The International Journal of Engineering and Science (IJES)
|| Volume || 6 || Issue || 4 || Pages || PP 21-31 || 2017 ||
ISSN (e): 2319 – 1813 ISSN (p): 2319 – 1805

Improving Phishing URL Detection Using Fuzzy Association Mining

[1.] Ms. S. Nivedha, [2]. Mr. S. Gokulan, [3.] Mr. C. Karthik, [4.] Mr.R.Gopinath, [5.] Mr.R.Gowshik

[1]*Assistant Professor, Department of Computer science and Engineering Sri Krishna College of Technology Coimbatore, India*
[2]*Department of Computer Science and Engineering Sri Krishna College of Technology Coimbatore, India*
[3]*Department of Computer Science and Engineering Sri Krishna College of Technology Coimbatore, India*
[4]*Department of Computer Science and Engineering Sri Krishna College of Technology Coimbatore, India*
[5]*Department of Computer Science and Engineering Sri Krishna College of Technology Coimbatore, India*

----------ABSTRACT----------

Phishing is the process to obtain sensitive information such as usernames, passwords, and credit card details by disguising as a trustworthy entity by the use of an electronic communication. Phishing attack continues to pose a solemn risk for web users and annoying threat within the field of electronic commerce. The Phishing detection using fuzzy and binary matrix construction method focuses on discerning the significant features that discriminate between legitimate and phishing URLs. The significant features are extracting the number of dots, length of the host etc., from each URL. These features are then subjected to associative rule mining-apriori and predictive apriori. The rules obtained are interpreted to emphasize the features that are more prevalent in phishing URLs. The key factors for the phished URLs are number of slashes in the URL, dot in the host portion of the URL and length of the URL. The pitfall of binary matrix method is the time complexity. So it impacts the overall speed of the system. The fuzzy based logic association rule mining algorithm was proposed to classify the legitimate and phishing URLs based on the features. The extracted features are converted to fuzzy membership values as "Low",' Medium' and "High". By applying association rule mining algorithm the rules are generated to detect the phishing URLs. The fuzzy based methodology provides efficient and high rate of phishing detection of URLs.

Keywords: *legitimate fuzzy pitfall*

Date of Submission: 07 March 2017 Date of Accepted: 10 April 2017

I. INTRODUCTION

Data mining is the computational process of discovering design in large data related to methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. Data mining plays a significant role in all fields that converts the data into valuable information. It is usually used in a wide range of profiling practices, such as scientific discovery, fraud detection, surveillance, and marketing. Data mining can be utilized to find out patterns in data but is often carried out only on samples of data. Consider if the samples in the data are good representation of the larger body of data then the data mining technique is effectual otherwise data mining process is ineffective. Web pages, commercial reports and scientific publications are available on Information Retrieval (IR) systems and the internet, high-quality document clustering participates more and more significant role in the applications such as web data management, web data mining, and filtering.

TASKS OF DATA MINING

1. **Classification** – Data mining classification arranges the data into predefined groups. For example, an email program might attempt to categorize an email as authorized or spam. Common algorithms include neural networks, naïve Bayesian Classification, decision tree learning and nearest neighbor.
2. **Clustering** – It is the classification of groups which are not predefined, so the algorithm will try to group similar items together.
3. **Regression** – Regression tries to discover a function which organizes the data with the minimum error.

4. **Association rule learning** – Association rule learning is a rule-based machine learning method for discovering relations between variables in large databases. It is intended to construct rules discovered in databases using some relations.

PHISHING

Phishing is a form of social engineering in which an attacker is known as a phisher. Phisher attempts to fraudulently acquire authorized users confidential or sensitive credentials by mimicking electronic communications from a trustworthy or public organization in an automated fashion. The word "phishing" appeared, when Internet scammers were using email lures to "fish" for passwords and financial information from the sea of Internet users; "ph" is a common hacker replacement of "f", which comes from the primary form of hacking, "phreaking" on telephone switches.

Early phishers copied the code from the AOL website and crafted pages that looked like they were a part of AOL, and sent takeoff emails or instant messages with a link to this fake web page, asking potential victims to reveal their passwords.

