Detection and Prevention of Phishing Attacks

Abu Saad Choudhary
Department of Information Technology
Shree L R Tiwari College of Engineering
Thane-401107, India
chaudharyabusaad @gmail.com

Rucha Desai
Department of Information Technology
Shree L R Tiwari College of Engineering
Thane-401107, India
rucharockx@gmail.com

Lavkush Gupta
Department of Information Technology
Shree L R Tiwari College of Engineering
Thane-401107,India
Lavkushgupta9172@gmail.com

Madhuri Gedam
Department of Information Technology
Shree L R Tiwari College of Engineering
Thane-401107, India
madhuri.gedam@gmail.com

***Abstract*- Phishing is one amongst the main issuesvisaged by cyber-world and ends up in monetarylosses for each industries and people. Detection ofphishing attack with high accuracy has forever been a difficult issue. At present, visual similarities-based techniques square measure terribly helpful for police work phishing websites expeditiously. Phishing web site appearance terribly similar in look to its corresponding legitimate web site to deceive users into basic cognitive process that they are browsing the right web site. Visual similarity primarily based phishing detection techniques utilize the feature set like text content, text format, HTML tags, Cascading sheet (CSS), image, then forth, to form the choice. These approaches compare the suspicious web site with the corresponding legitimate web site by victimisation numerous options and if the similarity is larger than the predefined threshold price then it is declaredphishing [2].**

Keywords— Phishing Attack; URL; Real Time Model; Phishing Detection

I. INTRODUCTION

Phishing could be a crime within which a wrongdoer sends the faux e-mail, that seems to return from widespread and trusty complete or organization, asking to input personal certification like bank positive identification, username, number, address, master card details, so forth. The faux e-mails usually look astonishingly legitimate, and even the web site wherever the net user is asked to input personal data additionally sounds like legitimate one. Phishing messages propagate over e-mail, SMS, instant messengers, social networking sites, VoIP, so forth, however e-mail is that the widespread thanks to perform this attack and phishing attack is achieved by visiting the link hooked up to the e-mail. Moreover, spear phishing attack is changing into widespread these days. Business e-mail compromise (BEC) is discovered as a serious net threat in 2015.In BEC, the persona non grata uses spear phishing ways to fool organizations and net persons [1]. More subtle spear phishing attacks targeted individual or teams inside the organization. Phishing is metaphorically like fishing within the water, however rather than attempting to catch a fish, attackers attempt to steal consumer's personal data. once a user opens a faux webpage and enters the username and guarded positive identification, the credentials of the user area unit noninheritable by the aggressor which may be Phishing websites look terribly similar in look to their corresponding legitimate websites to draw in sizable amount of net users. Recent developments in phishing detection have junction rectifier to the expansion of diverse new visual similarities- based approaches. Visual similarity-based approaches compare the visual look of the suspicious web site to its corresponding legitimate web site by exploitation numerous parameters [1].

II. RELATED WORK

A. Protecting user against phishing using Antiphishing: -

AntiPhishing is employed to avoid users from exploitation fallacious websites that successively could cause phishing attack.Here, AntiPhishing traces the sensitive data to be stuffed by the user and alerts the user whenever he/she is trying to share his/her data to a untrusted computing machine.The abundant effective elucidation for this can be cultivating the users to approach just for trusty websites [2]

B. Learning to Detect Phishing Emails: -

An alternative for police investigation these attacks could be a relevant method of reliableness of machine on a attribute supposed for the reflection of the enclosed deception of user by. This approach is utilized in the detection of phishing websites, or the text messages sent through emails that area unit used for stable gear the victims [3].

C. Phishing detection system for e-bankingusing fuzzy data mining: -

Phishing websites, primarily used for e-banking services, area unit terribly advanced and dynamic to be known and classified.because of the involvement of varied ambiguities within the detection, sure crucial data processing techniques could prove a good means that to keep the e-commerce websites safe since itdeals. with considering numerous quality factors instead of precise values [4].

