International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-9 Issue-1, May 2020

URL Based Phishing Detection

Neeraja Koppula,Vrinda Prabhakaram Ganti, Pranith Gandhe

Abstract: ***Damage caused due to phishing attacks is that which targets the user's personal information. Phishing includes sending a user an email, or causing a phishing page to steal personal information from a user. Blacklist-based detection techniques can detect this form of attack; however, these approaches have certain limitations, and the number of people affecting have continued to grow. The aim of a technique for phishing detection using machine learning to identify each URL into either a legitimate URL or a phished URL. Data availability here in this proposed solution is the key to executing the solution and if there is any issue with data availability it can cost the project accuracy. Data used for model testing must be reliable and appropriate to almost identify all the websites that the user wants to check. Model consistency is another factor that may trigger project failure so the model has to be accurate in determining a true identity of URLs. This technique employs features of a standardized resource locator (URL). The features have been defined which contain URLs for the phishing site. The suggested approach employs certain characteristics to detect phishing. The strategy was tested with a data collection of 3,000 URLs for the phishing site and 3,000 valid URLs for the site. The findings show that more than 90 percent of phishing sites can be identified by the proposed technique.***

Keywords : Phishing, ID3, PhishTank, Legitimate URLs, Phishing URLs.

I. INTRODUCTION

Web attacks have fully grown in amount and improved in quality with the fast development of net|the net} climate and therefore the diversification of obtainable web services. Phishing may be a sort of social attack which targets the delicate info of a user by causing a phishing email via a pretend web {site} that seems like a legitimate site.[1] in step with analysis administered by the opposed Phishing social unit (APWG), within the second quarter of 2010;85,062 phishing sites were found globally; by the second quarter of 2014; 128,978 were known. These statistics show a growth of 1.5 times the worth in an exceedingly quarter that has the phishing attack that occurred. In addition, phishing-caused annual harm was calculable at $5.9 billion. therefore phishing may be a dangerous observe worldwide that tends to develop.[2,3]

In response to the current growth in attacks, the main focus of respectable analysis was on phishing detection techniques. Typical techniques for phishing detection embody the blacklist detection approach. The technique keeps a typical link surveyor (URL) list of websites known as phishing sites; if there's a page in this list requested, the association is blocked[3]. This system is widely used and contains a lesser false-positive rate; but, the consistency of the list that's maintained defines its accuracy. so it's the disadvantage of not having the ability to identify temporary phishing sites.

Phishing could be a web site forgery technique that aims to watch and steal on-line users 'sensitive data. The hacker tricks the user with manipulation techniques like SMS, voice, email, web site and malware. numerous ways are developed and enforced to notice completely different phishing attacks like the utilization of blacklists and whitelists to call many.

In this proposed idea of solution, there is a tendency to area a unit proposing a desktop application referred to as PhishDetect that focuses on the phishing webpage's URL and web site content[4-5]. With the support of a package application referred to as PhishDetect a tendency to aim to notice phishing websites have achieved. To notice a range of phishing attacks, PhishDetect uses a mixture of blacklist and a range of tools. GoogleApiServices have been used for blacklist, that is Google's safe blacklist browsing, since it is consistently monitored for updates and maintained by Google[9]. PhishDetect will|can even|may also|may} be run as a daemon method that means it will be able notice the attacks in real time whereas a user will browse the web. PhishDetect takes uniform resource locator as its input data and creates uniform resource locator standing as a phishing or legitimate website. The properties accustomed find phishing area unit null price footer links, zero links within the hypertext markup language body, copyright data, title data and identity of the web site. PhishDetect is capable of sleuthing hour phishing attacks that will not be blacklisted, and is faster than the visual-based finding ways accustomed detect phishing. We tend to note that PhishDetect has obtained the next exactitude rate and it encompasses a broader spectrum of phishing attacks leading to less false negative and false positive performance.

