Available online at www.elixirpublishers.com (Elixir International Journal)

Information Technology

Elixir Inform. Tech. 86 (2015) 35183-35186

ISSN: 2229-712X

Phishing URL Detecting Using ANN Classification in Online Social Network

Vaibhav V. Satane and Rohit N. Devikar

Department of Information Technology, AVCOE Sangmner,District:-A.Nagar Maharashtra, India.

ARTICLE INFO

Article history:

Received: 1 August 2015;

Received in revised form:

12 September 2015;

Accepted: 18 September 2015;

Keywords

Phishing,
Data Mining,
Classification
Artificial Neural Network,
Phishtank,
Online Social Network.

ABSTRACT

Phishing is considered a form of web threats that is defined as the art of mimicking a website of an honest enterprise aiming to obtain user's confidential credentials such as usernames, passwords and social security numbers. Social engineering and technical tricks are commonly combined together in order to start a phishing attack. The phishing attack starts by uploading the post on social network site which contain fraud URL that seems authentic to potential victims advising them to meet site and update or validate their information by following a URL link. The fraud user posts their comment on social network service that contains URL which are malicious. These malicious URL some time it contains malware or virus. When user clicks on these fraud URL Then viruses enter into user system and system gets affected. The proposed system Extraction of feature set for detecting fraud URL using ANN in social networking services effectively detect fraud URL with high accuracy.

Introduction

Phishing is type of online attack whose target is influencing the online user as well as found the weakness in system processes as caused by online users. Worldwide Phishing is one of the cyber security threats. Phishing is the act of convincing users to give up their personal information, either. Phishing is an action of stealing user's financial details. In website phishing the fraud user creates website which is similar to original famous websites such as Gmail, Twitter, Dropbox, Facebook and Paypal. The fraud user creates URL with including the name of these famous websites and URL send to the user through social networking sites due to the use of social networking site increased nowdays so social networking sites are most popular to phisher. According to PhishTank, the definition of phishing attack is that "Phishing is a wrong effort, commonly made with the help of email, to pinch user personal information. According to Colin Whittaker, phishing is accessing website without permission and use behalf of a third party with the purpose of baffling viewers into performing an action with which the viewer would only trust a true agent of the third party [10].

Data mining is extraction of useful patterns from data source like data warehouse, data repository and database. Patterns are effective, unique and reasonable. Data mining are also known as knowledge extraction, data dredging, knowledge discovery and data mining.

Phishy URL and Social networking site

Online Social network services such as YouTube, Facebook, Twitter, and MySpace, have recently popular due to supportive information platforms that allow users to share and to interact with other user. Social networking sites are one of the main ways for users to keep track and communicate with their friends online. The increase in popularity of social networking sites allows them to collect a huge amount of personal information about the users, their friends, and their lifestyles

Twitter Social Network

Twitter is a much simpler social network than Facebook and MySpace. It is designed in such a way that where users send short text messages (i.e., tweets) that appear on their friends' pages. Unlike Facebook and MySpace, no personal information is shown on Twitter pages by default. Users are identified with the help of name of user. Only register users are able to read the tweets and post the tweets. A Twitter user can start "following" any other user such as famous and respectable people from different area like sport, politics, movie industry, businessman, different private or public organization people and their friends also. Authorized user are the register user who has twitter account. Authorized user receives tweets the other user's on his page. In twitter social networking site tweets are visible to everyone by default. Other authorized user can retweet through twitter their website.

Related work

Chia-Mei Chen et. al proposed a suspicious URL identification system for use in social network environments is proposed based on Bayesian classification. The proposed system consist 3 modules the first module is data collection, posts are collected including time and content. Posts that lack URL information are considered benign. In the second module, feature extraction, the features are retrieved and a feature vector is constructed for classification. In the third module, the Bayesian classification model, posts are classified based on a pertained classification model [1].

