Volume 6, Issue 6, June 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Enhancement in Phishing Detection Using Features Clustering

Amol C. Jadhav*, A. M. Pawar
Computer Department, Zeal College of Engineering, SPPU, Pune,
Maharashtra, India

Abstract— In day today's life without internet we can't do anything so the web services are mostly used for the day today's activity .So for phishing a lure technique came in to picture. Forged website such as Phishing website looks similar to realistic so the fatality have been increasing due to the un availability of using blacklist to finding the phishing so in this paper we have proposed the new technique that apply fuzzy logic based The technique can detect 98% phishing sites this technique is based on URL feature based we compared this with the technique with the classifies URLs automatically based on the host-based features. The usability, scalability problems, learning. The system achieves 93-97% accuracy and detected a large number of phishing hosts, at the same time maintaining a modest false positive rating. The data is examining which is raw data, and the value of various feature subsets is assessed. The relevance of bigrams is useful for assessing, and increased strength with the use of the chi-squared and information gain attribute evaluation methods used for finding the some features which are complex as cluster labels from the basic features is to be considered . So in this paper we also extended the discussion not only the classification of characteristics but also the clustering of the characteristics, how to draw complex features into cluster labels from the basic features is also taken into consideration.

Keywords— Phishing, Detection, Web Features, Clustering, Fuzzy logic.

I. INTRODUCTION

Phishing is the challenge to attain susceptible information such as passwords, usernames and atm card, credit card details (indirectly, money), often for malicious reasons, by hidden as a reliable entity in an electronic communication .Phishing is one of the techniques used by phishers in the purpose of getting the personal information. Phishing website is a counterfeit website that looks similar to genuine site in terms of interface and uniform resource locator (URL) address. Therefore, the numbers of fatality have been increasing due to non-sufficient methods using blacklist to detect phishing. In this paper we are proposing a new technique that will apply fuzzy logic based techniques. On the features of URL to detect phishing sites. This proposed technique evaluation with the dataset of 11,660 phishing sites and 5,000 genuine sites. The results show that the technique can detect over 98% phishing sites. Now a day's Web services are available for social networking, gaming, and online banking have rapidly evolved in the performing everyday tasks by people . So the result of all these, a large information is uploaded on a daily basis to the Web. Obviously These are the opportunities for criminals to upload malicious content. For these the extensive research, email based spam filtering techniques are used but are unable to protect other web services. Therefore, a counter measure should be taken against the generalization of web services to protect from phishing hosts. There is increasingly growth of the phishing websites seems to be astonishing. In United States, phishing causing billion loss in past years. In 2011, near about 85% of Americans and of Europeans regularly shopped online (Fortune Magazine, 2011). Meanwhile, phishers increases rapidly phishing websites in terms of quantity and quality. Therefore, the risk of theft user information is tremendously lofty. Because of these things only, detection of phishing web sites is very urgent, complicated work and extremely important problem in the web environment. Recently, there have been many techniques which are against phishing based on the site characteristics, as URL of website, content of website, combining both the website URL and that the algorithm TF-IDF is used which is based on 27 different features of webpage. The technique is useful to detect 97% phishing sites with 6% false positives. Although this technique is efficient, the time taken by the system for extracting 27 features of webpage is too much to meet real time demand and performance of the system in terms of the speed is not up to the mark as not in prescribed time and some features are not useful for improving the fake websites detection accuracy. Similarly, Cantina+ [6] In this paper used machine learning techniques which is based on 15 features only of webpage and only six of 15 features are efficient for phishing detection such as, Bad action, bad form fields, Non-matching URLs, Page in top search results, Search copyright brand plus domain and Search copyright brand plus hostname so the problem of real time demand is reduced. In [7], the author used the search engine directly so the URL to detect phishing sites automatically by verifying and extracting different terms of a URL through search engine. As in this paper proposed a new interesting technique not focused on the detection rate so, the detection rate is quite low (54. 3%). The technique [8] developed a content based Technique named CANTINA to detect phishing called which considers the Google Page Rank value of a page; however, the evaluation dataset is quite small. The source code characteristic is used to detect phishing sites in [9]. The authors in [10] have proposed same technique as discussed in the paper [5] fuzzy technique based on 27 features of webpage, only difference is the classified into 3 layer. Each feature has three linguistic values: low, moderate, high. The fuzzy technique has built a rule set, trapezoidal, triangular membership

