Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 10, October 2014, pg.505 – 510

RESEARCH ARTICLE

A NOVEL APPROACH FOR PREDICTING PHISHING WEBSITES USING THE MAPREDUCE FRAMEWORK

Hima Sampath Rao[1], SK Abdul Nabi[2]
M.Tech, Department of CSE, AVNIET, JNTUH, Hyderabad, AP, India
Professor and HOD, Department of CSE, AVNIET, JNTUH, Hyderabad, AP, India

***Abstract:** In this paper, we have proposed a new approach named as " A Novel Approach for Predicting Phishing Websites using Map Reduce Framework " to overcome the difficulty and complexity in detecting and predicting phishing website. We proposed an efficient, resilient and effective approach that is based on using MapReduce framework, classification Data Mining algorithms and cluster methodology. Detecting, Predicting & Identifying phishing websites are a tedious work. Several attributes are needed to be taken into consideration & finally using the data mining algorithms, an efficient and novel approach is defined. A map-reduce concept is involved followed by clustering and data mining algorithm which affects the entire process of detection and prediction of phishing websites to get the most effective and both original and genuine websites also increasing both speed & efficiency of the system. This system is very trustful, which surely guarantees that we will not miss a phishing website, even if it is a newborn.*

***KEYWORDS:** Phishing, MapReduce Framework, Phishing Detection, clustering*

1. INTRODUCTION

Now a day's many phishy websites are created by the attackers to extract all the personal information from the users. As the technology improves our security towards it also should be improved accordingly. Phishing is an attempt by an individual or a group to thieve personal, confidential information such as passwords, credit card information, etc. from unsuspecting victims of identity theft, financial gain and other fraudulent activities. Many people use mobile internet through mobile phones. The proposed anti-phishing tool to detect the phishing websites is very helpful to save the users from many fraud websites. Detecting and identifying phishing websites is a really complex and dynamic problem involving many factors and criteria. Phishing is a fraudulent attempt, usually made through email, to steal your personal information. Phishing web sites are forged websites created by malicious people to mimic real websites.

Generally, however, it may be divided into one of two types of categories: (1) crimes that target computer networks or devices directly and (2) crimes facilitated by computer networks or devices, the primary target of which is independent of the computer network or device. Examples for cybercrimes are fraud, spam, cyber terrorism and phishing. The whole process is lengthy, so relying totally on it might prove dangerous. The solution is to check the website attributes, as many possible, in the real time environment. The idea to use data mining concept is very useful than other applications, because it says mining a data or information from a huge database. Data mining can provide a way how to find these phishing websites with the help of many applications and algorithms. Likewise, in anti-phishing also machine learning is very useful. It is needed to speed up in finding attribute values which can be given by means of MapReduce.

MapReduce using is also having significant application in many ways. Phishing is a type of online fraud in which a scam artist uses an e-mail or website to illicitly obtain confidential information. It is a semantic attack which targets the user rather than the computer.

2. RELATED WORK

Phishing web pages are forged web pages that are created by malicious people to mimic Web pages of real websites. Phishing is a direct attack on the identity of a user, attacker steals the identity of user and impersonate as that victim user. So it is way too different than the virus, malware attacks. It is more of a user's specific attack so security needs to be provided at user level. Most of the websites Most of these kinds of web pages have high visual similarities to scam their victims. Some of these kinds of web pages look exactly like the real ones. Victims of phishing web pages may expose their bank account, password, credit card number, or other important information to the phishing web page owners. It includes techniques such as tricking customers through email and spam messages, man in the middle attacks, installation of key loggers and screen captures.

These popular technologies have several drawbacks:

i. The similarity assessment based technique is time-consuming. It needs too long, time to calculate a pair of pages, so using the method to detect phishing websites on the client terminal is not suitable. And there is a low accuracy rate for this method depends on many factors, such as the text, images, and similarity measurement technique. However, this technique (in particular, image similarity identification technique) is not perfect enough yet.

ii. Blacklist-based technique with low false alarm probability, but it cannot detect the websites that are not in the blacklist database. Because the life cycle of phishing websites is too short and the establishment of blacklist has a long lag time, the accuracy of the blacklist is not too high

iii. Heuristic-based anti-phishing technique, with a high probability of a false and failed alarm, and it is easy for the attacker to use technical means to avoid the heuristic characteristics detection.

