Detecting Suspicious URLs using Bayesian Classification in OSN Data

Bargal Varsharani D. [1], Barhate Apeksha S. [2],Shewale Rekha V. [3], Suralkar Rupali H.[4]

Prof. S. A. Abhang

Department of Information Technology,

SRES COE, Maharashtra, India

ABSTRACT

The main concept behind this project is to develop Social network services (SNSs) are increasing popular. Using SNSs communicating with friends that can be used to share information with friends. There are so many unwanted web sites are cornerstone of internet. As a result, there has been broad interest for developing system to prevent the end user from visiting such sites. So is uses approach to prevent the system is based on URL classification, using statistical methods to discover lexical and host based properties of malicious web sites URLs. In this a feature set is presented that combines the features of traditional social networking. Further a suspicious URL identification system for use in social network environments is proposed based on Bayesian classification.

Keyword:- *SNSs Social Network Services, URLs Uniform Resource Locators, TLD Top Level Domain.*

1. INTRODUCTION

Most spam filters analyze the textual part of email messages, and these filters are content-based filters which use text classification. The spam filter algorithms are mostly based on the bag-of-words model. Handling new spam tactics is difficult and prone to high misclassification rate is a social attack that exploits the human using a given system, and therefore user awareness training programs fight against phishing attacks. Phishing attacks are the social engineering attack through which the victim is applicable to perform certain actions, such as submitting personal information directly to the phisher, or executing malware. The classification of junk emails and Bayesian classification has been widely used. The other well-known approaches are Support Vector Machine (SVM) and k-Nearest. Naive Bayesian method is simple and could be easily implemented as an incremental learning model. Moreover, Naive Bayesian requires linear training time whereas SVM requires quadratic training time and k Neighbour requires more testing time. Phishing URLs and domain names have very different lengths compared to other URLs and domain names in the Internet. Based on advances in information technology, websites offer various convenient web services such as information retrieval, chat rooms, Web 2.0-based services, blogs, albums, and multimedia sharing. 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.

Lexical features: These features allow us to capture the property that malicious URLs tend to “look different” from benign URLs. For example, the appearance of the token ‘.com’ in the URL ‘www.ebay.com’ is not unusual. However, the appearance of ‘.com’ in ‘www.ebay.com.phishy.biz’ or ‘phish.biz/www.ebay.com/ index.php’ could indicate an attempt by criminals to spoof the domain name of a legitimate commercial Web site. To implement these features, we use a bag-of-words representation of tokens in the URL, where ‘/’, ‘?’, ‘.’, ‘=’, ‘-’, and ‘ ’ are delimiters. We distinguish tokens that appear in the hostname, path, the top-level domain (TLD), primary domain name (the domain name given to a registrar), and last token of the path (to capture file extensions). Thus, ‘com’ in the TLD position of a URL would be a different token from ‘com’ in other parts of the URL. We also use the lengths of the hostname and the URL as features.

Host-based features: These features describe properties of the Web site host as identified by the hostname portion of the URL. They allow us to approximate “where” malicious sites are hosted, “who” own them, and “how” they are managed.

All users can access fan pages without requesting permission from the page owner. A websense survey indicated that 10% of URLs posted to in Facebook were malicious links; thus, users who access popular fan pages may risk security breaches. Hackers can spread attacks by simply posting messages containing malicious links on the most popular fan walls; multiple fans are likely to click on such links. Compared with spam, posts containing malicious URLs are faster and more effective. In this study, a set of heuristic features and Bayesian classification are proposed for detecting malicious URLs in SNSs. According to the findings, malicious web links in a post exhibit domain and social anomalies that differ from those of typical links. The proposed detection method involves using a naïve Bayesian model to detect social network posts that contain malicious URLs based on anomalies in the URL domain and unusual posting behavior.

2. LITERATURE SURVEY

Tina R. Patil, Dr.V. M. Thakare and Dr. S. S. Sherekar[1] In this paper Many content based on spam filters are rule based or trained online. Handling new spam technique is difficult and prone to high misclassification rate. To increment adaptive spam mail filtering use Nave Bayesian classification will give good performance, simplicity and adaptability is that phishing email messages contain URLs that point to phishing websites, and lexically analyzing the URLs can enhance the classification accuracy of email messages. This system removes oldest emails but keep the features to train new incoming emails. Lexical URL analysis is applied on incoming email after preprocessing. It detects and classify the host website and reports with email is ham or spam.

Aurangzeb Khan, Baharum Baharudin, Lam Hong Lee, Khair- ullah khan , Tronoh[2]. This paper Phishing emails are a real threat to internet communication and web economy. Attackers are trying to convince unsuspecting online users to reveal passwords, account numbers, social security numbers or other personal information. Filtering approaches using blacklists are not completely effective as new phishing spam is created. We investigate the statistical filtering of phishing emails, where a classifier is trained on characteristic features of existing e mails and is able to identify new phishing emails with different contents. Email features generated by adaptively trained on a publicly available test classifiers using these features are able to reduce the number of misclassified mails. Comparing to recently proposed more expressive evaluation method these results are statistically significant.

