

Clustering Categorical Data for Internet Security Applications

Sapna V. Ambadkar¹, S. P. Akarte²

^{1,2}Department of Computer Sci. and Engg., Prof Ram Meghe Institute of Technology & Research, Badnera, Amravati, India-444701

Email address: ¹sapnaambadkar@gmail.com

Abstract—Clustering categorization data for internet security application developed a system for malware categorization, phishing website detection and secure E-mails using keywords. These are now a day has great interest. Phishing website detection has been new internet crime than malware categorization. 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.

In this approach can detect malware categorization, phishing website detection as well as detect spam mails in one system.

Keywords— Malware categorization; phishing website detection; feature extraction; categorization.

I. INTRODUCTION

Clustering is a division of data into group of similar objects. The detection process is generally divided into steps: Malware Categorization, Phishing Website Detection, secure spam mails using keyword.

Malware Categorization

Malware is the one of the major internet security threat. 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. Malware such as virus, worms, Trojan Horses, spyware, backdoors, and root kits has presented a serious threat to the security of computer systems. Given a collection of malware samples, these vendors 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.

Phishing Website Detection

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 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' Web site 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.

Secure E-mails

Secure emails from various or spam emails are also a security issue. For securing emails use keyword like spam and it automatically detect whether it is a spam mail or not and it is a spam mail then it automatically goes to spam folder, not spam mail then it is go to inbox.

II. RELATED WORK

S. Arun, D. Anandan, T. Selvaprabhu, B. Sivakumar, P. Revathi, H. Shine [1] proposes a proactive method to shut down a Phisher's operation by using a Pguard. This effectively stops a phishing attack at its source thereby protecting a significant number of other innocent users from being duped in the future. Most phishing attacks begin with spam. Spam is mass unsolicited email. The email message typically contains some sort of socially engineered message enticing the recipient to venture to a web site or to reply to the message. proposed techniques to prevent phishing attacks, Phishers are becoming increasingly sophisticated in their approaches. Phishing attacks often involve rigorous planning and incorporate strategies to bypass existing anti-phishing tools. This technique does not prevent an initial phishing email from being sent, once the phishing page has been removed, all future victims are essentially protected from the phisher. Experimental results show that this approach can be an

effective way to remove phishing pages hosted on servers around the world.

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 detect malware families by combining frequent subgraph mining and concept analysis to synthesize discriminative specifications in paper Synthesizing near-optimal malware specifications from suspicious behaviors [2].

Christodorescu, M., Jha, S., Seshia, S.A., Song, D. and R.E.Bryant proposed in paper Semantics-aware malware detection that semantics-aware malware detection that attempts to detect polymorphic malware by identifying semantically-equivalent instruction sequences in the malware categories. This work, described the malicious behavior, for example decryption loop with a template of instruction sequences, where a matching algorithm is applied on the disassembled binaries to find the instruction sequences that match the predefined malicious template [3].

Xuxian Jiang, proposed Stealthy Malware Detection Through VMM-Based “Out-of-the-Box” Semantic View Reconstruction where they presented VM watcher, a novel VMM based approach that enables out-of-the-box malware detection by addressing the semantic gap challenge. Identified the category of malware, malware classification algorithms, malwares activities and ways of preventing and removing malware if it eventually infects system [4].

Anugrah Kumar, Sarvesh SS Rawat proposes an approach towards Phishing Detection Using Rough Set Theory. For phishing thirteen basic factors directly responsible, they are grouped into four Strata. Reliability Factor is determined on the basis of the outcome of these strata, using Rough Set Theory. Reliability Factor determines the possibility of a suspected site to be valid or fake. Using Rough set Theory most and the least influential factors towards Phishing are determined.

A rough set based algorithm to determine a reliability factor of a given website. Reliability factor indicates the authenticity of the website. The paper proposes Rough Set theory for better analysis of the vagueness in parameters involved in Phishing detection in real time and thereby producing more accurate results for Phishing detection [5].

To protect the users against phishing attacks, various anti-phishing techniques have been proposed that follows different strategies like client side and server side protection.

