International Journal of Early Childhood Special Education (INT-JECSE)
DOI:10.9756/INTJECSE/V14I5.885 ISSN: 1308-5581 Vol 14, Issue 05 2022

An Overview of Fraude Usage and Phishing Attacks over the Internet

Karim Khcheesh Ibrahim, Ahmed J. Obaid

Faculty of Computer Science and Mathematics / University of Kufa

Abstract

This paper discussed the Web usage data extraction process, starting with data collection and cleaning, pattern discovery, and analysis. The weblog contains a lot of unnecessary data. The focus is on data related to user behavior, and URLs. This is illustrated with examples of URLs, and the most important topic of the study (fraud) is introduced. Focus on the most important type, which is phishing has two main types, either social engineering attacks or technical stunt attacks, and each type has many aspects of the attack. With regard to countering phishing sites, the most important method for phishing detection, the first is to train users to be aware of phishing attacks. The second is to counter the attacks technically. Then the characteristics of scam sites were explained, in terms of most of the features that these sites have. Deception detection methods are the topic that occupies the bulk of the class.

1. Introduction to Data collection & Preprocessing (Behavior User)

Most web usage mining systems use weblog data as the primary source of data. The weblog file records activity information when a web user places an order to the webserver. The log file can be placed in three Various places: 1) web servers, 2) web proxy servers, And 3) client browsers. In our research we used a data set collected from different sources. After performing the data cleaning process Data is ready for subsequent pre-processing stags [1].Pre-Processing it is a very important step in the use of web usage mining. is a stage consisting of several steps starting from User and Session Identification, Path Completion , Pattern Discovery and Pattern Analysis. as follows as follows [2].

1.1 Data Cleaning

The following figure is a sample of a weblog file, containing outlier data. It contains a large number of data not important in detecting phishing or web usage mining.

http://www.phishtank.com/phish_detail.php?phish_id=7232332 combuyit.pw/unlock20583273,http://www.phishtank.com/phish_detail.php?phish_id=7232325,http://www.phishtank.com/phish_detail.php?phish_id=7232297http://www.phishtank.com/phish_detail.php?phish_id=7232264,https://steancommunytiu.ru/profiles/76561198848491530,www.phishtank.com/phish_detail.php, phish_id=7232222http://anaozn.arabms.com/signin/?openid.pape.max_auth_age=0&openid.return_to=https:/www.amazon.co.jp/?ref_=nav_em_hd_re_signin&openid.identity=http:/specs.openid.net/auth/2.0/identifier_select&openid.assoc_handle=jpflex&openid.mode=checkid_setup&key=a@b.c&openid.claimed_id=http:/specs.openid.net/auth/2.0/identifier_select&openid.ns=http:/specs.openid.net/auth/2.0&ref_=nav_em_hd_clc_signin,http://www.phishtank.com/phish_detail.php?phish_id=7232221,2021-07-17T14:07:18+00:00,yes,2021-07-17T14:16:56+00:00,yes,Other7232220,

Figure Error! No text of specified style in document.**1 Log Data**

From unnecessary records that are not related to our work. Like not modified messages. This is determined when the visitor requests any content, the server sends a status code according to the data, and according to the type of this code, data cleaning is performed Entries containing images, graphics, etc. (jpg, jpeg, gif) are also deleted. All icons greater than 299 and less than 200 are removed. Because it is considered invalid Table shows some of the status codes [3].

Table.1 HTTP server status codes

Code	Description	Code	Description
200	OK	400	Bad Request
201	Created	401	Unauthorized
202	Accepted	403	Forbidden
301	Moved Permanently	404	Not Found
303	See Other	410	Gone
304	Not Modified	500	Internal Server Error

307	Temporary Redirect	503	Service Unavailable

1.2 User and Session Identification:

It is not important to know the identity of the visitor. But there is a need to uncover and characterize visitor behavior. The server records multiple client sessions, and as a visitor can visit certain sites frequently. This is done without any authentication mechanisms, in many web servers, because some users disable the cookies feature. because it shows the private information. This results in the IP address alone is insufficient to identify the unique visitor. Therefore, other criteria are used with the IP address such as user agent and referrer [3]. As for defining the user's session, define it. A group of pages visited by the same visitor in a certain period of time. On some webservers the session time is limited to 30 minutes. After this period, the second session begins.

