Mitigating E-Mail Threats - A Web Content Based Application

R. Dhanalakshmi, C. Chellappan

Abstract—The World Wide Web is a very powerful and interactive medium and its surveillance is unavoidable for information dissemination. Extracting valuable information from the vast unstructured data is a challenging and critical issue. Web content mining plays an important role in solving these issues. The applications of WWW are widespread and one among it is E-Mail communication. Due to its simple and inherently vulnerable nature, e-mail communication is harmed for various purposes. E-mail spamming, phishing, relay hijacking, Denial of service attacks, cyber bullying, child pornography, and sexual harassment are some common E-mail mediated cyber crimes. It becomes very essential to provide a protective mechanism for securing E-Mail systems. Content based web mining plays a vital role in the detection of E-Mail threats by examining the contents of suspected e-mail accounts to gather evidence during malicious behavior. There are various techniques available with different approaches and this paper is based on Content based filtering methods with machine learning algorithms.

Index Terms —Content Mining, Data Leakage, E-Mail threats, Spam, Phishing

I. INTRODUCTION

Web mining is a multi- disciplinary effort that draws techniques from fields like information retrieval, statistics, machine learning, natural language processing, and others. Web mining is commonly divided into the following three sub-areas:Web content Mining, Web Structure Mining and Web Usage Mining.

Web Content Mining deals with the application of data mining techniques to un-structured or semi structured text, typically html documents. Web Content Mining is a process designed to explore large amounts of data by capturing consistent patterns and relationships between data objects. Its finds its applications in various areas such as information retrieval from large databases, search engines, Automated answering systems and E-Mail Sorting applications etc. Web Structure Mining uses the hyper-link structure of the web analyzing its structural properties. Web Usage Mining performs the analysis of user interactions or behavioral patterns with a web server. There are various web threats which are emerging in day-today life and some among them are financially motivated, Identity theft and confidential data leakage. The most significant trend is towards targeted Attacks on both individuals and businesses. Some of the Web threats find its source from E-Mail communication and it is depicted in the following figure 1.

R.Dhanalakshmi is with the Department of Computer Science and Engineering at Anna University ,Chennai, India. (Email: dhanalakshmisai@gmail.com)
C. Chellappan is a Professor in the Department of Computer Science and Engineering at Anna University, Chennai, India. (Email: drcc@annauniv.edu)

As E-mail becomes a popular means for communication over the Internet, the problem of receiving unsolicited and undesired E-mail's, called spam or junk mails, arise severely. The volume of E-mail received and the amount of spam is constantly growing. Spam mails are defined as electronic messages posted to thousands of recipients usually for advertisement or profit. Some of the Spam E-Mails transform as Phishing E-Mails seeking users' confidential data and accessing their bank accounts for financial fraud. It is a form of identity theft misused by the hackers and malicious users.

The identified E-Mail threats are classified as In-Bound and Out-Bound E-Mail threats and are explained as follows and focused in this paper

A. IN-BOUND E-MAIL THREATS- SPAM/PHISHING E-MAIL

A form of Un-Solicited E-Mail called as Spam an in-bound E-Mail threat which fills the users' inbox thereby causing consumption of computer and network resources, bandwidth and large amount of storage space on mail servers. It also causes loss of Legitimate mail and time in reading junk unsolicited mails. Some of its issues are

- The growing sophistication of phishing attacks.
- Attachment based Spam/Phishing emails.
- Image-based Spam.
- Use of botnet for delivering spam.

The emerge of phishing E-Mails is a form of identity theft motivated for financial fraud by obtaining the users credentials. E-mails with fake Web site links are also sent in an effort to extract confidential information from a user such as financial account information, social security information, credit card numbers, and so on. A user may be directed to a phishing site via email or from another site. Some of the E-mails may contain attachment based spam and phishing content to fool the spam filters which usually looks for the content in the body of the Email alone. Apart from examining the attachments, the content (body of the email) alone may also be prone to word obfuscation technique to fool the spam filters.

