International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 9 (2017) pp. 1893-1898


# Two phase Classification Model to Detect Malicious URLs

**N.Jayakanthan[1], A.V.Ramani[2] and M.Ravichandran[3]**

*[1,] Department of Computer Applications, Kumaraguru College of Technology, Coimbatore, Tamilnadu, India.*

*[2,3] Department of Computer Science, Sri Ramakrishna Mission Vidyalaya College of Arts and Science, Coimbatore- 641020 Tamilnadu, India.*

***[1]ORCID: 0000-0002-4202-9122***

**Abstract.**
Malicious websites poses a challenging threat in cyber space which snips the user's critical information. Hence detecting the URLs of the malicious website is an essential task in the present scenario. Though various methods are available to identify malicious and genuine website, the existing phishing techniques have challenging issues. The current approaches only detect the familiar phishing websites using blacklisted profile. The main drawback is that non-blacklisted phishing sites are non-recognizable. The attackers are using multiple URLs pointing to the same website to bypass the detection techniques, thereby it is essential to detect such websites. The proposed work detects malicious URLs in two phases. The Enhanced Probing Classification algorithm to detect Malicious URL (EPCMU) is developed and Navie Bayes classifier is used to detect malicious URLs set. The proposed algorithm is effective in detection of malicious websites.

**Keywords:** Phishing, Malicious URL, Feature Analysis, Navie Bayes.

## INTRODUCTION

The web applications are being deployed across the globe which brings ease among the people, but on the other hand, various attacks are being created by the unauthenticated and unauthorized organizations. Web security being a challenging issue, people and organizations need to protect their data. Kaspersky internet security reports phishing attack increased by 87% in the year of 2015 -2016 when compared to 2014-2015.According to RSA online fraud report, the number of phishing attack launched in 2015 is higher than 2014.

The statistics of the Anti-Phishing work group's phishing activity trend report 2015 illustrates that the payment services are the primary target, reported 39.08% phishing attack followed by financial services 20.20% and retail services 16.53%. The banking, e-commerce and social networking sectors are facing most attacks. The phishing growth rate of the phishing trends is also high. At present various phishing tools are available to create malicious websites and snip the passwords. For an example "Wifiphisher" is an open source tool to access the WPA protected "Wi-Fi" network to steal username and passwords. So detecting malicious URLs which leads to a phishing website is an essential task in the current scenario.

The malicious URL directing to the phishing website is a general threat to an innocent user. Logistic regression over 18 selected features was used to classify malicious URLs by Garera et al [1]. Zheng et al. [2] proposed a technique to categorize phishing URLs by thresholding a weighted sum of 8 features, including 3 lexical features, 4 content-related features, and 1 WHOIS-related feature. Abubakr Sirageldin et al[3] proposed a frame work to identify the malicious web pages using artificial neural network learning techniques based on URL lexical and page content features.

Fette et al. [4] classified the phishing emails using Random forest machine learning algorithm. It analyzes the various URL features to detect the suspicious emails. This method is additionally enhanced through text classification to examine the email content by Bergholz et al. [5]. Rakeshverma et al [6] developed a natural language based scheme to detect the phishing emails. Such problems can be primarily identified and solved if knowledge about the attributes of the target URL is known earlier, in particular, whether it is safe or not. Several methods have been proposed to sense phishing websites, some of which are customized by the industry. These methods are based on blacklist which is a list of malicious URLs released by agencies like Google Safe Browsing [7], Microsoft IE9 anti-phishing protection [8], and Site Advisor [9]. The Gartner report [10] in October 2016 reveals that one in every 4,500 emails today is the phishing attack.

The Kaspersky Security Bulletin reports 1,966,324 malware infections that aimed to steal through online access to bank account. Yue Zhang et al [11] proposed the design, implementation, and evaluation of CANTINA, a novel, technique based on content formalization to detect the phishing web sites, based on the TF-IDF information retrieval algorithm.

