International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 3 Issue 5, May 2014

An Efficient Approach To Detecting Phishing A Web Using K-Means And Naïve-Bayes Algorithms With Results.

Ms. Nilam Badadhe [1], Ms. Sneha More [2], Mrs. N. V. Puri [3]
Computer Department[1, 2, 3], Universal College Of Engineering and Research, Pune[1, 2, 3]

Abstract- Since last Decade, Phishing is becoming one of the serious crime in the network. Phishing is a type of attack that results in online theft. In phishing, a combination of social engineering and web site spoofing techniques track a user into revealing confidential information with economic value. Over the past few years, different methods are designed for the detection of phishing web by using known as well as new features. In previous approach, Phishing sites are detected by using blacklist-based approach. The drawback of this approach is that non-blacklisted phishing sites cannot be detected using this approach. This paper presents an efficient approach for detection of phishing a web based on features of legitimate and phishing webs. In this paper, we have proposed two algorithms that are K-Means and Naïve Bayes. Using these algorithms, we have checked blacklist for detection of phishing sites as well as their behavior.

Index Terms- Phishing; K-means; Naïve Bayes; Spoofing; Blacklist-based approach.

I. INTRODUCTION

Now a days, there are lots of people using the internet services. These people are also known as 'Netizens'. These internet services are convenient to all Netizens and also beneficial because it saves MTE i.e., Money Time and Efforts. Unfortunately, the benefits of online services has been overshadowed by large-scale phishing attacks rised against the Netizens. Phishing is the type of identity theft in which attackers try to acquire personal information and financial credentials of online consumers for doing some illegal transactions or other illegal activities.

There are 4 phases in typical phishing attack such as preparation, mass broadcast, mature and account hijack. Phishing is a form of identity theft that occurs when a malicious website impersonates a legitimate one in order to access confidential information such as passwords, account details or credit card numbers.

Phishing is a fraud technique to collect confidential information by masquerading as a trustworthy person or business in an electronic communications

According to the antiphishing working group, there were 18480 unique websites. Unique phishing attacks and 9666 unique phishing sites reported in march 2006. The total no of phishing attacks launched in 2012 was 59% higher than 2011. It appears that phishing has been able to set another record year in attack volumes, with global losses from phishing estimated at $1.5 billion in 2012. This represents a 22% increase from 2011