Neda Abdelhamid et.al proposed the Multi-label Classifier based Associative Classification algorithm to detect phishing website also want to identify features that distinguish phishing websites from legitimate ones.it consist no of steps (1) the end-user clicks on a link within an email or browses the internet. (2) Then directed to a website which is genuine or phishy. (3) A script written in PHP that is embedded within the browser starts processing to extract the features of the test data and saves them in a data structure. (4) Now, the intelligent model will be active within the browser to guess the type of the website based on rules learnt from historical websites (previous data collected). The rules of the classifier are utilized to predict the type of the test data based on features. If the website is recognized as genuine no action will be taken. On the other hand, when the

Tele:
E-mail addresses: satanevaibhav@gmail.com

website finds phishy, the user becomes warned by the intelligent method that he is in risk [2].

Isredza Rahmi A. Hamid et.al an approach for email born phishing detection based on profiling and clustering techniques also formulate the profiling problem as a clustering problem using various features present in the phishing emails as feature vectors it produces profiles with the help of clustering predictions. These predictions are additional used to generate complete profiles of the emails and compared the performance of the proposed approach against the Modified Global K-means approach [3].

Gowtham Ramesh et. al proposed novel approach that not only overcomes many of the difficulties in detecting phishing websites but also identifies the phishing target that is duplicate website and proposed an anti-phishing technique that groups the domains from hyperlinks direct or indirect linked with the fraud website. The domains gathered from the directly linked webpages are compared with the domains gathered from the indirectly related webpages to reach at a targeted domain set. Applying Target Identification (TID) algorithm on this domain set, zero-in the target domain then perform third-party DNS lookup of the fraud domain and the target domain and by comparison identify the correctness of the fraud page [4].

Mingxing He et. at present a heuristic method to determine whether a webpage is a legitimate or a fraud page. This scheme detects new fraud pages which are blacklisted with the help of nonphishing tool. Initially webpage convert into number of features which are selected depending on the normal and phishing pages chartrastics. A training set of web pages containing normal and fraud web pages are input for a support vector machine to do training. A testing set is finally inserted into the trained model for the testing. Compared to the existing methods, results show that the proposed phishing detector can achieve the high accuracy rate with relatively low false positive and low false negative rates [5].

Santhana Lakshmi et. al employs Machine-learning technique for modeling the prediction task and supervised learning algorithms namely Multi-layer perceptron, Naïve Bayes classification and Decision tree induction are used. the decision tree classifier predicts the phishing website more accurately compare to other learning algorithms [7].

Haijun Zhang et. al proposed A novel framework using a Bayesian approach for content-based phishing web page detection. The proposed model takes into account textual and visual contents to calculate the similarity between the protected web page and suspicious web pages. This is required in the classifier for determining the class of the web page and identifying whether the web page is phishing or not. The text classifier and naive Bayes rule is used to calculate the probability that a web page is phishing.[8]

Arun Vishwanath et.al proposed an integrated information processing model of phishing susceptibility grounded in the prior research in information process and social fraud. this system improve and validate the model with the help of a sample of intended victims of an actual phishing attack [9].

Gaurav Gupta et. al proposed a model that hybrid of blacklisting, heuristics, and moderation based phishing prevention. The scheme can be implemented as a plug-in P into the browser B. The browser perform initial heuristic and blacklisting checks on the URL before loading webpage which block the page. If the URL is black or white-listed, no further checks need to be performed with the page blocked in the former, and loaded in the latter case. Otherwise, the page is fetched and secondary heuristic checks are performed to eliminate further chances of phishing. If the pagemakes this far, it implies that the first stage tests have been inconclusive and moderators are contacted to classify this page as Phish or Genuine [10].

Kuan-Ta Chen et.al an effective image based antiphishing scheme based on discriminative key point features in Web pages. The Contrast Context Histogram (CCH), computes the similarity degree between doubtful and genuine pages with high accuracy and low error rates [11].

Anthony Y. Fu et. al proposed An effective approach to phishing Web page detection is proposed, which uses Earth Mover's Distance (EMD) to measure Web page visual similarity. first convert the involved Web pages into low resolution images and then use color and coordinate features to represent the image signatures. use EMD to calculate the signature distances of the images of the Web pages. then train an EMD threshold vector for classifying a Web page as a phishing or a normal one [12].