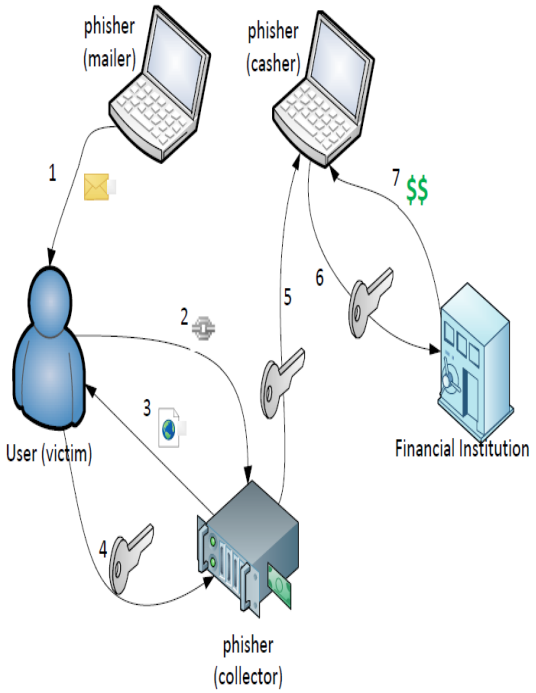


Figure 1.1: Process of Phishing

1. Mailers send out a large number of fraudulent emails which direct users to fraudulent websites.
2. Collectors set up fraudulent websites (usually hosted on compromised machines), which actively prompt users to provide confidential information.
3. The users provide the information to the fraudulent website.
4. The Collector sends the confidential information to the casher.
5. Cashers use the confidential information to achieve a pay-out.
6. Monetary exchanges often occur between those phishers.

A complete phishing attack involves three roles of phishers was shown in Figure 1.1 The latest statistics reveal that banks and financial institutions along with the social media and gaming sites continue to be the main focus of phishers. Some loyalty programs are also becoming favorite among phishers because with them phishers can not only opening the financial information of victim but also use existing reward points as currency. U.S. remains the largest host of phishing, accounting for 43% of phishing sites reported. A study of demographic factors suggests that women are more susceptible to phishing than men and users between the ages of eighteen and twenty five are more susceptible to phishing than other age groups. Phishing attacks that initially target general consumers are now evolving to include high profile targets, aiming to take intellectual property, corporate secrets, and excitable information caring national security.

TYPES OF PHISHING

Phishing has spread beyond email to include VOIP, SMS, instant messaging, social networking sites, and even multiplayer games. The categories of phishing are as follows.

Clone Phishing
Clone phishing creates a cloned email. User does this by getting information such as content and recipient addresses from a authorized email which was delivered previously, then user sends the same email with links replaced by malicious ones. User also employs address spoofing so that the email looks to be from the primary sender. The email can claim to be a re-send of the original or an updated version as a trapping strategy.

Spear Phishing
Spear phishing targets at a specific group. Instead of casting out thousands of emails randomly, spear phishers target selected groups of people with something in common, for example group from the same organization. Spear phishing is also being used against high-level marks, in a type of attack called "whaling".

Phone Phishing
Phone phishing refers to messages that demand to be from a bank asking users to dial a phone number paying attention to the problems with their bank accounts. Traditional phone equipment has dedicated lines, Voice over IP, being easy to manipulate, becomes a good choice for the phisher. Once the phone number, closely-held by the phisher and provided by a VoIP service, is dialed, voice prompts tell the caller to enter her account numbers and PIN. Caller ID spoofing, which is not impermissible by law, can be used along with this so that the call appears to be from a trusted source.

WEB SPOOFING
A phisher could forge a website that looks similar to a legitimate website, so that the users may think this is the genuine website and enter their passwords and personal information, which is collected by the phisher. Modern web browsers have certain built-in security signaling that can assist users from phishing scams, including domain name highlighting and https indicators and they are often neglected by careless users.

Process of Web Spoofing
Creating a forged website It's trivial to clone the look of a website by copying the front-end code; a little bit of web programming is necessary to redirect user's input into a file or database. Attracting traffic to forged website Once a forged website is online, the phisher must make potential users visit it.
Send spoofed emails with a link to the forged website.
Register a domain that is a common typo of a popular website. For example, register
Paypel.com and create a forged paypal.com.
Register the same domain name in a different TLD. Sometimes people will type in their country-specific TLD and expect to get a \localized" version of the website. For example, register gmail.com.cn and create a simplified-Chinese forged version of gmail.com.

