Comparative Analysis of Features Based Machine Learning Approaches for Phishing Detection

Ankit Kumar Jain
Department of Computer Engineering, National Institute of Technology, Kurukshetra, India
ankit.jain2407@gmail.com,

B. B. Gupta
Department of Computer Engineering, National Institute of Technology, Kurukshetra, India
gupta.brij@gmail.com

Abstract –Machine learning based anti-phishing techniques are based on various features extracted from different sources. These features differentiate a phishing website from a legitimate one. Features are taken from various sources like URL, page content, search engine, digital certificate, website traffic, etc, of a website to detect it as a phishing or non-phishing. The websites are declared as phishing sites if the heuristic design of the websites matches with the predefined rules. The accuracy of the anti-phishing solution depends on features set, training data and machine learning algorithm. This paper presents a comprehensive analysis of Phishing attacks, their exploitation, some of the recent machine learning based approaches for phishing detection and their comparative study. It provides a better understanding of the phishing problem, current solution space in machine learning domain, and scope of future research to deal with Phishing attacks efficiently using machine learning based approaches.

Keywords – Phishing, Machine learning, soft computing, neural networks, Support vector machine, Domain Name System.

I. INTRODUCTION

Phishing is one of the dangerous cyber security threats which contains fake web page that pretends to be truthful. This fake web page is used to performed phishing attack using social engineering techniques [1]. The motive of the attacker behind such attacks may be identity theft, financial gain or notoriety (i.e., to get recognition) [20-23]. Detection and prevention from phishing attacks is a big challenge to scientist and researcher because attacker performs these attacks in such a way that they bypass the existing anti-phishing techniques and even an educated and experience user may fall under these attacks. Usually phishing attack is performed by sending a fake email which appears to come from popular and trusted brand or organization, asking to input credential like bank login, password, etc [19]. Phishing messages are spread over using emails, SMSs, instant messages, social networking sites, VoIP, etc, however, email is the popular way to perform this attack. 65% of the total phishing attacks are performed by sending the malicious hyperlinks within the e-mail. Phishing attack lifecycle is shown in figure 1.

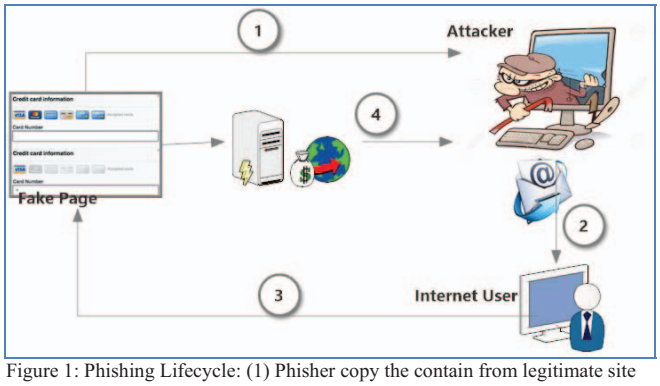

Figure 1: Phishing Lifecycle: (1) Phisher copy the contain from legitimate site and construct the phishing site (2) Phisher sent link of phishing URL to internet user (3) User open the link and fill personal on fake site (4) Phisher steal personal information of user.

According to the anti-phishing working report in the first Quarter of 2014, second highest numbers of phishing attacks are observed between January-March 2014 [2] and payment services are the most targeted by these attacks. The total number of phishing attacks notice in Q1 (first quarter) of 2014 was 125,215. Recent developments in phishing detection have led to the growth of various new machine learning based approaches. Machine learning based anti-phishing solutions are based on various features extracted from different sources [17]. These features differentiate a phishing website from a legitimate one. The features are taken from various sources like URL [18], page content, search engine, digital certificate, website traffic, etc, of a website to detect it as a phishing or legitimate. The accuracy of the anti-phishing solution depends on features set, training data and machine learning algorithm.

Rest of the paper is organized as follows. Section 2 presents the taxonomy of various types of Phishing detection and filtering techniques; specially, this section focuses on machine learning approaches in details. Section 3 presents the various heuristics used by machine learning approaches to train the system. Section 4 presents the open issues and challenge in phishing detection and protection. Finally, section 5 concludes the paper.

