A Framework for Multi Features based Phishing Information Identification using NB and SVM Approach

Jyoti S. Kharat
Professor
JSPM NTC
RSSOER, Narhe,
Pune

Snehal S. Shinde
Professor
JSPM NTC
RSSOER, Narhe
Pune

Anjali P. Deore
Professor
Sandip Institute of Technology &
Research Center Nashik

ABSTRACT

Criminal organizations around the world use the technique known as phishing to extract information from innocent citizens in order to access their bank details, to steal identities, to launder money and more. There are Different types of statistical learning based classification methods are available to differentiate the phishing webpage's from the original. Feature extraction method is the concept, which has been implementing into the development of web phishing information detection technique. Naïve Bayes and SVM statistical algorithms are used for feature extraction of URL and source code respectively. In contrast to other proposals, this scheme has a high detection rate and a low false negative rate as well as can achieve high detection accuracy, the lower detection time and performance with the small sample of a classification model training set.

General Terms

Phishing, SVM, Stemming, Web page classification, Feature fusion strategy, Naïve Bayes,URL.

Keywords

SVM, Phisher, Naïve Bayes, Multi features

1. INTRODUCTION

Automatic detection of phishing web pages has attracted much attention from security and financial institutions, software providers, to academic researchers [1]. Basic Methods for detecting phishing web pages can be classified into user-interface-base data-phishing, industrial toolbar based anti-phishing, and web page content-based anti-phishing. Typical phishing attacks leverage four common weapons: a database of email addresses, bulk email capabilities, a phishing email used to lure victims into the scam, and the phishing Web site used to collect personal information.
The flow of information illustrated in Fig 1, below, is the same for all phishing attacks.

In general, phishing attacks are performed with the following four steps:
1. A fake web site which looks exactly like the legitimate Web site is set up by phisher.

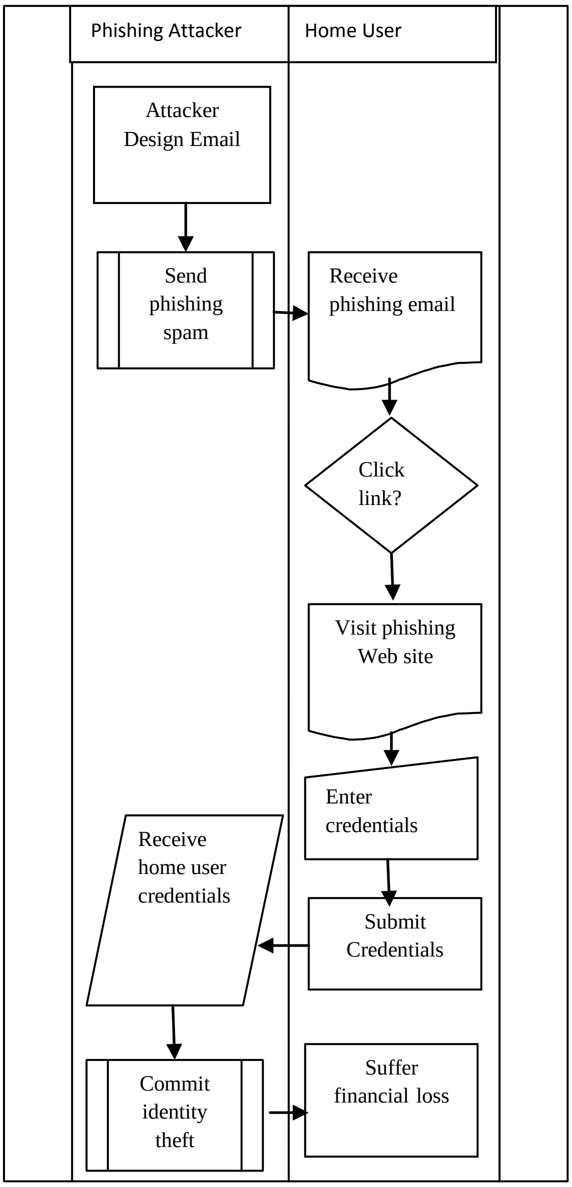

Fig 1: Information Flow of a Typical Phishing Attack

2. Phisher then send link to the fake web site in large amount of spoofed e-mails to target users in the name of legitimate companies and organizations, trying to convince the potential victims to visit their web sites.
3. Victim visits the fake web site by clicking on the link and input its useful information there, *e*.g. Password and Bank

details.
4. Phisher then steal the personal information and perform their fraud such as transferring money from the victims' account.[2]

