ISSN (ONLINE): 2395-695X
ISSN (PRINT): 2395-695X

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Volume 2, Special Issue 19, October 2016

A Study on Detection of Intelligent Phishing url using Association rule mining

Mrs. B. Umamaheswari, working as a Assistant professor, Department of Computer Science, P.S.G College of Arts and Science, Coimbatore, Tamil Nadu, India.

Dr. P. Nithya, working as a Assistant professor, Department of Computer Science, P.S.G College of Arts and Science, Coimbatore, Tamil Nadu, India.

Miss.Nair Sarika Chandran, Pursuing M.Phil Research Scholar, Department of Computer Science, P.S.G College of Arts and Science, Coimbatore, Tamil Nadu, India.

Abstract :

Phishing is an online criminal act that occurs when a malicious webpage impersonates as legitimate webpage so as to acquire sensitive information from the user. Phishing attack continues to pose a serious risk for web users and annoying threat within the field of electronic commerce. This paper focuses on discerning the significant features that discriminate between legitimate and phishing URLs. These features are then subjected to associative rule mining—apriori and predictive apriori.

The rules obtained are interpreted to emphasize the features that are more prevalent in phishing URLs. Analyzing the knowledge accessible on phishing URL and considering confidence as an indicator, the features like transport layer security, unavailability of the top level domain in the URL and keyword within the path portion of the URL were found to be sensible indicators for phishing URL. In addition to this number of slashes in the URL, dot in the host portion of the URL and length of the URL are also the key factors for phishing URL.

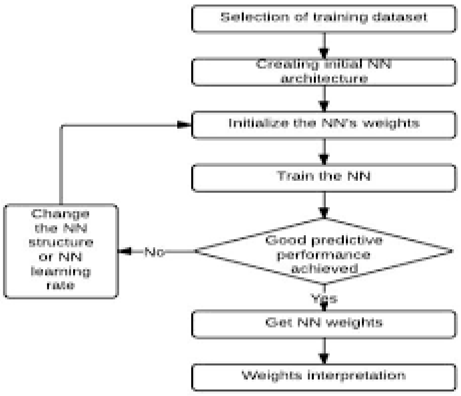
Keywords : Phishing, Web security, Association rule mining.

Introduction :

Phishing is a malicious website that impersonates as a legitimate one to get sensitive data like credit card number or bank account password. A phisher uses social engineering and technical deception to fetch private information from the web user. The phishing web pages generally have alike page layouts, blocks and fonts to mimic legitimate webpages in an endeavor to influence web users to obtain personal details such as username and password. Over the last few years, online banking has become very popular as more financial institutions have begun to offer free online services. With the increase in online theft, financial crimes have changed from direct attacks to indirect attack. Phishing [1] is a quickly growing type of fraud and is taken into account as one of the foremost dangerous threats within the web which cause folks to mislay guarantee [2] in on-line transactions. It is relatively a current web crime as compared with virus, hacking and remains an ominous threat to client and business round the world. According to the RSA's online fraud report [3], the year 2013 has been confirmed to be a record year where many phishing attacks have been launched globally. Additiona1ly, RSA estimates that over USD $5.9 billion was lost by global organizations due to phishing attacks at the same period. The Internet Security Threat Report 2014 [4] reports that cybercrimes are prevailing and damaging threats from cybercriminals still emerge over businesses and customers. According to RSA monthly fraud report January 2014, the [5]big data analytics and broader intelligence will lead to faster detection resulting in lower financial losses. Data mining techniques are used to extract helpful information by ana- lyzing the past information then predicting the future incidents. In this paper, the rules are generated using association rule mining to detect phishing URL. The remaining section in the papers is organised as follows: The outline of literature survey is shown in second section. The system architecture is illustrated in third section. The features that are generated from the URL are discussed in fourth section. Fifth section explains the methodology used in detecting phishing URL. Sixth section presents an association rule mining technique to discover the rules concerning phishing URL and in the last section conclusions are presented.

System architecture for prediction of phishing URL :

Figure 1 shows the system architecture for detecting phishing URL. The foremost objective of this system is to identify an URL that is provided as input as a phished URL or not. The proposed method consists of two phases (1) URL search phase and (2) feature

