmful or terrorist usages from their facilities. Many methods are used to fulfill this desire [1], for example some governments have passed laws that mandate the ISP's to set some usage limitation points in their agreements with users. Firewall is another method that is very important and effective way to protect the network. Personal firewalls and proxies are two major types of firewalls. They act as the trenches. They omit blindly all the packets containing doubtful contents. Processing the contents of all packets requires a very high performance computing power that needs expensive processors. Therefore, some ISP's process them in some expertistic domains. Our approach is to employ pattern recognition techniques where we have presented a discriminate function that rapidly decides on the router logs using statistical data and it alarms the users according to their applications. This method can be used simply in adjacent with the Internet routers due to its low processing utilization.

The rest of our paper is organized as follows: We will introduce the firewalls briefly in the next section then our method will be explained and in the forth section we will explain our experimental results.

M. Rahmati is with the Department of IT and Computer Engineering, Amirkabir University of Technology, Tehran, Iran, (e-mail: rahmati@ce.aut.ac.ir).

S.M. Mirzababaei, also is with the Department of IT and Computer Engineering, Amirkabir University of Technology, Tehran, Iran, (e-mail: mirzababaei@morva.net).

II. Firewalls

Network Computer Security Association (NCSA) believes that a firewall is a system or a complex of several systems that does some limitation between two or several networks. As a matter of fact a firewall tries to protect one inner network from another one by limiting the accesses between them. The firewall as represented in the Figure 1 processes the packets and recognizes the unacceptable transfers according to its predefined security policy.

Of course the firewalls are not the total security solutions. They have some drawbacks, and some insecurities or intrusions are out of firewall abilities, so the administrators have to use some physical security issues or host dignities, and so on.

Firewall technology has many variations in the past twenty years. The first generation of firewalls introduced in 1985 [6]. Some conceptual ideas (Screening Process) formed from Cisco routers that were known as Internetworking Operating System (IOS). In 1989 the AT&T laboratories introduced circuit level firewalls and the first working model of the application level firewalls in the same year. Many researches developed in 1991 on proxies as the third generation of firewalls. The fourth generation of firewalls came in the late of 1991. Their concepts are based on the dynamic packet filtering. In 1996 the fifth generation was implemented in the kernel of the operating system that called kernel proxies [6].

The firewalls have some drawbacks and challenges in the performance, some compatibility issues and conflicts with some network protocols. They also omit blindly some packets that only resembles to harmful information. In some expertist domains, administrators do not want to limit their specialists while they do not want to let invalid access either. In this paper we introduced a mechanism that processes the router logs in the background and we allow the user to access all the Internet resources but when this mechanism finds several harmful accesses, it alarms the user. Our proposed design is inexpensive because it processes the Uniform Resource

Locators (URL) [10] in router logs instead of the whole packets. There are other works similar to our system, but they performed on the web logs; for example refer to [2,3,4].

We selected 3 different measurements in the time granularity. After finding the value of those three measurements we obtain an alarm criteria that illustrates the users' usages. It worth mentioning that there are some similarities between ordinary and harmful texts that this method can consider it, because of this method it uses uncertain decision-making and does not omit the packets blindly as explain in the next section.

III. Proposed Method

We propose a mechanism that based on a passive firewall in contrast to an active one as represented in Figure 2. The passive firewalls process the contents and do not make the decision about packet delivery so the router can transfer the packets as soon as possible and the firewall's process power will not be a bottleneck in the packet delivery.

