Global Journal of Computer Science and Technology: G
Interdisciplinary
Volume 14 Issue 5 Version 1.0 Year 2014
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

A Text Mining-based Anomaly Detection Model in Network Security

By Mohsen Kakavand, Norwati Mustapha, Aida Mustapha & MohdTaufik Abdullah

Putra University, Malaysia

Abstract- Anomaly detection systems are extensively used security tools to detect cyber-threats and attack activities in computer systems and networks. In this paper, we present Text Mining-Based Anomaly Detection (TMAD) model. We discuss n-gram text categorization and focus our attention on a main contribution of method TF-IDF (Term frequency, inverse document frequency), which enhance the performance commonly term weighting schemes are used, where the weights reflect the importance of a word in a specific document of the considered collection. Mahalanobis Distances Map (MDM) and Support Vector Machine (SVM) are used to discover hidden correlations between the features and among the packet payloads. Experiments have been accomplished to estimate the performance of TMAD against ISCX dataset 2012 intrusion detection evaluation dataset. The results show TMAD has good accuracy.

Keywords: Text Mining, IDS, Anomaly Detection, TMAD Model, HTTP.

GJCST-G Classification: C.2.1 C.2.0

A Text Mining-based Anomaly Detection Model in Network Security

Mohsen Kakavand [α], Norwati Mustapha [σ], Aida Mustapha [ρ] & Mohd Taufik Abdullah [Ѡ]

Abstract- Anomaly detection systems are extensively used security tools to detect cyber-threats and attack activities in computer systems and networks. In this paper, we present Text Mining-Based Anomaly Detection (TMAD) model. We discuss n-gram text categorization and focus our attention on a main contribution of method TF-IDF (Term Frequency, Inverse Document Frequency), which enhance the performance commonly term weighting schemes are used, where the weights reflect the importance of a word in a specific document of the considered collection. Mahalanobis Distances Map (MDM) and Support Vector Machine (SVM) are used to discover hidden correlations between the features and among the packet payloads. Experiments have been accomplished to estimate the performance of TMAD against ISCX dataset 2012 intrusion detection evaluation dataset. The results show TMAD has good accuracy.

Keywords: Text Mining, IDS, Anomaly Detection, TMAD Model, HTTP .

I. Introduction

Changes in network security can be seen as a reliable estimation of transforming trends in the computer science. Information security is a significant issue in present and future life. It also seems to become more important with the development of cyber attacks against social media and mobile devices.

Robertson et al. (2006) referred to several web applications written by individuals with limited knowledge on security. According to CERT/CC, the number of cyber-attacks has increased from 1998 to 2002. Although attacks were relatively few in the early 1990s, a major increase has been reported since 2000 with about 25,000 attacks in 2000 (Malek & Harmantzis, 2004). According to the Common Vulnerabilities and Exposures list (CVE) (Christey & Martin, 2007) and a recent survey on security threats dealing with security risks in digital network world, susceptibility of web application was 25% of the total security issues (Malek & Harmantzis, 2004).

In 1980 James Anderson introduced intrusion detection systems (IDS) (Anderson, 1980) as a counter-action to the dramatic increase of hackers' attacks. There are two kinds IDS (Khalilian, Mustapha, Sulaiman, & Mamat, 2011): misuse detection (MD) and anomaly detection (AD). The latter type generates a model of normal behavior, and removes skeptical behavior or any abnormality from the normal behavior. Anomaly detection is able to identify new attacks, but its main weakness is its vulnerability to false positive alarms. Misuse detection system or signature-based system utilizes knowledge to directly identify the effects of intrusion with high detection precision but is unable to detect new threats and attacks.

Some intrusions use the susceptibilities of a protocol; other attacks attempt to examine a site by probing and scanning. The attacks can be identified by analyzing the network packet headers, or controlling the network traffic connection affairs and session behavior. Patterns in the header fields are such as protocol ID (TCP, UDP, ICMP), Quality-of Service (QoS) flags, port number, and particular values at typical header fields, such as checksums, options, and time to live (TTL). Furthermore, attacks, such as worms, include the delivery of anomaly payload to a susceptible application or service. These attacks might be identified by checking the packet payload. Moreover, examples of payload patterns involve strings such as "GET" and "POST" in the payloads of the HTTP GET and POST request packets. Figure 1 shows the structure of the HTTP GET-request packet, involving the "GET" subsequence pattern.

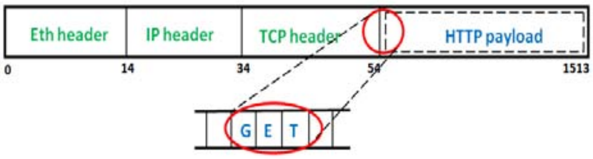


Figure 1: A HTTP GET-request packet

The present study raised the question that, how structural patterns can be identified and characterized? Thus, how can the packets matching those patterns be efficiently recognized? This paper presents a payload anomaly detection model, known as Text Mining-based Anomaly Detection (TMAD) based on data mining / machine leaning techniques.