B. OUT-BOUND E-MAIL THREATS - CONFIDENTIAL DATA LEAKAGE

Any information that is being transmitted over the Internet must be considered at risk from being seen or in some way tampered with. In an organization, E-Mail policies should be carefully designed and implemented that any confidential data

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

is masqueraded and sent outside in any form. For example, .exe files should ensure that it is not sent by changing it file type to .pdf or .doc which is allowable. There are other various approaches available such as source based filtering, White list/Blacklist, URL/IP address filtering based on Web structure and web usage mining based on the users' behavior. But a Web Content Mining proves to be better solution in mitigating In-bound and Out-bound E-Mail threats.

II. ANTI-SPAM/PHISHING TECHNIQUES

Anti-spam techniques may be classified as illustrated in Fig 1.

A. CONTENT BASED FILTERING

This technique attempts to computationally distinguish between a spam e-mail and a legitimate e-mail using machine learning techniques and the identification of spam and phishing emails are quite different. A spammer advertises a product while the phisher has to deliver a message that has an unsuspicious look and pretends to come from some reputable institution. Abu-Nimeh et al. [13] discussed above the various statistical filters such as SVM, K-NN and BART etc to identify phishing mails based on specific keywords. The machine learning techniques includes decision tree, naive Bayesian [1] [4], SVM (Support Vector Machine),BART (Bayesian Additive Regression Tress), K-Nearest neighbor algorithm [1] and artificial neural networks .Naïve Bayes Filters plays a very important role but it suffers from the limitation of Bayesian Poisoning (word obfuscation) by adding irrelevant words which are not in the Spam database.

B. SOURCE ADDRESS BASED FILTERING

A variety of techniques exist to trace the source of the E-Mail such as Blacklist/white list IP addresses and E-mail Addresses. By tracing the source of the E-mail w.r.to their sending domain and authorized senders. Some techniques involves identification of Spoofed E-Mails [2] based on various techniques such as SPF (Sender Policy Framework), Sender ID, Domain Keys Identified Mail (DKIM).

