

DETAILED STUDIES ON PHISHING WEBPAGE DETECTION IN DATA MINING

¹Sachin Sharma, ²R L Yadav

¹Dept Computer Science and Engg, Research Scholar, KITE, Jaipur

²Dept Computer Science and Engg, KITE, Jaipur

Abstract—As an application of data mining, businesses can learn more about their customers and develop more effective strategies related to various business functions and in turn leverage resources in a more optimal and insightful manner. This helps businesses be closer to their objective and make better decisions. Data mining involves effective data collection and warehousing as well as computer processing. For segmenting the data and evaluating the probability of future events, data mining uses sophisticated mathematical algorithms. Data mining is also known as Knowledge Discovery in Data (KDD). Data mining technique, one of the classifications of technical approach, has shown promising results in detection of phishing websites. As compared to non-technical approaches, data mining techniques can generate classification models which can make prediction on phishing websites in real-time. Phishing is a form of fraud in which the attacker tries to learn sensitive information such as login credentials or account information by sending. This paper studied about phishing detection and prediction using various efficient algorithms. Classification of phishing is defined in this study and phishing techniques as well.

Keywords—Data Mining, Phishing Detection, Phishing Attack, Phishing Techniques, Classification.

I. INTRODUCTION

Data mining is skill of retrieving information which is meaningful in nature from such large set of data. Data mining as well as discovery of knowledge technique has been functional to many areas those are including analysis of market, industrial retail, and decision support & financial analysis [1]. Phishing websites are the web pages those are fake in nature which are formed through the malicious form of the persons to duplicate the pages of web of the actual websites. Phisher create typically web pages those seem to be of same type to actual web pages so as to scam their form of victims. An unaware user be simply deceived through this type of scam. The phishing Web page Victims may represent the bank account of them, its code word, bank credit card number, /different essential info to the owners of Web page which is of phishing form. While relatively phishing is latest crime of Internet when matched to other forms (for eg, viruses & hacking), an increase of recognizable type in the no: as well as attacks of phishing severity are then informed [2].



Fig. 1: Flow of general phishing attack

In several phishing attacks emails were sent to user. These emails pretend to be from a legitimate organization and redirect the user to a website wherein user enters his or her personal and confidential information such as username, password, credit card details, bank details etc. These websites also called as phishing website now steal the user information. Phishers may sell this information or use it to carry out illegal tasks to obtain personal gain. Phishing attacks are mainly classified into four categories such as Phishing, Spear phishing, Clone phishing and Whaling [3].

II. PHISHING IN DATA MINING

Phishing is defined as fake acquisition of the secret data by recipients which is intended & the data misuse. The attack of phishing is done often by the electronic mail. An eg: of the phishing as if appearance of electronic mail may be as of the actual web-sites, beginning the bank of client, credit card corporation of user, user's electronic mail, /the provider of

Internet service. Usually, confidential info like credit card no/code word then requested to the accounts updating. These electronic mails are used to comprise a linkage of URL which guides a different website client. This site is being fake actually or customized form of the website. As soon as customers visit to this web-site, they are being requested to go into individual info to the phishing attacker those are forwarded. Phishing is being frequently being used to find out user's password /users credit card info. By the assistance of the electronic mail being ready as per if approach in gas of bank /institution which is official, users' processor is being fixed to the unreal web-sites. In the all-purpose, useful data which is taken through the attack of phishing is as:

- account number of clients
- passwords of client & name of client
- Credit card info
- information of the Internet banking

Simulator of Anti-Phishing that intended to avoid threats of solemn type such as, vicious circle malevolent electronic mail those are incoming at the addresses of electronic mail included within system. It as well offers regulate which is in the basis of URL. System includes the keywords in presented database and determines mail contents [4].

In general, attacks of phishing are being performed through below given 4 steps [5]:

- A web site of fake type that seems exactly like Web site which are legitimate is set up by phisher.
- Phisher which then send the link to unreal web site in the huge amount of e-mails of spoofed type to target the users in legitimate companies name & name of organizations, those trying to influence victims of potential type to access their web sites.
- Victims access the web site those are fake by nature by clicking on to link as well as enter its information which is useful.
- Then Phishers steal information of the individual as well as perform their fraud like transferring credit from account of victims.