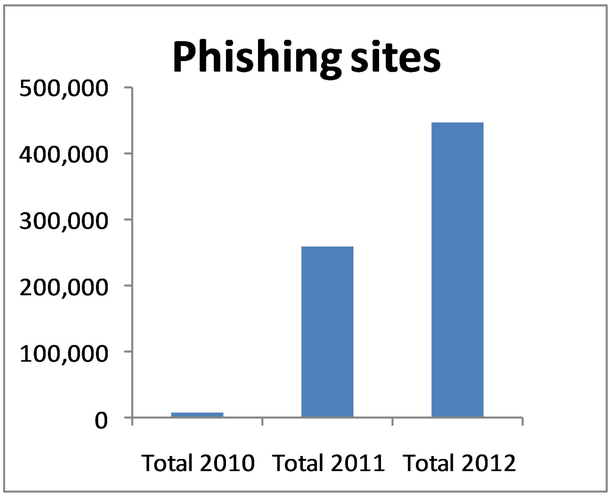


Figure 1. Unique phishing sites for three years

II. SYSTEM ARCHITECTURE

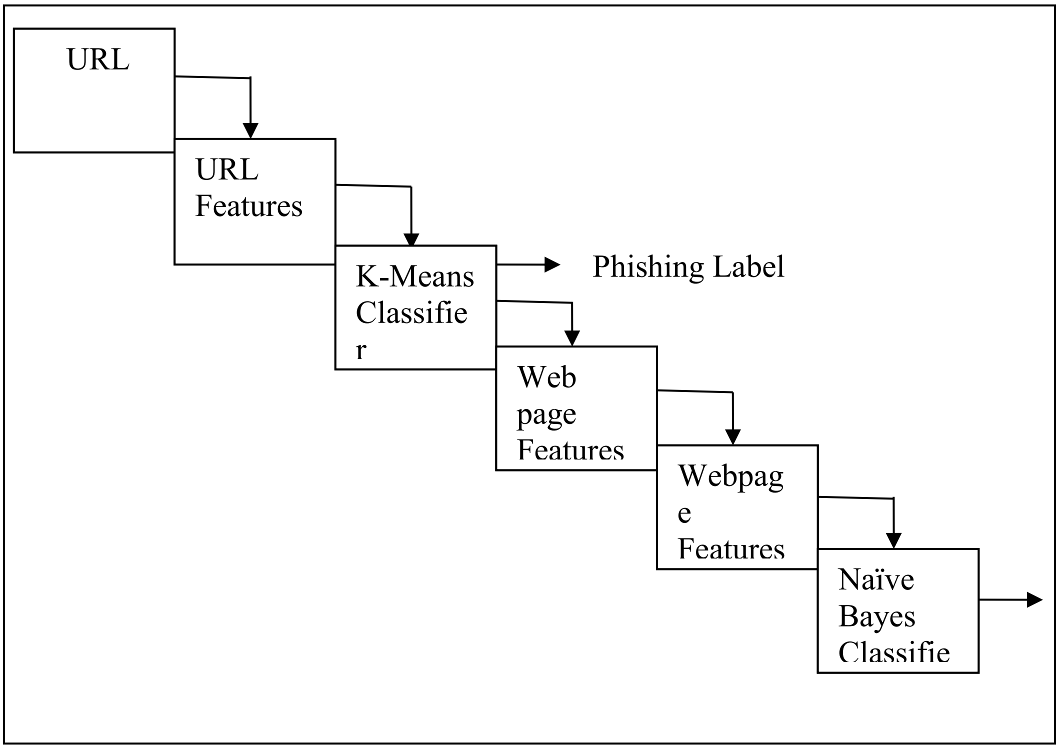


Figure 2: System Architecture

The phishing website attackers tricks victim by developing different types of engineering solutions such as scaring to stop victim or victim's account if they don't update their account details or due to any other purposes to attract the victims by visiting their phishing websites.

We have developed two algorithms for the detection of phishing websites using existing as well as new features. We have generated some features that can help to detect the phishing attacks with little amount of prior knowledge about the algorithms which we have used for discovering the phishing attacks. The system which we have developed can detect the phishing sites but it can't prevent the access to those phishing sites. This paper provides an efficient approach for the detection of the phishing websites.

This paper is classified as follows:
In section 3 we are presenting modules and data used. In section 4 we have generated experimental results. In section 5 we enhanced the future scope and conclusion remarks.

III. MODULES

We have developed two main modules for detecting the phishing webpages. The motive behind using these modules is to directly extract data and features from the URL. The modules used in our approach are described below:

A. Extraction of Features

Feature extraction plays an important role in the efficient detection of the phishing websites. In this module, features of the URL and website are described.

- URL's features:

Extraction of the URL with the trained model is less complicated operation as compared to the downloading of the complete webpage and using its data for extraction of features. There are number of URL features that are used to detect the phishing sites. This module shows the total count of the URL features after extraction, then that count of the URL features is given as the input to the K-Means algorithm for clustering of the data. The URL features are as follows:

- IP address:

Identity of the server is achieved through the use of an IP address. A legitimate website generally has the domain name for its verification and phishing sites use some unauthenticated Zombie system for hosting that particular site. For avoiding from the verification or domain registration, the IP address is the only way used for hiding from all identification and verification.

- No. of dots in URL:

Here number of dot appearance in the URL is checked. Phishing web uses fake or duplicate domain name to construct the authorised look of the URL with the help of extra dots put in URL. So the verification of dots in URL is being performed at this stage.

- No. of suspicious characters in URL:

When the phishing web try to trick the victims, the URLs of the phishing web may be modified to the pattern that is hard to check. '@' or '-' signs in suspicious URLs is checked which are symptoms of phishing URL. If '@' or '-' signs are present in the URL then that site is said to be a phishing site.

- No. of slashes in URL:

Generally, there should not be more number of slashes in the URL. More than five slashes in the URL indicates the URL to be a phishing URL.