Maher Aburrous [12] proposed an associative classification algorithm to detect the phishing website. This method is flexible and effective in detecting malicious websites. Neil Chou et. al. [13] justified some traits of common attacks and proposed a framework for client-side protection, a browser plug-in that examines web pages and alerts the user when data requirement may be part of a spoof attack. Spoof guard uses heuristic approach to assess whether a page has phishing characteristics. But it frequently generates false positives. T.Sharif [14] implemented a phishing filter in IE7 is a toolbar. It analyzes various features to detect the phishing site.

Google Safe Browsing, similar to IE 7 makes use of blacklists of phishing URLs to identify malicious sites. Various organizations offered blacklisting service based on various techniques including manual reporting. But many new phishing web sites are not detected and added in the list. Hence this approach is inefficient for classification. In addition to that user skips the security mechanism due to the time consumption of these techniques. Most of the client side protection schemes lacks in time efficiency. The attackers use multiple URLs pointing to the same website to bypass the detection techniques. Hence we propose two stage classifier which analyses the URL set of a website to detect whether the website is genuine or malicious.

The proposed work has the following contributions. It detects the malicious website with multiple URLs using two stage classification algorithms. The model performs an extensive probing to detect malicious URL using the proposed EPCMU algorithm. The significant features are used for classification. It improves the classification accuracy of the algorithm. The proposed algorithm is a light weight approach to accurately detect the malicious URLs.

The paper is organized as follows. Architecture of the proposed system is discussed in subsequent section. Methodology is explained in section 3. Section 4 presents the algorithms of the proposed mechanism. Section 5 discusses analysis of the URL. Finally Section 6 concludes the paper.

## ARCHITECTURE OF THE PROPOSED SYSTEM

The over view of the proposed system in given in the figure 1. The components are Browser, Feature Extractor, EPCMU – Detector, Navie Bayes Classifier.

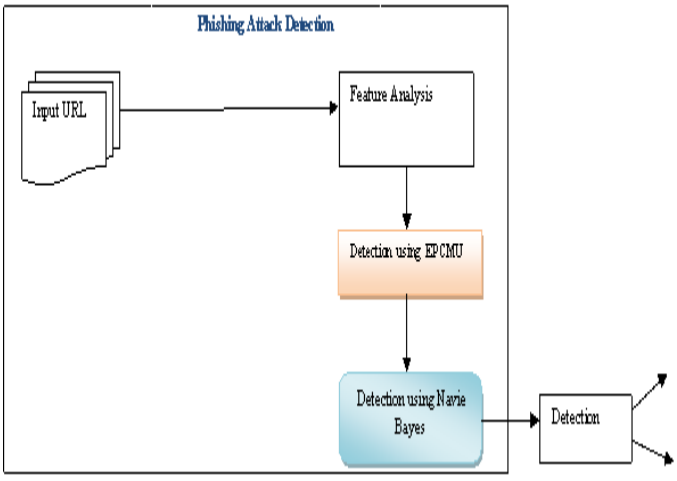


**Figure1.** Architecture Diagram

**Browser :**
The URL is entered as input in the browser of the system. The collected URL is send to feature analysis.

**EPCMU Detector :**
The detector model analyzes the features using the proposed EPCMU algorithm. If any malicious feature is present it marks the website as malicious and collects the malicious feature and reports it to the user.

**Navie Bayes Classifier :**
The classifier uses Naive bayes algorithm to analyze the features of the URLs. Based on the analysis, it detect where the URL set is genuine or malicious.

## METHODOLOGY

This paper proposes a novel solution to detect the malicious websites with multiple URLs. The proposed system contains two phases, detection and classification. In the first phase, the proposed EPCMU algorithm performs an extensive analysis of the URL and evaluates the various features. If any malicious feature exists, it detects the URL as malicious. Initially it checks the URL with the blacklisted profile of the system, which contains the list of malicious URLs already detected by the system. If match is found, it reports the URL as malicious and stops the process, else it checks the URL for the malicious special characters, number of dots, number of slashes and occurrence of the @ symbol. The set of URLs detected by the EMPCU algorithm is given as input to Navie Bayes algorithm in classification phase. It analyses the URL set to detect the website is genuine or malicious.

