International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue VII, July 2017- Available at www.ijraset.com

Detecting Phishing Problem in Websites Using Weka

Mr. M. Mangaleswaran
Assistant Professor, Department of Computer Science and Engineering, Jansons Institute of Technology, Coimbatore

Abstract: ***The phishing trouble is taken into consideration a crucial trouble in industries mainly e-banking and e-trade taking the variety of online transactions related to payments. Specific features related to valid and phishy web sites have been identified and accumulated from 1353 web sites from distinction assets. Phishing websites had been amassed from Phishtank statistics archive which is a loose network web site wherein users can post, affirm, tune and share phishing records. The valid websites have been gathered from Yahoo and place to begin directories the use of a web script advanced in personal home page. The PHP script changed into plugged with a browser and 548 legitimate websites out of 1353 websites have been amassed. There may be 702 phishing URLs, and 103 suspicious URLs. Whilst a website is taken into consideration suspicious its method could be either phishy or valid, which means the internet site held a few legit and phishy capabilities. In this paper, phishing problem is detected the usage of WEKA.***

Keywords: Phishing, Phistank, Suspicious URL, APWG, WEKA

I. INTRODUCTION

Phishing is a social engineering method this is used to pass technical controls implemented to mitigate security dangers in statistics systems. Humans are the weakest link in any protection program. Phishing capitalizes on this weak spot and exploits human nature so that it will gain admission to a device or to defraud someone in their belongings. The Anti-phishing Work Group (APWG) is a global organization focusing on "putting off the fraud, crime and identification of robbery that result from phishing, pharming, malware and electronic mail spoofing of every kind". The APWG issues reviews semi-yearly regarding present day traits and rising assault vectors. The APWG reports that phishing within the second 1/2 of 2012 remained at an excessive degree and multiplied from the primary half of 2012. This suggests the range of phishing sites detected via the APWG for the July via December 2012. This demonstrates a clean risk to groups and private information; Combating phishing calls for consciousness of phishing attack vectors and strategies. This can be used to enhance the content material of existing phishing consciousness applications that generally target big audiences in a "shot gun" method to studying in which it has a broad spread of statistics for many goals straight away. This method refines and narrows the concern into a "rifle shot" method in which the audiences contain less human beings, and the statistics is greater in particular.

II. DATA SELECTION

The most appropriate attributes for detecting phishing in websites are SFH, Pop Up Window, Final state of SSL, URL Request, Anchor URL, Web traffic, Length of URL, Domain age, Having IP Address and Result. Following is the characteristics of dataset considered for phishing.

Table 1: Data Characteristics

Data Characteristics	Multivariate
Attribute Characteristics	Integer
Associated Tasks	Classification
No: of instance	1353
No: of attributes	10

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue VII, July 2017- Available at www.ijraset.com

A. Decision Table Classification vs. Naive Bayes

The stratified cross validation of 10 folds yields the following result with WEKA. The below is a comparison of Decision Table and Naive Bayes algorithm.

Table 2: Decision Table vs. Naive Bayes

Instances	Decision Table	Naive Bayes
Correctly Classified Instances	84.4789 %	84.3311 %
In correctly Classified Instances	15.5211 %	15.6689
Kappa statistic	0.7195	0.7127
Mean absolute error	0.1718	0.1383
Root mean squared error	0.2672	0.2777
Relative absolute error	45.908 %	36.9589 %
Root relative squared error	61.7963 %	64.23 %

The performance measures for the two algorithms give the following result.

Table 3: Performance Measures

Performance Measures	Decision Table	Naive Bayes
TP rate	0.845	0.843
FP rate	0.11	0.118
Precision	0.835	0.82
Recall	0.845	0.843
F-Measure	0.839	0.828
ROC Area	0.954	0.948

B. Simple K-Means Clustering

In K-Means Clustering, we classify the total instances into 3 clusters namely Full data cluster, Cluster 0 and cluster 1 with 1353, 842 and 511 instances respectively.

