Phishing URL Detection: A Machine Learning and Web Mining-based Approach

Bhagyashree E. Sananse
Student, M.E
Thadomal Shahani Engineering College
Mumbai, India

Tanuja K. Sarode, PhD
Associate Professor
Thadomal Shahani Engineering College
Mumbai, India

ABSTRACT

There has been an abrupt development and use of online transactions over the past decade. The increased sophistication of cyber criminals has lead to proliferation of phishing attacks. The continuous expansion of World Wide Web has led to the rapid spread of phishing, malware and spamming.

This paper proposes a feature based approach to classify URLs into phishing or non-phishing category. The usage of a variety of URL features is done by studying the anatomy of URLs. For classification of URLs, two different algorithms have been used. Random Forest machine learning algorithm is used to build an efficient classifier which would decide whether a given URL is phishing or not. In addition, a novel scheme has been proposed to detect phishing URLs by mining the publicly available content on the URLs.

General Terms

URLs, Algorithm, Information Retrieval, Anti-phishing, Legitimate, Features

Keywords

Phishing URL, web mining, benign, machine learning

1. INTRODUCTION

The act of acquiring sensitive information by convincing the users to reveal their personal information such as usernames, passwords, credit card credentials, etc. by pretending as a trusty source in an electronic transmission is known as phishing. It is a criminal offense which targets both social engineering and technical tricks to steal personal identity or financial account information of user and is an automated form of identity theft. Phishing websites are affecting both individuals and financial organizations on the Internet, leading to a serious threat to electronic commerce. Every URL has this common syntax: <protocol>://<hostname><path>. Consider the following URL.

https://www.google.co.in/?webhp?@sourceid=chrome/_;=-instant&ion=1&espv=2&ie=UTF-8#q=baroquemusicusernamepassword

"https://www.google.co.in" indicates the base URL. The <protocol> part of the URL depicts which network protocol should be used to fetch the requested resource. HTTP or http (Hypertext Transfer Protocol), https (HTTP with Transport Layer Security) and FTP (File Transfer Protocol) are the most commonly used protocols. <hostname> is nothing but the identifier for the web server on Internet. In the above URL, www.google.com is the hostname, Google is the domain name and co.in is the TLD (Top Level Domain). <path> in the URL is similar to the path of any file on the computer. It contains different punctuation marks like slashes, dots, dashes, etc. The text after the first forward slash after the <hostname> indicates the path. The text after the first "?" indicates the query part of the URL. The text after the first "@" indicates the parameter part of the URL. The text after the first "#" indicates the fragment part of the URL.

A phishing URL is created with a malicious purpose to download malware, to perform phishing attacks or to manipulate search engine's results. The technical experience of criminals is increasing to build more survivable infrastructures that support phishing activities. Botnets are the main building blocks which are used to host phishing sites or send phishing emails. Internet is becoming a common place for information retrieval as information is easily accessible and available to all of the Internet users.

There has been a loss of user trust on the Internet due these disastrous phishing attacks posing a threat to the electronic commerce. Phishing has become a reason of both short term and long term economic damage and is a rapidly growing form of identity theft scam. Due to all these reasons, designing and implementing effective phishing detection techniques to withstand cyber crime and to ensure cyber security has become a major need.

In this paper, the following contributions are made: 1) An illustration, that by using information on the URL, it can be classified as phishing or non phishing. 2) Inspecting the significance of publicly available information on a URL to decide whether it is phishing or benign. 3) An illustration, that how the proposed methodology can be used in real-time applications for detecting phishing URLs. 4) Illustrating how the characteristics of phishing URLs differ over time.

2. LITERATURE REVIEW

2.1 Non-Machine Learning Techniques

B-APT is an anti-phishing toolbar to fight with phishing which uses Bayesian filter to filter spam emails and whitelist. The key advantage of Bayesian filter is its ability to detect never seen items before. B-APT is a good solution for zero day phishing sites [1]. SpoofGuard, a browser-based plugin monitors the Internet activity of users and calculates spoof index. If the index crosses the user selected level then it warns the user. It also uses domain name, URL, link, image checks and history of the user [1]. Some other anti-phishing tools are SiteAdvisor [2], Netcraft anti-phishing toolbar [3], and AVG Security toolbar [4].

2.2 Machine Learning Techniques

Whittaker et al. explained the scheme and execution characteristics of a scalable machine learning classifier which has been used to automatically maintain Google's phishing blacklist. Millions of pages a day are analyzed by their proprietary classifier. It also examines the URL and its webpage contents to decide whether a page is phishing or non-phishing. Web pages submitted by end users and URLs collected from Gmail's spam filters were classified by their system [5]. Garera et al. used logistic regression classifier for 17 hand-picked features for classification of phishing URLs. The features used included the existence of some red flag key words, some proprietary heuristics based on Google's PageRank and webpage quality instructions and the initially calculated features, based on pages which belonged to Google's proprietary infrastructure which they called as *Crawl Database* [6].

