IJCSC 0973-7391 Volume 7 • Number 1 Sept 2015 - March 2016 pp. 236-243 available online at www.csjournalss.com

A Machine Learning Approach to Improve the Efficiency of Fake Websites Detection Techniques

Anuj Dakwala, Kruti Lavingia
Nirma University,Ahmedabad, India
14mcen04@nirmauni.ac.in, kruti.lavingia@nirmauni.ac.in

Abstract: Phishing is a kind of cyber-attack in which perpetrators use spoofed emails and fallacious web sites to lure unsuspecting online users into giving up personal data. This paper takes a gander at the phishing issue totally by looking at different research works and their countermeasures, and how to increase detection. It makes out of two studies. In the main study, focus was on dataset assembling, feature extraction and preprocessing for the classification process. In the second study, focus was on metric evaluation of a set of classifiers (SVM, C4.5, LR and KNN) utilizing the precision, accuracy, f-measure and recall metrics. The output of the classifier study is utilized to pick the best performed classifier. The subsequent result of the study demonstrates that the classifier technique performed better with an accuracy of 99.37%. This outcome can be attributed to the small size of dataset utilized as it was appeared as a part of past scrutinizes that K-NN performs better with a decreasing size of dataset while classifiers like C4.5 and SVM performs better with increasing size of dataset.

Keywords: Classifier, Confidentiality, Cross-Validation, Fraud, Performance.

1. Introduction

Cybercrime introduces to crimes that objective to system or computer such that the PCs may or may not have been fully instrumental to the commission of the crime [1]. Personal Computers crimes comprise of a wide scope of conceivably criminal activities. However, it can be sorted into both of two sections [1]:

1. Crimes that directly target devices networks, or computers.
2. Crimes aided by devices, networks, or computers the main aim of which is not targeted at device or computer network.

Phishing in contrast with different types of internet threat, for example, hacking and virus is a quickly developing web wrongdoing. In the wide use of internet as a noteworthy type of correspondence, phishing can be executed in various routes for example [2]:

1. Email-to-website: when someone receives an email embedded with phishing web address.
2. Browser-to-website: when someone misspelled a legitimate web address on a browser and then referred to a phishing website that has a semantic similarity to the legitimate web address.
3. Email-to-email: when someone receives an email requesting sensitive information to be sent to the sender.
4. Website-to-website: when someone clicks on phishing website through a search engine or an online advert.

Distinctive sorts of anti-phishing measures are being utilized to counteract phishing, for example, Anti-Phishing Working Group is an industry gathering, which details phishing reports from various online incident resource and makes it accessible to its paying individuals [3]. In the meantime, anti-phishing measures have been executed as extra augmentation or toolbars for programs, as elements inserted in programs, and as a major aspect of site login operation. A large portion of these toolbars have been utilized as a part of the detection of phishing. [4] proposed Spoof Guard which warns users of phishing website user [5]. This tool makes utilization of URL, pictures, domain name, and link to evaluate the spoof likelihood.

Lucent Personalized Web Assistant (LPWA) is an apparatus that guards against wholesale fraud to secure client's personal information [6]. It uses a function to define user variables such as email address, username, and password for every server visited by the client. [7], proposed a similar approach in PwdHash.

[8] introduced TrustBar which is a third-party certification solution against phishing. Trusted Credentials Area (TCA) proposed by the authors. The TCA controls a noteworthy area, situated at the highest point of each browser window, and sufficiently substantial to contain exceptionally visible logos and other graphical symbols for credentials identifying a legitimate page. In spite of the fact that their answer does not depend on complex security elements, it doesn't prevent against spoofing assaults. In particular, since the logos of sites don't transform, they can be utilized by an attacker to make a look alike TCA in an untrusted web page.

DOI: 10.090592/IJCSC.2016.037

Because of the constantly expanding phishing sites springing up by the day, it is turning out to be progressively hard to track and block them as assailants are concocting imaginative techniques consistently to allure clueless clients into unveiling their own data [4].

2. Problem Background

As another kind of cyber security risk, phishing sites show up oftentimes in recent years, which have prompted incredible mischief in online money related services and information security [9].[10] claimed that the method used in carrying out phishing can be different across regions. Moreover, he also derived that the phishers in America and China region have diverse methodologies that he categorized into two on the premise of region.

1. The American phishers would rather deploy the phishing website using a hacked website.
2. The Chinese phishers prefer to register a new domain to deploy the phishing website.

