# A Survey on Machine Learning Approach for URL Based Web Phishing

Happy Chapla
Research Scholar
Marwadi Education Foundation
Rajkot, India
*happypatel2212@gmail.com*

Riddhi Kotak
Assistant Professor
Marwadi Education Foundation
Rajkot, India
*riddhi.kotak@marwadieducation.edu.in*

Mittal Joiser
Assistant Professor
Marwadi Education Foundation
Rajkot, India
*mittal.joiser@marwadieducation.edu.in*

***Abstract***— In this era, internet is become crucial part of our day to day life. Hence the security of the internet data is must. Phishing is the threat which is major issue of the web data and its security. Web phishing is well known assaults for acquiring the credential information from the users like security number, bank account number etc. Phishing detection is necessary to overcome this web problem. In this paper we discuss about the different technique of phishing, some classification techniques, and Supported algorithm for the better accuracy. And also give the literature survey of some papers.

***Keywords—****Data mining, Web mining, Classification,Feature Extraction, Phishing*

*****

## I. INTRODUCTION

In today's digital scenario people used to with the Internet and its online facilities for the better and speedy payments. However, the phishing attack is most probably used by the Identity Thieves to acquire the important information of the users. Phishers mainly focus on the credential information of the users like, the record of the company's employees, confidential information of the banks, Global data of the sites, governments' records, passwords and many more on the web. Website forgery is refers to the form of web threat that indirectly get the information of the website users. Generally, Phishers building fake version of the targeted original site which will may be success to get the user's trust. Then they were collecting the information and data from that fake site for further process. In this paper we will show the literature survey of other papers. More than one method used by the authors to increase the accuracy of the system developed by them. Some algorithm is also used some papers for the better performance, however the aim is to detect phishing sites and help the people for stop web fraud.

This article has following sections, Section II contains the overview of Web Mining and its types, Section III discuss the literature survey of papers, Section IV provides the different techniques of machine learning, Section V described tools used to identify the phishing sites. We conclude our work in last section VI that contains some statistics of phishing scenario.

## II. WEB MINING

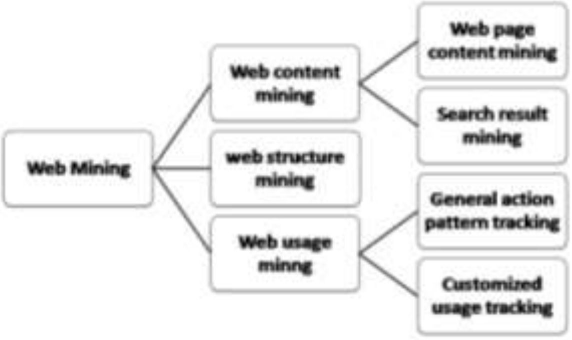

*Fig 1 Parts of Web mining*

**Web mining** – is one of the most important applications used to find out patterns from the World Wide Web.

**Web-content mining-** is the mining and extracting the data and information from the content which are available in the web in the form of websites or web pages. Content which are uploaded by the users are also mined when we need to survey or trying to do the analysis of content in the World Wide Web. It also include the mining of content which is in the webpage. It can also mine search result. When something is written on the web page for the search purpose at that time mining is needed and it is called the search result mining.

**Web structure mining-** is including the pattern structure mining. It is dividing in to two parts. One is used to extracting patterns of hyperlink from the web that contents the web pages mine at the different location. And second is used to mine the document structure that is include the analysis of the tree structure[7].

**Web usage mining-** is one type mining application used to find out the attractive patterns from the usage of webs. It can also used to provide the requests of web based application.

## III. LITERATURE SURVEY

Mustafa AYDIN at al. effectively distinguish in paper [1] a wide assortment of phishing pages, we removed and investigated various highlights identified with these pages. In this paper creators concentrate on two sorts of characterization classification, to be specific Naïve Bayes and Sequential Minimal Optimization (SMO). These calculations were keep running on the each datasets and utilized for an execution examination. The SMO method demonstrated better execution in both two element choice techniques when it is contrasted with the Naïve Bayes algorithm.

