Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 11, November 2015, pg.251 – 255

REVIEW ARTICLE

Clustering Categorical Data for Internet Security Applications: A Review

Sapna V Ambadkar[1], Prof. S.P. Akarte[2]
[1]Department of Computer Sci. and Engg.
Prof Ram Meghe Institute of Technology & Research, Badnera, Amravati, India
[2]Department of Computer Sci. and Engg.
Prof Ram Meghe Institute of Technology & Research, Badnera, Amravati, India
[1] sapnaambadkar@gmail.com

Abstract: Malware and phishing website detection has been the Internet security topics that are now a day has great interests.

Compared with malware attacks, phishing website fraud is a relatively new Internet crime. Over the past few years, many clustering techniques have been employed for automatic malware and phishing website detection. In these techniques, the detection process is generally divided into two steps:

1) Feature extraction, where representative features are extracted to capture the characteristics of the file samples or the websites.

2) Categorization, where intelligent techniques are used to automatically group the file samples or websites into different classes based on computational analysis of the feature representations.

This paper presents expository survey of malware categorization and phishing website detection.

Keywords - malware categorization, phishing website detection.

INTRODUCTION

Clustering is a division of data into group of similar objects. The detection process is generally divided into two steps: Malware Categorization and Phishing Website Detection.

Malware Categorization:-

Malware such as virus, worms, Trojan Horses, spyware, backdoors, and root kits has presented a serious threat to the security of computer systems. Currently, the most significant line of defense against malware is Internet security software products, which mainly use a signature-based method to recognize threats in the Clients. Given a collection of malware samples, these venders first categorize the samples into families so that samples in the same family shares some common traits, and generates the common string(s) to detect variants of a family of malware samples.

Malware is the one of the major internet security threat. Currently, Antivirus (AV) software product is used for providing protrude signature profile for detecting malware. Modern malware is very complex and many variants of the same virus with different abilities appear every day which makes the detection process more difficult. For many years, malware categorizations have been done by human analysts such as looking up description libraries, and searching sample collections.

The manual analysis is time consuming and subjective for handling huge data. An automatic categorization system is required for making malware detection more efficient.
Phishing website detection:-

The word 'Phishing' initially emerged in 1990s. The early hackers often use 'ph' to replace 'f' to produce new words in the hacker's community. Phishing is a new word produced from 'fishing', it refers to the act that the attacker allure users to visit a faked Website by sending them faked e-mails or instant messages.
Compared with malware attack, phishing website fraud is a relatively new Internet crime. Phishing is a form of online fraud, whereby perpetrators adopt social engineering schemes by sending e-mails, instant messages, or online advertising to allure users to phishing websites that impersonate trustworthy websites in order to trick individuals into revealing their sensitive information such as financial accounts, passwords, and personal identification numbers. Including credit card number, bank account information, social security number and their personal credentials in order to use these details fraudulently against them. Phishing has a huge negative impact on organizations' revenues, customer relationships, marketing efforts and overall corporate image.
Phishing is a new type of network attack where the attacker creates a replica of an existing Web page to fool users by using specially designed e-mails or instant messages into submitting personal, financial, or password data to what they think is their service provides' Website which can then be used for profit. To defend against phishing websites, security software products generally use blacklisting to filter against known websites. There is always a delay between website reporting and blacklist updating. Indeed, as lifetimes of phishing websites are reduced to hours from days, this method might be ineffective. Malicious code (or malware) is defined as software that fulfills the harmful intent of an attacker. The damage caused by malware has dramatically increased in the past few years. One reason is the rising popularity of the Internet and the resulting increase in the number of available vulnerable machines because of security-unaware users.

LITERATURE SURVEY

Over the past few years, many research efforts have been conducted on developing clustering techniques for automatic malware categorization. M. Fredrikson, S. Jha, M. Christodorescu, R. Sailer, and X. Yan proposed to detects malware families by combing frequent subgraph mining and concept analysis to synthesize discriminative specifications in paper Synthesizing near-optimal malware specifications from suspicious behaviors.[2]

Menahem, A. Shabtai, L. Rokach, and Y. Elovici and Y. Ye, T. Li, Q. Jiang, Z. Han, and L.Wan says that combining different classification methods using different learning methods with possible different feature representations from malware detection.Classification methods require a large number of training samples to build the classification models.[3]

M. Bailey, J. Oberheide, J Andersen, Z. M. Mao, F. Jahanian, and J. Nazario says Efficiently group large datasets of malware samples into clusters used locality sensitive hashing and hierarchal clustering on Automated classification and analysis of internet malware.