**Feature set :**
Features are significant to identify the malicious website. The list of features used by the proposed system is given below:

**Table 1 :** List of features

| Criteria | S.No | Features |
|---|---|---|
| URL & Domain Identity | 1 | Membership in Black listed profile |
| | 2 | Special characters |
| | 3 | Number of slashes in URL |
| | 4 | The number of dots in the URL. |
| | 5 | '@' symbol in URL |

***Membership in Black listed profile :***
The proposed system maintains a blacklisted profile. The detected URLs are added in the profile. It helps to avoid the repeated detection of the same URL.

***Special characters:***
The special characters are used in the URL to perform malicious activity. The occurrence of any one of the following special characters in URL is considered as malicious.
< > # % { } | \ ^ ~ [ ] ?
For example http://example.com/wp-admin/load-scripts.php?c=1&load[]=swfobject,jquery,utils&ver=3.5 is a malicious URL.

***Number of slashes in URL :***
In general, the URL should not include slashes more than five, then that URL is a phishing URL. For example consider, http://tohandtohome.com/module/fix/server/ajax/usaa.html.

***Number of dot symbols in URL:***
Phishing web uses false domain name to resemble the legitimate look of the URL with the help of additional dots placed in URL. The number of dots greater than '5', then the URL is considered as malicious. For example consider http://update-your-account.now.pp2.clickcom.com/cm is a malicious website.

***'@' symbol in URL :***
The phishing sites regularly have "@" symbol as an image in the URL to divert the user to a site similar in relation to what user expect. For example, for the URL http://www.western.com@www.maliciosweb.com, the real site to which this addresses points is not www.western.com, but http://www.maliciousweb.com which is a malicious one**.**

**ALGORITHM**
The two stage algorithm is used in this approach. In the first phase Enhanced Probing classification of Malicious URL, EPCMU algorithm is used to detect the given URL as malicious or not. In the second phase, the Navie Bayes algorithm is used to detect the nature of URL Set. Two stages of classifier improve the accuracy of the classification.

**Phase-I Detection Algorithm**

**Table 2.** Proposed algorithm

**Algorithm Enhanced Probing classification of Malicious URL – EPCMU(URL)**

//Input : The URL of the website

//Output : Classification of the URL as genuine or Malicious.

$M_T \leftarrow$ "Ø" // $M_T$ are set of Malicious features

Initialize decision variable S=Genuine //

1.For each URL, do the following

2. Check the URL for the Member ship in the blacklisted profile(BP).
If it matches with the profile
Set S ="Malicious" and stop
Set $M_T = M_T$ U BP : Go to step 8

3. Compare IP address with the blacklist
If match found then Set S="Malicious"

4.Compare the characters with the set of Malicious characters $M_C$
If Match found
Set S="Malicious"
Set $M_T = M_T$ U $M_C$

5.Count the number of slashes in the URL and store it in variable $S_c$
If ($S_c$ >=5) then
Set S="Malicious"
Set $M_T = M_T$ U $S_c$

6..Analyze the number of dots in the URL store it in variable $D_c$
If ($D_c$ >5) then
Set S="Malicious"
Set $M_T = M_T$ U $D_c$

7. check for @ symbol in URL
If @ symbol found then
Set S="Malicious"
Set $M_T = M_T$ U @

8.If S=="Malicious"
Display "URL is Malicious"
Display the set of malicious features in $M_T$
Block the URL and Alert the user
Else

Display URL is Genuine.
Allow the URL to proceed.
End for
**End EPCMU**

**Phase II The Naive Bays algorithm is used for classification.**
In the classification phase, the Navie Bays classification algorithm is used to detect the Malicious URL set. This algorithm is based on Bayes theorem with independent assumption between the features. The classifier has high classification capabilities. It detects whether the URL set is malicious or genuine.
Bayes formula is given below

$$P(C|X) = (P(X|C)*P(C)) / P(X) \quad (1)$$

P(C|X) is the probability of the attribute X belongs to the class C.
(P(X|C) is the probability of Class C having attribute X P(C) is Probability of the occurrence of the class C. P(X)= Probability of the occurrence of the attributes. The class C have two categories Malicious and Genuine in which an URL can be classified. By analyzing the feature vector, the classifier decides the category of the URL based on higher posterior probability using the following formulas.

