International Journal of Emerging Trend in Engineering and Basic Sciences (IJEEBS)
*ISSN (Online)*2349-6967 *Volume 2 , Issue 1(Jan-Feb 2015), PP1-5*

Double Filtering for Phishing Detection on Web

Ittam Rakesh[1], PhapaleMithun[2], Dherange Sanjay[3], SawaleNilesh[4], Anuradha Thakre[5]

(*IT, Dr. D. Y. Patil College of Engineering Ambi Pune India,)*

Abstract :- The E-Commerce and On-line Trade expand, phishing has already become major crime of the several forms of network crimes. This paper represents an automatic approach for intelligent phishing web detection based on learning from a large number of legitimate and phishing website. Using Nave Bayesian(NB) classifier Uniform ResourseLocator can analyse and classified its features when webpage are parsed and it is suspicious then using SVM classifier can classifie webpage features and perform result.

Experimental results show that our approach can achieve the accurately detection of website for purpose of security, in lower time and high performance with a small sample of the classification model training set. Phishing website detail study is largely still a time-consuming manual process of discovering potential phishing websites, verifying if distrustwebsites truly are parody, distributing their URLs to the appropriate blacklisting services. we present a method for quick detection and detail examination of phishing websites automatically.

Keywords:- Support Vector Machine(SVM)classifier, Naive Bayesian(NB),Uniform Resource Locator(URL).

I. INTRODUCTION

1.1 Introduction of Phishing

The phishing is the social crime in which a phisher, attempts to fraudulently retrieve authorised users' confidential or sensitive information by mimicking electronic communications from a trustworthy or public organization in an automated fashion .

Phishing is a form of identity theft that occurs when a malicious web site impersonates a legitimate one in order to acquire sensitive information such a, details of account , or credit card numbers passwords. Even though there are several anti-phishing applicationspresent and techniques for detecting phishing trials in electronic mails and detecting phishing web contents. phisher comes up with new and artificial method to circumvent application and methods whitch are present..

The phishing attackers trick users by employing different social engineering tactics such as threatening to suspend user accounts if they left incomplete the account updating task, provide other data to check validityof their accounts or some other reasons to get the users to visit their parody web pages. for detecting phishing and the newly invented method and to remove authorization of the phishing websites by applying different algorithms to detect them.

1.2 Phishing Information Flow

A complete phishing attack involves three roles perform attacker. Firstly, mailers send out a numberof faultyemails ,which direct users to fake websites. Secondly, collectors set up fake websites, which actively prompt users to provide credential data. Finally, cashers use the credential data to achieve a pay-out. primary exchanges often occur between those attackers. The data flow is shown in Figure 1.

In this sense, phishing webs are not isolated from their targets but have strong relationships with them, which can be used as clues to find their targets.