D. Collaborative Detection of Fast Flux Phishing Dom

Here, 2 approaches area unit outlined to search out correlation of evidences from multiple servers of DNS and multiple suspects of FF domain.real world examples is wont to prove that our correlation approaches expedite the detection of the FF domain, that area unit supported Associate in Nursing analytical model which mayquantify numerous DNS queries that area unit needed to verify a FF domain [5].

E. A Prior-based Transfer Learning Method for the Phishing Detection: -

A supplying regression is that the root of a priority primarily based transferrable learning technique, that is conferred here for our classifier of applied mathematics machine learning.it's used for the detection of the phishing websites counting on our elect characteristics of the URLs.because of the divergence within the allocation of the options within the distinct phishing areas, multiple model's area unit projected for various regions [6].

III. PROPOSED SYSTEM

Phishing has been a major security threat in which there is a huge loss for companies as well as customers. These phishing attacks are increasing day by day due to lack of efficient detection techniques and effective preventive measures. A comprehensive efficient detection technique should be developed in order to detect and inform the web users about the phishing attacks to make sure that their sensitive data will not be disclosed during these attacks. There are a unit varied techniques exists for detection of phishing, however it's still become a difficult work to note faux websites with the prevailing methodology. There is a unit numerous technique obtainable like blacklisting, white listing, heuristics and machine learning to observe phishing, however machine learning is being extensively used.to forestall this, data processing techniques is projected during this analysis work to spot the phishing website and alerting users from revealing their passwords.

IV. FLOWCHART

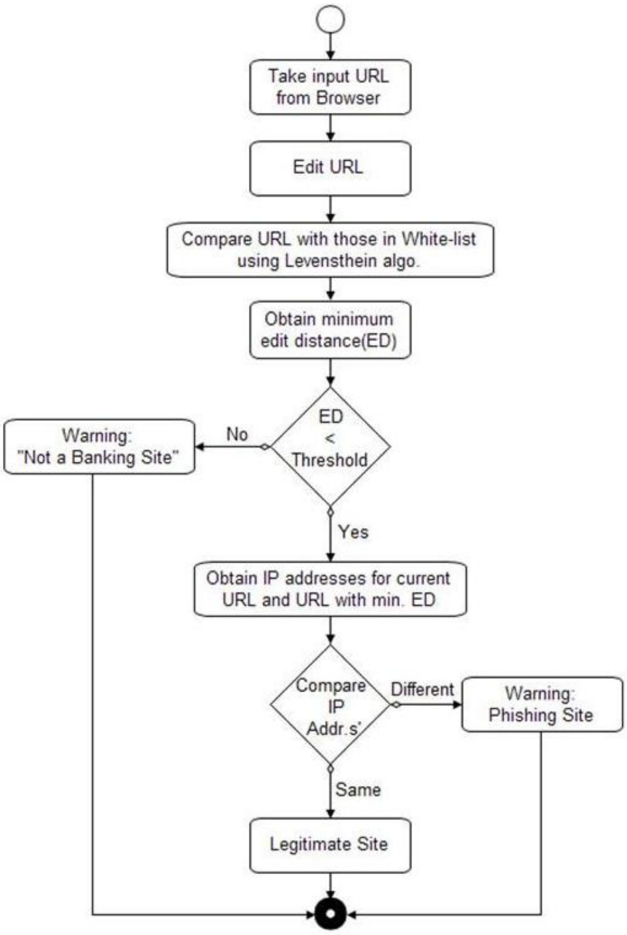
Take input URL from Browser
Edit URL
Compare URL with those in White-list using Levensthein algo.
Obtain minimum edit distance(ED)
ED < Threshold
No
Warning: "Not a Banking Site"
Yes
Obtain IP addresses for current URL and URL with min. ED
Compare IP Addr.s'
Different
Warning: Phishing Site
Same
Legitimate Site

Fig. 1. Flowchart

Phishing is one of the major problems faced by cyber-world and leads to financial losses for both industries and individuals. Detection of phishing attack is always a challenging issue. Phishing website looks very similar in appearance to its corresponding legitimate website to deceive users into believing that they are browsing the correct website. As the phishing sites uses the host name, that is incredibly concerning the legitimate website, the edit distance worth can clearly be low. So, the information processing address of the entered website is compared with the information processing address of the positioning within the white list that encountered the minimum edit distance. If each the addresses area unit same, then it's a legitimate. During this method the user is alerted [2].