Heng Yin, Dawn Song propose a novel approach for the detection and analysis of privacy-breaching malware. They observed that numerous malware categories, including spyware, key-loggers, network sniffers, stealth backdoors, and rootkits, share similar fundamental characteristics, which lies in their malicious or suspicious information access and processing behavior. Propose to use whole system, fine-grained taint tracking. The approach works by marking the sensitive information introduced in the tests as tainted and monitoring taint propagation over the whole system. Monitor the taint propagation at the hardware level. To perform

meaningful analysis, we also need a mechanism to extract operating-system level information. Proposed whole-system fine-grained taint analysis to discern fine-grained information access and processing behavior of a piece of unknown code [6].

Dynamic Security Skins proposes to use a randomly generated visual hash to customize the browser window or web form elements to indicate the successfully authenticated sites [7].

Wei Peng, Xukai Zou propose a general behavioral characterization of proximity malware which based on naive Bayesian model, which has been successfully applied in non-DTN settings such as filtering email spams and detecting botnets. And also proposed propose two extensions to look ahead, dogmatic filtering, and adaptive look ahead, to address the challenge of “malicious nodes sharing false evidence.” Real mobile network traces are used to verify the effectiveness of the proposed methods [8].

M. Archana, P. M. Durai Raj Vincent, Naveen Kumar Boggavarapu proposed anti phishing tool to detect the phishing websites is very helpful to save the users from many fraud websites. The wifi internet connection is used to deliver the internet service to the mobile device. The paper states that the phishing websites mostly get the e-banking sites and attack their passwords, credit card number, bank account and personal details of the user [9].

Sadia Afroz, Rachel Greenstadt proposes a phishing detection approach—PhishZoo—that uses profiles of trusted websites’ appearances to detect phishing. It includes a performance analysis and a framework for making use of computer vision techniques in a practical way investigate a new approach for phishing detection based on profiling the content of phished websites to determine when a user is being deceived by a false belief.

A detailed review has been conducted on the current situation of malware infection and the work done to improve anti-malware or malware detection systems. Thus, it provides an up-to-date comparative reference for developers of malware detection systems.

Dai et al. (2009) proposed a malware detection system, based on a virtual machine, to reveal and capture the needed features. The system constructs classification models using common data mining approaches. Yu et al. (2011) presented a simple method to detect malware variants. First, a histogram is created by iterating over the suspected file binary code. An additional histogram is created for the base sample (the known malware). Then, measures are calculated to estimate the similarity between the two histograms. Yu et al. (2011) showed that when the similarity is high, there is a high probability that the suspected file is a malware variant.

III. PROPOSED SYSTEM

Propose an approach to automatic identification of the phishing target of a given webpage by clustering the webpage set consisting of all its associated web pages and the given webpage itself. The associated web pages are those which are pointed by forward links of the given webpage and web pages

returned by a powerful search engine with certain representative keywords (e.g. brand, title and keywords of content) in the given webpage as queries. This approach first finds its associated web pages and 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 web pages. 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 and categorized the given samples.

In previous system malware detection, phishing website detection, securing emails these are the separate system for detection each security attack but in this system all are detect in one system.

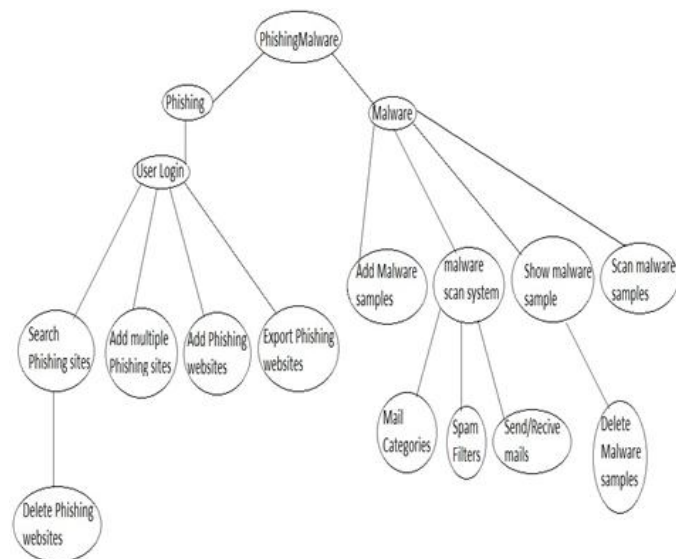


Fig. 1.