Zhang et al. proposed a content-based approach named as CANTINA to classify phishing websites, which is dependent on data retrieval known as TF-IDF algorithm and the Robust Hyperlinks algorithm. CANTINA uses 8 features out of which 4 are content-based, 3 are lexical, and 1 WHOIS-related [7]. Ma et al. explained a method to classify doubtful URLs by using a changeable number of lexical and host-related features of the URLs. A comparison is made to check the accuracy of online learning algorithms and batch algorithms by making use of those 8 features. They determined that Logistic Regression performs the best in their case [8].

2.3 Existing Approaches

Basnet et al. used a set of 138 URL based features to determine whether an URL is phishing or not. They grouped their features into 4 broad categories as lexical based, keyword based, search engine based and reputation based features [9]. James et al. proposed a method in which they used host based, page based and lexical features to classify URLs into phishing and non-phishing category [10]. Su et al. described a method that intends to learn the properties from multiple portions of URLs to assign the suspicion level of each portion. After the adjustment of suspicion threshold of each portion, the system would choose the most suspicious URL [11].

3. FEATURE ANALYSIS

The phishers confuse the e-citizens by many techniques. Confusing the host with an IP address- where the hostname of the URL is replaced with an IP address, confusing the host with another domain name- where the URL's host name includes a valid domain name, confusing the host name by concatenating a large string of words and domains after host name are some of the techniques used by phishers [6].

This paper deals with collection of various features by analyzing the benign and phishing URLs to detect whether a website is phishing or not. The keyword-based features used by Basnet et al. have a drawback. Those keywords used by them also occur in legitimate URLs also. Due to this feature a legitimate site may also get classified as phishing due to the presence of those keywords. This increases the number of false positives and false negatives. Other feature used by Basnet et al. i.e. the search engine based feature is machine and region dependent. In different countries same search engines may show different results for same query. As a result of usage of these two features, the difficulty level to decide a line of separation between phishing and legitimate URLs increases [9].

So in this paper, some robust features have been identified that can be used to differentiate between phishing and non-phishing URLs.

1. **Lexical features**
   These features point at the text based properties of URL, i.e. the publicly available information on URL rather than content of the web page that URL points to. The characteristics of features encountered in this type are length-based, character count-based, presence or absence of characters, strings or non-standard ports. There are 24 features in this category [9].
2. **WHOIS features**
   These properties explain "who" manages the sites, "where" they are hosted and "how" they are administered. They also give information about date of registration, update and expiry of websites [10][12]. There are total 48 features in this category.
3. **Page Rank**
   This feature shows how popular a web page is among the e-citizens. Google Search uses PageRank algorithm for ranking websites in the results of their search engine. PageRank is a technique of computing the significance of website pages and works by determining the number and quality of links to a page to decide an approximate estimate of how essential the website is. The underlying presumption is that if any page is more important, then websites will probably receive more links from other websites. Google is considered as the most reputed search engine at present. Google PageRank value is considered as one of the heuristic to detect whether a page is legitimate page or phishing page. Google's PageRank value is robust and updated frequently [13].
4. **Alexa Rank**
   It is ranking system set by alexa.com that basically audits the frequency of visits on numerous websites and makes it public. Alexa ranking is computed based on volume of traffic noted down from the users that have installed the Alexa toolbar for more than a period of 3 months. The parameters on which the traffic is based are reach and page views. The number of Alexa users who visit a particular site in one day is referred as reach. The number of times a particular URL is viewed by Alexa users is known as Page view [14].
5. **PhishTank-based feature**
   PhishTank produces statistical reports on phishing websites every month. The aim is to use the historical information on top IPs and domains which host phishing websites. If the host of the URL belongs to top IP or domain that hosts phishing websites and also many other phishing related heuristics are present, then the probability to classify the URL as phishing can be increased [9].

4. METHOD OVERVIEW

A feature-based approach has been proposed for classification of URLs into phishing or non phishing based on the details available on the URLs. This problem is considered as a binary classification problem where phishing URLs are labeled as positive class and benign URLs are labeled as negative class. Firstly phishing and legitimate sites are collected to build the dataset. Then a batch of code is run to collect a number of features on the URLs. Then two algorithms are applied as follows:

1. Random Forest algorithm, which is one of the most efficient machine learning algorithm to build prototypes from training data, which consists of pairs of features values and class labels. The prototypes are then fed with separate set of testing data and the data instance of the predicted class is compared with the actual class of data.
2. Content-based algorithm, (works on the publicly available data on the URLs) which focuses on the important features that distinguish phishing sites from legitimate ones.