functions. The achieved website phishing rate of the technique is 86%. However, it exist many drawbacks in [10]. In this the rule sets are not objective and greatly depend on the builder.

Secondly, the weight is used without any clarification. Finally, the proposed heuristics are Not most favourable and really effective. In the method previously used techniques, the URL has a minor role in detecting phishing websites. In this paper, we focus on URL features and apply the fuzzy logic technique to detect phishing sites. Our proposal differs with others in two aspects: i) URL has an important role in detecting phishing websites; ii) fuzzy logic is used as a mathematical method to detect phishing websites. Therefore, our contribution can be summarized as follows: first, we have proposed the new heuristics to detect phishing website more effectively and rapidly. Then, the new fuzzy-based approach has been proposed such that the rule set is not utilized. Hence, the result will be more precise and objective. Finally, the threshold values used in the membership functions are derived from the big data set so that the model is still equivalent for the new data set 530 [11] In this paper Author used a set of lexical as well as host-based features to perform classification of URL. Their classification accuracy of around 95%. However, their work uses raw hostname features [11] in additional bigrams the hostname part for each URL is characterized. Bigrams are the indicators for characterizing URLs based on the results related by Blum et al. [12]. In this paper the Author used idea of using n-grams it is supported by Krueger et al. [20] in this paper an N-gram Centroid Anomaly Detector used to analyze all possible n-grams beside byte sequences.

II. RELATED WORKS

The phishing detection techniques always classified such as 1] blacklist, 2] heuristic 3] machine learning. In blacklist approach, the phishing detection technique [I] [2] [3] [4] maintains a list of phishing websites so it can be called blacklist. However, the blacklist method is also not sufficient due to the rapidly increasing in the number of phishing sites. Therefore, the heuristic and machine learning technique approaches have received more attraction of researchers.978-1-4799-5051-5/14/$31.00 ©2014 IEEE Cantina [5] presented the webpage features in that the algorithm TF-IDF is used which is based on 27 different features of webpage. The technique is useful to detect 97% phishing sites with 6% false positives. Although this technique is efficient, the time taken by the system for extracting 27 features of webpage is too much to meet real time demand and performance of the system in terms of the speed is not up to the mark as not in prescribed time and some features are not useful for improving the fake websites detection accuracy. Similarly,

Cantina+ [6] In this paper used machine learning techniques which is based on 15 features only of webpage and only six of 15 features are efficient for phishing detection such as, Bad action, bad form fields, Non-matching URLs, Page in top search results, Search copyright brand plus domain and Search copyright brand plus hostname so the problem of real time demand is reduced. In [7], the author used the search engine directly so the URL to detect phishing sites automatically by verifying and extracting different terms of a URL through search engine. As in this paper proposed a new interesting technique not focused on the detection rate so, the detection rate is quite low (54. 3%). The technique [8] developed a content based Technique named CANTINA to detect phishing called which considers the Google Page Rank value of a page; however, the evaluation dataset is quite small. The source code characteristic is used to detect phishing sites in [9]. The authors in [10] have proposed same technique as discussed in the paper [5] fuzzy technique based on 27 features of webpage, only difference is the classified into 3 layer. Each feature has three linguistic values: low, moderate, high. The fuzzy technique has built a rule set, trapezoidal, triangular membership functions. The achieved website phishing rate of the technique is 86%. However, it exist many drawbacks in [10]. In this the rule sets are not objective and greatly depend on the builder.