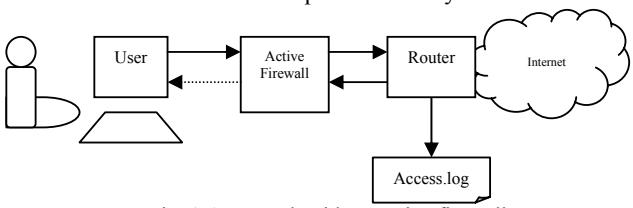

Fig. 1 A network with an active firewall

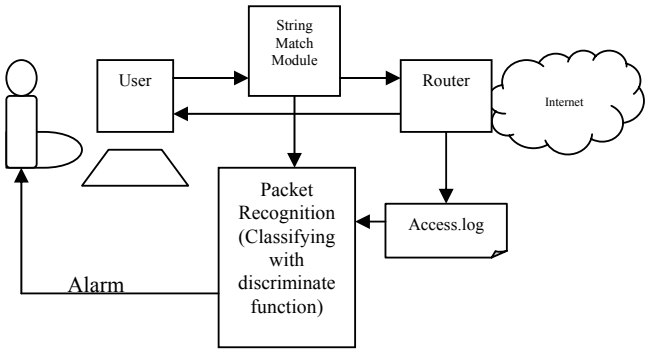

Fig. 2 A network with our proposed passive firewall

Our firewall processes the router log files. These log files contain some information about every transferred request. When users need a file, their computer informs the router about the file name and its address and then the router acquires it from the network. Figure 3 presents some of such request information in the router log files. The router log files have 11 fields as explained [9]:

1- System time in standard UNIX format since 1970.

2- Duration in milliseconds the transaction required.

3- Client address of the requesting browser.

4- Two entries separated by a slash. The former shows the result code about how the request was resolved or wasn't resolved if there was a problem. The latter shows the status code, which comes from a subset of the standard HTTP status codes.

5- The size of the data delivered to the client.

6- Request method of the HTTP, to obtain an object.

7- The requested Uniform Resource Locator [10].

9- RFC931 is the ident lookup information for the requesting client, if they are enabled in your router.

10- Hierarchy code with three items. A) a prefix of TIMEOUT_ if all ICP requests timeout. B) the code that explains how the request was handled. C) the name or IP of the host from which the object was retrieved.

11- The MIME type of object that was requested.

The Figure 3 shows that these log files does not distinguish between ordinary and harmful files so we must analyze the log contents to obtain effective data. We worked on six measurements for the analysis of the router log files that are represented in the Table 1. There is also another seventh measurement that is obtained from a string match program such as a module in the snort firewall that its importance is obvious and so does not shown in the Table 1.

1- The time measured in seconds.

2- The number of requests.

3,4- The size of requests: This measurement counted in two ways. A) String size of all the characters of the request. The technical properties of the software of the server has mentioned in the URL. B) The number of white spaces in the request. By this way we can find the levels and the depth of the server software.

5- Maximum number of the requests from one server.

6- Using the TCP protocol in the communication: The TCP protocol is a second order protocol and often brings some complexities and non-linearity. So always the TCP protocol is using for data transfer and does not use for multimedia transfers.

7- Number of special or harmful words that recognized by a string-matching program. Such as “bomb”, “terror”, “missile”, “porn” or “sex”.

```
•1044797404.687  16793 81.12.16.216 TCP_MISS/200 16266 GET http://www.digitalvisiononline.com/images/cdcases/515.jpg  -
DIRECT/62.189.247.137 image/jpeg
•1044797404.687  27943 81.12.16.249 TCP_MISS/200 38991 GET http://www.riyadh.ws/images/chatbanner.gif - DIRECT/66.220.30.18 image/gi
•1044797404.741      0 213.29.54.6 UDP_MISS/000 64 ICP_QUERY http://sa.windows.com/satasks/Engine271.xml - NONE/- -
•1044797404.741      0 213.29.54.5 UDP_MISS/000 46 ICP_QUERY http://www.yahoo.com/r/m1 - NONE/- -
```

Fig. 3 Samples of the router log records

The dependency of each pair of the measurements is evaluated and listed in Table 1. It can be seen the context of combining each pair that is earned in each element by dividing its column by its row. Some of them are suitable for recognition of other applications. For example when a user searches in the net the large amount of transfer in the shallow level happens. We studied the behaviors of the elements in both hazardous class of usages and ordinary ones. In three combinations those classes of usages was isolated from each other. So we chose the three combinations of measurements

for our application and implemented their program in the proposed firewall.

1- Continuous usage of a server: the number of the used servers during a predefined period.
2- The number of doubtful words.
3- Structured depth of the server: the count of URL parts (the address, folders...).

We formed the above measurements as a vector of three random variables and classified the users by Ward clustering algorithm in two classes, Normal and those that are abnormal. So the harmful usages will agglomerate in one class and the ordinary usages in the other cluster.