Figure 1 depicts overview of phishing URL detection system.

4.1 Data Collection

Phishing URLs were collected from PhishTank which is a community based phish confirmation system on Internet [15]. Developers and researchers are allowed to download verified phishing URL lists which are available in various file formats with the help of an API key but only after signing up. Non-phishing URLs were collected from various credible sources and Google search engine. Then the data was categorized into training and testing categories for their respective purposes.

4.2 Feature Extraction

In this phase, 24 lexical features, 48 WHOIS features, PageRank, Alexa Rank and PhishTank-based features are extracted.

4.3 Classification

In this phase, the URLs are classified using both Random Forest algorithm and Content-based algorithm.

4.3.1 Random Forest Algorithm

Random Forest machine learning algorithm has been evaluated, since it has many advantages. It is one of the most precise machine learning algorithms available and produces a highly accurate result for many data sets. It also works efficiently on large databases. It can manage thousands of input variables without the need of variable deletion. It has a productive method for determining missing data and maintains accuracy in case of large proportion of missing data. It has methods for stabilizing error in class population of unbalanced data sets. Generated forests can be used for future use on other data. Random Forest classifier implemented using WEKA (Waikato Environment for Knowledge Analysis) library has been evaluated using their default parameter values [16]

4.3.2 Content-based Algorithm

In this algorithm, some of the features are pruned down as they increased the false negative rate (FNR). So here 24 lexical features, 3 WHOIS features, PageRank and Alexa Rank are used.

1. Lexical features

For lexical features which depend on length count or character count, a particular threshold value is set for each of these features. If the threshold value exceeds the set threshold value for any URL, then that URL is marked as phishing. For lexical features that work on presence or absence of character or string or non-standard ports, the usage of database derived after training the Random Forest classifier is done. Those statistics helped us in deciding whether the presence or absence of certain character or string in any particular URL makes it phishing or non phishing.

2. WHOIS features

The WHOIS data contains various information about the registrar details, registrant details, admin details, tech details etc. which are least important in deciding whether a URL is phishing or not. So those features are pruned down to three important features which include domain name which should be the same as that in the URL, registrar whois server which should belong to some standard servers like godaddy, networksolutions, markmonitor etc., and the registrar URL which should contain the same domain name as that of registrar whois server.

3. PageRank

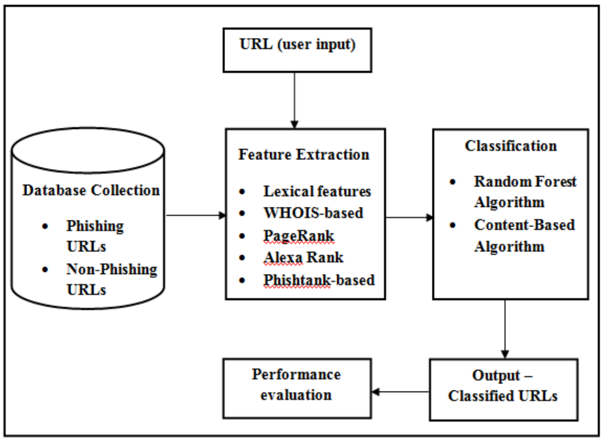


Figure 1: Block Diagram of Phishing URL Detection Framework

Google PageRank is one of the methods used by Google to estimate the relevance or importance of a page. Important pages are encountered to have higher PageRank and have higher probability to appear at the top of search results. If PageRank value for a given URL is less than 5 then the URL will be classified as phishing URL [17].

4. Alexa Rank

This feature evaluates how popular the website is by determining the number of visitors and the number of pages visited by them. Some phishing have short lifetime like zero day phishing websites. So they may not be acknowledged by the Alexa database. By analyzing the dataset, it is found that in worst-case legitimate websites ranked among the top 150,000. Hence, if the domain has no traffic or not being acknowledged by the Alexa database it is classified as “Phishy. So for Alexa rank the threshold is set to 150000. If the Alexa rank of URL exceeds this threshold value then it would be classified as phishing [18].