Fadi Thabtah, Mohammad et al. [2] explores some important highlights that are removed from the sites and utilizing that outcome they can choose site authenticity. The principle objective of that paper is to build up a group of highlight that can help to foreseeing phishing sites. Provide the group of extracted features as follows, Address bare based features, Abnormal based features, HTML and java script based features, Domain based features. For each removed component creators gives the rules to recognize the phishing site. Those principles give for the most part three outcomes named as "Legitimate", "Suspicious" or "Phishing". Creators principle is to achieve high accuracy. Creators motive are Request URL, Age of space and HTTPS and SSL feature has more astounding critical in distinguishing phishing sites. Were as, Disabling Right Click , URL having @ image highlight has least frequency .

R. Cooley et al. delineate the theory of the web mining and its procedure in [3]. Creators additionally characterize the scientific classification of the web mining. Paper likewise gives a diagram of apparatuses, strategies and issues related with it.

Right off the bat they partition the web mining in two sections named as Web Content Mining and Web Usage Mining. Moreover creators additionally isolate that part in another sub parts named as Agent Based Approach which is utilized for Information Filtering or Categorization and another part is Database Approach which is utilized for Web inquiry System and Multilevel databases.

V. Y. Kulkarni et al. clarify in the paper [4] give a calculation to recognize phishing sites from the URLs. This algorithm play out Google's blacklist check, Alexa Ranking, Google web index results and number of URLs in light of the component extractions. In the wake of finishing this procedure this will give the ready message like this is legitimate or phishing site. It will give the best execution when the site is old or known site for this framework.

This framework initially accepts URLs as an input. At that point it will look for in the Goggle's blacklist if that URL is found in the rundown than show the ready message generally again look in the Alexa Ranking if the rank is high than it is the unsafe or hunt in the URL based component framework if some element passed than it is protected else it will show the alert message of risk.

On the off chance that there is new URL will come than they check for two sides like Top Level Domain and Misspelled Domain. Framework can work with HTTP just, But when HTTPS URLs will come than framework will confused that how to manage this sort of URLs.

Yang Li et al. prepare a list of Vulnerable sites in [5] and another component recommend as the URL correlation. URL relationship depends on same URLs. They make the rundown on the bases of most viewed sites, most usable locales, most surfing locales, additionally pick the best locales which are typically assaulted and so forth. They pick URLs, whoes PageRank are in the highest point of Google's Page rank.

Two distinct criteria are there named as various separations, same separations. Out of that they characterize the best rate or best coordinating URL. Additionally give the rate of TF(True Positive), TN(True Negative), FP(False Positive) and FN(False Negative).

From the asking of the procedure creator's gather the original URLs and furthermore gather the Phishing URLs of the locales. At that point the typical highlights are removed from them. With the assistance of the classifier like Support Vector machine (SVM) and others they characterize and arrange them for aquiver the better precision. After that the culmination of the stapes same stapes are rehashed however the examination of the string and the information is expected to contrast and last outcome.

B.B. Gupta et al. overview diverse sorts of systems, distinctive apparatuses which can demonstrate the site as phishing in the paper [6]. Phishing life cycle has fundamentally 5 phases named as Planning and setup, Phishing, Break-in/data, Collection of information and Break-out/exfiltration. Creators give the list of datasets, list of Tools, sorts of phishing assaults and characterization of highlights as indicated by the productivity. Likewise give the scientific categorization of phishing location approach.

They additionally short show a few informational collections. (1)APWG's phishing chronicle informational index stores the record of phishing assaults which are accounted for or recognized by the APWG(Anti Phishing

Working Group). (2) Phish tank informational index stores the phishing locales URL detailed by the users.(3) Spamassassin's corpora which is a gathering of messages, fitting for testing spam separating framework. (4) TREC corpus informational collection give the standard evaluation of current and proposed SPAM separating approach.

