An Intelligent Rule Based Phishing Website Detection Model

Ogunmodimu Dupe Catherine
Department of Computer Science and Informatics
Federal University Otuoke
Otuoke, Nigeria.
dupefatoba@yahoo.com

Abstract

Phishing is a very serious challenge in web-based security these days. Phishing websites are forge web pages that are designed by noxious individuals to emulate pages of genuine sites. Phishers regularly make website pages that looks outwardly like the authentic and original site pages used in deceiving their victims. An uninformed internet customer is likely to be effortlessly tricked unknowing that this is a scam. Phishing webpages victims may uncover their financial details, secret key, and other sensitive data to the proprietor of phishing webpage. This research work is exceptionally significant in that it aimed at developing a model that will give security to users from phishers fraudulent tricks, and enable the users to distinguish the authenticity of websites. The benefit of this research work is to enable government institutes and financial organization to render variety of secured financial services to their various customers. The proposed model used association rule mining algorithm called Predictive Apriori to address phishing website issues that has been a global threat. The algorithm is proficient in analyzing and extracting phishing URL features and is able to perform URL feature classification to identify phishing websites. It then flags the website as either legitimate or malicious. Therefore the user will be able to open only legitimate links. Systems Analysis and Design Method and Prototype methodologies were applied during this research. Our results showed that the algorithm used in this research work generates more accurate predictions than that of the existing system.

Key Words: *Phishing; Association rule; URL.*

1.0 Introduction

Due to the developing utilization of smart cell phones, large numerous individuals are relying upon online services for the payment of bills, to carry out various financial transactions and even interact with family and friends, (both known and unknown). Most commercial and government establishments have also introduced more internet services to customers. While such exercises importantly affected the world wide economy, such expansive reliance on internet services builds security risks concerning customers alongside financial institutions.

Phishing is a falsified practice that utilizes social engineering with specialized deception tricks to obtain customer's personal account credentials and identity. “Phishing” is a newly introduced method of stealing personal identity. The mass media reports almost everyday cases concerning some organization that have their customers targeted by phishing scam. In February 2015, The Telegraph had announced that phishing might be viewed basically as Nigeria's greatest danger that year. While many financial organizations attempt dependably to improve on their security techniques to ensure customers' personal information is secured, phishers also grow considerably more modern attacking methods.

Phishing websites are forge web pages that are designed by noxious individuals to emulate pages of genuine sites (Fortune Magazine, 2011). Phishers regularly make website pages that looks outwardly like the authentic and original site pages used in deceiving their victims. An uninformed internet customer is likely to be effortlessly tricked unknowing that this is a scam. Phishing webpages victims may uncover their financial balance, secret key, and other sensitive data like Visa number to the proprietor of phishing webpage. Phishing could also be defined as a moderately new cybercrime when placed side by side with order internet crimes (e.g. hacking and viruses), an unmistakable increment in the number and seriousness of phishing attack is accounted for (Anti-Phishing Working Group, 2011).

In the cyber space, phishers are group of fraudsters disguising to steal customers financial account details, hurt monetary organizations and harm cyber security. Their exercises keep on affecting the general public and also the economy. Day by day, a relatively huge number of internet users get phishing email, popups, and links to spoof fake websites. Numerous fell (and keep on falling) for these traps.

A current report by Gartner (2011) also indicated that fifty seven million US web users have recognized receiving email connected to phishing, around 1.7 million among them were discover to have complied with the tricking attacks that deceived them into disclosing their personal data.

Investigations by a group known as Anti-Phishing Working Group (APWG) showed that Phishers are probably going to prevail with about 6% of all messages received (Anti-Phishing Working Group, 2011).

The main objective of phishing websites basically is to fraudulently have access to users' personal data via surfing and visiting a forged website page that resembles a genuine site of a true bank or organization and requests that the victim enter individual data, for example, their financial account number, personal identification number, credit card data… and so forth. The resulting effect is break in data safety through the tradeoff of confidential information and the internet customers may end up suffering the loss of cash or some of his or her very valuable assets.

2.0 Related Works

Neda et al. (2014): suggested a classification algorithm based on rules to predict phishing URL. However the rules that was utilized in this method rely on human understanding instead of smart data mining system.

Zhang et al. (2014): adopted domain feature improved classification model to detect Chinese e-business phishing websites.