V. MACHINE LEARNING IMPLEMENTATION

The following algorithms were chosen based on their performance on classification problems.

A. Random Forests

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. Random forest is a supervised learning algorithm. The "forest" it builds, is an ensemble of decision trees, usually trained with the "bagging" method. The general idea of the bagging method is that a combination of learning models increases the overall result. Random forest has nearly the same hyperparameters as a decision tree or a bagging classifier. Fortunately, there's no need to combine a decision tree with a bagging classifier because you can easily use the classifier-class of random forest. With random forest, you can also deal with regression tasks by using the algorithm's regressor. Random forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model. Therefore, in random forest, only a random subset of the features is taken into consideration by the algorithm for splitting a node. You can even make trees more random by additionally using random thresholds for each feature rather than searching for the best possible thresholds

B. Neural Networks

A neural network is structured as a set of interconnected identical units (neurons). The interconnections are used to send signals from one neuron to the other. In addition, the interconnections have weights to enhance the delivery among neurons. The neurons are not powerful by themselves, however, when connected to others they can perform complex computations. Neural networks are a set of algorithms, that are designed to recognize patterns. They interpret sensory data through a kind of machine perception, labelling or clustering raw input. The patterns they recognize are numerical, contained in vectors, into which all real-world data, be it images, sound, text or time series, must be translated. Neural networks help us cluster and classify. Machine learning algorithms that use neural networks generally do not need to be programmed with specific rules that define what to expect from the input. The neural networks learning algorithm instead learns from processing many labeled examples (i.e., data with "answers") that are supplied during training and using this answer key to learn what characteristics of the input are needed to construct the correct output.

C. Support Vector Machines

Support Vector Machine (SVM) is a supervised machine learning discriminative model, which conforms to the principle of drawing separating hyper-plane with maximum safety space, called margin, to minimize the risk of flawed predictions. Support Vector Machine” (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. Support vector machine is highly preferred by many as it produces significant accuracy with less computation power. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

D. Logistics Regression

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression is another technique borrowed by machine learning from the field of statistics. It is the go-to method for binary classification problems (problems with two class values). Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems. Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets. Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.

VI. Block Diagram

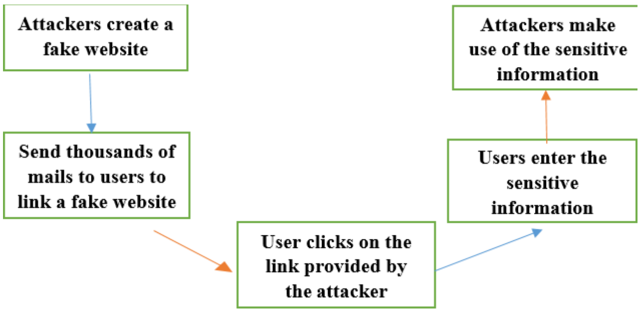
Attackers create a fake website
Send thousands of mails to users to link a fake website
User clicks on the link provided by the attacker
Users enter the sensitive information
Attackers make use of the sensitive information

Fig. 2. Block Diagram

A. Creating a fake website:

As part of phishing attack, attackers create a fake website which appears similar to original website. They use the main features of the original website such as logo, design of a website to create a fake website so that users cannot suspect such fake websites.

B. Linking a fake website through email:

Once creation of the fake website is done,attackers send thousands of e-mails to multiple users and make email recipients(users) to clicka URL which re-directs to the fake website.

C. Clicking a malicious URL:

The users who were not aware of the malicious URL provided in the email, clicks it which directs to the fake website provided by the attackers. This is where the phishing attack begins.

