Abstract

Phishing is an online crime that aims to create genuine looking websites to attract users and let them releasing their sensitive information on that fraud websites. Website phishing is one of the major attacks by which most of internet users are being fooled by the phisher. The best way to protect from phishing is to recognize a phish. Phishing emails usually appear to come from well-known organization and ask your personal information such as credit card number, security number, account number or passwords. What actually attacker does? The attacker creates the no of replicas of authenticate sites , and users are forced to direct to that websites by attracting them with offers. As standard mentioned in W3C (World Wide Web Consortium), I am proposing a system which can easily recognize the difference between authenticate site and phishing site. There are certain standards which are given by W3C (World Wide Web Consortium), based on these standards I am choosing some features which can easily describe the difference between legit site and phish site.

To protect you from phishing, I am proposing a model to determine the fraud sites. To
Web Phishing Detection System: Bayesian and Clustering Approach
determine the phishing attack, URL features and HTML features of web page are considered.
Clustering algorithm such as K-Means clustering is applied on the database and prediction
techniques such as Naive Bayes Classifier is applied. By applying this, probability of the web
site as valid Phish or Invalid Phish. To check the validity of URL, if still we are not able decide
the validity of web page then Naive Bayes Classifier is applied. Also training model is applied
for the extraction of HTML tag features of site and probability.

References

1. Rachna Dhamija, J. D. Tygar, and Marti Heast, “Why Phishing Works”, CHI-2006,
2. RSA Online Fraud Surveyor, “The phishing kit – the same wolf, just different sheep’s
clothing”, RSA Surveys, vol-1, February-2013.
3. Xiaojing GU, Hongyuan WANG, and Tongguang NI “An Efficient Approach to Detect
4. Haijun Zhang, Gang Liu, Tommy W. S. Chow, Senior Member, IEEE, and Wenyin Liu,
   Senior Member, IEEE “Textual and Visual Content-Based Anti-Phishing: A Bayesian Approach”,
5. Angelo P. E.Rosiello, Engin Kirda, Christopher Kruegel, Fabrizio Ferrandi, and Politecnico
di Milano “A Layout-Similarity-Based Approach for Detecting Phishing Pages”- unpublished
8. PHISHTANK.COM- The Online Valid Phish Sites Repository,
   http://data.phishtank.com/data/online-valid.csv
10. A hybrid model for detection of phishing sites using
11. clustering and Bayesian approach, 6th April 2014.

Index Terms

Computer Science

Information Sciences

Keywords

Anti Phishing, Bayesian technique, Data Mining, Database Clustering, and Phishing Attack.