International Journal of Recent Trends in Engineering & Research

ISSN (ONLINE) : 2455 – 1457

IMPACT FACTOR : 4.101

# URL THREAT DETECTION SYSTEM

**Prof. Ramachandra Rao Meka** [BE,M.Tech,Ph.D][1], **Adarsh J Hiremat**[2], **Chetan M Kuri**[3] , **Harsha G Deshpande**[4] , **Sanjay S Yadav**[5.]

[1,2,3,4,5] Computer Science and Engineering, AGMR College of Engineering and Technology, Hubli,

**Abstract:** The World Wide Web is emerging quickly and was many number of users and this end users all subjected to the many threats. Among these threats phishing is the most persistence of all attacks and detection of this is the most important task in order to secure the end. Performance can be applying the capability of machine learning to domain of phishing attacks.

## I. INTRODUCTION

PHISHING is an identity theft based internet fraud. In ecommerce domain, a constant spur is seen in this kind of attacks. An adversary sends a mail to the user posing as a financial institution. The mail comes with a message that urges the user to urgently click the given link and update the account information. The mail contains the URL of a web page. The user is tricked as the website put up looks exactly like the original site. It can include logos and images also. So the modus operandi of phishing attacker is, 1) Create a fake URL. 2) Create a fake page which resembles the original site and link it to the URL. 3) Send mail to the user posing as the genuine operator. 4) When the user enters the fake site and gives his credentials, then they are captured and fraud is committed. This type of attack comes under the category of impersonation attack. By superficially looking at the URL, it is not possible for a generic user to detect if the URL points to a phishing site.

## II. LITERATURE REVIEW

A literature survey is a text of scholarly paper which includes the current knowledge including substantive finding as well as theoretical and methodological contribution to a particular topic.

[1] Fonseca et al "have proposed a methodology where web application security mechanism is tested before deployment. Vulnerabilities are injected into the web applications, and their effect is reported. This way before deploying the system itself the system gets tested in real life scenarios and the vulnerable points can be fixed. A prototype has been built to automate the process of injection attack, analyzing the effect on the application and then publishing the results

[2] Ahmed Abbasi, et aI," have researched on the topic of detecting fake medical websites. The techniques used for detecting fake websites are graph-based classifiers and recursive trust labeling. They have analyzed the different features of the websites which will be able effectively to distinguish fake and genuine websites in the medical domain. proposed detection of phishing websites using similarity detection by using Earth Mover's Distance(EMD).The websites that are suspected to be fake are those, whose
URLs are present in apparent.

[3] Zhang, Liu, et altalk about two classifiers and an algorithm to fuse the result of both of them. One is a text classifier which uses naIve Bayes rule to perform classification, and the other is an image classifier which uses Earth Mover's distance. The algorithm that fuses the image and visual classifier

uses Bayes theorem. A Bayesian approach is used to calculate the threshold of both the classifiers through offline training.

[4] Xiang, Hong have worked on a feature rich machine learning framework named CANTINA+. They concentrate on the two important features of a phishing scam. 1) A website that is a mirror image of a financial institution's site. 2) A fake login page asking for sensitive information like password, credit card number, etc. on behalf of the financial institution. From these works, it is conspicuous that for the purpose of classification, the websites have been deconstructed. For each web page that has been considered, the corresponding features have been extracted and used for classification

A literature survey is a text of scholarly paper which includes the current knowledge including substantive finding as well as theoretical and methodological contribution to a particular topic.

