ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

# A Survey of Phishing Website Detection Techniques

Jyoti Vaibhav Jadhav, Prof. Krishna Kumar Tripathi

M. E Scholar, Department of Computer Engineering, Shree L.R. Tiwari College of Engineering, Mumbai University, Maharashtra, India

Assistant Professor, Department of Computer Engineering, Shree Shivajirao S. Jondhale College of Engineering, Mumbai University, Maharashtra, India

**ABSTRACT**: Fraud websites steals identity of the user automatically without user's awareness. These websites or web pages convince the users to reveal their personal or financial information like username, credit card number, password, account number and use this information for their monetary gain. Fraud websites looks exactly like their legitimate counterparts. So these anti phishing techniques are mostly machine learning based in which different features are used which are extracted from various sources. In this paper, we present the survey of fraud website detection approaches. This is the survey where most of the proposed techniques for detecting fraud websites are discussed.

**KEYWORDS**: machine learning, phishing, data mining, fraud websites, legitimate websites

## I. INTRODUCTION

Creating fraud website is one of the dangerous cyber security threats. In this the fake page is used by the attacker to obtain personal or financial information of the user for monetary gain or to get recognition. Such attacks are called as phishing attacks. Detecting and preventing such attacks is very difficult for researchers. Educated and experienced users also get victimized of such attacks. Usually attacker will perform such attacks by sending emails which has embedded fake web pages. Sometimes attacker uses big brand names in their fraud URLs to fool users. Usually attacker creates replica of the legitimate website and an email. The email looks like an email of a financial company or a reputed bank. The link of the fake web page is given in the email and when the victim clicks the link he will be redirected to a fake website. It will ask for passwords, account number, user id etc and this information is used against the victim.

There 3 major steps in phishing life cycle.

1. Cybercriminal creates the fake web page and send it to user.
2. User enters personal information assuming it as a genuine page.
3. Attacker steals this information.

Phishing attacks are broadly classified into four types:

Zero day: in this attack, the attacker attack the victim's website bypassing the built in security measures. Such attacks are not taken care or considered by anti-phishing techniques. These attacks are difficult to avoid.

Embedded Objects: original webpage is modified to create the fake webpage. Flash objects or images are embedded instead of html content so that these embedded objects will not be noticed by anti phishing techniques.

DNS: in this attack attacker exploits DNS vulnerabilities such that all the traffic will be rerouted to some fraud website.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

According to APWG (Anti Phishing working Group) second quarter report 2016[1], there are total 466,065 unique phishing sites observed in second quarter of 2016. So according to this report it was 61% higher than the previous quarterly record in fourth quarter 2015.

The Comodo Threat Research Labs (CTRL) team has discovered a phishing attack which was targeting ICICI bank's customers which is India's one of the largest private sector bank having branches all over the country [2]. The team identified that customer was getting an email which will appear legitimate. The email has a web link and customer has to click on the link. And customer was asked to enter financial information. The landing page was looking exactly like the real website as shown in the following figure.


## II. LITERATURE SURVEY

Phishing website detection techniques are broadly classified into two categories, user education and software. In user education approach user has to be educated about the safe browsing practices. Software approach has different machine learning based techniques. Some of these techniques are explained in detail here.

### A. CANTINA

Zhang et al [3] proposed CANTINA for detecting fake website. It used the frequency–inverse document frequency (TF-IDF) information retrieval algorithm. This algorithm is actually used for comparing, classifying and retrieving documents from a huge collection of documents. This algorithm is used for detecting phishing websites. TF (term frequency) is the number of times the term appeared in the specific document and IDF (inverse term frequency) is the general importance of the term in the whole set of documents. So if the TF-IDF of the term is zero then the term has no importance in the document that means it is very common.

CANTINA works in the following steps:

1. Calculate TF-IDF of each term on that web page.
2. Find first five terms with highest TF-IDF weights.
3. Generate lexical signature with these five terms.
4. Enter this lexical signature to Google.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

5. Compare the domain name of the web page to the top Google result domains.
6. If it matches then the web page is legitimate otherwise fraudulent.

CANTINA provides good accuracy but has a good number of false positives. To address this false positive problem a large set of heuristics has identified with proper weights.

Some of the heuristics are:

- Domain age: This heuristic defines the age of the domain. Most of the phishing websites are only two to three days old.
- Logos: This heuristic defines the standard logos defined by the brand.
- Suspicious URL: This heuristic check whether the URL has at (@) or dash (-) in the domain name.
- IP Address: This heuristic checks whether the domain name is an IP address.