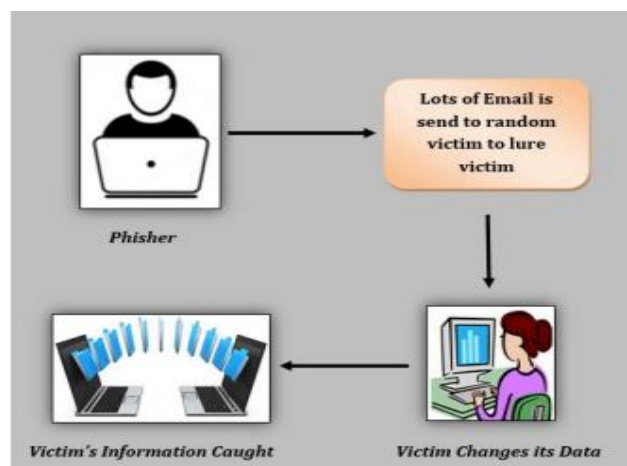


Fig. 2: Process of Phishing

III. PHISHING TECHNIQUES

There are various techniques of phishing being used to acquire information of individual from other users. As technology changes to be more advanced, techniques of phishing which is being required are much advanced. To overcome Internet phishing, customers may have info. of different kinds of ways of phishing as well as they be supposed to be get aware of techniques of anti phishing to secure itself from being phished. Now below discussed few phishing techniques [6]:

1. Email / Spam

Phishers those transfer similar email to several users, by requesting users to load up individual facts. These facts then used by the phishers for activities which are illegal by nature.

2. Web Based Delivery

Web based delivery is phishing method of sophisticated type. Also called as man in the middle attack where the attacker is situated in b/w real website & system phishing.

3. Instant Messaging

Instant messaging is ways into which customer gets a message through the link those to direct them to unreal phishing website that contain similar look & the feel as website which is legitimate.

4. Trojan Hosts

Invisible hackers are Trojan hosts those are demanding to be logged into the customers account to gather the testimonial by local machine. The information is present is after that transferred to the phishers.

5. Link Manipulation

Link manipulation is method into which the phisher forwards a link to the website. When customer clicks on deceptive link, that its opens up website of phisher's in place of website that is informed in link.

6. Key Loggers

Key loggers denotes malware those being required so as to analyze the link from keyboard. The useful data is to transfer the hackers those will make out the passwords & various kinds of the useful data.

7. Session Hacking

In the session hacking, phisher is to exploit the mechanism of web session to get the data from user. In procedure of session hacking called as sniffing of session, the phisher may require a sniffer to the intercept which is relevant to particular information so that they which can admittance illegal form of Web server.

8. System Reconfiguration

Phishers may transfer the message where user is being asked to configure it again the computer settings. The message come through the web address that has resembles of the reliable source.

9. Content Injection

Technique of Content injection of phisher keep on changes content part on page of a website which is reliable. This is to deceive user to go to outside the page of actual website where consumer is needed to put secret information.

10. Phishing through Search Engines

Few kind of phishing scams occupy search engines wherever the consumer is being aimed at to sites of products which might provide products of low cost / services of low cost. When the consumer tries so as to buy product of entering details of credit card, it's being composed by site of phishing.

11. Phone Phishing

In the phone phishing, phisher use to create phone calls to consumers who asks user to call number.

12. Malware Phishing

Involvement of malware by Phishing scams need to be a run on computer user's. The malware is generally being attached to send email so as to user done by the phishers.

13. Anti-Phishing Techniques

Anti-phishing is a method which is employed to detect & avert phishing attack. User gets protected against Antiphishing which is to protect from phishing. To safe oneself against phishing there is need to have antivirus which is installed as well as software of anti-phishing.

IV. CLASSIFICATION OF PHISHING

Some of the important types [7] of Phishing are as follows;

a) Deceptive phishing:

Broadcasting the links of the bogus site through emails, instant messengers with the information of attractive discounts for quick reply, account verification, apologies for system failure with a request to re-enter/reset the account details and attract a good number of people getting tricked to click the link to the fake web site, where their confidential information is collected; this is deceptive phishing.

b) Malware-based phishing:

In this type, malware is sent as an attachment, or hidden into a downloadable file hosted in the website, or injecting into the applications by exploiting their security vulnerability.

c) Hosts File Poisoning:

In the Internet, the web address is translated into its IP address, which gets transmitted to reach the destination. PCs running the Microsoft Windows operating system are more vulnerable to this because, the hostnames are checked into their hosts' file prior to undertaking DNS lookup. Hackers poison this host list, by inserting the address of the fake websites, where information is collected and stolen.