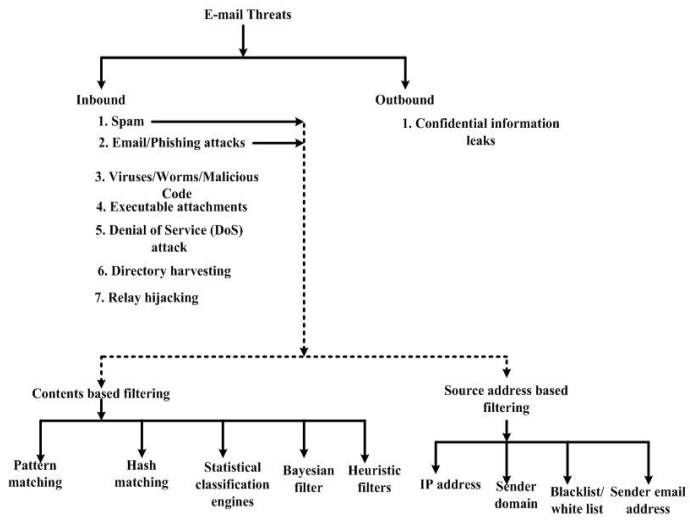

Fig 1: Classification of E-Mail Threats

III. RELATED WORK

There are many statistical filters available in the literature. Naïve Bayesian is a fundamental statistical approach based on probability initially proposed by Sahami et al. [14]. The Bayesian algorithm predicts the classification of new E-Mail by identifying an E-Mail as spam or legitimate. This is achieved by looking at the features using a 'training set' which has already been preclassified correctly and then checking whether a particular word appears in the e-mail. High probability indicates the new e-mail as spam e-mail. Androutsopoulos et al. [1] enhanced with the effect of attribute-set size, training-corpus size, lemmatization, stop-lists and improved the performance with cost sensitive evaluation. Graham [4] has used statistical filtering for spam detection using one corpus of spam and another one of non-spam emails. It obtained 99.5%spam with 0.03% false positives. Another classifier, k-Nearest Neighbor (KNN), maps a document to features and measures the similarity to the k-nearest training documents. Lobato et.al [7] used binary classification based on an extension of Baye's point machines. Yang and Elfayoumy [17] evaluated the effectiveness of feed forward back propagation Neural Network and Bayesian classifiers for spam detection. Meizhen et al. [11] proposed a model for spam behavior recognition based on fuzzy decision tree (FDT).

Previous research in identification of file types includes the use of file ''fingerprints'' by McDaniel& Heydari [9] and in this approach the ''fileprint'' of a collection of files is the histogram of byte frequencies together with their variance. Mehdi et al [10] uses Principle Component Analysis [PCA] and unsupervised neural networks for the automatic feature extraction and obtained 98.33% correct classification rate .In OSCAR method [6] the centroid, of a file type containing mean value vector and standard deviation vector is calculated by looking at the byte frequency distribution (BFD) of the byte stream. The centroid is then compared to a data sample by calculating the difference between the sample's BFD vector and the centroid's mean value vector. The differences are weighted by the standard deviation vector. A sum of squares distance metric is used and attained 99.2% accuracy but the approach is much limited with JPEG data.

Wei-Hen Li et al. [15] identify file types using n-gram analysis. They calculate 1-gram frequency distribution of files and build 3 different models of each file type: single centroid (one model of each file type), multi-centroid (multiple models of each file type), and exemplar files (set of files of each file type) as centroid using mean and standard deviation of 1-gram frequency distribution of files. They use Mahalanobis and Manhattan distance to compare these models with 1-gram distribution of given file to find the closest model. Calhoun and Coles [16] has build classification models (based on the ASCII frequency, entropy, and other statistics) and apply linear discriminant to identify file types. Irfan [5] proposed a recursive methodology for fast file type identification using the cosine similarity as a better metric than Mahalanobis distance in terms of classification accuracy, smaller model size, and faster detection rate.

IV. NAÏVE BAYES FILTERING

Bayesian spam filtering which is best suitable for spam detection finds a vital role in identifying phishing mails also. Phishing E-Mails are designed to look like legitimate E-Mails but aimed particularly for financial gain. To accurately catch phishing emails, Bayesian filters must be specifically designed for that purpose. Naïve Bayesian is a text classifier algorithm that analyzes textual features of an email to identify

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

it as a ham or spam or phish email based on probabilistic scoring of its textual attributes. The Naïve Bayesian approach consists of two phases – training phase and the classification phase as in Fig 2. The Naïve Bayes filter examines a set of known spam emails and a set of emails known to be legitimate. This filter is based on the Bayes theorem. Applied to Spam/Phish, it states that the probability of an E-Mail being Spam is equal to the probability of finding the same words in this E-Mail and Spam, times the probability that any email is spam, divided by the probability of finding those words in an arbitrary email. The same approach can be used for phishing email detection also and it is expressed in a conditional probability formula:

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$$

- $\Pr(A|B)$ is the probability that a message is spam/phishing which should contain the word B.
- Pr (B|A) is the probability of the word B in spam/phishing. This value is computable from the training collection.
- Pr (A) is the probability that the email is spam/phishing (i.e. the number of spam /phishing messages divided by the number of all emails in the training collection).
- Pr (B) is the probability of word B in the collection. Each word in the email contributes to the e-mail's spam probability.

A. TRAINING PHASE

The training phase scans an existing corpus of spam and ham and phish emails. It involves

Parsing - An email is parsed to identify different sections such as headers, body, to, from, subject, etc. Based on different filters different parsing techniques are used.

Tokenization - Tokenization consists of creating tokens from different sections of email. These tokens will be later used to classify emails.