Andr e Bergholz, Gerhard Paas Frank Reichartz, Siehyun Strobel Jeong-Ho Chang Konan[3] This paper Phishing emails are a real threat to internet communication and web economy. Attackers are trying to convince unsuspecting online users to reveal passwords, account numbers, social security numbers or other personal information. Filtering approaches using blacklists are not completely effective as new phishing spam is created. We investigate the statistical filtering of phishing emails, where a classifier is trained on characteristic features of existing emails and is able to identify new phishing emails with different contents. Email features generated by adaptively trained on a publicly available test classifiers using these features are able to reduce the number of misclassified emails.

Justin Ma Lawrence K. Saul Stefan Savage Geoffrey M. Voelker [4] This system explores online learning approaches for detecting malicious websites (those involved in criminal scams) using lexical and host-based features of the associated URLs. This application is appropriate for online algorithms in case of twitter, google, facebook the size of the training data is larger than can be efficiently processed in batch and because the distribution of features that typify malicious URLs is changing continuously. Using a real-time system we developed for gathering URL features. Combined with a real-time source of labeled URLs from a large web mail provider. We demonstrate that recently developed online algorithms can be as accurate as batch techniques, achieving more classification accuracy over a balanced data set..

Justin Ma Lawrence K. Saul Stefan Savage Geoffrey M. Voelker [5] The malicious websites are the main internet criminal activities. There has been broad interest in developing systems to prevent the end user from visiting such sites. In this system, we describe an approach to this problem based on automated URL classification, using statistical methods to discover the lexical and host-based properties of malicious website URLs. These methods are able to learn highly predictive models by extracting and automatically analyzing tens of thousands of features potentially indicative of suspicious URLs. The resulting classifier obtains greater accuracy, detecting large numbers of malicious websites from their URLs, with modest false positives.

3. PROPOSED SYSTEM

It provides the Security to online social networking users. It is Easy use of social networking by avoiding unwanted URL. Proposed approach achieves a high detection rate. In Proposed System Bayesian Classification Algorithm Support Vector Machine (SVM) Algorithm used.

A. Proposed System Architecture:

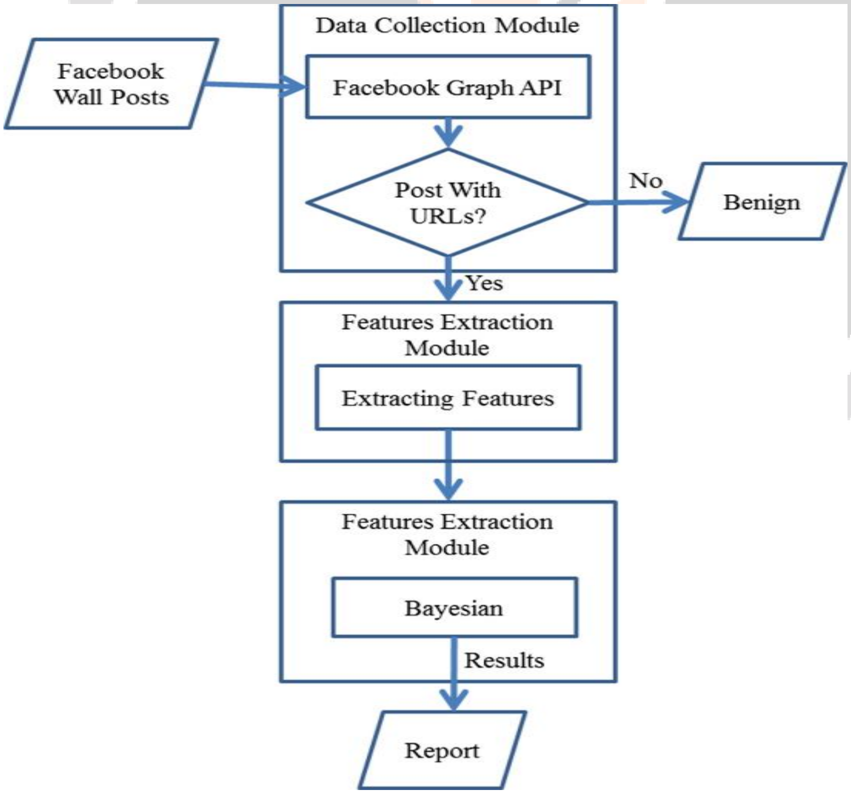

Figure 1: System Architecture

B . System Modules:

Proposed systemand overall process for detecting SNS-based malicious URLs.

1. Data Module: In the first module, data collection, posts are collected including time and content. Posts that lack URL information are considered benign.

2. Feature Extraction Module: In the second module, feature extraction, the proposed features elaborated in subsequent sections are retrieved and a feature vector is constructed for classification.