Hadi and Nawafleh (2012): projected associative classification algorithms for the discovery of phishy sites. The authors analyzed the contrast between projected algorithm, RIPPER, SVM, PRISM, and NB. From the outcome, it showed that associative classification algorithm performed more than all other conventional techniques.

Huang et al. (2012): projected SVM based system detecting phishing URL. The characteristics that were used are brand names, lexical and structural that are contained in such URL.

Aburrous, et al. (2010): suggested a procedure that combine association and the classifications rules mining algorithm, the sole purpose of said system is to classify all the unique characteristics that are applicable in identifying phishy website. The algorithm maximized over 20 phishing characteristics and categorizes them into 6 groups. And then 3 fuzzy set of values ("Genuine", "Doubtful" and "Legitimate") were utilized as the input values. The output looked at feature as the following as set of likely values ("Very Legitimate", "Legitimate", "Suspicious", "Phishy" "Very Phishy").

Zhang et al. (2011): projected CANTINA, a totally distinct HTML content technique for differentiating phishing websites. It analyzes the source-code of web- pages and proceed by utilizing (TF-IDF) to bring out the most outrageous ranking keywords. The keywords that are obtained are supply as inputs to Google- internet searcher and checked if the domain name of the (URL) corresponds with N- top search output and then considered as genuine. This methodology fully depends on Google search tool.

CANTINA+ suggested by Xiang et al. (2013): is an upgraded version of CANTINA, which came with some new features added in order to bring about better results. Specifically, the author incorporate HTML Document Object Model, Google web search and third party with machine-learning procedure to recognize phishing pages

However, both methods are dependable on Google web search tool and they also have their contents downloaded from the web-pages. In our research work, features associated with URL are looked into and therefore downloading the web-page content is eluded. Besides the prediction does not totally rely on search engine end result.

Cabanillas-Liebana et al. (2013): proposed totally unique technique to look for the variables which are commonly used in financial establishments in order to foresee the trust amid electronic banking.

Yuanchenga et al. (2013): suggested semi- supervised based technique for recognition of phishing page. The structures of the web picture structure and DOM characteristics were considered. Transductive Support -Vector Machine was applied to categorize and identify phishing website pages.

Islam et al. (2013): suggested passing the phishing email including the text of the e-mail and header through a filter using multi-level classification model.

3.0 Material and Methods

The existing system adopted a classification algorithm based on rules to predict phishing URL. However the rules that was utilized in this method rely on human understanding instead of smart data mining system. They actually uses a manually scales methods of extraction to detect genuine site and illegitimate site using the padlock symbols to detect phishing site which is prone to damage and attack. Figure 3.1 shows the architecture of the existing system.