T. Lee and J. J. Mody says in paper Behavioral classification adopted KM clustering approach to categorize the malware samples.[4]
M. Aburrous,M. A. Hossain, K. Dahal, and F. Thabtah, says in paper Predicting phishing websites using classification mining techniques with experimental case studies that Semantic attack targets the user rather than the computer is new significant security threat to the Internet in comparison with malware.[5]
M. Ester, H. P. Kriegel, J. Sander, and X. Xu propose in paper A density-based algorithms that for discovering clusters in large spatial databasewith noise, Clustering algorithm is to decide if there is a cluster in given webpage. If cluster is found, then webpage regarded as a phishing webpage. Otherwise it is identified as a legitimate webpage.[6]

In 2013 M.Madhuri1, K.Yeseswini2, U. Vidya Sagar3 says in paper Intelligent Phishing Website Detection And Prevention System By Using Link Guard ALGORITHM proposed a new end host based anti phishing algorithm usually called link guard algorithm utilizing the generic characteristics of the hyperlinks in phishing attacks. characteristics are derived by analyzing the phishing data archive provided by the Anti-Phishing Working Group (APWG).In this paper they verified that LinkGuard is effective to detect and prevent both known and unknown phishing attacks with minimal false negatives. They told with his experience that LinkGuard is light weighted and can detect and prevent phishing attacks in real time. They also told LinkGuard successfully detects 195 out of the 203 phishing attacks. They implemented Link Guard for Windows XP. They showed that LinkGuard is light-weighted and can detect upto 96% unknown phishing attacks in real-time. LinkGuard is useful for detecting phishing attacks, also can shield users from malicious or unsolicited links in Web pages and Instant messages.[7]

In 2014 B.Uma Maheswari Dr. P.Sumathi proposed in paper A New Clustering and Preprocessing for Web Log Mining A data preprocessing treatment system for web usage
mining has been analyzed and implemented for log data such as data cleaning, user identification, session identification and clustering. The Common log formats or Extended Log Formats only records the visitors browsing activities rather than the details of the visitor's identity. This means that different visitors sharing the same host cannot be differentiated. If there are proxy servers the problem became much severe. Users are identified easily by using Cookies or authentication mechanism. But users are not attracted by these types of sites due to privacy concerns. [8]

in 2011 R. Dhanalakshmi, C. Prabhu, C. Chellapan Detection Of Phishing Websites And Secure Transactions proposed the system for phishing website detection that is The client request the website for transaction process then this phishing detection system validate the website with four features such as WHOIS registration, IP address, domain and inter domain values of the corresponding requested website. The website related information's are stored in a WHOIS database for verification process. If website is phishing session key and send through the user's mobile. The phisher login into the original website means they can't get the session key so they enter a wrong session key then this system fferacdeny the access. The original user means the transaction is performed successfully.[9]

In 2012 Radha Damodaram,M.C.A,M.Phil, Dr.M.L.Valarmathi propose in paper Phishing website detection and optimization using Modified bat algorithm, presents an approach to overcome the difficulty and complexity in detecting and predicting phishing websites. Tested it against other heuristic algorithms, including Ant Colony optimization (ACO) and particle swarm optimization (PSO). There is a significant relation between the two phishing website criteria's (URL & Domain Identity) and (Security & Encryption) for identifying phishing website.[10]

In 2010 Maher Aburrous, M.A. Hossain, Keshav Dahal, Fadi Thabtah Intelligent phishing detection system for e-banking using fuzzy data mining proposed approach is to stop phishing. First approach is that the e-mail level (Adida, Hohenberge, & Rivest, 2005), current phishing attacks use broadcast e-mail (spam) to lure victims to a phishing website (Wu, Miller, & Garfinkel, 2006).Second approach is to use security toolbars. The phishing filter is a toolbar approach with more features such as blocking the user's activity with a detected phishing site. A third approach is to visually differentiate the phishing sites from the spoofed legitimate sites. Dynamic Security Skins (Dhamija & Tygar,2005) proposes to use a randomly generated visual hash to customize the browser window or web form elements to indicate the successfully authenticated sites.
A fourth approach is two-factor authentication, which ensures that the user not only knows a secret but also presents a security token. Many industrial anti-phishing products use toolbars in web browsers, but some researchers have shown that security tool bars do not effectively prevent phishing attacks. [11]

In 2005 Dhamija and Tygar proposed a scheme that utilises a cryptographic identity-verification method that lets remote web servers prove their identities.

In 2006 Liu, Deng, Huang, & Fu proposed a tool to model and describe phishing by visualizing and quantifying a given site's threat, but this method still would not provide an anti-phishing solution. The visual similarity between two web pages is then evaluated in three metrics: block level similarity, layout similarity, and overall style similarity, which are based on the matching of the salient block regions. The training and classification experiments have proven that it is possible to improve the categorization process.