Manuscript received on April 02, 2020.
Revised Manuscript received on April 15, 2020.
Manuscript published on May 30, 2020.
* Correspondence Author
Neeraja Koppula, Associate Professor, Department of Information Technology, MLR Institute of technology, Hyderabad, India. Email: kneeraja123@gmail.com
Vrinda Prabhakaram Ganti, Department of Information Technology, MLR Institute of technology, Hyderabad, India. Email: gantivrinda@gmail.com
Pranith Gandhe, Department of Information Technology, MLR Institute of technology, Hyderabad, India. Email: pranithgupta7@gmail.com

II. LITERATURE SURVEY

A. *"An Evaluation of Machine Learning-based Methods for Detection of Phishing Sites"* Based on matrix, the heuristic-based solution calculates the probability of a site becoming a phishing site and compares with the given threshold of discrimination. The rate of detection of current heuristic-based solutions is different from practical usage.

B. *"Heuristic-based Approach for Phishing Site Detection using URL Features"* The heuristic-based approach analyzes features of phishing sites and uses those features to create a classifier. The blacklist-based method has the plus points of easy implementation on and a low false-positive rate; however, it can not detect phishing sites, even temporary sites, that are not specified in the database.

C. *"Detecting Phishing Websites, a Heuristic Approach"* In legitimate websites, The number of links redirecting to their domain is high relative to the number of links redirecting to a foreign domain. Phishing identification based on visual and image similarities requires a mechanism for robustly retrieving the content of a website. Any distortion in the processing of web page information contributes to false positives.

D. *"Website Phishing Detection using Heuristic Based Approach"* The number of links redirecting to their domain is possible on legal websites as compared with the frequency of hyperlinks pointing to international domains. Device used SVM showing 96 per cent accuracy and very low false-positive rate.

E. "*Phishing detection: a literature survey*" This paper depicts few researches and also what the definitions of phishing are. It tells the various ways to detect the phishing sites.

F. *"Countermeasure techniques for deceptive phishing attack*", A description of a malicious phishing attack and its tactics is discussed in this article, which is called anti-phishing.

G. *"Beyond blacklists: learning to detect malicious web sites from suspicious URLs"* , This article talks about how malicious websites are detected by using supervised learning techniques.

H. "*Implementation of phishing URL detection using data mining techniques"* This paper talks about the losses of communication due to phishing and also talks about how phishing detection can be done with the help of data mining techniques.

I. "*Hybrid Client Side Phishing Websites Detection Approach*" This paper explains a client side solution for phishing as an browser extension. It gives the URL description.

J. *"A comparison of machine learning techniques for phishing detection."* There are several phishing detection applications available. Unlike predicting spam, however, there are few studies comparing ML techniques in predicting phishing. This research compares the predictive accuracy of various ML approaches, including Classification and Regression Trees (CART), and Sup- Port Vector Machines (SVM), Random Forests (RF), Bayesian Additive Regression Trees (BART), Logistic Regression (LR) and Neural Networks (NNet) for phishing email prediction. The comparative analysis utilizes a data collection of 2889 phishing and original emails. Additionally, 43 functions were used to train and evaluate the classifiers.

III. PROBLEM STATEMENT

PhishDetect takes universal resource locator as its input and creates universal resource locator standing as a phishing or legitimate website. The properties accustomed notice phishing square measure null price footer links, zero links within the HTML body, copyright data, title data and identity of the web site. PhishDetect is capable of sleuthing hour phishing attacks that will not be blacklisted, and is faster than the visual-based noticeion strategies accustomed to detect phishing. We tend to note that PhishDetect has obtained a better exactness rate and it encompasses a broader spectrum of phishing attacks leading to less false negative and false positive performance.

IV. PROPOSED SOLUTION

Malicious websites collect a variety of various black-market entities that square measure dangerous to go to, that is why numerous kinds of malicious websites assign specific risks to users. Once this way of threat is known, it'll be simple to objectively examine these sorts and establish their options which will facilitate monitoring the malicious web site and notice an answer to a specific form of danger. This paper explains the foremost common options wont to establish legitimate and phishing WebPages distinction supporting the practicality of the URLs. Through reviewing all of the options, one will decide that the web site represents the subsequent options that square measure thought of phishing. Common options square measure known to develop a legitimate web site and these options square measure compared with options of the phishing website. This can be done by the prediction formula. variations square measure known, Associate in Nursing formula is developed to differentiate and establish the legitimate web site from the phishing web site by considering these options.