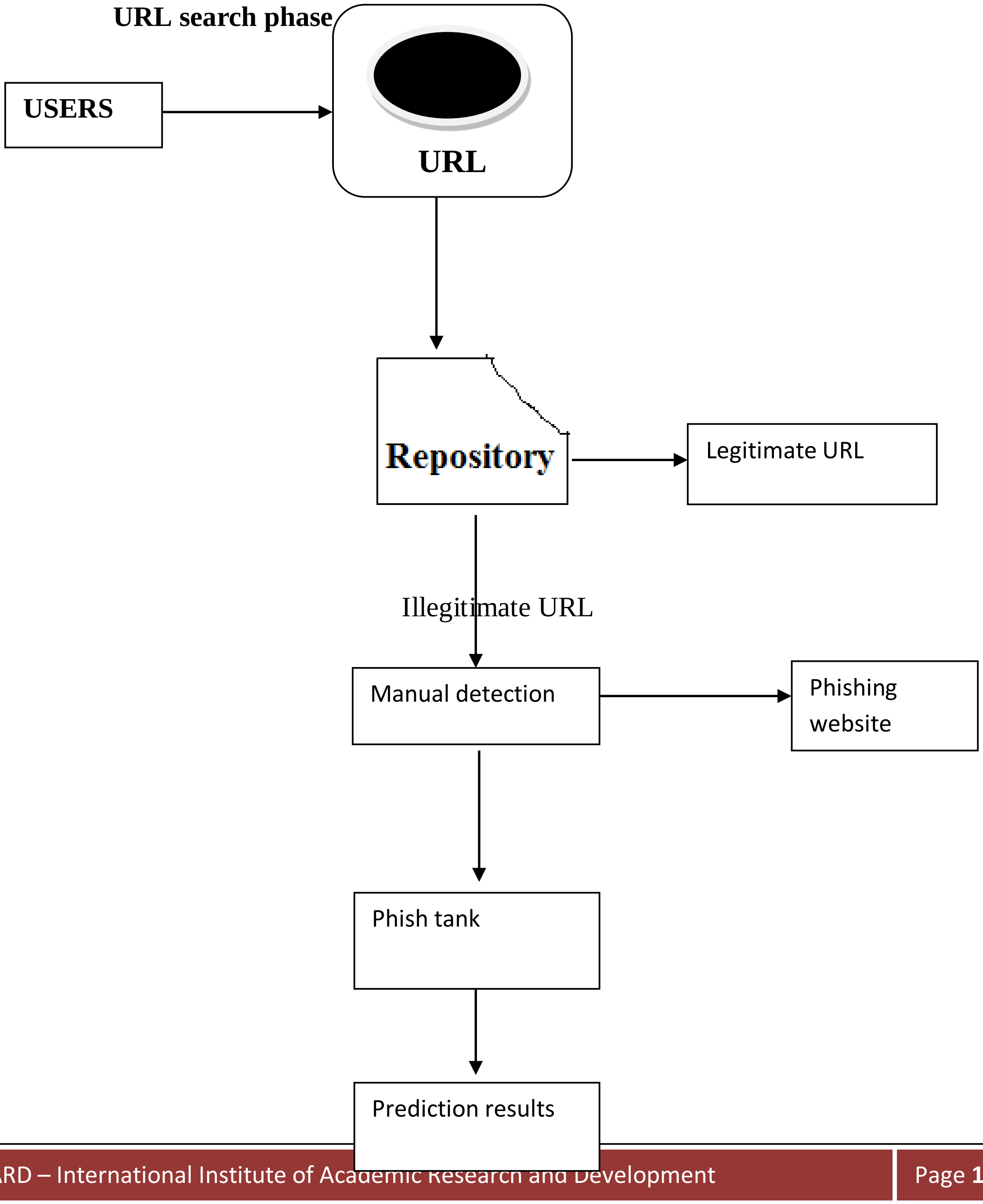


Figure3.1: Architecture of the Existing system (Neda et al. 2014)

Our proposed work is based on further study of Neda et al. (2014) which suggested a classification algorithm based on rules to predict phishing URL. However the rules that was utilized in this method rely on human understanding instead of smart data mining system.

The proposed system comprises of two stages:

(1) URL search stage and

(2) Feature extraction stage.

In the URL search stage, anytime a user requested a URL, a search is carried out to check weather that URL is in the warehouse of genuine URLs. If a match is found in the repository, the URL is then taken as genuine URL. Else the URL isn't a genuine URL and it will be considered to pass through the next stage. The main purpose of carrying out a search before the feature extraction is to decrease unnecessary computation in the feature extraction stage and enhances the model total response time.

When carrying out extraction, we defined some heuristics to mine ten characteristics from the URL. However, to ascertain the legitimacy of the URL, associations rule mining is applied on these URL features.

Feature Extraction

This proposed work is concern with recognizing the necessary features that distinguish phishing sites from real websites and afterward given them to the association rule mining, to distinguish major features. Our phished URL dataset is gotten from the phish tank (http://www.phishtank.com) and the legitimate URL dataset is obtained from 5 different sources. In view of the heuristics, ten features were characterized which are released to associations rule mining for it to effectively determine the real and phishing URL.

Algorithm of the Proposed System

The predictive aprior algorithm is being utilized in the proposed system, and it is stated as follows:

Rule 1: if {Transport layer security = http ∩ “keyword” in the path segment of the URL = Yes ∩ Top level domain = yes} =>class phishing (conf., 1).

Rule 2: if {Number of slash in URL ≥5 ∩ Transport layer security = http ∩ keyword in the path segment of the URL = Yes} =>class phishing (conf., 1).

Rule 3: if {Special characters = yes Ռ Transport layer security= http Ռ keyword in the path segment of the URL>4}=>class phishing (conf., 1).

Rule 4: if {dot is in the host URL>4 Ռ Transport layer security = http Ռ keyword is in the path segment of the URL>4}=> class phishing (conf., 1).

Rule 5: if {Number of slash in URL ≥5 Ռ Dot in the host URL>4Ռ length of URL>75} =>class phishing (conf., 1).