S.Sugantha,C. Ramasamy proposed in paper Link based Cluster Ensemble Framework - Clustering Categorical Data for Internet Security Applications that a categorization system for profiling signatures to improve the anomaly detection process more efficiently. A categorization system that uses a link based cluster ensemble for automatically categorizing security threats. Cluster ensemble aggregates different clustering algorithms producing different solutions for grouping malware samples and phishing websites. For detection ,the method first finds the associated webpages with the given page, then mines the features such as links relationship, ranking relationship, webpage text similarity, and webpage layout similarity relationship between the given webpage and its associated webpages, and finally applies DBSCAN (Density Based Spatial Clustering of Applications with Noise) clustering algorithm to decide if there is a cluster around the given webpage. If such cluster is found, the given webpage is then regarded as a phishing webpage; otherwise, it is identified as a legitimate webpage. Existing clustering methods usually apply a specific clustering method on a feature representation. Different clustering methods have their own advantages and limitations in malware detection. In our study, we use a link based cluster ensemble to aggregate the clustering solutions that are generated by both hierarchical and partitional clustering methods. Our ensemble framework is also able to incorporate the domain knowledge in the form of sample level constraints.[12]

In 2013 C.M.Geetha1, K.Sangeetha 2 ,Dr S.Karthik3 proposed in paper A Survey On Enhanced Approach For Categorical Link Based Clustering that Cluster ensembles are used as best alternative to the standard cluster analysis. The data set has been clustered by using any of the well known cluster algorithm and represented as a cluster ensemble. The cluster ensembles generate a final data partition based on incomplete information and the information is not prefect to make use of it. Existing cluster ensemble methods to categorical data analysis rely on the typical pairwise-similarity and binary cluster-association matrices, which summarize the underlying ensemble information at a rather coarse level.

The clustering result is improved by applying ranking to the cluster ensembles by finding similarity between the cluster data points.[13]

In 2008 Xun Dong, John A. Clark, Jeremy L. Jacob proposed in paper User Behaviour Based Phishing Websites Detection that a novel approach to detect phishing websites based on analysis of users' online behaviours that is the websites users have visited, and the data users have submitted to those websites. Such user behaviours cannot be manipulated freely by attackers, detection based on those data can not only achieve high but also is fundamentally resilient against changing deception methods.

Phishing website filters in Internet Explorer browsing in Firefox, and Netcraft toolbar are all blacklist anti-phishing websites detection systems. They check whether the URL of the current web page matches any identified phishing web sites before rendering the webpage to users. It has been designed and implemented to be hard to circumvent, and have discussed its unique strength in protecting users from phishing threats. [14]

In 2010-2011 Mona Ghotaish Alkhozae, Omar Abdullah Batarfi proposed in paper Phishing Websites Detection based on Phishing Characteristics in the Webpage Source Code that a phishing detection approach based on checking the webpage source code, extract some phishing characteristics out of the W3C standards to evaluate the security of the websites, and check each character in the webpage source code, if we find a phishing character, then decrease from the initial secure weight. Finally calculate the security percentage based on the final weight, the high percentage indicates secure website and others indicates the website is most likely to be a phishing website. Check two webpage source codes for legitimate and phishing websites and compare the security percentages between them, find the phishing website is less security percentage than the legitimate website. Detect the phishing website based on checking phishing characteristics in the webpage source code.[15]

METHODOLOGY

A cluster is a collection of phishing websites or malicious files that share some common traits between them and are "dissimilar" to the phishing websites or malware samples belonging to other clusters. The clustering algorithms are used to classify & categorized the given samples.

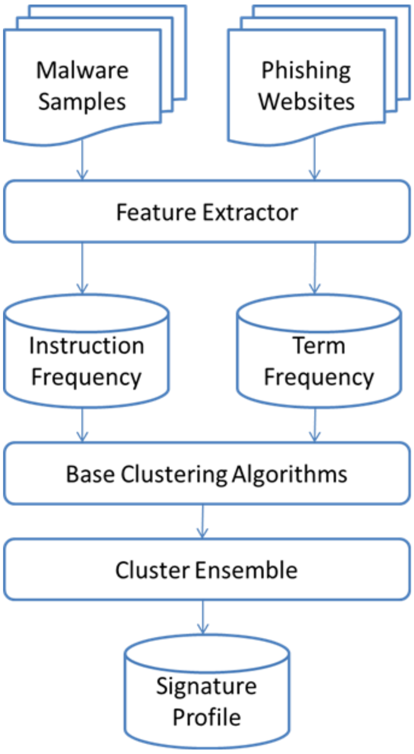

Fig.1 clustering algorithms

1) Term-frequency feature extractor:

For phishing website categorization, the ACS(Automatic categorization system) first uses the term-frequency feature extractor to extract the terms from the web pages of the collected phishing websites, and then transforms the data into term-frequency feature vectors. These vectors are stored in the database. The transaction data can also be easily converted to relational data if necessary.

