IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

# AN STUDY OF SIMILARITY MEASUREMENT BETWEEN PHISHING AND LEGITIMATE WEBSITES USING BAYESIAN CLASSIFICATION AND ITS PERFORMANCE EVALUATION

**Dr. Rajendra Gupta**
Department of Computer Science and Application
AISECT University, Raisen, India

**DOI**: 10.5281/zenodo.266735

## ABSTRACT

Safe web browsing and feeding confidential information into websites require the use of protected and secured websites. For the web security, a number of anti-phishing tools have been proposed which provide web user with a dynamic system of warning and protection against potential phishing attacks. Earlier study shows that there is no anti-phishing tool gives satisfactory result in identifying the phishing web pages. For the solution of this problem, in this paper a Bayesian classification approach is proposed to identify the phishing websites. Bayesian filter require two datasets in their approach; one is legitimate website details and second thing is phishing website parameters. A large set of legitimate transactional websites are needed in the study because the set of websites mostly resembles just like phishing websites and the filter must have numerous examples of legitimate transactional websites to achieve a low false positive rate. With the use of Bayesian Classification, some prominent results obtained by selecting phishing indicators.

**KEYWORDS :** Phishing and Anti-Phishing, Legitimate Webpage, Phishing Webpage, Bayesian Classifier

## INTRODUCTION

The term Phishing is emerged for spoofing websites which are used for stealing confidential information of the web user such as banking passwords, credit card credential and user's private information on the web. An unaware user about phishing, inter the confidential information in such type of websites and get lost their information. The research on the topic 'Phishing' is being continuing because of different phishing attacks are generating day-by-day with different techniques and software use. The status of the legitimate and phishing websites which are identified by the APWG in the third quarter of the 2016 is as given below in Table 1. [1]

***Table 1 : Statistical Highlights for 3rd Quarter 2016 by APWG***

| | **July** | **August** | **September** |
|---|---|---|---|
| Number of unique phishing websites detected | 1,55,102 | 1,04,349 | 1,04,973 |
| Number of unique phishing e-mail reports (campaigns) received by APWG from consumers | 93,160 | 66,166 | 69,925 |
| Number of brands targeted by phishing campaigns | 358 | 340 | 361 |

A number of techniques can be used to deceive the user and alter the user's data by spreading virus, malwares, worms etc. Apart from the spam attack, the cyber criminals are turning to the social networks to launch their phishing website attack. The varying nature of attacks, the network user incorrectly assuming that they are not at risky condition. The attacker takes benefit by using these sites to target new victims. The goal of this research study is to analyze the previously defined anti-phishing systems, its performance effectiveness and to provide the best possible solution to countermeasure the phishing attack. The Phishing Activity Trends Report, 3rd Quarter 2016 reported by APWG for the most targeted industry sectors is shown in Figure 1.

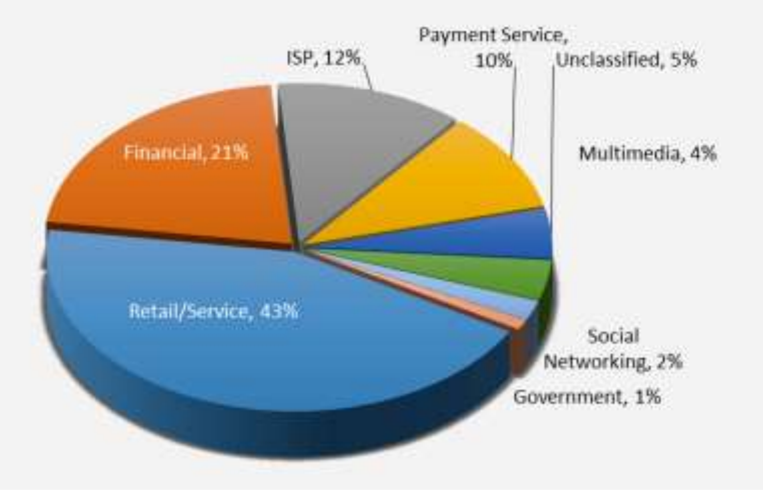


***Figure 1. Phishing Activity Trends Report, 3rd Quarter 2016 reported by APWG***

## CATEGORIES OF PHISHING ATTACKS

Since the attacker can use different methods to spoof the user, so it is necessary to find all the possible techniques and methods which attacker can use. A number of methods are exists to construct a phishing URL. Each method involves some different form of obfuscation technique. The phishing attacks can be categorised in the following forms [2] (Doshi, S., Provos, N., Chew, M., Aviel, & D. Rubin (December 2006).