1.3 Path Completion

It is an important step in pre-treatment. Mostly, it takes place after completing the session. The agent or client is mostly due to temporary caching to the loss of access references for some pages [4]. When the real URLs are more than recorded in the server log. This indicates a loss of access to references to these pages. It is possible to reveal this when a visitor requests a specific page that is not related to the previous page (the previous request) for the same visitor. The referrer can refer to the register to find out which page the above application contains. If the missing page is in the visitor's last, click log. It is a page not registered in the registry. This state indicates that the visitor browsed again using the (Back) button [5].

1.4 Pattern Discovery

It is one of the most important parts of Web Usage Mining extracting data from the web history. The main goal of this part is to discover interesting patterns [6].

1.4.1 Statistics

It is an important technique for finding useful information for any weblog. And to know the content of the web history and the number of visits to clients in that record. The number of visits is calculated on the basis of each valid entry in the weblog. Because it

helps improve system performance such as monitoring visitor activities, monitoring and checking pages and sites, and aggregating visitors based on their behavior [7].

1.4.2 Association Rule

This technique is used to find recurring rules and patterns in the data generated from the preprocessing stage of the weblog data such as the number of users' frequent visits to certain pages. The task of this technology is to understand the visitor's requirements. This is done by discovering the relationships between the pages visited by a particular visitor to a specific website. Several algorithms are used as Apriori algorithm, to find recurring association rules [8].

1.4.3 Clustering

Clustering is a method used to group certain elements (pages, users, etc.). Based on similar characteristics such as a grouping of webpages with similar content or Clustering a group of visitors with similar browsing behavior or collecting many users who visit similar sites and others. It is also possible to use the normalization process with Clustering. This will give better combination results, because there are different ranges in the data point for each area. Clustering technology helps in deducing customer stats in e-commerce market operations and provide customized web content based on individual visitors. Also, Clustering is useful in making indexes of websites on the Internet [7].

1.4.4 Classification

This technique categorizes data elements into distinct, predefined categories, which are related to a specific category. This technology requires extracting and selecting the distinct classes on which the classification. Then, the classification process is performed [7].The main goal of categorizing weblog data is to develop the log for visitors belonging to a specific category as opposed to aggregation, fraud detected and etc. Because classification is a directed learning method (supervised learning). Of the algorithms used by this technique naïve Bayesian classifiers, decision tree, Random Forest, etc. [9].

1.4.5 Sequential Pattern

It is the conduct of analysis to find patterns in sequence through serial sessions and by applying several algorithms such as SPADE, Apriori, etc. For example, a specific

visitor. visited link A and then link B one by one at the same time. Using analysis for such a pattern, we can predict the suspected visitor. When visited in a pattern similar to the previous one. Through psychology used to uncover crime, predict shopping, advertisements, fraud, etc. [7].

1.5 Pattern Analysis:

It is the last stage of web usage mining. In this respect, knowledge is found in the detected patterns. Interesting patterns. And that by getting rid of inappropriate patterns. This can be done by applying a validation rule. To get rid of inappropriate patterns and uncover appropriate patterns [10]. A commonly used technique in pattern analysis is the OLAP (Online Analytical Processing Technique). Visualization methods, administrative advertising deals, it uses graphic patterns to interpret the file results in an easier way. As well as (the mechanism of using the knowledge query SQL). Which is used to analyze several reasons for the abnormal patterns of your visitors like a fraud [7].