Most analysts have taken a shot at expanding the accuracy of website phishing location through various procedures. These classifiers can be ordered into two strategies: Machine learning or either probabilistic. Based on these algorithm, a few issues with respect to phishing site detection have been settled by various analyst. Some of these algorithm were assessed utilizing four metrics, precision, F1 Score, accuracy and recall.

Some studies have connected K-Nearest Neighbour (KNN) for phishing site classification. KNN classifier is a nonparametric grouping calculation. One of the characteristic of this classifier is that it sums up at whatever point it is generalizes. In addition, previous researchers have demonstrated that KNN can accomplish precise results, and infrequently more exact than those of the typical classifiers.

Meanwhile, Artificial Neural Network (ANN) is another mainstream machine learning strategy. It comprises of a gathering of preparing components that are profoundly interconnected and transform a set of inputs to a set of desired outputs. The significant detriment is in the time it takes for parameter choice and network learning. Then again, previous researchers about have demonstrated that ANN can accomplish exceptionally precise results contrasted with other learning characterization techniques.

3. Scope of Study

The scopes of this research are as follow:

1. The phishing dataset is obtained from phishtank (www.phishtank.com) whereas the legitimate website is obtained manually using Google webcrawlers.
2. First, the dataset is divided into three sets which are then used to train and test the algorithms; Decision Tree (C4.5), Support Vector Machine (SVM), Linear Regression (LR), and K-Nearest Neighbor KNN.
3. The performance metrics of the reference algorithms based on precision, recall, f1-score and accuracy of the three algorithms are compared.

4. Proposed Solution

Research framework will be for implementing the steps taken throughout the research. It is normally used as a guide for researchers so that they are more focused in the scope of their studies. Figure 1 shows an operational framework that will be followed in this study.

The study is divided into three phases and each phase’s output is an input to the next phase. Part-1 is based on dataset processing and feature extraction. Part-2 is based on evaluating individual reference classifiers that involve training and testing using precision, recall, accuracy, and F1-score.

These phases are depicted in the Figure 1.

The processing of dataset was carried out on the collected datasets to better refine them to the requirement of the study. Many stages are involved in processing, some of this are: feature extraction, normalization, dataset division, and attribute weighting. These are very necessary in ensuring that the classifier can understand the dataset and properly classify them into the reference classes.

Evaluation of classifiers is required in this research to measure the performance achieved by a learning algorithm. To do this, a test set consisting of dataset with known labels is used. Each of the classifier is trained with a training set, applied to the test set, and then measured the performance of by comparing the predicted labels with the true labels (that were not available to the training algorithm). Therefore, it is important to evaluate the classifiers by training and testing with the dataset obtained using the following performance metrics; precision, recall, f1-score, and accuracy.

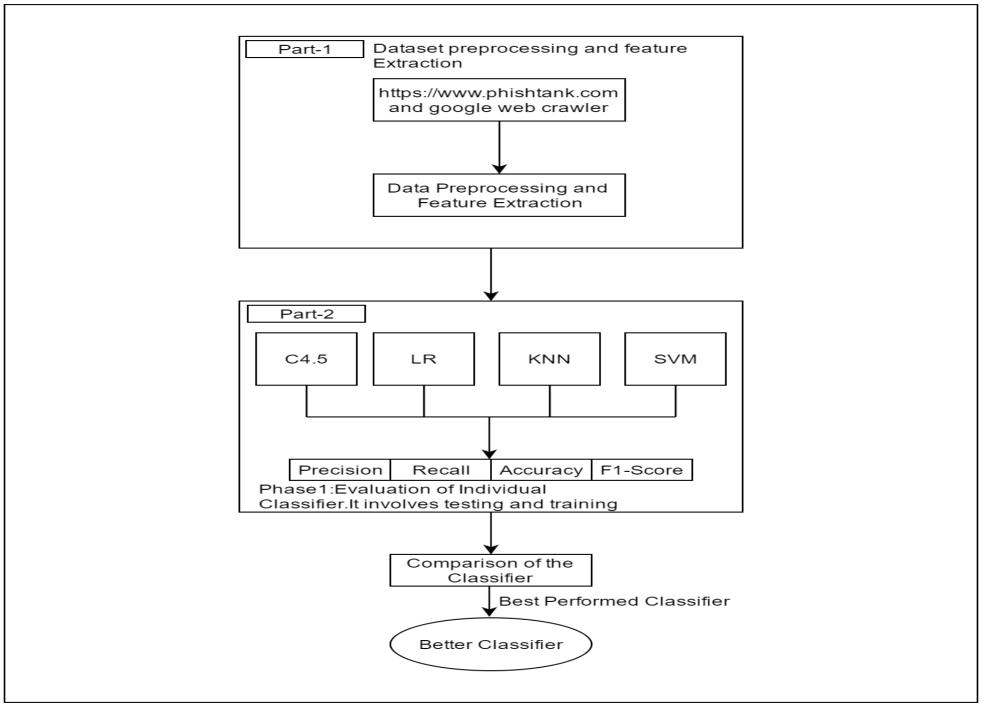
Part-1
Dataset preprocessing and feature Extraction
https://www.phishtank.com and google web crawler
Data Preprocessing and Feature Extraction
Part-2
C4.5
LR
KNN
SVM
Precision
Recall
Accuracy
F1-Score
Phase1:Evaluation of Individual Classifier.It involves testing and training
Comparison of the Classifier
Best Performed Classifier
Better Classifier