Probability of URL to be malicious is given below
$$P(C1|X) = (P(X|C1)*P(C1)) / P(X) \quad (2)$$

Probability of URL to be genuine is given below
$$P(C2|X) = (P(X|C2)*P(C2)) / P(X) \quad (3)$$

P(C1|X) > P(C2|X) then the URL is malicious else (4)
The URL is genuine.

**ANALYSIS OF THE URLS**
The URLs of the same website used for analysis is listed below:

1.http://www.racil1.org.ar@74.125.131.105

2.http://www.racil1.org.ar/wp-admin/espace-

caf.fr/0FZEFZEF0ZEFZEF0ZEF0ZEF0EZFEZFZE508F5ZE8F04EZF048ZEF48EZ0F48ZEF/3dsecureclient.service.caf.fr/id/eec5bc2a4a47e2941bbac5a62f3e979e/step1.htm

3.http://www.racil1.org

4.http://www.racil1.org.ar/ss-admin/espace-caf.fr.2dsecureclient.service.caf.fr

5.http://www.racil1.org.ar/Client/espace-caf.fr/0FZEFZEF0ZEFZEF0ZEF0ZEF0EZFEZFZE508F5ZE8F04EZF048ZEF48EZ0F48ZEF/3dsecureclient.service.caf.fr/id/2ee96247eb996860335d0d28bcf6cc09/step3.htm

6. http:// www.racil1.org /wp-admin/load-scripts.php?c=1&load[]=swfobject,jquery,utils&ver=3.5

7.http://www.racil1.org.ar/clientele/space-far.cs.3dsecureclient.asp

8.http://www.racil1.org/phqqwm.phqqwm.phqqwm.hirata.com.mx/img/mps/logo/506.png?alceo.balboni@mail.it

9.http://www.racil1.org/home.html

The proposed EPCMU algorithm analyzes each URLs and detect it as genuine and malicious. The features and detection of the each URL is given in the following table.

**Table 3.** Detection result of EPCMU algorithm

| SL.NO | Member ship in block listed profile (MP) | Malicious IP address (MIP) | Special characters (SP) | Number of slashes(NS) | Number of dots (ND) | @ Symbol(@) | Class |
|---|---|---|---|---|---|---|---|
| 1 | No | No | No | < 5 | < 5 | Yes | Malicious |
| 2 | Yes | No | No | <5 | ≥5 | No | Malicious |
| 3 | No | No | No | <5 | <5 | No | Genuine |
| 4 | No | Yes | No | ≥5 | <5 | No | Malicious |
| 5 | Yes | No | No | <5 | ≥5 | Yes | Malicious |
| 6 | No | No | Yes | <5 | <5 | No | Malicious |
| 7 | No | No | Yes | <5 | ≥5 | No | Malicious |
| 8 | Yes | Yes | Yes | ≥5 | ≥5 | Yes | Malicious |
| 9 | No | No | No | No | <5 | No | Genuine |

The EPCMU algorithm predict 7 URLs as malicious and 2 URLs as genuine. This URL set is given to Navie Bayes algorithm.

In phase-II Navie Bayes algorithm analyzes the 6 features to classify the nature of the website. The values of the features are listed below:

Member ship in blacklisted profile ϵ {Yes, No}
Malicious IP address ϵ {Yes, No}
Special characters ϵ {Yes, No}
Number of slashes ϵ { <5, ≥ 5}
Number of dots ϵ {<5, ≥ 5}
@ symbol in URL ϵ {Yes, No}

Navie Bayes algorithm analyze the values of each independent features.

