International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:1, No:4, 2007

Finding an Optimized Discriminate Function for Internet Application Recognition

E. Khorram, S.M. Mirzababaei

***Abstract*—** Everyday the usages of the Internet increase and simply a world of the data become accessible. Network providers do not want to let the provided services to be used in harmful or terrorist affairs, so they used a variety of methods to protect the special regions from the harmful data. One of the most important methods is supposed to be the firewall. Firewall stops the transfer of such packets through several ways, but in some cases they do not use firewall because of its blind packet stopping, high process power needed and expensive prices. Here we have 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. So an administrator can alarm to 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

Everyday there are many new uses with their computers that join to the Internet. They easily can access a world of information. Each user can search for every subject and also every user can present every bulk of information for all others. There is some available harmful information for all. So some ISP's and governments do not allow the harmful or terrorist usages from their prepared facilities. Many methods used for this desire [1], for example some governments have some laws that forces the ISP's to mention 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. So they omit blindly all the packets containing doubtful contents. Processing of all the contents of packets needs a very high performance computing power that needs expensive processors. So some ISP's pass up them in some expertistic domains. We have presented a discriminate function that rapidly decides on the router logs using statistical data and alarm 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 paper organized as follows: We will introduce the firewalls shortly in the next section then our method will be explained according the router logs in the third section and the fourth and fifth sections belong to the method and in the sixth section we will explain the results.

Manuscript received January 10, 2005.

E. Khorram, is with the Department of Mathematical and Computer Science, Amirkabir University of Technology, Tehran, Iran, (e-mail: eskhor@aut.ac.ir).

S.M. Mirzababaei, 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 form another one by limiting the accesses between them. The firewalls have three major tasks:

-They limit the users' access to a predefined controlled point.

-They protect the inner resources from hackers.

-They limit the users' retrieved information to a predefined controlled point.

These predefined controlled points are the same as the movable bridges that used as a gate in the ancient castles. The firewall has to be installed in the access point between the inner and outer network and the entire traffic to/from inner network will be done through the firewall. The firewall processes the packets and recognizes the unacceptable transfers according to its predefined security policy. We can define a firewall as a restrictor, splitter and analyzer. A firewall is a router with several access control lists or a software program on a PC or a special hardware box. Also there are more complicated firewalls that formed by several systems or multi-computer solutions or several routers. Some of the major benefits of firewalls are:

- A firewall concentrates security decisions in a point. As a matter of fact we can make several decisions about the network security and concentrate them in a special controlled point

- A firewall executes the network security policies. These different network security policies define their related presented services in the network. These policies define who can use which services. Finally the firewalls will execute many of these security policies in the form of executable rules. There are a few insecure and dangerous services that will be allowed with some limitations only in the inner network. This security policy differs for individual computers in the inner and the outer network.

- A firewall records the operations in some log files. The firewall is the best point for the recording of the transfer events between inner and outer network, because it acts as a gate. We can obtain good results by processing of these logs about network usages, Inner or outer intruders, hackers or some developing attacks.

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 is very young and fresh but it developed quickly and has many variations in the past twenty years. The first generation of firewalls introduced in 1985. 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 done in 1991 on proxies as the third generation of firewalls and the proxies came to market very quickly. The fourth generation of firewalls came in the late 1991 months. Their concept 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].

Now we need some quick, extensible, maintainable and flexible security systems. So many companies research on proper solutions to answer the user requirements. The firewalls have some drawbacks and challenge in the following areas. They need a high performance computing power that leads to expensive hardware costs. They have also some incompatibilities and conflicts with some network protocols. They omit blindly some packets that only resembles to harmful information. In some expertist domains, administrators do not want to limit their specialists whereas they do not want to let invalid access. In this paper we introduced a mechanism that processes the router logs in the background so we allow the user to access all the Internet resources but when our mechanism finds several harmful accesses, it alarms the user. Our design is also inexpensive because it processes the Uniform Resource Locators (URL) in router logs in substitution with the whole packets. As a matter of fact every requested file has an average size about 100KB but it's URL as a request has only about fifty bytes long. There are other works similar to our but they done on the web logs for example [2,3,4].

We selected three different, measurements in the time granularity. After the process of those three measurements we obtain an alarm criteria that illustrates the users' usages. Finally it has to be mentioned that there are some similarities between ordinary and harmful texts that this method can consider it, because of this method uses uncertain decision-making and does not omit the packets blindly.

III. Proposed Method

We propose to use a passive firewall in spite of an active one. 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.