This paper is organized as follows. In the second section, we review the related works and in the third section, we introduce text mining-based anomaly detection. In the fourth section, we present our system overview with its various subsections. Fifth section discusses the evaluation of TMAD model. Finally, we concluded the paper in the last section.

Author α σ : Faculty of Computer Science and Information Technology, University Putra Malaysia. e-mails : Kakavandirit@gmail.com, {norwati, aida_m, mtaufik} @upm.edu.my.

II. Related Work

Many studies have examined anomaly detection in network traffic. The major obstacles to its practicality are high false-positive rates and lack of clarity and transparency in the detection procedure. The previous methods did not represent adequate precision of false-positive rates. They also did not present diagnostic information to assist forensic analysis.

Some previous methods considered packet header information or statistical properties of sets of packets and connections. Packet header anomaly detection systems such as SPADE (Staniford, Hoagland, & Mcalerney, 2002), PHAD (Mahoney & Chan, 2001), Zhao et al., (2009) (Zhao, Huang, Tian, & Zhao, 2009), Guennoun et al. (2008) (Guennoun, Lbekkouri, & El-khatib, 2008) and Elbasiony et al. (2013) (Elbasiony, Sallam, Eltobely, & Fahmy, 2013) employing statistical approaches for anomaly network traffic and generate alarms when a huge deviation from the normal profile is observed. Feature selection from the packet headers has mostly been used in the mentioned systems. Table 1 summarizes the review on the packet header methods.

Analysis on packet header information typically reduces the data preprocessing necessities. Headers mainly constitute only a tiny part of the whole network data. Thus, processing needs less sources such as storage, memory and CPU. Additionally, features from packet header work quickly, with relatively low memory overheads and computation. Furthermore, they hinder some legal and privacy issues in regard to network packet analysis. Due to these benefits, several studies have utilized packet header as the major features in the intrusion detection systems. However, every request header feature set has its own features.

Therefore, they cannot be used to directly identify attacks bounded for application layer due to the attack bytes are often placed in the request body (Davis & Clark, 2011).

Table 1: Packet header approaches

Authors	Data input	Data preprocessing	Main algorithm	Method	Detection
SPADE, (Staniford et al.,2002)	Packet headers	Preprocessing hold packets with high anomaly score. Score is inverse of probability of packet event.	Entropy, mutual information, or Bayes network.	MD/AD	Probes
PHAD,(Mahoney and chan, 2001)	Ethernet, IP, TCP headers	Models each packet header using clustering	Univariate anomaly detection	AD	Probe, DoS
(Guennoun et al., 2008)	802.11 frame headers	Apply feature construction for 3 higher level features. Feature selection is used to find optimal subset.	K-means Classifier used to detect attacks	AD	Wireless network attacks.
(Zhao et al., 2009)	TCP sessions	Create separate dataset for each application protocol. Quantization of TCP flags within each session.	HMM for HTTP, FTP and SSH to model TCP state transitions.	AD	FTP anomalies
(Elbasiony et al., 2013)	Packet headers	Feature importance values calculated by the random forests algorithm are used in the misuse detection.	Random forests, k- means clustering, weighted k-means	MD/AD	DOS, R2L, U2R, probing

Currently, approaches such as (Estévez-Tapiador, Garcı́a-Teodoro, & Dı́az-Verdejo, 2004), PAYL (Wang & Stolfo, 2004), (Kruegel, Vigna, & Robertson, 2005), McPAD (Perdisci, Ariu, Fogla, Giacinto, & Lee, 2009), SensorWebIDS (Ezeife, Dong, & Aggarwal, 2008) and FARM (Chan, Lee, & Heng, 2013), were suggested for the analysis of packet payloads. These approaches were performed by defining features over payloads, and extracting models of normality based on these features. Packets being not fit into these models are anomalous and trigger alarms. These methods use fairly simple features computed over payload bytes.

Table 2 summarizes the reviewed on packet payload approaches. The table classifies payload anomaly detections works based on the kinds of data preprocessing, major algorithms and detection of various attack categories such as DoS, Buffer overflow, R2L, XML DoS, U2R, etc.

Payload data analysis seems to be more expensive than packet header data analysis as it needs deeper packet inspection, more computation and obfuscation analysis approaches. Due to payload data analysis is complicated, most studies consider tiny subsets of the payload data or only the client-side sections of web content (Davis & Clark, 2011).

This paper aims to examine language models derived from packet payload traffics. Additionally, this study attempts to develop a supervised and unsupervised learning algorithm that can be directly applied to extracted feature vectors. Thus, the present paper focuses on a major contribution of method TF-IDF (Term frequency, inverse document frequency) for improving an effective computation of measures between text categorization (n-grams).

