A MULTILAYERED PHISHING WEBSITE DETECTION USING ADABOOST AND SVM

Priyanka.c[1], Chitra.s[2], Sankari.v[3]

[1]priyankac026@gmail.com

Department of IT, MVIT

[2]chitrashanmougam@gmail.com

[3]tosankari@gmail.com

Assistant Professor, Department of IT, MVIT

***Abstract*— Phishing attack is an online method used by scam artists to steal money and personal information. Typically, a "phishing attack" is an e-mail masquerading as a message from a trusted source (bank, Credit Card Company, e-commerce retailer, and so on). The message typically asks you to verify your account information immediately with the threat of a negative consequence if you do not verify the information. Users are often tricked into providing the requested personal information, such as bank or credit card account numbers, social security numbers, passwords, and more. Phishing has become most popular practice among the web criminals .We propose here a robust methodology to detect phishing websites that employs for semantic analysis a topic modeling technique, Latent Dirichlet Allocation, and for classification, AdaBoost. The methodology developed is a content driven approach that is device independent and language neutral. The website content of mobile and desktop clients are collected by employing an intelligent web crawler. The website contents that are not in English are translated to English using Google's language translator. Topic model is built using the translated contents of desktop and mobile clients. The phishing website classifier is built using (i) distribution probabilities for the topics found as features using Latent Dirichlet Allocation and (ii) AdaBoost voting technique.**

***Keywords*—** Adaboost ,Dirichlet, Support Vector Machine ,Tesseract, Phishtank, Fmeasure, Naïve Bayes

I. INTRODUCTION

Many researches has been done towards protecting users from phishing attacks. They include firewalls, black listing certain domains and Internet protocol (IP) addresses, client toolbars, classifiers and user education. Each of these existing techniques has some advantages and some disadvantages. For example, the blacklist approach is harder to maintain with an expanding IP address/domain space. The warnings displayed by phishing toolbars are ignored by the user. Existing phishing website detection classifiers are built using features that are susceptible to technology changes.

For example, a classifier that uses long Uniform Resource Locator (URL) to distinguish a phishing website will fail for websites hosted at URL shortening services. The content classifiers that use term-frequency as features do not account for synonyms, words with similar meanings whose meaning changes according to the context. Moreover, the classifiers were not built for both mobile and desktop clients. According to PayPal, sixty seven percent of consumers are expected to use their mobile device for online purchase. The findings by Trustier conclude that mobile users are three times more vulnerable to phishing attacks than desktop users as mobile devices are always on, users are likely to check messages first on their devices, and devices do not have the same level of protection as desktops. Thus, it is critical for a phishing detection methodology to work not only on desktop clients but also on mobile devices. Furthermore, the past classifier evaluation was limited to English websites.

We propose an intelligent anti-phishing strategy model for phishing website detection and categorization through learning and training samples from large and real daily phishing websites. We first parse and analyze the webpage content and extract 10 different types of features such as title, keywords, description, alt and link text information to represent the webpage. Then we build heterogeneous classifiers according to the characteristics of different features. Finally, an ensemble method is used to combine the prediction results of these heterogeneous classifiers for phishing detection, and a hierarchical clustering algorithm is employed for categorizing the phishing websites. Experiments on real life datasets

demonstrate that our method outperforms existing popular detection methods and commonly used anti-phishing tools in phishing detection.