## III. METHODOLOGY

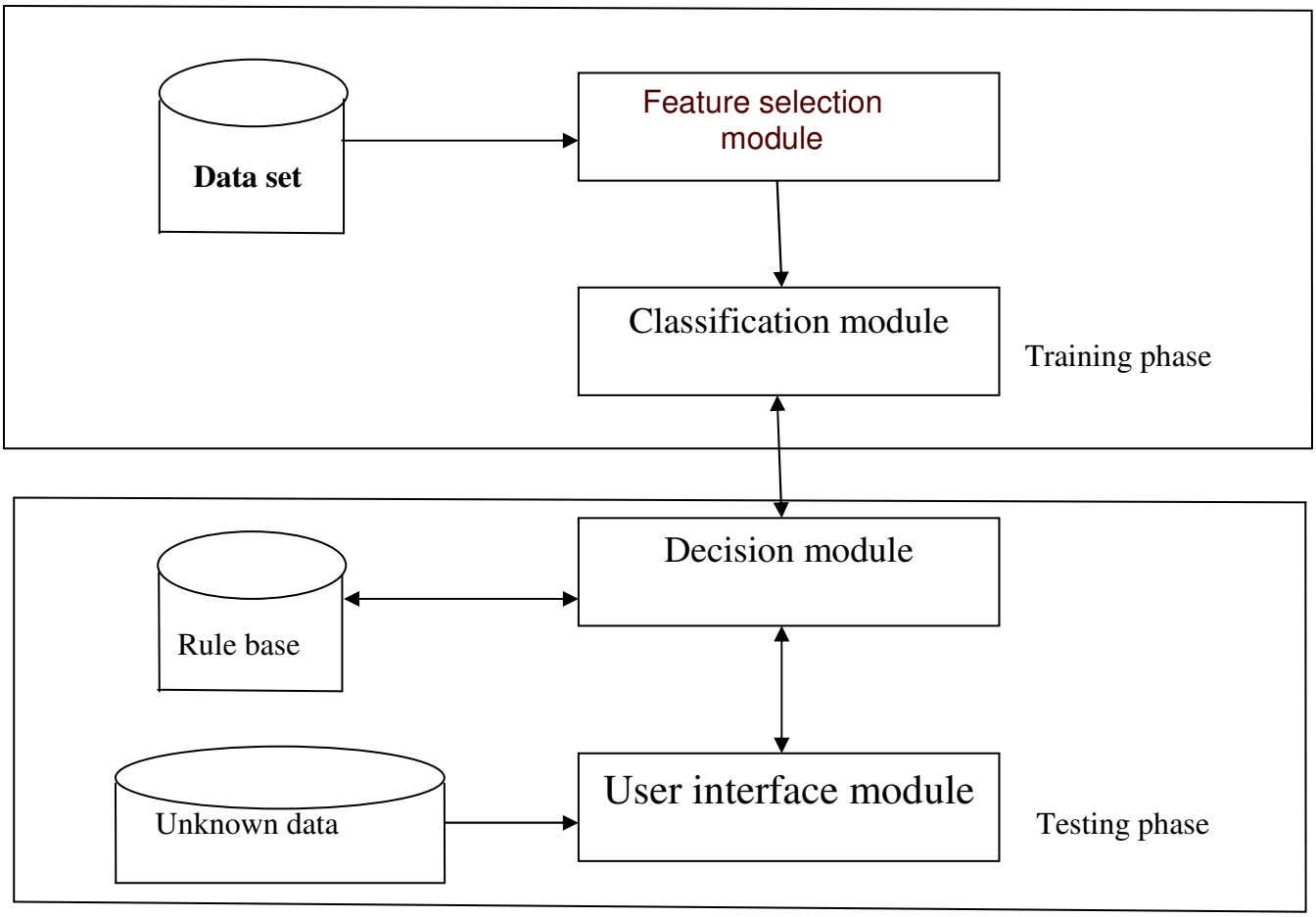


**Fig. System Architecture**

In the above system architecture the data collected is first passed through the training phase where it undergoes feature selection and classification. In the testing phase, when unknown data is presented through the user interface, the decision module along which the rule base classifies the unknown data based on the inferences made from the testing phase.

 

**1) Lexical features:**

• Presence of IP address:
E.g., http:// 185 .28.22. 67 Ibchileperfilam iento/Process?MI D=&AID=LOGIN-0004&RQI=500 1435125BE97. While creating a URL, IP address is not included in it. It is masked by the domain name. (IP addresses are not used on the internet environment but are frequently used in intranet). So the presence of IP address in the URL is a sign of the URL being a phishing one, because it might have been used to disguise the phishing domain name.

• Unknown Noun Presence:
E.g., http://emmhhh.ru:8080/forum/links/column.php. Domain names are not created by using some random letters. They use proper or common nouns.
A dataset has several URLs was collection for the purpose of performing this classification. Of these, 2500 are genuine URLs, and 2000 are phishing URLs. The genuine URLs have been collected from the DMOZ repository. DMOZ is considered one of the large human-edited directories on the web. It is applied by many famous sites like MacAfee, Yahoo mail, Kaspersky, Mozilla etc., for blacklisting URLs. URLs were collected from different domains to introspect The addition of this feature has increased accuracy considerably.
• Count of the number of dots present:
Previous studies show that the number of dots in the URL is more for phishing sites. So this parameter has also been included.

**2) URL based features:**
• Presence of security sensitive word: e.g.: confirm, account, banking, secure, websrc,login, and sign in.
• Suspicious symbol presence
e.g.: has a symbol like @ in the URL.
It is not general programming practice to use a symbol like @ in the URL. Whenever a @ symbol issued in the URL, all the text before it is ignored..e.g.: www.paypal.com@abc.com. Even though this looks like the link to paypal.com, the user is taken to abc.com.
• Misplaced top level domain.
e.g.:http://a9s7px4x2ys3ciy4x.Opu.ru/https/www.pay
pale.fr/Client17541982041 In the above URL, we can see that the word PayPal has been used. At a quick glance, it might look fine. But a closer look shows that the actual domain name is very weird and the word PayPal is present in the path section of URL, and it is also spelled wrong. Sowe say that the familiar top-level domain (here PayPal) is out of position.

