TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 9041 - 9046

Identifying Phishing URLs using Cosine Similarity

Bhawna Sharma[a], Dr. Parvinder Singh[b]
[a]Research Scholar, Deenbandhu Chhotu Ram University of Science & Technology, Murthal, Sonepat-131039, Haryana, India,
Email: bhawnash024@gmail.com
[b]Professor, Department of Computer Science &Engineering,DeenbandhuChhotu Ram University of Science & Technology, Murthal, Sonepat -131039, Haryana, India, Email: parvinder23@rediffmail.com

Article Info
Volume 82
Page Number: 9041 – 9046
Publication Issue:
January-February 2020

Article History
Article Received*: 18 May 2019*
Revised*: 14 July 2019*
Accepted*: 22 December 2019*
Publication*: 09 February 2020*

Abstract:
Phishing is one of the serious issues looked by digital world and prompts budgetary misfortunes for ventures and people. Discovery of phishing assault with high precision has consistently been a difficult issue. Phishing site looks fundamentally the same as in appearance to its relating genuine site to beguile clients into accepting that they are perusing the right site. In this article, we acquaint with cosine-similarity centered phishing identification technique which calculates cosine-similarity between test vectors and training vectors. A high value of cosine-similarity indicates more similarity between the two vectors. The proposed technique is highly efficient. We test our technique using 100 URLs in testing dataset and 300 URLs in training dataset. Experiments show that the proposed technique classified the test data with 98.7% accuracy.

Keywords: digital world, Phishing,.

I. INTRODUCTION

At present, phishing, a sort of social designing assault, is one of most regular assault types utilized by digital assailants to draw naivecustomers to disclose their subtleinformation, for example, client certifications or charge card data [15]. As per the report generated by phishing removal group, in the second quarter the quantity of phishing assaults of 2019 overshadowed the number found in the seventy five percent previously. The all out number of phishing destinations distinguished by APWG in April through June 2019 was 182,465. This beat the 180,768 found in first quarter of 2019, and was up remarkably from the 138,328 found in the final quarter of 2018 [20]. Additionally, these assaults have developed after some time and become progressively further developed as phishers endeavor to make the vibe of their phishing site pages and relating URLs as comparative as conceivable to target sites while using different avoidance strategies to go around existing phishing countermeasures.

Phishing site identification can be done using blacklists and whitelists. Web browsers coordinate either blacklists or whitelists to shield clients from the problem of phishing [14]. The popular search engine, Google, gives a blacklist of noxious sites that is persistently refreshed. Google Safe Browsing APIs can be used by the clients for checking URL security. List-based phishing identification is fast but their update process is very slow. Notwithstanding blacklist and whitelist, AI strategies are broadly utilized in phishing site recognition. The explanation is that malignant URLs or phishing pages have a few attributes that can be recognized from authentic sites, and AI can be powerful in such manner for preparing [13], [18]. Current standard AI strategies for phishing site discovery separate

January - February 2020
ISSN: 0193 - 4120 Page No. 9041 - 9046

factual highlights from the URL and the host or concentrate pertinent highlights of the site page, for example, the format, CSS, content, and afterward characterize these highlights [1], [3]. Be that as it may, these techniques just dissect the URL or concentrate highlights from a solitary point of view, which makes it difficult to extricate the total traits of phishing sites. Additionally, some preposterous highlights may decrease the precision of location. The character arrangement of the URL is common, consequently created highlight that maintains a strategic distance from the subjectivity of artificially chose highlights. Moreover, it doesn't need outsider help and any earlier information about phishing.

To discourseabove mentioned issues, we put forward a model that groups each URL in the test dataset as phishing or genuine. This model uses cosine similarity as the measure of evaluating similarity between the test URLs and the training URLs. It is based upon the idea that any test vector is considered as phishing if its average cosine similarity becomes less than the threshold value [6]. Otherwise, the test vector is considered as legitimate.

While there is a lot of work on phishing recognition, our work is novel in the accompanying manners:

1) Used char2vector as a tool for converting each phishing URL in the database to its corresponding vector form.
2) Evaluated average cosine-similarity of testing dataset with training dataset
3) Selected an appropriate threshold value for classifying testing dataset
4) Proposed an efficient model for identifying phishing URLs

The rest of the paper is designed in this manner: Next section, Section 2, agreements with survey based upon a lot of existing methods for solving the phishing problems. Section 3 talks about the proposed model. Experimentation and Results are exhibited in Section 4. At last, Section 5 accomplishes the paper.

