INTRUSION DETECTION SYSTEM BASED ON WEB USAGE MINING

Mahza Mabzool[1] and Mina Zolfy Lighvan[2]

[1]Department of Electrical and Computer Engineering, Tabriz University, Tabriz, Iran
[2] Department of Electrical and Computer Engineering, Tabriz University, Tabriz, Iran

ABSTRACT

This artical present a system developed to find cyber threats automatically based on web usage mining methods in application layer. This system is an off-line intrusion detection system which includes different part to detect attacks and as a result helps find different kinds of attacks with different dispersals. In this study web server access logs used as the input data and after pre-processing, scanners and all identified attacks will be detected. As the next step, vectors feature from web access logs and parameters sent by HTTP will derived by three different means and at the end by employment of two clustering algorithms based on K-Means, anomaly behaviour of data are detached. Tentative results derived from this system represent that used methods are more applicable than similar systems because this system covers different kinds of attacks and mostly increase the accuracy and decrease false alarms.

KEYWORDS

Anomaly detection, Intrusion detection system, log file, n-gram, K-means clustering, web mining

1. INTRODUCTION

Nowadays internet plays a vital role in communications and data transfer. Besides, web applications are widely used in this regard. Attacks against applications has been prevalent these days because of important information on them. So IDSs are much noticed by majority of researchers in recent decades. These traditionally implemented IDS system was based on signature-base detection in which add in new attacks is a must. The main disadvantage of these systems is that they are not able to detect new attacks while the new attack is not added to the signature of data-base yet. This was the cause of extension of IDSs based on anomaly detection [1].

IDS system proposed in this article takes advantages of combination of both methods. In signature-base detection, by considering patterns of few attacks, we detect them and in anomaly detection sector we derive the anomaly behavior by means of feature vectors and clustering methods. By this method we will be able to detect almost all unknown attacks. However, as all the anomaly traffic are not threats. We would have some false alarms.

The rest of the paper is structured as follows. In Section 2 we describe related works with a special emphasis on the use of data mining methods for intrusion detection. Details of this system structure and components described in section 3 then in section 4 we describe the examination results and finally section 5 contain a short conclusion and some suggestions.

DOI : 10.5121/ijcsea.2014.4101

2. RELATED WORKS

The first anomaly detection system based on HTTP protocol was represented on 2002[2]. This method detects malicious requests by evaluating three characters: request type, request length and payload distribution. Another system represented in this regard is converting malicious requests to signature by using anomaly extension [3]. Another system with the name of MINDS extended after [4]. MINDS was detecting different kinds of intrusion in network layers with different elements such as scan detector, anomaly detector and summarization components.

The other system represented is using hidden MARKOV pattern for detecting anomaly [5]. This system uses web server access logs in education phase and in order to detecting attacks, it uses components which formed for web applications and the system is able to detect attacks which happened to specific set of applications.

Each of these systems has their own advantage and disadvantages. One of the common problems in similar projects is the disability in covering different kinds of attacks. The other problem is immensity of false alarms. Also, implementation of these systems is based on a theory which says web access logs relevant to attack are dispensable in comparison with normal web access logs and it is ignored in that situation which the website has attacked in recent days continuously.

Another problem of such systems is adapting them with different systems. Our proposed system is Adaptable with other systems and tries to cover different kinds of attacks with various plenty in web access logs. Besides, this system reduces number of false alarms significantly with help of preliminary knowledge of the administration.

3. Proposed solution

Figure 1 shows the proposed system's operational profile. As can be seen in the figure, the system is composed of different steps; opration of these steps will be described.

- **Pre-processing**

At first the raw data is entered this stage. Pre-processing step is composed of several sub-steps and will be explained in more detail.

- **Preparation**

The first step of pre-processing is data preparation in which all parts of each record are separated by a comma.