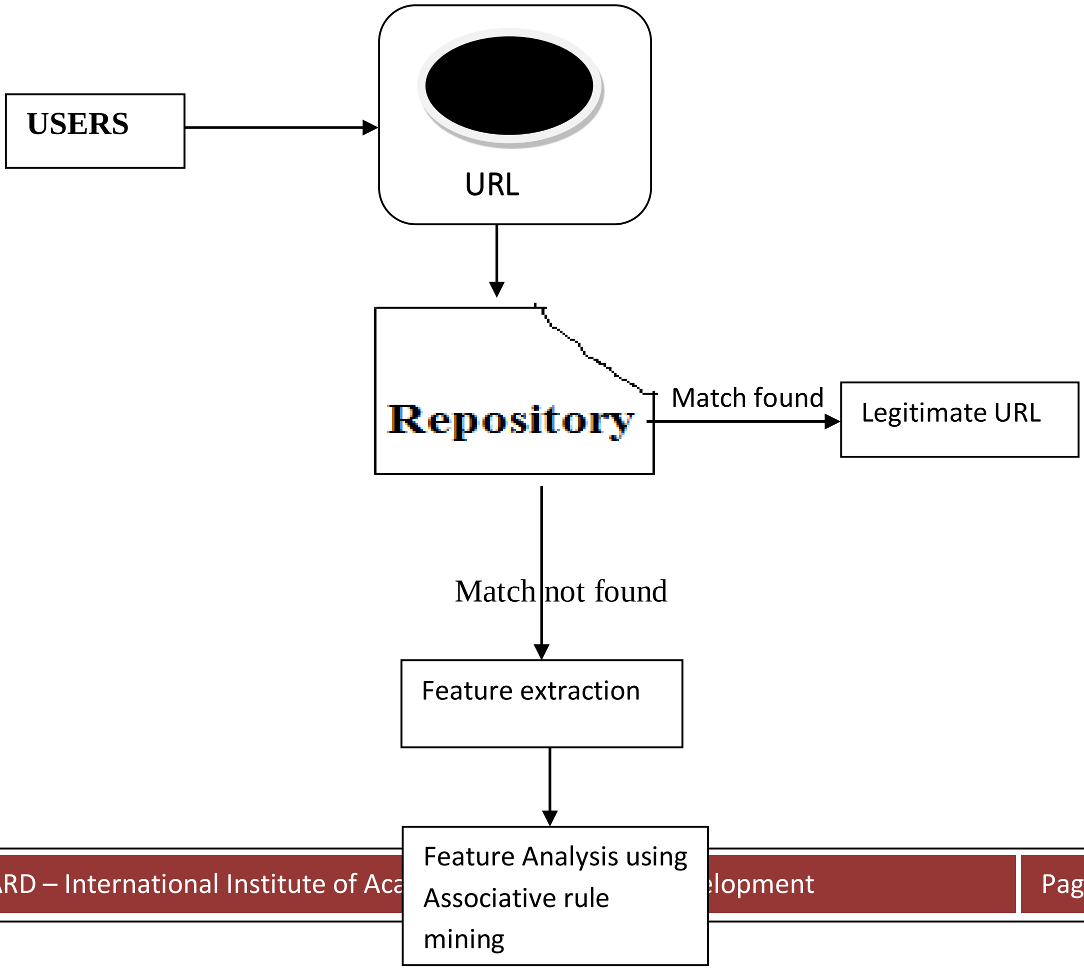
Rule 6: if {Special characters = yes Ռ Transport layer security = Yes Ռ Top level domain = yes} =>class phishing (conf., 1).

Rule 7: if {dot is in the host URL>4 Ռ Transport layer security = http Ռ keyword in the path area of the URL>4}=> class phishing (conf., 1).

Rule 8: if {Transport layer security = http Ռ keyword in the path segment of the URL = Yes}=>class phishing (conf., 1).

Rule 9: if {dot exist in the host URL>4 Ռ keyword in URL path segment = Yes Ռ Top level domain = yes}=> class phishing (conf., 1).

Figure 3.2 shows the architecture of the proposed system.