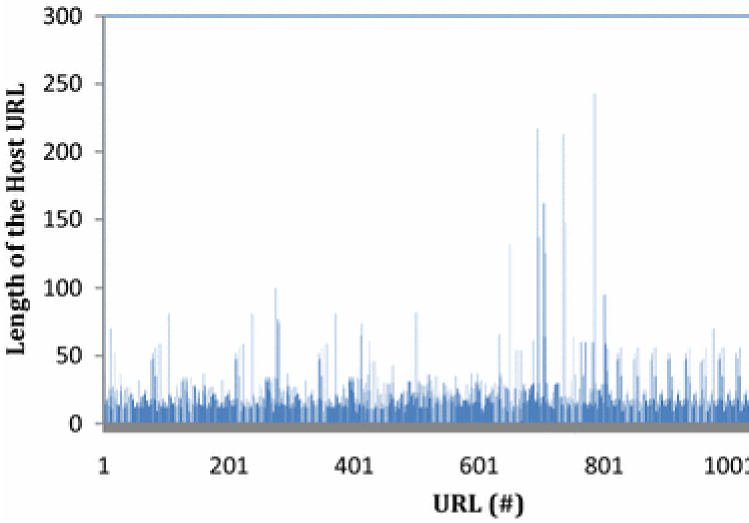

Selection of training dataset
Creating initial NN architecture
Initialize the NN's weights
Train the NN
Good predictive performance achieved
No
Change the NN structure or NN learning rate
Yes
Get NN weights
Weights interpretation

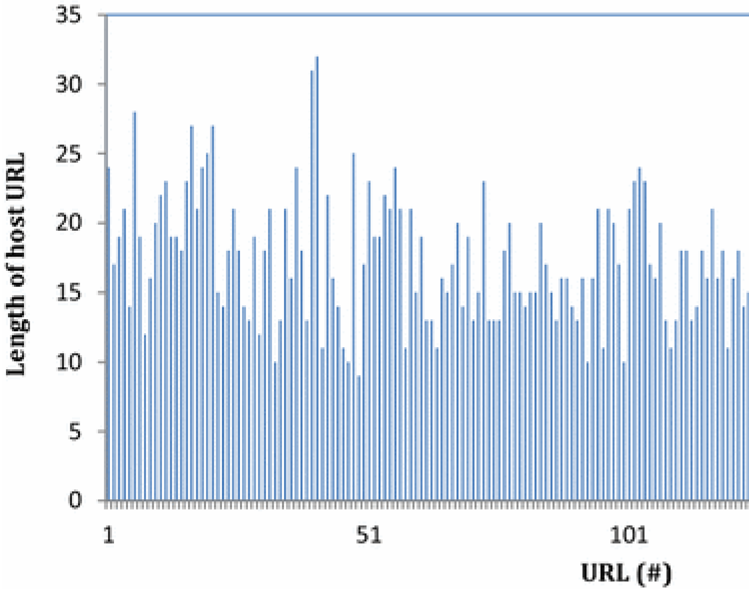
Feature extraction :

The proposed work focuses on identifying the relevant features that differentiate phishing websites from legitimate websites and then subjecting them to association rule mining. In order to identify the relevant features, certain statistical investigations and analysis were carried out on the phish tank and legitimate dataset. Based on the heuristics, fourteen features were defined and are subjected to association rule mining to effectively determine the legitimate and phished URL.

Heuristic 1: length of the host URL :

URL is a formatted text string utilized by internet users to recognize a network resource on the Internet. URL string consists of three elements such as network protocol, host name and path. For a given URL, the host name is extracted and host name length is examined. For the input data set (1200 phishing URLs and 200 legitimate URLs), domain name length is analyzed for phishing and legitimate URLs. The distribution of the domain name length for phished URL is plotted in Fig. 2 and the average length of the domain name (⌊) in phishing URL is found to be greater than 25 characters. The distribution of the domain name length for legitimate URL is plotted in Fig. 3 and the average length of the domain name (⌊) in legitimate URL is found to be 20 characters.