TABLE I
CONTEXTS OF COMBINING PAIRS OF MEASUREMENT

	Time	No of records	Records length	No of parts of the records	Max request to a server	TCP Connection
Time	-	More traffic	Server has more depth	More depth in the software of the server	More focus on a server	Complexity of the software in the server
No of records	Record rate	-	Server depth in the use	Software depth in the use	More interest in the use	More congestion
Records length	Server has more depth	Server depth in the use	-	Server depth in the software depth	More interest in the software depth	More downloadable data
No of parts of the records	More depth in the software of the server	Software depth in the use	Server depth in the software depth	-	More interest in the server depth	More downloadable data in the server depth
Max request to a server	More focus on a server	More interest in the use	More interest in the software depth	More interest in the server depth	-	Favorite file server
TCP Connection	Complexity of the software in the server	More congestion	More downloadable data in the server depth	More downloadable data in the server depth	Favorite file server	-

A. Design of the Discrimination Function

We can cluster the input variables as normal random distributed variables by finding their average and variances to get the discriminate function via the formula (1).

$$h(\underline{x}) = \underline{v}^T \underline{x} + \underline{v_0} \quad (1)$$

That wants v from formula (2) and s by (3,4)

$$v = \left[s\Sigma_1 + (1-s)\Sigma_2\right]^{-1}(m_2 - m_1) \quad (2)$$

$$s = \frac{\frac{\partial f}{\partial \sigma_1^{\,2}}}{\frac{\partial f}{\partial \sigma_1^{\,2}} + \frac{\partial f}{\partial \sigma_2^{\,2}}} \quad (3) \qquad \text{And} \qquad f = \frac{P_1\eta_1^{\,2} + P_2\eta_2^{\,2}}{P_1\sigma_1^{\,2} + P_2\sigma_2^{\,2}} \quad (4)$$

But Fischer [5] suggests to consider S equal to ½ and find η from the formula (5) $\eta = \mathrm{E}\{H(\underline{x})|w_i\}$ and v_0 from the formula (6) $v_0 = -v^T\left[P_1 m_1 + P_2 m_2\right]$. For more information about this method refer to [5,8].

B. Solution with Unknown Distribution Model

In the case which the random variables do not distributed normally and we want to find the discriminate function by the Bayesian relation, we need to estimate the probability density function. In this way Parzen [5] helps us with the formula (7)

$$\hat{\mathrm{P}}(\underline{x}) = \frac{1}{v^n}\left(\frac{1}{N}\sum_1^N \Phi\left(\frac{x_i - x}{v}\right)\right) \quad (7)$$

In this formula v has to be adjusted in a way that the total probability equals to one. In the above formula every sample has a rectangle window that its width equal to one and its height is equal to $1/v$. We can improve the estimation by changing the estimating window with an isosceles triangle. The central vertex of the triangle is above the sample and its edges are drawn by 0.5 to the left and right. In this research the random vectors are 3 dimensional, because we chose 3 combinations of measurements for the classification .The 3 dimensional vectors turn the isosceles triangle into conical so the closed formula will be very sophisticated. We represented this formula in a program function, and then we clustered both the harmful samples and ordinary samples in two classes. The priori probability of the classes is found by some simple

statistic calculations. So this leads us to determine the discriminate function by (8)

$$\varepsilon = \int_{R_2} \mathrm{P}_1 P_1(x)dx + \int_{R_1} \mathrm{P}_2 P_2(x)dx \quad (8)$$

In this stage we have to derive this formula. So we calculate the $\frac{\partial \varepsilon}{\partial x}$ and then we force its derivative to be equal with zero to find minimum error. But the above function has not a closed formula so we have to find optimum point by fibonacci algorithm [7] that does not need the derivative in the optimum point. The fibonacci algorithm is a numerical calculation method, which can be coded in a program to find the optimum value of a curve or function without need to the derivative of the function. The Figure 5 shows this function that has to be optimized. The earned optimum vector introduces a discriminate function by its triple values.