978-9-3805-4421-2/16/$31.00 ©2016 IEEE

II. TAXONOMY OF MACHINE LEARNING BASED PHISHING DETECTION APPROACHES

In this section, we present an overview of phishing detection approaches. Phishing detection schemes are broadly classified in two ways. First is based on user education and another is based on software. Software-based approaches are further classified into machine learning based, blacklist based and visual similarity based. A machine learning based approach uses common features of phishing and legitimate site such as the login form, uniform resource locator, DNS information, etc. The websites are declared as phishing sites if the heuristic design of the websites matches with the predefined rules. The negative aspect of the blacklist and visual similarity based approaches is that it usually doesn't cover newly launched (zero hour attack) phishing websites. Some of the machine learning based phishing detection and protection approaches are explained below.

A. CANTINA

Zhang et al [3] proposed a page content-based anti-phishing technique called CANTINA, which take a rich set of feature set from various field of a webpage. Proposed technique calculates term frequency-inverse document frequency (TF-IDF) of the content of a website and creates a lexical signature. Top 5 terms with highest TF-IDF values submit into the search engine. The top 'n' result is used to check the legitimacy of a website. Though, the performance of CANTINA is affected by the TF-IDF algorithm and language used on the website. To reduce the false positive rate author also take some heuristic like the age of domain, URL contain "-" or "@" symbol, dots(.) count in URL, etc.

B. CANTINA+

Xiang et al. [4] proposed CANTINA+ is an upgraded and modified version of CANTINA. In CANTINA+, author added 10 new features along with previous features to achieve better accuracy. System filters the website without login form in the first step to decrease false positive rate. The proposed approach takes the 15 features from URL, HTML DOM(Document object model), third party services, search engine and trained these features using SVM(Support vector machine) to detect phishing attack. True positive rate of CANTINA+ is 92% and low false positive rate is 0.4%.

C. Detection of Phishing attacks in Iranian e-banking

Montaze et al [5] proposed phishing Detection system to be used in e-banking system in Iran. Authors identified the 28 features used by the phisher to deceive the Irani banking sites. The accuracy to filter the phishing sites is 88% on Iranian banking system. System uses the fuzzy expert system to train the system.

D. A comparison of machine learning techniques

Abu-Nimeh et al. [6] compared six machine learning algorithms for phishing detection. Machine learning algorithms are Bayesian Additive Regression Trees, Logical Regression, SVM, RF, Neural Network, and Regression Trees. The result shows that there are no standard machine learning algorithms for phishing detection. Authors trained dataset using 43 features. There is the trade-off between false positive rate and false negative rate, if some algorithm has low false Positive rate then it have a high false negative rate. Logical Regression whose false positive rate is 4.9%, obtained a large number of false negative rate of 17%.

E. Associative Classification data mining

Neda Abdelhamid et al [7] proposed Multi-label Classifier based Associative Classification method for website phishing. Associate classification is effectively detecting phishing websites with high accuracy and MCAC are use to generate new rules and it also enhance its classifiers predictive performance.

F. classification mining techniques

Aburrous M[8] present a technique based on Classification Data Mining (DM) for detection on e-banking phishing websites. Auther implement the six different classification algorithm to measure the performance and accuracy on each of algorithm, the drawback of this algorithm is that the false positive rate is very high (13%).

G. Intelligent phishing detection system for e-banking

Aburrous M [9] present a technique based on fuzzy data mining for detection of phishing attack in banking. The authors used 27 features to construct a model based on fuzzy-logic. Author took the features from the URL and web to identify the difference between phishing and legitimate website. They explore the good applicability of fuzzy data mining in phishing detection. In this paper, authors focus only e-banking sites and did not discuss the performance of other type of websites. Furthermore, the website was classified into five different classes.

H. A SVM-based technique to detect phishing URLs

H Huang [10] et present an approach based on the URL based features. They have taken 23 features from URL and train the system using SVM. System takes Decision based on lexical and brand name features of URL.

I. Using Latent Dirichlet Allocation and AdaBoost

V. Ramanathan [11] presents an anti-phishing solution which extracts the feature of a webpage using Latent Dirichlet Allocation and AdaBoost use as a classifier.

J. A Framework for Detection and Measurement of Phishing Attacks

In [12], Garera et al. proposed a technique based on phishing URLs explained four different kinds of obfuscation techniques of phishing URLs. In proposed work, we have taken various URL features and suspicious keywords found in URL along with some addition and modification of this technique. We also taken the classifier used by this technique.