Effectiveness of Browser Security Indicators And Https
Browser security indicators are not as effective as one might think. A survey report explains about 23% of participants used only the content of a webpage to determine legitimacy; an identical looking clone under any domain name without https is enough to deceive them. Many users cannot distinguish between a padlock icon in the browser chrome and a padlock icon as the favicon or in the page contents.
Relying on HTTPS is also not sufficient. Malware can install the public key of a phisher's CA to local computer's trusted root CA list, so that certificate signed by this CA would be trusted. When the phishing website is using a similar-looking domain that is registered by the phisher, a real certificate can be requested after domain ownership verification. CAs could be hacked to issue fraudulent certificates. Moreover, if a government is involved in phishing, it can order a CA under its control to issue a certificate for the phishing server.

Other Counter measures
Dynamic Security Skins seems to be a good method. The idea is that the website server create a unique abstract image for each user, and the web browser also independently computes the same image. The algorithm ensures that a phisher cannot predict this image. The user just needs to compare these two images; if both are identical, then the server is legitimate.

ASSOCIATION RULE MINING
Association rule mining is a procedure which is meant to find frequent design, correlations, associations, or causal structures from data sets found in different kinds of databases such as relational databases, transactional databases, and other forms of data repositories.

Association Rule Problem

By the introduction in last section, the formal statement of association rule mining problem was firstly stated. Let I=I1, I2, · · · , Im be a set of m distinct attributes, T be the transaction that include a set of items such that T ⊆ I, D be a database with various transaction records Ts. An association rule is an implication in the form of X->Y, where X, are sets of items called itemsets called antecedent while Y is called consequent, the rule means X implies Y. There are two important elemental measures for association rules, support(s) and confidence(c). Since the database is large and users care about only those often purchased items, usually thresholds of support and authority are predefined by users to drop those rules that are not so interesting or helpful. The two thresholds are called minimal support and minimal confidence respectively, additional restriction of interesting rules also can be nominal by the users. The two primary parameters of Association Rule Mining (ARM) are: support and confidence. Support(s) of an association rule is outlined as the percentage/fraction of records that include X U Y to the total amount of records in the database. The count for each item is enhanced by one every time the item is encountered in different dealings T in database D during the scanning process. It substance the support count does not take the amount of the item into account. For example in a transaction a customer buys three bottles of beers but we only increase the assist count number of {beer} by one, in another word if a transaction contains a item then the support count of this item is increased by one. Support(s) is calculated by the following formula

$$Support(XY) = \frac{Support\ count\ of\ XY}{Total\ number\ of\ transaction\ in\ D} \quad \textbf{… (1.1)}$$

From the definition, the item is a statistical significance of an association rule. Suppose the support of an item is 0.1%, it means only 0.1 percent of the dealings hold buying of this item. The retailer will not pay much attending to such kind of items that are not bought so frequently, apparently a high support is wanted for more interesting association rules. Before the mining process, users can undertake the minimal support as a threshold, which means they are only interested in definite association rules that are yield from those item sets whose supports exceed that threshold. Withal, sometimes even the item sets are not so frequent as defined by the threshold, the association rules return from them are still essential. For example in the supermarket some items are very costly, accordingly they are not purchased so often as the threshold needed, but association rules between those costly items are as important as other often bought items to the retailer.

Confidence of an association rule is defined as the percentage/fraction of the number of transactions that hold X*[*Y to the total number of records that hold X, where if the percentage exceeds the threshold of confidence an interesting association rule X*)*Y can be generated.