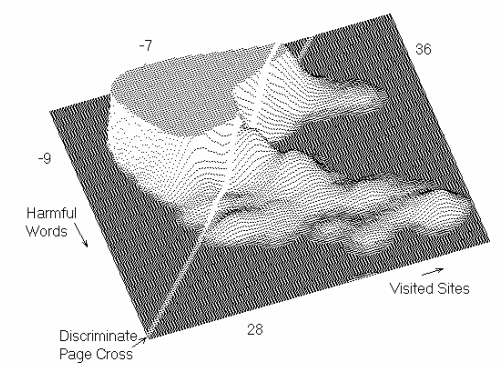

Fig. 4 The cross of the discriminate function (a page) with the probability density function

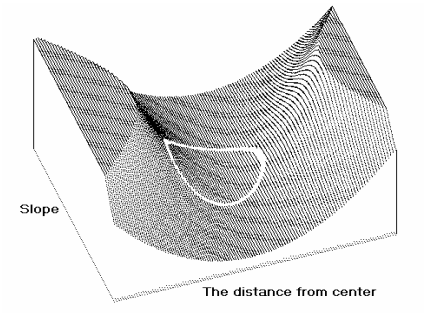

Fig. 5 The optimum saddle point that its coordination is equal to the parameters of the discriminate function.

Now the discriminate function can classify the new samples by putting their vector in the discriminate function. This method has shown in the Figure 4. The Figure 4 shows the probability density function of 500,000 samples that used by the Parzen formula (7) for estimation. The estimating window changed to isosceles triangles in the Parzen formula and turned to conical because of 3 dimensions. This point has to be mentioned that we omitted the URL parts count dimension to obtain a viewable three-dimensional perspective (the least valuable measurement).

IV. Conclusions

As explained, we found an optimized discriminate function with our proposed method. We can use this discriminate function as follows. While a router reports in the log file, our passive firewall measures the router reports in a certain time period and samples the measurements to form the random vector variables. Then it puts the random vector variable in the discriminate function to find the class of the usage. If the usage was harmful, it alarms the user by an email and notifies the administrator.

Our method decides discerning, in uncertain conditions and even can be used with ordinary method of firewalls that act blindly. Our proposed method has three important benefits against the ordinary method by using the active firewalls. The first is that it decides uncertainly and does not act with a single ambiguous packet. Its second advantage is that the firewall processes quickly. If we consider the average volume of the requested files equal to n (ordinary average is equal to 50 Kbytes) and consider also that the user requested m times, then the order of the search over the whole packets will be equal to O(n*m). But according to the mentioned matters we searched in the users sent requests with an extra offline or online stage of learning with the discriminate function in the proposed algorithm. This method is from O(k) order that k is a constant and is equal to the average size of the recorded requests (the average value of the k is about 100 bytes). So the processing speed is not a bottleneck in the network traffic. The third benefit is its low price because of its light-processing algorithm so we can use this algorithm in any network.

References

[1] P. Gupta, N. McKeown, "Algorithms for Packet Classification", IEEE Networks, Mar/Apr 2001.
[2] A. Benczur, K.Csalogany, A.Lukacs, B. Racz, C.Sidlo, M.Uher, L.Vegh, "An Architecture for Mining Massive Web Logs with Experiments", Project Report Data Riddle & OTKA & AKP ,2003.
[3] Q. Yang, H. Wang, W. Zhang, "Web-log Mining for Quantitative Temporal-Event Prediction", IEEE Computational Intelligence Bulletin, 2002.
[4] Z. Su, Q. Yang, H. Zhang, X. Xu, Y. Hu, "Correlation-based Document Clustering using Web Logs", Microsoft Research China Report, 1999-2000.
[5] K. Fukunaga, "Statistical Pattern Recognition", Academic Press Inc.
[6] W. Stallings, "Data and Computer Communications", Prentice Hall.
[7] E. Chong, S. Zak, "An Introduction to Optimization", John Wiley & Sons Inc.
[8] E. Khorram, S.M. Mirzababaei, "Finding an Optimized Discriminate Function ", Proceeding of ALDM'05, 2005.
[9] J. Cooper, "The Book of Webmin",avilable at: http://www.swelltech.com/ support/ webminguide/ ch12.html, Amazon.ca,2003.
[10] IETF Standard Track Category: IETF Uniform Resource Locators (URL) Specification: RFC 1738.