K. Learning to detect phishing emails

Fette et al. [13] proposed a technique to classify phishing and legitimate emails. They used a large publicly available corpus of genuine and fake emails. Their learning classifiers examine 10 different features such as the presence of hyperlink an e-mail, IP Address, age of link, etc.

L. phishGILLNET

To improve the accuracy of phishing detection system, Ramanathan et. Al [14], proposed a multi-layered approach phishgillnet. Proposed machine learning based system contain algorithm in three layers. The first layer is Probabilistic Latent Semantic Analysis (PLSA), which is used to build a topic model. The second layer is AdaBoost, which is used to build a robust classifier; and the third layer is Co-Training, which is used to build a classifier from labeled and unlabeled examples.

M. A comprehensive and efficacious architecture for detecting phishing webpages

Gowtham et al [15] proposed heuristics to extract 15 features from webpages. These heuristic results were fed as an input to a trained machine learning algorithm to detect phishing sites. The system used two preliminary screening modules in this system. The first module, the preapproved site identifier, checks webpages against a private white-list maintained by the user, and the second module, the Login Form Finder, classifies webpages as legitimate when there are no login forms present.

TABLE I. LIMITATION AND ADVANTAGE OF VARIOUS ANTI-PHISHING SOLUTIONS

Approach	**Machine Learning Algorithm Used**	**Advantage**	**Limitation**
CANTINA[3]	TF- IDF algorithm used to fetch keywords	Fast and not dependent on prior training	Dependent on accuracy of TF-IDF algorithm, fail to detect phishing page if embedded objects use
CANTINA+[4]	J48 Decision Tree , Adaboost , SVM, LR, Bayesian Network (BN), Random Forest (RF)	Some novel heuristic are used to detect phishing attack	Fail to detect DNS poisoning cannot detect the compromised domain IP address based legitimate URLs
Phishing attacks in Iranian e-banking [5]	fuzzy–rough hybrid system	Novel features proposed to detect bank site	False Positive Rate is high(12%)
Associative Classification data mining[7]	Multi-label Classifier based Associative Classification (MCAC)	Proposed applicability of MCAC in phishing detection	Accuracy of system is not explained
Classification mining techniques[8]	C4.5, JRip, PART, PRISM, CBA and MCAR	Design to detect the e-banking site	False positive rate is very high
Intelligent phishing detection system [9]	fuzzy data mining	Applicability of fuzzy data mining in phishing detection	Focus only on e-banking phishing detection
A SVM-based technique to detect phishing URLs[10]	SVM	Its content independent so it can detect if embedded object present in website	Based on URL based features so accuracy is low
Using Latent Dirichlet Allocation and AdaBoost[11]	semantic analysis Latent Dirichlet Allocation, and for classification, AdaBoost	Can detect zero hour attack	Justification is not given to choose features
A Framework for Detection and Measurement of Phishing Attacks[12]	Logistic Regression	Novel URL features proposed	Based on URL based features so accuracy is low
Learning to detect phishing emails[13]	Random forest and support vector machine (SVMs)	10 different features including whois query	phishing and legitimate emails not well classified.
phishGILLNET [14]	PLSA, AdaBoost, Co-Training algorithm	Accuracy of approach is high	Take more memory and computation time
A comprehensive and efficacious architecture [15]	Support Vector Machine	Use two preliminary filter to improve accuracy	Cannot detect embedded object and compromised domain

III. FEATURES OF PHISHING WEBPAGE

The accuracy of a phishing detection system is depending on the features set which distinguishes the phishing and legitimate site. In this section, we explore the important

features which are used to train the dataset. The features are taken from various sources; moreover, we also present some new features and update the some features.

A. URL based features :-

IP Address: IP address can be use in URL in place of the domain name. A phisher may use the IP address in place of the domain name to hide the identity of a website.

Sub Domain: Phishing sites may contain more than two sub-domains in URL. Each domain is separated by the dot symbol (.). If any URL contains three or more than three dot then the probability of the suspicious site is more.

URL contain "@" Symbol: the presence of "@" symbol in the URL ignore everything previous to it.

Number of dash (-) in host name: To looks like genuine URL, phisher adds some prefix or suffix to the brand name with the dash symbol. e.g. www.amazon-india.com.

Length of URL: To hide the domain name, phisher uses the long URL. In our experiment, we found the average length of phishing website URL is 74. 30.3% of phishing URL contain more than 80 characters.

Suspicious words in URL: Phishing URLs contain suspicious words such as token, security, paypal, login, signin, bank, account, update, blog etc to gain the trust on the website.