1.6 Advantages and Disadvantages of WUM:

Table Error! No text of specified style in document.**:** Comparison Advantages Vs Disadvantages of WUM

Advantages of WUM[32]	**Disadvantages of WUM[32]**
1- WUM technology is used by government agencies and others to classify and combat terrorist threats.	1- When using WUM on personal information, some concerns and negative consequences appear.
2- Through WUM companies can better understand the requirements of their visitors. It provides the fastest response to visitor requests.	2- A concern for users when some companies collect data from People for a specific purpose, such as a job or business. Because some companies sell personal data.
3- It helps companies a lot in attracting and retaining useful customers who can save the best production costs.	3. Classifies individuals based on controversial characteristics such as race, sexual orientation, or religion.

1.7 Data collection & Pre-processing (phishing URLs)

The selection and collection of data samples was a critical stage in this study. First, I selected the appropriate data sources for collecting and selecting domain data on the Internet. After selecting these sources. Automated scripts were used to download datasets, and to extract Features of domains, paths, and Features website URLs. The selection process was based on Characteristics of domains and certificates. These characteristics are noted While generating the dataset and introduction patterns about phishing/legitimate Categories. These patterns made it easy to analyze how attackers create phishing sites[11]. I tried to use all the freely available data in this study.

1.7.1 Characteristics of Domains

1.7.1.1 **Top-level domains:** \.com", \.net", \.org", \.uk",\.ca" from the characteristics that are found a lot in the legitimate domains \.site", \.xyz", \.icu", \tk", \.online", \.live" were observed in phishing domains. These top-level phishing domains are confirmed by Proofpoint Domain Fraud Report. Although these TLDs of phishing domains are also used in legitimate domains but frequency is quite less

1.7.1.2 Length

We often find phishing areas to be long. For example, {"\ applied support- update-account-supportupdate42299codeanyapp.com"} whereas for legitimate domains, short length was observed. Sub-domains were observed in phishing domains.

1.7.2 Features Extraction

After analyzing the related information in domains, URLs, and paths, to identify fraudulent user behavior, 81 features were considered. Hopefully, these features contain some content that indicated to how attackers created phishing domains. Features are categorized as text, integer, and Boolean. For some features, a separate script has been implemented to create features for all phishing and legitimate data saved in a setup CSV file System data set. Domain, Port, Host Type, Query, Having IP ,Having Subdomain ,URL Length ,URL Length Threshold ,URL Depth ,Redirections ,SSL Type Shortening Services ,Prefix & Suffix , URL Have Sign -, _, /,?, =, &,!, ~, +, *, # $, %, @, Domain Have Sign -,Domain Have Sign _FTC the symbol ("_") in a Domain, Domain Have

Sign /FTC the symbol ("/") in a Domain. , Domain Have Sign ? FTC the symbol ("?") in a Domain. Domain Have Sign = FTC the symbol ("=") in a Domain. , Domain Have Sign & FTC the symbol ("&") in a Domain. , Domain Have Sign ! FTC the symbol ("!") in a Domain. , Domain Have Sign FTC the number (" ") in a Domain ,Domain Have Sign , FTC the number (",") in a Domain. , Domain Have Sign ~ FTC the number ("~") in a Domain .,Domain Have Sign + FTC the symbol ("+") in a Domain. Domain Have Sign * FTC the symbol ("*") in a Domain., Domain Have Sign # FTC the symbol ("#") in a Domain. ,Domain Have Sign $ FTC the symbol ("$") in a Domain. ,Domain Have Sign % FTC the symbol ("%") in a Domain. ,Domain Have Sign @ FTCH domain contains)"@"(symbol. ,Path Have Sign . FTCH path contains (".")symbol. ,Path Have Sign _ FTCH path contains ("_")symbol. Path Have Sign / FTCH path contains ("/")symbol .,Path Have Sign ? FTCH path contains ("?")symbol. ,Path Have Sign = FTCH path contains ("=")symbol. ,Path Have Sign & FTCH path contains ("&")symbol .,Path Have Sign ! FTCH path contains ("!")symbol .Path Have Sign FTCH path contains (" ").,Path Have Sign , FTCH path contains (",")symbol. ,Path Have Sign ~ FTCH path contains (".")symbol .,Path Have Sign + FTCH path contains ("+")symbol., Path Have Sign * FTCH path contains ("*")symbol .Path Have Sign # FTCH path contains ("#")symbol. ,Path Have Sign $ FTCH path contains ("$")symbol. ,Path Have Sign % FTCH path contains ("%")symbol. ,Path Have Sign @ FTCH path contains ("@") symbol. And other important features in identifying phishing sites.