$$Confidence(X|Y) = \frac{Support(XY)}{Support(X)} \quad \textbf{…(1.2)}$$

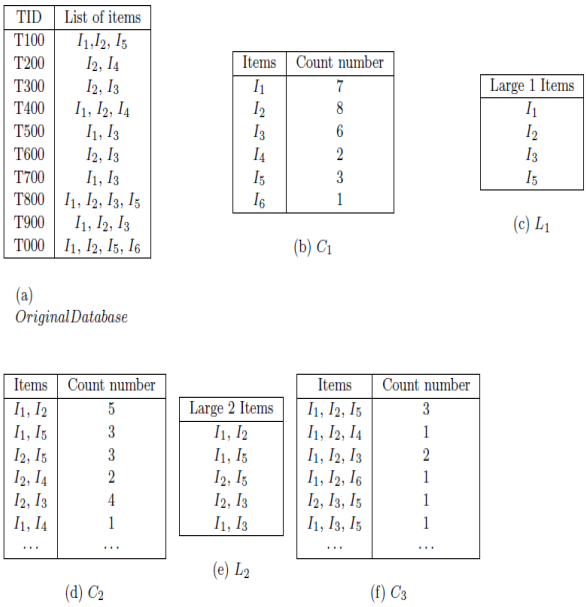
Confidence is a activity of strength of the association rules, imagine the confidence of the association rule X*)*Y is 80%, it way that 80% of the transactions that contain X also contain Y together, likewise to assure the interestingness of the rules mere minimum confidence is also predefined by users.

Association rule mining is to find out association regulation that fulfill the predefined minimum support and confidence from a given database. Suppose one of the large itemsets is *Lk*, *Lk= fI*1, *I*2, ¢ ¢ ¢, *Ik_j*1,*Ikg*, association rules with this itemsets are created in the following way: the first rule is *fI*1, *I*2, ¢ ¢ ¢, *Ik_j*1*g) fIkg*, by checking the confidence this rule can be determined as interesting or not. Then other rule are generated by deleting the last items in the antecedent and inserting it to the consequent, further the confidences of the new rules are checked to determine the interestingness of them. Those processes iterated until the antecedent becomes empty. Since the second subproblem is quite straight forward, most of the researches focus on the first subproblem.

Association Rule Mining Approaches

Association rule mining is a well explored research area, the introduction of basic and classic approaches for association rule mining. The second subproblem of ARM is straight forward, most of those approaches focus on the first subproblem. The first subproblem can be further divided into two subproblems: candidate large itemsets generation process and frequent itemsets generation process. The support of itemsets that exceeds the support threshold as large or frequent itemsets are taken, those itemsets that are expected or have the hope to be large or frequent are called candidate itemsets. The Apriori series approaches another milestone, tree structured approaches will be explained. Finally this section will end with some special issues of association rule mining,

including multiple level ARM, multiple dimension ARM, constraint based ARM and incremental ARM. In order to make it easier for us to compare those algorithms we use the same transaction database, a transaction database from a supermarket, to explain how those algorithms work. This database records the purchasing attributes of its customers.



TID	List of items
T100	I_1, I_2, I_5
T200	I_2, I_4
T300	I_2, I_3
T400	I_1, I_2, I_4
T500	I_1, I_3
T600	I_2, I_3
T700	I_1, I_3
T800	I_1, I_2, I_3, I_5
T900	I_1, I_2, I_3
T000	I_1, I_2, I_5, I_6

(a) *OriginalDatabase*

Items	Count number
I_1	7
I_2	8
I_3	6
I_4	2
I_5	3
I_6	1

(b) C_1

Large 1 Items
I_1
I_2
I_3
I_5

(c) L_1

Items	Count number
I_1, I_2	5
I_1, I_5	3
I_2, I_5	3
I_2, I_4	2
I_2, I_3	4
I_1, I_4	1
…	…

(d) C_2

Large 2 Items
I_1, I_2
I_1, I_5
I_2, I_5
I_2, I_3
I_1, I_3

(e) L_2

Items	Count number
I_1, I_2, I_5	3
I_1, I_2, I_4	1
I_1, I_2, I_3	2
I_1, I_2, I_6	1
I_2, I_3, I_5	1
I_1, I_3, I_5	1
…	…

(f) C_3

Figure 1.2: Support and Confidence value Computation using Association mining

OBJECTIVES

Features are extracted from the URL, then apriori algorithm is used to generate binary matrix method.Association rule mining is to find out association rules that satisfy the predefined minimum support and confidence from a given database. From the confidence values rules are generated. The rules are used to detect whether the given URL is phishing or not.