Position of Top-Level Domain: This feature checks the position of top level domain at the proper place in URL.

Example-
http://xyz.paypal.com.accounts.765issapidll.xtmll.ebmdata.com
.

Embedded Domain in URL: It checks this by checking for the occurrence of "//" in the URL. Phishing URL may contain more than one domain in URL. Two or more domains are used to forward address.

HTTPS Protocol: HTTPS protocol is used for security. Phishing does not start with https while legitimate URL provides security. (Only 388 phishing sites contain https protocol)

Brand Name in URL: Most of the phishing websites contain the brand of the targeted domain in somewhere in the URL. According to current report of Q3 2014 APWG [17] the 45% of phishing sites contain brand name of targeted site in URL. If the URL is matched with some brand and position of domain is not right place then declare as phishing. Out of total targeted

brand 32% share of payment services where 27% share is the financial organization and 13 % of ISPs. [17].

Ex. www.example.com/ ebay.in

B. Lookup based features:-

DNS lookup: The life of phishing site is very short, therefore; this DNS information may not be available after some time. If the DNS record is not available then website is phishing. The median uptime of phishing attacks is 8 hours and 42 minutes [2].

Whois database: If the domain name of suspicious webpage is not match with the WHOIS database record, then webpage consider as phishing.

Age of Domain: Life of the phishing sites are very short and they most of the phishing site not working after 48 hour of their launching. If age of website is less than 3 month then chances of fake webpage is more.

C. Search Engine based features

Page Rank : Google Pagerank is an important factor search engine optimization (SEO). The value of page rank is between 0 and 1. Page rank measures the important of a website in search engine by checking content quality. Higher page rank means site is on top in search result. The rank of phishing site is 0.

Search Result: To check legitimacy of a suspicious webpage, the phishing detection approach uses the search engine as a tool. Identity of any webpage can easily checked by the search engine. If a user put a query in a the search engine, then it always gives the link of genuine and relevant webpages in its top results, search engine never return phishing webpage in top result.

D. HTML DOM based features :

Server form handler ''SFH'': Phishing WebPages usually contains login form to steal credentials of online users. When a user inputs his\her personal information in a fake site, which is forwarded to the attacker, afterwards attackers use the personal information of the user for financial or some other benefits. Attackers either use the different domain (other than visited domain) or null (hyperlink in footer section) in the phishing sites.

Hyperlinks: In the legitimate site, most of the links are pointing to some domain but in the phishing site most of the links point to different domains. If the ration of links pointing to the same domain and total domain present is less than 50% then it is likelihood the phishing site. Hyperlinks are extracted from page source using <a href>, <link> , <meta> and <script> tag. From our study we observed that in phishing page most of hyperlinks are null, i.e they are not redirecting to any other webpage .

Example: <a href="#">

E. Certificate based Features

TLS/SSL Certificate: This certificate is used to prevent an attacker from impersonating a secure website. All legitimate websites contain the TLS/SSL certificate but phishing websites do not contain this certificate from fake Certificate authority. We recommended to check if this certificate provided by a trusted issuer such as ''Comodo, Symantec, GoDaddy, GlobalSign and DigiCert.

F. Website Traffic features

Website Traffic: Phishing websites are having very low traffic as compare to legitimate websites because legitimate websites visited very frequently. If the web traffic in a website is very low then it may a phishing site but if the traffic rate is high then the website is the legitimate one.