Secondly, the weight is used without any clarification. Finally, the proposed heuristics are not most favorable and really effective. In the method previously used techniques, the URL has a minor role in detecting phishing websites. In this paper, we focus on URL features and apply the fuzzy logic technique to detect phishing sites. Our proposal differs with others in two aspects: i) URL has an important role in detecting phishing websites; ii) fuzzy logic is used as a mathematical method to detect phishing websites. Therefore, our contribution can be summarized as follows: first, we have proposed the new heuristics to detect phishing website more effectively and rapidly. Then, the new fuzzy-based approach has been proposed such that the rule set is not utilized. Hence, the result will be more precise and objective. Finally, the threshold values used in the membership functions are derived from the big data set so that the model is still equivalent for the new data set 530 [11] In this paper Author used a set of lexical as well as host-based features to perform classification of URL. Their classification accuracy of around 95%. However, their work uses raw hostname features [11] in additional bigrams the hostname part for each URL is characterized. Bigrams are the indicators for characterizing URLs based on the results related by Blum et al. [12]. In this paper the Author used idea of using n-grams it is supported by Krueger et al. [20] in this paper an N-gram Centroid Anomaly Detector used to analyze all possible n-grams beside byte sequences.

III. OVERVIEW OF FEATURES

All paragraphs must be indented. All paragraphs must be justified, i.e. both left-justified and right-justified.

1. Lexical Features

URLs [Phishing] and domains exists characteristics which are different from benign URLs and domains [13]. Website Criminals are known to use various innovative methods to lure unsuspecting users. In recent times, criminals have some methods which is also known as URL hijacking called Typo squatting. The method is useful for finding the users who incorrectly type the address of a website on their web browser, such users gets targeted. Just consider, the user type www.pavpal.com, www.paypayk.com or instead of www.paypal.com. In this type of cases, users led to an alternative

website that closely duplicates the original website, and duplicated websites will ask users to enter login details as well as financial credentials. For the analysis of such Phishing URLs lexical content captured to find incorrectly spelt tokens. The result of this is the user can be alerted. Host-based content helps in identifying such websites by using network infrastructure information of other phishing campaigns that are closely related to the website being analyzed. Phishers are known to target specific brands (such as Amazon, PayPal) during specific times. An online learning algorithm can retrain the model continuously and update itself based upon the emerging trends in phishing URLs. In Maetal.'s work. In some work length of the URL is considered the authors extract continuous valued features, and the number of dots present in the URL. Various other authors also support using URL length as a feature since phishing URLs exhibit a certain pattern of URL length. They usually tend to have different lengths when compared to other URLs and domains on the Internet [13]. In the current work, token based bigrams such as Google .com, are used as hostname part characterizing the domain. The usage of bigrams has gained popularity as 242 suggested by Blum et al.'s work [15]. From an information retrieval perspective, there exists a lack of standardization in the writing system of URLs [16]. URLs are not necessarily constructed using proper terms in English and often times they are just a random string of characters. The motivation for adding bigrams to Ma et al.'s [14] work emerges from the idea that phishing URLs can exhibit a certain pattern of random character strings occurring in certain combinations. The subtlety of such random occurrences in character strings can be captured by the usage of bigrams.