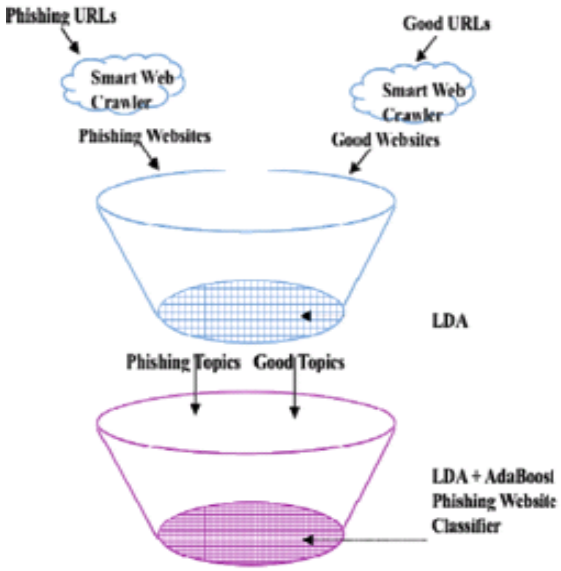

Fig.1.1 Phishing Website Detection Methodology

The main contribution of this research is a content driven phishing detection method that is robust to technology changes, robust to changes in word usage, can be applied to mobile and desktop clients, and is language neutral. The phishing detection method employs the topic modeling technique, Latent Dirichlet Allocation (LDA), to extract features, and classification technique, AdaBoost, to build the classifier. The paper is organized as follows. We first review state of the art phishing website detection technique. The modeling techniques employed, namely LDA and AdaBoost.

1.1 Adaboost

ADABOOST is a classifier ensemble technique that combines predictions of multiple classifiers and produce a single and robust classifier. The prediction result from the combined classifier is usually better than those of individual classifiers.

1.2 Latent-Dirichlet Location

Latent Dirichlet Allocation (LDA) is a natural language processing technique that discovers topics from a collection of documents. Documents are represented as random mixtures over latent topics and each topic is represented by a distribution over words.

II. ARCHITECTURE

The architectural components of the phishing detection methodology are presented in this section. A schematic representation of the architecture is shown.

2.1 Feature Extractor

Firstly, IPDCM uses the feature extractor to extract the terms from the WebPages of the collected phishing websites, and then converts the terms to a group of 32-bit global IDs as the features of the data collection. For training samples, these integer vectors are transformed into term frequency features and stored in the database.

2.2 Classifier Training Module

Ten heterogeneous classifiers are built according to the characteristics of different features, improved NBC(Naïve Bayes Classifier) and SVM(Support Vector Machine) algorithm were employed for the training.

The basic idea of the linear SVM algorithm is find the best classification hyperplane between the two class, which should meet the following conditions:

$$\text{Min}(1/2 \parallel w \parallel^2 + c \sum_{i=1}^{n} i)\, y_i\, (wx_I - b) > 1 - i$$

Where i=1,2,….n,n is the dimensionality of the feature, x is the input vector to the hyperplane. Considering there may be some samples cannot be correctly to reduce the degree of misclassification. We calculate TF-IDF score for each word in the "string" feature like this:

$$TF(Xi) = count(j, X_i)/Count(j)$$

$$DF(Xi) = countfile(Xi)/countfile$$

$$TF\text{-}IDF(Xi) = TF(Xi)/DF(Xi)$$

Where TF(Xi) is the term frequency of Xi in document frequency for Xi. Then we convert all words in "String" feature into TF-IDF score vector for SVM input.

2.3 Ensemble Classification Module

Ensemble classification method is used to combine all the prediction results from heterogeneous classifiers, which has better detection performance than each individual classifier. Internet security experts can look at the partitions and manually generate some website-level constraints.

The major architectural components are URL fetcher, we crawler, parser, language translator,

LDA topic modeler and LDA + AdaBoost classifier. The URL fetcher has access to good website URLs from two public available websites, DMOZ and Alexa. DMOZ maintains a directory of the web organized into several categories. The URL fetcher fetches website URLs in business, internet, banking, and games, as these categories are the most targeted by attackers.

ID	Feature Name	Description
1	Title	Title of the webpage.
2	H1-H6	Content in the <h1> to <h6> tags.
3	Keyword	Keyword information in the Meta tag.
4	Description	Page description in the Meta tag.
5	Copyright	Copyright info in the Meta tag.
6	Link text	Corresponding text of the link.
7	Frame	Url address of the Frame.
8	Img	Url address of the Image.
9	Alt	Description text of the image.
10	String	All the other visible string of the page.