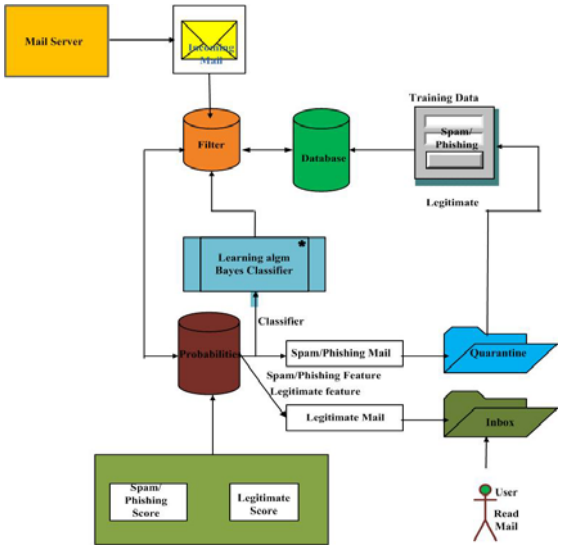

Fig 2: Naïve Bayes Classifier for Spam/Phishing Emails

B. CLASSIFICATION PHASE

In classification phase an incoming email is classified as a spam or ham or phish. An incoming email is first tokenized to get individual tokens. The corresponding probabilities for each token are retrieved .Naïve Bayesian formula is used to classify this email as Ham or Spam using these probabilities.

C. PERFORMANCE & COST EVALUATION - FALSE POSITIVES & FALSE NEGATIVES

A false positive is falsely classifying a legitimate email as a spam, and a false negative is falsely classifying a spam as a legitimate email. The cost of a false positive is much higher than that of a false negative. False positives may lead to the loss of legitimate mails to be lost by the user as they may be misclassified as spam. In that case, it is acceptable to allow some false negatives rather than having any false positives.

Let L→S be false positive error type and S→L be false negative error type. Assuming that L→S is λ times costlier than S→L, we classify a message as spam if:

$$\frac{P(C = spam \mid X = x)}{P(C = legitimate \mid X = x)} > \lambda$$

If we are considering a Naïve Bayesian filter's independency, the assumption holds. Therefore,

P(C=spam | X=x) = 1 - P(C=legitimate | X=x),

This leads to:

P(C=spam | X=x) > t,

Where t = threshold value

Thus $t = \lambda / (1+\lambda)$ as $\lambda = t / (1-t)$.

Lower values of λ are acceptable depending on the different configurations made available for the spam folder. If the configuration is set up to resend the email back to the sender asking him to send it to a private unfiltered email address of the recipient, then λ = 9 (t=0.9) seems to be reasonable. Even λ =1 (t=0.5) is acceptable if the recipient happens to go through every email in the bulk folder before manually deleting them. Two factors could be used in the context to measure the performance of a filter, namely, spam precision and spam recall. Let $n_{L\to S}$ and $n_{S\to L}$ be the numbers of L→S and S→L errors, and let $n_{L\to L}$ and $n_{S\to S}$ count the correctly treated legitimate and spam messages respectively. Spam recall (SR) and spam precision (SP) are defined as follows:

$$SR = \frac{n_{S\to S}}{n_{S\to L} + n_{S\to S}}$$

$$SP = \frac{n_{S\to S}}{n_{L\to S} + n_{S\to S}}$$

D. TOTAL COST RATIO

The evaluation factors that are frequently used in case of classification are accuracy (Acc) and the error rate (Err = 1 – Acc). Accuracy can be defined as the number of correct classifications, i.e. spam correctly classified as spam and legitimate messages as legitimate out of the total messages. The error rate is the ratio of the sum of false positives and false negatives out of the total messages.

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

$$Acc = \frac{n_{S\to S} + n_{L\to L}}{N_L + N_S} \quad Err = \frac{n_{L\to S} + n_{S\to L}}{N_L + N_S}$$

Where N_L = Number of Legitimate Messages.
N_S = Number of Spam messages.