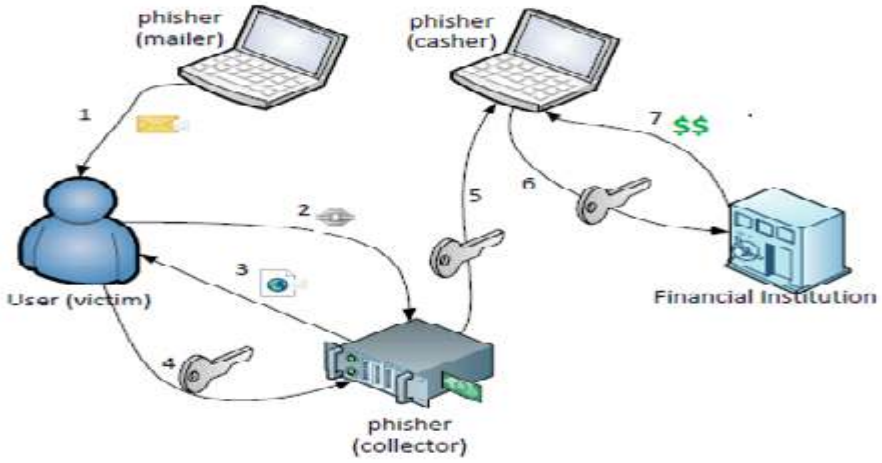
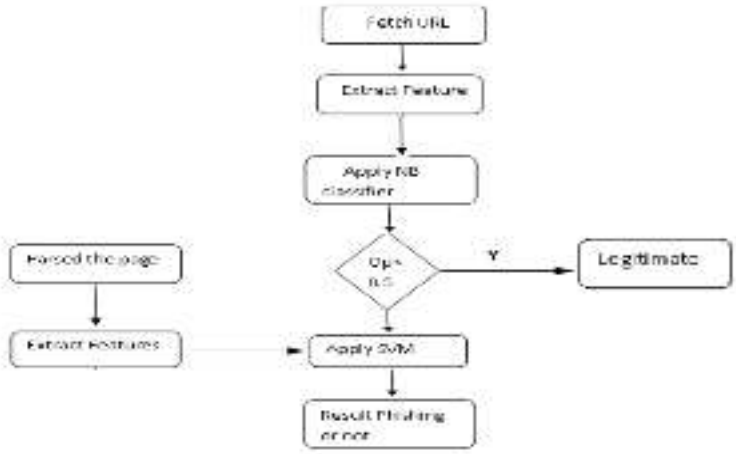


Figure 1: Phishing information flow

II. RELATED WORK

In this paper, a fast and accurate approach is proposed to detect phishing web. Our approach determines whether a webpage is a phishing web or a legitimate one, based on its Unified Resource Locator and webpage features, and is merely a combination of NB and SVM. The NB classifier used to detect the URL is that NB is a rapid detection method for classification and URL features can be easily acquired. If the NB classifier cannot judge the given web's legality definitely, the SVM classifier is used to detect it based on its webpage's importants aspects. Also our method may work together with a blacklist-based technique to provide a efficient protection.

2.1System Approach:-
System Architecture-



Step 1:- Given web A, then fetch URL.

Step 2:-extract URL identity and generate features.

Step 3:-classify A by NB classifier and return result(0:suspicious,-1:phishing, ,+1:legittimate)

Step 4:- if output is < 0.5 then it is legitimate URL otherwise it is phishing and goto step 5.

If it is exact zero then output is suspicious.

Step 5:- Now web page has no input and the output is phishing label then parsed the web page.

Step 6:-extract webpage identity and generate features.

Step 7:-apply SVM classifier and generate result in the form of -1:phishing and +1:legal.

2.2 Identity extraction

Classing a URL with a trained model is a lightweight operation compared to first downloading the webpage and using its content for classification. For our purposes, URL reputation is treated as a binary classification problem where legitimate examples are benign URLs and phishing examples are malicious URLs. Significantly, webs are classified based only on the context of the URL and the relationship between URLs and the lexical.

The webpage identity is retrieved from two sources; one is from the content of a webpage and the other is from the structure of a webpage. Therefore these features are useful to find the identity of the web page. Features extracted in identity extraction phase include META Keyword, META Description, META title and HREF of < *a* >tag. Here the webpage is parsed into the Document object model (DOM) tree.

META Tag: The <*meta*>tag gives metadata about the html file. Meta elements are typically used to specify page author, description, keyword of the document, last modified. The Meta description tag is a snippet of HTML code that comes in the Head section of a web page. It will be placed before the Meta keywords tag. The identity relevant object is the value of the content attribute in Meta tag. It consists of a description about the web.

HREF Tag: The href attribute specifies the destination of a link. When a href text is selected, it has to direct to the anxious web page. Phishers wont make any change in the destination site link. So it points to the origional web. The value of the href attribute is a URL in which the domain name has high probability to be the identity of the web.

III. URL FEATURES AND NB CLASSIFIER

3.1 URL features

Internet Protocol Address: For avoiding from domain signup or user validating , the IP address is a simple way used to hinder from verification.