Fig.2.1. Feature Extraction.

In addition to DMOZ, the top 500 websites published by Alexa were also fetched by the URL fetcher. The good website URLs are downloaded by the fetcher once. Phishing website URLs are fetched from phishtank 104 website every hour of the day. Phishtank provides a dump of confirmed phishing URLs that are online at a given instant.

As phishing URLs are short lived, these URLs are fetched periodically. Both good and phishing URLs are stored in a URL database using mySQL. The web crawler fetches the contents of the underlying URLs, both phishing and good ones. Requests from the web crawler are proxy through Tor anonymous network. This prevents a) crawler's IP address from getting blocked by the website, and, b) to capture the actual contents instead of fake content that the attackers' website sometime displays, This type of implementation is a unique and novel development compared to the state-of-the-art research on phishing. The generated HTML pages of the website are rendered using the headless browser (i.e., browser with no user interaction). This ensures capturing rendered web contents instead of the raw HTML which sometimes contain nothing more than references to javascript code in the phishing web sites. The good websites requests are also proxy through Tor for latency measurement.

The rendered website contents are stored in in a different database on the same MYSQL server. Contents are generated and stored for mobile (iPhone, iPAD, Android) and desktop (Windows, Mac) clients. The parser component parses rendered website contents and extracts hyperlinks, text and images. Images are further converted to text using optical character recognition (OCR) tool. The open source Tesseract was used as the OCR tool. The hyperlinks are converted to text by removing all non alphanumeric characters. The combined parsed HTML text, hyperlink text and text from image is stored in website text database.

The language translator converts text that are not in English to English. It uses Google's language detection API to classify the language of the underlying text and calls the translation API for language translation. The translated data is stored in another database for subsequent processing. This is yet another unique and novel development advanced by the research described here. The LDA topic modeler builds the topic model from the translated text contents of both phishing and good websites.

LDA model discovers topics and employs Gibbs sampling for parameter estimation. The Stanford topic modeling toolbox is used to implement the topic modeler component. The term document frequency matrix is built after tokenizing the text into words. The standard stop word filter is applied to the tokenized text. the corpus.

Finally, the classifier is built using LDA topic distributions as input and AdaBoost classification technique. Several weak learners (as detailed in section V) are used to build a robust classifier for phishing website detection. The WEKA open source software is used to build the final Adaboost classifier.

III.CONCLUSION

A multi-layered phishing detection methodology is proposed and evaluated for phishing website detection. The methodology employs a smart web crawler for capturing rendered website content. The methodology captures contents of desktop clients and mobile devices and applies language translation for content that are not in English. The methodology builds a LDA topic model from the rendered website content. The topic model that yields the best generalization performance is then used to build a