- K-means Algorithm:

K-means is one of the algorithms which is very useful in solving the well known clustering problem. K-means clustering is used as a method of the cluster analysis in the data mining and statistics. In that k-means is used for partitioning of n observations into k clusters in which each observation belongs to the cluster with the nearest mean. In K-means algorithm mean is recognized It is similar to the expectation-maximization algorithm which is used for mixtures of Gaussians in which, both attempt to find the centers of the natural clusters in the data as well as in an iterative refinement approach employed by both the algorithms.

Description:

Given a set of observations (x1, x2, …, xn), where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k sets (k ≤ n) S = {S 1 , S 2 , …, S k } so as to minimize the within-cluster sum of squares (WCSS):

$$J = \sum_{j-1}^{k} \sum_{i-1}^{x} ||x_i^{(j)} - C_j\, ||^2$$

(1)

Where xi is the mean of points in Si.

The algorithm is represented by following steps:

Step 1: Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.

Step 2: Assign each object to the group that has the closest centroid.

Step 3: When all objects have been assigned, recalculate the positions of the K centroids.

Step 4: Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

In our approach, after applying the K-Means algorithm on URL's features, the predictor predicts one of the three states. Those three states are 'yes', 'no', 'may be' Predictor is used for showing the prediction states. It is behavioural response to the phishing risk.

- If predictor predicts 'Yes' then the site is phishing.
- No is for the non-phishing web.
- May be is for suspicious webs.

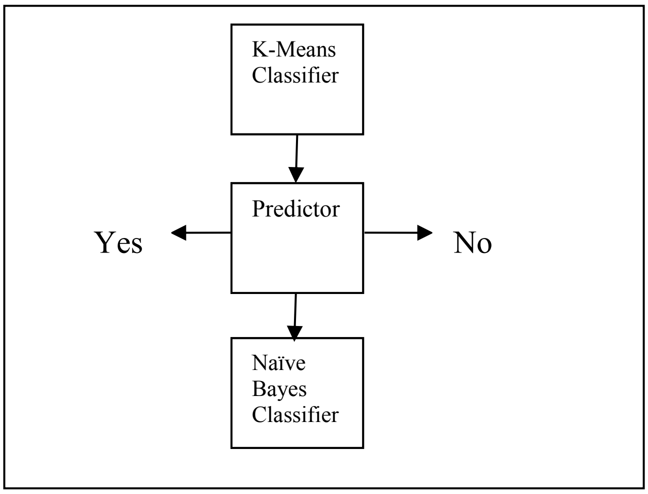


Figure 3: Working of K-Means

If the result is 'may be', then naive bayes algorithm will be applied on complete website to detect phishing.

- Website Features:

Given the suspicious web T and its identity terms would determine the feature value of that webpage. First of all, the whole suspicious webpage is downloaded then the features of that webpage are extracted. After that the webpage is parsed into Document Object Model (DOM) tree. Document Object Model (DOM) is a cross-platform and language independent convention for representing and interacting with objects in HTML, XHTML, XML documents. After parsing, feature vector generated would be passed to the Naïve Bayes Classifier to detect whether webpage is phishing or non-phishing. The website features are discussed below-

- Forms:

If a webpage consists of any HTML text entry form asking for personal or confidential data from users such as account number, password and credit card number. Most phishing sites contain such forms for acquiring personal information from users. Then this is risky for users.

- Nil Anchors:

A nil anchor is an anchor that doesn't point anywhere. If a webpage has more number of nil anchors, then that webpage becomes more suspicious.

- Foreign Anchors:

Foreign anchor is the anchor which contains the link which is present on another page. This anchor tag contains href attribute. This attribute has value is on URL to which the page is linked with domain name in URL of pages are not similar in foreign anchors in webpage, but too many foreign anchors increase the doubt on suspicious web.

- Foreign Requests:

As like foreign anchor, when there are many foreign requests present in webpage then that web could be suspicious or less credible.

- Naïve Bayes algorithm:

The Naïve Bayes Classifier technique is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods. Naïve Bayes model identifies the characteristics of patients with heart disease. It shows the probability of each input attribute for the predictable state.