II. RELATED WORK

2.1 CANTINA: A Content-Based Approach to Detecting Phishing Web Sites

CANTINA, A Content-based approach proposed by Hong et al.[1] to detect the phishing websites based on the TF-IDF information retrieval algorithm. In this paper the design and evaluation of several heuristics are also discussed. It was developed to reduce the false positives. This experiments show that CANTINA is good at detecting phishing sites, correctly labeling approximately 95% of phishing sites.

Hong et al. [1] presented the design, implementation, and evaluation of CANTINA, 1 a novel content-based approach for detecting phishing web sites. CANTINA examines the content of a web page to determine whether it is legitimate or not, in contrast to other approaches that look at surface characteristics of a web page, for example the URL and its domain name. CANTINA makes use of the well-known TF-IDF algorithm used in information retrieval, and more specifically, the Robust Hyperlinks algorithm previously developed by Phelps and Wilensky for overcoming broken hyperlinks. The results show that CANTINA is quite good at detecting phishing sites, detecting 94-97% of phishing sites. It is shown that the CANTINA can be used in conjunction with heuristics used by other tools to reduce false positives, while lowering phish detection rates only slightly. A summary evaluation is presented, comparing CANTINA to two popular anti-phishing toolbars that are representative of the most effective tools for detecting phishing sites currently available. The experiments show that CANTINA has comparable or better performance to SpoofGuard with far fewer false positives, and does about as well as NetCraft. CANTINA combined with heuristics is effective at detecting phishing URLs in users' actual email, and that its most frequent mistake is labeling spam-related URLs as phishing.

ISSUES:

- CANTINA is used to detect the small scale versions of the websites.
- It does not support the larger data websites.

2.2 CANTINA+: a feature-rich machine learning framework for detecting phishing web sites

CANTINA+: a feature-rich machine learning framework proposed by Cranor et al. [2] that aims at exploiting the expressiveness of a rich set of features with machine learning to achieve a high True Positive rate (TP) on novel phish, and limiting the FP to a low level via filtering algorithms.
CANTINA+, the most comprehensive feature-based approach in the literature including eight novel features, which exploits the HTML Document Object Model (DOM), search engines and third party services with machine learning techniques to detect phish. They designed two filters to help reduce FP. The first is a near-duplicate phish detector that uses hashing to catch highly similar phish. The second is a login form filter, which directly classifies Web pages with no identified login form as legitimate. Finally CANTINA+ has been demonstrated to be a competitive anti-phishing solution.

ISSUES:
- It relies only on google search engine and the contents are downloaded from the web pages.
- The system prediction is exclusively based on querying search engine result.

2.3 A semi-supervised learning approach for detection of phishing web pages
A new phishing webpage detection approach proposed by Zhao et al. [3] based on a kind of semi-supervised learning method-transductive support vector machine (TSVM).
The features of web image are extracted for complementing the disadvantage of phishing detection only based on document object model (DOM); includes gray histogram, color histogram, and spatial relationship between sub graphs. The features of sensitive information are examined by using page analysis based on DOM objects. In contrast to the drawback of support vector machine (SVM) algorithm which simply trains classifier by learning little and poor representative labeled samples, this method introduces the TSVM to train classifier that it takes into account the distribution information implicitly embodied in the large quantity of the unlabeled samples, and have better performance than SVM.
Zhao et al. [3] extracted the features of web page image in addition to that the disadvantage of detection based on DOM objects and properties are complemented. The features include web image, DOM objects that reflect the characteristics of web pages absolutely. The tradition SVM algorithm trains labeled samples to form classifier that it cannot reflect the features of samples set totally, instead of SVM, the TSVM is utilized to classify the features vectors of page that it takes into account the distribution information implicitly embodied in the large quantity of the unlabeled examples and have better performance than SVM. As an independent scheme, the experiments show that the method is effective at detecting phishing sites, and can achieve higher accuracy of phishing pages detection.

ISSUES:
- The detection rate of this method is a little lower.
- It relies only on google search engine and the contents are downloaded from the web pages.

2.4 A Multi-tier phishing detection and filtering approach.
The system proposed by Islam et al. [4] is a new approach called multi-tier classification model for phishing email filtering. It is an innovative method for extracting the features of phishing email based on weighting of message content and message header and select the features according to the priority ranking and examined the impact of rescheduling the classifier algorithms in a multi-tier classification process to find out the optimum scheduling. The empirical evidence is that this approach reduces the false positive problems substantially with lower complexity.