This system work is based on machine learning approaches to determine the relevant features to extract from phishing web pages, and data mining techniques to determine hidden patterns associated to the relationship between the extracted features. There are two main types of machine learning algorithms.

1. **Supervised Learning Algorithms**

The algorithms are given labeled examples for the various types of data that need to be learned.

2. **Unsupervised Learning Algorithms**

Data is unlabeled and the algorithms attempt to find patterns within the data or to cluster the data into groups or sets. Proposed system is basically using SVM (Support Vector Machine) and NB (Naive Bayesian) classifier which are belong to supervised learning approach for classification of feature.

2. RELATED WORK

Different anti-phishing technique has been implemented such as Toolbar and Web Wallet are belonging to automatic anti-phishing technique.Binay Kumar et al.[3] proposed DC Scanner is an email scanner which identifies malicious URLs received in the email message opened by users. This work digs html contents of emails and web pages referred. Also domains and domain related authority details of these links,script codes associated to web pages are analyzed to conclude for the probability of phishing attacks. This work in two phases: first to get html contents of every links of the email body and mine the contents to verify domains of every links. For detailed verification of URLs they use domain registration information and compare it with the domain authority unique infonnation displayed in the web pages. The second phase checks for malicious URLs in the script codes of the web pages. Also it has been checked whether the phishers have tried to modify the html tags and their attributes so that they could not be traced. For bringing efficiency and correctness they have suggested some standards for web pages' design.

Vimal Kumar and Rakesh Kumar [4] approach in ad hoc environment, which is based on visual cryptography. According to this approach a user generates two shares of an image using (2, 2) visual cryptography scheme. Client stores the first share of this image and second share is uploaded to the website at the time of user registration. After this, website asks for some other information like second share of the image, user name, and password. These credentials of a particular user can change once per login. During each login phase, a user verifies the legitimacy of a website by getting secret information with the help of stacking both shares. There are many existing approaches based on cryptographic technique but they all suffer from False Positive notification. However, proposed approach does not suffer from False Positive (FP) notification and outperforms all existing approaches.

Luong Anh Tuan Nguyen et al.[5] proposed a new phishing detection approach based on the features of URL. Specifically, the proposed method focuses on the similarity of phishing site’s URL and legitimate site’s URL. In the proposed technique, the system model is built to detect phishing sites by using six heuristics (primarydomain, subdomain, pathdomain,pagerank,alexarank, alexareputation). In addition, the ranking of site is also considered as an important factor to decide whether the site is a phishing site. The proposed technique is evaluated with a dataset of 11,660 phishing sites and 5,000 legitimate sites. The results show that the technique can detect over 97% phishing sites.