Table 4: Simple K-Means Clustering

Attribute	Cluster number		
	Full data	Cluster 0	Cluster 1
SFH	1	1	-1
PopUpWindow	0	0	-1
SSLFinal_State	1	1	-1
Request_URL	-1	0	-1
Anchor_URL	-1	1	-1
Web_Traffic	0	-1	1
URL_Length	0	0	-1
Age_of_Domain	1	1	-1
Having_IP_Address	0	0	0
Result	-1	-1	1

C. Visualization

The chart below shows the relation among three major attributes of the Phisihng dataset namely Request_URL, Web_Traffic and Having_IP_Address. The different colors denote three different clusters discussed above.

Fig 1: Visualization

III. CONCLUSION

Phishing will never be totally killed. In any case, the risk can be diminished through a mix of client and corporate protections and server-side measures. Client instruction remains the most grounded and in the meantime, the weakest connect to phishing countermeasures. It is likewise a scholarly commitment to the representative profession development and at last to the advancement of the host associations as more secure, phishing free working environments. Associations giving web benefits additionally have a part to play.

The best answer for phishing is preparing clients not to indiscriminately take after connections to sites where they need to enter delicate data, for example, passwords. Nonetheless, expecting that all clients will comprehend the phishing risk and surf in like manner is impossible. There will dependably be clients that are deceived into going by a phishing site. Along these lines, it is essential for analysts and industry to give answers for the phishing risk.

REFERENCES

[1] J.A Chaudhry, S.A Chaudhry, R.G Rittenhouse, R.G, "Phishing: Classification and Countermeasures", In: The 7th International Conference on Multimedia, Computer Graphics and Broadcasting. SERSC, Jeju, South Korea (2015).
[2] G. Ollmann, "The Phishing Guide--Understanding & Preventing Phishing Attacks", (2007).
[3] M. Jakobsson , S. Myers, "Phishing and Countermeasures: Understanding the Increasing Problem of Electronic Identity Theft", Wiley, Hoboken, NJ, USA (2006).
[4] J. Hong, "The state of phishing attacks. Commun", ACM. vol. 55, no. 74, (2012).
[5] Anti-Phishing Working Group: Phishing Activity Trends Report. (2014).
[6] A. Litan, "Phishing attack victims likely targets for identity theft", Gartner First Take FT-22. (2004).
[7] R. Dhamija, J.D Tygar, M. Hearst, "Why phishing works. In: Proceedings of the SIGCHI conference on Human Factors in computing systems" - CHI '06. p. 581. ACM Press, New York, New York, USA (2006).
[8] S. Sheng, M. Holbrook, P. Kumaraguru, L.F Cranor, J. Downs,"Who falls for phish? In: Proceedings of the 28th international conference on Human factors in computing systems" - CHI ' ACM Press, New York, New York, USA , vol. 10, (2010), p. 373.
[9] E. Earley, "Understanding social engineering", http://www.net-security.org/article.php?id=1403.
[10] T.N. Jagatic, N.A Johnson, M. Jakobsson, F. Menczer, "Social phishing", Commun. ACM. vol. 50, (2007), pp. 94–100.
[11] F. Zhou, "Phishing Sites and Prevention Measures", Int. J. Secur. Its Appl.vol. 9, (2015), pp. 1–10.
[12] F. Howard, O. Komili, "Poisoned search results: How hackers have automated search engine poisoning attacks to distribute malware", Sophos Tech. Pap. (2010).
[13] S.A Robila, J.W Ragucci, "Don't be a phish", ACM SIGCSE Bull. vol. 38, no. 237, (2006).
[14] P. Kumaraguru, S. Sheng, A. Acquisti, L.F Cranor, J. Hong, "Teaching Johnny not to fall for phish", ACM Trans. Internet Technol. vol. 10, (2010), pp. 1–31.
[15] Caputo, D.D., Pfleeger, S.L., Freeman, J.D., Johnson, M.E.: Going Spear Phishing: Exploring Embedded Training and Awareness. IEEE Secur. Priv. vol. 12, (2014), pp. 28–38.