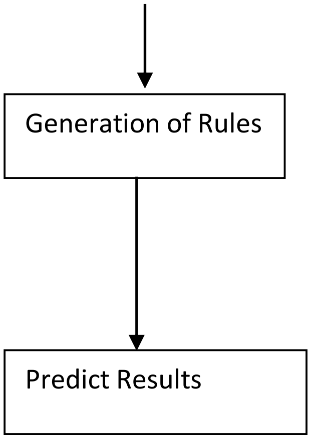
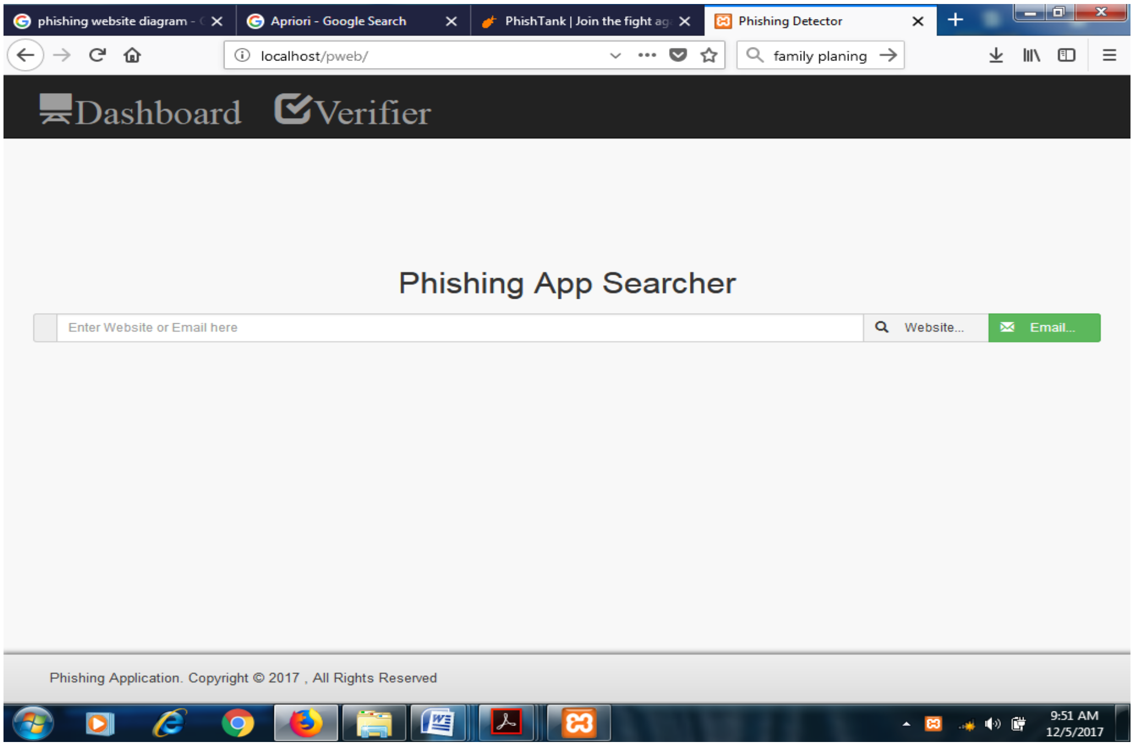