- Bayes Rule:

A conditional probability is the likelihood of some conclusion, C, given some evidence/observation, D, where a dependence relationship exists between C and D.

This probability is denoted as $P(C|D)$ where

$$P(D/C)=[P(D/C)P(C)] /[P(D)]$$

- NB Classifier:

Naïve Bayes classifier is one of the high detection approach for learning classification of text documents. Given a set of classified training samples, an application can learn from these samples, so as to predict the class of an unmet samples.

The features (n1, n2, n3, n4) which are present in URL are independent from each other. Every feature ni($1<=i<=4$) text binary value showing whether the particular property comes in URL. The probability is calculated that the given web belongs to a class r(r1: Non-phishing and r2: Phishing) as follows:

$$P(r1/N)= (P(r1)*P(N/ri))/P(N)$$

Where all of P(N) are constant meanwhile $P(ni|r1)$ and P(ri) can be easily calculated from training. The proportional to $P(r1|N)$, $P(r2|N)$ is calculated and the results are as follows:

$P(r1|N)P(r2|N) > b$ ($b>1$), Non-phishing Web.

$P(r2|N)P(r1|N) > b$, Phishing Web

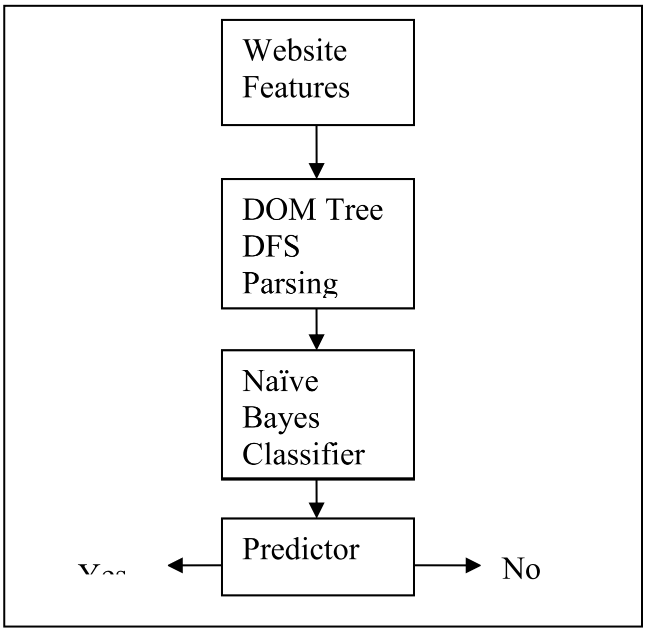
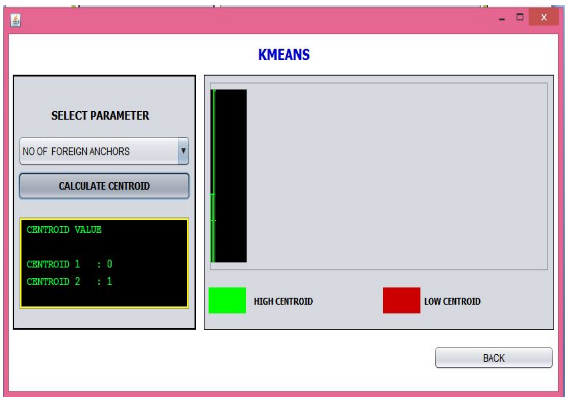


Figure 4:Working of Naive-Bayes

IV EXPERIMENT RESULTS

The aim of our experiment is to demonstrate the effectiveness of our approach. We used K-Means for clustering the URL features.So the clustering result is as follows:



```
CENTROID VALUE

CENTROID 1   : 0
CENTROID 2   : 1
```

Figure 5:Clustering K-Means

We tested the datasets with legitimate websites and phishing websites. All datasets which are used for learning are collected from phishtank.com. This is known site which collects real phishing webs. The dataset with 200 legitimate webs and 400 phishing webs are used for implementation of experiment. For implementation of experiment training set and testing set is required. Collected website search results are training set and rest of all are testing set. Here training set are taken by100 legitimate and 100 phishing webs and the rest of 100 legitimates and 300 phishing webs are taken for tasting sets.