M. Aburrous et al. [6] states that Fuzzy Data Mining Techniques can be an effective tool in assessing and identifying e-banking phishing websites since it offers a more natural way of dealing with quality factors rather than exact values. Gaston L’Huillier et al. [7] proposed Latent semantic Analysis and text mining methodology for characterization of such strategies, and further classification using supervised learning algorithm. All feature set were evaluated using SVM’s, Naïve Bayes, and logistic regression classification algorithm [8]. Heidy M. Marin-Castro et al. [9] is used Web Query Interfaces (WQIs) which allow to access databases in the Web and retrieve information that is not reachable by traditional search engines. Eric Medvet [10] presents a novel technique to visually compare a suspected phishing page with the legitimate one. The goal is to determine whether the two pages are suspiciously similar. No false positives were raised and only two phishing attempts (that actually did not resemble the legitimate web page) were not detected. Li Yue, et al. [11] proposed DOM-Based block text identification method to improve navigation pages detection. According to the purpose, web pages can be classified into navigation page and content page. Rudy AG., et al. [12] reviews the implementation of web data extraction and stages in making a Mashup. In this paper proposed system implements web data extraction by visually extract targeted data from data sources (web pages). Afterward, system combines web data extraction with the stages of making a Mashup, e.g. data retrieval, data source modeling, data cleaning/ filtering, data integration and data visualization [13]. Suzhi Zhang, et al [14] proposes the design and implementation of a kind of the Web Wrapper which based on pre-defined schema. Wrapper is a program that is able to provide software applications with a structured view of a semistructured Web source [15]. Jingqi Wang, et al. [16] proposes a method to use HTML (Hyper Text Markup Language) line break tags to identify basic semantic units with granularity between those of nodes in DOM tree and content blocks for web page content segmentation. A basic semantic unit (or “BSU” for abbr.) is the one piece of continuous text separated by html line break tags. Nodes in a BSU have closer semantic relevance. Paragraph information in the text content is also maintained. Ram Basnet, et al. [17] applied different methods for detecting phishing emails using known as well as new features. This paper provides insights into the effectiveness of using different machine learning algorithms for the purpose of classification of phishing emails.

3. PROPOSED SYSTEM

The specific objectives of the research have been to:

- Investigate Different anti-phishing Technique.
- Design an effective anti-phishing Technique.
- Implementation of proposed anti-phishing Technique.
- Validation of anti-phishing Technique.

It is observed that phishing is a major security threat to the online community. Normally phishing leads to financial loss. The perfect secure system, three main anti-phishing techniques has been used;

- Black listing and white listing
- Network and encryption based countermeasures
- Content based filtering.

This system concentrates on the content-based anti-phishing, which is referred to as using the features of web pages, consists of surface level characteristics, textual content, and visual content.

The system architecture is shown in Fig. 2. Proposed system is performing in the following procedures

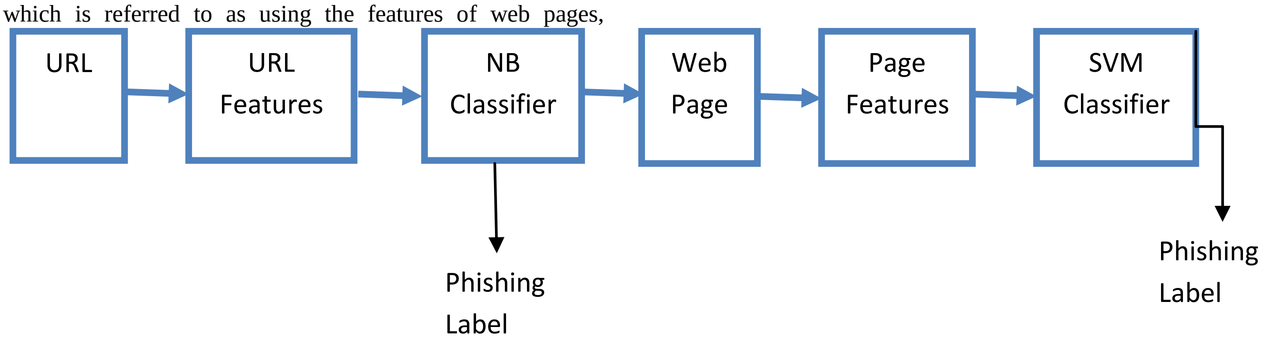

Fig 2: System Architecture

Working of proposed architecture has been described in some steps which are explained as follow.

Step 1: Given a webpage P, extract its URL identity and generate features.

Step 2: Classify P by NB classifier and return result (0, 1, -1).