d) Content Injection Phishing:

This is done by way of replacing a part of the website with the content that is injected with the malicious code, which misleads the users to part with their confidential information.

e) Man-in-the middle phishing:

This type of the Phishing is very hard to notice, as the hackers position themselves somewhere in-to between the user & the website of legitimate kind or system. Here the hackers intercept the information, to record it silently without interrupting the flow, so that the users' transactions are not affected. The recorded information will be used later either to gain entry into the server or to sell in the market.

f) Search engine phishing:

Here, the fake websites are indexed legitimately with search engines; this will get them displayed in the top search results, leading to more traffic during the search. This will attract more people to the website, to fool around.

g) Tabnabbing:

An innocent-looking page being opened in a browser tab covers itself as page of login of web application to popular, when user's focus is on various tab (Philippe et al 2013).

The two main classifications of phishing techniques [7] are blacklist based phishing and heuristics based phishing.

A. Blacklist and Whitelist

This is probably very solution of straightforward type for anti-phishing. The list of known phishing sites is known as the Blacklist, and the list of legitimate sites is known as the Whitelist. Most of the current phishing methods rely on a combination of both these lists. The representative of the blacklist or systems based on whitelist includes, SiteChecker, Google safe Browsing, FirePhish (adons.mozilla.org) and CallingId Link Advisor. These anti-phishing solutions work by installing extensions to web browsers or as toolbars, and warn the users whenever they visit a fake site. But the challenge is the time taken to update these lists. Sometimes it is updated, thereby creating a window of vulnerability (Liu et al 2010).

Covering the entire Internet in listing the legitimate websites is highly impossible for the administrator; but it is otherwise required to maintain the global white list. This creates confusion in the minds of the users, when they receive the alert over the genuineness of any website they visit (Weili et al 2012). Everyday, the Phishtank is loaded with hundreds of phishing sites and the company allows the community to verify the same by voting yes or no. This list is not so effective in protecting the users, since the updation happens at different speeds and coverage whereas most of the victims are caught in less than 20% of phishing at zero hour. Collecting and allocating the blacklist is a step by step process (Steve et al 2009).

B. Heuristics based Phishing

Heuristics-based methods employ the common characters of phishing sites, like characteristic keywords being used in the Web pages / URLs to identify new sites of phishing that are not yet listed in blacklists. The heuristic approach recognizes the fraudulent web sites by analyzing the characteristics, similarity and relationship with the legitimate websites. This is possible by the computation of the information collected on the legitimate websites known as feature library (Chen et al 2009).

Visual similarity based phishing: It is the look and feel that deceives the victims. Therefore, the owner of the legitimate website or its agent should be able to compare the suspicious URLs and the corresponding web pages with the legitimate ones in visual aspects.

V. LITERATURE SURVEY

Seow Wooi LIEW et al. [2019] Phishing Social Engineering is a type of delinquency that usages victims to cheat them through pointing a deceitful web-site, wherever their personal & secret info is composed further to the illegitimate activities. The attacks of phishing take now focused on clients at the Online Social Networks (OSN)s, for example, Myspace, Facebook, Twitter, and so on that generally, focusing on electronic mail clients. Due to its tremendous data dispersion and difficult due to not being generally distinguished for email, Twitter has become so turbid for spreading attacks of phishing these days. In that capacity, the viability of security caution to provoke Twitter clients are constantly important for tweets with the Phishing Uniform Resource Locator (URL). Several arrangements have been implemented but their viability is inadequate & doubtful. Within this article, we implement a compelling safety ready system creating utilization of the grouping archetypal got as of a directed AI strategy of the Random Forest (RF) as well as recognized 11 top order highlights produced the 94.75% precision greater than the 94.56% produced through different analysts those utilized in excess of 11 highlights prepared over the equivalent data-set gathered as of Twitter. On the way to decide its adequacy, we utilized the 200 phishing's URLs gathered from the Twitter & Phish-Tank individually. As of our test, we can legitimize that such proposed security ready instrument figured out how to provoke 97.50% successfully the security alarm to twitter clients progressively [8].