- **Clean Up**

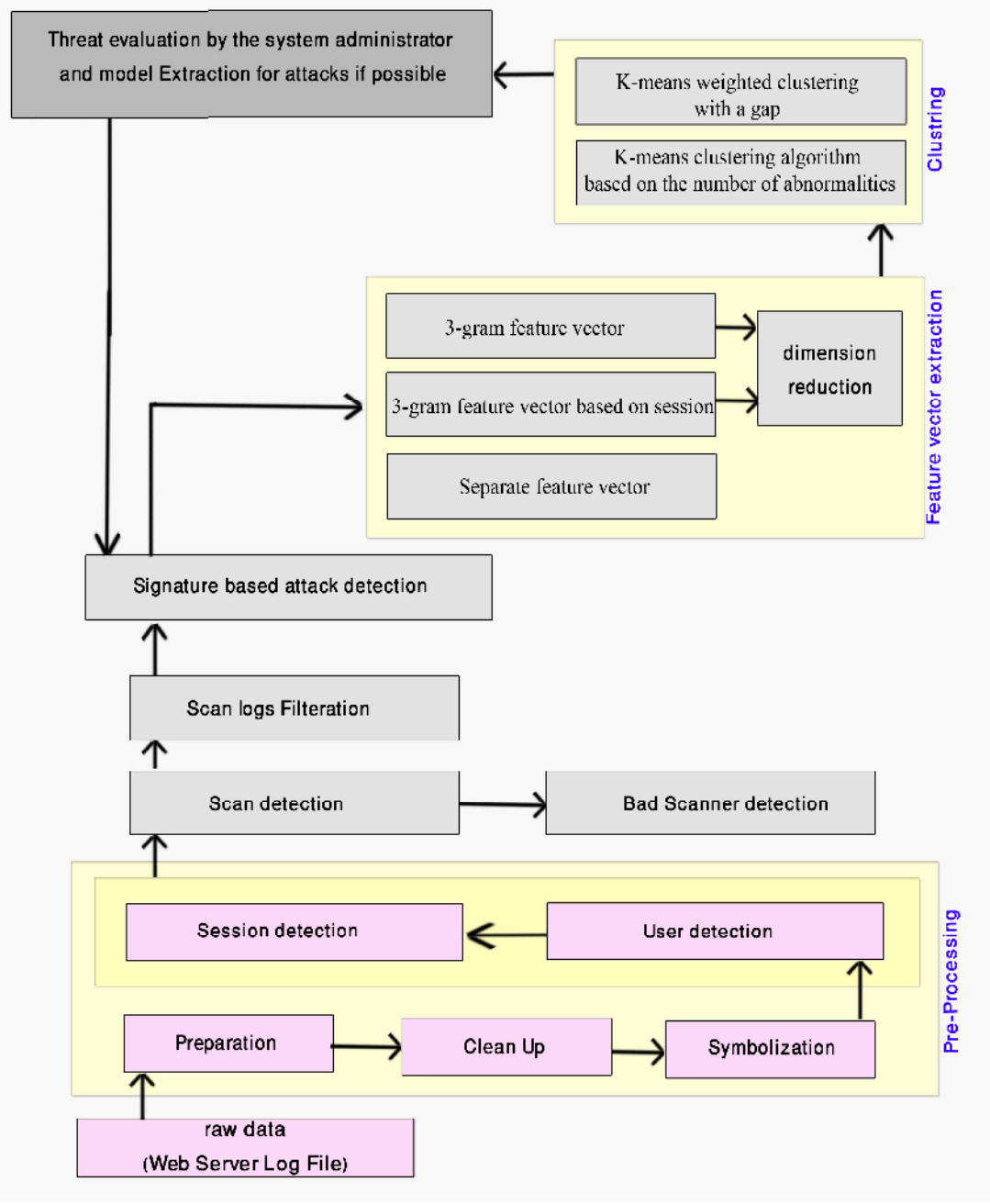
In the Clean Up, the information not required at a later stage will be deleted; they include the time difference with GMT, the size of the requested object and the user identity.

- **Symbolization**

To perform calculations faster, some long strings of information are symbolized.

- **User identification**

At this point, we are going to identify users. Records that have the same IP and user agent are recognized as a unique user.



1Figure. The proposed IDS systemoperational profile

- **Identifying Session**

We have used heuristic method based on the time to identify the sessions. To do this, we consider a threshold time of 30 minutes, this means that if we do not have a user request during any 30 minute period, then another request from this user will be considered as a new session [7].

