An approach for malicious URL detection through URL heuristic

[1]Shruti Pankajkumar Parekh, [2]Shyamal Pandya
[1]PG Student, [2]Technical Evangelist
[1]Information Technology,
[1]Silver Oak College of Engineering & Technology, Ahmedabad, India

***Abstract*—Malicious URLs become major concern for internet users. Most of the time, Users use resources like emails, search engines, financial transactions, gaming services, etc. The replicas of legitimate websites are created to fool users. This paper gives an approach for finding malicious URLs by combining blacklisting approach and feature extraction approach. In addition, the Alexa rank heuristic is also considered as an important factor to decide whether site is phishing or not.**

***Index Terms*—Malicious URL, Heuristic**

I. INTRODUCTION

In phishing attack, the phishers induce internet users to click on a link to get user's confidential information like username, password, bank account number, credit card number, etc. Phishing is the word that is derived from the word 'Fishing' by replacing F with Ph. Most affected sectors are retail or service. After that, second highly affected sector is financial sector in which bank transaction details and bank account information are included. Then other sectors are there such as multimedia, social networking, government, unclassified, etc. Users must identify the URLs of surfing webpages, so that, they can be safe from the phishing attack by criminals. In some cases, the hackers try to send spoofed emails to the users from the well-known third party. The user assumes that the email is from the authenticated and known organization. But he or she does not know that it is going to be a data breach.

II. BACKGROUND THEORIES

Many researches have different approaches. Some of them use the TF-IDF (Term Frequency – Inverse Document Frequency) weighting system which assigns initial weight to the extracted words from the website content. The TF-IDF algorithm computes initial weights of all words in plain text extracted from HTML content. URL weighting system computes further weight and is then added to initial weight of words to get final weights. Some approach uses classification of URLs based on their lexical and host-based features. Other approach uses URL ranking, which includes three mechanisms: URL clustering, URL classification and URL categorization. In URL clustering, clustering is performed on entire dataset using K-means algorithm. It assigns a unique cluster ID for each URL. In URL classification, the URLs are classified using its features. In last mechanism, that is, URL categorization, all categories are extracted from Microsoft Reputation Service(MRS) and placed into three separate bags: Severe, Moderate and Benign. The URLs are given ranks using URL reputation services.

There are various heuristics of an URL. They are as:

- Page Rank
- Alexa Rank
- Age of Domain
- DNS record
- Abnormal URL
- Long URL
- Prefix Suffix
- Subdomains
- Http/Https
- IP address

Blacklisting is the most common technique used by major browsers such as Internet Explorer, Mozilla Firefox, Chrome, Opera, etc. When users of the browser try to load a URL that is in the browser's blacklist, the browser will warn the user by prompting the alert message that the site accessed by you can be harmful.

III. PROPOSED APPROACH

The proposed approach is based on identifying malicious URLs efficiently. The flow of the proposed approach is:

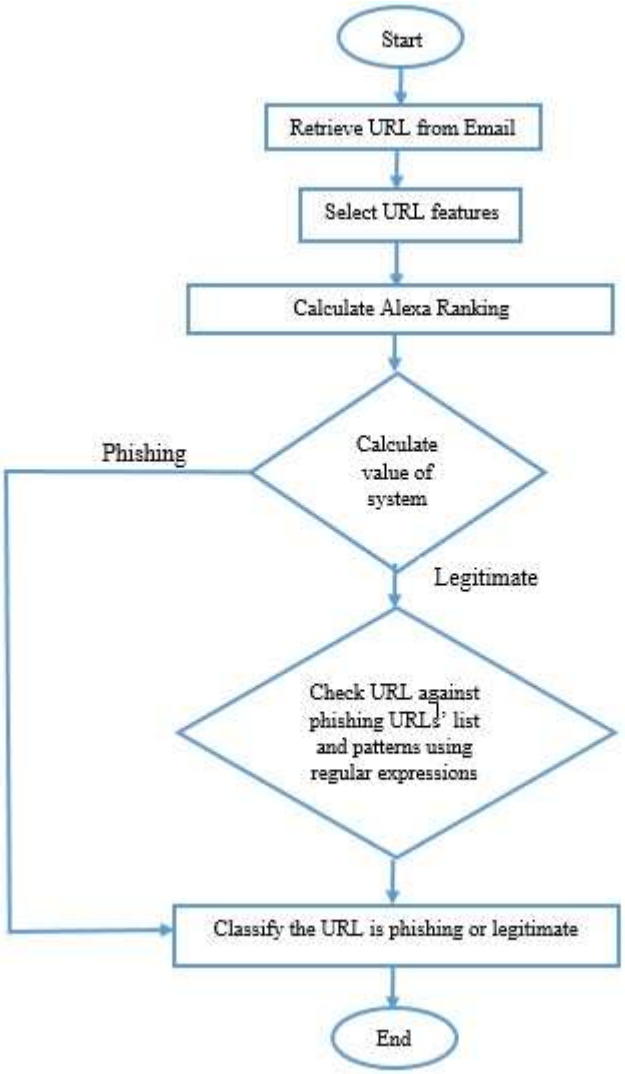

Fig. 1 Flow diagram of proposed approach

The Figure 1 shows the proposed approach. It starts process by retrieving URL from Email. There are 1000 URLs in the collection. All links from the email are collected. The features of collected URLs are selected and Alexa rank of URL is calculated. After calculating all these values, the whole value of system is calculated. The value of system is compared to threshold value. According to that, the URL is checked if it is phishing or legitimate. The filtered URLs from this process will be given to the next step. The next step is to check URL against phishing URLs list and also checked against patterns using regular expressions. By combining these two approaches, more accurate results can be obtained.

IV. RESULTS

By selecting URL features, the URL is divided into primary domain, sub domain, path domain, TLD and protocol. Figure 2 shows the result of the URL features selection:

URL	Protocol	PrimaryDomain	TLD	SubDomain	PathDomain
http://allibabaj.ultimatefreehost.com/login.alibaba.com/alibaba/Alibaba.html/	http	ultimatefreehost	com	allibabaj	/login.alibaba.com/alibaba/Alibaba.html/
http://mobilemusiclessons.com/on/dbfile/dbfile/best	http	-	com	mobilemusiclessons	/on/dbfile/dbfile/best
http://www.find-my-iphone.it/	http	find-my-iphone	it	www	/
http://shopplussizelingerie.com/dfghj/55425555555554242/	http	-	com	shopplussizelingerie	/dfghj/55425555555554242/
http://www.paypal.com.web-scr.us/paypal/cgi-bin/webscrcmd.php	http	paypal.com.web-scr	us	www	/paypal/cgi-bin/webscrcmd.php
http://www.paypa.com.index-php-cgibin.com/	http	paypa.com.index-php-cgibin	com	www	/
http://heroeswm-server-com.phpzilla.net	http	phpzilla	net	heroeswm-server-com	/

Fig. 2 URL features selection

After selecting features, all feature values are calculated using algorithms. Figure 3 shows the result of calculated feature values for number of URLs:

URL	alexaRankValue	primaryDomainValue	subDomainValue	pathDomainValue
http://allibabaj.ultimatefreehost.com/login.alibaba.com/alibaba/Alibaba.html/	0	0	1	1
http://mobilemusiclessons.com/on/dbfile/dbfile/best	1.00333333333333	0	0	1
http://www.find-my-iphone.it/	1.00333333333333	0	1	0
http://shopplussizelingerie.com/dfghj/55425555555554242/	1.00333333333333	0	1	1
http://www.paypal.com.web-scr.us/paypal/cgi-bin/webscrcmd.php	1.00333333333333	1	1	1
http://www.paypa.com.index-php-cgibin.com/	1.00333333333333	1	0	0
http://heroeswm-server-com.phpzilla.net	0	0	1	0

Fig. 3 Feature values calculation

The whole value of system is calculated according to values of Alexa rank, primary domain, sub domain and path domain. The value of system is then compared to threshold value. If value of system is less than the threshold value, then the URL will be phishing. If the value of system is greater than the threshold value, the URL will be legitimate.