Chuan Pham et al. [2018] utilize URL features & web traffic features to spot phishing sites dependent on an expected neuro-fuzzy structure. Contingent upon the new strategy, mist count as invigorated through Cisco, we plan an anti-phishing model to straightforwardly screen and guard fog clients from phishing attacks. The testing results of our proposed strategy, relies upon the huge scale dataset together from authentic phishing cases have demonstrated that our structure can productively avert phishing attacks and get better the safety of the n/w [9].

Shweta Dasharath Shirsat [2018] The aim is to show phishing recognition utilizing fuzzy logic and translating results utilizing distinctive defuzzification strategies. This paper exhibits the identification of phishing site utilizing the fuzzy logic approach with five heuristic parameters(Highly Legitimate, Legitimate, Suspicious, Phished and Highly Phished).This method is effectively executed and tried on a dataset containing 300 unique URLs [10].

Murat Karabatak and Twana Mustafa [2018] phishing site dataset taken from UCI was explored. Its measurement was decreased and the execution correlation of classification algorithms is examined on diminished phishing site dataset. Through the UCI AI store, the phishing site dataset was involved. This dataset has 11055 records and 31 structures. Feature selection algorithms were connected to diminish the component of phishing site dataset and to accomplish high characterization execution. At that point, the execution of classification algorithms is contrasted with other DM classification algorithms. At last, a similar grouping act on the diminished dataset by utilizing the basic characterization algorithms is specified [11].

Mehek Thaker et al. [2018] proposes a framework which will identify old just as recently created phishing URLs that have totally no past practices to make a decision after, utilizing Data Mining. A cloud-based grouping model will be made for the equivalent wherein different extricated qualities through the URL will be utilized as input data. The model will be trained with a thorough dataset in order to guarantee the most extreme precision [12].

Dyana Rashid Ibrahim and Ali Hussein Hadi [2018] develop the capacities of classification procedures on PWP, & acquaint an approach which defends clients from the attackers. The boycott methodology is anything but a sufficient method to remain safe from the cybercriminals. Accordingly, the phishing site indicators should be considered for this reason with the presence and use of machine learning algorithms. To assess their efficiency in terms of RAE value for PWP with accuracy and feature selection process and five different classification techniques have been used for each of them. The WEKA tool was used to classify these classifiers on open datasets from the NASA Store. The motivation to run this test is to use different DM calculations due to phishing sites and to see their accuracy and their viability in the context of RAE, where DM classmates have proved themselves well in such problems [13].

Akansha Priya and Er. Meenakshi [2017] Using the WEKA tool, the C4.5 (J48) DM algorithm former analyzed. C4.5 is a benchmark DM innovation that can precisely recognize phishing sites. A trained dataset of 750 URLs has been made to training the Algorithm J48, which is the usage of the C4.5 algorithm in WEKA. Tested datasets of 300 URLs are utilized to make a forecast utilizing the classifier made following J48 training. After testing process of TPR, TNR, FPR, FNR, success rate, error rate and accuracy are estimated. The consequence illustrates that the accuracy of C4.5 is 82.6% [14].

Shuichiro Haruta, Hiromu Asahina, and Iwao Sasase [2017] Indicator provides visual equality-related phishing detecting systems by pictures & CSS using the scene. To address the first deficiency, we demand the way that genuine sites are frequently connected to different sites and consider such a site to be genuine and store screen captures and CSS in the database. Since CSS is a document that portrays the substance of the site, attackers frequently take real CSS to copy real sites. In this way, by detecting the website that validates the presence of legitimate websites or CSS, we find out the phishing website and its goals simultaneously. In addition, we can reduce the 2 cuts by utilizing CSS in light of the fact that almost certainly, the sites which have an alternate appearance, utilize the equivalent CSS at the neighborhood level. By PC reproduction with a genuine dataset, they improve the accuracy of detecting the plan while searching for a phishing goal [15].

Mahmood Moghimi and Ali Yazdian Varjani [2016] introduce a newer approach which is based on rule for detecting phishing attacks from Internet banking. Here 2 newer set of features were used to verify the identity of the webpages in this approach. For evaluating Identity of Page Resource and identifying Access Protocols of Page Resource Elements uses four features to each from the proposed features set. They utilized surmised sequence coordinating calculations to decide the connection among the substance and the page URL list of capabilities. These features are free from outsider administrations, for example, web indexes result and additionally internet browser history. SVM algorithm utilized to categorize webpages. The analyses demonstrate that the implemented model can distinguish phishing pages in web keeping money with precision of 99.14% TP and just 0.86% FPR. Yield of affectability examination shows the critical effect of proposed features above conventional features. They separated the concealed information from the SVM show by embracing a connected strategy. They implanted the removed principles into a program expansion termed PhishDetector to make proposed strategy progressively useful and simple to utilize. Assessing of the actualized program augmentation shows that it can identify phishing assaults in web managing an account with high precision and unwavering quality. PhishDetector can identify zero-day phishing assaults as well [16].