II. RELATED WORK

Jain and Gupta [1] proposed a review on phishing identification approaches dependent on visual similitude. This study gives a superior comprehension of phishing site, different arrangement, and future extension in phishing recognition. Numerous methodologies are talked about in this paper for phishing location; anyway a large portion of the methodologies still have constraints like precision, the countermeasure against new phishing sites, neglecting to recognize installed objects, etc. These methodologies utilize different highlights of a website page to recognize phishing assaults, for example, content comparability, text style shading, text dimension, and pictures present in the site page. Abutair and Belghith [2] can efficiently distinguish web phishing assaults because of the ceaseless alteration and the tinyexistence cycle of fake sites by utilizing Case-Based Reasoning Phishing Detection System. PhishMon [3] is based upon AI structure which is used to distinguish phishing site pages and mainly depends upon fifteen novel highlights that can be registered without requiring outsider administrations, for example, web indexes, or WHOIS servers. Further theses featuresfocus on various qualities of web applications for authenticationfor their hidden web frameworks and copying of these highlights required to invest more energy and exertion on their hidden frameworks and web applications. Peng Yang, Guangzhen Zhao, PengZang [4] proposed anapproach based upon multidimensional component. In the first phase of approach, evacuated the URL character gathering features that is used for classification based upon learningtechnique.This movement need not require any help from thirdparty or old data for phishing.Henceforth joining the URL authentic

January - February 2020
ISSN: 0193 - 4120 Page No. 9041 - 9046

featuresi.e code of page, contect etc.Christopher et. al [5] developed a SAFE-PC framework for distinguish latest phishing and uses real world phishing. Further it removes each message's header features and body. Hossin, M. also, Sulaiman, M.N [6] efficiently investigated the related assessment measurements that are explicitly planned as a discriminator for improving generative classifier. Prof. Sarikaet. al [7] suggested a novel system of figuring cosine comparability. The system is approved by watching exploratory outcomes which demonstrate to be more financially savvy and effective. SaharSohangir and Dingding Wang [9] portrayed the utilization of cosine closeness, which is described by the correlation of comparative analysis of two vectors. This exploration utilized two methodologies: (1) word2vec and (2) Bag-of-Words (BoW) for separating every single applicable tweet. PhiDMA [17] proposed a Phishing Detection frameworkbased upon Multi-layered Approach. Proposed framework used whitelist, feature extractorprocess of URL, generatorof lexicalsignatureand calculate the accessibility score that were further used to identified phishing goals.

III. THE PROPOSED MODEL

A. Cosine-Similarity

Given two vectors p and v and state θ is the point between these two vectors, at that point the cosine likeness [7], [8], denoted bycos(θ), is spoken to utilizing a speck item and extent as

$$\text{cosine} - \text{similarity} = \cos(\theta) = \frac{\text{p.v}}{\|\text{p}\|\|\text{v}\|} \quad (1)$$

B. Description of Proposed Model

The proposed model is designed to classify all URLs as either legitimate or phishing. This model is isolated into two stages, specifically, Training Phase and Classification Phase[16], [19]. Training initiates with the uploading of training dataset.Every character present in the URL is converted to its ASCII value. This in turn forms a vector corresponding to the URL. This process is repeated for every URL in the training dataset. All the generated vectors are kept in char2vector database. Average cosine similarity of each training vector with the training dataset (avgTraining) is computed. Algorithm 1 describes the training phase.

In classification phase, testing dataset is uploaded first. Then every test URL is converted to its corresponding vector using char2vector conversion. Average cosine similarity of each test vector with the training dataset (avgTest) is computed. Algorithm 2 describes the classification phase. If the absolute value of the difference between avgTraining and avgTest comes out to be less than the threshold value t, than the test vector is declared as phishing. Otherwise, the test vector is categorized as legitimate.

Algorithm 1: Training Phase

1) Upload training dataset.
2) Generate char2vector for every training instance and store all vectors in char2vector database.
   //cosine-similarity works on vectors
3) foreach v in vector
   foreach p in vector
   if (v!=p) calculate cosine similarity for v and p //use equation (1)
   else choose next vector pair
4) Calculate average cosine similarity of each training vector (avgTraining).

Algorithm 2: Classification Phase

1) Upload test data.

January - February 2020
ISSN: 0193 - 4120 Page No. 9041 - 9046

2) Convert each test URL into its corresponding vector by using char2vector conversion.
3) Calculate cosine similarity of test vector with all training vectors.
4) Calculate average cosine similarity (avgTest) of test vector.
5) If (abs (avgTest-avgTraining)< t)then declare the test URL as phishing; // t is threshold value
   Else the URL is declared as legitimate.
6) Repeat steps 2-4 for all test vectors.