Creators enroll a few highlights utilized for distinguishing proof of phishing tricks named as body-based highlights, subject-based highlights, URL-based highlights, URL-based highlights, content based highlights and sender-based highlights.

## IV. Different Machine Learning Technique

Different techniques of machine learning techniques like, SVM, SMO, Naive bayes, Fuzzy Logic

### *A. SVM(Support Vector Machine)*

Support vector machines are supervised learning models with associated learning algorithms used to analyze data for classification and regression analysis. It has maily two types of learning SVM named as supervised and unsupervised. Supervised learning is used when data is labeled. Unsupervised learning is used to find natural clustering of the data group. Improvement in the support vector machines is called support vector clustering.

### *B. SMO(Supprt Machine Optimizing)*

Sequential minimal optimization (SMO) is an algorithm for solving the quadratic programming (QP). SMO is used when problem arises during the Support Vector Machine. Support Machine Optimization is mostly used for training support vector machines. LIBSVM tool is used to implement Support Machine Optimizing problems. The SMO algorithm is publish in 1998.

### *C. Fuzzy Logic*

Fuzzy logic is a form of many-valued logic. This has truth values of variables it can be any real number from 0 to 1. It can handle the concept of partial truth. Truth value is fall between the completely true and completely false. Though it has a Boolean logic, the truth values of variables may only be the integer values 0 or 1. When linguistic variables are used, degrees may be managed by specific (membership) functions. Fuzzy logic has been applied to many fields, like control theory and artificial intelligence.

## V. Tools[8]

**MATLAB** – It has in-built math functions. It can get the faster result.

**WEKA –** It is the set of machine learning algorithms. Those sets of algorithms can applicable to databases, after that it can help to preprocess the data, to classify the data and can do the clustering of datasets also.

**Rapid Miner** – Client- server model can implement by Rapid Miner. We can extend the functionalities of tool using plug-ins.

**Open NN** – In the name itself NN is stands for the Neural Network. To implement the neural network it has open source library.

**Rattle** – It is the Open-source software of data mining. It is written in R statistical programming language.

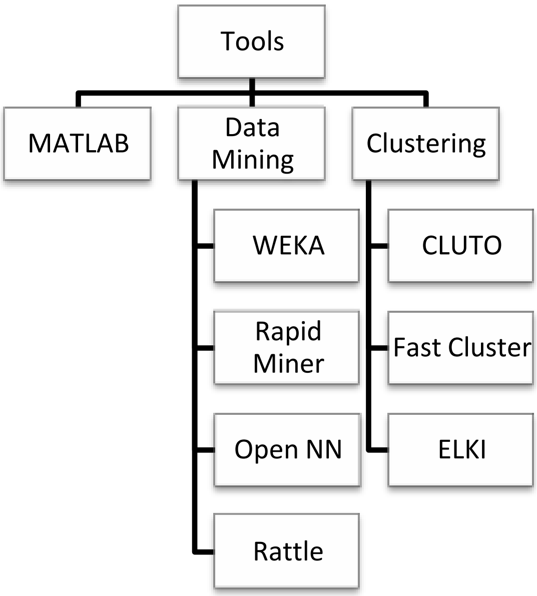

*Fig.2 Tools used for identifying phishing sites[6]*

**CLUTO** – It is a software toolkit. It can be used for analyzed and cluster the high and low dimensional type of data sets.

**Fast cluster** – It can provide the interface to R and Python types of software standards.

**ELKI –** It can support to the indexed structures and KDD-applications. As well as provides data structures such as the R*-tree Environment.