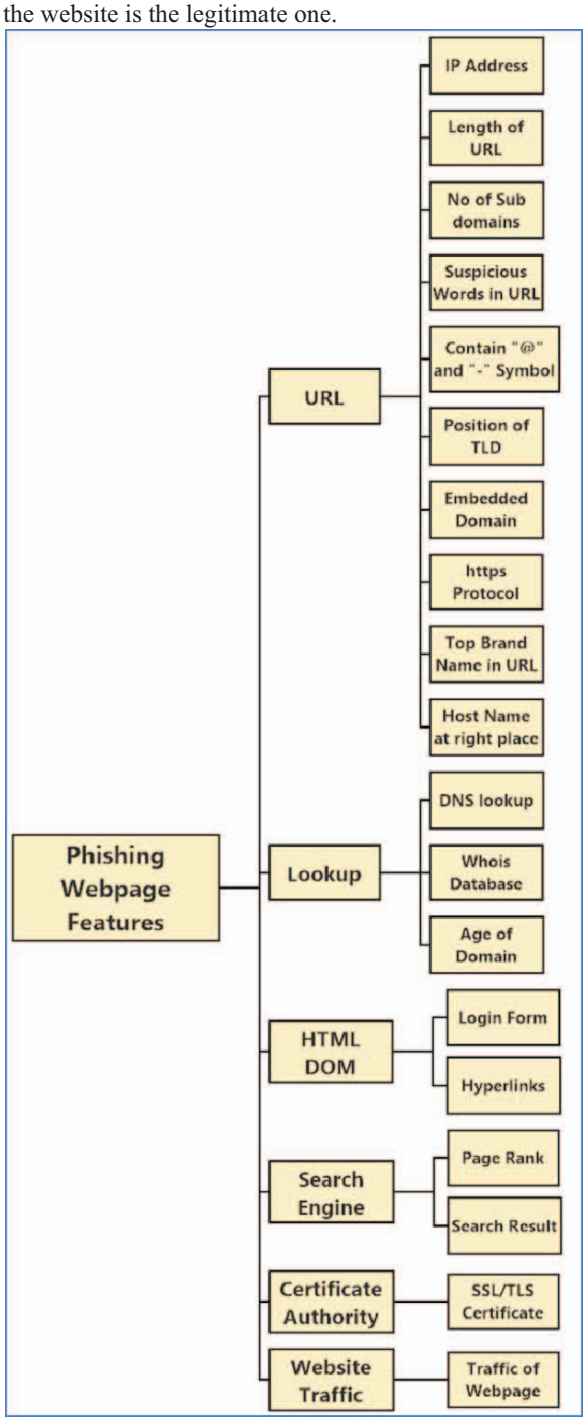

Figure 2: Phishing website features set

IV. OPEN ISSUES AND CHALLENGES

We have discussed various anti-phishing techniques based on machine learning. However, still there exists no single technique that can detect various types of phishing attacks. Detection of phishing attack is the example of supervised learning, therefore, accuracy of the solution depends on features set, machine learning algorithm and training data. The first issue is the zero hour attack as most of the anti-phishing techniques compare the features of the suspicious website and predefine feature set. Therefore, the accuracy of the system depends on features set and how accurately these features are selected. Most of the e-commerce and banking sites have different text languages in various countries e.g. Amazon, EBay, Citibank, while the layout is almost similar. Machine learning based techniques are based on heuristics, which include the frequently appearing keyword in the phishing site. If these techniques can detect the keyword written in the English language, these techniques may not be able to detect text written in another language, e.g. Chinese, Hindi etc. Moreover, another issue present in the phishing webpage is embedded objects, which can bypass the anti-phishing solution. The attacker uses images, JavaScript, etc, in place of text to bypass the anti-phishing system. Therefore, detection of phishing site which uses embedded object is still an open challenge. Machine learning based techniques require a high computational power to extract and compute the features in real time environment.

V. CONCLUSION AND FUTURE SCOPE

Phishing is the one of the most serious threat in the cyber world. In this attack, the user inputs credential to a bogus website which looks like a genuine one. In this paper, we have presented a survey on phishing website detection based on machine learning approaches. These approaches use various features of a web page to detect phishing attacks, such as text, URL, website traffic, login form, certificate, etc. Some of these approaches still have limitations in terms of accuracy, embedded objects, zero hours attack, etc. Detection of phishing site which uses embedded object is still an open challenge. Moreover, machine learning based techniques require a high computational power to extract and compute the features in real time environment.

REFERENCES

[1] A. Almomani, B. B. Gupta, S. Atawneh ,A. Meulenberg and E. ALmomani, "A Survey of Phishing Email Filtering Techniques," IEEE Communications Surveys & Tutorials, vol. 15, no. 4, pp. 2070-2090, 2013.

[2] APWG Q1 report, Available at: docs.apwg.org/reports/apwg_trends_report_q1_2014.pdf (Last accessed on 6 November 2015)

[3] Y. Zhang, J. Hong, and L. Cranor, "Cantina: a content-based approach to detecting phishing web sites," in Proceedings of the 16th international conference on World Wide Web, 2007.

[4] G. Xiang, J. Hong, C. Rose, and L. Cranor, "Cantina+: A feature-rich machine learning framework for detecting phishing web sites," ACM Transactions on Information and System Security, vol. 14, no. 2, 2011.