IV. EXPERIMENT AND RESULTS

To consider our methodology, we gathered an enormous dataset containing 500 real and 500 one of a kind phishing URLs from PhishTank. In this dataset, authentic pages are webpages of the well-known sites chosen arbitrarily from the Alexa top one million area name list, and phishing pages are unmistakable confirmed phishing occurrences chosen from PhishTank 12,769 URLs, a network driven site for sharing and approving phishing URLs [17].

At first, we chose haphazardly 500 from Alexa top one million websites. We at that point rejected the websites showed up in malwaredomains.com and networksec.orgblacklists. Next, we visited the site of the rest of the domains with our web scrubber to frame the dataset of real website pages. Further, we evacuated pages that contain certain expressions demonstrating the webpage is under development, not useful, or not supporting the search engine utilized by our web scrubber. Along these lines, we acquired 100 real webpages. Out of 12,769 phishing URLs, we first selected 300 URLs that were most recent ones. Then, we excluded the duplicate URLs and finally selected 205 phishing URLs.

A trial was intended to assess the adequacy of our proposed technique. It used 205 phishing instances and 100 legitimate instances taken from PhishTank and Alexarespectively. The proposed method was implemented in MATLAB. The training dataset contained 305 URLs (205 phishing and 100 legitimate). The test dataset contained 100 URLs (50 legitimate and 50 phishing). The description of dataset is given in Table 1. Figure 1 represents the training and phishing instances used in the experiment. Figure 2 shows the detection results for the experiment using three performance measures, namely, precision, recall and f-measure [12].Precision is utilized to quantify the positive examples that are accurately anticipated from the absolute anticipated examples in a positive class. Recall is utilized to gauge the portion of positive examples that are accurately identified. F-measure is calculated by finding the harmonic mean between precision and recall [10], [16]. Values obtained for precision, recall and F-measure are 1, 0.988 and 0 .994respectively.

Table 1. Description of dataset

Dataset	Legitimate Instances	Phishing Instances
Training	100	205
Testing	50	50

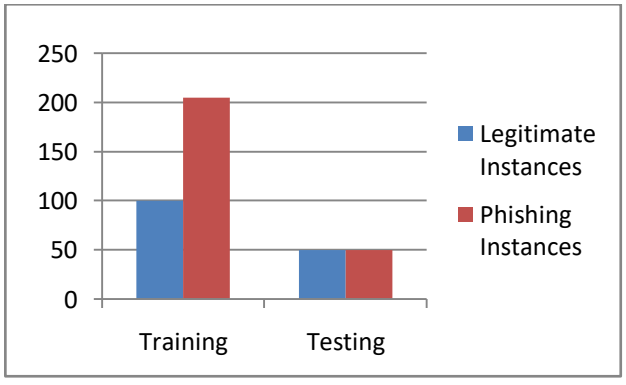

Figure 1.Training and Testing instances

January - February 2020
ISSN: 0193 - 4120 Page No. 9041 - 9046

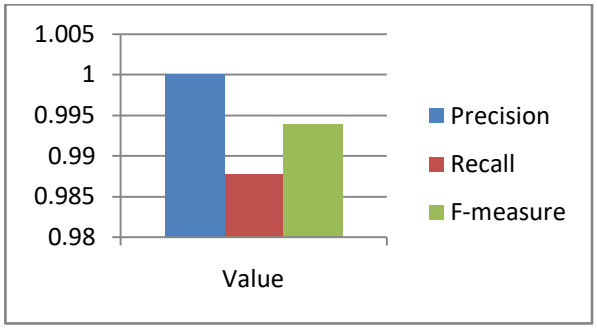

Figure 2. Performance Measures

V. CONCLUSION

Distinguishing phishing URLs stays an exceptional investigation. Our model predicts phishing attacks with a high accuracy. The proposed model considers cosine-similarity as the measure for comparing similarity between test set URLs and training set URLs.The higher the value of cosine-similarity, the more will be the degree of similarity between two URL vectors. We used char2vector as a tool for converting URL to vector. It is versatile to different informational collection estimates and can adjust proactively. In this paper, we collected a training dataset having305 URLs out of which 205 are legitimate and 100 are phishing URLs. We tested it upon a test dataset containing 50 legitimate and 50 phishing URLs. The values obtained for precision, recall and F-measure are1, 0.988 and 0 .994 respectively. For future scope, we can integrate our proposed model with machine learning framework.

