A Proposal on Phishing URL Classification for Web Security

Sonam Saxena
Department of Computer Science Engineering
Swami Vivekanand College of Engineering
Indore, India

Amit Shrivastava
Assistant Professor,
Department of Computer Science Engineering
Swami Vivekanand College of Engineering
Indore, India

Vijay Birchha
HOD Professor,
Department of Computer Science Engineering
Swami Vivekanand College of Engineering
Indore, India

ABSTRACT
Data mining and machine learning is one of the most essential tools in new generation technology. That is used in a number of applications i.e. security, banking and decision making. In this paper, data mining application of web data security is described in details. In this context the domain of phishing URL detection and classification is key aim of the proposed work. This paper includes the different aspects of phishing and recently made contributions for accurately classification of phishing URLs. In addition of that a data mining based model is also proposed that is help to classify the phishing URLs more accurately. Finally the paper provides the future extension of the work.

Keywords
Data mining, machine learning, classification, Phishing URL, web security.

1. INTRODUCTION
The main aim of the proposed work is to classify the phishing URLs, which can harm someone socially and financially. In this context first of all need to understand the issue of phishing and web security. Basically phishing is a kind of cybercrime, in this act an attacker is trying to extract the personal and confidential information from the end client. For that purpose the attacker can creates the fake and malicious web pages and their web URLs, and try to influence the end user to visit that malicious web page or URL. When user visit that page or URL then, user provide information on that page. By using the user given details attacker hack their bank accounts and can do other acts [1].

In this presented work, a survey on web phishing URL technique is presented. In addition of that provide a phishing URL classification model for accurately classifying the phishing URLs. The survey is dedicated to find an effective way for classifying the URLs effectively and efficiently therefore the recently developed different techniques are surveyed in this paper. In addition of that based on the outcome of literature the final objectives and their modeling is provided.

2. BACKGROUND
This section involves the key terms and their relevant descriptions which are used in this research paper writing. These keywords and their definitions are used for solution design and system development.

1. *Phishing:* phishing is an act of cybercrime, on which the attacker try to recover the end client's confidential and sensitive information, related to the bank account or any social media account to trap the end client. User when visit the designed URL then he/she provide their confidential information and this act can harm the client in various different ways [2].
2. *Classification:* The classification is supervised technique of learning in data mining. The classification technique first learn on the predefined patterns and then able to identify the trained patterns. Therefore the classification technique are accurate the unsupervised learning technique [3].
3. *URL classification:* In most of the time in classification the data patterns are available in structured manner. But the URL data is not available in a fixed pattern. Therefore to apply the classification techniques or machine learning techniques in URL data. therefore additional approaches are need to be employ for handling the URLs such as feature extraction, URL encoding and others [4].
4. *Phishing detection:* Phishing is a crucial issue in web security. The phishing detection techniques are enabling us to identify the phishing URLs by evaluation of the URLs. In order to evaluate the URLs, a number of techniques are available i.e. black list and white list based technique, statistical analysis based techniques and machine learning based techniques. Among the available techniques the machine learning techniques are more effective and accurate. In such kind of techniques the malicious URL patterns are learnt by classification algorithms and when required it is identified the URL types (phishing or legitimate) [5].
5. *Malicious URLs:* The URLs that contains fake information or duplicate web page of any authentic and reputed web page, additionally that act or deviate from the actual URL behavior is termed here as malicious or phishing URL [6].
6. *Phish tank database:* The phish tank database is a kind of standard repository which keeps records of the phishing reported URLs by different web security agencies. This database contains a number of different attributes among some essential attributes (features) are listed as [7]:

a. Reporting date
b. Phishing target
c. URL
d. Reported agency and others

This section provides the basic overview of different keywords that are going to be used in further system design and description.

3. LITERATURE SURVEY

This section involves the different recently made contributions and research articles that are developed for optimizing the traditional phishing classification techniques.

The increasing volume of pernicious content in social media requires automate methods to detect and eliminate such content. ***Parth Parekh et al [8]*** describes a superintend machine learning classification model that will be built to detect the distribution of pernicious content in online social networks/medias. Multisource features have been used to detect social network posts that contain vitriolic Uniform Resource Locators. These URLs could direct users to websites that contain malignant content, drive-by download attacks, phishing, spam, and scams. For the data collection stage, the Twitter streaming application programming interface was used and Virus Total was used for labeling the dataset. The fraudulent practice of sending emails is a criminal scheme to get the user's personal data and other login and confidential information. It is known as phishing that acquires user's private information such as password, bank account detail, credit card number, financial username and password etc. and later it can be mistreat by attacker .We aim to use fundamental visual features of a web page's appearance as the basis of detecting page similarities. We propose a novel solution, to efficiently detect phishing web pages. Note that page layouts and contents are fundamental feature of web pages' appearance. Since the standard way to specify page layouts is through the style sheet, we develop an algorithm to detect similarities in key elements related to CSS. In this paper, we proposed a system that uses SVM technique along with Image Spam Filtering, spam map reduce archetype to achieve a higher accuracy in detection of the spam urls and image spamming.

