Phishing URL detection using URL Ranking

Mohammed Nazim Feroz
Texas Tech University
Computer Science
Lubbock, USA
mohammed.n.feroz@ttu.edu

Susan Mengel
Texas Tech University
Computer Science
Lubbock, USA
susan.mengel@ttu.edu

***Abstract*— The openness of the Web exposes opportunities for criminals to upload malicious content. In fact, despite extensive research, email based spam filtering techniques are unable to protect other web services. Therefore, a counter measure must be taken that generalizes across web services to protect the user from phishing host URLs. This paper describes an approach that classifies URLs automatically based on their lexical and host-based features. Clustering is performed on the entire dataset and a cluster ID (or label) is derived for each URL, which in turn is used as a predictive feature by the classification system. Online URL reputation services are used in order to categorize URLs and the categories returned are used as a supplemental source of information that would enable the system to rank URLs. The classifier achieves 93-98% accuracy by detecting a large number of phishing hosts, while maintaining a modest false positive rate. URL clustering, URL classification, and URL categorization mechanisms work in conjunction to give URLs a rank.**

***Index Terms*—Clustering, Feature Vector, Classification, Web Categorization, URL Ranking**

I. INTRODUCTION

Over the last few years, the Web has seen a massive growth in the number and kinds of web services that include social networking, forums, blogs, and video sharing sites [1]. As these web services drive new opportunities for people to interact, they also create new opportunities for criminals. The fact that Google detects about 300,000 malicious websites per month is a clear indication and proof that these opportunities are extensively used by criminals [2]. Phishing websites are employed to steal personal information, such as credit cards and passwords, and to implement drive-by downloads. Criminals also employ phishing in order to lure users into visiting their fake websites [3].

The vector that is used to trap users is a Uniform Resource Locator (URL). The user must perform sanity checks before clicking a URL, such as paying close attention to the spelling of the website's address and evaluating the associated risk that might be encountered by visiting the URL [4]. Recent work has shown that security practitioners have developed techniques, such as blacklisting, to protect users from phishing websites. In blacklisting, a third party compiles the names of known phishing websites. Besides having a minimal query overhead, the technique does not provide complete protection as no blacklist can be comprehensive and up-to-date. As a result, the user may click on a link that directs to a phishing website before the link appears on a blacklist [4]. Security researchers have done extensive research to detect accounts on social networks used for spreading messages that lure the recipient to a potential phishing website [6]. Their technique of using honey-profiles to identify single spam bots as well as large-scale campaigns does not provide complete protection for users. Social networks consist of real-time interaction, but the technique for detecting fraudulent accounts can incur delays due to the need for building a profile of inappropriate activity. The current work proposes a system capable of clustering, classifying, categorizing, and ranking URLs in real-time while adapting to new and evolving trends in URL characteristics.

An attempt to improve existing techniques for detecting phishing URLs is made by adding novel features (i.e. predictor variables), and making the learning process more efficient by using Mahout, an Apache project comprising of scalable machine learning algorithms [8]. Online algorithms adapt to changes quickly and are a better choice since the values for predictor variables of URLs can change over time [9]. One major contribution of the current paper is the implementation of a hybrid approach combining both clustering and classification. Clustering is performed prior to classification, and cluster IDs from the clustering step are added as a predictor variable to expand each feature vector present in the dataset. The process of performing clustering and using cluster IDs (or cluster labels) helps in obtaining a hyperplane with a larger margin which in turn results in higher classifier efficiency due to better generalization. Another major contribution of the current paper is URL categorization by Microsoft Reputation Services (MRS) [18]. The categories obtained from MRS are used to perform ranking of URLs by using a novel URL Ranking approach, which consists of using the classification result and an internal scale (Red/Yellow/Green) of severity derived for each URL.