Table 2. Packet payload approaches

Authors	Data input	Data preprocessing	Main algorithm	Method	Detection
(Tapiador et al., 2004)	Payload	Statistical analysis payload length and mean probability density and standard deviation	Markov chains	AD	HTTP attacks
PAYL(Wag et al., 2004)	Payload	1-g used to compute byte-frequency distribution models for each network destination	Mahalanobis Distance Map (MDM)	AD	Worms, Probe, DoS, R2L, U2R
(Kruegel,et al. 2005)	Web Requests	Construct content-based features from user supplied parameters in URL	Models of normal usage created for each web app. Compare requests to models.	AD/MD	Buffer overflow, Directory traversal, XSS, input validation, Code red
McPAD, (Perdisci, el al., 2009)	Payload	2v-grams extracted from payload. Feature clustering used to reduce dimensionality	One-Class SVM	AD	Shellcode attacks to web servers
SensorWeb IDS (Ezeife, el al., 2008)	Web Requests	network sensor for extracting parameters and the log digger for extracting parameters from web log files	Association Rule Mining (ARM)	AD/MD	XSS, SQL injection, DoS, buffer overflow, cookie Poison
FARM (Chan, el al.,2013)	Payload	Validating User ID, password, service request's input values, input size and SOAP size to from associative patterns and then matching these patterns with interesting rules	Fuzzy Association Rule Mining (FARM)	AD	SQLinjection, XML injection, XML content , SOAP oversized payload, coercive parsing, XML DoS

III. Text Mining-Based Anomaly Detection

Data mining seeks for patterns in data. Similarly, text mining seeks for patterns in text: Analyzing text involves extracting information useful for special goals(Ian H. Witten, Eibe Frank, 2011). Text mining seeks for patterns in natural language texts that are unstructured. Generally, text mining is influential in environments where huge numbers of text documents are managed. Knowledge discovery from text (KDT) considers the machine supported analysis of text. KDT employs methods from information retrieval, information extraction and natural language processing (NLP). Then, it links these three to the algorithms and approaches of Knowledge discovery from data (KDD), data mining, machine learning and statistics (Hotho, Andreas, Paaß, & Augustin, 2005). Text mining is a popular area in anomaly detection which commonly involves the tasks such as clustering, classification, semi-supervised cluster, and so on.

Classification-Based Anomaly Detection analyzes a set of data and generates a set of grouping rules that can categorize future data or predict future data trends called supervised learning. There are different sorts of classification approach such as decision tree induction, Bayesian networks, k-nearest neighbor classifier. In this case, main idea is build a classification model for normal and anomalous events based on labeled training data with require knowledge of both normal and anomaly (attacks) class. The learned model is then applied on the test dataset in order to classify unlabeled records into normal and anomalous records in order to classify each new unseen event (Amer, Goldstein, & Abdennadher, 2013).

Clustering-Based Anomaly Detection is second learning approach called unsupervised learning. Here, the data have no labeling information. Furthermore, no separation into training and testing phase is given. Unsupervised learning algorithms assume that only a small fraction of the data is anomaly and that the attacks exhibit a significantly different behavior than the normal records.

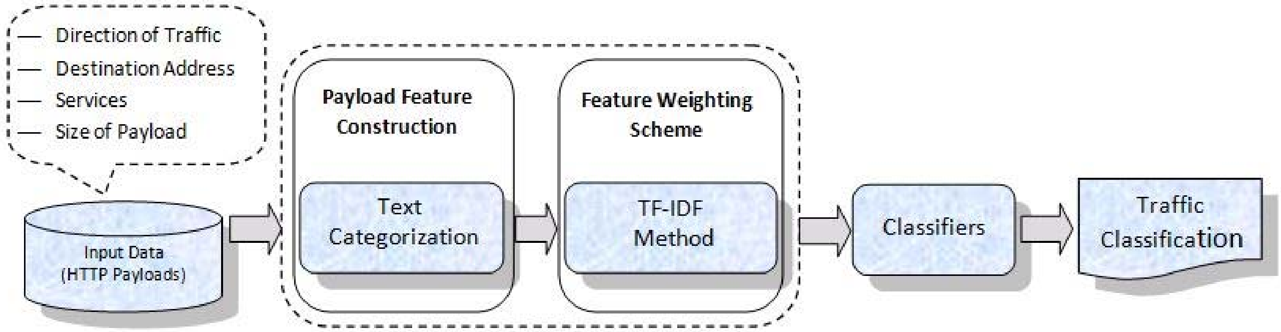


Figure 2 : Overview of text mining based-anomaly detection model

In many practical application domains, the unsupervised learning approach is particularly suited when no labeling information is available. Moreover, in some applications the nature of the anomalous records is constantly changing. Thus, obtaining a training dataset that accurately describe anomaly is almost impossible. On the other hand, unsupervised anomaly detection is the most difficult setup since there is no decision boundary to learn and the decision is only based on intrinsic information of the dataset (Amer et al., 2013).