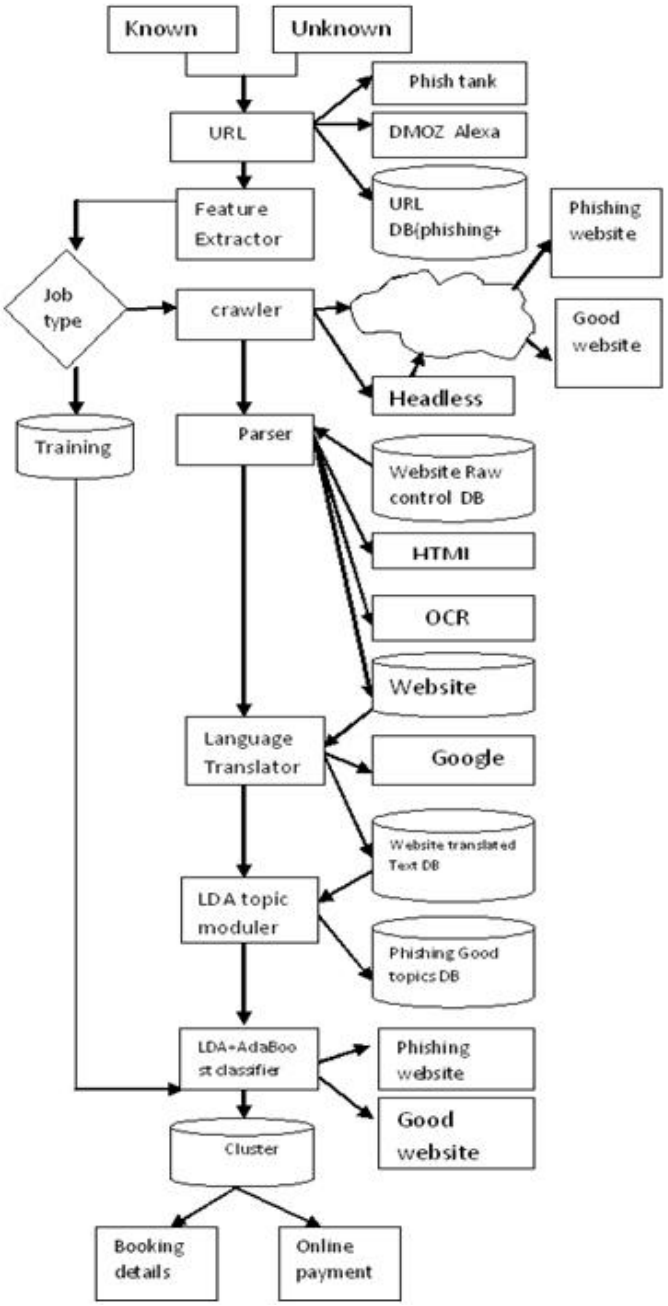

Fig:2.2.phishing website detection architecture

robust classifier using AdaBoost. Experimental results show that language translation lowers the perplexity to 1/10th of non-translated content. The AdaBoost classifier with random forest as the weak learner provides the best classification performance. The true positive rate and Fmeasure obtained with 5% phishing content in the training set yielded were 99% which equaled state-of-the-art research. However, our method was evaluated on a much large corpus, it is device neutral and language independent, and hence a significant new research contribution to phishing website detection.

REFERENCES

[1] Phishing Activity Trends Report, http://www.antiphishing.org/phishReportsArchive.html, Last accessed Feb 20, 2012.

[2]MobileCommerce Daily,http://www.mobilecommercedaily.com/2011/11/28/how-consumerswill-use-their-mobile-devices-during-the-holidays, Last accessed Feb 20,2012.

[3]Trusteer,http://www.trusteer.com/blog/mobile-users-three-times-morevulnerable- phishing-attacks, Last accessed February 20, 2012.

[4] M. Aburrous, M. A. Hossain, K. Dahal and F. Thabtah, "Associative classification techniques for predicting e-Banking phishing websites",International Conference on Multimedia Computing and Information Technology, Sharjah, 2010.

[5] L. Wenyin, G. Liu, B. Qiu and X. Quan, "Anti-phishing by discovering phishing target", IEEE Internet Computing, Issue 99, 2011.

[6] M. He, S. J. Horng, P. Fan, M. K. Khan, R. S. Run, J. L. Lai, R. J. Chen and A. Sutanto, "An efficient phishing webpage detector", Expert Systems with Applications, Elsevier, Vol. 38, 2011.

[7] G. Xiang, J. Hong, C. P. Rose and L. Cranor, "A feature-rich machine learning framework for detecting phishing websites", ACM Transaction on Information Systems Security, Vol. 14, Article 21, September, 2011.

[10] C. H. Hsu, P. Wang and S. Pu, "Identify fixed path phishing attack by STC", Proceedings of 8th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference, Perth, Australia, September 2011.

[11] A P. Felt and D. Wagner, "Phishing on mobile devices," Web 2.0 Security & Privacy 2011, Oakland, Calfornia, May, 2011.

[13] D. M. Blei, A. Y. Ng, M. I. Jordan, "Latent Dirichlet Allocation" Learning Research, Vol. 3, pp. 993-1022, 2003.