IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Volume-3, Issue-9 E-ISSN: 2347-2693

A Survey on Detecting Suspicious and Malicious URLs in E-mail and Social Networks

Vispute Dhanashri[1*], Vispute Bhagyashri[2], Sonawane Monika[3], Nikam Seema[4]

[1*,2,3,4]*Department, Of Computer Engineering,*
Savitribai Phule Pune University, Maharashtra, India

www.ijcseonline.org

Received: Aug/21/2015 Revised: Aug/30/2015 Accepted: Sep/20/2015 Published: Sep/30/2015

***Abstract*—** These days, Email is also one of the advertising medium. Though it is a healthy medium for advertising, this is getting misused also. It gets really inconvenient to attend all those unnecessary emails. It is also very distracting. Here we are proposing a solution as email classifier. It will classify the inbox emails into various categories. A selected category of emails can be blocked considering it spam. In this study, the features of traditional heuristics and social networking are presented by combining them in feature set. This is done with Bayesian algorithm, know very helpful in such text classification tasks. The experimental result shows that the high detection rate is achieved by proposed approach. In this by using reduced feature set method we identify malicious URLs in email.

***Keywords*—**Social network, URL detection, Bayesian classification, Decision tree, Feature set extraction.

I. INTRODUCTION

The web server is better medium for a large number of malicious activities such as Spam attacks, Phishing attacks, DDos attacks and etc. motivated under financial aspects. These attacks attract the users to click on links attached in legitimate URLs and spam emails and make them to visit the malicious sites. Browser may not display the legitimacy Of the website which outlooks the phishing websites as legitimate. In some cases, the user also overrides the browsers decision [2].

Social Network Services (SNSs), such as Facebook, Twitter, and MySpace, have recently proliferated offering interactive information platforms that allow users to share and to interact. In other words, the convenience of SNSs facilitates potential cyber-attacks on SNS platforms. Malware programs often leverage short URL and blog services are often used by malware to disguise original URLs and evade security inspections, such as blacklist filtering applications. URL shortening transfers original URLs through shortened URLs by using redirection. Because URL shortening is often abused, these providers may find themselves blacklisted. The proposed detection method involves using a naïve Bayesian model to detect email that contain malicious URLs based on anomalies in the URL domain and unusual email contents behaviour's [1].

Now days, most of the machine learning and data mining techniques have been applied in spam email classification, such as Naive Bayes, Support Vector Machines (SVM) and rule learning. Decision tree is very popular and powerful tool in data mining community and the rules generated by decision tree are simple and accurate for most problems. In this study, a c4.5 classification method based decision tree and Bayesian classification is introduced to classify the spam email effectively [3].

Corresponding Author: Dhanashri Vispute ,*dhanashri1895@gmail.com*
Department of Computer Engineering, University of Pune, India

II. LITERATURE SURVEY

Zhang et al. proposed a content-based method for observing phishing web pages, saying that phishing sites are created based on minor modifications from the authenticated sites and show low page list ranks in the Google quest outcome. A set of heuristics was proposed based on domain name, lexical trademark of web links, and the HTML entry-content of websites. Five keywords were extracted from each web page based on TF-IDF (Term Frequency-Inverse Document Frequency) (phrase frequency/opposite record frequency) algorithm and the Google search were applied to verify the website legitimacy. According to McGrath et al., a brand name should appear in the URL of a web site. The URLs of phishing and non-phishing websites are collected and analysed, determining that different countries host phishing sites, in the registered countries phishing domains are rarely hosted, and phishing domains last approximately 3 days [1]. According to Xu et al., the behaviour of social network worm attacks and internet worm attacks are same; however, the methods of detecting internet worms are different so they do not apply to social networks. Detecting internet worms often involve observing abnormal behaviour in network traffic or the infected host; by contrast, on the social network sites each particular client host and the traffic or host activity cannot be monitored. In addition, infected clients might behave similarly to uninfected users, updating

personal account information, posting messages, or joining new groups. Therefore, detecting suspicious social network attacks is intensively challenging [1].

Blacklists may be in the form of IP addresses or websites used by email refined and block the actors through a feasible list of IP addresses or websites. Phish Net (Pawan et al 2010) potentiate existing blacklists by detecting related malicious URLs. Modification process of phishing URLs is insufficiently fast due to this one of the major problem with blacklists is that they fail to identify phishing URLs in the previous hours of a phishing [2].