[5] G. A. Montazer, S. Yarmohammadi, “Detection of phishing attacks in Iranian e-banking using a fuzzy–rough hybrid system,” Applied Soft Computing, Vol. 35, pp. 482-492, 2015.
[6] S. Abu-Nimeh, et al, “A comparison of machine learning techniques for phishing detection,” in Proceeding of eCrime Researchers Summit,, Pittsburgh, ACM Conf, Pittsburgh, PA, pp. 60-69, 2007.
[7] Neda Abdelhamid, Aladdin Ayesh, Fadi Thabtah, “Phishing detection based Associative Classification data mining,” Expert Systems with Applications, Volume 41, Issue 13, pp. 5948-5959, 2014
[8] M. Aburrous, MA Hossain,K Dahal ,T. Fadi , “ Predicting phishing websites using classification mining techniques,” In:Seventh international conference on information technology, Las Vegas, Nevada, USA, 2010.
[9] M. Aburrous, MA Hossain, K Dahal, F. Thabtah, “ Intelligent phishing detection system for e-banking using fuzzy data mining,” Expert Syst Appl Int J vol 37 no. 12, pp. 7913–7921, 2010.
[10] H. Huang, L. Qian, Y. Wang, “A SVM-based technique to detect phishing URLs,” Inf.Technol. J. vol. 11, no. 7, pp. 921–925, 2012.
[11] V. Ramanathan, H. Wechsler, “Phishing website detection using Latent Dirichlet Allocation and AdaBoost,” IEEE International Conference on Intelligence and Security Informatics. Cyberspace, Border, and Immigration Securities, Piscataway, NJ, pp. 102–107, 2012.
[12] S. Garera, N. Provos, M. Chew, and A. D. Rubi, “A Framework for Detection and Measurement of Phishing Attacks,” In WORM ’07.
[13] I. Fette, N. Sadeh, and A. Tomasic, “Learning to detect phishing emails,” In proceedings of the 16th International Conference on World Wide Web, Banff, Alberta, Canada, 2007.
[14] V. Ramanathan and H. Wechsler, "phishGILLNET phishing detection methodology using probabilistic latent semantic analysis, AdaBoost, and co-training, " EURASIP Journal on Multimedia and Information Security, vol. 2012, no. 1, pp. 1-22, 2012.
[15] R. Gowtham, Ilango Krishnamurthi, “A comprehensive and efficacious architecture for detecting phishing webpages,” Computers & Security, Vol. 40, pp. 23-37, 2014.
[16] APWG 2014 H2 Repoet Available at : http://www.antiphishing.org/download/document/245/APWG_Global_Phishing_Report_2H_2014.pdf
[17] A. Almomani , B.B. Gupta, T.Wan, A. Altaher "Phishing Dynamic Evolving Neural Fuzzy Framework for Online Detection Zero-day Phishing Email" , Indian Journal of Science and technology , vol. 6 , no. 1 , 2013.
[18] A. Jain, BB Gupta, “PHISH-SAFE: URL Features based Phishing Detection System using Machine Learning,” In proceeding of CSI-2015, New Delhi, India, December 2015.
[19] A. Mishra, BB Gupta, “Hybrid Solution to Detect and Filter Zero-day Phishing Attacks,” In proceeding of ERCICA 2014.
[20] S. Tripathi, B. B. Gupta, A. Almomani, A. Mishra, S. Veluru , "Hadoop based Defence Solution to Handle Distributed Denial of Service (DDoS) Attacks," Journal of information Security (JIS), Scientific Research, vol. 4, no. 3, pp. 150-164, 2013.
[21] A. Mishra, B. B. Gupta, “Hybrid Solution to Detect and Filter Zero-day Phishing Attacks,” Second International Conference on “Emerging Research in Computing, Information, Communication and Applications” (ERCICA-14), Aug., 2014.
[22] B. B. Gupta, S. Gupta, S. Gangwar, et. al, “Cross-Site Scripting (XSS) Abuse and Defense: Exploitation on Several Testing Bed Environments and its Defense, Journal of Information Privacy and Security, Taylor & Francis, Vol. 11, Issue 2, pp. 118-126, 2015.
[23] E. Alomari, S. Manickam, B. B. Gupta, et. al., “Design, Deployment and use of HTTP-based Botnet (HBB) Testbed,” in proceedings of 16th International conference on Advance Communication Technology (ICACT-2014), Phoenix Park, PyeongChang, South Korea Feb. 16-19, 2014.