In the last decade, numerous fake websites have been developed on the World Wide Web to mimic trusted websites, with the aim of stealing financial assets from users and organizations. This form of online attack is called phishing, and it has cost the online community and the various stakeholders hundreds of million Dollars. Therefore, effective counter measures that can accurately detect phishing are needed. Machine learning is a popular tool for data analysis and recently has shown promising results in combating phishing when contrasted with classic anti-phishing approaches, including awareness workshops, visualization and legal solutions. ***Neda Abdelhamid et al [9]*** investigate ML techniques applicability to detect phishing attacks and describes their pros and cons. In particular, different types of ML techniques have been investigated to reveal the suitable options that can serve as anti-phishing tools. More importantly, we experimentally compare large numbers of ML techniques on real phishing datasets and with respect to different metrics. The purpose of the comparison is to reveal the advantages and disadvantages of ML predictive models and to show their actual performance when it comes to phishing attacks. The experimental results show that covering approach models are more appropriate as anti-phishing solutions, especially for novice users, because of their simple yet effective knowledge bases in addition to their good phishing detection rate.

Hemali Sampat et al [10] describe the detection of Phishing website is an intelligent and effective model that is based on using classification or association Data Mining algorithms. These Algorithms were used to identify and characterize all rules and factors in order to classify the phishing website and relationship that correlate them with each other so we detect them by their performance, accuracy, number of rules generated and speed. Proposed system implements both algorithms which is Classification and Association that optimizes the system which is more efficient and faster than existing system. By using these two algorithms with WHOIS protocol the error rate of the existing system decreases by 30% so by using this method proposed system create an efficient way to detect the phishing website. Although there does not exist a system which can detect the entire phishing website but using these methods it will create a most efficient way to detect the phishing website.

Anna L. Buczak et al [11] depicts an engaged writing review of machine learning (ML) and data mining (DM) techniques for digital investigation in help of interruption location. Short instructional exercise depictions of every ML/DM strategy are given. In light of the quantity of references or the importance of a rising strategy, papers speaking to every technique were distinguished, perused, and outlined. Since information are so vital in ML/DM approaches, some outstanding digital informational collections utilized in ML/DM are depicted. The unpredictability of ML/DM calculations is tended to, dialog of difficulties for utilizing ML/DM for digital security is introduced, and a few proposals on when to utilize a given strategy are given.

In this paper, ***Mahmood Moghimi et al [12]*** present another standard based strategy to identify phishing assaults in web managing an account. Our standard based strategy utilized two novel capabilities, which have been proposed to decide the website page personality. Our proposed capabilities incorporate four highlights to assess the page assets character, and four highlights to recognize the entrance convention of page asset components. We utilized rough string coordinating calculations to decide the connection between the substance and the URL of a page in our first proposed list of capabilities. Our proposed highlights are free from outsider administrations, for example, web indexes result and additionally internet browser history. We utilized help vector machine (SVM) calculation to characterize site pages. Our investigations show that the proposed model can distinguish phishing pages in web keeping money with exactness of 99.14% genuine positive and just 0.86% false negative caution. Yield of affectability examination shows the critical effect of our proposed highlights over customary highlights. We extricated the concealed information from the proposed SVM show by receiving a related technique. We implanted the separated tenets into a program augmentation named Phish Detector to make our proposed technique progressively useful and simple to utilize. Assessing of the executed program expansion demonstrates that it can distinguish phishing assaults in web managing an account with high exactness and unwavering quality. Phish Detector can distinguish zero-day phishing assaults as well.

4. PROPOSED SYSTEM

This section provides the overview of the proposed working model for classifying the phishing URLs.

The proposed data model for classifying the phishing URLs from the legitimate URLs are demonstrated in figure 1. In addition of that their component level description is given as:

The proposed system architecture contains five layers of data processing. In first layer the phish tank data base is provided as input to the system. The phish tank data base is cross verified for finding any missing attribute in dataset. If there are not any attributes are missing then the next phase is called. In the next phase the system processes the data for removing additional attributes from the dataset. Only the URL list is extracted from

the data base and remaining data is removed. In the same layer the URLs are evaluated against the heuristic for computing the URL features. After evaluation of the URLs 14 features from each URL is computed and the URLs are transformed on 2D vector based on the computed properties of URL. By using the threshold values the computed 2D vector is encoded into binary format (0, 1). The encoding is performed for employing the rule mining algorithm, therefore for mining rules two popular algorithms are implemented in this phase namely apriori and FP-tree.