Figure 3.2: Architecture of the Proposed System

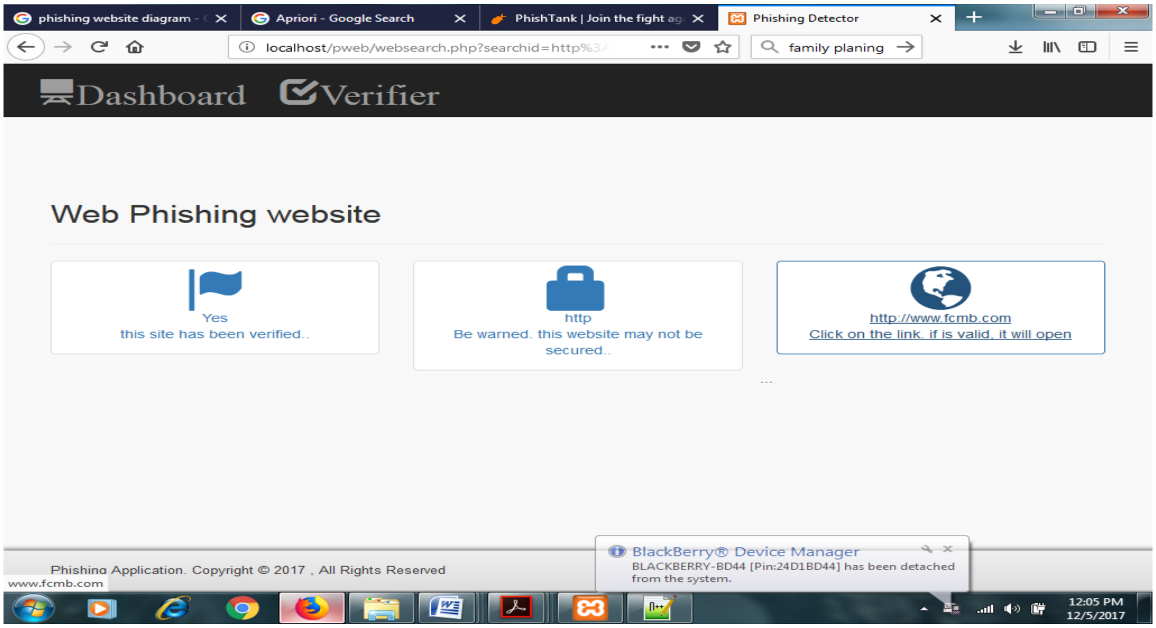
4.0 Result and Discussions

The result of the new model is presented in figure 4.1, figure 4.2 and figure 4.3 below.



Phishing Application. Copyright © 2017 , All Rights Reserved

Figure 4.1 Home Page


phishing website diagram -
Apriori - Google Search
PhishTank | Join the fight ag
Phishing Detector
localhost/pweb/websearch.php?searchid=http%3A
family planing
Dashboard
Verifier

Web Phishing website

Yes
this site has been verified..

http
Be warned, this website may not be secured..

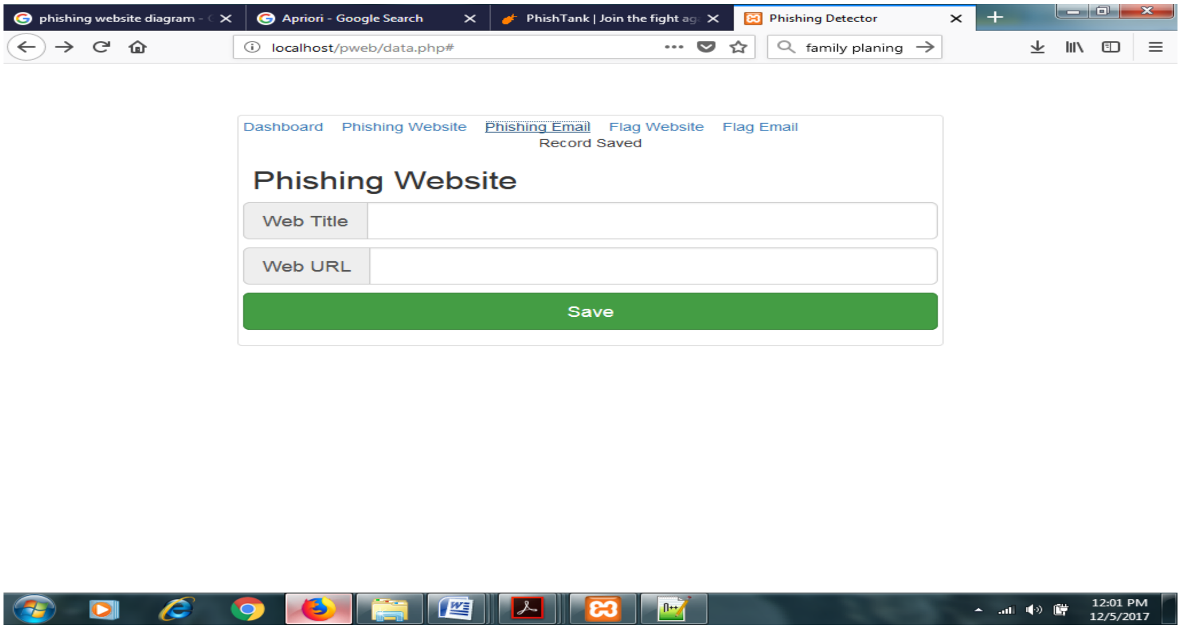
http://www.fcmb.com
Click on the link, if is valid, it will open

Phishing Application. Copyright © 2017 , All Rights Reserved

www.fcmb.com

BlackBerry® Device Manager
BLACKBERRY-BD44 [Pin:24D1BD44] has been detached from the system.