2. Host-Based Features

Previous works by researchers has shown that host-based features are to be used in order to characterize phishing URLs effectively. The author's gets motivated for using URL host-based features came from Ma et al.'s work [17]. They collect important metadata for a URL, such as A (IP address of the URL), MX (IP address of the mail exchanger), NS (IP address of the name server), and PTR (pointer) records from the Domain Name System (DNS).The idea behind using these records is that phishing websites have exhibited a pattern of being hosted in a particular "bad" portion of the Internet [14]. The pointer record enables reverse Domain Name System lookups. The presence of a PTR record indicates that the hostname is well established [14]. Autonomous System (AS) numbers for these records would further indicate the presence of ISPs (Internet Service Providers) that are known to host phishing websites. Mark at al.'s research [20] observes that criminals who register domains on the Internet for malicious purposes often operate the domains using related sets of name servers. As a result, identifying the name servers commonly used by perpetrators would serve as a useful indicator for identifying a phishing website. Ma et al. [14] some papers support the idea of identifying related sets of name servers as name server records represent the DNS infrastructure that leads the user to a phishing website. Further, the infrastructure may also be hosted on ISP's that are known to host phishing websites. MX records for known phishing sites are composed. A fresh web site in the future has an associated mail server that is present in the collected set, and then the new website could be classified as phishing [14]. Blum et al. [15] Author of this paper completely discard using host-based features and restricted their research to lexical features and as they consider it to be a vulnerable loom. Although these can identify a large number of phishing URLs, they could have identified a larger percentage of phishing URLs had they included a host-based feature, which has been supported by the results obtained from many researchers.

The model built in this work and the models built by related work are all susceptible to criminals who can possibly evade the system by studying it in detail and crafting websites and URLs in a manner that the system would fail to detect it accurately. To overcome this drawback, drawing complex features such as cluster labels from the basic features would be considered. So here we will discuss drawing the drawing complex features such as cluster labels from the basic features would be considered.

IV. FEATURE EXTRACTION OF PHISHING WEBSITES

Phishing website representation URL of the website uses several feature extraction methods, user interface, associated Webpages of the website , webpage block, page layout, and whole style, term f given webpage with the TF-IDF score etc. considering the expression ability of the website and the complexity for the categorization inputs. So in these different we can extract features as per different authors in this paper, we extract the term frequencies from the Webpages of their relating to website.

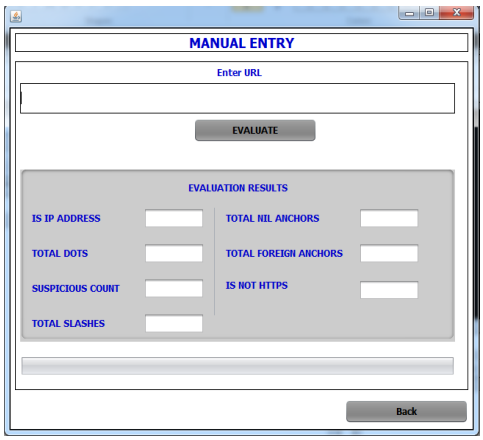

Fig. 1 Feature Extraction

Firstly extract the terms form the website html pages on basis of java script. The description of the extraction is illustrated as follows.

- In the first step select script between tags in the HTML file, this is in the code form.
- Transform the all script in the case upper alphabets. Then remove the extra symbols except the upper case alphabets.
- After that define the minimum word length l1 and maximum word length l2. And removing all the extras word, which are not between l1 and l2.
- Then let there be a fix term frequency f which can be greater than f in files.
- Convert these term frequency into TF and TFIDF.

V. SYSTEM ARCHITECTURE

Show and we briefly describe each component below. The architecture of the detection based on kernel k-means clustering shown in fig 4.1Term-frequency feature extractor: In phishing website categorization, the system first uses the term-frequency to extract the terms form the Webpages feature extractor of the collected phishing website and transforming the data in feature vector of term- frequency. These vectors are stored in database. Transaction data can be simply converted to relational data if necessary.