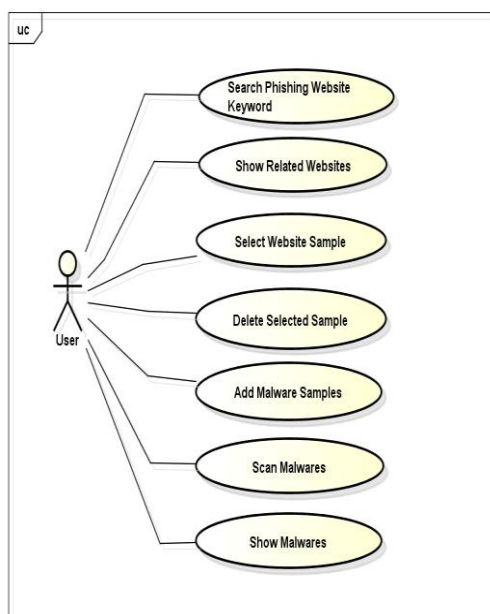


Fig. 2.

K-Medoids Clustering Approach

Clustering algorithm for categorization is squared error-based partitioning clustering, such as K-means assigns a set of data points into clusters using an iterative relocation technique. A cluster is represented by one of its real data point called medoids or by the mean of its data points called centroid in KM and K-means methods, respectively. They are very simple, but effective and widely used in many scientific and industrial applications. Considering that the distributions of phishing websites and malware samples are typically skewed, irregular, and of densities, in order to well deal with the outlier problem, we use KM instead of K-means for categorization. The algorithm procedure for KM is described in Algorithm. Clustering algorithms are valuable tools for malware categorization.

IV. RESULT

In this project developed three applications for internet security applications. Our first task is phishing website detection system for this write any phishing site into white bar go on it. If entered site is a phishing site then it shows a message ‘This is a phishing site’ and entered site is not a phishing site then open webpage for entered site, when data connection on the system.

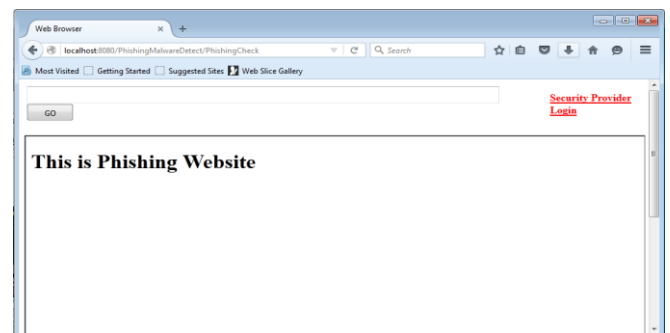


Fig. 3.

Using security provider login provides us additional facilities like search phishing site, add phishing site, delete phishing site, upload multiple phishing sites and malware scan system.

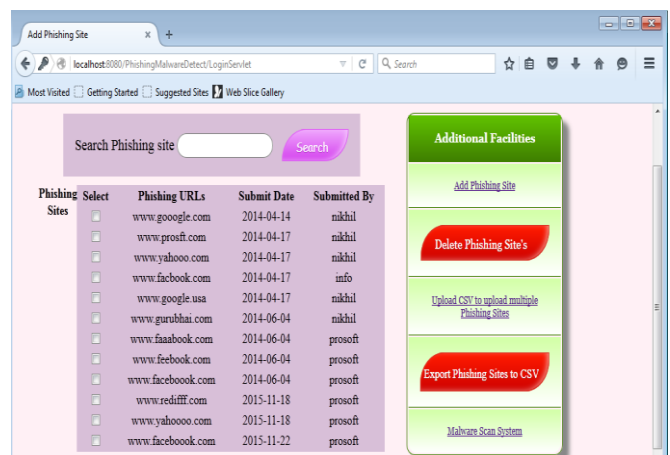
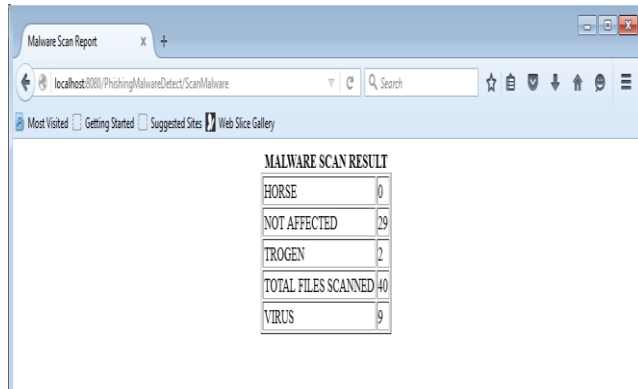