- **Identifying Scanners**

Scanners are considered a threat to the website because the attackers usually scan the websites to find its vulnerabilities. For separating the normal people from scanners, we use the nine characteristics: lots of clicks, HTML high rate to image, high percentage of requesting PDF / PS files, a high percentage of requests with 4xx error, HTTP requests of HEAD, a high percentage of requests with unknown source, the 'Robots.txt' file request, the standard deviation of the depth of the sent requests and high percentage of HTTP consecutive requests. Some scanners are abuse and some are well-behaved. To detect abuse scanners, we use the user agent and attributes requested object [8].

- **Signature-based detection**

In this step, we focus on signature-based attacks detection. Therefore, our system should include a pattern of the known attacks. We modeled the four attack (XSS, CSRF, SQLi and BSQLi) signatures and added them to the system.

- **Feature extraction**

Feature vectors extracted should be able to separate attacks records from normal records and help separate types of attacks .In this study, we extract the feature vectors by three methods. One of these methods named 3-gram is based on previous research in intrusion detection systems [6] and the others are new approaches that have been suggested in this study.

In 3 gram method, at first step we extract all the uploaded parameters: then we examine the number of occurrences of each triple substring for all parameters .The second feature vector proposed in this paper is like the previous vector with this difference that each vector is specific to each unique session and contains of all ternary substring sent by a unique session . The third proposed feature vector which we have named it ***Separate*** **feature vector** is based on a Session that has seven properties. Each of these features is useful for detecting certain types of threats and anomalies, these features include the character described below :

• The first characteristic is the average length of the sent addresses in a single Session. This characteristic is proper to detect attacks such as code injection and SQL injection that usually send large number of long queries through the parameters. In addition it is proper for attacks such as buffer overflow, which may send a very long string. The problem to detect attacks with this feature this is that the pages of the Site that send parameters, may have very long URL than other pages .Thus, if a user visits pages with long parameters repeatedly; number of this feature would be more than the normal state and may be supposed that user is a threat mistakenly.

• Another characteristic is the mean length of the 10 longest user requests in a unique session. This characteristic in terms of helping identify threats, acts like the above character, except that it has the limited detection range, but its accuracy is higher.
For example, a user who has a long session with visiting much than 10 pages with long URL may consider as a threat, while at the features above because of the being long

session and averaging over all pages requested through the session, the average number would be like to the normal state.

• The number of requests that their reference is null or "-"or external websites. This characteristic is important in this regard which can help in identifying sessions involving robots and hacking tools that often have reference "-", Also it is proper to distinguish cross site request forgery attacks, such as submitting a request from the website to the victim site.

• Request that their response code is 4xx. This response code occurs when an invalid request is sent to the server, such as, request of accessing a particular folder or sending a URL that does not exist. Thus for detection of threats attempts to gain unauthorized access to confidential files or for testing web site penetration for injection attacks through special characters is suitable.

• The next item is the average time between user requests. This characteristic in a normal meeting is usually more than abnormal meeting; for example in some attacks such as denial of services many request will be sent during one minute that a normal user cannot send these requests in one minute.

• The number of sent request from a unique session that the number of its requests by other sessions divided by the total number of requests is less than 2%. This characteristic is useful for detecting threats done by sending requests not to be among the normal website requests and rarely going to happen. For example, injection attacks and unauthorized access to files are one of these cases. A problem that can occur in this case is that it is possible the web site has a page that is rarely observed and rate of this page request be lower than other pages rate ;so this makes that visit such a page beconsider as anomaly behavior.

Also if a web site has member and member names send by URL parameters, according to the username is always unique for each user, so the URL that has the username is unique for each session, so sending such unique request to the Web server is considered an anomaly.

• The number of sent request from a unique session that the number of its repetition in that session divided by the total number of session requests is more than 2%. This characteristic is useful for detecting efforts to unauthorized access to website files. This is also suitable for attacks such as denial of service.

A problem that can occur to anomaly detection based on this attribute is that sometimes due to user's high interest to the specific information ؛ user request some Website page frequently and numerous of requests from these pages send to Web server.

After extracting feature vectors, we use dimension reduction in order to reduce the system operational costs.

- **Clustering**

The first method used for clustering is based on one of the approaches is K-means (K-means--); which the number of alerts (L) are determined by the system administrator and is given to the system as entries. In all iteration of the algorithm, clusters' new centers - without taking the L numbers of the farthest data points are calculated and finally, L numbers of records with the farthest distance than other records are determined as the most abnormal points.

