Anti-Phishing based on Text Classification using Bayesian Approach

Pankaj H. Gawale
P. G. Student
Department of Computer Engineering,
R. C. Patel Institute of Technology, Shirpur,
Dist. Dhule, Maharastra, India

D. R. Patil
Assistant Professor,
R. C. Patel Institute of Technology, Shirpur,
Dist. Dhule ,Maharastra, India

ABSTRACT
Phishing is an act of cracking by single person or group of persons to stolen the personal confidential information such as credit card detail, bank account detail, passwords etc., from unknown sufferer for illegal activities. In this paper we have implemented the text classifier using Bayesian approach for phishing detection. Text classifier works on textual content for measuring the similarity between the real web page and untrustworthy web page. Stemming is used for simplicity of our model. For generating threshold we used probabilistic approach with large data set of homepage URLs. The experimental result gives phishing pages detection ratio is 98.87% also for FAR is nearly equal to zero.

Keywords
Uniform Resource Locator (URL), Web Pages, Phishing, Text Classifier, Bayesian approach, Correct Classification Ratio (CCR), F-score, Matthews Correlation Coefficient (MCC), False Negative Ratio (FNR), False Alarm Ratio (FAR).

1. INTRODUCTION
Phisher are generating phishing web page which mostly similar to real web page. The most commonly in phishing collects information such as bank account, credit card details, user name, password, social security numbers, mobile numbers, and birthdates by masquerading as honest entity in an electronic communication. Now days e-mail contains much type of links of different web sites. Phishing is carried out by quick messaging or e-mail spoofing and also fake web pages or phishing web sites directs users to enter details which is useful for phishers. A phishing web site is felt and look likes the real or legitimate web site. Phishing is responsible for large amount of personal data loss and money loss [1].

In general phishing attacks are done in following way:

- Firstly phisher set up the fake web site which is identical to legitimate web site.
- Phisher then send links to the phishing web site in the large amount of spoofed e-mails to the target user. Due to that the phisher are trying to convince the victims to visit their fake web sites.
- By clicking on the link the victims visits the fake websites and inputs its confidential information there.
- Phisher then steal the confidential information and use into their fraud such as transferring money from victim's account.

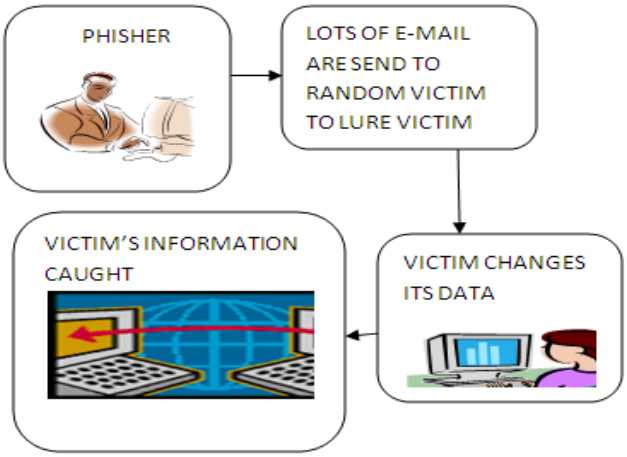

Fig. 1 A General way of phishing attack [1].

Phishers are technically introducing many new ideas and cam affords to invest money in technology. It is common misunderstanding that phishers are amateurs. Phishers can afford investment in technology adequate with illegal benefits gained by their crimes. Phishing is done in many different ways. Some of them are: Deceptive Phishing, Malware Based Phishing, Key loggers and Screen loggers, Session Hijackers Web Trojans Hosts, File Poisoning, System Reconfiguration Attack, Data Theft, DNS-Based Phishing ("Pharming"), Content-Injection Phishing, Man-in-the-Middle Phishing, Search Engine Phishing Software and security provider, financial institution and academic researcher gives the much attention on finding phishing web page. Anti-phishing refers to methods that derived to detect and prevent phishing attacks. It provides security from phishing attacks. There are large amount of work has been done on anti-phishing for deriving the various anti-phishing techniques. From these techniques some works on URL of web sites, some on e-mails, some works on attributes of web sites and some works on content of web sites. In general anti-phishing can classify in some categories such as black listing, symptoms based prevention and domain binding [1].