//0: legitimate, 1: phishing, -1: suspicious

Step 3: If result=0 or 1, output the phishing label,

If result= -1, go to Step 4.

Step 4: If P has not a text input, output the phishing label (0).

If P has a text input, go to Step 5.

Step 5: Extract its webpage identity and generate features.
Step 6: Classify P by SVM classifier and output the phishing label.

3.1 URL Features

Feature extraction plays an eminent role for the efficient prediction of phishing web. The features are described as the following:

- IP Address: For escaping from domain registration or user checking, 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 or use a redirect script to bring the victim to another site. Here the number of dots in a page's URL is checked.
- Suspicious 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 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.

The specific objectives of the research have been to:

- Investigate Different anti-phishing Technique.
- Design an effective anti-phishing Technique.
- Implementation of proposed anti-phishing Technique.
- Validation of anti-phishing Technique.

It is observed that phishing is a major security threat to the online community. Normally phishing leads to financial loss. The perfect secure system, three main anti-phishing techniques has been used;

- Black listing and white listing
- Network and encryption based countermeasures
- Content based filtering.

This system concentrates on the content-based anti-phishing, which is referred to as using the features of web pages, consists of surface level characteristics, textual content, and visual content.

3.2 NB Classifier

Naive Bayes or Bayes' Rule [18] is the basis for many machine-learning and data mining methods. The rule (algorithm) is used to create models with predictive capabilities. It provides new ways of exploring and understanding data. Why preferred naive bayes implementation:

- When the data is high.
- When the attributes are independent of each other.
- When user want more efficient output, as compared to other methods output.

The features described are used to encode webs' URLs as high dimensional feature vectors. The NB classifier is considered one of the most effective approaches for learning how to classify 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 sample. Each URL is represented by features set(x1, x2, x3, x4) are independent from each other. Each feature xi($1 \leq i \leq 4$) takes a binary value(0 or 1) indicating whether the corresponding property appears in the URL. The probability is calculated that the given web belongs to a class c(c1:legitimate and c2:phishing) as follows:

$$p(C_i|X) = \frac{p(C_i) \times p(X|C_i)}{p(X)}$$

$$= \frac{p(C_i) \times \prod_{\mathrm{i}=1}^{4} (\mathrm{X_i}|\mathrm{C_i})}{\mathrm{p(X)}} \qquad (1.1)$$

where all of p(X) are constant, meanwhile $P(x_i|c1)$ and $P(C_i)$ can be calculated easily from training. The proportional to $\frac{P(C1jX)}{P(C2jX)}$ is calculated, and the results are as follows:

$\frac{P(C_1|X)}{P(C_2|X)} > \alpha(\alpha > 1)$, α legitimate web.

$\frac{P(C_2|X)}{P(C_1|X)} > \alpha$, α Phishing web.

$1/\alpha \leq \frac{P(C_1|X)}{P(C_2|X)} \leq \alpha$, α suspicious web , need to be detected further.

3.3 Web Page 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 gathered for web's content as follows:

- **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 URL is not similar to the domain in page URL then it is called as foreign anchor. For any web, it is normal to link to the foreign domains, but too many foreign anchors would decrease the credibility of the web.
- **HTTP/HTTPS:** Web browsers such as Internet Explorer and Firefox display a padlock icon to indicate that the website is secure, as it also displays https:// in the address bar. When a user connects to a website via HTTPS, the website encrypts the session with a Digital Certificate. A user can tell if they are connected to a secure website if the website URL begins with https:// instead of http://. Here this is the strong parameter which is being checked to identify phishing web page.
- **Visual Layout:** Visual layout is one of the parameter for classification. User should observe the visual layout of webpage and supply manual input according to accuracy of web page design.

3.4 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 4- dimension feature vector produced from the feature generation step $V = (< F1, F2, F3, F4 >)$ 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

There are many linear classifiers (hyper planes) that separate the data. However only one of these achieves maximum separation. The reason we need it is because if we use a hyper plane to classify, it might end up closer to one set of datasets compared to others and we do not want this to happen and thus we see that the concept of maximum margin classifier or hyper plane as an apparent solution.