The second proposed method for clustering is based on K-means where you may wish to give more weight to one dimension. In addition to the desired weight of a feature vector, the largest gap is calculated between the values of each dimension. Then, max gap that is the maximum of the numbers among all gaps is added to the numbers after the local gap that is the largest gap in the desired dimension weight. These calculations have been shown in the formula (1).

$$\begin{cases} V = V \times V + \text{maxgap} & \text{if } V \text{ is after than localgap} \\ V = V \times V & \text{else} \end{cases} \quad (1)$$

By doing so, the possibility of clustering can be achieved with an emphasis on that particular dimension. After performing these steps, the results are presented to the system administrator and the manager can identify and solve the website problems based on them. In addition, if the administrator can identify a specific pattern for some of the attacks, he could add this pattern to the intrusion detection based on signatures. Measure of distance in both methods is Euclidean distance, which is expressed by the formula (2) [9].

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \quad (2)$$

In this formula p and q are two points in Euclidean n-space, then the distance from p to q, or from q top is given by d.

We use two criteria to evaluate system performance. The first criterion is care that is the total number of intrusions correctly diagnosed to the all intrusions and the second measure is the false alarm rate that is the number of diagnosed normal cases as intrusion to the total number of intrusions has given recognition [10].

5. EXPERIMENTAL RESULTS

Tables 1 and 2 show the results of testing the implemented system. Since the standard data sets of standard log file that contains the sent parameter and attacks are not available and the available data sets used in related studies are not be available due to security issues and maintaining confidentiality, we use its own data sets and we use the "www.zarcommerce.ir" log files as entries, also for the exact accuracy of results, we produce labeled log file generated for a hypothetical site. Both used input files are related to requests in PHP language.

Accuracy of the weighted clustering algorithm with a gap on all vectors is higher than clustering algorithm based on the number of abnormalities and percentage of mistake alarm is also less.

Separate feature vector with the weighted clustering base on gap, in a case that records related to attack in input data isn't negligible (manufactured file) acts better than other vectors and provides better resolution capability of a variety of attacks based on various criteria. In a small number of abnormalities, 3-gram vectors based on session with cluster based on the number of abnormalities are more efficient. ***Separate feature vector***, in both clustering algorithms will act better on manufactured logs and is less efficient on the real logs in which abnormalities are noises; because ***Separate feature vector*** has some characteristics that express data cluster separation and is less efficient in showing the difference between the noise data. Among these three feature vectors, section based 3-gram vector performs the best discrimination on the real data in both clustering algorithms.

Table 1 Test of means-K weighted clustering algorithm basedon gap

Separate feature vectors		3 gram session on based		3 gram		Input
False alarm	accuracy	False alarm	accuracy	False alarm	accuracy	
%50	%70	%22	%75	%25	%65	zarcommerce file log
0	%98	0	%80	%10	%66	manufactured log file

Table 2 Test of means-K clustering algorithm based on number of abnormalities

Separate feature vectors		3 gram session on based		3 gram		Input
False alarm	accuracy	False alarm	accuracy	False alarm	accuracy	
%48	%45	%19	%75	%32	%58	zarcommerce file log
%30	%61	%31	%65	%45	%50	manufactured log file

The results of the implementation of the system can generally expressed as follows:

- In both clustering, in the generated files respectively، the ***Separate feature vector***s, 3-gram based on session feature vectors and 3-gram feature vectors show greater accuracy and fewer false alarms.
- In both clustering, in the Zarcommerce files Respectively، 3-gram based on session feature vectors ,3-gram feature vectors and the ***Separate feature vector***s, greater accuracy and fewer false alarms
- In the case of noise abnormalities (such as Zarcommerce log file) in both clustering feature vector, gives better results according to both evaluation criteria.
- In the case of high abnormalities (such as created log file), in both clustering, separated feature vector, gives better results according to both evaluation criteria.
- Results execute weighted clustering algorithm based on the gap on the manufactured file in terms of accuracy and number of false alarms, gives better results than Zarcommerce file.
- ***Separate feature vector*** on the created log file works better and on the noise events have less efficiency in both the clustering algorithm.
- Intrusion detection accuracy and false alarm, in the weighted clustering algorithm based on the gap, usually is better than clustering algorithm based on the number of abnormalities.
- Highest precision in weighted clustering algorithm based on gap, is related to the separated feature vector (98% accuracy detection-0% false alarm).
- In the weighted clustering algorithm based gap, created log file have more accuracy and less false alarm rate than Zarcommerce log file.
- In the clustering algorithm based on anomaly, Zarcommerce log file (in the most cases) has more precise and less false alarm.)