II. RELATED WORK

This section discusses related methodologies used by researchers who have tried to solve the problem of phishing URL detection and classification. The work by Ma et al. [7] is most closely related to the current work. Their work achieves classification accuracy of around 95% by extracting lexical and host-based features from URLs. The set of lexical features used by Ma et al. [7] is closely related to the set of lexical features used in the current work with the inclusion of bigrams for

978-1-4673-7278-7/15 $31.00 © 2015 IEEE
DOI 10.1109/BigDataCongress.2015.97

characterizing the hostname portion of each URL. Bigrams are known to be useful indicators for characterizing URLs based on the results related by Blum et al. [11]. The current work builds on Ma et al.'s work in [7] by using online learning algorithms rather than batch learning algorithms, and by using a hybrid approach consisting of clustering and classification, followed by a URL ranking mechanism.

Blum at al. [11] use online learning in order to perform URL classification. Their work uses a similar set of lexical features. However, the Blum et al.'s [11] work totally discards the use of host-based features. Their classifier achieves an accuracy of around 97% if quality training data can be provided. The current work builds on Blum et al.'s [11] work by adding host-based features. Host-based features are known to be reliable indicators for characterizing phishing URLs [10].

The current work is a continuation of previous research [19] where a robust classification system was built and compared against other classification algorithms. The datasets were examined and the J48 tree [17] was used to assess the effectiveness of lexical, bigram, and host-based features. Chi-Squared and Information Gain attribute evaluation were performed to assess and strengthen the relevance of bigrams. Rule generation was included to emphasize the inner workings of the classification system. The current work builds on the existing system [19] by introducing URL clustering, categorization, and ranking mechanism.

The problem of online fraud is exacerbated by the fact that most end users make security decisions, based on a very rudimentary understanding of risk [20]. Srikwan et al. [20] state that phishing is both a matter of technology and psychology and most people want to trust what they see. URL Feedback that is too simple fails to capture the problem well, while that which is too complex faces the risk of not being understood by a novice end user. To overcome the drawback and make feedback more meaningful, a combined system that gives succinct URL classification feedback along with a color band for the URL rank is implemented.

III. OVERVIEW OF FEATURES

A. Lexical Features

Phishing URLs and domains are known to exhibit characteristics that are different from other benign URLs and domains [5]. For instance, in Typosquatting, the user might type www.paypak.com instead of www.paypal.com where users might be led to an alternative website that closely duplicates the original website. The duplicated websites may ask users to enter login details or financial credentials. Phishing URLs of this type are usually captured by analyzing the lexical content to find incorrectly spelt tokens; so, the user can be alerted. Criminals 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.

B. Host-Based Features

The motivation behind using URL host-based features comes from Ma et al.'s work [7]. They collect significant 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 [10]. The PTR record enables reverse DNS lookups. The presence of a PTR record indicates that the hostname is well established [10]. Autonomous System (AS) numbers for these records would further indicate the presence of ISPs (Internet Service Providers) that are known to host phishing websites. AS numbers and associated BGP prefixes are extracted for the corresponding A, CNAME, MX, and NS records.

C. Cluster Label Feature

Clustering is performed on the entire dataset using the K-means algorithm [8]. The highest accuracy is yielded by the classification system when the number of clusters is 6, and is used as the chosen size throughout the rest of the paper. K points are chosen at random as cluster centers, and new objects are assigned to their closest cluster center depending on the outcome of the Euclidean Distance function. The centroid or mean of each cluster is calculated using all the objects in the cluster including the newly added ones. Clustering continues until the centroids do not change significantly in consecutive rounds [12]. Clustering extracts a "structure" for the whole dataset [22] and provides a new predictor variable; i.e., the cluster ID assigned to each example.