4.4 Performance Evaluation

As the phishing URL detection problem is binary classification problem, every URL falls into one of four possible categories: true positive (TP, correctly classified phishing URL), true negative (TN, correctly classified non-phishing URL), false positive (FP, non-phishing URL wrongly classified as phishing), and false negative (FN, phishing URL wrongly classified as non- phishing). Standard measures such as false positive rate (FPR), false negative rate (FNR), precision, recall, and F-measure were determined using the following equations [19]:

1. $$FPR = \frac{|FP|}{\#\ legitimate\ \ URLs} \qquad (1)$$

2. $$FNR = \frac{|FN|}{\#\ phishing\ URLs} \qquad (2)$$

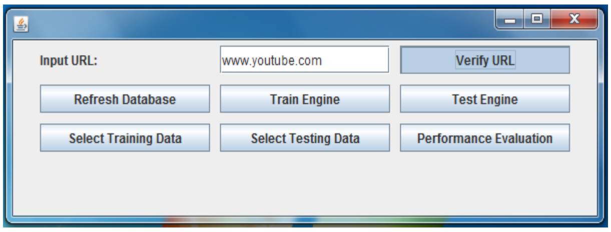
3. $$precision = \frac{|TP|}{|TP| + |FP|} \qquad (3)$$

4. $$recall = \frac{|TP|}{|TP| + |FN|} \qquad (4)$$

5. $$F = \frac{2.precision\ .recall}{precision\ +recall} \qquad (5)$$

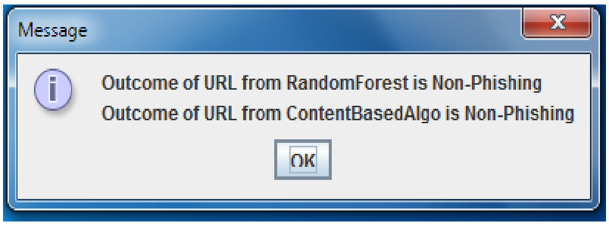
5. RESULTS

Random Forest classifier has been trained using a set of 500 URLs and tested the classifier using a set of 100 URLs. Figure 2 and Figure 3 shows the output when legitimate URL is fed as input to the system. In Figure 2, *www.youtube.com* is fed as input to the system and verified for its result. Figure 3 shows the output as “non-phishing” by both the algorithms, i.e. Random Forest algorithm and Content-based algorithm.



Input URL: www.youtube.com

Figure 2: Legitimate URL for Verification

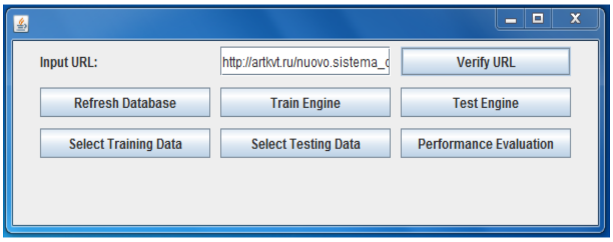


Message

Outcome of URL from RandomForest is Non-Phishing
Outcome of URL from ContentBasedAlgo is Non-Phishing

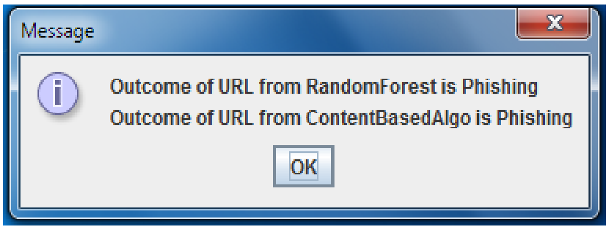
Figure 3: Output after Verification as Non-Phishing

Figure 4 and Figure 5 shows the output when phishing URL is fed as input to the system. In Figure 4, *http://artkvt.ru/nuovo.sistema_di_sicurezza.web.modulo-cliente80174.sh...* is fed as input to the system and verified for its result. Figure 5 shows the output as “phishing” by both the algorithms, i.e. Random Forest algorithm and Content-based algorithm.



Input URL: http://artkvt.ru/nuovo.sistema_d

Figure 4: Phishing URL for Verification



Message

Outcome of URL from RandomForest is Phishing
Outcome of URL from ContentBasedAlgo is Phishing

