A Hybrid Approach to Detect Suspicious URLs

Saurabh Muthal[1], Ameya Pawar[2], Saurabh Harne[3]

[1] Student, Computer Department, Vishwakarma Institute of Information Technology, Maharashtra, India
[2] Student, Computer Department, Vishwakarma Institute of Information Technology, Maharashtra, India
[3] Student, Computer Department, Vishwakarma Institute of Information Technology, Maharashtra, India

ABSTRACT

Social network services are increasing trend now days. Communicating between people forms a social network. Personal information is usually collected through Social network services in targeted attacks and craft attacks based on a specific user profile. Online payment systems can be used to create a fake payment portal by using a similar gateway URL (Online fraud). Because users are curious and unknown, they generally click on suspicious URLs without verification. A feature set is presented that merged the features of traditional heuristics and lexical features of URLs. A suspicious URL identification system for use in all over web environments is proposed based on K-means clustering algorithm and Bayesian classification.

Keywords: - *Bayesian classification, Malicious websites, K-means clustering.*

1. INTRODUCTION

In a matter of very few years, the internet which is defined as worldwide interconnection of individual networks are growing fast. From 1994 it has been increasing continuously to serve millions of users for business, communication and various other uses. Most of the users don't know how it works or what safety precautions should be taken while using it. Some hackers take advantage of it and they create some fake links to steal user's data. These links are the URL's (Uniform Resource Locater). It is the address of the webpage. We are not able to tell whether the URL is legitimate or not by looking towards it with naked eye. The URL consists of protocol identifier and resource name. The protocol identifier tells which protocol is used and resource name specifies the IP address or the domain name. By which we can able to extract various features from URL. It is possible to predict whether it suspicious or not from the characteristics of features.

There were various techniques introduced to detect suspicious URL's. They were resource and time consuming. They are analytic tool more than detection system. Most of them were for social sites like Facebook, Twitter etc. There is need to build the URL detection system for all links on the internet. The system must be able to differentiate between suspicious URL and legitimate URL on the webpages.

This paper is mainly focuses on the methodology of detection. We are proposing the hybrid approach means combining the two methods clustering and classification for better prediction. This hybrid method shows the suspicious feature values of URLs.

2. LITERATURE SURVEY

This section introduces the previous existing detection models for spam URLs. In recent past lots of work were done in spam URL detection.

Jason Hong et al. [1] proposed a content based detection system where five keywords were extracted based on the term-frequency/inverse document frequency (TF/IDF) algorithm from the web page and for verifying the authenticity, Google search was applied.

McGrath and Gupta [2] analysed collected URL's as phishing and non-phishing websites. They extracted semantic features from the URL such as brand name. They mainly focus on understanding modus operandi of the phisher.

Kurt Thomas, David Nicol [3] observed the detection and infection rate of Koobface worm. The blacklist services required 4 days to react and identified very few among them. There is need of efficient detection method to react dynamic changing URLs.

Robertson et al. [4] combined social network information and security heuristics. It searches for the friend list of the user and the news feed generating source and all posts from the user. However it yields only 62% accuracy rate.

Chen et al. [5] uses Bayesian algorithm for classification of spam URLs on the basis of extracted features. They categorized the features as social features; lexical features, host features. They give manual threshold values to form the group of similar feature values. It is difficult to face dynamic changing structure of URLs.

S. Lee and J. Kim [6] proposed the robust system for spam detection in the Twitter which is known as WarningBird. The dynamic redirections are not handled by this system.

D. Canali et al. [7] found that JavaScript features, HTML features can be used for the efficient detection of malicious websites. It can easily adapt to changes in malicious websites and their features over time.

According to C. Whittaker et al. [8], a machine learning classifier can be used to maintain the blacklist of phishing websites.

3. PROPOSED SYSTEM

We are preparing the data set to train the machine. We are using two data classes; one is of malicious URLs and other one containing legitimate URLs. We are using two machine learning techniques; K-means for clustering and Naïve Bayes for classification. There are five models in our system described as follows.

3.1 Data Collection

We are creating training data set. This data set later on load into the system to train the module. The data set save into the serialize file. The comma separated values are stored into that file. We can add and remove data dynamically.

3.2 Feature Extraction

- IP address :

  It checks whether the given URL is an IP address or DNS address. Many times it happens that URL containing IP address considered as spam but it is not always true.

- Dot count in host name :

  This lexical feature is important for detecting malicious URLs. J. Hong et al. [1] supported that various malicious URLs contain more dots, whereas legitimate URLs include three or four dots.