Proposed System

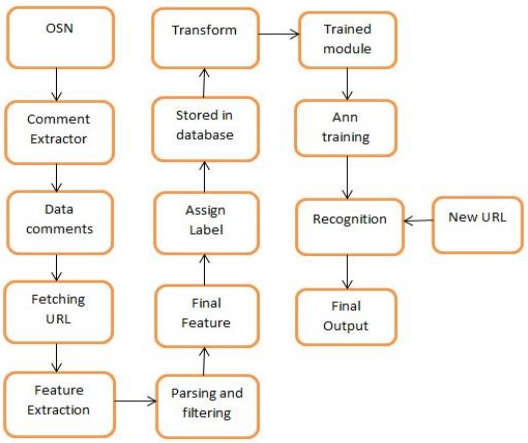
The proposed system consist four modules

- OSN Comment Extractor
- Feature Extraction Module
- Training Using ANN Model
- Recognition and Result
- OSN Comment Extractor

Nowadays there are number of online social network available such as Twitter Facebook. The proposed system developed for twitter. First user need to create account on twitter, then sign in the twitter account and create application in twitter so twitter generates following credential.

Consumer Key
Consumer Secret
Access Token
Access Token Secret

The first module include number of steps such as extractor, data comments and fetching URL.These credentials are used to access the twits that post by other user. When user post the tweets sometime user post the URL which are used as input for our system. The twitter provide API by using API extract twits and also The comment data downloaded from online social networking site by using java code. Once the comments data set downloaded, it will be passed and sent to next modules.



Feature Extraction Module

The feature module include number of steps feature extraction, parsing and filtering, assign label and stored in database. The URL is used as input for second module. By using java code separate the each term that include in URL,applying parsing and filtering technique on the feature of URL, finally assign label to that feature and stored in database .from any

URL extract different feature like number of slash,no of dots, no. of special character, foreign link and anchors. The user assign label to different data set known as labeled data set. Labeled data set means that once comments are downloaded, we need to divide them into two groups

1. Training set
2. Testing set

Testing set does not need any labeling but training set needs labeling. Labeling is the process in which we inform that which of the URLs that are extracted from comments are good and which are bad. The output of labeling is used for training the classifier. The labeled data set store into database.

Training Using ANN Model

In third module the comment data set which store in to the database is used. This module applies transform technique on comment data set and then using URL filter those labeled data set. The transform technique we will be using is thresholding or clustering either of these techniques shall change the continuous data (from feature set) to bounded discrete data. Use some URL to apply the ANN training, we use multilevel feed forward network and backpropagation algorithm to build ANN trained module.

E.g. number of symbols in URL will can be anywhere between 0 to 20 or even more. but we can cluster or threshold them into two groups of <5 and >=5 or we can cluster them into three groups like 0-5, 6-beyonds way the new transformed values will only be between 0 and 1 for 2-class and 0,1,2 for 3-class clusters.

less the clusters, more the robustness with less accuracy, more the clusters, more the accuracy with less robustness.

In next step apply ANN algorithm on data set to create model known as ANN model. ANN training is the back propagations part

Recognition and Result

Once training is done. We can use different kinds of new URL to recognize phishing and nonPhishing URL. In fourth module recognition of new feature, classify the output and analysis of final result will obtain. IN this module Ann recognition step used to recognize the new comment features which are posted by user on online social network, classify the output and show the final result whether URL is legal or illegal. Artificial neural network recognition is the FEED FORWARD part of ANN.

Dataset

We have collected phishing URL from website Phishtank and some non-phishing URL. We use around 45 phishing URL and 20 non phishing URL.