Email	Legitimacy	Relavant features of email and sites
NASA	Real	Sender is known person Addressed to user Link in email: “this” URL:antwp.gsfc.nasa.gov/apod/astropix.html
Cognix	Real	Regarding work details Link in email: www.cognix.com URL in status bar: http://www.cognix.com

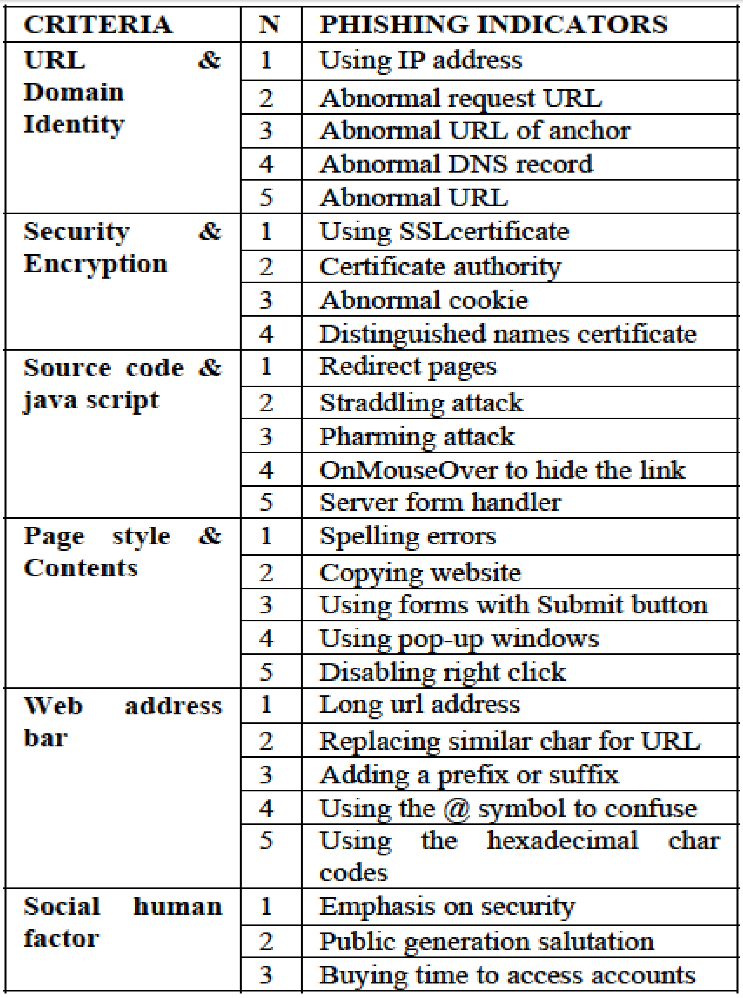
Paypal	Phishing	Urgent request Lock image in body of webpage Link: “Click here to activate your account” URL:http://payaccount.me.uk/cgibin/ webscr.htm?cmd=_login-run
eBay	Real	Registered name “Pat Jones” displayed Link in email” “PAY [Click to confirm…]” URL:http://payments.ebay.com/ws/ eBayISAPI.dll?item=6600378513
Laptop	Spear Phishing	Generic message about eBay item Link: www.set-ltd.net URL: www.set-ltd.net

Many of the phishing websites uses PayPal and eBay for the transactions and shopping’s. The table listed below is the features of five emails and corresponding web sites for email and web role play. The social phishing is fully based on the social network database. It has the public data where it finds the social networks and can collect all the emails to check and to authenticate the web logs. The proposed method also prevents man-in-the middle attacks since the response is obtained from the executable which is called by the browser and third man's interruption is impossible.

Detecting and identifying any phishing website in real-time, particularly for e-banking, is really a complex and dynamic problem involving many factors and criteria. Because of the subjective considerations and the ambiguities involved in the detection, Data Mining (DM) Techniques can be an effective tool in assessing and identifying phishing websites since it offers a more natural way of dealing with quality factors rather than exact values. “Modelling a Phishing Detection System using the MapReduce framework and Data Mining algorithms along with clustering concepts”, a novel approach to overcome the ‘phishing’ propose an intelligent resilient and effective model for detecting phishing websites.