VI. CONCLUSION

In simple words, data mining is defined as a process used to extract usable data from a larger set of any raw data. It implies analysing data patterns in large batches of data using one or more software. Data mining has applications in multiple fields, like science and research. Information theft or phishing is a security challenge which is usually carried out by sending spoofed emails and spams. Data mining possesses a number of methods for extracting useful information from different unorganized data. Here, we have given detailed studies about phishing detection using data mining with various phishing classifications. Since several works has been done in this field which is discussed in related work but

still some issues have remained on that work will be done. Also, different kinds of phishing techniques have discussed in the paper.

REFERENCES

- [1] Divya V , Vivitha Vijay, “Phishing Website Detection Using Associative Classifiers”, International Journal on Recent Researches in Science, Engineering & Technology (IJRRSET), Volume 5, Issue 11 November 2017.
- [2] Hemali Sampat, Manisha Saharkar, Ajay Pandey, Hezal Lopes, “Detection of Phishing Website Using Machine Learning” , International Research Journal of Engineering and Technology (IRJET), Volume: 05 Issue: 03 | Mar-2018.
- [3] Anti-Phishing Working Group (APWG), “Phishing activity trends report—first quarter 2013.<http://antiphishing.org/reports/apwgtrendsreportq12013.pdf>, accessed September 2014.
- [4] P. Liu and T. S. Moh, "Content Based Spam E-mail Filtering," 2016 International Conference on Collaboration Technologies and Systems (CTS), Orlando, FL, pp. 218-224, 2016.
- [5] Hicham Tout, William Hafner “Phishpin: An identity-based anti-phishing approach” in proceedings of international conference on computational science and engineering, Vancouver, BC, pages 347-352, 2009.
- [6] <http://www.phishing.org/phishing-techniques/>.
- [7] Gaurav Varshney, Manoj Misra and Pradeep K. Atrey, “A survey and classification of web phishing detection schemes”, Security Comm. Networks 2016; 9:6266–6284 Published online 26 October 2016 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/sec.1674
- [8] LIEW, S. W., SANI, N. F. M., ABDULLAH, M. T., YAAKOB, R., & SHARUM, M. Y. (2019). An Effective Security Alert Mechanism for Real-Time Phishing Tweet Detection on Twitter. Computers & Security, pp. 1-14.
- [9] Pham, C., Nguyen, L. A. T., Tran, N. H., Huh, E.-N., & Hong, C. S., “Phishing-Aware: A Neuro-Fuzzy Approach for Anti-Phishing on Fog Networks”, IEEE Transactions on Network and Service Management, 2018, pp. 1–14.
- [10] Shirsat, S. D., “Demonstrating Different Phishing Attacks Using Fuzzy Logic”, Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, pp. 57-61.
- [11] Karabatak, M., & Mustafa, T., “Performance comparison of classifiers on reduced phishing website dataset”, 6th International Symposium on Digital Forensic and Security (ISDFS), 2018, pp. 1-5.
- [12] Thaker, M., Parikh, M., Shetty, P., Neogi, V., & Jaswal, S., “Detecting Phishing Websites using Data Mining”, Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2018, pp. 1876-1879.
- [13] Dyana Rashid Ibrahim and Ali Hussein Hadi, “Phishing Websites Prediction Using Classification Techniques”, International Conference on New Trends in Computing Sciences, 2017, pp. 133-137.
- [14] Akansha Priya and Er. Meenakshi, “Detection of phishing websites using C4.5 data mining algorithm”, 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), 2017, pp. 1468 – 1472.
- [15] Haruta, S., Asahina, H., & Sasase, I., “Visual Similarity-Based Phishing Detection Scheme Using Image and CSS with Target Website Finder”, GLOBECOM 2017 -IEEE Global Communications Conference, 2017, pp. 1-6.
- [16] Moghimi, M., & Varjani, A. Y. (2016). New rule-based phishing detection method. Expert Systems with Applications, 53, 231–242.