**3) Network based features:**
• Number of sites linked to the URL: If the URL has been used or linked to many other pages which are genuine, then it is most likely that it is also genuine.
• Traffic received: The amount of web traffic that each website gets is
measured in some sites like Alexia (subsidiary of Amazon.com). This data can help to identify phishing sites as once blacklisted, phishing sites generate very less traffic.
4) Domain-based features:
• Domain Age: The date of creation of the domain is taken from the
WHOIS properties. We include this feature because phishing sites are taken down within days of their generation once their identity is revealed. So the older the site, the lesser is the chance for it to be phishing.

 

**4) URL based features:**

• Presence of security sensitive word: e.g.: confirm, account, banking, secure, web src, login, and sign in.

• Suspicious symbol presence e.g.: has a symbol like @ in the URL.

It is not general programming practice to use a symbol like @ in the URL. Whenever a @ symbol is used in the URL, all the text before it is ignored.

E.g.: www.paypal.com@abc.com. Even though this looks like the link to paypal.com, the user is taken to abc.com.

• Misplaced top level domain e.g.:http://a9s7px4x2ys3ciy4x.Opu.ru/https/www.paypale.fr/Client17541982041 In the above URL, we can see that the word PayPal has been used. At a quick glance, it might look fine. But a closer look shows that the actual domain name is very weird and the word PayPal is present in the path section of URL, and it is also spelled wrong.

**5) Network based features:**

• Number of sites linked to the URL:If the URL has been used or linked to many other pages which are genuine, then it is most likely that it is also genuine.

• Traffic received: The amount of web traffic that each website gets is measured in some sites like Alexia (subsidiary of Amazon.com). This data can help to identify phishing sites as once blacklisted, phishing sites generate very less traffic .

## IV. SYSTEM IMPLEMENTATION

In our proposed system, we recognize phishing URLs just by analyzing the URL structure. We do not click on and enter the phishing site. Hence, the time taken to analyze is drastically because we are not going to look at the URL content or page details. depicts the architecture of propose work Unknown data Selection Module Classification Module Decision Module User Interface Module

The data collected, is first passed through the training phase, where it undergoes feature selection and classification. In the testing phase, when unknown data is presented through the user interface, the decision module along which the rule base classifies the unknown data based on the inferences made from the testing phase.. It is an open content, multilingual directory of the web links. All entries in this directory are manually tested by volunteer editors. The 2000 phishing URLs have been picked from PHISHTANK It is a community-based phishing site where users submit a phishing URL and other users vote to ascertain if the URL is indeed a phishing site or not. It is applied by many famous sites like MacAfee, Yahoo mail, Kaspersky, Mozilla etc., for blacklisting URLs.

## V. CONCLUSION

Security analysts throughout the world are constantly challenged by the phishing community as new and advanced methods are developed each day. In this evolving environment, it's every researcher's main responsibility to deceive a system that can tackle the situation. This system is used to secure users from phishing attacks and data confidentiality will be increased for the users.

## VI. FUTURE WORK

- The future scope of this project will be with respect to its scalability. We intend to enhance the system performance further by incorporating an online learning mode.
- High and sophisticated graphics can be utilized for user interface development.
- This will further improve the accuracy and help to achieve better performance as the system becomes dynamic.

 

## REFERENCES

I. V. Praveen, A. Jenifer, T. Kiruthika, R. Akalya, Dr. P. Gomathi, "Modern Agriculture Development System Using Android Application", International Journal of Science Research in Computer Science Engineering and Information Technology, issue 3, vol 3, 2456-3307, 2018.

II. Krishna Chaitanya S, K. T. Ilayarajaa, Koti Muni Teja Reddy, "Android Based IOT for Agriculture Automation", International Journal of Pure and Applied Mathematics, vol 117, no. 21, pp169-176, 2017.

III. Constantina Costopoulu, Maria Ntaliani, Sotiris Karetsos , "Studying Mobile Apps for Agriculture", IOSR Journal of mobile computing and Application, issue 6, vol 3 , pp 44-99, Nov-Dec 2016.

IV. Anusha P, Dr. Shobha K R, "Design and Implementation of Wireless Sensor Network for Precision Agriculture", International Journal of Scientific Engineering and Applied Science, issue 4, vol 1, ISSN 2395-3470, July 2015.

 