Fig. 4.

Now next part is scan malware detection system, here we provide path for malware sample file. After scanning result of scan system shown.



MALWARE SCAN RESULT	
HORSE	0
NOT AFFECTED	29
TROGEN	2
TOTAL FILES SCANNED	40
VIRUS	9

Fig. 5.

In malware scan system we can add malware virus definition, also can delete this definition and the final part of our project is secure e-mails from spam mails using keyword. . For securing e-mail separate Email login is there. Recognizing spam mail using spam keyword, and it recognise mail is spam mail then it automatically goes to spam folder. If not a spam mail then it automatically goes inbox.

V. CONCLUSION

This project proposed the identification phishing website and categorizing malware samples and secure emails are the challenging task in internet security threads. In the proposed system Clustering algorithm for categorization is used, for clustering categorization, phishing website categorization, but also for categorizing malware samples into families that takes

input as fish tank database and malware categories, detect phishing site as well as detect the malware, identify it, show which type of virus and how many times it occurs and then remove it. Shows some result for it. For securing E-mails use keyword, using keyword automatically checks whether it is spam mail or not.

REFERENCES

- [1] S. Arun, D. Anandan, T. Selvaprabhu, B. Sivakumar, P. Revathi, and H. Shine, "Detecting phishing attacks in purchasing process through proactive approach," *Advanced Computing: An International Journal (ACIJ)*, vol. 3, no. 3, 2012.
- [2] M. Fredrikson, S. Jha, M. Christodorescu, R. Sailer, and X. Yan, "Synthesizing near-optimal malware specifications from suspicious behaviors," in *Proceeding IEEE Symp. Secur. Priv., Washington, DC IEEE Computer Society*, pp. 45–60, 2010.
- [3] M. Christodorescu, S. Jha, S.A. Seshia, D. Song, and R. E. Bryant, "Semantics-aware malware detection," in *Proceedings of the IEEE Symposium on Security and Privacy*, pp. 32–46, 2005.
- [4] A. O. Surajudeen, M. A. Mabayoje, A. Mishra, and O. Oluwafemi, "Malware detection, supportive software agents and its classification schemes," in *Proceeding International Journal of Network Security & Its Applications (IJNSA)*, vol. 4, no. 6, pp. 33–49, 2012.
- [5] A. Kumar, S. S. Roy, S. S. S. Rawat, and S. Saxena, "Phishing Detection by determining Reliability Factor using Rough Set Theory," 2013.
- [6] H. Yin and D. Song, "Panorama: capturing system-wide information flow for malware detection and analysis," *Proceedings of the 14th ACM Conference on Computer and Communications Security*, pp. 116–127, 2007.
- [7] R. Dhamija and J. D. Tygar, "The battle against phishing: dynamic security skins," *Proceedings of the 2005 Symposium on Usable Privacy and Security*, pp. 77–88, 2005.
- [8] Wei Peng and Xukai Zou, "Behavioral Malware Detection in Delay Tolerant Networks," in *Proceeding IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 1, 2014.
- [9] M. Archana, P. M. D. R. Vincent, and N. K. Boggavarapu, "Architecture for the detection of phishing in mobile internet," *International Journal of Computer Science and Information Technologies (IJCSIT)*, vol. 2, no. 3, pp. 1297–1299, 2011.