TABLE I. CLASSIFIER ACCURACIES FOR VARIOUS NUMBER OF CLUSTERS

Number of Clusters	Run #1	Run #2	Run #3	Accuracy (Avg. Run #1, #2, & #3)
6	95.34375%	95.28125%	94.96875%	94.86%
10	94.75%	93.78125%	93.84375%	94.12%
9	94.46875%	93.1875%	94.125%	93.92%
7	93.625%	93.8125%	93.25%	93.56%
8	92.0625%	94.34375%	94.125%	93.51%
3	94.5%	93.8125%	90.8125%	93.04%
5	93.78125%	91.34375%	94.0625%	92.72%
4	93.09375%	93.9375%	90.6875%	92.57%
2	90.3125%	91.125%	93.46825%	91.63%

IV. FEATURE VECTOR ENCODING

Classification requires proper encoding for different types of values associated with predictor variables to obtain a feature vector that accurately describes the URL. For instance, feature hashing is used in order to encode raw feature data into feature vectors [8]. After carefully analyzing the dataset, the size of the feature vector was chosen to be in the 59,000 dimension space which reduced feature collisions as shown in Table II.

V. DATA

For this paper, benign URLs are collected from the DMOZ open directory project [13]. Phishing URLs for experimentation are collected from PhishTank [14]. Obtaining phishing URL data for classification is a challenging task, as most phishing URL campaigns are short lived.

TABLE II. Classifier Accuracies For Various Feature Vector Sizes

Feature Vector size	Run #1	Run #2	Run #3	Accuracy (Avg. Run #1, #2, & #3)
59,000	95.0625%	95.3125%	95.25%	95.20%
60,000	94.6875%	94.46875%	95.625%	94.92%
50,000	94.40625%	94.84375%	93.375%	94.20%
70,000	95.53125%	93.59375%	93.21875%	94.11%
100,000	95.28125%	91.96875%	95.03125%	94.09%
10,000	93.78125%	93.21875%	94.375%	93.79%
58,500	93.53125%	95.625%	92.09375%	93.75%
59,500	94.4375%	95.59375%	93.15625%	93.72%
69,000	94.03125%	92.59375%	93.90625%	93.51%
25,000	93.25%	92.9375%	93.8125%	93.33%
61,000	91.65625%	93.0%	94.5%	93.05%
65,000	92.75%	93.6875%	91.46875%	92.63%
58,000	91.875%	94.6875%	88.28125%	91.61%
3,000	88.71825%	89.9375%	93.78125%	90.8125%
300	90.34375%	89.125%	89.21875%	89.5625%
100	89.125%	88.34375%	87.4375%	88.30%

VI. URL Categorization and URL Ranking

All categories from the Microsoft Reputation Service (MRS) [18] are extracted and placed into three separate bags – Severe, Moderate, and Benign. Multiple categories for the URL may be returned. If a particular URL has two categories, one of which is severe such as 'Phishing' and the other is benign such as 'Education', then preference is given to the bag with a higher threat as depicted below.

Extracted Categories: #1 Phishing, #2 Education
Category #1 → Severe, Category #2 → Benign
Hypothesis: Preference given to bag with higher threat
Result: Severe

An internal scale gives the categorization of each URL.

Internal Scale: Red/Yellow/Green
If Result = Severe, Then Scale → Red
Else If Result = Moderate, Then Scale→ Yellow
Else If Result = Benign, Then Scale→ Green

URL ranking is divided into five categories. The probability of phishing P(Ph) and benign P(Bn) URLs is obtained directly from the result of the classifier.

Rules at 'Internal Scale=Red':

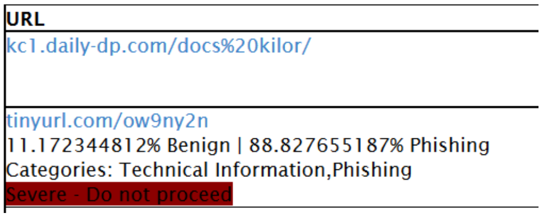
Rule 1: (P(Ph)>0.8) **&** (P(Bn)<0.2) → Severe
Rule 2: (0.6< P(Ph)<0.8) **&** (0.2<P(Bn)<0.4) → Dangerous
Rule 3: (0.4< P(Ph)<0.6) **&** (0.6<P(Bn)<0.8) → Pot.Threat