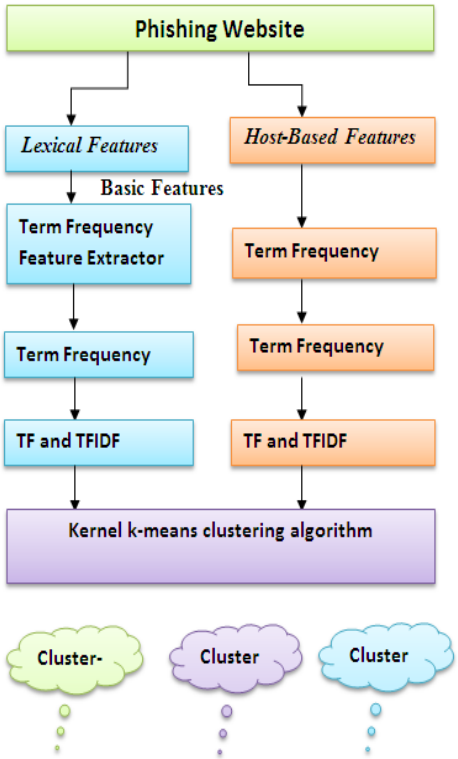

Fig. 2. System Architecture

VI. PROPOSED METHOD

We proposed A system in that Kernel k-means clustering to categorize features of phishing in this paper, website. Kernel k-means Tzortzis et al. is a generalization and also standard k-means clustering algorithm where vector are mapped from vector space to a higher dimensional feature space through a kernel function and then k-means is applied into feature space. The outcome of the kernel k-means clustering algorithm is in linear separators of feature space which exchange in to nonlinear separators in vector space. Thus, kernel k-means avoids the disadvantage of linearly separable clusters in vector space which is not in k-means. The objective function that kernel k-means minimize clustering error in feature space. We can identify a kernel matrix KϵRN×N , where Kij = Ø(xi)T Ø(xj) and taking advantage of the kernel trick, we can calculate the squared Euclidean distances in (2) without explicit knowledge of the transformation using (3) any positive-semi definite matrix (PSD) can be used as a kernel matrix. Observe that in this case cluster centers mk in feature space cannot be calculated. In the kernel k-means clustering, the kernel function K(xi, xj) is used to directly present the inner products in feature space without explicitly defining transformation, therefore Kij = K(xi, xj). Kernel k-means is described in Algorithm.

VII. EVALUATION OF THE PROPOSED METHOD FOR PHISHING WEBSITE CATEGORIZATION

Using phishing websites and their corresponding 1500 webpage's collection obtained from the Phishload - Tables explained.html. We construct kernel k-means clustering on TF and TF-IDF. We examine that the phishing website categorization result are best, than methods which are used in earlier work in this work features which are complex as categorized as cluster labels and found from the basic features is to be considered . So in this paper we also extended the discussion not only the classification of characteristics but also the clustering of the characteristics, how to draw complex features into cluster labels from the basic features is also taken into consideration.

Ensemble clustering algorithm. It should be pointed out that in some cases, categorizing a phishing website to a certain family is still the prerogative of Internet security experts. For example, some of the phishing websites are prize-winning fraud websites and share similar shape of term-frequency patterns, thus may be categorized to the identical family, according to their exact intents divided into different families. On the contrary, there are some metamorphic phishing websites, like selling and social networking sites fraud which may differ from term representations but they are in the same family. The result of the kernel based system compare with Ensemble clustering (Hierarchical clustering and k-method) the proposed method on basis of the accuracy and error rate for phishing website categorization.