Their square measures several limitations on the present systems that Phishing detection will overcome. Phish's principal benefits over commonplace phishing detectors square measure as follows:

1. It is based on a property identification method capable of detecting Zero hour phishing attacks which are fairly new phishing attacks
2. The system comprises only 5 modules which function as strainers to determine the URL's legitimacy. People who use simply need to include the website's URL, the authenticity of which must be decided. The consumer wants nothing else to do.
3. The algorithm can easily determine even complex phishing attacks. It has comparatively lower false positive and lower false negative levels.
4. Its estimated accuracy rate for detecting phishing websites is 96.57 percent.

   The key benefit is that it will detect phishing sites that deceive people by placing content with pictures that are not visible by most of the current anti-phishing techniques, even though they do, they take longer than our application to execute.

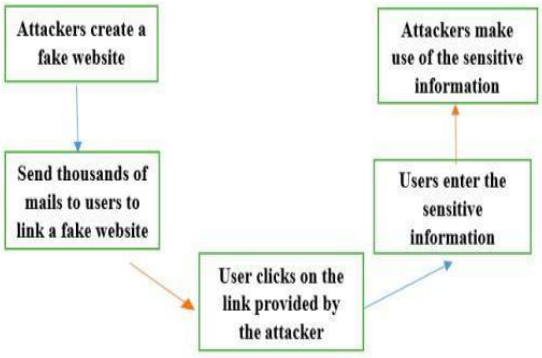

Fig 1. Block diagram for Phishing Attack

V. RESULT AND DISCUSSION

"PhishDetect " is predicated on uniform resource locator content and uniform resource locator website like footer links, and copyrights and title info, and secure blacklist browsing API. The API was accustomed to analyze online page hypertext mark-up language contents and extract content from htmls. ChromeDriver may be a driver tool for launching a Chrome browser from within the app. For blacklists we have a tendency to use Google's Secure Browsing API to conduct an internet search that is the first means of detection. For analysis functions, we have a tendency to use an internet site named PhishTank for phishing URLs. PhishTank is an associate degree anti-phishing web site that enables everybody to send, verify, monitor or share the phishing info. It contains a phishing info containing legitimate phishing pages, on-line or offline. A complete of 250 legitimate, invalid, offline, on-line phishing websites uniform resource locators were gathered from this site to assess the potency of the PhishDetect application in detection phishing websites.

VI. CONCLUSION

Phishing attacks via URLs is one in every of the key problems the web community faces because of on-line everyday transactions. The phishing attacks may cause substantial damages to enterprises, consumers and internet users. The phishing victims were social interacting sites, as well as Facebook, LinkedIn and Twitter. Anti-phishing tools conjointly exist that may facilitate users establish and stop phishing attacks.

The Phishing Detection System detects harmful uniform resource locators and defines the aim for analyzing a URL as a phishing which will facilitate users remember of malicious and suspicious URLs like this. The program offers sixteen properties for the user to apply on input uniform resource locators once the user has entered a uniform resource locator to make a decision if a uniform resource locator is phishing or legitimate. It saves all the information input URLs which might be called for the later use. additionally the program conjointly shows the results of phishing within the style of graphs.

The analysis is administered on all legitimate websites further as on malicious websites obtained from phish tanks. The analysis is administered on the mix of many properties further as individual properties to make sure machine economical practicality. From a group of URLs checked, the program has classified most URLs as acceptable. The system assessment is conducted employing a uncertainty matrix containing verity True Positives, True Negatives, and False Negatives and False Positives. Once all of this data is gathered, the program measures the accuracy and also the recall. The preciseness and recall dissent consequently, looking on the properties chosen by the user. The false negatives and false positives are often decreased for a more robust accuracy and recall, which is able to improve classification accuracy.