Figure 1: A network with an active firewall

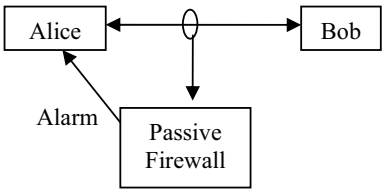

Figure 2: A network with a passive firewall

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. This log file has a record about fifty bytes for every file that has about 100K bytes. The low process volume leads to lower costs for the hardware of the firewall. The figure 3 presents some of such request information in the router log files.

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 found some measurements for the analysis of the router log files [8].

1- URL length: we obtained this measurement in two ways:

A- The length of the URL characters in bytes that contains the technical features of the server and its folders in the network.

B- Count of white-space characters in the URL. This measurement shows the levels and the depth of the server folders and also its software structure.

2- The maximum number of request to an individual server in the network. This measurement counted in a pre-specified duration.

These measurements find the scatter of the users usage. The measurements used in the proposed mechanism with the following concepts:

1- Professional use from a site: The counted number of the met sites during a pre-specified time period for a specified user.

2- Harmful counter: Number of harmful items recognized by a string matching routine during a pre-specified time period for a specified user.

3- Software depth: This is the depth of the structure of the server levels and folders. This measurement obtained by counting a request parts (address, folders, other request parts) during a pre-specified time period for a specified user.

•1044797404.687 5092 81.12.17.241 TCP_MISS/200 0 GET http://adserver.yahoo.com/a? - DIRECT/216.136.232.177 application/x-javascri

•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 29542 81.12.16.237 TCP_MISS/200 17962 GET http://www.gomaldives.com/Images/Resorts/22/AboutResort3.jpg - DIRECT/207.195.39.252 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 imag

•1044797404.687 23354 81.12.16.249 TCP_MISS/200 19774 GET http://www.ashge.net/qs2a5e.JPG - DIRECT/207.44.194.51 image/jpeg

•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/- -

•1044797404.741 0 213.29.54.5 UDP_MISS/000 151 ICP_QUERY http://rd.yahoo.com/M=224039.1984929.3466713.1922510/D=yahoo/P=m1r1f9qb110o0800/S=2766679:HEAD/A=1030697/R=0/*http://www.yahoo.com - NONE/- -

•1044797404.741 0 213.29.54.6 UDP_MISS/000 58 ICP_QUERY http://128.242.100.84/images/top4.gif - NONE/- -

•1044797404.741 0 213.29.54.6 UDP_MISS/000 149 ICP_QUERY http://srd.yahoo.com/hpt1/ni=18/ct=modem/sss=1044797387/t1=1044797483748/d1=5031/d2=5984/d3=6328/d4=9375/0.027002808806568 4 - NONE/- -

•1044797404.741 0 213.29.54.6 UDP_MISS/000 94 ICP_QUERY http://www.dotukdirectory.co.uk/search_boxes/dotyoursite_selected_tab NONE/- -

•1044797404.741 0 213.29.54.6 UDP_MISS/000 93 ICP_QUERY http://www.dotukdirectory.co.uk/images/dotukdirectory_unselected_tab.g NONE/- -

•1044797404.741 0 213.29.54.5 UDP_MISS/000 82 ICP_QUERY http://x.ddfprod.net/1byday/galleries/014/thumbnails/tn07.jpe - NONE/- -

Figure 3: Samples of the router log records

Now if we draw the coordinates of these measurements, we find that the proper usages and harmful usages agglomerate in two separate colonies. This fact has showed in the figure 4. Now we have to find a discriminate line between two parts (the proper usage in one side and the harmful usages in the other side) in the run time. We can put the measured properties of the URLs during a period in the line formulae and decide about the usage by the sign of the result. There are several methods to find this discriminate line. Often the Bayesian method has been used to find the discriminate function. In the Bayesian method the probability density function of the samples often premises to be Gaussian and then we try to minimize the error value. But as represented in figure 4 the samples in this case did not distribute normally. So we used two other methods. At the former we estimated the probability density function to be used in the Bayes method and in the latter we found the discriminate function by the Euclidean distances.

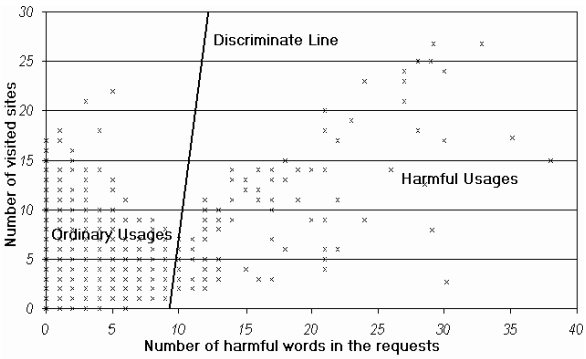

Figure 4: The distribution of usages with the discriminate function found by Euclidean distance