2. ANTI-PHISHING TECHNIQUES
There are some techniques of anti-phishing as below:

2.1 Attribute based Anti-Phishing Technique
Proactive and reactive anti phishing defenses implemented by attribute based techniques. PhishBouncer tool is used attribute based anti-phishing approach. In this multiple checks are perform such as image attribution check, HTML cross link check. Due to multiple checks the response time is increases.

2.2 Genetic Algorithm based Anti-Phishing Technique

Genetic algorithm is used to detect phishing web pages. Genetic algorithm evolve some simple rules for differentiate normal web sites from phishing web site. These techniques are more complex.

2.3 An Identity based Anti-Phishing Technique

This technique follows the mutual authentication methodology where both user and online entity validates each other's identity during handshakes.

2.4 Character based Anti-Phishing Technique

Hyperlink characteristics are used to detect the phishing link character based technique is used in link Guard tools. This technique is effectively detecting known & unknown attacks. In real time the link guard detected 96% unknown attacks [9].

2.5 Content based Anti-Phishing Technique

In this visual content, textual content and surface level characteristics of web pages are used to find phishing web sites. Content of web page includes the whole information of web page such as image, text, terms, URL, forms, hyperlink & domain name. Spoof guard acquires the characteristics of phishing web page using well known logos, links, URL & domain [1] [2].

We work on content based anti-phishing technique. In this we used textual content for classifying whether the given page is real or phishing one. Our main focus on text present in the web page.

3. RELATED WORK

Fu et al. have developed earth movers distance method to calculate the similarity of the images. In this method they have firstly converted web page in to finding visual similarity. These approach only detecting phishing web page at pixel level not at text level [2].

Zang et al. have proposed effectiveness of anti-phishing toolbars from their study & analysis. They worked on browsers security indicators and two user studies of three security toolbar. They conducted two studies for detecting which attacks are more effecting than others & generate the anti-phishing tool bar [4].

Liu et al. have worked on use of semantic link network to identify phishing web page. They proposed the novel approach to finding the phishing web page by calculating the reasoning on the semantic link network .they firstly find the associated web pages of given web page & the constructing a semantic link network for all web pages [5] .

Zang et al. have proposed a novel content based approach to finding phishing web sites based on TF_IDF information retrieval algorithm CANTINA i.e.Carnegie mellon anti phishing & Network analysis tool takes robust hyperlinks. This method first evaluating TF_IDF of each term then an retrieval algorithm is used in information retrieval to generate a lexical signature provides to a search engine & then matches the domain name of current web page is phishing or not [6].

Likarish et al. have developed B-PAT anti phishing tool bar that helps users to identify phishing websites using Bayesian approach. B-PAT developed to finding the phishing websites by using open source Bayesian filter on the basis of taken which are extracted from document object module analyzer [7].

Liu et al. have developed concept of visual approach to phishing detection which is oriented by document object module based visual similarity. Firstly decomposes the HTML Web pages in to visually differentiable block regions. There are three metrics namely over all style, block level similarity and layout used to evaluate the visual similarity between two web pages [8].

Chandrasekaran et al. have proposed a novel approach to developed classification method based on structural characteristics of phishing e-mails. In this method they used support vector machine for phishing classification. This technique is compared with other widely used machine learning techniques. Identification of phishing e-mail was based on a number of structural features such as domain name, presence of form tag, presence of JavaScript [9].

Zang et al. have developed new content based anti-phishing system using Bayesian approach. New features like text classifier, image classifier and fusion algorithm was developed. For generation of text classifier naive Bayes rule was used and for image classifier earth mover's distance method was used [10].

4. METHODOLOGY

4.1 Web Page Content

Phisher responded by compiling web pages with non HTML component such as images, flash objects Java applets. An important feature of phishing web pages on the basis of visual similarity is red, blue and green values of each pixel. Visual similarity of web pages uses three metrics as layout, block level and style. Most of web browsers implement the concept of security zones, where the security setting of web browser can vary based on the location of web page being viewed.