REFERENCES

1. Khonji, Mahmoud, Youssef Iraqi, and Andrew Jones, "*Phishing detection: a literature survey*", Communications Surveys & Tutorials, (2013): 2091-2121.
2. Huang, Huajun, Junshan Tan, and Lingxi Liu. *"Countermeasure techniques for deceptive phishing attack*", New Trends in Information and Service Science, 2009. NISS'09. International Conference on, 2009.
3. Ma, Justin, et al. *"Beyond blacklists: learning to detect malicious web sites from suspicious URLs"*,Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009.
4. Jeeva, S. Carolin, and Elijah Blessing Rajsingh. "*Intelligent phishing url detection using association rule mining*", Human-centric Computing and Information Sciences 6.1 (2016): 1-19
5. Hall, Mark, et al. "*The WEKA data mining software: an update*", ACM SIGKDD explorations newsletter 11.1 (2009): 10-18.
6. Kausar, Firdous, et al. "*Hybrid Client Side Phishing Websites Detection Approach*", International Journal of Advanced Computer

Science and Applications (IJACSA) 5.7 (2014).
7. Sunil, A. Naga Venkata, and Anjali Sardana. "*A pagerank based detection technique for phishing web sites"*, Computers & Informatics (ISCI), 2012.
8. Mohammad, Rami M., Fadi Thabtah, and Lee McCluskey."*Intelligent rule-based phishing websites classificatio.*", Information Security, IET 8.3 (2014): 153-160.
9. Canali, Davide, et al. *"Prophiler: a fast filter for the large-scale detection of malicious web pages."* Proceedings of the 20th international conference on World wide web. ACM, 2011.
10. Xiang, Guang, et al. "*Cantina+: A feature-rich machine learning framework for detecting phishing web sites."* ACM Transactions on Information and System Security (TISSEC) 14.2 (2011): 21.
11. WANG, Wei-Hong, et al. *"A Static Malicious Javascript Detection Using SVM*." strings. Vol. 40. 2013.
12. L. Ladha and T. Deepa, "*Feature selection methods and algorithms*," International journal on computer science and engineering, vol 3, no 5, 2011.
13. Hou, Yung-Tsung, et al. *"Malicious web content detection by machine learning*", Expert Systems withApplications 37.1 (2010): 55-60.
14. Cao, Ye, Weili Han, and Yueran Le. "*Anti-phishing based on automated individual white-list*", Proceedings of the 4th ACM workshop on Digital identity management. ACM, 2008.
15. Huh, Jun Ho, and Hyoungshick Kim. "*Phishing detection with popular search engines: Simple and effective*", Foundations and Practice of Security. Springer Berlin Heidelberg, 2012. 194-207.
16. Abela, Kevin Joshua, et al. "*An automated malware detection system for android using behavior-based analysis AMDA*." International Journal of Cyber-Security and Digital Forensics (IJCSDF) 2.2 (2013): 1-11.
17. Abu-Nimeh, Saeed, et al. *"A comparison of machine learning techniques for phishing detection."* Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit. ACM, 2007.
18. A. Y. Fu, L. Wenyin and X. Deng, "*Detecting Phishing Web Pages with Visual Similarity Assessment based on Earth Mover's Distance (EMD)*", Transactions on Dependable and Secure Computing, vol. 3(4), pp. 301–311, (2006)
19. J. Mao, P. Li, K. Li, T. Wei and Z. Liang, "*Baitalarm: Detecting Phishing Sites using SimilarityinFundamental Visual Features*", In 5th International Conference on Intelligent Networking and Collaborative Systems, INCoS 2013, pp. 790–795, September (2013).
20. L. Wenyin, G. Huang, L. Xiaoyue, Z. Min and X. Deng, "*Detection of Phishing Webpages based on Visual Similarity"*, In Special Interest Tracks and Posters of the 14th International Conference on World Wide Web, ACM, pp. 1060–1061, May (2005).
21. G. Xiang, and J. I. Hong, *"A Hybrid Phish Detection Approach by Identity Discovery and Keywords Retrieval*", In Proceedings of the 18th International Conference on World Wide Web, ACM, pp. 571–580, April (2009).

AUTHORS PROFILE

Neeraja Koppula, Associate Professor, Department of Information technology, MLR Institute of Technology, Hyderabad, India. Email: kneeraja123@gmail.com

Vrinda Prabhakaram Ganti, Department of Information Technology MLR Institute of technology, Hyderabad, India. Email: gantivrinda@gmail.com

Pranith Gandhe, Department of Information technology, MLR Institute of Technology, Hyderabad, India. Email: pranithgupta7@gmail.com