ISSN:1991-8178

**Australian Journal of Basic and Applied Sciences**

Journal home page: www.ajbasweb.com

# New Discriminative Features for Phishing Filtering

[1]Hiba Zuhair, [2]Ali Selamat and [2]Mazleena Salleh

[1]*Faculty of Computing, Universiti Teknologi Malaysia, Johor, Malaysia Al-Nahrain University, Baghdad, Iraq.*
[2]*Faculty of Computing, Universiti Teknologi Malaysia, Johor, Malaysia*

**ARTICLE INFO**

***Article history:***
*Received 13 June 2015*
*Accepted 5 August 2015*
*Available online 12 August 2015*

***Keywords:***
*Phishing, Hybrid Features, Recursive Validation, Base Feature Classifier*

**ABSTRACT**

Rapidly evolving phish websites with new deceptive features makes phishing more sophisticated and complicated cybercrime's form day by day. More specifically, phish websites becomes a serious problem for financial organizations and their customers on cyberspace whenever the existing phish website filters misclassify those features and being bypassed by phishers who exploiting them. Towards picking up new phishing features and avoiding the existing misclassifications, this paper presents new hybrid features and a recursive evaluation approach to investigate their utility for phish website filtering. With the aid of a feature selection criterion and a machine learning classifier, these hybrid features were extracted, recursively trained and validated, and tested through several steps. Experimentally, investigating these features on the collected dataset revealed that they could be potential in terms of classification performance for better phish website filtering.



**To Cite This Article:** Hiba Zuhair, Ali Selamat and Mazleena Salleh., New Discriminative Features for Phishing Filtering. ***Aust. J. Basic & Appl. Sci.,*** *9(26): 83-88, 2015*

## INTRODUCTION

Phishing is a form of cybercrime to deceive users and steal their sensitive information such as password, username, financial information via cyberspace. Phishers usually attack customers of financial organizations like online payment websites and banks; and users of social networks like Facebook and Twitter (Khonji, M., 2012). Therefore, Anti-Phishing Work Group (APWG) reported that phishing causes severe risks economy and cyber-security due to the use of social engineering methods, exploiting poor usability of existing security approaches, and deceive victims directly as phish websites and indirectly as phish emails including links of phish websites. Typically, phish is a fake website looks like a legitimate one that it impersonates. It is created by replicating either the whole or a part of a legitimate website and exploiting its logo, images and URL where it is hosted to trick users for their credentials. Whenever users respond to phishing, they may lose their digital identities and then their money.

For best phishing, phishers exploit various types of deceptive features such as login form features, HTML tags and attributes, JavaScript codes, cookies, embedded objects, SSL certificates and specific indicators in URL address (Khonji, M., 2011; Rajalingam, M., 2012). Furthermore, phishers continually exploit new deceptive features and exploit them to bypass existing phish filters that have not yet identified them before. Evolving such new features over time may become a serious risk for both users' and enterprises defenses and then it may cause tremendous consequences on cybersecurity and economy (Islam, R. and J. Abawajy, 2013; Gowtham, R. and I. Krishnamurthi, 2014).

For the aforesaid observations, this paper mainly aims to improve phish website filtering by eliminating missing ones. Thus, it involves presenting new hybrid features and estimating their utility for better phish website filtering in terms the classification performance over the collected datasets. For this purpose, a recursive evaluation approach with the aid of the feature selection criterion Gain Ratio (GR) and the machine learning classifier namely Support Vector Machine (SVM) is presented. The organization of this paper involves four sections: section II presents a background of phishing filtering approaches and their related studies. Section III describes in details the proposed hybrid features and their recursive evaluation approach. While, section IV presents the application and experimental results of presented approach. Finally, section V draws some conclusion remarks with suggestions for future research direction.

**Corresponding Author:** Hiba Zuhair, Faculty of Computing, Universiti Teknologi Malaysia, Johor, Malaysia Al-Nahrain University, Baghdad, Iraq.

***Related Work:***

In the literature, almost researchers proposed non-classification based and classification based approaches for phish filtering (Whittaker, C., 2014). Non classification based approaches commonly deployed blacklists, whitelists and rule-based filters as toolbars, plug-in and extensions on popular web browsers like Internet Explorer and Firefox. Such filters are Netcraft toolbar, AIWL and PhishCatch (Khonji, M., 2013; Prakash, P., 2010; Basnet, R.B., 2011; Han, W., 2012; Yu, W.D., 2009). These phish websites filters were used to check URLs of examined websites based on sets of valid phish URLs, valid legitimate URLs and standard rules as blacklist, whitelist and rule-based approaches respectively. Even though they are simple short in run time and easy to use by users, those filters require human labor and time for maintenance, update and training to adapt newly emerged phishes. Therefore, they can be easily bypassed by new phishes which makes the user defense more risky.