**Member ship in block listed profile**

| P( MP=Y/N \| Class Play=Y/N ) | Frequency | | Probability in Class | |
|---|---|---|---|---|
| Membership in block listed profile (MP) | Class=Malicious | Class=Genuine | Class=Malicious | Class=Genuine |
| Yes | 3 | 0 | 3/9 | 0/9 |
| No | 4 | 2 | 4/9 | 2/9 |
| | Total=7 | Total=2 | | |

**Malicious IP address (MIP)**

| P( NS=/N \| Class Play=Y/N ) | Frequency | | Probability in Class | |
|---|---|---|---|---|
| Membership in block listed profile (MP) | Class=Malicious | Class=Genuine | Class=Malicious | Class=Genuine |
| Yes | 2 | 0 | 2/9 | 0/9 |
| No | 5 | 2 | 5/9 | 2/9 |
| | Total=7 | Total=2 | | |

**Special characters**

| P( SP=Y/N \| Class Play=Y/N ) | Frequency | | Probability in Class | |
|---|---|---|---|---|
| Special characters (SP) | Class=Malicious | Class=Genuine | Class=Malicious | Class=Genuine |
| Yes | 3 | 0 | 3/9 | 0/9 |
| No | 4 | 2 | 4/9 | 2/9 |
| | Total=7 | Total=2 | | |

**Number of slashes**

| P( NS=≥5/ <5 \| Class Play=Y/N ) | Frequency | | Probability in Class | |
|---|---|---|---|---|
| Number of slashes NS | Class=Malicious | Class=Genuine | Class=Malicious | Class=Genuine |
| ≥5 | 2 | 0 | 2/9 | 0/9 |
| <5 | 5 | 2 | 5/9 | 2/9 |
| | Total= 7 | Total=2 | | |

**Number of dots**

| P( ND=≥5/ <5 \| Class Play=Y/N ) | Frequency | | Probability in Class | |
|---|---|---|---|---|
| Number of dots ND | Class=Malicious | Class=Genuine | Class=Malicious | Class=Genuine |
| ≥5 | 4 | 0 | 3/9 | 0/9 |
| <5 | 3 | 2 | 3/9 | 2/9 |
| | Total= 7 | Total=2 | | |

**@ Symbol**

| P( Q=Y/N \| Class Play=Y/N ) | Frequency | | Probability in Class | |
|---|---|---|---|---|
| Number of dots ND | Class=Malicious | Class=Genuine | Class=Malicious | Class=Genuine |
| Yes | 3 | 0 | 3/9 | 0/9 |
| No | 4 | 2 | 4/9 | 2/9 |
| | Total= 7 | Total=2 | | |

**Classification**

$P(Class_{Malicious=Yes \mid X}) = [\ P(MP='Y' \mid Class_{Malicious=Yes}) * P(MIP='Y' \mid Class_{Malicious=Yes}) * @(SP='Y' \mid Class_{Malicious=Yes}) * P(NS\geq 5 \mid Class_{Malicious=Yes}) * P(ND\geq 5 \mid Class_{Malicious=Yes}) * P(@='Y' \mid Class_{Malicious=Yes})$

$= 3/7 * 2/7 * 3/7 * 2/7 * 4/7 * 3/7 = 0.0037$

$P(Class_{Geniune=Yes \mid X}) = [\ P(MP='Y' \mid Class_{Geniune=Yes}) * P(MIP='Y' \mid Class_{Geniune=Yes}) * @(SP='Y' \mid Class_{Geniune=Yes}) * P(NS\geq 5 \mid Class_{Geniune=Yes}) * P(ND\geq 5 \mid Class_{Geniune=Yes}) * P(@='Y' \mid Class_{Geniune=Yes})$

$= 0/2 * 0/2 * 0/2 * 0/2 * 0/2 * 0/2 = 0$

$0.0037 > 0$

*$P(Class_{Malicious=Yes \mid X}) > P(Class_{Geniune=Yes \mid X})$ so the website is malicious and is blocked*

## CONCLUSION

In this paper, we propose a new methodology EPCMU integrated with Navie Bayes to detect the malicious website. It has two stages in which EPCMU algorithm is used initially to detect the malicious URL and then Navie Bayes based classification approach to detect the URL set. The two stage classifier improves the accuracy of the system. This approach also detects the malicious website with multiple URLs. The goal is to develop real-time malicious webpage detection system with an enhanced classification technique.

## REFERENCES

[1] S. Garera, N. Provos, M. Chew, and A. D. Rubin. A Framework for Detection and Measurement of Phishing Attacks. In Proceedings of the ACM Workshop on Rapid Malcode (WORM), Alexandria, VA, Nov. 2007.