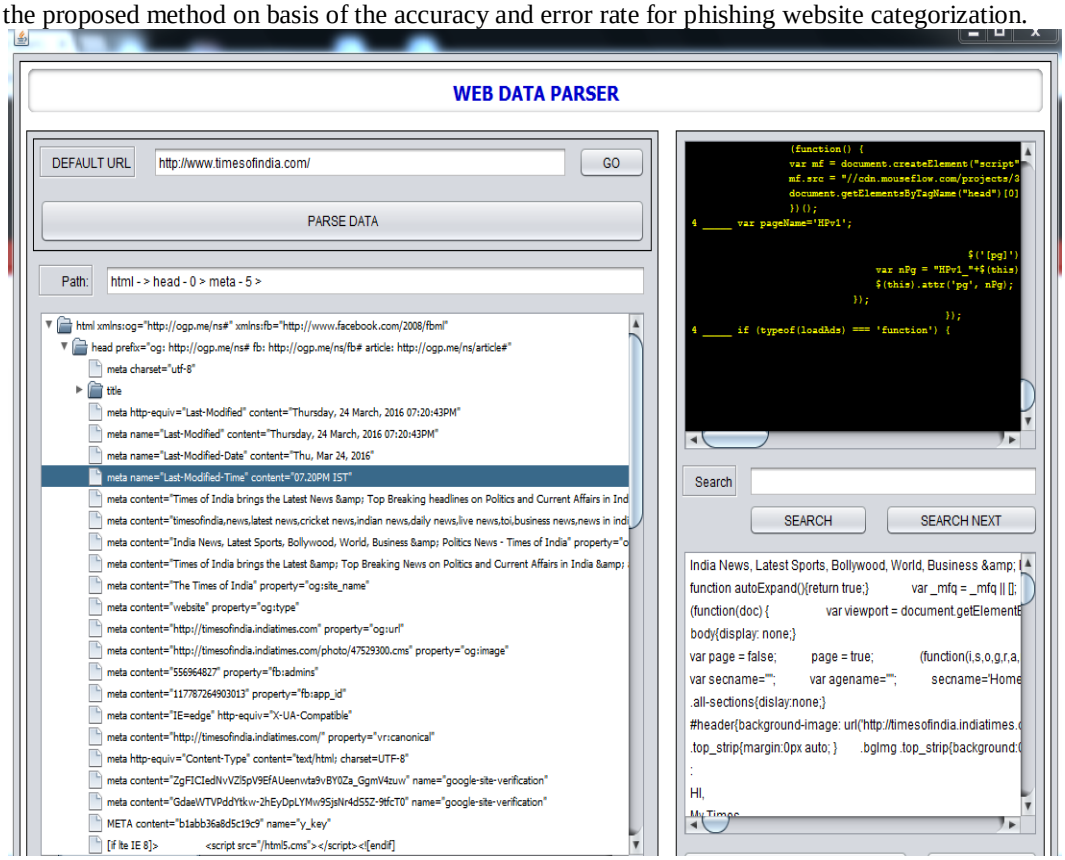

Fig. 3. Web Data Extraction

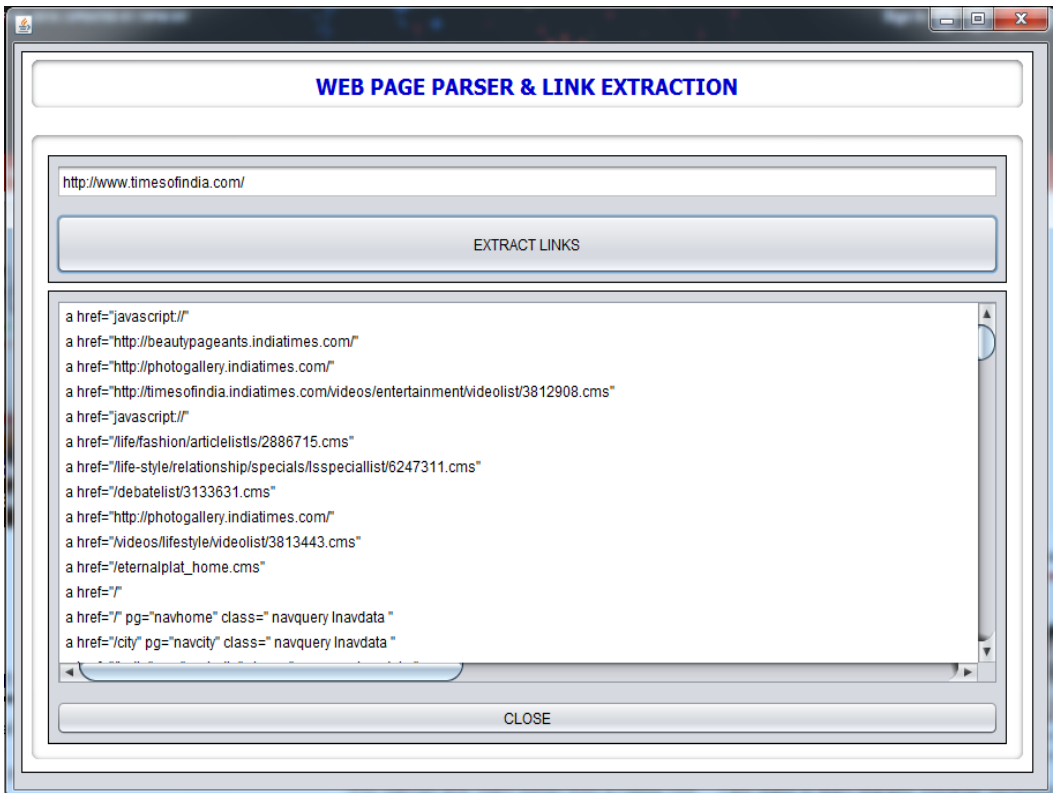

Fig. 4. Web Link Extraction

VIII. CONCLUSION

In this paper, firstly extract feature such as lexical and host-based from the phishing website sample through term frequency and then we have developed a system which is applied sample categorization or phishing website categorization into families that share some common traits by kernel k-means clustering method. The studies on large and standard data set collected from data center and Phish load our system will performs well for real phishing website feature categorization applications. In this technique the categorization accuracy and error rate of phishing website sample improved as compare to ensemble clustering technique. The accuracy and error rate will improve 10 to 20 % phishing website feature categorization. In the future work that can enhance with three fields firstly in this work the feature extracted by different method for phishing website, second there are many clustering algorithm which can be applied to malware and phishing website categorization and third one is that can include anomaly detection with phishing website feature categorization.

ACKNOWLEDGMENT

We are pleased to express our sentiments of gratitude to all who rendered their valuable guidance to us. We express our appreciation and thanks to the Principal of our college. We are also thankful to the Head of Department and guide Prof. A. M. Pawar. We thank to the reviewers for their valuable comments.

REFERENCES

[1] PhishTank. (2013, Nov.) Statistics about phishing activity and phishtank usage. [Online]. Available: http://www.phishtank.com/stats/2013/01/, IEEE Std. 802.11, 1997.