IV. Estimating The Probability Density Function

We have to estimate the probability distribution function in order to be used to obtain the discriminate function from the Bayes rule. The Parzen method [5] can help us to estimate the probability distribution function with the following formula:

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

The N is the measured samples, the x is the measurement, and the v in this formula will be determined in such a way to keep the entire probability equal to one. As showed in the above formulae we have a rectangle over every sample. The width of the rectangle is equal to one and its height is equal to 1/v. Isosceles triangles (the Φ) that their equal angles drawn by 0.5 to left and right and their height is equal to 2/v can give a better estimation in spite of rectangles. The formula of this probability density function cannot be written in the closed form so it needs a program function to calculate the results. Now we can design a discriminate line by the Bayes formula for minimizing the error value according its equation.

$$\varepsilon = \int_{R_2} \mathrm{P}_1 P_1(x)dx + \int_{R_1} \mathrm{P}_2 P_2(x)dx \qquad (2)$$

$$\text{And} \qquad \frac{\partial \varepsilon}{\partial x} = 0$$

In this formula, P_1 is the probability of the harmful usages and P_2 is the probability of the ordinary usages. R_1 is the domain of harmful usages and the R_2 is the domain of ordinary usages. $P_1(x)$ and $P_2(x)$ are the priori probability functions. The error value ε is a concave function and it has a minimum point. Since the probability density function is a program function, its derivative has not closed formula too. It is better to use an optimization (error minimization) method that does not need the derivatives. In this research we used the fibonacci method. In this stage we find an R with three properties that illustrates average of R_1 and R_2. This point is the extermum saddle point. Now we have to find the passenger line from this point that its sheer vector direction is toward the gradient of the probability distribution function. We work in 3D so we need the average of the gradient of probability distribution functions.

V. Discrimination With Euclidean Distance

In this method we obtain the minimum distance of the discriminate line to all the samples by finding the sheer line from every sample to the mentioned line and find the maximum of its value. Figure 5 shows that the answer can be in the infinity but the only acceptable answer is a relative maximum in a saddle point that has showed in figure 5. The figure 6 is the same as figure 5 but it has a better precision in the saddle point that is represented by the circle in the figure.

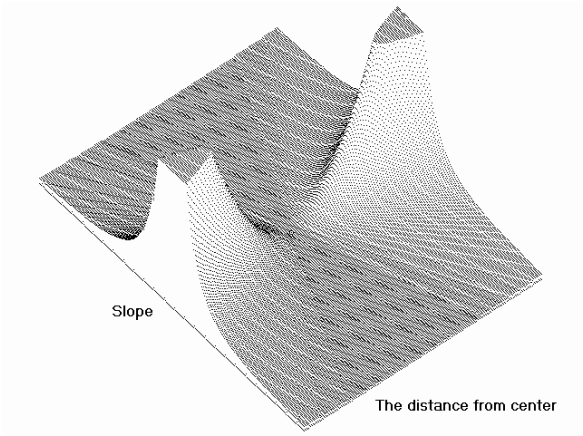

Figure 5: The space of discriminate lines and their related distance values

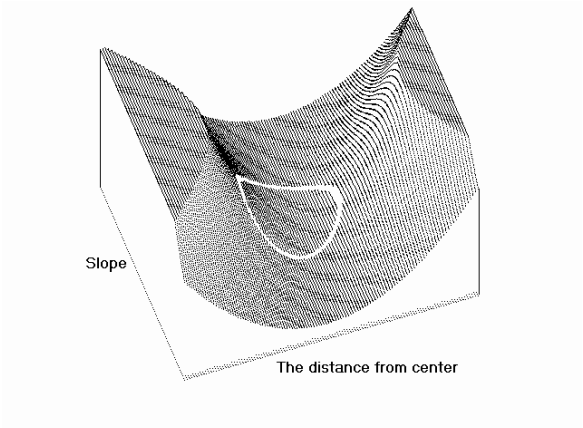

Figure 6: The optimized saddle region with its limitation for the convergence

Solving the problem with its limitation needs Kanush-Kuhn-Tucker (KKT) method [7] with inequality situations that helps finding the answer.

VI. Conclusions

According to the mentioned matters, a discriminate function that found in a learning stage can search in the user's sent requests to find harmful requests in the Internet. This method has O(k), that k is a constant and is equal to the average size of the requests. The k is about fifty. The usual method of firewalls need processing power of O(n) that is equal to user requested information volume.

References

[1] P. Gupta, N. McKeown, "Algorithms for Packet Classification", IEEE Networks, 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] M. Rahmati, S.M. Mirzababaei, "Router Logs Data Mining", Project Report, 2004.