Fig 1.Proposed Model

5. Training and Testing Model

Training and testing model is also termed as baseline model in this paper. This model serves as a baseline for selecting the best classifier discussed in next section. Furthermore the baseline model output serve as one of the input for the process discussed in next section Figure 2 shows the procedure of training and testing model.

In this design, the “retrieve dataset” process will retrieve the one of the three datasets at a time and pass it over to the “training and validation” process where *x*-validation used and the model applied for training. The most important component of this model are the reference classifiers used for each loop from the “performance metric” to “training and validation.” Also, the “performance metric” loop back to “retrieve dataset” after every complete rotation of obtaining performance metrics until all the three datasets have been passed through the model.

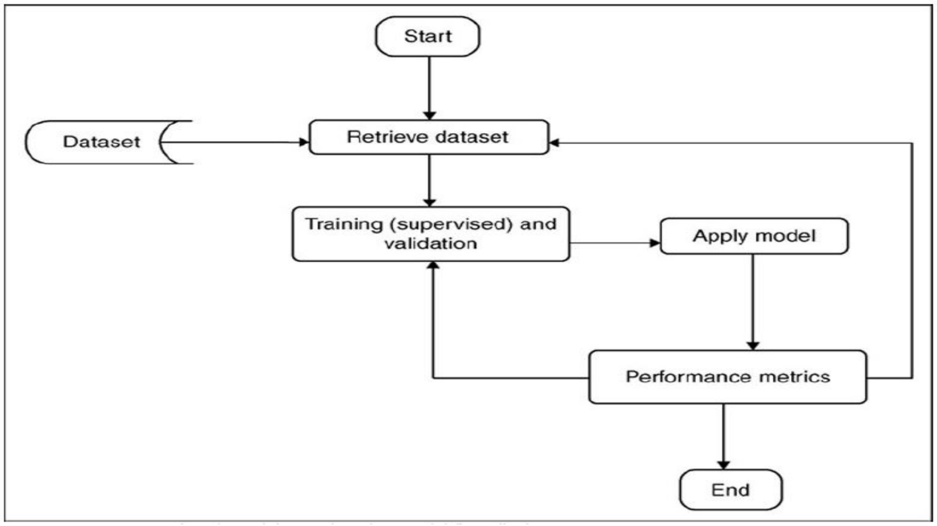

Fig 2. Procedure for training and testing model (baseline)

DOI: 10.090592/IJCSC.2016.037

6. Performance Evolution

TABLE 1. Accuracy Result for Validation Numbers Used Respectively

CV	C4.5	LR	K-NN1	K-NN2	SVM
10	99.09%	99.03%	99.37%	99.26%	99.03%
20	99.08%	99.03%	99.37%	99.26%	97.88%
30	98.97%	99.03%	99.37%	99.26%	99.03%
40	98.97%	99.03%	99.37%	99.26%	99.03%
50	99.03%	99.03%	99.37%	99.26%	99.03%
60	98.98%	99.03%	99.37%	99.26%	98.80%
70	99.09%	99.03%	99.37%	99.26%	98.63%
80	98.97%	99.03%	99.43%	99.32%	99.03%
90	99.03%	99.03%	99.37%	99.25%	98.62%
AVG	99.02%	99.03%	99.38%	99.27%	98.79%
STD	0.00050111	2.22045E-16	0.00018856	0.00019500	0.00360648