[2] D. Goodin. (2012) Google bots detect 9,500 new malicious websites every day. [Online]. Available: http://arstechnica.com/security/2012/06/google-detects-9500-newmalicious-websites-daily/

[3] S. Sheng, B. Wardman, G. Warner, L. Cranor, J. Hong, and C. Zhang. (2009) an empirical analysis of phishing blacklists. [Online]. Available:http://ceas.cc/2009/papers/ceas2009-paper-32.pdf

[4] Google. (2011, Aug.) Google safe browsing api. [Online]. Available:http://code.google.com/apis/safebrowsing/

[5] McAfee. (2011, July) Mcafee site advisor. [Online]. Available:http://www.siteadvisor.com

[6] Y. Zhang, J. I. Hong, and L. F. Cranor, "Cantina: a content-based approach to detecting phishing web sites," in The 16th international conference on World Wide Web, 2007, pp. 639–648.

[7] G. Xiang, J. Hong, C. P. Rose, and L. Cranor, "Cantina+: a feature-rich machine learning framework for detecting phishing web sites," ACM Transactions on Information and System Security, vol. 14, no. 2, pp.1–28, Sept. 2011.

[8] M. E. Maurer and D. Herzner, "Using visual website similarity for phishing detection and reporting," in CHI '12 Extended Abstracts on Human Factors in Computing Systems, 2012, pp. 1625–1630.

[9] A. Sunil and A. Sardana, "A pagerank based detection technique for phishing web sites," in IEEE Symposium on Computers & Informatics, 2012, pp. 58–63.

[10] M. G. Alkhozae and O. A. Batarfi, "Phishing websites detected based on phishing characteristic in the webpage source code," in International Journal of Information and Communication Technology Research, vol. 1,no. 6, Oct. 2011, pp. 283–291.