D. Entering sensitive information:

Once the user is redirected to the fake website, the sensitive information such as login credentials and other details are entered by theuser in order to access the website created by the attacker.

E. Compiling the stolen data and using it:

Once the user enters the sensitive information, all the sensitive data is collected so that the attacker can sell the data or use it for his/her own purpose [4].

VII. ADVANTAGES

1. It does not depend on the phishing technique.
2. It can detect pharming attacks, which are undetectable by manyexisting systems.
3. Some system tries to detectphishing webpage. Some system detects phishing webpage when user opens new webpage [7].

a) Hardware requirements: -

- 4 GB RAM
- 10GB HDD
- Intel 1.66 GHz Processor Pentium

b) Software requirements: -

- Windows 7
- Python 3.6.0
- Visual Studio Code

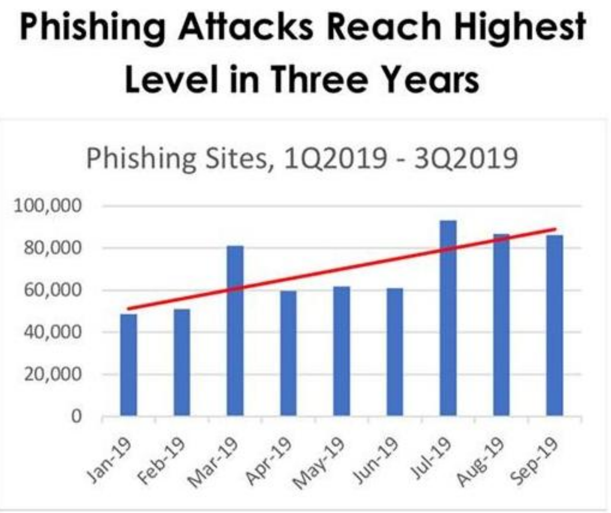

Fig. 3. Emergence of new malicious attacks.

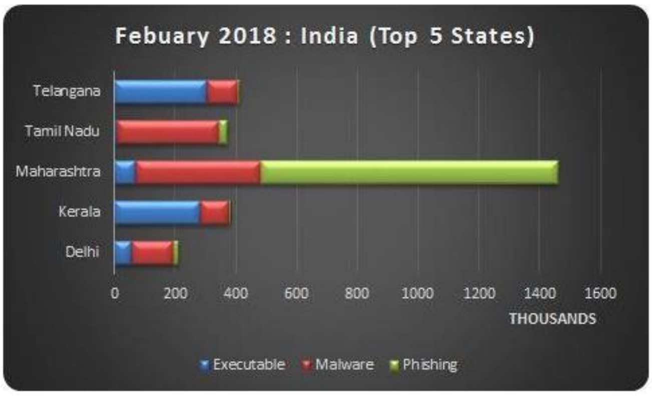

Fig. 4. Emergence of malicious attacks in India.

There has been a sharp increase in theransomware family and its variants, however, the phishing and other attacks are consistent and there has been no sharp rise. India, however, has seen its fair share of attacks with Maharashtra being the most targeted by Phishing attacks. Users from Maharashtra havealso been attacked by Malware of various types[4].

Although there are many methods exist to prevent phishing attacks, still its wings are spread over the entire network, the web and duping individuals, organization, and the society. This research work aims to presentsa data mining method to construct a model to protect against phishing attacks. The architectural model provides a powerful approach to identify phishing sites without inducing high overhead over the browser and work effectively. Identifying different features helped in recognizing the different E-mails into different clusters and able to detect the cluster specially designed by thephishers.

The results by using Gemini as a browser extension for Firefox, Chrome and Internet Explorer are shown as an example. This is to conclude that Phishing attacks are very dangerous threat to individuals, organizations, and the society. The proposed work is very efficient methodology in terms of complexity and overhead to detect phishing attacks.