V. EXPERIMENTAL RESULTS

A sample of 1000 spam and legitimate messages was collected from various sources. A simple JAVA application was used to fetch the messages using IMAP and transmit using SMTP Gateway (PORT 25) which delivers messages to the user's inbox. The comparisons of various Spam Mail Filters are given below:
Total No. of Mails Sent – legitimate = 1000
Total No. of Mails Sent – Spam = 1000

TABLE I : Comparison of Legitimate Sample Messages across Various Spam Filters

Classified as	Legitimate Sample Messages		
		Bogo Filter	Spam Assasin(Bayes Enabled)
Legitimate	986(98.6%)	976(97.6%)	1000(100%)
suspected spam	14(1.4%)	24(2.4%)	0(0.0%)
Spam	0(0.0%)	0(0%)	0(0.0%)
	Spam Sample Messages		
Legitimate	300(30%)	3(0.3%)	32(3.2%)
suspected spam	279(27.9%)	103(10.3%)	80(8%)
Spam	421(42.1%)	894(89.4%)	888(88.8%)

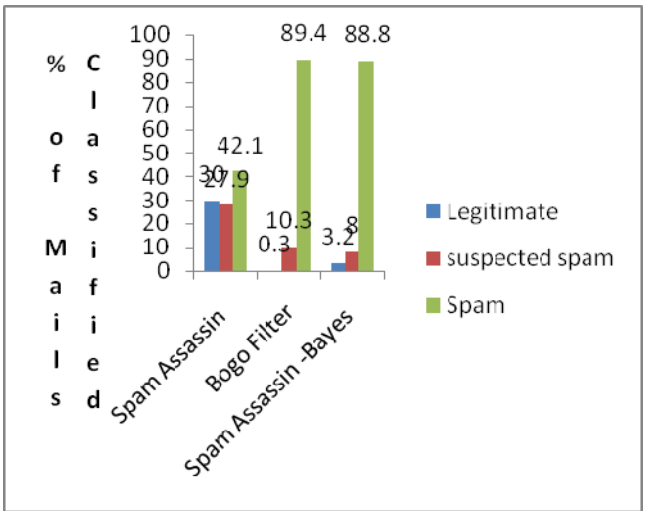

Fig 4 : Comparison of Spam Sample Messages Across Various Spam Filters.

VI. LEGITIMATE ATTACHMENTS

By attaching legitimate file formats such as Ms-office and PDF, spammers avoid contents analysis by spam filters and the message is passed which has got commercial advertisements with more spam keywords. On the other hand, it draws the user attention to open the attachment which usually does not have spam subject or content in order to pass the spam filter.

Fang et al [3] proposed a new method for extracting information from PDF files by parsing them to get text and format information and injects tags into text information to transform it into semi-structured text. To extract the contents from the Microsoft office documents Apache POI is used which creates and maintains Java APIs for manipulating various file formats based upon the Office Open XML standards (OOXML) and Microsoft's OLE 2 Compound Document format (OLE2). For each MS Office application there exists a component module that attempts to provide a common high level Java API to both OLE2 and OOXML document formats. This is most developed for Excel workbooks (SS=HSSF+XSSF). The extraction from the pdf files is done with a library called PDFBox. Apache PDFBox is an open source Java PDF library and allows creation of new PDF documents, manipulation of existing documents and the ability to extract

VII. CONFIDENTIAL DATA LEAKAGE

In digital forensic, there are numerous file formats in use. Criminals have started using either non-standard file formats or changing extensions of files while storing or transmitting them over a network. In an organization confidential data may be sent out in allowable different file type in scenarios of file type being changed by the malicious user.