Surface level content of web page is defined as the characteristics that are used by the users to access to a web page or to connect to other web pages. Features of surface level content consist of the domain name, URL, and hyperlinks which are involved in a given web page.

Textual content in this paper is defined as the terms or words that appear in a given web page, except for the stop words a set of common words like "a," "the," "this," etc. Firstly separate the main text content from HTML tags and apply stemming to each word. Stems are used as basic features instead of original words. For example, "program," "programs," and "programming" are stemmed into "program" and considered as the same word.

Visual content refers to the characteristics with respect to the overall style, the layout, and the block regions including the logos, images, and forms. Visual content also can be further specified to the color of the web page background, the font size, the font style, the locations of images and logos, etc. In addition, the visual content is also user-dependent. On the other hand the web page at the pixel level, i.e., an image that enables the total representation of the visual content of the web page [10].

4.2 Text Classification System

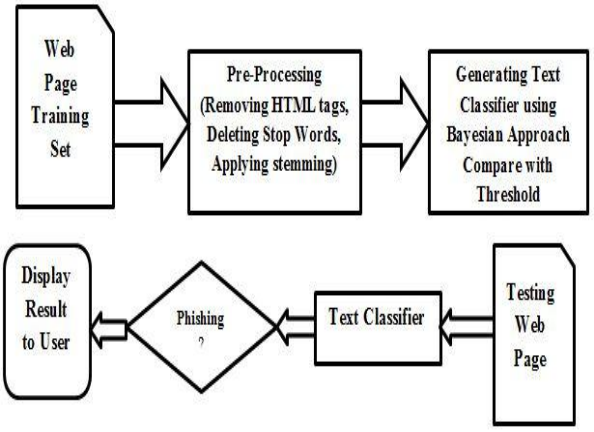

Fig.2 Architecture of system

In preprocessing firstly extracted all the text i.e. words from web page by removing HTML tags. Then we built the vocabulary to form the histogram vectors for each word from given web page. After extracting all words we apply stemming process to each word. Then stem words are used as basic feature for generation of text classifier instead that extracted words. From given web page we have to form a histogram vectors (h1, h2, . . . , hn), where each component represents the term frequency (a term appears in the web page) and n denotes the total number of components in the vector. We consider three points here for generation text classifier.

i. For the simplicity, we do not use any feature extraction algorithms in the process of vocabulary construction.

ii. Do not extract words from all the web pages in a dataset to build the vocabulary, because phishers usually only use the words from a targeted web page to scam unwary users.

iii. Do not take the semantic associations of web pages into account, because the sizes of most phishing web pages are small.

In this paper, we have used the Bayes classifier to classify the text content of web pages. In the classifying process, the Bayes classifier outputs probabilities that a web page belongs to the corresponding categories. These probabilities also can be examined as the similarities or dissimilarities that given web pages have with the protected web page. Let G = {g1, g2, g j, . . . gd } denote the set of web page categories, where d is the total number of categories. For anti-phishing problem only two categories are included: the phishing web page category *g*1 and the normal web page category g2. Given a variable vector (v1, v2, . . . , vn) of a web page, the classifier is employed to determine the probability P(g j |v1, v2, . . . , vn) that the web page belongs to category g j . Applying the Bayes rule, the posterior probability P(g j |v1, v2, . . . , vn) is calculated by equation 1.

$$P(gj \mid v1, v2, ..., vn) = \frac{P(v1, v2, ...vn \mid gj)P(gj)}{P(v1, v2, ...vn)} \quad (1)$$

Let Cj = {c j,1, c j,2, . . . , c j,K j } be the set of training web pages belonging to category g j, where K j is the number of web pages in set C j, and let Hl = (hl,1, hl,2, . . . , hl,n) (l =1, 2, . . . , K j) denote the histogram vector of the l[th] web page in Cj corresponding to the word vocabulary (u1, u2, . . . , un). Conditioning on category g j. Thus, given a testing web page T, the probability P (g j |T) that the web page T belongs to category g j is calculated by equation 2.