Whereas classification based approaches mainly deployed hybrid phishing features with the aid of machine learning algorithms to outperform other approaches (Khonji, M., 2013; Xiang, G., 2011; Ramesh, G., 2014; Pan, Y. and X. Ding, 2006; Rajalingam, M., 2012; Gowtham, R. and I. Krishnamurthi, 2014; Whittaker, C., 2010; He, M., 2011; Neumann, J., 2005) And they deployed numerous algorithms of machine learning like Support Vector Machine (SVM), Random Forest (RF) boosting, Bayesian, Decision Tree (DT), etc. such as those in. However they outperform non-classification based approaches substantially; they so far suffer from the misclassification of legitimate websites and phish websites. Furthermore, up to date they fail to adapt newly emerged phishes, provide non-optimum phish filtering, and cause potential risks for users.

On the other hand, researchers studied different categories of features like website content features and URL features. And numerous hybrid features have been deployed along with the development of anti-phishing solutions. For instance, authors in (Gastellier-Prevost, S., 2011) have defined 20 potential hybrid features for both legitimacy and phishness classification. And authors of at Carnegie Mellon University proposed an effective anti-phishing solution "CANTINA+" with the presence of 15 hybrid features and a machine learning classifier.

Furthermore, authors in (Shahriar, H. and M. Zulkernine, 2012) have analyzed XSS-based phishing webpages to define and categorize some JavaScript features into: JavaScript code output, JavaScript code behaviour and JavaScript code content. Also some HTML tags and sets of hyperlinks of phishing webpages hosted in non-English have identified for phishing detection in (Ramesh, G., 2014). Whilst researchers in proposed an anti-phishing solution by using eight features out of World Wide Web Consortium (W3C) standards which have been extracted from the webpage source code file. Authors of have studied and grouped 15 hybrid features into webpage content, webpage identity and login form features.

***Proposed Approach:***

Through previous studies we found that almost hybrid features are related and support each other, to take this phenomenon we develop an evaluation approach to investigate some of these features for their utility for phish website filtering. Based on the literature, hybrid features were extracted from different parts of websites such as URL, HTML document and scripts of JavaScript (Gastellier-Prevost, S., 2011; Gowtham, R. and I. Krishnamurthi, 2014) That implies in many phish websites the webpage content features or URL features are not significant if they were used solely. And exploring additional features can be important and needs further attention (Pan, Y. and X. Ding, 2006; Shahriar, H. and M. Zulkernine, 2012).

Furthermore, features discriminability should be considered as another factor in phish filtering as well as in the determination of an adequate classifier for better phish classification due to the relation between classification accuracies and learning these features [6-8, 13, 14, 19, 20]. To implement this aim, we investigated new hybrid features for their discriminative power in phish website filtering by using a recursive evaluation approach through several steps: features extraction, recursive validation and testing as described in the following subsections. Fig. 1 shows the operational flow of our proposed method in the presence of feature matrix generation and features induction steps. Then two experiments were conducted based on this evaluation approach to identify some which features could be recommended for better phish website classification.

***A. Features Extraction Step:***

On the basis of features listed in Table I, 30 features were extracted as feature vectors for each website in the training dataset and they were kept in features matrix for further steps. Then, all features included in the feature matrix would be fed to the validation step for recursively trained and evaluated with the aid of SVM classifier. Table II list these extracted features with their features indexes.

***B. Features Validation Step:***

Given the feature matrix all features were recursively trained and evaluated by using SVM classifier and Gain Ratio (GR) criterion. This could help to decisively set their utility and whether they could distinguish an input unknown website as phish or non-phish. Furthermore, this evaluation would imply whether the classification with the presence of these features could be with performed with fewer rates of misclassifications. In this step, two metrics

were used to investigate each feature's utility for phishing filtering. These metrics are involved in *Gain Ration (GR)* that are *Information Gain (IG)* and *Entropy* criterion as mentioned in Equations 1 and 2. *IG* partitions samples in the training datasetaccording to a specifically investigated feature to measure the expected reduction of *Entropy*. And *Entropy* identifies the utility of a features set for classification by evaluating the impurity of that set.