A. CONTENT BASED FILE TYPE DETECTION - FILE HEADER/TRAILER ANALYSIS(FHT)

If the patterns are not easily identifiable, the file headers and file trailers can be analyzed and used to strengthen the recognition of many file types. The file headers and trailers are patterns of bytes that appear in a fixed location at the beginning and end of a file. If H is the number of file header bytes to analyze, and T is the number of trailer bytes to analyze, then two two dimensional arrays are built, one of dimensions H X 256 and the other of dimensions T x 256. For each byte position in the file header (trailer), all 256 byte values can be independently scored based upon the frequency with which the byte value occurs at the corresponding byte position.

It takes the input file or a fragment as input and analyses its contents to determine the file format. A file print is constructed using the header and trailer of various file formats. Given an input file, it is compared with existing file prints and a score generated for the input file with each file format. The format which gives the maximum score is the resultant format.
Training Phase:

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Various files are given for each file format and header/trailer scores are calculated and tabulated.

$$\overline{x_n} = \frac{n_0 x_{l-1} + x_l}{n_0 + 1}$$

- $\overline{x_n}$ = Updated average.
- x_0 = Old fingerprint Array Entry
- x_i = New Array Entry
- n_0 = previous number of files.

Identification Phase:
When an input file is given, the header array is extracted and scores are calculated with respect to scores of each format stored in the database. The format that gives the maximum score is the resultant format. Generate the score using the following equation.

$$S = \frac{\overline{x_1}.\overline{y_1} + \overline{x_2}.\overline{y_2} + \overline{x_3}.\overline{y_3} + \ldots\ldots + \overline{x_n}.\overline{y_n}}{\overline{x_1} + \overline{x_2} + \overline{x_3} + \cdots + \overline{x_n}}$$

X -> correlation strength for the byte value extracted from the input file for each byte position.
Y->correlation strength of the byte value in the fingerprint array with the highest correlation strength for the corresponding byte position.
Compare the unknown file with fingerprint and cross-correlation values and pick out the best match. It takes the input file or a fragment as input and analyses its contents to determine the file format.

B. MULTIPLE DISCRIMINANT ANALYSIS(MDA)
File type – The overall type of a file. This is often indicated by the application used to create or access the file.
Data type – Indicative of the type of data embedded in a file. Thus, a single file type will often incorporate multiple data types. Thus, when attempting to locate relevant files the goal becomes the location of relevant data types. Matrices are constructed with ASCII, low, entropy and correlation values computed for each file format. Discriminant analysis is performed and the results tabulated. Given an input file, the specified statistical measures are computed and a score is generated. This score on comparison with existing file formats gives the format the file belongs to.

- Statistical Measures Used:

AVERAGE
The average is taken by averaging the byte values for each window i and averaging the set of window averages. N denotes the number of bytes in the window.

$$AM = \frac{1}{n}\sum_{i=1}^{n} a_i$$

DISTRIBUTION OF AVERAGES
The probability that an average chosen from all the averages of a memory block is of value B in the range of 0-255. The goal with mapping the distribution of the statistics, i.e. measuring the probability of a statistical value occurring, is to provide a summary of the type of data in a file, providing an overview of the components of a file.

Dx = Pr ((B + 1) > Xj _>= B)

STANDARD DEVIATION
The standard deviation of the byte values of a window from the average for the window. This essentially identifies how chaotic elements values within a window are and how tightly knit the elements are to the median; i.e. are there many outliers in the window or are the values mostly consistent?

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2},$$

DISTRIBUTION OF STANDARD DEVIATION
The probability that a standard deviation chosen from all the standard deviations of a file is the value B.

Ds = Pr ((B + 1) > Sj >= B)

KURTOSIS
The 'peakedness' or consistency of the data calculated from two different modifications of the standard deviation, the numerator is the standard deviation squared with a fourth power instead of a square power and the denominator is the standard deviation squared.
DISTRIBUTION OF BYTE VALUES
The probability that a byte chosen from all the bytes in a window is the value of B, only unique values are used in the analysis. These statistical characteristics are then utilized in the algorithmic analysis of the digital data to uniquely identify data of each data type. In various cases, the other statistics mentioned in can be used to increase accuracy and differentiate between very similar data types.