4. MATHEMATICAL MODEL

There are different types of terminologies used to describe the implementation of proposed system such as 'S' is set of functions and variables. The flow of a system divided into two phases, one is 'Training Data' and second is 'Testing Data'. In training phase, system extract the features 'fs' of webpage 'P' and apply Naïve bayes NB_C to check threshold condition of URL. The Naïve bayes algorithm classifies webpage into phishing, non-phishing or suspicious. If webpage is suspicious then system extracts the feature set of source code and applies SVM_C and webpage is being saved with label phishing or Legitimate. In this way, user can create dataset of phishing and non-phishing webpages. Testing phase includes common steps, only it differs from training at last step. Instead of labeling it automatically declares that webpage is either phishing or legitimate.

```
S= {P, DE, VDE, FS, DS, $NB_C$, $SVM_{C,}$ R}
While Training Data:
Open webpage, P1, P2ϵ P
fs1=DE (P1)
fs2=VDE (P1)
Where DE: Data Extractor, VDE: Visual Data Extractor
Extract Feature set, FS=fs1 U fs2
Design Data Set DS,
To Test Data:
Apply Naïve bayes ($NB_C$ ) classifier and extract feature of
URL
R=$NB_C$ (URL)
IF R =Not Suspicious then
Label P=Phishing/legitimate
Else
Extract Feature
fs1=DE (P1)
fs2=VDE (P1)
Combine Text and visual Data Feature in FS,
FS=fs1 U fs2
Apply SVM Classifier on web content and
and store in Result R
R=$SVM_C$ (FS, DS)
Declare P =phishing/legitimate.
```

Here the webpage is parsed into the Document object model tree [16]. DOM is a platform and language-neutral interface that will allow programs and scripts to dynamically access and update the content, structure and style of documents. Similar to approach, the (tf-idf) technique is applied to extract term identity set from a webpage.

5. PERFORMANCE EVALUATION

The proposed technique which helps to classify the legitimate webpage from phishing WebPages. This approach applies a threshold point to each feature. Feature set store the count of occurrence of features which should not cross the threshold point. Here are some performance parameters to check threshold point. Those parameters are as follows:

- **FAR**: False Acceptance Rate.

FAR= (Number of false acceptances) / (Number of identification attempts)

• **FRR**: False Rejection Rate
FRR= (Number of false rejections) / (Number of identification attempts)
• **EER**- Equal Error Rate.
The point where FAR equals to FRR.

Since both the FAR and the FRR depend on the threshold, they are strongly related to each other: increasing the FAR will reduce the FRR and vice-versa.

At initial stage, 10-20 samples tested to check threshold point and tested samples classified as phishing or legitimate. Fig 3.1 represents affect of threshold point on classification and that impact can describes using FAR and FRR parameters

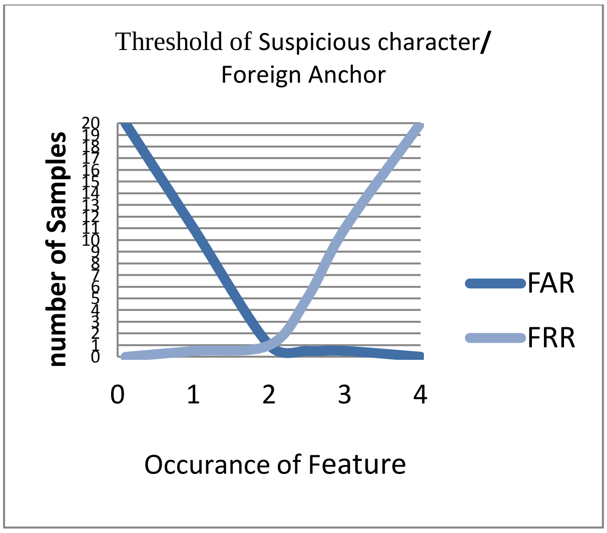

Fig 3: a) Effect of Threshold Point=2 during Classification of Samples (Web Pages)