$$P(gj \mid T) = \frac{P(gj)\prod_{i=1}^{n} P(ui \mid gj)^{\frac{hiT}{R}}}{\sum_{s=1}^{d} P(gs)\prod_{i-1}^{n} P(ui \mid gs)^{\frac{hiT}{R}}} \quad (2)$$

Where, hi,T represents the frequency of the i[th] word appearing in the web page T, and R is the total number of words extracted from the protected web page. Here, the term R is used to enlarge the value of the term P (ui |g j) hi,T /R such that the denominator of above equation will not be close to zero, because for most phishing cases the phishing web pages include much more term frequencies than the normal web pages. We then compare the probability P (g1|T) of the web page T belonging to the phishing category g1 to a threshold θ_T which is estimated later by using the Bayesian theory. If the probability P (g1|T) exceeds the threshold θ_T , the web page is classified as phishing or normal [10].

4.3 Threshold Generation

Below Formula from [10] is used to find threshold. Where K(s > di, O) and K(s > di, N) denote the numbers of phishing and normal web pages, the similarities of which exceed di , respectively, K(O) and K(N) denote the number of phishing and normal web pages in the training set, respectively, and KT = K(O) + K(N) denotes the total number of web pages in the training set. It is noted that the posterior probability P(O|s > θ) and P(O|s > θ) <= 1. If P (O|s > θ) =1, then P(s > θ |N) = 0, i.e., K(s > di , N) = 0. It indicates that we can select a threshold as large as possible to let K(s > di , N) = 0. But it is noted that K(s > di , O) also decreases when the threshold saturates to 1.

$$\theta = \arg\max_{\hat{d}} \left(\frac{K(s > \hat{d}i, O}{K(s > \hat{d}i, O) + K(s > \hat{d}i, N)} \right)$$
$$\left(\hat{d}i \in \left\{ \arg\max_{di} K(s > di, O) \right\} \right) \quad (3)$$

5. EXPERIMENTAL RESULT

We conduct a large-scale experiment to evaluate the performances of the text classifier. We randomly used 50% data set for training our system and 50% data set used for testing purpose. All the experiment were performed on a PC with Intel(R) Core(TM) i3-2370M CPU @ 2.40Ghz 4Gb RAM. Using 26 keywords as queries, 10 272 homepage URLs. The entire dataset can be downloaded at below link

www.ee.cityu.edu.hk/twschow/Phishing_CityU.rar.

The entire dataset consists of eight sub-datasets corresponding to the real web pages. The web page distribution of the phishing and normal categories for different sub-datasets used in this paper are ebay, PayPal, Rapidshare, HSBC, Yahoo, Alliance-Leicester, Optus, and Steam.

We calculate the result on basis of different classifiers based on five criteria:-

1. Correct Classification Ratio (CCR): It is the ratio of number of correctly classified web pages and total number of web pages,
2. F-score: It is a weighted average of the precision and recall where the score reaches its best value at 1 and worst value at 0.
3. Matthews Correlation Coefficient (MCC): It is a balanced measure that describes the confusion matrix of true/false positives and negatives, such measure can be used even if the classes are of very different sizes,
4. False Negative Ratio (FNR): It is the ratio of number of false negatives and number of phishing web pages.
5. False Alarm Ratio (FAR): It is the ratio of number of false alarms and number of normal web pages.

In Table 1 shows the text classification result using probabilistic Bayesian approach for finding phishing web page. We compare the result of Bayesian text classifier using Bayesian threshold estimation and predefined threshold method. It is simple to set predefined threshold for text classifier using equation (3).