1.7.3 Data Normalization

Normalization is the process of creating numeric values in a data set on a common scale. In this thesis, datasets that contain text data and integers were used in very different ranges. In order to normalize these two classes, I used them. MIN-MAX normalization method .This normalization method rescales the range of features between [0,1]. The main advantage of data usage normalization was to increase numerical stability and reduce training time and also increasing the accuracy of classification models .Use the following equation to calculate the MIN-MAX Normalization[12]. $Z = \frac{x - \min(x)}{\max(x) - \min(x)}$ (2.1)

In the figure (2.2), it is shown how the values of the linear data in the vector were changed from values of large dimensions to values that are close and of higher accuracy.

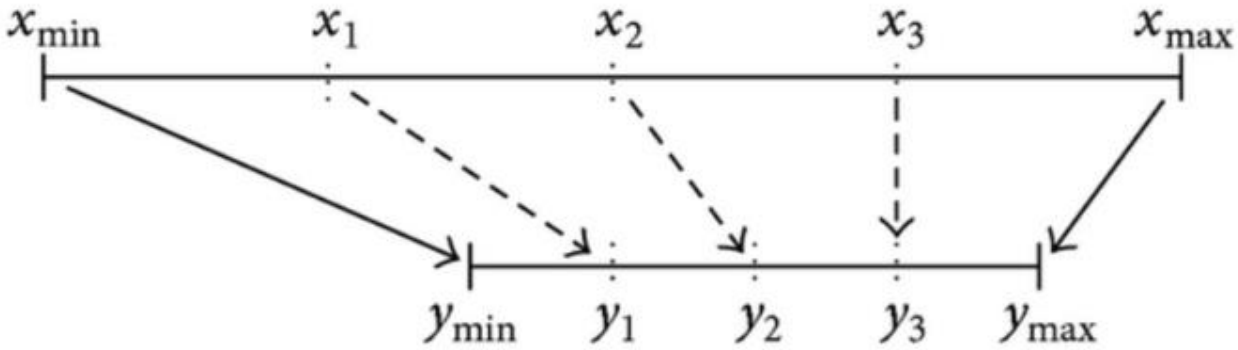

Figure 2: Min.-max. method of normalization[35]

1.7.4 Feature Selection

The process of feature selection is to identify the most useful features in a data set in terms of a particular format or access to a particular type of data such as detecting user requirements in a field or detecting phishing in websites and others. Choosing features from a very large data set is important because it reduces data volume, reduces training speed, and increases model accuracy. Therefore, some software is used in this technique, such as Weka which is effective in it. Some features that are unnecessary are removed from the data set. Weka offers an option to add and remove features. Like features that have a unique value give one value for all cases it is given a zero. These features will not be useful in machine learning. As for the features that are important, removing them will have a significant impact on the performance of the model.

1.8 Types of URL Fraud

there are five different types of URL fraud, where URLs are hidden through the process of keyword shuffling in paths, queries, and low-level domains are listed with examples in Table (2.6): [13].

1: Obfuscation of other areas

2: Complication with keywords

3: Typo-squatting domains

4: Obfuscate the IP

5: Complication with URL shorteners

Table 2: Samples of URLs

Samples	Examples
1	http://school511.ru/333/www.paypal.com/29546374287905815
2	http://quadrodeofaster.com.br/www1.paypal-com/encrypted/ss4578
3	http://cgi-6.paypalsecure.de/info5/kdgvnchdit.html
4	http://69.25.415.96/javaseva/https://paypal.com/uk/twopagepaypal.htm
5	http://goo.g1/HQx7h

References

[1] P. Verma and N. Kesswani, “Web Usage mining framework for Data Cleaning and IP address Identification,” 2014.