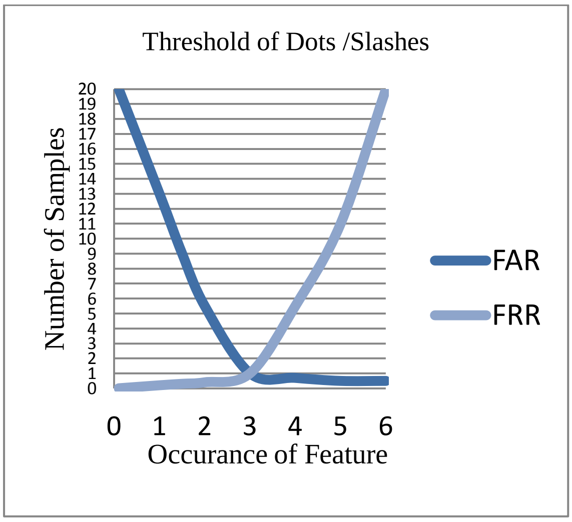

Fig 3: b) Effect of Threshold Point=3 during Classification of Samples (Web Pages)

Fig. 3 (a) shows the threshold point for suspicious character and foreign anchor as 2. It means URL should not contain more than 2 suspicious characters and source code of webpage should not have more than two foreign anchors otherwise input webpage considered as phishing webpage. Fig. 3 (b) shows the threshold point used for Dots and Slashes occur in URL. Threshold point is 3 that mean URL should not contain more than 3 Dots and slashes. If it contains more than three slashes or Dots then the URL is considered to be a phishing URL. Threshold point is 1 applies to Nil Anchor; it means if webpage contains single nil anchor then it is classified as phishing.

The dataset used for learning is collected from PHISHTANK [19]. The dataset with 400 phishing webs and 200 legitimate webs is developed for implementation. 50 legitimate and 50 phishing webs are taken as the training set, and the rest of 150 legitimate and 350 phishing pages compose the testing dataset. The feature vector corresponding to phishing web is assigned a class label 0 and 1 is assigned to legitimate webpage and phishing respectively.

The basic goal behind our approach is to Increase accuracy level for TP and reduce probability of occurrence for FN

- True Positive (TP) - The phish pages which were classed as phish page.
- False Positive (FP) - The legitimate pages which were classed as phish page.
- True Negative (TN) - The legitimate pages which were classed as legitimate page.
- False Negative (FN) - The phish pages which were classed as legitimate page.

6. CONCLUSION

A novel approach is presented to identifying the potential phishing target of a given web. Every web claims a webpage identity, either real or fake. Proposed system divided into two phases; one is ‘Training Phase’ and another one is ‘Testing Phase’. In training phase user can create the dataset as well as manage dataset and in a testing phase, 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. Webpage features include anchor tags to find out valid hyperlink of input webpage. The experimental results show that this approach has a high detection rate and a low false negative rate. This approach is used to find the difference between final probabilities which decides the accuracy level i.e. if the difference of Bayes classifier is high then accuracy automatically increased. This approach can achieve high detection accuracy, the lower detection time and performance with small sample of classification model training set Strong features given to SVM.

7. REFERENCES

[1] Haijun Zhang, Gang Liu, Tommy W. S. Chow, and Wenyin Liu, “Textual and Visual Content-Based Anti-Phishing: Bayesian Approach”, IEEE transactions on Neural networks, Vol. 22, No. 10, October 2011.

[2] Gaurav Mishra et. al, “AntiPhishing Techniques: A Review”, International Journal of Engineering Research and Applications (IJERA), Vol. 2, Issue 2,MarApr 2012, pp.350-355

[3] Binay Kumar, Pankaj Kumar, Ankit Mundra, Shikha Kabra,” DC Scanner: Detecting Phishing Attack”,2015 IEEE.