Rules at 'Internal Scale=Yellow':

Rule 1: (P(Ph)<0.2) **&** (P(Bn)>0.8) → Unsafe
Rule 2: (0.2<P(Ph)<0.4) **&** (0.6<P(Bn)<0.8) → Pot.Threat
Rule 3: (0.4< P(Ph)<0.6) **&** (0.4< P(Bn)<0.6) → Pot.Threat
Rule 4: (0.6<P(Ph)<0.8) **&** (0.2<P(Bn)<0.4) → Pot.Threat
Rule 5: (P(Ph)>0.8) **&** (P(Bn)<0.2) → Dangerous

Rules at 'Internal Scale=Green':

Rule 1: (P(Ph)<0.2) **&** (P(Bn)>0.8) → Safe
Rule 2: (0.2<P(Ph)<0.4) **&** (0.6<P(Bn)<0.8) → Unsafe
Rule 3: (0.4< P(Ph)<0.6) **&** (0.4< P(Bn)<0.6) → Pot.Threat



URL
kc1.daily-dp.com/docs%20kilor/

tinyurl.com/ow9ny2n
11.172344812% Benign | 88.827655187% Phishing
Categories: Technical Information,Phishing
Severe - Do not proceed

Fig. 1. Example of Severe – Do not proceed.

For the example in Fig. 1, URL categorization retrieved [18] two categories 'Technical Information', and 'Phishing'; yielding *Internal Scale = Red*. The URL receives the rank 'Severe' based on *Rule1* from the *Rules at Internal Scale=Red*. For the URL above, AS numbers of the website, mail server, and name server are different. IP prefixes of the website, mail server, and name server are different. These among other factors contributed to the URL being classified as phishing.

VII. Assessment

Table III shows a comparison between the current work described in this paper and the approaches used by other researchers. Ma et al. [7] used a combination of lexical and host-based features. However, their research [7] did not include the usage of bigrams, clustering, or URL Ranking. Research conducted by Blum et al. [11] consisted of lexical features and the usage of hostname bigrams. However, their research [11] discarded the usage of host-based features which are proven to be useful for detecting phishing URLs on the Internet [1], [4], [5], [7], and [10]. Using many types of features as with the proposed approach in this paper and not just a subset will eventually make the system more reliable as it is not dependent on any particular subset of features [19].

TABLE III. Comparison with other Methodologies

Technique	Lexical Features	Bigrams	Host	Accuracy	URL Ranking Feedback
[7]	•		•	97%	
[11]	•	•		97%	
Current work	•	•	•	98.46%	•

An attempt was made to implement the system proposed by Ma et al. [7]. Features that were not implemented include uplink connection speed for the URL, and city/country of the URL. Using the current dataset, the simulated implementation of the work done by Ma et al. [7] yielded classification accuracy of 85%. The classification accuracy of the current work is 98.46%. The increase in accuracy indicates the benefit of using novel features (i.e. bigrams, cluster labels).

CONCLUSION

Microsoft Reputation Services (MRS) [18] returns one or more categories for each individual URL submitted with varying threat levels. The end user is left to make the crucial judgment of determining to which category a particular URL belongs based on a very rudimentary understanding of risk [20]. The current work proposes a framework that identifies the returned categories from MRS and derives an internal threat scale. Another Microsoft product, Microsoft Forefront Threat Management Gateway (Forefront TMG) uses the results from MRS in a similar manner [21], but with a hardcoded category precedence list. The Severe, Moderate, and Benign bags in the current work are analogous to the precedence list. The internal threat scale filters the classification score to derive a URL rank using a novel URL ranking algorithm that helps derive more meaningful feedback for the end user particularly with the addition of the color band. URLs are ranked by using URL clustering, classification, and categorization demonstrating that cluster labels increase the classifier accuracy from 97.08% [19] to 98.46%.