$$IG(S,A) = E(S) - \sum_{v \in V(A)} \frac{|S_v|}{|S|} \cdot E(S_v) \quad (1)$$

$$E(S) = \sum_{c \in C} \frac{|S_c|}{|S|} \cdot \log_2 \frac{|S_c|}{|s|} \quad (2)$$

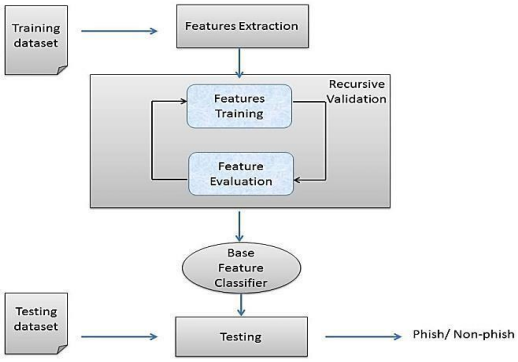


**Fig. 1:** Workflow of the proposed recursive evaluation approach.

**Table I:** New hybrid features with their categories, extraction sources and informative tags.

| Category | Extraction Source | Informative Tags and Indicators |
|---|---|---|
| | | |
| Webpage content | HTML and JavaScript parts | <body> and </body>, <a> and </a>, <base> and </base>, <form |
| | | and </form>,<input type="">,<meta>, <embed>, <object>, |
| (Group 1) | | <applet>, <iframe>, <frame>, <head>, <title>, </head>, </title>, |
| | | <script>,  and herf attribute of <link>, <script> and ActiveX |
| | | controls. |
| Language Independent | URLs | Basename, subdomain, pathdomain, anchors. |
| (Group 2) | | |

In Equation (1), IG metric computes the expected reduction of *Entropy* with respect to feature *A*, where the set *V* includes all possible values of *A*. whilst equation (2) computes the *Entropy* of a set of items *S*
such that all subsets *S* in *C* are represented by *Sc* (Neumann, J., 2005; Shabtai, A., 2009).

A classifier, Support Vector Machine (SVM), is publically used in the areas of pattern recognition and data classification. It is usually used to obtain the optimal separating hyper plane between two classes. And it guaranteed the lowest level of true error because of its generalization ability and handling of high dimensional feature space. SVM estimates a decision function by constructing a linear classification model by non-linearly mapping of input vectors into a feature space with high dimension. Let $\Omega = \{x_i\}_{i=1}^{n}$ be a set of *n* training vectors, and xi is *m*-dimensional vector labelled as $y_i \in \{1, -1\}$ with $y_i = 1$ and $y_i = -1$.

Where $\alpha_i$ and b are obtained by a quadratic algorithm, x is the unlabeled instance and $x_i$ is the training vector of Instances. The function $K(x, x_i)$ maps the space of input instances to higher dimensions where instances are learned individually. For this step, recursive training by SVM produced the base feature classifier that would be forwarded to the next step, testing step. Base feature classifier would decisively distinguish an unknown website as either phish or non-phish.

These labels indicate that *xi* is a member in class 1 and class 2 by using Equation (3) (Huang, H., 2012; He, M., 2011).

$$f(x) = \sum_i \alpha_i \gamma_i K(x, x_i) + b \quad (3)$$

### C. *Testing Step:*

To decisively classify unknown website as phish or non-phish, two class labels were used to represent the classification result: +1 and -1 respectively. In this step, the produced base feature classifier is tested on the feature vector extracted from the input website. Experimentally, a machine learning tool found among those of Waikato Environment for Knowledge Analysis (WEKA) is used to implement SVM algorithm for the purpose of this step. The results of classification would prove the utility of the extracted hybrid features for future predictions of phishness.

**Table II:** Extracted hybrid features in terms of their indexes Group 1.

| index | Features | Index | Features |
|---|---|---|---|
| F1 | Number of (Scripting.FileSystemObject) | F11 | Number of functions' calls |
| F2 | Number of Excel.Application | F12 | Number of script lines |
| F3 | Presence of WScript.shell | F13 | Script line length |
| F4 | Presence of Adodb.Stream | F14 | Existence of long variable names |
| F5 | Presence of Microsoft.XMLDOM | F15 | Existence of long function names |
| F6 | Number of <embed> | F16 | Number of fromCharCode() |
| F7 | Number of <applet> | F17 | Number attachEvent() |
| F8 | link length in <embed> | F18 | Number of eval() |
| F9 | The number of <iframe> | F19 | Number of eval() |
| F10 | The number of <frame> | F20 | Number of escap() |
| Group 2 | | | |
| index | Features | index | Features |
| F21 | Multiple TLD | F26 | Typos in Base name |
| F22 | Brandname | F27 | Shift domain name |
| F23 | Special symbols | F28 | Misleading subdomain |
| F24 | Encoded URL | F29 | Number of dots |
| F25 | IP address instead of domain name | F30 | Post path domain |