TABLE 2. Precision Result for Validation Numbers Used Respectively

CV	C4.5	LR	K-NN1	K-NN2	SVM
10	99.75%	99.92%	99.76%	99.76%	99.92%
20	99.76%	99.92%	99.76%	99.76%	99.83%
30	99.68%	99.92%	99.76%	99.76%	99.93%
40	99.68%	99.92%	99.76%	99.76%	99.92%
50	99.76%	99.92%	99.76%	99.76%	99.92%
60	99.68%	99.92%	99.76%	99.76%	99.92%
70	99.92%	99.52%	99.76%	99.76%	99.92%
80	99.77%	99.92%	99.77%	99.77%	99.92%
90	99.77%	99.93%	99.77%	99.77%	99.93%
AVG	99.75%	99.88%	99.76%	99.76%	99.91%
STD	0.0007036062	0.00126139	4.2E-05	4.2E-05	0.00028846

TABLE 3. Recall Result for Validation Numbers Used Respectively

CV	C4.5	LR	K-NN1	K-NN2	SVM
10	98.94%	98.69%	99.35%	99.18%	98.69%
20	98.94%	98.70%	99.35%	99.18%	97.14%
30	98.85%	98.69%	99.35%	99.18%	98.69%
40	98.86%	98.70%	99.35%	99.19%	98.70%
50	98.86%	98.69%	99.35%	99.19%	98.69%
60	98.87%	98.70%	99.36%	99.19%	98.38%
70	98.76%	98.67%	99.33%	99.17%	98.08%
80	98.77%	98.69%	99.43%	99.27%	98.69%
90	98.85%	98.68%	99.34%	99.18%	98.08%
AVG	98.86%	98.69%	99.36%	99.19%	98.35%
STD	0.0005871042	0.00009428	0.0002708	0.00028197	0.00493929

TABLE 4. F-Score Result for Validation Numbers Used Respectively

CV	C4.5	LR	K-NN1	K-NN2	SVM
10	99.34%	98.60%	99.55%	99.47%	99.30%
20	99.34%	99.30%	99.55%	99.47%	98.32%
30	99.25%	99.30%	99.55%	99.46%	99.30%
40	99.26%	99.29%	99.55%	99.46%	99.29%
50	99.30%	99.29%	99.55%	99.47%	99.29%
60	99.25%	99.29%	99.56%	99.46%	99.12%
70	99.31%	99.30%	99.53%	99.45%	98.90%
80	99.24%	99.28%	99.58%	99.50%	99.28%
90	99.28%	99.28%	99.54%	99.47%	98.87%
AVG	99.29%	99.29%	99.55%	99.47%	99.07%
STD	0.0003654863	0.00007857	0.0001247	0.00013147	0.00312769