Figure 4.2 Verifier Page



Record Saved

Phishing Website

Web Title

Web URL

Figure 4.3 Adding Phishing Website to the Data Base

Figure 4.1 shows how the model searches URL, Figure 4.2 shows how the model alerts the user whether the URL is a legitimate or phishing URL while figure 4.3 shows how the model saves phishing URL to the data base.

Discussion of Result

The method of recognizing a URL type is generated by means of association rules which used different heuristics to get hidden knowledge. These rules are utilized for recognizing the type of URL any time a client accesses it. We have defined some heuristics mined from some URLs and more than 20 URLs from different sources are obtained. Our genuine URLs are gotten from five different sources and about 50 phishing URLs are dig out from the phishtank database (http://www.phishtank.com). The feature extraction is carried out with PHP, MySQL connector is used in order to fetch the data set in the data base of the system. This investigation was carried out using predictive apriori rules generation algorithms. The investigation is carried out to establish the rules centered on phishing URLs. Table 3.1 shows legitimate data source.

5.0 Conclusion

In our research work, the characteristics of the URLs are analyzed and associative rules mining algorithm is used on them. The rules which are generated are translated for highlighting some of the features which are more common in phishing URLs. The outcomes got uncover that the Associative rules with data mining classification methods beat every other algorithms as for accuracy. The outcomes obtained from the mining of rules shows the helpful features which are present in phished URL.

6.0 Contribution to Knowledge

1. We have develop a model with high prediction accuracy for detecting phishing websites.

2. The model save guard users from online crimes.

7.0 Recommendations

Asides from financial institutions and government organizations, this model is recommended to other organizations such as educational institutions that have academic webpages. It is likewise recommended for software engineers to be included in new software so that fake website can easily be recognized and new website could be secured.

8.0 References

Abdelhamid N, Ayesh A., Thabtah F. (2013) Phishing Detection using Associative Classification Data Mining. ICAI'13 - The 2013 International Conference on Artificial Intelligence, pp. (491-499). USA.

Abdelhamid N., Ayesh A., Thabtah F. (2014) Phishing detection based associative classification data mining. Expert Systems with Applications 41 (13) Pages 5948–5959, Oct 2014.

Aburrous, M. Hossain, M. Dahal, K. and Thabtah, F. (2010): “Predicting phishing websites using classification mining techniques with experimental case studies,” in ITNG: Seventh International Conference on Information Technology: New Generations. 176–181.

Chen X, Bose I, Leung ACM, Guo C (2011) Assessing the severity of phishing attacks: a hybrid data mining approach. Expert System Appl 50:662–672.

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

Islam R, Abawajy J (2013) A multi-tier phishing detection and filtering approach. J Netw Comput Appl 36:324–335.

Xiang G, Hong J, Rose CP, Cranor L (2011) CANTINA+: a feature-rich machine learning framework for detecting phishing web sites. ACM Trans Information System Security14:21.

Zhang D, Yan Z, Jiang H, Kim T (2014) A domain-feature enhanced Classification model for the detection of Chinese phishing e- business websites. Inf Manag 51:845–853.

Han, J.; and Kamber, M., (2001). Data mining: concepts and techniques.

Morgan Kaufmann. p. 5. ISBN 978-1-55860-489-6. Thus, data mining should have been more appropriately named "knowledge mining from data," which is unfortunately somewhat long.

Chen KT, Chen JY, Huang CR, Chen JY (2009) Fighting phishing with discriminative key point features of webpages. IEEE Internet Comput 13:56–63.

Chen X, Bose I, Leung ACM, Guo C (2011) Assessing the severity of phishing attacks: a hybrid data mining approach. Expert Syst Appl 50:662–672.

Clifton G. and Christopher H (2010). "Encyclopædia Britannica: Definition of Data Mining". Retrieved 2010-12-09.

Dhamija R. and Tygar J.D.,(2005):"The Battle Against Phishing: Dynamic Security Skins," System Usable Privacy and Security, 2(5)23-44.

Fu AY, Wenyin L, Deng X (2006) Detecting phishing web pages with visual similarity assessment based on earth mover's distance. IEEE Trans Dependable Secure Comput 3(4):301–321.

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.r