***Experimental Results:***

Totally 2000 websites including phish and legitimate were collected by using *PhishTank* and Alexa's top sites during January 2013 to June 2014. Out of those 2000 websites, 130 websites were phish and 1870 websites were legitimate. To experimentally demonstrate the utility of the proposed hybrid features, the results are stated with respect to the feature space and primarily used measurements for performance evaluation. Such measurements are: *TP, FP, FN, Precision, Recall* and *F1-measure*. *TP,True Positive,* indicates the rate of correctly classified phish instances. *FP*, *False Positive* refers to the rateof wrongly classified legitimate instances as phishing ones. Contrarily, *False Negative* (*FN*) indicates the wrongly labeled phish instances as legitimate ones. On the other hand, high *Precision* value states the maximal positively classification of websites. Whilst, the maximal *Recall* value denotes minimal prediction error. Then, *F-measure* is used to harmonically compute the mean of both aforesaid measurements and it denotes the initial phishness indication of the extracted features. Each of *Precision, Recall* and *F-measure* are computed in terms of *TP, FP* and *FN* as parameters. Equations 4, 5 and 6 describe them as follows:

$$Precision = \frac{|TP|}{|TP|+|FP|} \quad (4)$$

$$Recall = \frac{|TP|}{|TP|+|FN|} \quad (5)$$

$$F-measure = 2\times\frac{Precisio\times Recall}{Precision+Recall} \quad (6)$$

Regarding to the above mentioned measurements, the 5-folded classification with all extracted features was conducted and its performance was recursively evaluated over five sets of training and testing phishes. The preliminary classification results are shown in Figure **2**. Next, a set of features were selected with respect to their high values of *GR* to be trained and tested in the presence of the same collected dataset. The selection of highly *GR* valued features is conducted to show whether the selected features could be discriminative enough to obtain the best base feature classifier. **Figure 3** illustrates those features in the form of their GR values.

In **Figure 2,** if these features were ranked according to their calculated *GR* values then the highest features in ranking would be features *F21, F1, F2, F3, F24, F29, F4, F13, F5, F5, F8, F15 and F13*. That implies the maximal feature in *GR* value could be utilized for better phish website filtering. Whereas features with minimal *GR* values would be the least ones in their utility for the same purpose and they may possibly cause noise to the classifier and then inaccurate classification results. Unlike the

original set of extracted features, when those highest 12 features in ranking were applied for recursive training and testing, they outperformed their competitors by producing more efficient base feature classifier as presented in **Figure 2**. Consequently, these features could train the base feature classifier quicker than the original set because of low dimensionality and high significance.

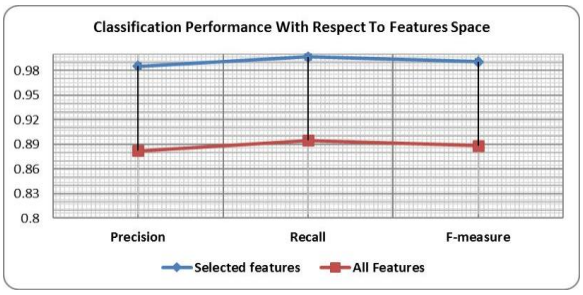


**Fig. 2:** Results of classification performance with the use of all extracted features and the selected features in terms ofPrecision, Recall and F-measure.

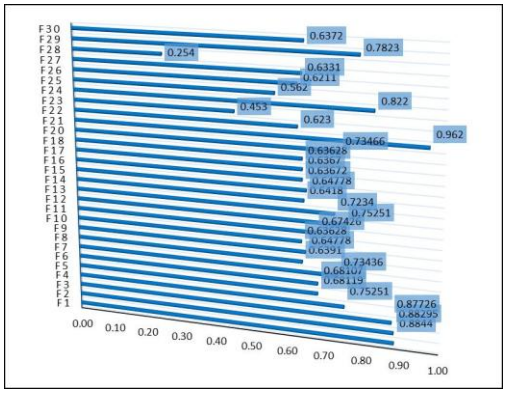


**Fig. 3:** Extracted features with their*GR*values.

***Concluding Remarks And Future Work:***

This paper presents an experimentally recursive evaluation approach to provide some supplementary hybrid features that could help in phish website filtering with the aid of machine learning classifier and feature evaluation criterion. The extracted 30 hybrid features include webpage content features and URL features. Experimentally, the utility of these features was investigated and the results concluded that a set of ten selected features in 5-fold validations outperforms the original set of extracted features. The results motivate us for further work to propose an automatic feature selection approach to investigate these features by using distinct features selection criteria.