Dx= Pr ((B + 1) > Xj >= B)

C. TEST CASE AND RESULTS
The proposed algorithm has been tested with the following algorithms and the results are in Table II.

TABLE II - Performance of FHT and MDA algorithms.

Test File type	Identified File type Case 1		Identified File type Case 2		Identified File type Case 3	
	FHT	MDA	FHT	MDA	FHT	MDA
DOC	DOC	DOC	DOC	DOC	RTF	DOC
RTF	RTF	RTF	RTF	RTF	RTF	RTF
TXT	TXT	TXT	TXT	TXT	TXT	TXT
PPT	PPT	PPT	PPT	PPT	PPT	PPT
GIF	GIF	GIF	JPG	GIF	GIF	JPG
JPG	JPG	JPG	JPG	JPG	GIF	JPG
EXE	EXE	EXE	EXE	EXE	EXE	EXE
ZIP	ZIP	ZIP	RAR	ZIP	RAR	ZIP
MP3	MP3	MP3	MP3	MP3	MP3	MP3
RAR	RAR	RAR	ZIP	RAR	RAR	RAR
HTML	HTML	HTML	HTML	HTML	EXE	HTML
PDF	PDF	PDF	PDF	PDF	DOC	PDF
LOG	DOC	LOG	LOG	LOG	LOG	LOG
JAVA	JAVA	JAVA	JAVA	JAVA	JAVA	JAVA
WAV	MP3	MP3	WAV	WAV	MP3	WAV

Total No. of Files: 60
No. of Files classified by FHT: 54(90% Accuracy)
No. of Files classified by MDA: 58(96.66% Accuracy)

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Header Scores per File Type:

The following table shows the sample header values generated for a particular file type "GIF" which has been compared with the various file types. The range of file sizes for all the file types are listed which shows the file type identification based on the Correlation frequency has no effect towards the identification of the file types.

TABLE III : FHT Scores for the sample File Type "GIF"

File type (GIF)	Score for Sample File Type 1 Chess.gif (3.47 KB)	Score for Sample File Type 2 Ha.gif (249 bytes)	Score for Sample File Type 3 Sun.gif (4.36 KB)
HTML(1.88 KB – 25.7 KB)	0.775	0.738	0.463
JPG(3.08 KB – 5.33 MB)	0.829	0.888	0.960
EXE(41.5 KB – 3.60 MB)	0.564	0.6	0.764
RTF(2.85 KB – 116 KB)	0.805	0.769	0.466
TXT(0KB –138 KB)	0.861	0.825	0.463
PDF (10KB – 4.82MB)	0.377	0.413	0.666
DOC (21.5 KB – 305 KB)	0.233	0.269	0.541
PPT(365 KB – 2.68 MB)	0.950	0.916	0.724
GIF(249 Bytes – 4.36 KB)	**1.0**	**0.941**	**1**
MP3(172 KB – 5.57 MB)	0.827	0.863	0.683
RAR(26.6 KB - 3.21 MB)	0.335	0.372	0.590
LOG(111bytes – 0.97MB)	0.825	0.819	0.463
JAVA(2.42 KB – 73 KB)	0.875	0.839	0.463
MP(463 KB – 700 KB)	0.75	0.715	0.463

VIII. CONCLUSION

Effective content analysis is the cornerstone of successful email monitoring and control operations. The success depends on the well defined and designed corporate email policies, and then detects them within the messages and attachments flowing through the mail network through effective and sophisticated content analysis capabilities. The proposed system deals with the applications of Web Content Mining towards the mitigation of In-bound and Out-bound E-Mail threats. The web content mining approach can also applied to detect and prevent other threats such as Embedding malicious code/E-mail Malware, DDos attacks due to Spam E-mails and illegal distribution of protected documents.