VIII. CONCLUSION

Although there are many methods exist to prevent phishing attacks, still its wings are spread over the entire network, the web and duping individuals, organization, and the society. This research work aims to presents a data mining method to construct a model to protect against phishing attacks. The architectural model provides a powerful approach to identify phishing sites without inducing high overhead over the browser and work effectively. Identifying different features helped in recognizing the different E-mails into different clusters and able to detect the cluster specially designed by the phishers. The results by using Gemini as a browser extension for Firefox, Chrome and Internet Explorer are shown as an example. This is to conclude that Phishing attacks are very dangerous threat to individuals, organizations, and the society. The proposed work is very efficient methodology in terms of complexity and overhead to detect phishing attacks.

REFERENCES

[1] Choon Lin Tan, Kang Leng Chiew, San Nah Sze , "Phishing Webpage Detection Using Weighted URL Tokens for Identity Keywords Retrieval", in the proceedings of 9th International Conference on Robotic, Vision, Signal Processing and Power Applications, pp 133-139, Springer Singapore, 2017.

[2] U Gürtürk, M Baykara, M Karabatak, "Identifying the Visitors with Data Mining Methods from Web Log Files", International Journal of Emerging Technologies in Engineering Research (IJETER), 5(3), 243- 249, 2017.

[3] B. Gupta, A. Tewari, A. K. Jain, and D. P. Agrawal, "Fighting against phishing attacks: state of the art and future challenges," Neural Computing and Applications, vol. 28, no. 12, pp. 3629–3654, 2017.

[4] A. Aleroud and L. Zhou, "Phishing environments, techniques, and countermeasures: A survey," Computers & Security, vol. 68, pp. 160 – 196, 2017. [Online]. Available: http:// www.sciencedirect.com/science/article/pii/S01 67404817300810.

[5] Dipesh Vaya, Sarika Khandelwal, Teena Habpawat, "A Review on Visual Cryptography", International Journal of Computer Applications, Volume.174 (Issue 05), ISSN: 0975- 8887, September 2017.

[6] The biggest phishing attacks of 2018 and what companies can dot prevent them in 2019, available at: https://www.techrepublic.com/article/the- biggest-phishingattacks-of-2018-and-what- companies-can-do-to-prevent-themin-2

[7] P. Yi, Y. Guan, F. Zou, Y. Yao, W. Wang and T. Zhu, "Web Phishing Detection Using a Deep Learning Framework", Wireless Communications and Mobile Computing, vol. 2018, pp. 1-9, 2018

[8] K. L. Chiew, J. S.-F. Choo, S. N. Sze and K. S. C. Yong, "Leverage Website Favicon to Detect Phishing Websites", Security and Communication Networks, vol. 2018, pp. 1- 11, 2018.

[9] A. Tewari, A. K. Jain, and B. B. Gupta, "Recent survey of various defense mechanisms against phishing attacks," Journal of Information Privacy and Security, vol. 12, no. 1, pp. 3–13, 2016.

[10] A. K. Jain and B. B. Gupta, "A novel approach to protect against phishing attacks at client side using auto-updated white- list," EURASIP Journal on Information Security, vol. 2016, article 9, 11 pages, 2016.

[11] M. Moghimi and A. Y. Varjani, "New rule- based phishing detection method," Expert Systems with Applications, vol. 53, pp. 231– 242, 2016.

[12] G. A. Montazer and S. Yarmohammadi, "Detection of phishing attacks in Iranian e- banking using a fuzzy-rough hybridsystem," Applied Soft Computing, vol. 35, pp. 482–492, 2015.

[13] A. Mishra and B. B. Gupta, "Hybrid solution to detect and filter zero-day phishing attacks," in Proceedings of the Emerging Research in Computing, Information, Communication and Applications (ERCICA '14), Bangalore, India, August 2014.

[14] K. L. Chiew, E. H. Chang, S. N. Sze, and W. K. Tiong, "Utilisation of website logo for phishing detection," Computers & Security, vol. 54, pp. 16–26, 2015.

[15] K. Parsons, A. McCormac, M. Pattinson, M. Butavicius, and C. Jerram, "The design of phishing studies: challenges for researchers," Computers & Security, vol. 52, pp. 194–206, 2015.