Dots in URL: Many dots appearance may be caused by an attempt that the phishing web use sub-domain to construct a legitimate look of the URL Here the number of dots in a page's URL is checked.

Suspicious URL: When the phishing web try to guess the victims, the Uniform ResourseLocaters of the phishing web may be altered to the pattern that is hard to check. '@' or '-' signs in suspicious URLs is checked which are often used to modify the URL.

Slash in URL: The URL should not contain more number of slashes. If it contains more than five slashes then the URL is considered to be a phishing URL .

3.2 NB classifier

The features described are used to encode webs' URLs as high dimensional features. The NB classifier is considered more effective than other methods for learning how to classify text documents Given a set of classified process samples, an application can learn from these samples so as to guess the class of an unexpected sample. Each Uniform ResourseLocator is represented by features(**X**1, **X**2, **X**3, **X**4) are independent from each other. Each feature **X***i*($1 \le i \le 4$) takes a binary value(0 or 1) indicating whether the corresponding property appears in the URL. The probability is calculated that

$$p(Ci|\mathbf{X}) = \frac{p(C1)\times p(\mathbf{X}|Ci)}{p(\mathbf{X})} = \frac{\mathrm{p(c1)}\times \Pi p(\mathbf{x}i|C1)}{p(\mathbf{X})}$$

where all of $p(\mathbf{X})$ are constant, meanwhile $P(\mathbf{x}i|c1)$ and $P(Ci)$ can be calculated easily from training.The proportional to $\frac{P(C1j\mathbf{X})}{P(C2j\mathbf{X})}$ is calculated, and the results are shown below:

$$\begin{cases} \frac{P(C_1|\mathbf{X})}{P(C_2|\mathbf{X})} > \alpha \ \ (\alpha > 1), & a\ legitimate\ web, \\ \frac{P(C_2|\mathbf{X})}{P(C_1|\mathbf{X})} > \alpha \ , & a\ phishing\ web, \\ 1/\alpha \le \frac{P(C_1|\mathbf{X})}{P(C_2|\mathbf{X})} \le \alpha, & a\ suspicious\ web,\ need\ to\ be\ detected\ further. \end{cases}$$

For the suspicious web, SVM is used to detect it according to web's content features.

IV. WEBPAGE FEATURES AND SVM CLASSIFIER

4.1 Webpage features

Given a suspicious web P and its term identity generation step would determine the features value of the webpage. The feature vector generated in this step would then be inputted into a SVM classifier to determine whether a web is a phishing or a legitimate web. The features are categorized that are collected for web's content as follows:

Forms: If a page contains any HTML text entry forms asking for personal data from people, such as credit card number and password. Most phishing webs contain such forms requesting for personal information, otherwise the criminals dare not getting the personal information they want.

Nil anchors: A nil anchor is an anchor that points to nowhere. The more nil anchors a page has, the more suspicious it becomes.

Foreign Anchor: An anchor tag contains href attribute whose value is an URL to which the page is linked with. If the domain name in the Uniform ResourseLocator is not similar to the domain in page URL then it is called as foreign anchor.

Foreign requests: Similar to the foreign anchors, requests to the foreign domains are also a normal behavior. When there are too many foreign requests, the web could be less credible.

SSL Certificate: SSL is an acronym of socket layer. It creates an encoded connection between the web server and the user's web browser allowing for private information to be transmitted without the problems of neglecting. All legitimate webs will have SeccureSocket Layer certificate. But phishing webs do not have Secure Socket Layer certificate. Based on above features, the Forms feature is used to be a filter for dataset selection. The reason is that the dangerous pages causing users lost their information must contain forms with input block. If a webpage has not a text input, the detection is not required since users do not have a way to enter their secret information.