For classification, we assigned legitimate webs with positive answers(P) and phishing webs with negative answers (N). True positive and false positive can be summarized below:

- True positive (TP) – The legitimate websites are correctly classed as legitimate websites.
- False positive (FP) – The legitimate websites are incorrectly classed as phishing websites.

Using this the accuracy of three classifiers are compared. – K-Means, Naive Bayes and our approach.

Table 1: Performance Comparison

	K-Means	Naive Bayes	Our Approach
Training set time	30s	55s	50s
Testing set time	50s	101s	80s
TP(%)	90.78	95.05	97.08
FP(%)	6.21	2.98	1.15

It shows, True positive and false positive rate with training and testing set time for each classifier. Based on this comparison, our approach has lower false positive rate than other approach. The performance comparison shows that Naive Bayes has high accuracy.

V.FUTURE SCOPE

As a future work, we plan to use different efficient algorithms to compare accurate rates. We also plan to adjust existing and known features extraction methods and use more relevant features that produces better result. After applying both algorithms if web's legality is still suspicious, then we can apply fast and high detection algorithms. And we will plan for implementing anti-phishing tools for different environments.

VI. CONCLUSION

Phishing is becoming major crime in the network. In this project, an efficient approach is proposed to identifying the potential phishing target of a given web. Every web claims a webpage identity, either legal or illigals. If a web claims a fake identity, abnormality may exist in a network space; therefore our approach could detect and differentiate between a legitimate and a phishing web. Our approach first extracts the URL features and then test whether the page is phishing or not using K-Means algorithm. When the web's legality is still suspicious, then categorize its webpage features and test whether the page is phishing or not using Naive Bayes algorithm.The experimental results show that our approach is very efficient than other approach. And it has high accuracy, high detection rate and low false positive rate.

ACKNOWLEDGEMENT

This work was supported by Universal College of Engineering and Research, pune under grants Prof. N.V.Puri and Prof. A.D.Londhe for building grant is great fully acknowledged.

REFERENCES

[1] Angelo P. E. Rosiello∗ , Engin Kirda‡, Christopher Kruegel‡, and Fabrizio Ferrandi Politecnico di Milano angelo@rosiello.org,ferrandi@elet.polimi.it ‡Secure Systems Lab, Technical University Vienna {ek,chris}@seclab.tuwien.ac.at A Layout-Similarity-Based Approach for Detecting Phishing Pages.

[2] F. Schneider, N. Provos, R. Moll, M. Chew, and B. Rakowski. Phishing Protection Design Doc- umentation. http://wiki.mozilla.org/Phishing_ Protection:_Design_Documentation, 2007.

[3] Gartner Press Release. Gartner Says Number of Phishing E-Mails Sent to U.S. Adults eearly Doubles in Just Two Years . http://www.gartner.com/it/ page.jsp?id=498245, 2006.

[4] Google Inc, Google safe browsing for Firefox, http://www.google.com/tools/firefox/ safebrowsing/

[5] Journal of Computational Information Systems 9: 14 (2013) 5553–5560 Available at http://www.Jofcis.com Xiaoqing GU, Hongyuan WANG∗ , Tongguang NI, An Efficient Approach to Detecting Phishing Web.

[6] Microsoft. Sender ID Home Page. http://www. microsoft.com/mscorp/safety/technologies/ senderid/default.ms%px, 2007.

[7] NetCraft. Netcraft anti-phishing tool bar. http:// toolbar.netcraft.com, 2007.

[8] PhishtankInc, phishing dataset, http://www.phishtank.com/.

[9] R. Dhamija and J. D. Tygar. The battle against phishing: Dynamic security skins. In Proceedings of the 2005 symposium on Usable privacy and security, New York, NY, pages 77–88. ACM Press, 2005.

[10] SpoofGuard. Client-side defense against web- based identity theft. http://crypto.stanford.edu/ SpoofGuard/, 2005.

First Author : Nilam Ashok Badadhe, BE Computer (2013-14), Universal College Of Engineering.

Second Author: Sneha Kisan More, BE Computer (2013-14), Universal College Of Engineering.

Third Author: Prof. N.V.Puri, M-Tech (IT), Bharati Vidyapeeth College Of Engineering.