## REFERENCES

Khonji, M., Y. Iraqi, A. Jones, 2013. "Phishing detection: a literature survey," Comm. Surveys & Tutorials, 15(4): 2091–2121.

Phishing archive. anti-phishing working group, http://www.apwg.org

Khonji, M., Y. Iraqi, A. Jones, 2011. " lexical url analysis for discriminating phishing and legitimate websites. In proceedings of the 8th annual collaboration, electronic messaging, anti-abuse and spam conference. Acm.

Prakash, P., M. Kumar, R.R. kompella and M. Gupta, 2010. "Phishnet: predictive blacklisting to detect phishing attacks," in infocom, proceedings ieee, 1-5.

Gastellier-Prevost, S., G.G. Granadillo and M. Laurent, 2011. "Decisive heuristics to differentiate legitimate from phishing sites," in Network and Information Systems Security (SAR-SSI), Conference on, 1-9.

Xiang, G., J. Hong, C.P. Rose and L. Cranor, 2011. "CANTINA+: a feature-rich machine learning framework for detecting phishing web sites," ACM Transactions on Information and System Security (TISSEC), 14-21.

Shahriar, H. and M. Zulkernine, 2012. "Trustworthiness testing of phishing websites: A behavior model-based approach," Future Generation Computer Systems, 28: 1258-1271.

Ramesh, G., I. Krishnamurthi and K. Kumar, 2014. “An efficacious method for detecting phishing webpages through target domain identification”. Decision Support Systems, **61**: 12-22.

Alkhozae, M.G., O.A. Maratfi, 2011. "Phishing websites detection based on phishing characteristics

in the webpage source code," Int. J. Inform. Communication Technology Research.

Pan, Y. and X. Ding, 2006. "Anomaly based web phishing page detection," in Computer Security Applications Conference, ACSAC'06. 22nd Annual, 381-392.

Rajalingam, M., S.A. Alomari and P. Sumari, 2012. "Prevention of Phishing Attacks Based on Discriminative Key Point Features of WebPages". International Journal of Computer Science and Security (IJCSS), **6**(1): 1.

Islam, R. and J. Abawajy, 2013. "A multi-tier phishing detection and filtering approach," Journal of Network and Computer Applications, 36: 324-335.

Gowtham, R. and I. Krishnamurthi, 2014. "A comprehensive and efficacious architecture for detecting phishing webpages," Computers & Security, 40: 23-37.

Whittaker, C., B. Ryner and M. Nazif, 2010. "Large-Scale Automatic Classification of Phishing Pages," in NDSS.

Basnet, R.B., A.H. Sung and Q. Liu, 2011. "Rule-based phishing attack detection," in International Conference on Security and Management (SAM 2011), Las Vegas, NV.

Han, W., Y. Cao, E. Bertino and J. Yong, 2012. "Using automated individual white-list to protect web digital identities," Expert Systems with Applications, 39: 11861-11869.

Yu, W.D., S. Nargundkar and N. Tiruthani, 2009. "Phishcatch-a phishing detection tool," in Computer Software and Applications Conference, COMPSAC'09. 33rd Annual IEEE International, 451-456.

Huang, H., L. Qian and Y. Wang, 2012. "A SVM-based technique to detect phishing URLs," Information Technology Journal, 11: 921-925.

He, M., 2011. "An efficient phishing webpage detector. Expert Systems with Applications", **38**(10): 12018-12027.

Neumann, J., C. Schnörr, G. Steidl, 2005. "Combined SVM-based feature selection and classification" Machine Learning, 61(1-3): 129-150.

Shabtai, A., R. Moskovitch, Y. Elovici, C. Glezer, 2009. "Detection of malicious code by applying machine learning classifiers on static features: A state-of-the-art survey". Information Security Technical Report, 14(1): 16-29.

PhishTank. Available from:http://www.phishtank.com/what_is_phishing.php

Alexa Top Sites available in http://www.alexa.com/site/ds/top_sites?ts_mode=global&lang=none