Semi-Supervised Cluster Analysis, in contrast with classification, is without direction from users. Thus, it may not produce highly valuable clusters. The quality of unsupervised can be highly developed through some weak forms of supervision. Such a clustering is called semi-supervised that is based on user's feedback or guidance constraints is called. (Jiawei Han and Micheline Kamber, 2011). In semi-supervised anomaly detection method, the algorithm models are the only normal records. Records that do not fit into this model are called outliers in the testing stage. Advantages of this semi-supervised anomaly detection can be easily understood of Models as well as normal behavior can be accurately learned but possible high false alarm rate - previously unseen (yet legitimate) data records may be recognized as anomalies (Amer et al., 2013).

However, TMAD model uses a machine learning method and statistical anomaly detection approach to determine an anomaly behavior in the network.

IV. System Overview

This section, represents a comprehensive introduction about the TMAD applying text mining techniques into payload-based anomaly detection. Additionally, the most important contribution of TMAD is the combination of TF-IDF approach and payload-based anomaly detection systems, which have not been investigated in previous studies. This intrusion detection system uses statistical analysis and machine learning algorithms, respectively and then comparison among supervised and unsupervised classifiers. Characters come into network traffic will be analyzed by Mahalanobis Distances Map (MDM) and Support Vector Machine (SVM) to recognize abnormal traffic data from normal ones. Figure 2 shows the overview of TMAD model process.

Text categorization: the text categorization functionality was primarily used. n-Gram text categorization (Wang & Stolfo, 2004), (Banchs, 2013) is in charge of feature construction and request feature analysis. It extracts raw payload features using n-gram (n=1) text categorization method from packet payload and transform observations into a series of feature vectors. Each payload is represented by a feature vector in an ASCII character (256-dimensional).

TF-IDF method: This study investigates the geometrical framework of language modeling. Furthermore, the vector space model and term frequency inverse document frequency (TF-IDF) weighting scheme are examined (Banchs, 2013). Term weighting schemes are often employed to develop the performance. In these schemes, the weights show the significance of a word in a particular document of the selected collection. Huge weights are appointed to terms often used in relevant documents but seldom in the whole document collection (Hotho et al., 2005). Thus, the data resources are processed and the vector space model is set up in order to represent a convenient data structure for text classification. This method is employed to explore similarity between the normal behaviors with the novel input traffic data. The vector space model presents documents as vectors in m-dimensional space, i.e. each document d is explained by a numerical feature vector $W(d) = (x(d, t_1), \ldots, x(d, t_m))$. Therefore, a weight for$W(d, t)$ a term t in document d is computed by term frequency tf (d,t) time inverse document frequency idf (t), describing the term specificity within the document collection. In addition to term frequency and inverse document frequency — defined as (1), a length normalization factor is employed to guarantee that all documents have equal chances of retrieving independent of their lengths (2). Where N is the size of the document collection D and nt is the number of documents in D containing term t.

$$idf\ = (t) := \ log(N/n_t) \tag{1}$$

$$W(d,t) = \frac{tf\ (d,t) log\ (N/n_t)}{\sqrt{\sum_{j=1}^{m} idf\left(\mathrm{d,t_j}\right)^2 (\ log\ (N/n_{t_j}))^2}}, \tag{2}$$

In the following, the application of geometrical framework model in payload-based anomaly detection is explained.

Classifiers: two distinctive types of algorithms such as Mahalanobis Distances Map (MDM) (Wang & Stolfo, 2004) and Support Vector Machine (SVM) (Scholkopf et al., 1996) are used. Mahalanobis Distances Map (MDM), some factors such as mean value and standard deviation are applied to each byte's frequency. For a payload model, the feature vector that is a set of relative frequencies is the occurrences of each ASCII character to the total number of characters that appear in the payload. Generally, each feature vector can be presented as (3). Then, the mean value and standard deviation of each byte's frequency are computed and explained as (4), (5), (6) and (7) respectively. The mean value and standard deviation vectors, and , are stored in a model M.

$$X = [x_0\, x_1\ \dots\ x_{255}] \tag{3}$$

$$\bar{X}\ = [\bar{x}_0\, \bar{x}_1\ \dots\ \bar{x}_{255}] \tag{4}$$

$$\bar{\sigma}\ = [\bar{\sigma}_0\, \bar{\sigma}_1\ \dots\ \bar{\sigma}_{255}] \tag{5}$$

$$\bar{x}_i\ = \frac{1}{n} \sum_{k=1}^{n} x_{i,k} \quad (0 \le i \le 255) \tag{6}$$

$$\bar{\sigma}_i\ = \sqrt{\frac{1}{n} \sum_{k=1}^{n} (x_k\ - \ \bar{x}_i)^2} \quad (0 \le i \le 255) \tag{7}$$