Result

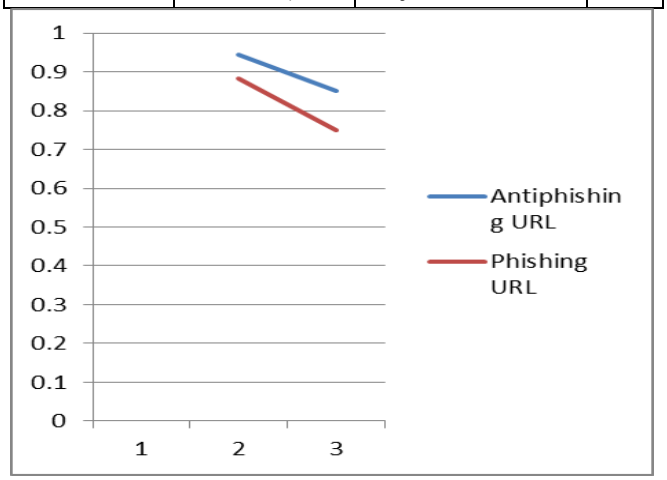
Phishers usually try to convince the victims into clicking on fraud URLs pointing to phishing sites. A URL obfuscation technique is used by phisher and has aim of hiding the real host and registered domain. Mixing the original domain name or phishing keywords into the remaining part of the URL. The keywords are such as famous brand. The goal of our system is to detect phishing URL that post by fraud user on social networking service over previous system on the basis of the classification technique that used to classify the features of URL.For the performance evaluation of phishing detection in online social networking sites using ANN classification. we established a system configuration which allows us to evaluate phishing detection. We conduct number of experiment to measure accuracy of our system. For our experiment we use data set in which total phishing URL are 44 and nonPhishing URL are 20. The phishing URL are obtained from website PhishTank.

Precision

Antiphishing URL	(Relevant Retrieved)	Correct Retrieved Object	0.97
Phishing URL	(Relevant Intersect Retrieved) / Retrieved	Correct Retrieved Object / Retrieved Objects	0.96

Recall

Antiphishing URL	(Relevant Intersect Retrieved) / Retrieved	Correct Retrieved Object / Retrieved Objects	0.95
Phishing URL	(Relevant Intersect)	Correct Retrieved Objects	0.96



Conclusion

Fraud user use social networks are often used to collect personal information as well as collect financial data of user. Extraction of feature set for finding mistrustful URL using ANN classification in social networking site, we can extract URLs,cluster the different set of features, apply transformation technique, apply ANN classification to identify the fraud URL post. We can also extract data of website as well as different foreign link present on that website. ANN is generally considered to be an adaptive system that changes its structure in response to external or internal information that flows through the network during the learning phase.This study has found that by using Extraction of Feature set for finding phishing URL using ANN Classification in

Reference

[1] Chia-Mei Chen et. al "Feature set identification for detecting suspicious URLs using Bayesian classification in social networks", Information Sciences 289, 0020-0255 Elsevier.

[2] Neda Abdelhamid et. al "Phishing detection based Associative Classification data mining", Expert Systems with Applications 41, 0957-4174 Elsevier,2014.

[3] Isredza Rahmi A. Hamid. "An approach for profiling phishing activities", computers & security, 0167-4048Elsevier 2014.

[4] Gowtham Ramesh et. al "An efficacious method for detecting phishing webpages through target domain identification", Decision Support Systems 61, 0167-9236 Elsevier2014.

[5] Mingxing He et.al "An efficient phishing webpage detector", Expert Systems with Applications 38 Elsevier 12018–12027,2014.

[6] Mahmoud Khonji et al. "Phishing Detection: A Literature Survey", IEEE Communication Surveys & Tutorials, vol. 15, No. 4. 2013.

[7] Santhana Lakshmi et. al "Efficient prediction of phishing websites using supervised learning algorithms", International Conference on Communication Technology and System Design, Elsevier 2011.

[8] Haijun Zhang et. al "Textual and Visual Content-Based Anti-Phishing:A Bayesian Approach", IEEE Transaction On Neural Networks, VOL. 22, NO. 10, OCTOBER 2011.

[9]Arun Vishwanath et. al "Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model", Decision Support Systems 51, 0167-9236 Elsevier 2011

[10]Gaurav Gupta et.al "Socio-technological phishing prevention" information security technical report 16, Elsevier 2011.

[11] Kuan-Ta Chen et. al "Fighting Phishing with Discriminative Keypoint Features", IEEE Internet Computing, 1089-7801, 2009.

[12] Anthony Y. Fu et. al "Detecting Phishing Web Pages with Visual Similarity Assessment Based on Earth Mover's Distance (EMD)", IEEE Transactions on Dependable and Secure Computing, VOL. 3, NO. 4, 2006