Spam email has caused many problems such as wasting network bandwidth and taking recipient time. It is time consuming and if there are too many spam email in mailbox then it is difficult to remove spam email by hand. Thus, it has become very important to automatically classify the spam emails from legitimate emails. In machine learning community decision tree and altogether learning are two famous and powerful techniques. In this study, a decision tree and altogether learning introduced to organize the spam email effectively based on the novel classification [3].

Email spam is one of the major problems of the today's Internet, which brings damage to companies and to individual users. Among the approaches developed to stop spam, ltering is an important and popular one which contrasts of disparate ltering design. In this proposal we also provide the in-detailed description of other branches of anti-spam protection and the use of various approaches in commercial and non-commercial anti-spam software solutions [4].

III. PROPOSED SYSTEM

This study involved detecting suspicious URLs in social network environments. Based on relevant outfit and incident analysis, attackers may uses SNSs (Social Network Services) as vehicles, using user accounts to post messages that contain malicious URLs. Posts and feeds that are submitted on social networking sites are easily shared by the users, as they trust on their friends; and thus, they become the victims of social networking attacks. Therefore, Social networking heuristics should addressed the accessibility to identify malicious URLs in social networks in addition to the traditional attributes of the malicious URLs.

Certain malicious URLs are disguised using blogs or URL shortening service; which increases the trouble of detecting malicious posts and feeds. Malicious URLs used in social networks may make the use of trust and social relationships which correspond to phishing websites in spam; thus, multiple sets of features are proposed for detecting spam or malicious URLs. In this study, Facebook was used as the social networking environment and posts were collected using Facebook API.

Malicious URLs can be analysed based on the lexical sfeatures and host based features of the URL. The lexical feature analyses the structure of the URL. URLs contain the path and the host name. For example, consider 'www.annauniv.edu/emmrc/emmrc.html', the host name is www.annauniv.edu and emmrc/emmrc.html is the path. Host based features such as Page rank and age of domain, various lexical based features such as URL encoding, presence of malicious characters, hexadecimal character or malicious IP addresses to hide them and analysis of the word probabilities to find whether the email contains any doubtful links etc. are analysed in the proposed methodology.

It is helpful when illegitimate users hide their identities, pass authentication tests and during content analysis also it may get escaped by avoiding spam keywords. Some emails contain only malicious links instead of some information in it and it motivate the users to click on them which lead to fraudulent websites.

IV. SYSTEM ARCHITECTURE

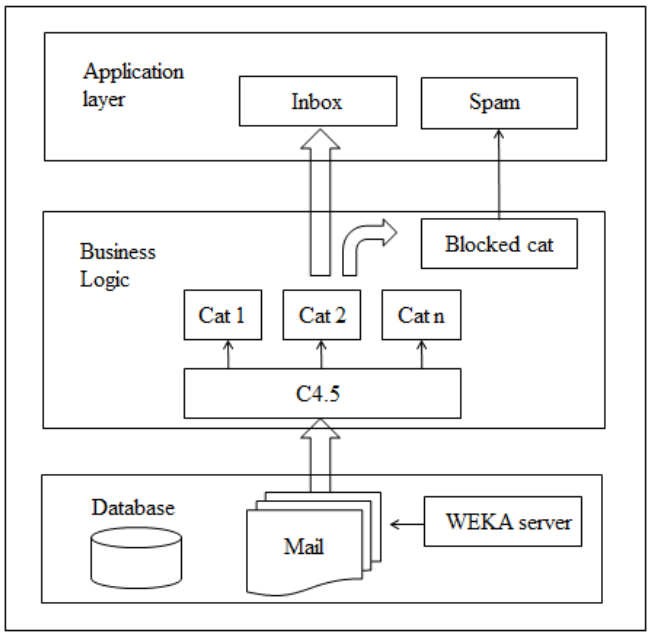

Figure 1. System Architecture

A. Lexical Features:

Lexical features analyses the URL format which includes the length of the host name, length of the URL, the number of dots, suspicious characters present such as @ symbol, hexadecimal characters and other special binary characters such as ('.', '=', '$', '^' and etc.) in the host or path name. To hide the actual URLs IP addresses and the hexadecimal characters are used.

B. Feature Selection:

1. Dash Count in Hostname:

A preliminary analysis on the blacklists indicated that numerous malicious URLs contain dashes, whereas dashes are rarely included in legitimate URLs. Therefore, the number of dashes in the host name was used as a lexical feature in this study.