* **Form – I : Use of IP address for obfuscating the Host**
The IP address can be used to obfuscating the user by using it in place of URL's host name and usually the organization being phished with the use of this obfuscating path. Since the IP address is having four field uses decimal form, but the field can use the hexadecimal numbering to spoof the user.

- **Form – II : Changing the domain name for obfuscating the Host**
In this type of attack, the URL's host name uses the similar name of the legitimate website with minor changes in the URL. This form of attack usually tries to imitate URLs containing a redirect page so as to make it appear valid.

- **Form – III : Use of Large Host Names**
This type of attack includes a number of letter and special symbols to make the long URL. The aim of this type of attack is to confuse the user about the URL. The webpage direct the user to phishing webpage. This attack is used to target the organisations and uses large string of words and domains after the hostname.

- **Form – IV : Spelling change in Domain Name**
If the user does not check the proper address, he/she can be redirected to the spoofed webpage. The technique uses changes in spelling of the URL name e.g. 'www.google.com' can be used as 'www.goo1e.com'. Both the URLs are looking similar, but in the second URL, one ('1') is written in place of letter 'l'.

## DATA MINING APPROACH

The following table shows the earlier work on the topic of phishing and anti-phishing system designing approach, the features, mechanism and algorithms selected by researcher to tackle the phishing problem. The table also shows the drawback of the proposed method.

*Table 2 : Comparison of Data Mining Approach for Phishing Website Detection*

| Author | Features Selected | Mechanism | Algorithms | Drawback |
|---|---|---|---|---|
| Bergholz et.al. [114] [3] | Study of phishing e-mails by statistical filtering method | Based on the trained classifier of features obtained | Markov Chain Model | Time consuming, Large number of features, Much memory requirement |
| Bazargani Gilani [115] [4] | Heuristic Selection Method for Text Classification of Phishing e-mails | Working in 5 steps | Nave Bayes Classification | Accuracy is low as compared with other techniques |
| Chandrasekaran et.al. [116] [5] | Structural features | The Prototype Implementation | Support Vector Machine (SVM) | Small datasets, only 200 e-mails testing, time consuming |
| Ganger et.al. [117] [6] | Training of Smart Screen | Using the feedback of the users | Bayesian Statistics Method | Lower level of recall measurement, fix number of features are used |
| Chandrasekaran Chinchani et.al. [118] [7] | PHONEY : Mimicking user response system | Proposed technique is installed at user's system | Response received from Mimicking user | Less data collected, time consuming method |
| Maofoghi et.al [119] [8] | Robust classifier model | 7 hybrid features, model consists of five stages | Information gain algorithm, Decision Tree Algorithm | Using a few numbers of features, non standard dataset |
| Fettlsadeh et.al. [120] [9] | PILFERS prototype method | 10 features included for WHOIS query | Random Forest and SVM | e-mails does not classified properly |
| Proposed Study | 15 Phishing criteria | Classification | Bayesian Classification | Tested better performance as compared to other methods |

## DISADVANTAGES OF EXISTING SYSTEMS

On the basis of the study of earlier work and the report of Computer Associate Internationals Inc. published in September 2012, the author have identified some weak point, which are given below [10] :

a. Blacklist-based technique with low false alarm probability. This type of system does not detect the phishing website if the website is not stored in the blacklist database. Because the life cycle of phishing websites is too short and the establishment of blacklist has a long lag time, the accuracy of blacklist is not too high.

b. Heuristic-based anti-phishing technique, with a high probability of false and failed alarm. The attacker can take the benefit by finding technical means to avoid the heuristic characteristics detection.

c. The proposed similarity assessment based techniques are time-consuming. There is low accuracy rate for this method depends on many factors, such as the text, images and similarity measurement technique.