2) Instruction-frequency feature extractor:

For malware categorization, the ACS first uses the instruction-frequency feature extractor to extract the function-based instructions from the collected Portable Executable (PE) malware samples, converts the instructions to a group of 32-bit global IDs as the features of the data collection, and stores these features in the signature database. These integer vectors are then transformed to instruction frequencies and stored in the database. The transaction data can also be easily converted to relational data if necessary.

3) Base clustering algorithms:

Base clustering solutions are generated by applying different clustering algorithms that are based on the feature representations. The HC algorithm and KM partitioned approach are applied on the Term-frequency vectors or instruction-frequency vectors with the TF-IDF and TF weighting schemes, which are widely used for document representation in IR (information retrieval.

4) Cluster ensemble with constraints:

Cluster ensemble is used to combine different base clustering. The cluster ensemble is also able to utilize the domain knowledge in the form of website-level/sample-level constraints.

CONCLUSION

Phishing website and malware categorisation are the big issues in internet security. The identification phishing website and categorizing malware samples are the challenging task in internet security threads. Phishing website detection and malware categorisation detection on internet security threads have been extensively studied. As well as the methodology of model also studied.

REFERENCES

[1] WeiweiZhuang, Yanfang Ye, Yong Chen, and Tao Li,” *Ensemble Clustering for Internet Security Applic*” *Ieee Transactions On Systems, Man, And Cybernetics—Part C: Applications And Reviews, Vol. 42, No. 6, November 2012*

[2] M. Fredrikson, S. Jha, M. Christodorescu, R. Sailer, and X. Yan, “Synthesizing near-optimal malware specifications from suspicious behaviors,” in Proc. IEEE Symp. Secur. Priv., Washington, DC IEEE Computer Society, May 2010, pp. 45–60.

[3] E. Menahem, A. Shabtai, L. Rokach, and Y. Elovici, “*Improving malware detection by applying multi-inducer ensemble,*” J. Comput. Stat. Data Anal., vol. 53, no. 4, pp. 1483–1494, Feb. 2009.

[4] T. Lee and J. J. Mody, “*Behavioral classification*,” in Proc. EICAR, May 2006.

[5] M. Aburrous,M. A. Hossain, K. Dahal, and F. Thabtah, *“Predicting phishingwebsites using classificationmining techniqueswith experimental case studies*,” in Proc. 7th Int. Conf. Inf. Technol., 2010, pp. 176–181.

[6] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databasewith noise,” in Proc. ACM Int. Conf. Knowl. Discovery Data Mining, 1996, pp. 226–231.

[7] M.MADHURI1, K.YESESWINI2, U. VIDYA SAGAR3,” *INTELLIGENT PHISHING WEBSITE DETECTION AND PREVENTION SYSTEM BY USING LINK GUARD ALGORITHM*” in proc. International Journal of Communication Network Security, ISSN: 2231 – 1882, Volume-2, Issue-2, 2013

[8] Uma Maheswari,Dr. P.Sumathi,” A NewClustering and Preprocessing for Web Log Mining” in proc 2014 World Congress on Computing and Communication Technologies.

[9] R. Dhanalakshmi, C. Prabhu, C. Chellapan,” Detection Of Phishing Websites And Secure Transactions”in proc Detection Of Phishing Websites And Secure Transactions

[10]Radha Damodaram,M.C.A,M.Phil.*, Dr.M.L.Valarmathi,” Phishing website detection and optimization using Modified bat algorithm” in proc Vol. 2, Issue 1, Jan-Feb 2012, pp. 870-876

[11] Maher Aburrous, M.A. Hossain, Keshav Dahal, Fadi Thabtah,” Intelligent phishing detection system for e-banking using fuzzy data mining”in proc journal homepage: www.elsevier.com/locate/eswa

[12] S.Sugantha,C. Ramasamy,”Link based Cluster Ensemble Framework - Clustering Categorical Data for Internet Security Applications”

[13] C.M.Geetha1, K.Sangeetha 2 ,Dr S.Karthik3,” *A SURVEY ON ENHANCED APPROACH FOR CATEGORICAL LINK BASED CLUSTERING*”,in proc IJARCET Volume 2, Issue 1, January 2013.

[14] Xun Dong, John A. Clark, Jeremy L. Jacob,” *User Behaviour Based Phishing Websites Detection*”,in proc 978-83-60810-14-9/08/$25.00 c 2008 IEEE

[15] Mona Ghotaish Alkhozae, Omar Abdullah Batarfi,” *Phishing Websites Detection based on Phishing Characteristics in the Webpage Source Code*”, in proc Volume 1 No. 6, October 2011 ISSN-2223-4985