## VI. Statastics

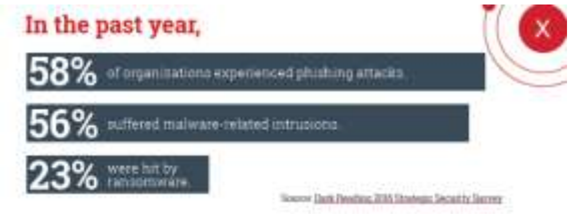

*Fig. 3 Statistics*

A survey of past year(2016) shows some persentage of the attack in which popular organizations and brands suffer due to the phishing attack. Survey is conclude by the company "Dark Reading 2016 Strategic Security Survey" in which highest ratio is 58% in which organizations experienced phishing attack, another 56% ratio is due to malware-related attack, and 23% victims suffer due to ransomware.

The following best practices recommends by Hamer to stay safe online. [8]

- **Do not click on attachments or links in doubtful emails.** If you don't know that the email is legitimate or not then skips the links. Phishing URLs, frequently hyperlinked use friendly language like "click here," and attached documents that may lead to a fake website developed by attackers.
- **Be confident on your intuition.** Sometimes, fake websites and phishing emails can look like the original. Fake emails may even appear to come from a known sender. If something seems surprising, be distrustful: It may possibly be phishing.
- **Do not be scared.** A common phishing approach is lead to, loss of service, or other cost for not acting quickly. Slow down and look at the message carefully. It might be a phishing attack.
- **Reach out when you are in doubt.** If you don't trust an email, go back from that site and open a browser and search in the legitimate website URL to learn more.
- **Don't try to give up your username and password to anyone.** Legitimate organizations, including Harvard IT support staff, will never ask for your password or username, especially via email.

## Conclusion

From last few years' phishing threat increases. Due to that reason we need the protection as well as safety. Different ways are available to overcome this problem but the attackers found another way to create threat. In this paper we can conclude the tools which are used for the phishing as well as some literature survey of the papers. We also suggest some safety precautions for online surfing. In future we will implement different methods or technique to identify the phishing websites.

## References

[1] M. Aydin and N. Baykal, "Feature extraction and classification phishing websites based on URL," in *IEEE Conference on Communications and Network Security (CNS)*, Florence, 2015, pp. 769-770.

[2] Mohammad, Rami, McCluskey, T.L. and Thabtah, Fadi Abdeljaber, "An Assessment of Features Related to Phishing Websites using an Automated Technique," in *International Confrece For Internet Technology And Secured Transactions. ICITST*, London, UK, 2012, pp. 492-497.

[3] R. Cooley, B. Mobasher and J. Srivastava, "Web mining: information and pattern discovery on the World Wide Web," *Proceedings in 9th IEEE International Conference on Tools with Artificial Intelligence*, Newport Beach, CA, 1997, pp. 558-567.

[4] Varsharani Ramdas Hawanna, V. Y. Kulkarni, R. A. Rane, "A novel algorithm to detect phishing URLs," in *International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT)*, Pune, 2016, pp. 548-552.

[5] Ying Xue1 Yang Li1, Yuangang Yao, Xianghui Zhao, Jianyi Liu, Ru Zhang, "Phishing sites detection based on urlcorrelation," in *4th International conference on cloud Computing and Intelligence system(CCIS)*, Beijing,China, 2016, pp. 244-248.

[6] B. B. Gupta, Aakanksha Tewari, Ankit Kumar Jain, Dharma P. Agrawal, "Fighting against phishing attacks: state of the art and future challenges," in *Neural Comput & Applic*, vol. 28(12), pp.3629-3654, Mar 2016.

[7] R. Cooley, B. Mobasher and J. Srivastava, "Web mining: information and pattern discovery on the World Wide Web," *Proceedings in 9th IEEE International Conference on Tools with Artificial Intelligence*, Newport Beach, CA, 1997, pp. 558-567.

[8] Hamer. "Best Practices to Stay Safe Online. Internet: https://news.harvard.edu/gazette/story/2017/11/harvard-expert-offers-best-practices-to-thwart-phishing-attacks/ , Nov. 2017[Nov. 18 2017].