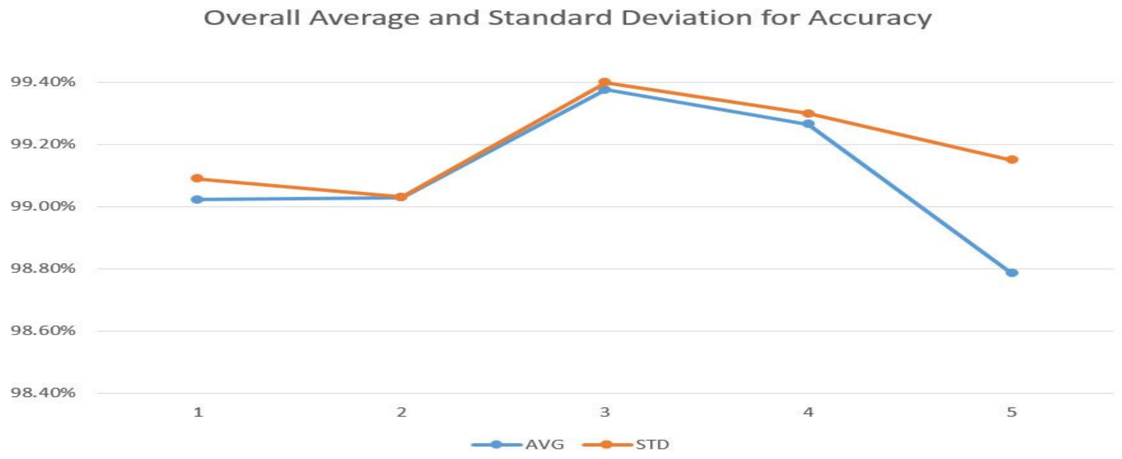

Fig 3. Overall average and standard deviation for Accuracy

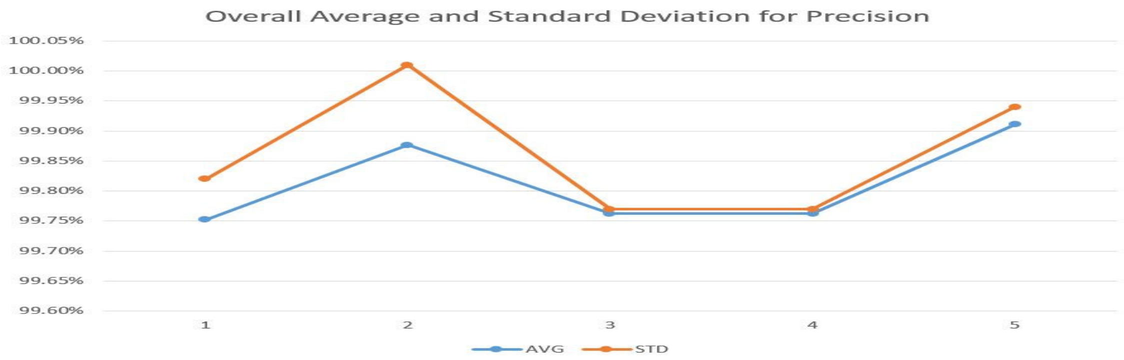

Fig 4. Overall average and standard deviation for Precision

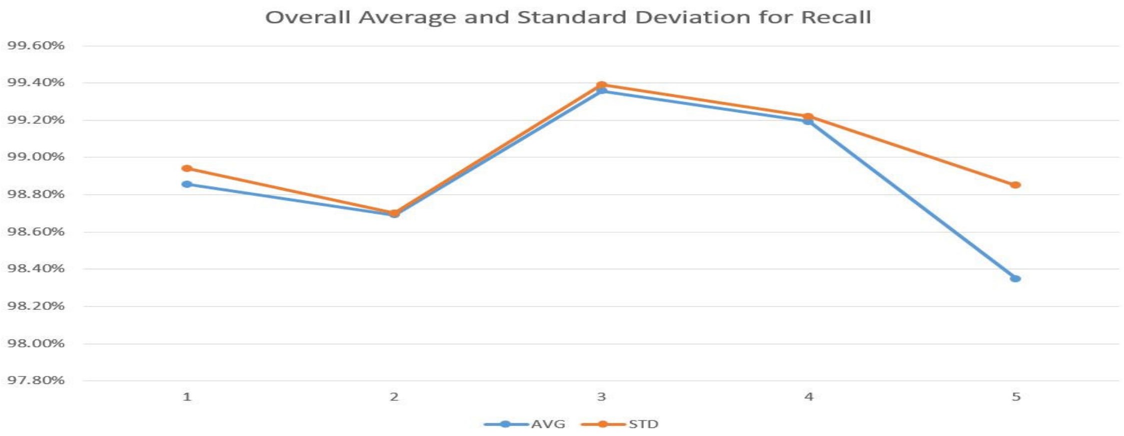

Fig 5. Overall average and standard deviation for Recall

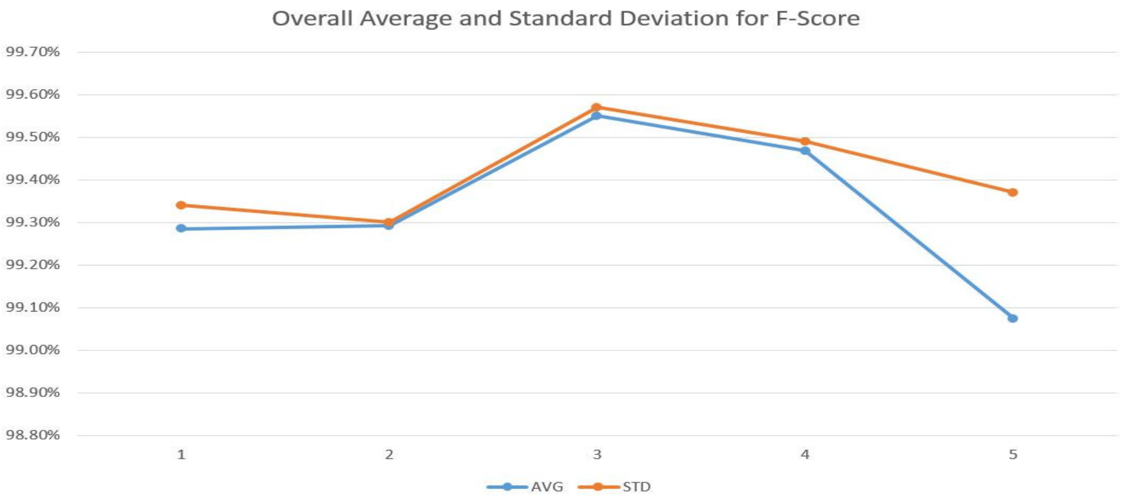

Fig 6. Overall average and standard deviation for F-measure

Looking at the accuracy of K-NN1 and K-NN2 shown in Table 1, it is obvious to conclude that K-NN1 performs better than K-NN2 and as such K-NN1 is chosen over K-NN2 in the further implementation phases discussed later on in this chapter. Based on the justification discussed for number of validation, each of the reference algorithms was trained and tested across the three sets of dataset and the resulting output of this process is shown in Tables 1–4. Corresponding charts of the result obtained are shown in Figures 3–6.

TABLE 5. Best Performed Classifier

SET	KNN
Accuracy	99.37%
Precision	99.76%
Recall	99.35%
F-Score	99.55%

Scrutinizing the results obtained from individual classifier performance across the varying dataset used, it was observed that K-NN perform best with Set A based on accuracy and f-measure. Perhaps, considering both precision and recall may give a confusing interpretation to the results without considering the f-measure which is the harmonic mean of combined precision and recall. Therefore, investigating the f-measure of individual classifiers across varying dataset as shown in Table 5, it is obvious that K-NN f-measure is the highest at 99.55%. Hence, the best performed classifier out of all the reference classifiers is chosen as K-NN (Table 5).

Conclusion

Part 1 focuses on dataset gathering, preprocessing, and feature extraction. The objective is to process data for use in Part 2. The gathering stage is done manually by using Google crawler and Phishtank, each of this data gathering methods were tested to ensure a valid output. The dataset is validated first after gathering, then normalized, features extraction and finally dataset division. Selected Features for this, to ensure an optimum result from the classifiers and also, since using a small feature set will invariably speed up processing time for training and for classification of new instances.