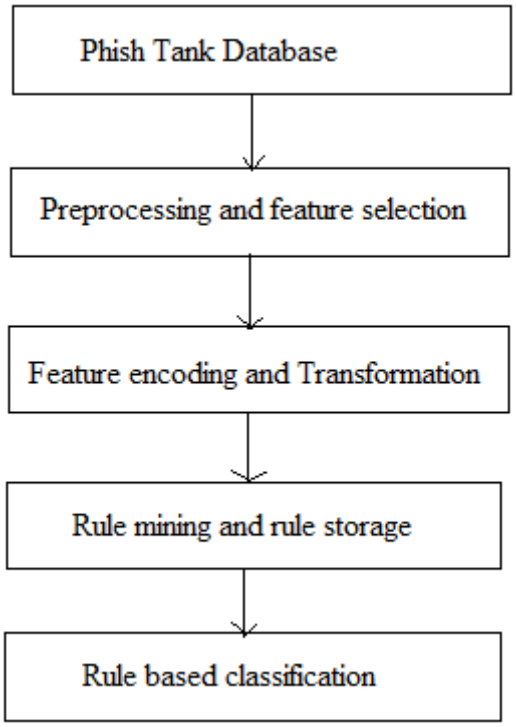

Figure 1 proposed system architecture

Both the algorithms are used for generating the association rules. These rules are stored separately for utilizing the rule in classification of URLs in terms of legitimate and phishing URLs.

5. CONCLUSION

The main aim of the proposed study is to explore the domain of web security more specifically phishing detection and their identification approaches. In this context a rich literature is explored for obtaining the relevant methodologies and techniques of classification. Finally some relevant techniques are obtained and on the basis of available literature an accurate classification model is proposed for design and implementation. The proposed data model is based on the concept of rule mining and rule based classification technique. Therefore two popular rule mining techniques namely apriori algorithm and FP-Tree is proposed for system employment. This model is implemented and their performance evaluation is demonstrated in near future.

6. REFERENCES

[1] B. B. Gupta, Nalin A.G. Arachchilage, Konstantinos E. Psannis, "Defending against Phishing Attacks: Taxonomy of Methods, Current Issues and Future Directions", https://arxiv.org/ftp/arxiv/papers/1705/1705.09819.pdf

[2] G.Parthasarathy, D.C.Tomar, K. Christina Praisy, "AN ENHANCEMENT OF ASSOCIATION CLASSIFICATION ALGORITHM FOR IDENTIFYING PHISHING WEBSITES", Indian Journal of Computer Science and Engineering (IJCSE) Vol. 7 No. 4 Aug-Sep 2016

[3] R. Sathya, Annamma Abraham, "Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification", International Journal of Advanced Research in Artificial Intelligence, Vol. 2, No. 2, 2013

[4] Patrick Dave P. Woogue, Gabriel Andrew A. Pineda, and Christian V. Maderazo, "Automatic Web Page Categorization Using Machine Learning and Educational-Based Corpus", International Journal of Computer Theory and Engineering, Vol. 9, No. 6, December 2017

[5] Zheng Dong, Apu Kapadia, Jim Blythe and L. Jean Camp, "Beyond the Lock Icon: Real-time Detection of Phishing Websites Using Public Key Certificates", 978-1-4799-8909-6/15/$31.00 c 2015 IEEE

[6] Peter F. Likarish, "Early detection of malicious web content with applied machine learning", PhD (Doctor of Philosophy) thesis, University of Iowa, 2011.

[7] https://www.phishtank.com/

[8] Parth Parekh, Kajal Parmar, Pournima Awate, "Spam URL Detection and Image Spam Filtering using Machine Learning", International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056, Volume: 05 Issue: 07 | July 2018

[9] Neda Abdelhamid, Fadi Thabtah, Hussein Abdel-jaber, "Phishing Detection: A Recent Intelligent Machine Learning Comparison based on Models Content and Features", 978-1-5090-6727-5/17/$31.00 ©2017 IEEE

[10] Hemali Sampat, Manisha Saharkar, Ajay Pandey, Hezal Lopes, "Detection of Phishing Website Using Machine Learning", International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056, Volume: 05 Issue: 03 | Mar-2018

[11] Anna L. Buczak, and Erhan Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection", IEEE COMMUNICATIONS SURVEYS & TUTORIALS, VOL. 18, NO. 2, SECOND QUARTER 2016

[12] Mahmood Moghimi, Ali Yazdian Varjani, "New rule-based phishing detection method", Expert Systems With Applications 53 (2016) 231–242