- Suspicious count :

  It is the count of number of special symbols appears in the URL. Generally malicious URLs contains more special symbol than the legitimate URL.

- Slash count :

  The malicious URLs are generally having long length and path. So the number of slash in the URL is more. Sometimes the legitimate URLs also have long length but machine learning train the module relatively.

- Nil anchor count :

  It is the count of redirecting URL but it is not going on other webpage. It redirects to the nil page.

- Foreign anchor count :

  This feature counts the number of foreign anchors means it counts the number of times it is redirecting to the other webpage.

- Protocol Identifier :

  It checks which protocol it is using. Whether is it HTTP or HTTPS? The HTTPS is secure protocol than the HTTP.

3.3 Manage and View

This module shows the entire dataset. We can alter it anytime. It shows the URL with the extracted features and their count.

3.4 Clustering

In this module, we create the similar group within the features as low and high to set the threshold value. K-means is used as clustering method.

- K-means :

  In K-means, we are forming the K clusters by calculating the distance measure of feature values. The formulae for calculating the distance measure is as follows,

$$d=\sqrt{\sum_{i=0}^{n}\left(x_i - y_j\right)^2} \qquad (11)$$

  First, we select the random centroids and then calculate the distance of each point to the centroid. The closest distance points are group under that centroid. The average has being taken to select new centroid. Again it calculates the distance between new centroid and all points. This process repeats until no new centroid is obtains. We get the final clusters after completion of above process.

3.5 Classification and Detection

In this model, the real time classification and detection take place. The Bayesian classifier is used to train the data. The transformed data is passed to the module.

- Bayesian Classifier :

  It is the probabilistic model. The independent conditional probability is calculated by the model. The formula for calculating the probability is,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad (12)$$

- Detection :

  The system will detect malicious URL using classification and it will also tell the prediction in percentage.

4. CONCLUSIONS

The previous URL detection systems are mainly for social networking sites. Therefore the methods are not efficient for URL on other websites than social networking sites. The proposed system is extracting lexical features from the URL. It does not depend on whitelist or blacklist mechanism. The proposed system uses K-means algorithm for clustering the feature values as two clusters. It is used as threshold values. The system uses Bayesian classifier to calculate the independent probability and classification of the feature. This method is useful for the noisy data also. Overall it is the robust and dynamic method for real time detection.

5. FUTURE SCOPE

The browser extension for detecting malicious URLs can be done in future. The continuous changing dynamic structure of URLs needs to be addressed.

6. ACKNOWLEDGEMENT

First of all we thank the computer department and Head of the department Professor Dr.S.R.Sakhare for giving the support to complete the project. Next we express our gratitude to our respected Professor Mrs.M.P.Karnik for giving us knowledge, courage and following us to do the project work intentionally.

7. REFERENCES

[1]. Y. Zhang, J. Hong, L. Cranor, "CANTINA: a content-based approach to detecting phishing web sites," Proc. of the International World Wide Web Conference (WWW), 2007.

[2]. D. Kevin McGrath, M. Gupta, "Behind Phishing: an examination of phisher modi operandi," Proc. of the USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET), 2008.

[3]. K. Thomas, D. Nicol, "The Koobface botnet and the rise of social malware," in Malicious and Unwanted Software (MALWARE) 5th International Conference, 2010.

[4]. M. Robertson, Y Pan, B. Yuan, "A social approach to security: using social networks to help detect malicious web content," International Conference on Intelligent Systems and Knowledge Engineering (ISKE), 2010.

[5]. C.M. Chen, D.J. Guan, Q.K. Su, "Feature set identification for detecting suspicious URLs using Bayesian classification in social networks" Information Science 289 (2014) 133-147.

[6]. S. Lee, J. Kim, "WarningBird: Detecting Suspicious URLs in Twitter Stream," in IEEE Transactions on Dependable and Secure Computing, May/June 2013.

[7]. D. Canali, M.Cova, G. Vigna, C. Kruegel, "Prophiler: A Fast Filter for the Large-Scale Detection of Malicious Web Pages," Proc. 20th International World Wide Web Conference (WWW), 2011.

[8]. C. Whittaker, B. Ryner, M. Nazif, "Large-Scale Automatic Classification of Phishing Pages," Proc. 17th Network and the Distributed System Security Symp.2010.

[9]. www.google.com

[10]. www.internetworldstats.com/emarketing.html