## CRITERIA TO FIND PHISHING WEBSITES

A number of keyword can be used to identify the websites whether it is phishing or legitimate. In this study, only 15 essential phishing indicators have been taken which can decide the website category. It is analysed that when we select less number of indicators, the system tool take less time to send the feedback to the user about the type of website. The functioning of the proposed method is based on the checking and testing of the following indicators/check points:

1. Number of ' . ' present in the URL
2. Number of '@' present in the URL
3. Number of ' // ' present in the URL
4. Existence of IP address in the URL
5. Port Number in the URL
6. The websites which are having HTTPs protocol

7. Number of Phishing Keywords present in the URL
8. Country Code present in the URL
9. Title Tag
10. Form Tags
11. Image Tags
12. href Tags
13. Login/Password evaluation
14. Script Tags
15. Link Tags

The spoofing website remains almost similar to the legitimate website so that the user can be spoofed easily. The spoofed website matches almost 90 to 99% to the legitimate website. As per the previous study, it is found that the above mentioned points are generally used to design the phishing websites.

## BAYESIAN CLASSIFIER FOR PHISHING WEBSITES DETECTION

The Bayesian theorem is generally used to solve the prediction problems. According to this classification algorithm, two data sets can be cross checked on the basis of probability measure. If stored legitimate web page data in database is denoted by 'A' and the hitting website by the user which has to be cross checked with legitimate site from its database is denoted by 'B', than the Bayesian algorithm functions as follows:

Suppose we have related events (target websites) denoted by 'B' and other mutually exclusive events (stored legitimate websites) denoted by A1, A2, A3, … A*i* . To find the probability of B when a randomly selected target website (suppose A5) is

$$P(A5 \mid B) = \frac{P(A5 \& B)}{P(B)} \qquad (1)$$

Where the symbol ' | ' denote the term "given".

We can also find the probability of B with respect to each phishing feature sampled A5 with the following formula :

$$P(B \mid A5) = \frac{P(A5 \& B)}{P(A5)} \qquad (2)$$

Here we can multiply P(A5) and substitute it for P (A5 & B) to find

$$P(A5 \mid B) = \frac{P(A5)P(B \mid A5)}{P(B)} \qquad (3)$$

To find the probability of phishing for B, the P(B) can be written as

$$P(B) = P(A1)\,P(B \mid A1) + \ldots\ldots + P(A10)\,P(B \mid A10) + P(Ai)\,P(B \mid Ai)$$

So, to get the final form of Bayesian theorem, we can substitute P(B) for P(A5|B) like

$$P(A5 \mid B) = \frac{P(A5)P(B \mid A5)}{P(A1)P(B \mid A1) + \ldots\ldots P(A10)P(B \mid A10)} \qquad (4)$$

On the basis of above formula, the probability of the phishing websites can be measured with following situations :

- The probability that the features of both the websites A and B is not matching is P(B=0) = 0/15 = 0; that means the target website is legitimate.
- The probability that the website is phishing is P(B=1) = 1 – P(B=0) = 1 – 0 = 1

- The probability that the features of A and B are matching above 50 percentage is P(A=1 | B=1/2) = 0.50; that means the target website is suspicious.
- The probability that A is legitimate and B is phishing is P(A=1 | B=0) = 0.25; which means the target website can be kept for the further checking for remaining matching situations.

For each value of a phishing features (point 1-15 in previous section), in discriminating phishing from non-phishing can be determined by examining the Odds Ratio for that feature. By using each phishing feature, we can define hypothesis. By comparing different hypothesis, we can estimate the odds ratio. The higher the Odds Ratio, the higher chances of being a phishing website. The Odds Ratio can be calculated as