B. CANTINA+

Xiang et al. [9] proposed CANTINA+ is an upgraded and modified version of CANTINA. It is feature based approach. CANTINA+ includes 8 novel features which uses HTML, DOM, search engines, third party services with machine learning techniques. Author tested the system with a true positive over 99%. Zhang et al [3] tested CANTINA with 100 legitimate and 100 fraud URLs. But CANTINA+ is evaluated on a larger set of URLs. This system increases the speed of the system and reduces the false positive. CANTINA+ works with different machine learning algorithms.

CANTINA+ system works in the following steps:
Training Stage:

1. Extract 15 features from each URL from the training data set.
2. Arrange the feature values in proper format.
3. Forward it to machine learning engine.
4. Build the classifier for phishing detection.

Testing stage:

1. Check whether the incoming page is already phishing webpage using hash filter
2. Use login form detector to classify webpage
3. Classify the webpage as legitimate if no login form is present.
4. Send the webpage to feature extractor if login form is present.
5. Run pre trained models to classify the webpage.

C. ASSOCIATIVE CLASSIFICATION DATA MINING

Neda Abdelhamid et al [4] proposed Multi-label Classifier based Associative Classification (MCAC) method for website phishing. Associative classification is used in detecting phishing websites with good accuracy. MCAC is used to generate new rules for enhancement. It does not consider the content based features of the website. MCAC method create hidden rules which could not be generated by previous algorithms.

MCAC algorithm works in following steps:

1. Look for hidden relation between the attribute values and the class attribute in the training data set.
2. Create association rules using this relation.
3. Sort the rules using an efficient sorting algorithm based on the confidence and support.
4. Duplicate rules are ignored.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

5. Measure the accuracy in terms of false positive and true positive.

D. CLASSIFICATION MINING TECHNIQUES

Aburrous M [5] proposed a technique based on Classification Data Mining (DM) for detection on e-banking phishing websites. In this technique author has implemented six different classification algorithms. And all these algorithms are tested to calculate their accuracy. The classification algorithms implemented are C4.5, JRip, PART, PRISM, CBA and MCAR. Each algorithm has its own strategy to learn rule from data set. C4.5 algorithm uses divide and conquer strategy. RIPPER algorithm uses separate and conquer strategy. PART algorithm uses combination of both. PRISM algorithm handles only nominal attribute. CBA algorithm uses association rule mining. MCAR algorithm works in two steps. In first step is for rules generation and second step is for classifier builder. This approach is called as associative classification (AC) approach.

Associative Classification works in following steps:

1. Scan training data set to find the frequent data items.
2. Combine the items recursively to find more items.
3. Generate the rules.
4. Rank the rules.
5. Store the rules.
6. Generate the classifier by using the rules.
7. Test the test data using the classifier.

E. A SVM APPROACH

H Huang [6] et proposed an approach based on the URL based features. Author has taken 23 features from URL and trained the system using SVM. The vector consists of 23 features which is used to model SVM. The vector consists of 4 features which are structure features of URL, 9 features are lexical and 10 features are brand name of the website.

Structure features: Structure features are IP address, number of dots in the URL, length of host name, number of dash in the URL.
Lexical features: Author selected tokens such as http, banking, login, secure, sign in as lexical features. Brand name features: 9 brand names which are selected for the SVM approach are eBay, PayPal, sulake, orkut, facebook, Santander, visa, warcraft, bradesco.

SVM based technique works in following steps:

1. Extract 23 features from training dada instance.
2. Organize feature vector in LIBSVM format.
3. Feature values and vector format are same as training stage in classifier stage.
4. Output label value specifies whether the URL is legitimate or fake.

F. URL-ASSISTED BRAND NAME WEIGHTING SYSTEM

C. L. Tan et al[6] proposed phishing website detection using URL-assisted brand name weighing system. A brand name is a name given to a legally registered product, service or business and is used as a unique identity [7]. Along with brand logo user always look for brand name in the URL. The brand name appearing in the URL has more importance if it appears in the beginning of the URL.

URL-assisted brand name weighing system works in following steps:

1. Extract words from web page content.
2. Assign TH-IDF weights to them.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

3. URL-assisted system calculates further weights which will be added up to the initial weights.
4. Sort all words by weights and identify most weighted words as brand names.
5. Submit the brand name to search engine.
6. Perform WHOIS lookup to obtain domain name owner.
7. Successful match shows the page is legitimate.

G. NOVEL APPROACH USING AUTOUPDATED WHITE LIST

Ankit Kumar Jain et al [8] proposed a novel approach to detect phishing attack using auto updated white list. Author has designed the system with two modules, URL and DNS module. DNS module has while list and white list contain two attributes domain name and its corresponding IP address. It checks the legitimacy of the website by using hyperlink features. This approach is suitable to real time environment. It detects almost all types of attacks.