3. PHISHING INDICATORS WITH ITS CRITERIA

Phishing in e-banking is prevalent nowadays. The phishing problem is a hard problem because of the fact that it is very easy for an attacker to create an exact replica of a good banking site, which looks very convincing to users. Phishing is a continual threat that keeps growing to this day. The risk grows even larger in social media such as Facebook, Twitter, Myspace etc. Hackers commonly use these sites to attack persons using these media sites in their workplace, homes, or public in order to take personal and security information that can affect the user and the company (if in a workplace environment). Phishing is used to portray trust in the user since you can usually not tell that the site or program being visited/ used is not real, and when this occurs is when the hacker has the chance to access the personal information such as passwords, usernames, security codes, and credit card numbers among other things.



CRITERIA	N	PHISHING INDICATORS
URL & Domain Identity	1	Using IP address
	2	Abnormal request URL
	3	Abnormal URL of anchor
	4	Abnormal DNS record
	5	Abnormal URL
Security & Encryption	1	Using SSLcertificate
	2	Certificate authority
	3	Abnormal cookie
	4	Distinguished names certificate
Source code & java script	1	Redirect pages
	2	Straddling attack
	3	Pharming attack
	4	OnMouseOver to hide the link
	5	Server form handler
Page style & Contents	1	Spelling errors
	2	Copying website
	3	Using forms with Submit button
	4	Using pop-up windows
	5	Disabling right click
Web address bar	1	Long url address
	2	Replacing similar char for URL
	3	Adding a prefix or suffix
	4	Using the @ symbol to confuse
	5	Using the hexadecimal char codes
Social human factor	1	Emphasis on security
	2	Public generation salutation
	3	Buying time to access accounts

4. SYSTEM OVERVIEW

The proposed Framework has shown below in figure (1). The user will enter the URL of the webpage, she wishes to visit. Using that URL, we will download the source code of the webpage & then decide the values of the attributes. For finding these values we will make use of MapReduce. This will speed up the process of attribute value assignment. Basic word count example of MapReduce is used to search sensitive words in web pages.

In the same way wherever required help of Data Mining algorithms along with clustering concepts are taken. These calculated attributes are the input to the Prediction module. Based on the records stored from phishtank.com database, training data is prepared using feature selection and extraction concepts by data mining algorithms.

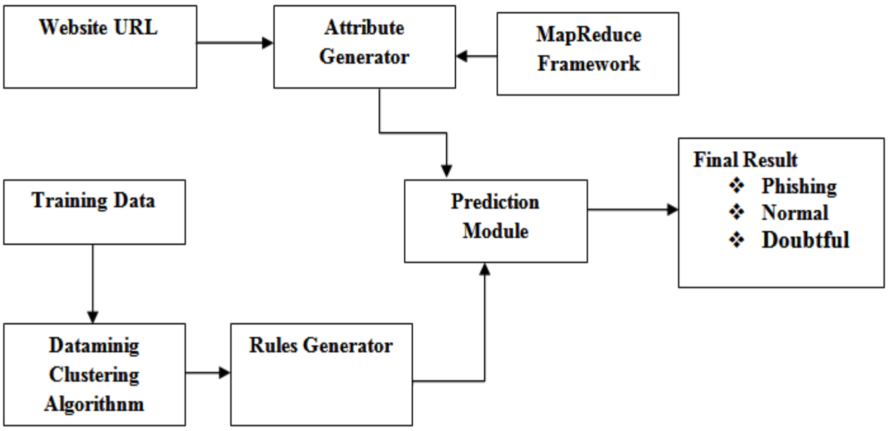


Fig.1 Proposed Framework

All the characteristics of reported phishing website at phishtank.com corpus are studied and based on that attribute are decided and training data for the machine learning algorithm is prepared. Using training data machine learning algorithm generates set of rules based on which decision is to be made. Prediction module gets two input rules generated by machine learning algorithm and attribute found from requesting a URL. Prediction module finally predict URL falls under which category that are Phishing, Legitimate, and Doubtful.

a. Obtain URL

The URL is the website address which the user enters in the address bar. When the mobile handset is connected to the internet, it extracts the URL from the address bar and gives that URL to next addressed links.

b. Attributes Consideration

Initially, all the phishing website details are collected and stored in the phishing website attributes. Then it is sent to a preprocessor to convert into a machine understandable format.

c. Website Phishing Training Data Sets

For our implementation we plan to use two publicly available datasets to test: the “phishtank” from the phishtank.com. The PhishTank database records the URL for the suspected website that has been reported, the time of that report, and sometimes further detail such as the screenshots of the website, and is publicly available.