Part 2 focuses on design and implementation of training and validating model using single classifier. A predefined performance metrics is used as a measurement of accuracy, precision, recall, and f-measure. The objective of this phase is to test the performance of individual classifiers in the pool of varying dataset as divided and select the most performed of all the reference classifiers. An accuracy of 99.37% was obtained from K-NN which is the highest as compared to other classifiers referenced.

DOI: 10.090592/IJCSC.2016.037

References

[1] Martin A, Anutthamaa N, Sathyavathy M, Francois MMS, Venkatesan DVP. A Framework for Predicting Phishing Websites Using Neural Networks. CoRR. 2011:1074.

[2] Alnajim A, Munro M. An Approach to the Implementation of the Anti-Phishing Tool for Phishing Websites Detection. Intelligent Networking and Collaborative Systems, 2009. INCOS'09. International Conference on, 2009. IEEE. 2009:105–112.

[3] RSA. Phishing special report: What we can expect for 2007? White Paper. 2006.

[4] Garera S, Provos N, Chew M, Rubin AD. A framework for detection and measurement of phishing attacks. Proceedings of the 2007 ACM workshop on Recurring malcode. ACM; 2007:1–8.

[5] Chou N, Ledesma R, Teraguchi Y, Boneh D, Mitchell JC. Client-side defense against web-based identity theft. San Diego, USA: 11th Annual Network and Distributed System Security Symposium (NDSS'04); 2004.

[6] Gabber E, Gibbons PB, Kristol DM, Matias Y, Mayer A. Consistent, yet anonymous, Web access with LPWA. Commun. ACM. 1999; 42:42–47.

[7] Ross B, Jackson C, Miyake N, Boneh D, Mitchell JC. A browser plug-in solution to the unique password problem. Proceedings of the 14th Usenix Security Symposium. 2005.

[8] Herzberg, A., Gbara, A., 2004. Trustbar: Protecting (even naive) web users from spoofing and phishing attacks. Computer Science Department Bar Ilan University, 6.

[9] Zhuang W, Jiang Q, Xiong T. An Intelligent Anti-phishing Strategy Model for Phishing Website Detection. Distributed Computing SystemsWorkshops (ICDCSW), 2012 32nd International Conference on, 2012. IEEE. 2012:51–56.

[10] Zhang J, Ou Y, Li D, Xin Y. A prior-based transfer learning method for the phishing detection. J. Networks. 2012; 7:1201–1207.

DOI: 10.090592/IJCSC.2016.037