Auto updated white list based technique works in following steps:

1. Match the domain name of the current website to the white list.
2. If domain name matched then match the IP address.
3. If IP address matches that means the current website is legitimate.
4. Start second module (if the user is accessing the website first time then the it is not there in white list) which checks whether the webpage is phishing.
5. Examine features of hyperlinks using phishing detection algorithm.
6. Warn the user if the website is fake.
7. Update the while list with new website name if it is legitimate.

## III.COMPARISON

TABLE1 ADVANTAGES AND LIMITATIONS OF THE PHISHING WEBSITE DETECTION TECHNIQUES

| Technique | Advantages | Limitations |
|---|---|---|
| CANTINA | Fast | Accuracy fully dependent on the TF-IDF algorithm |
| CANTINA+ | True positive is very high over 99% and the system speedy | Can not detect DNS compromised URLs |
| ACdata mining | MCAC applicability in phishing detection | Accuracy is not calculated |
| Classification mining techniques | Designed for financial phishing detection | False positive is high |
| SVM Approach | Content independent | Accuracy is low |
| URL-assisted | True positive rate is | Brand logo can not be incorporated as a |

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

| | | |
|---|---|---|
| brand name weighing system | very high | search attribute |
| Novel approach using auto updated white list | False positive rate is very less | Running time is higher if other features are added with hyperlinks. |

## IV. CHALLENGES IN FUTURE

We have surveyed different phishing detection techniques in detail. There is no technique which can detect all types of phishing websites. The accuracy of the techniques depends on so many parameters. It depends on features selection, algorithm used and the training data set. Zero hour attack cannot be detected with these techniques. Now days, cybercriminals are innovating new ways to fool users so researchers also have to try out new and innovative ways to protect the system. One more barrier is the language of the website e. g. Amazon is in different languages. These techniques may not be able to detect keywords from different language website. So this issue has to take care in future. One more thing is embedded objects. Attacker uses objects which are embedded in the page itself so it is very difficult to detect in comparison to text. These techniques need large computational power.

## V. CONCLUSION

Phishing is the most dangerous cybercrime in today's world. It is the serious threat to internet users. Phishing websites asks user to update their personal and financial information like account number, credit card number, pin number etc. The fake website captures and steals this information for illegal use. Cybercriminals are practising different strategies every day to crack the security system. So there should be phishing detection techniques which will detect all these attacks. We have surveyed different phishing website detection techniques in this paper. These different techniques use different features of web page like text, URL, certificates etc. Approaches discussed in this paper have different limitations on accuracy and performance. All these techniques require high computational power in real time environment. So in future we should come up with a solution with a high accuracy and moderate computational power.

## REFERENCES

[1] https://docs.apwg.org/reports/apwg_trends_report_q2_2016.pdf

[2] https://blog.comodo.com/it-security/new-phishing-attack-targets-icici-bank/

[3] Y. Zhang, J. Hong, and L. Cranor, "Cantina: a content-based approach to detecting phishing web sites," in Proceedings of the 16th international conference on World Wide Web, 2007.

[4] Neda Abdelhamid, Aladdin Ayesh, Fadi Thabtah, "Phishing detection based Associative Classification data mining," Expert Systems with Applications, Volume 41, Issue 13, pp. 5948-5959, 2014

[5] M. Aburrous, MA Hossain,K Dahal ,T. Fadi , " Predicting phishing websites using classification mining techniques," In:Seventh international conference on information technology, Las Vegas, Nevada, USA,2010.

[6] C. L. Tan, K. L. Chiew and S. N. Sze , "Phishing website detection using URL-assisted brand name weighting system," *2014 International Symposium on Intelligent Signal Processing and Communication Systems(ISPACS)*,Kuching,2014,pp.054-059.doi: 10.1109/ISPACS.2014.7024424

[7] J. McLaughlin. (2014, June) What is a brand, anyway? Forbes.com. [Online]. Available: http://www.forbes.com/sites/jerrymclaughlin/2011/12/21/what-is-a-brand-anyway/

[8] Ankit Kumar Jain, B.B. Gupta, "A novel approach to protect against phishing attacks at client side using auto updated white list," EURASIP Journal on Information Security (2016) 2016:9

DOI 10.1186/s13635-016-0034-3

[9] G. Xiang, J. Hong, C. 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, 2011.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

## BIOGRAPHY

**Jyoti Vaibhav Jadhav** is a ME Scholar in the computer engineering department, Shree L.R. Tiwari College Of Engineering, Mumbai University.

**Prof. Krishna Kumar Tripathi** is assistant professor in the computer engineering department, Shree Shivajirao S. Jondhale College Of Engineering, Mumbai University.