Table 1: Text classification result

Protected Web page	Thr	CCR	F-score	MCC	FNR	FAR
eBay	0.20	0.986	0.914	0.890	130/818	1/4145
PayPal	0.25	0.983	0.981	0.975	33/1275	8/4146
RapidShare	0.10	0.999	0.959	0.952	33/226	1/4146
HSBC	0.10	0.993	0.818	0.911	34/226	0/4145
Yahoo	0.05	0.998	0.518	0.564	66/102	1/4145
Alliance-Leicester	0.05	0.986	0.920	0.856	26/91	0/4146
Optus	0.05	0.972	0.836	0.791	16/50	1/4145
Steam	0.20	0.989	0.827	0.794	47/48	0/4145

In this system we find that the result of threshold changing from 0 to 1. The classification result for different sites is shown in above table, in which 'Thr' best value for predefined threshold. It is clearly observed that the text classifier using Bayesian approach to determine threshold has better performance on CCR than other parameters like F-score, MMC, FNR, FAR. Which indicate that our system is more correctly classify the web page either phishing or original web page. For the "PayPal" sub-dataset the CCR of the text classifier is slightly better, but it delivers larger number of false alarms. For "RapidShare", "HSBC", and "Alliance-Leicester" sub-datasets, we note that the classifier with statistically estimated threshold performs better on the CCR, F-score, MCC, and FNR but at the cost of increasing false alarms.

6. CONCLUSION

In this paper, we have implemented text classifier for phishing web page detection using Bayesian approach. We consider only text of web page. By using text similarity we differentiate the web page into two categories either phishing or original one. The threshold matching used in text classification is effectively estimated by using Bayesian approach. We have worked on large scale experiments in which 10272 homepage URLs and result also shows that this system is efficient of improving the correctness of phishing detection rate up to 98% and also FAR rate decreases. The textual based anti-phishing system for phishing detection can be enhanced by adding more features into the content representations into current system.

7. REFERENCES

[1] A. Emigh. (2005, Oct.). *Online Identity Theft: Phishing Technology, Chokepoints and Countermeasures. Radix Laboratories Inc*., Eau Claire, WI [Online]. Available: http://www.antiphishing.org/phisging- dsh-report.pdf

[2] A. Y. Fu, W. Liu, and X. Deng, "Detecting phishing web pages with visual similarity assessment based on earth mover's distance (EMD)", *IEEE Trans. Depend. Secure Comput.*, vol. 3, no. 4, pp. 301–311, Oct.-Dec. 2006.

[3] N. Chou, R. Ledesma, Y. Teraguchi, and D. Boneh, "Client-side defense against web-based identity theft", in *Proc. 11th Annu. Netw. Distribut. Syst. Secur. Symp.*, San Diego, CA, Feb. 2005, pp. 119–128.

[4] Y. Zhang, S. Egelman, L. Cranor, and J. Hong, "Phinding phish: Evaluating anti-phishing tools", in *Proc. 14th Annu. Netw. Distribut. Syst. Secur. Symp.*, San Diego, CA, Feb. 2007, pp. 1–16.

[5] W. Liu, N. Fang, X. Quan, B. Qiu, and G. Liu, "Discovering phishing target based on semantic link network", *Future Generat. Comput. Syst.*, vol. 26, no. 3, pp. 381–388, Mar. 2010.

[6] Y. Zhang, J. Hong, and L. Cranor, "CANTINA: A content-based approach to detecting phishing web sites", in *Proc. 16th Int. Conf. World Wide Web*, Banff, AB, Canada, May 2007, pp. 639–648.

[7] P. Likarish, E. Jung, D. Dunbar, T. E. Hansen, and J. P. Hourcade, "B-APT: Bayesian anti-phishing toolbar", in *Proc. IEEE Int. Conf. Commun.*, Beijing, China, May 2008, pp. 1745–1749.

[8] W. Liu, X. Deng, G. Huang, and A. Y. Fu, "An antiphishing strategy based on visual similarity assessment", *IEEE Internet Comput.*, vol. 10, no. 2, pp. 58–65, Mar.–Apr. 2006.

[9] M. Chandrasekaran, K. Narayanan, and S. Upadhyaya, "Phishing email detection based on structural properties", in *Proc. 9th Annu. NYS Cyber Secur. Conf.*, New York, Jun. 2006, pp. 2–8.

[10] H. Zang , G. Liu, Tommy W., S. Chow ,"Textual and visual content based anti-phishing : a Bayesian approach", *IEEE Transaction of neural network*, 1532-1446, 2011.