The model considers the correlations among various features (256 ASCII characters). Therefore, for each network packet, a feature vector is defined by (3). Here, there are the average value of features in the 1-gram model (8) and the covariance value of each feature (9). To investigate the association among the characters, where μ is the average frequency of each ASCII character presented in the payload, Σi is the covariance value of each feature. Next, the classifiers between two characters (10) are presented:

$$\mu\ = \frac{1}{256} \sum_{i=0}^{255} x_i \tag{8}$$

$$\sum \boldsymbol{i}\ = (x_i - \mu)\, (x_i - \mu)' \ \ (0 \le i \le 255) \tag{9}$$

$$\boldsymbol{d}_{(\boldsymbol{i},\boldsymbol{j})} = \frac{(x_i - x_j)\, (x_i - x_j)'}{\sum \boldsymbol{i} + \sum \boldsymbol{j}} \quad (0 \le i \le 255) \tag{10}$$

MDM is a statistical method or a model-based method that is created for the data. The objects of the study are evaluated with respect to how well they fit into the model. According to the above evaluation, the MDM of a network packet is made as matrix D. Then, the Mahalanobis distance between two distributions of D and the model *M* is evaluated. Then, the weight w is calculated using (11), if the weight is larger than a threshold, the input packet is considered as an intrusion.

$$W\ = \sum_{i,j}^{255,255} \frac{(\,d_{ob\,(i,j)} - \ \bar{d}_{nor\ (i,j)})2}{\sigma^2_{nor\ (i,j)}} \tag{11}$$

In the following, the Support Vector Machine (SVM) algorithm is chosen. This algorithm is originally proposed by Scholkopf et al. in (Scholkopf et al., 1996). SVM have been shown to achieve good performance in text mining classification problems (Sebastiani, 2002). It is also known as a supervised classification algorithm that is able to process feature vectors of high dimensions for providing a fast and influential approach for learning text classifiers (LEOPOLD & KINDERMANN, 2002). Typically, document d is presented by a – possibly weighted – vector $(t_{d1}, \dots t_{dN})$ of the counts of its featur. A SVM can only separate two groups — a positive group L1 (shown by y = +1) and a negative group L2 (shown by y = -1). In input vectors, a hyperplane might be defined by setting y = 0 as follows:

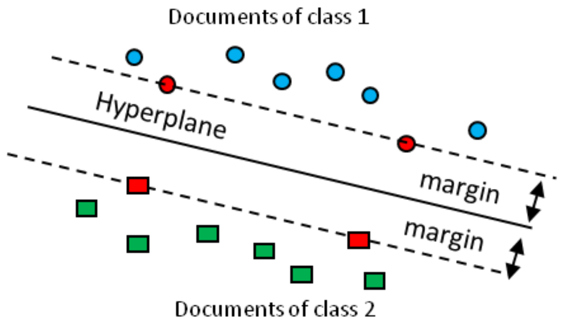


Figure 3 : Decision boundary and margin of SVM

The SVM algorithm identifies a hyperplane placed between the positive and negative examples of the training set. Figure 3 shows the parameters bj are adapted in a manner that the distance ε– called margin – between the hyperplane and the closest positive and negative example packet payload is maximized. This considers a limited quadratic optimization issue which can be resolved for a huge number of input vectors. The documents that have distanceε from the hyperplane are known as support vectors and identify the actual location of the hyperplane (12).

$$y\ = \int \left(\vec{t}_{\,d}\right)\ = b_0\ + \sum_{j=1}^{N} b_j t_{dj} \tag{12}$$

Typically only a tiny fraction of payloads are considered as support vectors. A new document with term vector $\vec{t}_{\,d}$ is classified in L1 if the value $\int \left(\vec{t}_{\,d}\right) > 0$ and into L2 otherwise. If the packet payload vectors of the two classes are not linearly separable, a hyperplane is selected. Classifier approaches identify patterns of packet payloads in network traffic data. These are undertaken in extracting the hidden correlations between features and the correlations among network packet payloads.

V. Experimental Results

This section represents results of the extensive experiments performed. This study examined TMAD model on the ISCX dataset 2012 (Shiravi, Shiravi, Tavallaee, & Ghorbani, 2012), reflecting current trends traffic patterns and intrusions. This is in contrast to static datasets that are widely used today but are outdated, unmodifiable, inextensible, and irreproducible. ISCX dataset 2012 is considered as a new standard data set for evaluation of intrusion detection systems.

a) ISCX Dataset 2012

To evaluate TMAD model, ISCX dataset 2012 (Shiravi et al., 2012) were collected under the sponsorship of Information Security Centre of Excellence (ISCX). All the network traffic of the data set was included in both normal network traffic and attack traffic for system evaluation of the proposed approach for text mining based-anomaly detection. ISCX dataset 2012 contains categories of attacks including scan, DoS, R2L, U2R and DDoS.