VI. REFERENCES

[1] Ankit Kumar Jain, B.B.Gupta, "Phishing Detection: Analysis of Visual Similarity Based Approaches", Security and Communication Networks, Vol. 2017, pp. 1-20, 2017

[2] Hassan Y. A. Abutair, AbdelfettahBelghith, "Using Case-Based Reasoning for Phishing Detection", The 8th International Conference on Ambient Systems, Networks and Technologies, Volume 109, 2017, pp. 281-288, 2017

[3] AmirrezaNiakanlahiji, Ehab Al-Shaer, "PhishMon: A Machine Learning Framework for Detecting Phishing Webpages", IEEE International Conference on Intelligence and Security Informatics, pp. 220-225, Nov. 2018

[4] Peng Yang, Guangzhen Zhao, PengZang, "Phishing Website Detection Based on Multidimensional Features Driven by Deep Learning", IEEE Access, Vol. 7, pp. 15196 – 15209, Jan. 2019

[5] Christopher N. Gutierrez, Taegyu Kim, Raffaele Della Corte, Jeffrey Avery, Dan Goldwasser, Marcello Cinque, SaurabhBagchi, "Learning from the Ones that Got Away: Detecting New Forms of Phishing Attacks", IEEE Transactions on Dependable and Secure Computing, , Vol. 15, pp. 988-1001, Nov.-Dec. 2018

[6] Hossin, M. and Sulaiman, M.N., "A Review on Evaluation Metrics for Data Classification Evaluations", International Journal of Data Mining & Knowledge Management Process, Vol.5, No.2, pp. 1-11, Mar. 2015

[7] Prof. Sarika N Zaware, Mr. AsmitGautam, Ms. SumedhaNashte, Ms. PuneetKhanuja, "An Effectual Approach for Calculating Cosine Similarity", International Journal of Advance Engineering and Research Development, Vol. 2, No. 4, pp. 13-18, Apr. 2015

[8] SaharSohangir, Dingding Wang, "Improved sqrt-cosine similarity measurement", Journal of Big Data, Vol. 2017. No. 1, pp. 1-13, July 2017

[9] Carolina Focil-Arias, Jorge Ziiniga, GrigoriSidorov, IldarBatyrshin, Alexander Gelbukh, "A tweets classifier based on cosine similarity",CEUR Workshop Proceedings, Vol. 1866,pp. 1-10, 2017

[10] MarinaSokolova, GuyLapalme, "A systematic analysis of performance measures for classification tasks", Information Processing & Management, Vol. 45, No. 4, pp. 427-437, July 2009

[11] Tommy Chin, KaiqiXiong, Chengbin Hu, "PhishLimiter: A Phishing Detection and Mitigation Approach Using Software-Defined Networking", IEEE Access, Vol. 6, pp. 42516 - 42531, 2018

[12] Murat Karabatak, Twana Mustafa, "Performance Comparison of Classifiers on Reduced Phishing Website Dataset", IEEE, pp. 1-5, 2018

[13] Sophie Le Page, Guy-Vincent Jourdan, Gregor v. Bochmann, Jason Flood, Iosif-ViorelOnut, "Using URL Shorteners to Compare Phishing and Malware Attacks", IEEE, pp. 1-13, 2018

January - February 2020
ISSN: 0193 - 4120 Page No. 9041 - 9046

[14] Ankit Kumar 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, pp. 1-11, 2016

[15] Rami M. Mohammad ,FadiThabtah and Lee McCluskey, "Tutorial and critical analysis of phishing websites methods", Elsevier, pp. 1-24, 2015

[16] Mahmoud Khonji, Youssef Iraqi,and Andrew Jones, "Phishing Detection: A Literature Survey", IEEE Communications Surveys & Tutorials, vol. 15, No. 4, 2013

[17] GunikhanSonowal, K.S. Kuppusamy, "PhiDMA - A Phishing Detection Model with Multi-filter Approach", Journal of King Saud University - Computer and Information Sciences, pp. 1-14, July 2017

[18] YasinSönmez, TürkerTuncer, HüseyinGökal, EnginAvcr, "Phishing Web Sites Features Classification Based on Extreme Learning Machine", 6th International Symposium on Digital Forensic and Security, pp.,1-5,2018

[19] MuhammetBaykara ,ZahitZiyaGürel, "Detection of phishing attacks", 6th International Symposium on Digital Forensic and Security, pp. 1-5, 2018 https://www.businesswire.com/news/home/20190912005943/en/APWG-Q2-2019-Report-Phishing-Attacks-Maintaining