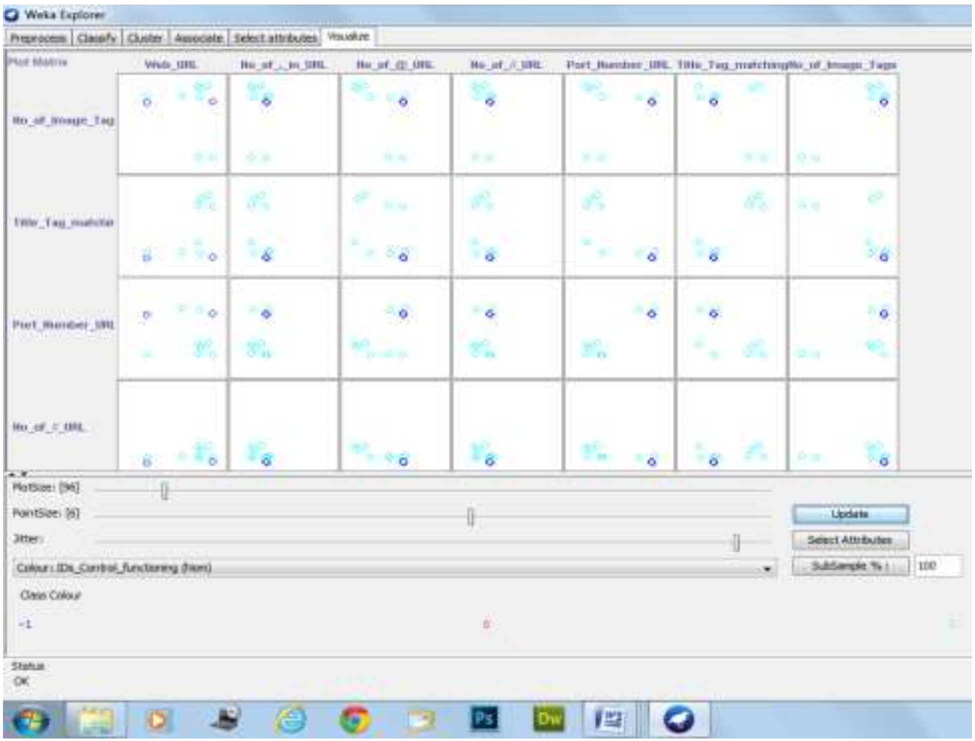
$$OddsRatio = \frac{P(H1|Phishing)}{P(H2|Non-Phishing)}$$

Where *H1* and *H2* denote the Hypothesis-1 and Hypothesis-2 for different phishing features present in the hitting website.

## RESULT AND DISCUSSION

To find the Odds Ratio, the author hit 512 phishing and 780 non-phishing websites which are declared by Advanced Phishing Working Group in the third quarter of 2016. To extract the selected indicators, a program was written in Java language, and the algorithm was implemented in WEKA. We have evaluated the performance of classifier on the basis of selected indicators.

Figure 2 shows the graphical representation of the phishing indicator in the form of plot matrix under Weka Visualize. The plot shows matching of phishing indicators with the legitimate webpage datasets. By choosing selected attribute (phishing indicator), we can compare the legitimate data set with the phishing data set.



PlotSize: [96]
PointSize: [6]
Jitter:
Colour : IDs_Control_functioning (Nom)
SubSample % : 100
Status
OK

*Figure 2 Weka Visualize result for four phishing criteria showing matching of phishing dataset with legitimate dataset*

The phishing features has been analysed using WEKA Bayesian classifier and obtained the odds ratio for selected key features. The following table shows the phishing features as keywords and its Odds Ratio.

*Table 3 : Odds Ratio for Bayesian Classification based on the phishing features with its description*

| Keyword | Description | Odds Ratio |
|---|---|---|
| No_links | Number of links in the webpage | 122 |
| Link_IP | Links that contains IP address | 18 |
| No_IntExt | Number of Internal and External links in the website | 36.5 |
| Link_Image | Number of links for images in the webpage | 24.0 |
| User name | The field, that contains the entry of user's name | 80.0 |
| Login | This field ask for user's ID | 88.6 |
| Password | It requires the password of the user | 140.4 |
| Copy right | It shows the security notation on the webpage | 25.6 |
| FAQ | Simple link 'Frequently Ask Question' for the user | 28.2 |
| Contact us | Simple link | 16.8 |
| Privacy Statement | Security conditions for the webpage | 8.4 |
| Terms & Conditions | Simple link which redirect the user at another page | 7.5 |
| HTML, Java Scripting and Forms | Types of method to write the webpage coding | 6.3 |
| Check_date | Checking the Current website visited previously or not | 2.5 |
| Home | Simple page which shows the general information of the webpage | 2.6 |

Based on the above example, the odds ratio is higher for the keyword 'No_links' and 'Password' which shows strongest indicator of phishing. When the odds ratios of different keywords are considered as together, the probability for a particular website can be estimated effectively whether it is phishing or not.