[2] HTUN ZAW OO, “Pattern Discovery Using Assocation Rule Mining on Clustered Data Htun Zaw oo Me,” no. August, p. 55, 2018.

[3] T. A. Al-Asadi and A. J. Obaid, “Discovering similar user navigation behavior in web log data,” International Journal of Applied Engineering Research, vol. 11, no. 16. pp. 8797–8805, 2016.

[4] M. Gayatri, “Review of Current Trends in Web Usage Mining,” no. October 2018, 2020, doi: 10.14419/ijet.v7i3.20.22972.

[5] Yehia Helmy, “Empirical-Study-of-Data-and-Web-Mining-in-Education.pdf.” International Journal of Scientific & Engineering Research, p. 8, 2020.

[6] M. J. Hamid Mughal, “Data mining: Web data mining techniques, tools and algorithms: An overview,” Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 6, pp. 208–215, 2018, doi: 10.14569/IJACSA.2018.090630.

[7] B. Bhavani, V. Sucharita, and K. V. V. Satyanarana, “Review on techniques and applications involved in web usage mining,” International Journal of Applied Engineering Research, vol. 12, no. 24. pp. 15994–15998, 2017.

[8] A. J. Obaid, A. S. Abdulbaq, S. A. Najy hilmi and S. A. AL-Ameedee, "Role of Information Technology in Structuring and," International Journal of Early Childhood Special Education, vol. 14, no. 3, pp. 567-573, 2022. 10.9756/INT-JECSE/V14I3.72

[9] Y. YANG, “Effective Phishing Detection using Machine Learning Approach Case Western Reserve University Case School of Graduate Studies,” p. 85, 2019.

[10] Anandkumar R, Dinesh K, Ahmed J. Obaid, Praveen Malik, Rohit Sharma, Ankur Dumka, Rajesh Singh, Satish Khatak,Securing e-Health application of cloud computing using hyperchaotic image encryption framework, Computers & Electrical Engineering, Volume 100, 2022, 107860, ISSN 0045-7906, https://doi.org/10.1016/j.compeleceng.2022.107860.

[11] Z. Alkhalil, C. Hewage, L. Nawaf, and I. Khan, “Phishing Attacks: A Recent Comprehensive Study and a New Anatomy,” Front. Comput. Sci., vol. 3, no. March, pp. 1–23, 2021, doi: 10.3389/fcomp.2021.563060.

[12] A. J. Obaid, T. Chatterjee and A. Bhattacharya, "Semantic Web and Web Page Clustering Algorithms: A Landscape View," EAI Endorsed Transactions on Energy Web, vol. 8, no. 33, 2020.

[13] F. Aburub and S. Alhawari, “A new fast associative classification algorithm for detecting phishing websites,” Appl. Soft Comput. J., vol. 48, pp. 729–734, 2016, doi: 10.1016/j.asoc.2016.08.005.

[14] I. V. I. P. A. V Preethi and G. Velmayil, “Automated Phishing Website Detection Using URL Features and Machine Learning Technique.”

[15] P. Patil and P. P. R. Devale, “A Literature Survey of Phishing Attack Technique,” vol. 5, no. 4, pp. 198–200, 2016, doi: 10.17148/IJARCCE.2016.5450.

[16] Research in Intelligent and Computing in Engineering. Advances in Intelligent Systems and Computing, vol 1254. Springer, Singapore. https://doi.org/10.1007/978-981-15-7527-3_36

[17] Ahmed J Obaid, A. S. A. (2022). Status, Challenges, and Future Views of DeepFake Techniques and Datasets. Mathematical Statistician and Engineering Applications, 71(2), 225 –. https://doi.org/10.17762/msea.v71i2.81.

[18] Saeed, , M.M., Hasan, , M.K., Obaid, , A.J., Saeed, , R.A., Mokhtar, , R.A., Ali, , E.S., Akhtaruzzaman, , M., Amanluo, , S., Hossain, , A.K.M.Z.: A comprehensive review on the users’ identity privacy for 5G networks. IET Commun. 00, 1– 16 (2022). https://doi.org/10.1049/cmu2.12327