Figure 5: Output after Verification as Phishing

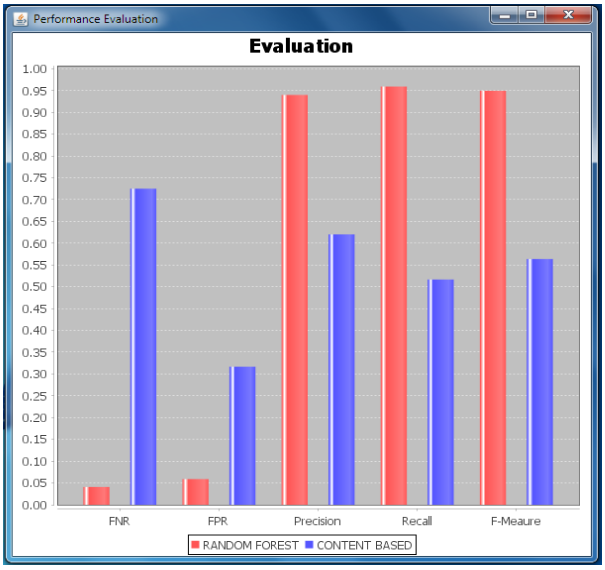


Figure 6: Graph of Performance Evaluation

Figure 6 shows the graph for performance evaluation for different parameters like FNR, FPR, Precision, Recall and F-Measure. On observation it can seen that Random Forest algorithm performs well than Content-based algorithm.

6. CONCLUSION

In this paper, a system has been proposed that uses lexical features, WHOIS features, PageRank and Alexa rank and PhishTank-based features for Random Forest algorithm to classify phishing URLs. It has been demonstrated that by applying web mining heuristics on Random Forest algorithm, a precision of more than 90% has been achieved and FNR and FPR rates less than 1%. But in case of Content-based algorithm the precision achieved was less than 65%.

As future work, there is a need to work on selection of more efficient features for Content-based algorithm to increase the precision and decrease the FNR and FPR. Also webpage content based features can be integrated to make the system more robust.

7. REFERENCES

[1] Namrata Singh, Nihar Ranjan Roy, “A Survey of Phishing Website Detection Techniques”, IRAJ International Conference-Proceedings of ICRIEST-AICEEMCS, 2013, Pune India.

[2] McAfee SiteAdvisor Software- Website Safety Ratings and Secure Search, http://www.siteadvisor.com, accessed on June 25, 2015.

[3] Netcarft Anti-Phishing Toolbar, http://toolbar.netcraft.com, accessed on June 25, 2015.

[4] AVG Security Toolbar, http://www.avg.com/product-avg-toolbar-tlbrc#tba2, accessed on June 25, 2015.

[5] C. Whittaker, B. Ryner, M. Nazif, "Large-Scale Automatic Classification Of Phishing Pages", In: Proc 17th Annual Network and Distributed System Security Symposium, NDSS'10, San Diego, CA, USA, 2010.

[6] S. Garera, N. Provos, M. Chew, A.D. Rubin, "A Framework For Detection And Measurement Of Phishing Attacks". In: Proc. 5th ACM Workshop on Recurring Malcode, WORM'07, ACM, New York, NY, USA, 2007.

[7] Y. Zhang, J. Hong, L. Cranor, "CANTINA: A Content-Based Approach To Detecting Phishing Web Sites", In: Proc. 15th Int. Conf. World Wide Web, WWW"07, Banff, Alberta, Canada, 2007.

[8] J. Ma, L.K. Saul, S. Savage, G.M. Voelker, "Beyond Blacklists: Learning To Detect Malicious Web Sites From Suspicious URLs", In: Proc. 15th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, Paris, France, 2009.

[9] Ram B. Basnet, Andrew H. Sung, Quingzhong Liu, "Learning To Detect Phishing URLs", IJRET: International Journal of Research in Engineering and Technology", 2013.

[10] Joby James, Sandhya L, Ciza Thomas, "Detection Of Phishing URLs Using Machine Learning Techniques", International Conference on Control Communication and Computing (ICCC), 2013.

[11] Ke-Wei Su, Kuo-Ping Wu, Hahn-Ming Lee, Te-En Wei, "Suspicious URL Filtering Based On Logistic Regression with Multi-view Analysis", Eight Asia Joint Conference on Information Security, 2013.

[12] http://www.whois.com/, accessed on December 17th 2014.

[13] Anjali Sardana, A.Naga Venkata Sunil, "A PageRank Based Detection Technique for Phishing Web Sites", IEEE Symposium on Computer and Informatics (ICSI), 2012.

[14] http://developers.evrsoft.com/find-traffic-rank.shtml, accessed on December 17th 2014.

[15] https://www.phishtank.com/, accessed on November 30th, 2014.

[16] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, "The WEKA Data Mining Software: An Update", SIGKDD Explorations, 2009.

[17] Sana Ansari, Jayant Gadge, "Architecture For Checking Trustworthiness Of Websites", International Journal of Computer Application, 2012.

[18] Rami m. Mohammad, Fadi Tabhtah, Lee McCluskey, " Intelligent Rule based Phishing Websites Classification, Information Security, IET, 2014.

[19] Ram B. Basnet, Andrew H.Sung, "Mining Web to Detect Phishing URLs", 11th International Conference on Machine Learning and Applications, 2012.