The entire ISCX labeled dataset comprises nearly 1512000 packets and covered seven days of network activity. However, the ready-made training and testing dataset is not available. Thus, ISCX HTTP/GET traffic was randomly divided into two parts: a training set made of approximately 80% of the HTTP/GET traffic and a testing set made of the remaining 20% of the traffic. The present study focuses on the anomalous coming through inbound HTTP requests. HTTP-based attacks are mainly from the HTTP GET request at the server side. Request bodies carry data to the web server but sometimes the request message has no body, because no request message data is needed to give GET a simple document from a server (Davide & Brian, n.d.). Therefore, we removed all HTTP request packets begin without request message.

b) Analysis And Result

We directed experiments for our TMAD model with the extracted data from the ISCX dataset 2012. In the first part of our experiments, we present the model generation for normal HTTP traffic. Afterwards, we evaluate the accuracy of our TMAD model in detecting various attacks coming through HTTP services including scan, DoS, R2L, U2R and DDoS. We trained the TMAD model on the ISCX dataset 2012, and then evaluate the model on test dataset, which contains different attacks. For port 80, the attacks are often malformed HTTP requests and are very different from normal requests.

For our research, we have plan on the anomalous incoming through inbound HTTP requests. HTTP-based attacks are mostly from the HTTP GET request at the server side. Figure 4 shows histogram of HTTP request distributions.

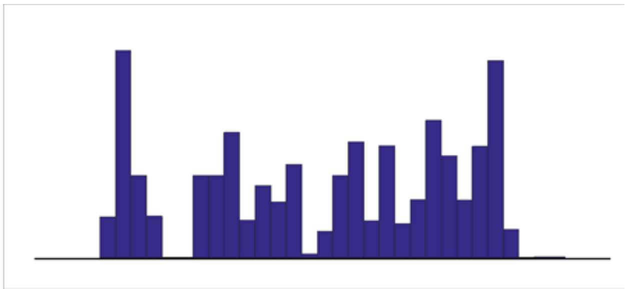


Figure 4 : Histogram of HTTP request distributions

Figure 5 provides an example displaying the variability of the frequency distributions from port 80. The plot represents the characteristic profile for that port 80 and flow direction (inbound) full length payloads.

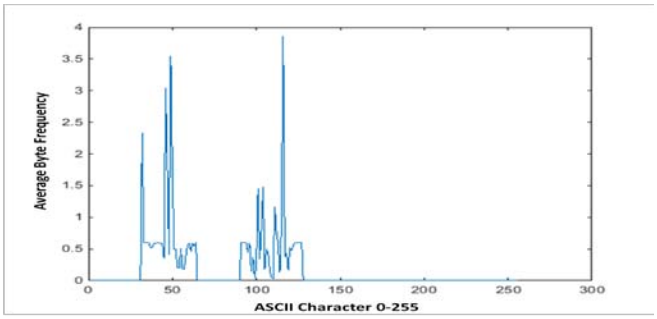
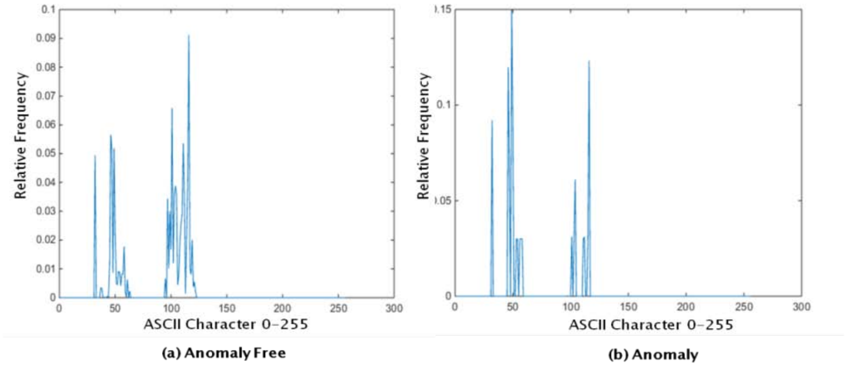


Figure 5 : Example byte distribution for port 80

As illustrated in figure 6, (a) and (b) indicate the anomaly free and anomaly character relative frequencies. We can see the character relative frequencies of anomaly packet payloads are very different from the normal packet payloads, which can distinguish anomalous from normal packet payloads.