[11] MA, J., Saul, L.K., Savage, S., and Voelker, G.M., "Learning to Detect Malicious URLs". In ACM Transactions on Intelligent Systems and Technology. 2, 3, Article 30 April 2011, ACM New York, NY, USA, pp. 1245-1254. DOI=http://dl.acm.org/citation.cfm?id=1961202.

[12] Blum, A., Wardman, B., Solorio, T., and Warner, G., "Lexical Feature Based Phishing URL Detection Using Online Learning". In AISec '10 Proceedings of the 3rd ACM workshop on Artificial Intelligence and Security, Illinois, USA, 2010, ACM New York, NY, USA, pp. 54-60. DOI=http://dl.acm.org/citation.cfm?id=1866423.1866434&coll=DL&dl=ACM&CFID=237444071&CFTOKEN=87140042.

[13] McGrath, K., and Gupta, M., Behind Phishing: "An Examination of Phisher Modi Operandi". in LEET'08 Proceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats, California, USA, 2008, USENIX Association Berkeley, CA, USA. DOI=http://dl.acm.org/citation.cfm?id=1387713.

[14] MA, J., Saul, L.K., Savage, S., and Voelker, G.M., "Learning to Detect Malicious URLs". in ACM Transactions on Intelligent Systems and Technology. 2, 3, Article 30 April 2011, ACM New York, NY, USA, pp. 1245-1254. DOI=http://dl.acm.org/citation.cfm?id=1961202.

[15] Felegyhazi, M., Kreibich, C., and Paxson, V., "On the Potential of Proactive Domain Blacklisting". in LEET'10 Proceedings of the 3rd USENIX conference on Large-scale exploits and emergent threats: botnets, spyware, worms, and more, California, USA, 2010, USENIX Association Berkeley, pp. 6-6. DOI= http://dl.acm.org/citation.cfm?id=1855692

[16] Blum, A., Wardman, B., Solorio, T., and Warner, G., "Lexical Feature Based Phishing URL Detection Using Online Learning". in AISec '10 Proceedings of the 3rd ACM workshop on Artificial Intelligence and Security, Illinois, USA, 2010,ACM New York, NY, USA, pp. 54-60. DOI=http://dl.acm.org/citation.cfm?id=1866423.1866434&coll=DL&dl=ACM&CFID=237444071&CFTOKEN=87140042.

[17] Manning, C., Prabhakar, R., and Schutze, H. Introduction to Information Retrieval. Cambridge University Press, NY, 2008.

[18] Ma, J., Saul, L., Savage, S., and Voelker, G., "Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs". in KDD'09 Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, 2009, ACM New York, NY, USA, pp. 1245-1254. DOI= http://dl.acm.org/citation.cfm?id=1557019.1557153&coll=DL&dl=ACM&CFID=236703599&CFTOKEN=16695054.

[18] S. Haykin, "Neural networks: A comprehensive foundation," in The Knowlegde Engineering Review Vol 13, 1999, pp. 409–412

[8] "A Proposal of the AdaBoost-Based Detection of Phishing Sites", Daisuke Miyamoto, Hiroaki Hazeyama, and Youki Kadobayashi (Internet Engineering Laboratory, Graduate School of Information Science, Nara Institute of Science and Technology, Japan),In JWIS, August 2007.

[9] Tyler Moore, Richard Clayton, and Henry Stern. "Temporal correlations between spam and phishing websites". In Proceedings of the 2nd USENIX conference on Large-scale exploits and emergent threats: botnets, spyware, worms, and more (LEET'09). USENIX Association, Berkeley, CA, USA, 5-5, 2009.

[10] R. Dhamija and J. D. Tygar. "The battle against phishing: Dynamic security skins". In Proceedings of the 2005 symposium on Usable privacy and security, New York, NY, pages 77–88. ACM Press, 2005.