4.2 SVM classifier

SVM as a well-known data classification technique is applied to classify webpage features. The SVM classifier input in our approach is a 6-dimension feature vector produced from the feature generation step ($VP =< F1, F2, F3, F4, F5, F6 >$). Since a webpage is only considered as a legitimate or a phishing, it is naturally a binary classification problem. The SVM would produce output in two classes: -1 means phishing, and +1 means legitimate. Here the least squares support vector machine(LS-SVM)is applied, and the optimal model is as follows

$f(\mathbf{x}) = \Sigma n \alpha i K(\mathbf{x}, \mathbf{x}i) + b$

Where $K(\mathbf{x}, \mathbf{x}i)$ is the RBF kernel(SVM-rbf), and the form is $K(\mathbf{x}, \mathbf{x}i) = e^{-\|\mathbf{x} - \mathbf{x}i\|^2}$, $\mathbf{x}$, $\mathbf{x}i$ are webpage features. The value of α and b can be obtained by solving the following equations:

$$\begin{vmatrix} 0 & 1 & \cdots & 1 \\ 1 & K(\mathbf{x}_1, \mathbf{x}_1) + 1/\gamma & \cdots & K(\mathbf{x}_1, \mathbf{x}_n) \\ \vdots & \vdots & & \vdots \\ 1 & K(\mathbf{x}_n, \mathbf{x}_1) & \cdots & K(\mathbf{x}_n, \mathbf{x}_n) + 1/\gamma \end{vmatrix} \begin{vmatrix} b \\ \alpha_1 \\ \vdots \\ \alpha_n \end{vmatrix} = \begin{vmatrix} 0 \\ y_1 \\ \vdots \\ y_n \end{vmatrix}$$

Before the classifying of SVM, it should undergo a training process to develop a classification model. $\mathbf{x}i$and $yi(i= 1, \ldots, n)$ indicate the Feature vectors and class label of the web samples which are identified classified. If $f(\mathbf{x}) = +1$, the giving web is considered to be a legitimate one, and if $f(\mathbf{x}) = -1$, the giving web can be considered to be a phishing one.

V.EXPERIMENTS AND RESULTS

The dataset used for learning is collected from PHISHTANK [16]. The dataset with Six Hundred phishing webs and Four Hundred legitimate webs is developed for implementation. Hundred legitimate and Hundred phishing webs are taken as the training set, and the rest of Three Hundred legitimate and Five Hundred phishing pages compose the testing dataset. The robustness of the classifiers is evaluated using 10-fold cross validation. The feature vector corresponding to phishing web is assigned a class label -1 and +1 is assigned to legitimate web. Two experiments have carried out to evaluation of our method. In the first experiment, the optimal value of α is searched for. True Positive (TP)-The phishing webs that were classed asphishing web. False Positive (FP)-The legitimate webs that were classed as phishing web.

Table 1: The value of α and performance

α	TP (%)	FP (%)	α	TP (%)	FP (%)
1.1	90.68	4.78	2.1	95.34	1.32
1.2	91.32	4.31	2.2	96.63	0.74
1.3	91.91	2.93	2.3	97.35	0.31
1.4	92.30	4.08	2.4	97.71	0.13
1.5	92.99	3.69	2.5	97.71	0.13
1.6	93.06	3.78	2.6	97.71	0.13
1.7	93.58	3.15	2.7	97.72	0.13
1.8	93.89	2.78	2.8	97.71	0.13
1.9	94.20	3.01	2.9	97.71	0.12
2.0	94.88	1.73	3.0	97.71	0.13

The performance result is shown in Table 1. The value of α is tested between 1.1 and 3.0. The TP rate of NB classifier is raising to maximum 99.73% when α is equal to 2.4 and nearly keep the same when between 2.5 and 3.0. The FP rate of NB classifier is dropping to minimum 0.134% when α is equal to 2.4. So our approach provides the best performance with α= 2.4. In the second experiment, the accuracy of the three classifiers is compared: NB, SVM and our approach. The features describing the properties of URL and webpage are both used in NB classifier and in SVM classifier, including 11 features in all that are described in section 3 and 4.