ACKNOWLEDGMENT

This work is supported by the NTRO, Government of India. NTRO provides the fund for collaborative project "Smart and Secure Environment" and this paper is modeled for this project. Authors would like to thanks the project coordinators and the NTRO members.

REFERENCES

[1] Androutsopoulos, J. Koutsias, V. Chandrinos, and D. Dpyropoulos, "An experimental comparison of naïve Bayesian and keyword-based anti-spam filtering with personal e-mail messages," In Proc. of the 23rd Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, 2000.

[2] Dhanalakshmi R,L.Kavisankar,C.Chellappan , "Enhanced Enhanced E-Mail Authentication Against spoofing Attacks To Mitigate Phishing, European Journal of Scientific research Vol . Issue 2011

[3] Fang Yuan, Bo Liu , A New Method Of Information Extraction from PDF Files , Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, 18-21 August 2005.

[4] Graham, P. (2002), A plan for spam. http://www.paulgraham.com/spam.html, 2003-11-13.

[5] Irfan Ahmed, Kyung-suk Lhee, Hyunjung Shin and ManPyo Hong, Fast File-type Identification, Proceedings of the 25th ACM Symposium on Applied Computing (ACM SAC 2010), ACM, Switzerland, March 2010.

[6] Karresand Martin, Shahmehri Nahid "File type identification of data fragments by their binary structure". In: Proceedings of the IEEE workshop on information assurance; 2006

[7] Lobato DH, Lobato JM (2008). Bayes Machines for binary classification. Pattern Recognition Letters. Elsevier, 29: 1466-1473.

[8] L. Wenyin, N. Fang, X. Quan, B. Qiu, and G. Liu, "Discovering Phishing Target based on Semantic Link Network," Future Generation Computer Systems, Elsevier, Volume 26, Issue 3, March 2010, pp. 381-388.

[9] Mason McDaniel and M. Hossain Heydari, "Content Based File Type Detection algorithms" ,IEEE Proceedings of the 36th Hawaii International Conference on System Sciences,2003

[10] Mehdi Chehel Amirani Mohsen Toorani Ali Asghar Beheshti Shirazi "A New Approach to Content-based File Type Detection" IEEE 2008.

[11] Meizhen W, Zhitang L, Sheng Z (2009). A Method for Spam Behavior Recognition Based on Fuzzy Decision Tree. IEEE, Ninth International Conference on Computer and Information Technology, pp. 236-241.

[12] Robert Erbacher ,John Muholland, "Identification and Localization of DataTypes within large scale file systems, Systematic Approaches to Digital Forensic Engineering ",IEEE 2007.

[13] Saeed Abu-Nimeh, Dario Nappa, Xinlei Wang, and Suku Nair, "A Comparison of Machine Learning Techniques for Phishing Detection" APWG eCrime Researchers Summit, October 4-5, 2007, Pittsburgh, PA, USA.

[14] Sahami, M., Dumais, S., Heckerman, D., and Horvitz, E. A Bayesian Approach to Filtering Junk EMail. In Learning for Text Categorization – Papers from the AAAI Workshop, pp. 55–62, Madison Wisconsin. AAAI Technical Report WS-98-05, 1998.

[15] Wei-Jen Li, Ke Wang, Salvatore J. Stolfo, Benjamin Herzog, "Fileprints: Identifying File Types by n-gram Analysis ", Proceedings of the IEEE Workshop on Information Assurance 2005.

[16] William C. Calhoun, Drue Coles "Predicting the types of file fragments ", Digital Forensic Research Workshop,Elsiever , Science Journal 2008

[17] Yang Y, Elfayoumy S (2007). Anti-Spam Filtering Using Neural Networks and Baysian Classifiers. Proceedings of the 2007 IEEE International Symposium on Computational Intelligence in Robotics and Automation, pp. 272-278