Threshold	Phishing Sites
0.1	122
0.2	189
0.3	246
0.4	550
0.5	680
0.6	770
0.7	812
0.8	997

Table 1 Number of URLs at threshold value

At threshold value 0.7 and 0.8, best results can be obtained. So, the value of system of an URL can be compared to these two threshold values. By comparing value of system to these threshold values, the malicious URL identification can be done efficiently. The blacklist is prepared using standard dataset of Phish Tank. The filtered URLs are compared to this blacklist to get more results.

V. CONCLUSION

In this paper, we use the approach of combining two approaches, such as, blacklisting and feature extraction. Using this approach, the malicious URLs can be identified more accurately. The system model is checked for 1000 URLs which are collected from emails. The technique is experimented with primary domain, sub domain, path domain and Alexa ranking. Thus, combination of two approaches works efficiently.

REFERENCES

[1] L. A. T. Nguyen, B. L. To, H. K. Nguyen and M. H. Nguyen, "A novel approach for phishing detection using URL-based heuristic," 2014 International Conference on Computing, Management and Telecommunications (ComManTel), Da Nang, 2014, pp. 298-303.

[2] C. L. Tan, K. L. Chiew and S. N. Sze, "Phishing website detection using URL-assisted brand name weighting system," 2014 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), Kuching, 2014, pp. 054-059.

[3] M. N. Feroz and S. Mengel, "Phishing URL Detection Using URL Ranking," 2015 IEEE International Congress on Big Data, New York, NY, 2015, pp. 635-638.

[4] R. B. Basnet and A. H. Sung, "Mining Web to Detect Phishing URLs," 2012 11th International Conference on Machine Learning and Applications, Boca Raton, FL, 2012, pp. 568-573.

[5] Y. S. Chen, Y. H. Yu, H. S. Liu and P. C. Wang, "Detect phishing by checking content consistency," Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014), Redwood City, CA, 2014, pp. 109-119.

[6] S. Kaui and A. Kaur, "Detection of phishing webpages using weights computed through genetic algorithm," 2015 IEEE 3rd International Conference on MOOCs, Innovation and Technology in Education (MITE), Amritsar, 2015, pp. 331-336.

[7] J. James, Sandhya L. and C. Thomas, "Detection of phishing URLs using machine learning techniques," 2013 International Conference on Control Communication and Computing (ICCC), Thiruvananthapuram, 2013, pp. 304-309.
[8] R. Patil, B. Dasharath Dhamdhere, K. S. Dhonde, R. G. Chinchwade and S. B. Mehetre, "A hybrid model to detect phishing-sites using clustering and Bayesian approach," International Conference for Convergence for Technology-2014, Pune, 2014, pp. 1-5.
[9] MA, J., Saul, L.K., Savage, S., and Voelker, G.M., "Learning to Detect Malicious URLs". in ACM Transactions on Intelligent Systems and Technology, New York, NY, USA, Article 30 April 2011, ACM, pp. 1245-1254. DOI=http://dl.acm.org/citation.cfm?id=1961202.
[10] S. Sheng, B. Wardman, G. Warner, L. Cranor,J. Hong, and C. Zhang. (2009) An empirical analysis of phishing blacklists. [Online]. Available: http://ceas.cc/2009/papers/ceas2009-paper-32.pdf
[11] J. Ma, L. K. Saul, S. Safage, G. M. Voelker, "Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs," Proc. ACM SIGKDD, Paris, France, 2009, pp. 1245-1253.
[12] R. B. Basnet, A. H. Sung, Q. Liu, "Rule-Based Phishing Attack Detection" Proc. International Conference on Security and Management (SAM'11), Las Vegas, NV, 2011, pp. 624-630.
[13] Xiaoqing GU, Hongyuan WANG, and Tongguang NI "An Efficient Approach to Detect Phishing Web" Journal of Computational Information Systems, 2013, pp.5553-5560.
[14] M. G. Alkhozae and O. A. Batarfi, "Phishing websites detected based on phishing characteristic in the webpage source code," in International Journal of Information and Communication Technology Research, vol. 1, no. 6, Oct. 2011, pp. 283–291.
[15] R. Patil, B. Dasharath Dhamdhere, K. S. Dhonde, R. G. Chinchwade and S. B. Mehetre, "A hybrid model to detect phishing-sites using clustering and Bayesian approach," International Conference for Convergence for Technology-2014, Pune, 2014, pp. 1-5.