2. Longest Domain Label:

Legitimate websites typically use meaningful, short, and easy-to-remember terms as domain names; which are compared with the domain names of websites, those of malicious websites are typically longer, and may not have meaningful terms. Therefore, feature extracting a term that shows the meaning of the website.

For example considered the URL, ''www.facebook.com'' having domain label as ''Facebook,'' which has a feature value equal to eight; thus, the length is ''Facebook'' is eight. The longest value is used to calculate the feature value among multiple URLs that are listed in a post.

3. Domain Rank:

Guan's study indicated that the Google search reputation or rank of a website produced strong classification results. Because the API returns a limited number of search results, a website was considered as normal if it is ranked within the first four search results; and when multiple URLs were listed in a post it has lowest rank.

4. Domain Age:

A normal domain typically has a long history and extended domain registration. Relevant studies have analyses that, the malicious URLs have domains registered at a future date, or lack registration dates. Thus, the domain age feature is considered, as most of the malicious domains are promptly taken down.

5. URL count:

Attackers may post multiple URLs in a post, to target the interests of users. Attackers compose to flood a wall, whereas the normal users do not compose posts that contain multiple URLs.

6. Similar Message Count from a User:

A virus may present in a restricted number of meaningful sentences, and it is possible that malicious messages involve similar content. In contrast to accounts or viruses, a normal user rarely posts similar content several times. Therefore, this feature represents unusual behavior. To compute the similarity of message content a fuzzy string comparison was used.

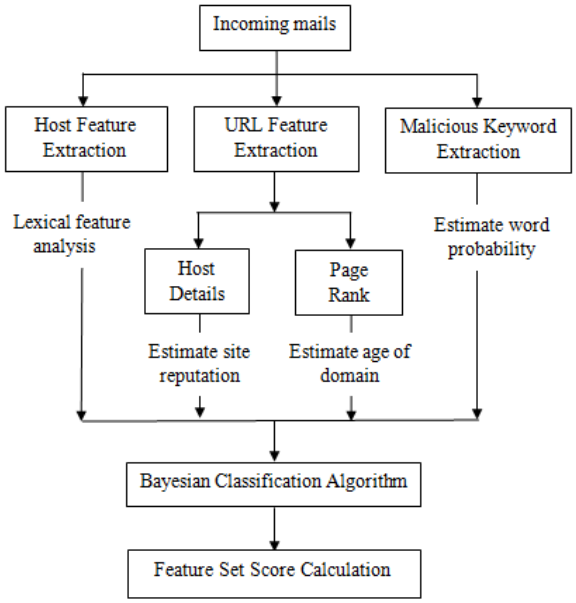

Figure: URL Extraction Process

7. Similar Message Count from Different Users:

Attackers may use multiple accounts to post suspicious posts and highly connected users might receive the same or similar information multiple times. Therefore, To count the number of similar messages that distinct user accounts post on a wall, this feature is used.

C. Bayesian Algorithm:

The Bayesian algorithm is a set of rules that uses data to change your beliefs. Naïve Bayes classifiers are a family of impels probabilistic classifiers in machine learning, based on Bayes' theorem with strong (naive) independence assumptions between the features. Popular (baseline) method for text categorization, judging documents is one of the problem which belongs to one category or the other (such as spam or legitimate, sports or politics, etc.) with word frequencies as the features. It is competitive in this domain with more advanced methods including support vector machines with appropriate preprocessing.

a. What is Bayes' theorem?

The conditional probability given that B has occurred, where A and B are two events in sample space is defined as follows;

$$P(A|B) = \frac{P(B-A)}{P(B)} \quad (1)$$

As long as P(B) ≠ 0. Here P(B ∩A) is the probability that both A and B occur and P(A|B) is the probability that A occurs only when B has occurred. Equation (1) is true, with A and B interchanged so that we also have P(B∩A) = P(B|A)P(A). By substituting this expression into (1) it then gives the following equation;

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad (2)$$

The definition written in this format is called as **Bayes' theorem**. It is a mathematical proposition which is accepted universally.

b. Algorithm Steps:-

1. On different values for a model parameter θ ∈Θ formulate your economic model as a collection of probability distributions condition.
2. Organize beliefs about θ into a probability distribution over Θ.
3. Then into the family of distributions collect and insert data given in step 1.
4. Calculate your new beliefs about θ using Bayes theorem.
5. Then criticize your model.