6. CONCLUSIONS

This paper present an IDS system which provides a new approach for cyber threats detection by using web mining technology. The research uses three different methods to extract feature vectors from the users' sent requests and the applies tahn to two clustering algorithms based on K-means of abnormal behaviors. The provided system works well on real data and ideal manufactured log file and the error alarms are significantly reduced. This system is so useful to detect unknown influences. The main challenge of the proposed system determines the input clustering algorithms and relies to these values. The obtained experimental results show that the technology used in detecting attacks greatly increase accuracy and reduce false alarms.

REFERENCES

[1] Ariu, D. (2010). Host and Network Based anomaly detector for HTTP attacks. Cagliari: Ph.D. dissertation, Department of computing science, University of Cagliari.

[2] Axelsson, S. (1999). The base-rate fallacy and its implications for the difficulty of intrusion detection. In Proceedings of the 6th ACMConference on Computer and Communications Security ,pages1.7-

[3] C.Krugel, T. E. (2002). Service Specific Anomaly detection for network intrusion detection. Vienna: Distributed Systems Group, Technical University of Vienna.

[4] Dusan Stevanovic, A. A. (2012). Feature evaluation for web crawler detection with data mining techniques. Toronto, Ontario, Canada: Department of Computer Science and Engineering, York University, 4700 Keele St., Toronto, Ontario, Canada M3J 1P3.

[5] Goverdhan Singh, F. M. (2010). Mining Common Outliers for Intrusion. Advances in Knowledge Discovery and Management, SCI 292, pp. 217–234.Springer-Verlag Berlin Heidelberg.

[6] Robert.Cooley, B. M. (1997). Web mining: Information and Pattern Discovery on the World Wide Web. In International conference on Tools with Artificial Intelligence, pages 558-567.

[7] Teknomo, K. (2007). K-Means Clustering Tutorial. Teknomo, Kardi. K-Means Clustering Tutorials. http:\\people.revoledu.com\kardi\ tutorial\kMean\.

[8] Tuomo Sipola, A. J. (2012). Dimensionality Reduction Framework for Detecting Anomalies from Network Logs. Jyväskylä, Finland: Department of Mathematical Information Technology, University of Jyväskylä.

[9] Varun Chandola, E. E. (2006). Data Mining for Cyber Security. Minnesota: Department of Computer Science, University of Minnesota.

[10] W.Robertson, G. C. (2006). Using Generalization and characterization techniques in the anomaly based detection of web attack. san diego: 13th annual network and distributed system security symposium.

Authors

Mahza Mabzool received her B.S.c degree in information technology Engineering from Electrical and Computer Engineering faculty, Tabriz University, Tabriz, Iran in 2011. She is currently M.Sc. student in Computer Engineering (Artificial Intelligent) from Electrical and Computer Engineering faculty of Tabriz University, Iran. Her research interests include Web Programming, Web Security, Algorithm Design, Data Mining and Intrusion Detection Systems.

Mina Zolfy Lighvan received her B.Sc degree in Computer Engineering (hardware) and M.Sc. degree in Computer Engineering (Computer Architecture) from ECE faculty, university of Tehran, Iran in 1999, 2002 respectively. She received Ph.D. degree in Electronic Engineering (Digital Electronic) from Electrical and Computer Engineering faculty of Tabriz University, Iran. She currently is an assistant professor and works as a lecturer in Tabriz university. She has more than 20 papers that were published in different national and international conferences and Journals. Dr. Zolfy major research interests include Text Retrieval, Object oriented Programming & Design, Algorithms Analysis, HDL Simulation, HDL Verification, HDL Fault Simulation, HDL Test Tool VHDL, Verilog, hardware test, CAD Tool, synthesis, Digital circuit design & simulation.