(a) Anomaly Free

(b) Anomaly

Figure 6 : Relative Frequencies of Characters of Normal and Anomaly

Concurrently, the experimental results illustrated the acceptable performance of the TMAD model in detecting HTTP anomaly on web services. That is the knowledge from the TF-IDF method explaining the correlation among 256 ASCII characters. The results showed that the Text Mining-Based Anomaly Detection (TMAD) is able to detect new attacks with different detection rate and false positive rate in MDM and SVM algorithms. We also used Receiver Operating Characteristic (ROC) curve method to compare the performance of our model with MDM and SVM. The ROC curve showed incorrectly flagging non-attack requests as an attack (false positives) and detection rate attack. TP and FP rate are shown in figure 7 in ROC curve.

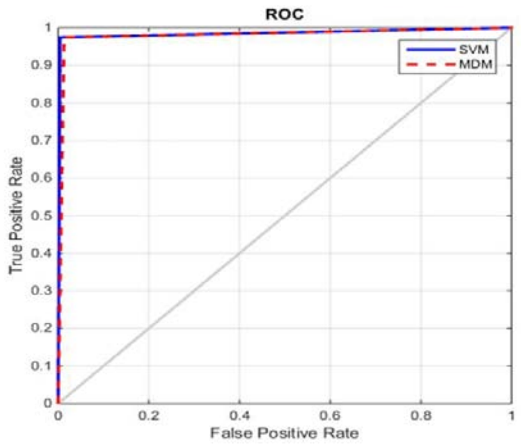


Figure 7 : ROC Curve for the Accuracy of the TMAD model

The results obtained for the model are very encouraging. TMAD has a detection rate around 97.44% and 1.3% false positive rates for MDM and detection rate around 97.45% with 0.4% false positive rates for SVM. Performance obtained by TMAD model in training and test data is presented in Table 3.

Table 3 : MDM and SVM using Training and Testing Dataset

Algorithm	Training Data		Testing Data	
	Detection Rate	False Positive	Detection Rate	False Positive
MDM	93.46%	3.3%	97.44%	1.3%
SVM	97.61%	1.6%	97.45%	0.4%

VI. Conclusions

This study presented TMAD (Text Mining-Based Anomaly Detection), a new approach to detect HTTP attacks in the network traffic. TMAD is an anomaly detector based upon text categorization and TF-IDF method. The use of TF-IDF improves the performance usually term weighting schemes are used, where the weights reflect the importance of a word in a specific document of the considered collection.

The experimental result indicated that the method is effective at detection rate, but the notable detection accuracy suffered from high level of false positive rate for ISCX dataset 2012 (Shiravi et al., 2012) collected under the sponsorship of Information Security Centre of Excellence (ISCX). For port 80, it achieved almost 97.44% detection rate with around 1.3% false positive rate in unsupervised learning method, and 97.45% detection rate with around 0.4% false positive rate in supervised learning. In future research, we intend to reduce the dimensionality of feature space and false positive rate by applying data-mining preprocessing techniques to ISCX dataset 2012.

In our future work we aim to evaluate the performance of TMAD model on 1999 DARPA/MIT Lincoln Laboratory, which produced the most prominent datasets for testing IDS. Moreover, we will try to reduce of false positive rate and improve detection rate.

VII. Acknowledgement

This research work supported by the Fundamental Research Grant Scheme (FRGS) under Ministry o Higher Education, Malaysia and it is gratefully appreciated.