D. C4.5 algorithm:

It is an algorithm used to generate a decision tree. At each node of the tree, C4.5 select the attribute of the data that most effectively divides its set of samples into subsets enriched in one class or the other. The splitting criteria are nothing but the normalized information gain (difference in the entropy). To make the decision, attribute with highest normalized information gain is selected. Then C4.5 algorithm recurs on the smaller sub lists.

The decision tree is one of the important tool of decision-making theory. Decision tree is a classifier in tree like structure to show the process of reasoning. Each node in decision tree structures is called as a leaf node or a decision node. The values of the target attribute of instances are indicated by the leaf node.

The leaf node or a decision node indicates two or more branches and each branch shows values of the attribute to be tested. When we classify an unknown instance, they are routed down according to the values of the attributes in the successive nodes in the tree.

The most popular decision trees algorithm is C4.5. According to the strategy of the splitting nodes, C4.5 builds decision trees from training data set. C4.5 chooses one attribute that effectively splits the set of instances into subsets at each node of the tree. The C4.5 algorithm recursively visits each leaf node or decision node and split until no further splits are possible. Following Figure 3 shows the structure of decision tree built by the C4.5

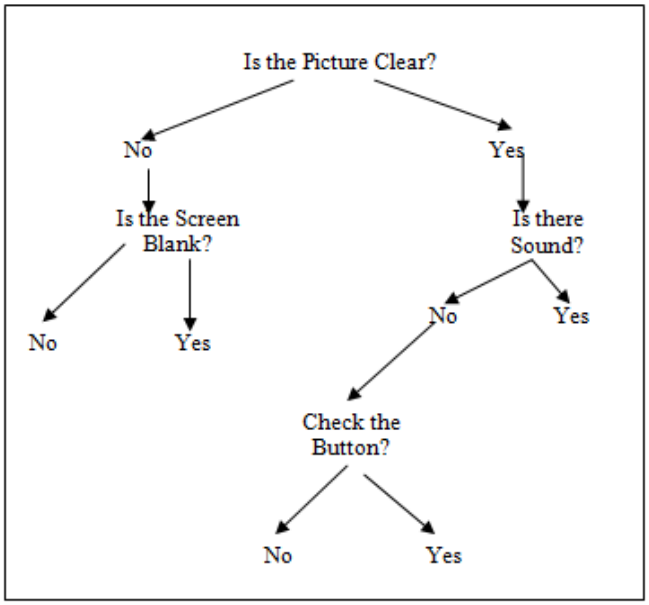

Figure 3: Example of Decision tree

In above figure, 'Is the Picture Clear?', 'Is the Screen blank?', 'Is there Sound?', 'Check the Button?' of the tree are condition attributes in inner nodes and Yes and No are the values of decision attribute in the dataset.

V. CONCLUSION

Blacklist mechanisms are not safe for blocking malicious URLs in social network environments. The proposed solution does not support on blacklist or whitelist mechanism, but rather it uses URL information and the social behaviour of users. The proposed detection method uses Bayesian classification indicated the efficient performance levels in various social network environments. By using this classification method we are checking each and every mail that is entering into the users account, and discarding it if found as malicious. Because of this system, there are very less chances of user to click mistakenly on that malicious link. Thus, detection mechanism should be able to identify these avoidance techniques.

REFERENCES

[1] Chia-Mei Chen, D.J. Guan, Qun-Kai Su, National Sun Yat-sen University, Kaohsiung, Taiwan, ROC.Feature set identification for detecting suspicious URLs using Bayesian classification in social networks, 133-147, **2014**.

[2] Dhanalakshmi ranganayakulu, Chellappan C, "Adhiparasakthi Engineering College, Melmaruvathur

603319, INDIA. Anna University, Chennai 600025, INDIA. Detecting malicious URLs in E-mail-An implementation, 125-131, **2013**

[3] Lei SHI, Qiang WANG, Xinming MA, Mei WENG, Hongbo QIAO, College of Information and Management Science, HeNan Agricultural University, Zhengzhou 450002,China.Spam Email Classification Using Decision Tree Ensemble, 949-956, **2012**

[4] Enrico Blanzieri University of Trento, Italy Anton Bryl, Italy Create-Net, Trento, Italy. A Survey of Learning-Based Techniques of Email Spam Filtering, 1-35, October **2007**.

[5] Xin Jin et.al "Social Spam Guard: A Data Mining Based Spam Detection System for Social Media Networks", *37thInternational Conference on Very Large Data Bases*, August 29th **2011**, Washington.