d. MapReduce Framework

Complex problems such as the one being taken in this report must often be done in multiple MapReduce steps. Each step takes as input the output from a previous step MapReduce. Data Preparation: This data set is a collection of huge amount of files each containing data for a single record. Each of these files contains the record identification number of the record as the first line.

e. Datamining Algorithm-Clustering

As we have proposed new approach, which is An Efficient MapReduce Algorithm using Clustering concept. A new module is introduced in the first step is a Map, which can split the existing data clusters and another module is reduce can merge the intermediate training data which is successful results of map phase or module. Then the training data can be processed in reduce phase can give the efficient features. When map function is predicted, it processed the clustering with key and value pairs and when the reduction is addressed, then the intermediate data can be processed with grouped key-value collection, and merger function can be performed.

For generating the intermediate datasets while finding phishiness of any site we need to simply put the “?” for the last attribute in every layer attribute so that weak tool rule generator will add the predicted value at the place of “?” in dataset given to it, which will be used for the further process in hierarchically next level and so on. At least we get the final result as pushy, Legitimate or Doubtful. Efficiency goes on increasing as the correctly classified instances percentage increases. For that accurate priority based dataset is provided to the rule generator. As the numbers of records in the data set are increasing the correctly classified instances are increased.

5. CONCLUSIONS

The prediction of phishing websites is essential and this can be done using neural networks. The main goal of the system is to achieve speeds up with existing anti-phishing system by using a MapReduce approach. For the production of phishing websites, earlier works were done using various data mining, classification algorithms were used, but the error rate of those algorithms was very high. Using Data mining algorithms and MapReduce approach in integration with anti-phishing technique we have achieved considerable time speedup. Even if the phishing webpage is not showing phishing characteristics very clearly at first layer it might show characteristics in the next layer so that no phishing webpage will pass through our system. This system is very effective in securing the network from phishing attached even at its best. As per type of organization we are protecting from phishing attach change the attributes to be considered for making effective decision about the phishiness of the system. We believe that this framework works better and gives a lower error rate and also the proposed methodology is also useful to prevent the attacks of phishing websites on financial web portal, banking portal, online shopping market.

ACKNOWLEDGEMENT

We would like to thank everyone who has motivated and supported us for preparing this manuscript.

REFERENCES

1. Tianyang Li.; Fuye Han.; Shuai Ding and Zhen Chen.; "LARX: Large-scale Anti-phishing by Retrospective Data-Exploring Based on a Cloud Computing Platform", in Proceedings of IEEE- 20th International Conference on Computer Communications and Networks, 2011. Qingxiang Feng.; Kuo-Kun Tseng.; Jeng-Shyang Pan.; Peng Cheng and Charles
2. Fu, A.Y.; Liu Wenyin; Xiaotie Deng, "Detecting Phishing Web Pages with Visual Similarity Assessment Based on Earth Mover's Distance (EMD)," Dependable and Secure Computing, IEEE Transactions on , vol.3, no.4, pp.301,311, Oct.-Dec. 2006.
3. I. H., & Frank, E. (2005). Data mining: Practical machine learning tools and techniques (3rd ed.). San Francisco, CA: Morga Kaufmann
4. Hong Bo; Wang Wei; Wang Liming; Geng Guanggang; Xiao Yali; Li Xiaodong; Mao Wei, "A Hybrid System to Find & Fight Phishing Attacks Actively," Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on , vol.1, no., pp.506,509, 22-27 Aug. 2011.
5. Elena Tsiporkova,Veselka Boeva,Elena Kostadinova, MapReduce and FCA Approach for Clustering of Multiple-Experiment Data
6. Maher Aburrous, M.A. Hossain, Keshav Dahal, Fadi Thabtah “Intelligent phishing detection system for e-banking using fuzzy data mining”, Expert Systems with Applications: An International Journal Volume 37 Issue 12, December, 2010
7. Intelligent phishing detection system for e-banking using fuzzy data mining by Maher Aburrous, M.A. Hossain, Keshav Dahal, Fadi Thabtah in 2010.
8. JungMin Kang, DoHoon Lee, "Advanced White List Approach for Preventing Access to Phishing Sites", 2007 International Conference onConvergence Information Technology, ICCIT 2007, p 491-496, 2007