3. Bayesian Classification Module: In the third module, the Bayesian classification model, posts are classified based on a pertained classification model.

4. Feature selection Module:

(1) F1: Dash Count in Hostname:

It is supported the importance of lexical features in detecting malicious URLs. A preliminary analysis on he blacklists indicated that numerous malicious URLs contain dashes, whereas legitimate URLs rarely include dashes. Therefore, the number of dashes in the host name was used as a lexical feature in this study.

(2) F2: Longest Domain Label:

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

For example, the longest domain label of ‘‘www.facebook.com’’ is ‘‘facebook,’’ which has a feature value of eight thus, the length is ‘‘facebook’’ is eight. The longest value is used to compute the feature value when multiple URLs are listed in a post.

(3) **F3: Domain Rank:**

This indicated that the Google search reputation or rank of a website yields strong classification results. Because the API returns a limited number of search results, a website was considered normal if it ranked within the first four search results; the lowest rank was used when multiple URLs were listed in a post.

(4) F4: Domain Age:

A normal domain typically exhibits a long history and extended domain registration. Relevant studies have suggested malicious URLs exhibit domains registered at a future date, or lack registration dates. Most malicious domains are promptly taken down; thus, the domain age feature is considered.

(5)F5: URL count:

Attackers may post multiple URLs in a post, targeting the diversified vulnerabilities or interests of users. Attackers seek to flood a wall, whereas normal users do not typically compose posts that contain multiple URLs used a similar feature for detecting phishing; thus, the URL count is used a feature in the proposed detection method.

(6) F6: Similar Message Count from a User:

It indicated that news feeds posted by worm are automatically generated. A virus may present a limited number of meaningful sentences, and it is highly possible that malicious messages involve similar content. In contrast to compromised accounts or viruses, a normal user rarely posts similar content several times. Therefore, this feature represents anomalous behaviour. A fuzzy string comparison was used to compute the similarity of message content.

Similarity is defined as the length of the longest common subsequence (LCS) of two feeds compared with the average of their string lengths, as follows:

(7) F7: Similar Message Count from Different Users:

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

6. CONCLUSION

In this systemdata will be collected from Facebook. Various ratios of malicious to benign sample URLs will be selected to simulate various social network environments and will facilitate identifying malicious URLs in social networks. And accordingly security will be provided to the facebook users by avoiding unwanted URLs.

REFERENCES

1] Chia-Mei Chen , D.J. Guan, Qun-Kai Su member of Science Direct. Feature set identification for detecting suspicious URLs using Bayesian classification in social networks" Year 2014.

[2] Tina R. Patil, Dr. V. M. Thakare and Dr. S. S. Sherekar member of International Conference .A Combined Naive Bays and URL Analysis Based Adaptive Technique for Email Classification" year 2014.

[3] Aurangzeb Khan, Baharum Baharudin, Lam Hong Lee, Khairullah khan , Tronoh, Malaysia member of Science Engineering and Technology .A Review of Machine Learning Algorithms for text-Documents Classification" year 2010.

[4] Andre Bergholz, Gerhard Paas Frank Reichartz, Siehyun Strobel Jeong-Ho Chang Konan Technology member of International Conference .Improved Phishing Detection using Model-Based Features" year 2008.

[5] Justin Ma Lawrence K. Saul Stefan Savage Geoffrey M.Voelker member Computer Science Engineering \Identifying Suspicious URLs: An Application of Large-Scale Online Learning" year 2009.

[6] Justin Ma, Lawrence K. Saul, Stefan Savage, Geoffrey M. Voelker member of computer Science and engineering. Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs" year 2009.

[7] Ritesh Kumar Shital Ghadge G.S. Navale member of International conference .Spam Detection using Approach of Data Mining for Social Networking Sites " year 2014.

[8] Ugo Fiore a,n, Francesco Palmieri b, Aniello Castiglione c, Alfredo De Sant is member of science direct.Network anomaly detection with the restricted Boltzmann machine" year 2013.

[9] Shobeir Fakhraei James Foulds Madhusudana Shashanka member of International Conference. Collective Spammer Detection in Evolving Multi-Relational Social Networks" year 2015.

[10] D. Kevin McGrath, Minaxi Gupta is member of science direct \Behind Phishing: An Examination of Phisher Modi Operandi" year 2008.

BIOGRAPHIES

	Bargal Varsharani D. is pursuing B.E. Information Technology in SRES COE, Kopargaon. Her area of research interest include Data Mining.
	Barhate Apeksha S. is pursuing B.E. Information Technology in SRES COE, Kopargaon. Her area of research interest include Data Mining.
	Shewale Rekha V. is pursuing B.E. Information Technology in SRES COE, Kopargaon. Her area of research interest include Data Mining.
	Suralkar Rupali H. is pursuing B.E. Information Technology in SRES COE, Kopargaon. Her area of research interest include Data Mining.