[4] Vimal Kumar and Rakesh kumar,”Detection of Phishing Attack Using Visual Cryptography in Ad hoc Network”,2015 IEEE.

[5] Luong Anh Tuan Nguyen, Ba Lam To, Huu Khuong Nguyen and Minh Hoang Nguyen,” Novel Approach for Phishing Detection Using URL-Based Heuristic”

[6] M. Aburrous, M.A. Hossain, F. Thabatah, K. Dahal, “Intelligent phishing website detection system using fuzzy techniques”, Information and Communication Technologies: From Theory to Applications, 2008. ICTTA 2008. 3rd International Conference on, pp. 1 – 6, 7-11 April 2008.

[7] Hevia, A. ; Weber, R. ; Ríos, S., “Latent semantic analysis and keyword extraction for phishing classification”, Intelligence and Security Informatics (ISI), 2010 IEEE International Conference on, pp. 129 – 131, 23-26 May 2010.

[8] Saeed Abu-Nimeh, Dario Nappa, Xinlei Wang and Suku Nair, “A comparison of machine learning techniques for phishing detection”, anti-phishing working groups 2nd annual eCrime researchers summit, pp. 60-69, 2007.

[9] Sosa-Sosa, V.J. ; Lopez-Arevalo, I., “A strategy for identification of Web query interfaces using supervised learning”, Next Generation Web Services Practices (NWeSP), 2011 7th International Conference on, pp. 233 – 237, 19-21 Oct. 2011.

[10] Eric Medvet, Engin Kirda, Christopher Kruegel, “Visual-Similarity-Based Phishing Detection”, 4th international conference on Security and privacy in communication networks, Article No. 22, 2008.

[11] Dong Shou-bin ; Zheng Xiang ; Ma Bin-Hua, “Improving Navigation Page Detection by Using DOM-Based Block Text Identification”, ICT and Knowledge Engineering (ICT & Knowledge Engineering), 2012 10th International Conference on, pp. 129 – 134, 21-23 Nov. 2012.

[12] Sari, R.F. ; Budiardjo, B., “Implementing web data extraction and making Mashup with Xtractorz”, Advance Computing Conference (IACC), 2010 IEEE 2nd International, pp. 385 – 393, 19-20 Feb. 2010.

[13] Rattapoom Tuchinda, Pedro Szekely, and Craig A. Knoblock, “Building Mashup by Example”, 13th international conference on Intelligent user interfaces, pp. 139-148 ,2008.

[14] Suzhi Zhang, Peizhong Shi, “An Efficient Wrapper for Web Data Extraction and its Application”, Computer Science & Education, 2009. ICCSE '09. 4th International Conference on, pp. 1245 – 1250, 25-28 July 2009.

[15] Juan Raposo, Alberto Pan, Manuel Alvarez, Justo Hidalgo, “Automatically maintaining wrappers for semi-structured web sources[J]”, Data & Knowledge Engineering.,Vol 61: pp.331-358,2007.

[16] Jingqi Wang, Qingcai Chen, Xiaolong Wang, “Basic Semantic Units Based Web Page Content Extraction”, Systems, Man and Cybernetics, 2008. SMC 2008. IEEE International Conference on, pp. 1489 – 1494, 12-15 Oct. 2008.

[17] Ram Basnet, Srinivas Mukkamala, Andrew H. Sung, “Detection of Phishing Attacks: A Machine Learning Approach”, Book Title-Soft Computing Applications in Industry, Vol 226 pp. 373-383,2008.

[18] Mrs.G.Subbalakshmi, Mr. K. Ramesh, Mr. M. Chinna Rao, “Decision Support in Heart Disease Prediction System using Naive Bayes”, Indian Journal of Computer Science and Engineering (IJCSE), Vol. 2 No. 2 Apr-May 2011.

[19] PhishtankInc,phishingdateset, http://www.phishtank.com/.