Table 2: Performance comparison

–	NB	SVM	Our Approach
Train Time	50s	92s	71s
Test Time	80s	109s	90s
TP(%)	90.08	94.41	96.90
FP(%)	4.80	3.98	1.25

Table 2 shows the training, testing times and detection accuracy for each algorithm of classifier. Based on the comparison in Table 2, the accuracy of our approach outperforms the other approach while its false alarm rate is much lower than the other classifier or method. NB classifier training and testing time is shorter, but the accuracy is less. SVM classifier training and testing time is largest, and the accuracy is more than NB.

VI. CONCLUSION

In this paper, a novel approach is presented to identifying the potential phishing target of a given web. Every web claims a webpage identity, either genuine or fraud. If a web claims a fraudulent identity, not expected may exist in a network space; therefore our approach could detect and differentiate between a legitimate and a phishing web. Our approach first categorizes the URL features and test whether the page is phishing or not using NB. When the web's legality is still suspicious, then categorize its webpage features and test whether the page is phishing or not using SVM. The
experimental results show that our approach has a high detection rate and a low false positive rate. In future works, the plan is to adjust existing feature extraction methods and seek for more relevant features to get a better result.

VII. ACKNOWLEDGEMENT

Every engineering student looks toward the final year project as an opportunity by which he can implement the skill that he has eventually nurtured in the year by hard work dedication the milestone of completing the project would have been intractable without the help of few people who need to be acknowledge. We owe this moment of satisfactions with a dear sense gratitude to our internal guide **Prof. AnuradhaBobade** who guided us at every stage. Whose technical support and helpful attitude give us high moral support. We would also like to extend our sincere thanks to our **H.O.D. Prof. Ravi Patki** for his guidance and constant encouragement. We are highly obliged to the entire staff of the information technology department and principal**Dr. S.D.Shirbahadurkar** for their kind co-operation and help. We also take this opportunity to thank all our colleagues who baked our interest by giving useful suggestions and also possible help. At last but not least we are thankful to our friend colleagues and all the people directly or indirectly concerned with this project.

REFERENCES

[1] J. S. Downs, M. B. Holbrook, Decision method and easily influence to phishing, in: Proc. The second symposium on usable privacy and security 2006.
[2] Google Inc, Google browsing for Firefox, http://www.google.com/safebrowsing/
[3] NetcraftInc, Netcraftanti-phishing toolbar, http://toolbar.netcraft.com/
[4] Y. Pan, Anomaly based web phishing page detection, in: Proc. Twentysecond annual computer security applications conference
(ACSAC'06), 2006, pp. 381-392.
[5] I. He, S.J. Horng, An efficient phishing webpage detector, Expert Systems with Applications 38 (2011) 12018-12027.
[6] X. Chen, I. Bose, Assessing the severity of phishing attacks: A hybrid data mining approach, Decision Support Systems 50 (2011) 662-672.
[7] H.Wang, B.Zhu, C.WANG, A Method of Detecting Phishing Web Pages Based on Feature Vectors Matching, Journal of Information and Computational Systems 2012 Vol. 9 (15): 4229-4235.
[8] W.Zhuang, Q. Jiang, Intelligent Anti-phishing Framework Using Multiple Classifiers Combination, Journal of Computational Information Systems 2012 Vol. 8 (17): 7267-7281.
[9] Y. Zhang, J.Hong, CANTINA: A content-based approach to detecting phishing web sites, in: Proc. the international World Wide Web conference(WWW), 2007, pp. 639-648.
[10] D. K. McGrath,M. Gupta, Behind Phishing: An Examination of Phisher Modi Operandi, in: Proc. the USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET), 2008, pp. 1123-1136.
[11] J. Ma, L. K.Saul, S.Savage, Identifying Suspicious URLs: An Application of Large-scale Online Learning, in: Proc. of International Conference on MLP, 2009, .