ACKNOWLEDGMENTS

Our thanks go to Everett McArthur and Brian Nordstrom.

REFERENCES

[1] Thomas, K., Grier, C., Ma, J., Paxson, V., and Song, D, “Design and Evaluation of a Real-Time URL Spam Filtering Service”. in Proceedings of the 2011 IEEE Symposium on Security and Privacy, California, USA, 2011, IEEE Computer Society, pp. 447-462. DOI= http://dl.acm.org/citation.cfm?id=2006781.

[2] Mills, E. Google finds 9,500 new malicious Web sites a day. *CNET News.* http://news.cnet.com/8301-1009_3-57455614-83/google-finds-9500-new-malicious-web-sites-a-day/. 2012.

[3] Siciliano, R. What is Typosquatting?. *McAfee Blog Central.* http://blogs.mcafee.com/consumer/what-is-typosquatting. 2013.

[4] Ma, J., Saul, L., Savage, S., and Voelker, G., “Identifying Suspicious URLs: An Application of Large-Scale Online Learning”. in Proceedings of the 26th International Conference on Machine Learning, Montreal, Canada, 2009, ACM New York, NY, USA, pp. 681-688. DOI= http://dl.acm.org/citation.cfm?id=1553462.

[5] 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.

[6] Stringhini, G., Kruegel, C., and Vigna, G., “Detecting Spammers on Social Networks”. in ACSAC’10 Proceedings of the 26th Annual Computer Security Applications Conference, Florida, USA, 2010, ACM New York, NY, USA, pp. 1-9. DOI= http://dl.acm.org/citation.cfm?id=1920263.

[7] 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.

[8] Owen, S., Anil, R., Dunning, T., and Friedman, E. *Mahout In Action*. Manning Publications, NY, 2010.

[9] Orr, G., and Muller, K. *Neural Networks: Tricks of the Trade*. Springer, NY, 1998.

[10] 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.

[11] 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.

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

[13] Netscape. DMOZ Open Directory Project. http://www.dmoz.org.

[14] OpenDNS. PhishTank. http://www.phishtank.com.

[15] Team Cymru Community Services. http://www.team-cymru.org.

[16] Whittaker, C., Ryner, B., and Nazif, M., "Large-Scale Automatic Classification of Phishing Pages”. in NDSS’10 Proceedings of the NDSS Symposium 2010, San Diego, California, USA, 2010. DOI= http://www.internetsociety.org/doc/large-scale-automatic-classification-phishing-pages.

[17] Quinlan, J. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., CA, 1993.

[18] Microsoft Reputation Services. Feedback and Error Reporting. http://www.microsoft.com/security/portal/mrs/.

[19] Feroz, M., and Mengel, S., "Examination of Data, Rule Generation and Detection of Phishing URLs using Online Logistic Regression”. in 2014 IEEE International Conference on Big Data, Washington DC, USA, 2014.

[20] Srikwan, S., and Jakobsson, M., "Using Cartoons to Teach Internet Security”. in Cryptologia, Volume 32 Issue 2, April 2008, pp. 137-154. DOI= http://dl.acm.org/citation.cfm?id=1451180.

[21] Grote, M. Overview of the Microsoft Reputation Service (MRS), Microsoft Malware Protection Center (MMPC) and other techniques. ISAserver.org. http://www.isaserver.org/articles-tutorials/general/Overview-Microsoft-Reputation-Service-MRS-Microsoft-Malware-Protection-Center-MMPC-other-techniques.html.

[22] Kyriakopoulou, A., and Kalamboukis, T., "Text Classification using Clustering”. in 2006 Proceedings of the ECML-PKDD Discovery Challenge Workshop. http://www.ecmlpkdd2006.org/kyriakopoulou.pdf3