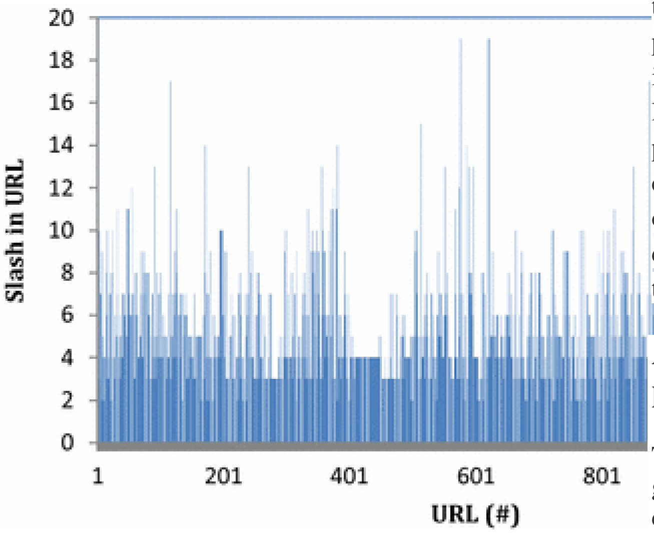


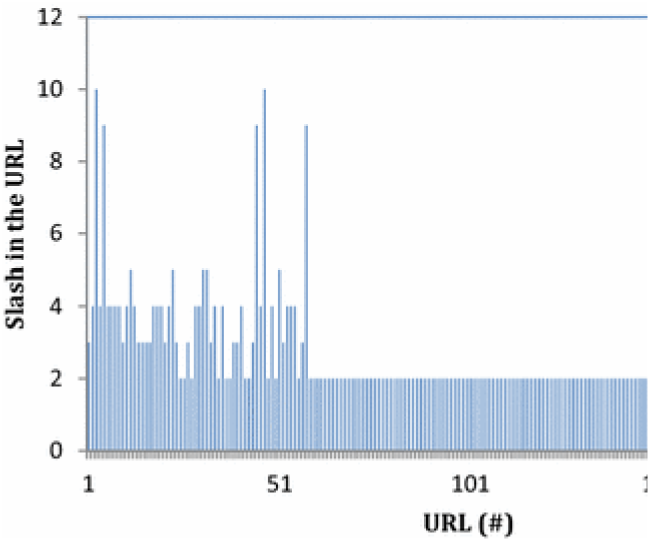
Therefore, the heuristic is defined as

$$H_1 = \begin{cases} if\ length(host) > \lfloor \rightarrow Phishing\ URL \\ \quad else, Legitimate\ URL \end{cases}$$

Heuristics 2: number of slashes in URL :

The phishers try to trick web users by mimicking the doubtful URL look legitimate. One such technique used in scamming is the addition of slashes in URL. The present study, therefore, considers the number of slashes in URLs as a feature of identification of phishing and examines the number of slashes (μ) in legitimate and phishing URLs. For the input data set (1200 phishing URLs and 200 legitimate URLs), the number of slash is analyzed for phishing and legitimate URLs. The distribution of number of slashes in the phishing URLs and legitimate URLs are analyzed and are plotted in Figs. 4 and 5. The result shows that the average number of slashes (μ) in phishing URL is found to be greater than or equal to five and the average slash (μ) in legitimate URL is found to be three.





Therefore, the heuristic is defined as

$$H_2=\{\text{if(SlashinURL)}\geq\mu\rightarrow\text{Phishing URLelse,LegitimateURL}$$

Rule extracted from predictive apriori

The detailed study and analysis of phishing URL were carried out and the results indicate that in predictive apriori algorithm few different rules are generated apart from apriori. The rules generated by predictive apriori are based on accuracy. The result obtained from predictive apriori is shown in Table 4. The strong rule generated by the predictive apriori with accuracy level above 99 % has been considered for further analysis and the other rules are discarded. An investigation on the itemsets reveals that the features like transport layer security, unavailability of the top level domain in the URL and keyword within the path portion of the URL were found to be sensible indicators for phishing URL. In addition to this number of slashes in the URL, dot in the host portion of the URL and length of the URL are also the key factors for phishing URL. Other features such as special characters in the URL, Unicode in the URL, length of the URL is greater than 75 and more than four dots in the host name of the URL were also found to be significant features of phishing URL.

Association rule mining to detect phishing URL :

The process of identifying the type of a URL is generated using association rules in which the different heuristics are utilized to acquire unknown knowledge. This rule is used to ascertain the URL type when a user accesses it. We have recognized different heuristics extracted from the URLs and collected over 1400 URLs from several sources. Legitimate URLs is acquired from five sources and is shown in Table 2 and we collected 1200 phishing URLs from phishtank database. The feature extraction is implemented in PHP. The experiments were performed using WEKA. WEKA a data mining tool incorporates collection of machine learning algorithms. The experiments have been performed with apriori and predictive apriori rule generation algorithms. The experiment is done to discover the rules based on phishing URLs. Detail of this experiment is provided in the following subsections.

Conclusion :

In this paper, the features of the URL are analyzed and are subjected to associative rule mining—apriori and predictive apriori. The rules obtained are interpreted to emphasize the features that are more prevalent in phishing URLs. The results obtained from rule mining have highlighted the useful features available in the phished URL. Analyzing the information available on phishing URL and considering confidence as indicator, the features such as transport layer security, unavailability of the top level domain in the URL and keyword within the path portion of the URL were found to be sensible indicators for phishing URL. In addition to this number of slashes in the URL, dot in the host portion of the URL and length of the URL are also the key factors for phishing URL.

References :

1.https://ers.trendmicro.com/guide/en_us/AG/App A/Phish_Attack.htm. Accessed May 2015.

2. Zhang Y, Hong JI, Cranor LF (2007) CANTINA: a content based approach to detecting phishing web sites. In: Proceedings of the 16th international conference on world wide web, Banff, p 639–648.

3. Huang H, Qian L, Wang Y (2012) A SVM based technique to detect phishing URLs. Int Technol J 11(7):921–925Google Scholar.

4. Liu W, Deng X, Huang G, Fu AY (2006) An antiphishing strategy based on visual similarity assessment. IEEE Computer Society 1089-7801/06 IEEE, IEEE Internet Computing

5. Shah R, Trevathan J, Read W, Ghodosi H (2009) A proactive approach to preventing phishing attacks using Pshark. In: Sixth international conference on information technology: new generations. IEEE, Las Vegas, pp 915–921.