References Références Referencias

1. Amer, M., Goldstein, M., & Abdennadher, S. (2013). Enhancing one-class support vector machines for unsupervised anomaly detection. Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description- ODD '13, 8–15. doi:10.1145 /2500853.2500857
2. Anderson, J. P. (1980). Computer Security Threat Monitoring And Surveillance (pp. 1–56). Washington.
3. Banchs, R. E. (2013). Text Mining with MATLAB® (pp. 1–347). New York, NY: Springer New York. doi:10.1007/978-1-4614-4151-9
4. Chan, G.-Y., Lee, C.-S., & Heng, S.-H. (2013). Discovering fuzzy association rule patterns and increasing sensitivity analysis of XML-related attacks. Elsevier Science, Journal of Network and Computer Applications, 3
5. Christey, A. S., & Martin, R. A. (2007). Vulnerability Type Distributions in CVE. MITRE, Common Weakness Enumeration (CWETM). Retrieved May 22, 2007, from http://cwe.mitre.org/documents/vuln-trends/index.html
6. Davide, G., & Brian, T. (n.d.). HTTP The Definitive Guide (2nd ed., p. 617). United States of America: Publisher: O'Reilly Media.
7. Davis, J. J., & Clark, A. J. (2011). Data preprocessing for anomaly based network intrusion detection: A review. Computers & Security, 30(6-7),353–375. doi:10.1016/j.cose.2011.05.008
8. Elbasiony, R. M., Sallam, E. a., Eltobely, T. E., & Fahmy, M. M. (2013). A hybrid network intrusion detection framework based on random forests and weighted k-means. Ain Shams Engineering Journal, 4(4), 753–762. doi:10.1016/j.asej.2013.01.003
9. Estévez-Tapiador, J. M., García-Teodoro, P., & Díaz-Verdejo, J. E. (2004). Measuring normality in HTTP traffic for anomaly-based intrusion detection. Elsevier, Computer Networks, 45(2), 175–193. doi:10.1016/j.comnet.2003.12.016
10. Ezeife, C. I., Dong, J., & Aggarwal, A. K. (2008). SensorWebIDS: a web mining intrusion detection system. Emerald, International Journal of Web Information Systems, 4(1), 97–120. doi:10.1108/17440080810865648
11. Guennoun, M., Lbekkouri, A., & El-khatib, K. (2008). Selecting the Best Set of Features for Efficient Intrusion Detection in 802 . 11 Networks. In in:Information and communication technologies: from theory to applications, ICTTA 2008. 3rd International Conference. (pp. 1–4).
12. Hotho, A., Andreas, N., Paaß, G., & Augustin, S. (2005). A Brief Survey of Text Mining. LDV Forum - GLDV Journal for Computational Linguistics and Language Technology, 1–37.
13. Ian H. Witten, Eibe Frank, M. A. H. (2011). Data Mining Practical Machine Learning Tools and Techniques (Edition, T., Vol. 40, pp. 1–629). Elsevier. doi:10.1002/15213773(20010316)40:<923 ::AID-ANIE9823>3.3.CO;2-C.
14. Jiawei Han and Micheline Kamber. (2011). Data Mining: Concepts and Techniques (Third Edit.). San Francisco, CA, USA: Morgan Kaufmann.
15. Khalilian, M., Mustapha, N., Sulaiman, N., & Mamat, A. (2011). Intrusion Detection System with Data Mining Approach: A Review. Global Journal Of Computer Science & Technology, 11(5).
16. Kruegel, C., Vigna, G., & Robertson, W. (2005). A multi-model approach to the detection of web-based attacks. Elsevier Science, Computer Networks, 48(5),717–738. doi:10.1016/j.comnet.-20 05.01.009
17. LEOPOLD, E., & KINDERMANN, J. (2002). Text Categorization with Support Vector Machines .How to Represent Texts in Input Space ? Kluwer Academic Publishers Machine Learning, (22), 423–444.
18. Mahoney, M. V, & Chan, P. K. (2001). PHAD : Packet Header Anomaly Detection for Identifying Hostile Network Traffic. Florida Institute of Technology Technical Report, (1998), 1–17.
19. Malek, M. ;, & Harmantzis, F. (2004). Data mining techniques for security of web services. In Proc. International Conference on E-Business and Telecommunication Networks (ICETE) (Vol. 1–14, pp. 1–14). Retrieved from http://www.stevens-tech.edu/perfectnet/publications/Papers/ICETE2004-paper.pdf
20. Perdisci, R., Ariu, D., Fogla, P., Giacinto, G., & Lee, W. (2009). McPAD : A Multiple Classifier System for Accurate Payload-based Anomaly Detection. Elsevier Science, Computer Networks, 5(6), 864–881.
21. Robertson, W., Vigna, G., Kruegel, C., & Kemmerer, R. A. (2006). Using Generalization and Characterization Techniques in the Anomaly-based Detection of Web Attacks. In: Proceedings of the Network and Distributed System Security Symposium (NDSS), 14. Retrieved from http://iseclab.org/papers/webfuzzing.pdf
22. Scholkopf, B., Sung, K., Burges, C., Girosi, F., Niyogi, P., Poggio, T., & Vapnik, V. (1996). Comparing Support Vector Machines with Gaussian Kernels to Radial Basis Function Classifiers. IEEE Transactions on Signal Processing, (1996-12-01). doi:10.1109/78.650102
23. Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34(1), 1–47. doi:10.1145/505282.505283
24. Shiravi, A., Shiravi, H., Tavallaee, M., & Ghorbani, A. a. (2012). Toward developing a systematic

approach to generate benchmark datasets for intrusion detection. Computers & Security, Elsevier, 31(3), 357–374. doi:10.1016/j.cose.2011.12.012

25. Staniford, S., Hoagland, J. A., & Mcalerney, J. M. (2002). Practical automated detection of stealthy portscans. Journal of Computer Security, 10, 105–136.
26. Wang, K., & Stolfo, S. J. (2004). Anomalous Payload-based Network Intrusion Detection. Springer Berlin Heidelberg,7th International Symposium, RAID, Sophia Antipolis, France, September 15 - 17, Proceedings, pp 203–222.
27. Zhao, J., Huang, H., Tian, S., & Zhao, X. (2009). Applications of HMM in Protocol Anomaly Detection. In 2009 International Joint Conference on Computational Sciences and Optimization (pp. 347–349). Ieee. doi:10.1109/CSO.2009.51

This page is intentionally left blank