To perform the calculation for a Bayesian filter for identifying the type of website, only two datasets are required; a dataset of phishing and a dataset for legitimate websites which are already declared by the authority. Table 4 shows the test performed on the phishing and legitimate websites and the obtained results in the form of False Positive (FP) rate and False Negative (FN) rate. The less value of False Positive and False Negative means higher accuracy in the tested dateset.

***Table 4 : Test result performance of websites by Bayesian Classification Method***

| | **Phishing** | **Legitimate** | **No Outcome** |
|---|---|---|---|
| **Hitting Websites** | **512** | **780** | **16** |
| False Positive | 16 (3.23%) | 18 (1.98%) | 12 |
| False Negative | 12 (2.5%) | 10 (1.1%) | 4 |

The above results shows that the probability of False Positive rate is around 3.23% for identifying phishing websites which means that in the Bayesian Classification, the number of phishing websites can be calculated around 96.77% while the probability of False Negative rate is around 2.5% which shows that legitimate websites finding percentage is around 97.5%. During the experiment, 12 websites doesn't found as phishing or legitimate, because of non availability of the web contents or may be no existence of these websites. Since the Bayesian Classification is based on the probability of the situation, the result varies little bit by hitting the target website repeatedly.

## CONCLUSION

The term Phishing is a kind of spoofing website which is used for stealing confidential information of the user. The Bayesian classifier approach is based on the probabilistic relationships between the attribute set and the class variable. The phishing websites are designed to looks like legitimate transactional correspondence and almost always work to foolish and steam the confidential information of the user. The result of *Bayesian Classification* shows that the False Positive rate is around 3.23% which means that in the Bayesian Classification, the number of phishing websites can be calculated around 96.5 percentages while the legitimate website finding percentage is around 98 percentages.

## ACKNOWLEDGEMENTS

The author would like to acknowledge the Anti-Phishing Working Group for providing database of phishing and anti-phishing websites. My sincere gratitude to the research committee for providing me the guidance for data collection, valuable guidance, suggestions and encouragement throughout the work.

## REFERENCES

[1] APWG Third Quarter Report 2016, https://docs.apwg.org/reports/apwg_trends_report_q3_2016.pdf

[2] S. Doshi, N. Provos, M. Chew, Aviel, D. Rubin "A Framework for Detection and Measurement of Phishing Attacks", A Technical Report from Johns Hopkins University, December 2006

[3] A. Bergholz "Improved phishing detection using model-based features" in Proc. Conference on E-mail and Anti-Spam (CEAS), Mountain View Conference, CA, August 2008.

[4] M. Bazarganigilani, "Phishing E-Mail Detection Using Ontology Concept and Nave Bayes Algorithm" International Journal of Research and Reviews in Computer Science, Vol. 2, No. 2, 2011.

[5] M. Chandrasekaran, "Phishing email detection based on structural properties", in New York State Cyber Security Conference (NYS), Albany, NY 2006.

[6] D. L. Ganger, "Gone phishing: Evaluating anti-phishing tools for Windows - A Technical report", September 2006.

[7] M. Chandrasekaran, "Phoney: Mimicking user response to detect phishing attacks", In: Symposium on World of Wireless, Mobile and Multimedia Networks, IEEE Computer Society, pp. 668-672, 2006

[8] L. Ma, "Detecting phishing emails using hybrid features", IEEE Conference, pp. 493-497, 2009

[9] I. Fette, "Learning to detect phishing emails" in Proc. 16th International World Wide Web Conference (WWW 2007), ACM, New York, NY, USA, pp. 649-656, May 2007

[10] A Report by 'Computer Associate Internationals Inc.', http://www.scientificcomputing.com/ company-profiles/computer-associates-international-inc, September 2012