[2] Y. Zhang, J. Hong, and L. Cranor. CANTINA: A Content-Based Approach to Detecting Phishing Web Sites. In Proceedings of the International World Wide Web Conference (WWW), Banff, Alberta, Canada, May 2007.

[3] Abubakr Sirageldin, Baharum B. Baharudin, Low Tang Jung. Malicious Web Page Detection: A Machine Learning Approach. International Journal of Advances in Computer Science and its Applications. Volume 279, pp 217-224 2014.

[4] I. Fette, N. Sadeh, and A. Tomasic. Learning to Detect Phishing Emails. In Proceedings of the International World Wide Web Conference (WWW), Banff, Alberta, Canada, May 2007.

[5] A. Bergholz, J. H. Chang, G. Paaß, F. Reichartz, and S. Strobel. Improved Phishing Detection using Model-Based Features. In Proceedings of the Conference on Email and Anti-Spam (CEAS), Mountain View, CA, Aug. 2008.

[6] Rakesh Verma, Narasimha Shashidhar, Nabil Hossain. Detecting Phishing Emails the Natural Language Way. 17th European Symposium on Research in Computer Security.September 10-12, 2012.

[7] Google Safe Browsing, https://developers.google.com /safe-browsing/, January 2013.

[8] Microsoft phishing filter, http://www.microsoft.com/, January 2013

[9] McAfee Site Advisor, http://www.siteadvisor.com/ , January 2013

[10] I A. Litan, (2009),"The War on Phishing Is 16far From Over", [Online] Available: http://www.gartner.com/ it/page. jsp?id=936913

[11] Yue Zhang, Jason Hong, Lorrie Cranor,"CANTINA: A Content-Based Approach to Detecting Phishing Web Sites", WWW '07 proceedings of the 16th international conference on world wide Web ACM New York, 2007.

[12] Maher Aburrous, M. A. Hossain, Keshav Dahal,"Associative Classification Techniques for predicting e-Banking Phishing Websites", Multimedia Computing and Information Technology, 2010.

[13] Neil Chou Robert Ledesma Yuka Teraguchi Dan Boneh John C. Mitchell, Client-side defense against web-based identity theft, In proc. NDSS 2004, 2004.

[14] T.Sharif (2006),"Phishing Filter in IE7", Second symposium on Usable privacy and security, [Online] Available: http:// www.blogs,msdn.com/ie/archive/ 2005/09/09/463204, aspx

[15] He, G., Zou, F., Tan, D., and Wang, M. (2011). Phishing Detection System Based on SVM Active Learning Algorithm. Computer Engineering. 37(19), 126-128.

[16] Jayadeva, R. Khemchandani, and S. Chandra, "Twin support vector machines for pattern classification,"IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 5, pp. 905–910, 2007.

[17] M. A. Kumar and M. Gopal, "Least squares twin support vector machines for pattern classification,"Expert Systems with Applications, vol. 36, no. 4, pp. 7535–7543, 2009.

[18] D. Tomar, S. Singhal, and S. Agarwal, "Weighted least square twin support vector machine for imbalanced dataset," International Journal of Database Theory and Application, vol. 7, no. 2, pp. 25–36, 2014.

[19] Min-Yen Kan , Hoang Oanh Nguyen Thi, Fast webpage classification using URL features, Proceedings of the 14th ACM international conference on Information and knowledge management, October 31-November 05, 2005, Bremen, Germany .

[20] Guan, D. J., Chen, C.-M., and Lin, J.-B. 2009. Anomaly based malicious url detection in instant messaging. In

[21] Proceedings of the Joint Workshop on Information Security (JWIS).

[22] C. Chong, D. Liu, and W. Lee. Malicious url detection. http://cs229.stanford.edu/proj2012/ChongLiu-MaliciousURLDetection.pdf.

[23] Byung-Ik Kim, Chae-Tae Im, Hyun-Chul Jung."Suspicious Malicious Web Site Detection with Strength Analysis of a JavaScript Obfuscation.Internationa Journal of Advanced Science and Technology. Vol 26, Januar 2011.

[24] P. Baldi, S. Brunak, Y. Chauvin, C. A. Andersen, and H. Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics, 16(5):412–424, February 2000.