ISSUES:
- It is an open challenge to develop a robust malware detection method retaining accuracy for future phishing emails.
- Feature retrieval is inefficient.

2.5 Assessing the severity of phishing attacks: a hybrid data mining approach.
The system proposed by Guo et al. [5] assessed the severity of phishing attacks in terms of the risk levels and the potential loss in market value suffered by the targeted firms. The supervised classification techniques are used, which is a major stream of data mining, to assess the severity of phishing attacks. At the same time, the key antecedents that contribute to a high risk level or a high financial loss generation by a phishing attack are identified. Guo et al. [5] used a hybrid approach which combines key phrase extraction and supervised classification methods that make use of the textual data description of the phishing attack with the financial data of the targeted company to assess the severity of a phishing attack according to its risk level or financial loss generating potential.

The neural network (NN), decision tree (DT), and support vector machine (SVM) are used to classify the risk levels. The three classifiers have different characteristics. NN consists of three interconnected layers, namely, input layer, hidden layer, and output layer. Each layer contains interconnected nodes than can process the data. The results show that the key identifying variables for risk level and potential financial loss of phishing attacks are different from each other. High risk level is associated with phishing emails that ask users of large firms to update the accounts whereas high financial loss is characterized by phishing attacks targeted to users of large firms that have high total liabilities.

ISSUES:
- This approach supported only for equal misclassification cost. Researchers can't conduct the experiments using unequal misclassification cost.
- The impact of misclassifying a high risk or high CAR phishing attack can be quite severe.

2.6 Soft computing based imputation and hybrid data and text mining: the case of predicting the severity of phishing alerts by Nishanth KJ, Ravi V, Ankaiah N, Bose I.

Nishanth et al. employed a novel two-stage soft computing approach for data imputation to assess the severity of phishing attacks. The imputation method involves K-means algorithm and multilayer perceptron (MLP) working in tandem. The hybrid is applied to replace the missing values of financial data which is used for predicting the severity of phishing attacks in financial firms. After imputing the missing values, mine the financial data related to the firms along with the structured form of the textual data using multilayer perceptron (MLP), probabilistic neural network (PNN) and decision trees (DT) separately.

First, replace the missing values in the financial data using the soft computing based data imputation approach. Then, apply text mining on the textual (unstructured) data of phishing alerts. Thus, textual data is converted into structured data. Finally, predict the risk level of phishing attacks using the combined financial data from the financial statement of the companies and textual data using MLP and DT separately. The overall classification accuracies for the three risk levels of phishing attacks using the classifiers MLP, PNN, and DT are superior.

Issues:
- For financial data alone the overall accuracy using PNN is not the best.
- Accuracy is varied with different risk levels.

2.7 Visual-similarity-based phishing detection by Medvet E, Kirda E, Kruegel C.

Kruegel et al. presented an effective approach to detect phishing attempts by comparing the visual similarity between a suspected phishing page and the legitimate site that is spoofed. When the two pages are "too" similar, a phishing warning is raised. In this system, they consider three features to determine page similarity: text pieces (including their style-related features), images embedded in the page, and the overall visual appearance of the page as seen by the user (after the browser has rendered it). They quantified the similarity between the target and the legitimate page by comparing these features, computing a single similarity score.

Kruegel et al. chose to perform a comparison based on page features that are visually perceived. This is because phishing pages mimic the look-and-feel of a legitimate site and aim to convince the victims that the site they are visiting is the one they are familiar with. Once trust is established based on visual similarity, there is a higher chance that the victim will provide her confidential information. Typically, a victim's visual attention focuses both on the global appearance of the page and on salient details such as logos, buttons, and labels. And proposed a novel comparison technique that eliminates the shortcomings of Anti-Phish and DOM Anti-Phish. This solution can be used together with these tools, but can also be integrated into any other anti-phishing system that can provide a list of legitimate sites that can be potential targets of phishing attempts.

Issues:
- DOM Anti-Phish is ineffective against phishing pages that use mostly images.
- These phishing attempts were not detected in all websites.

2.8 Detecting phishing web pages with visual similarity assessment based on earth mover's distance by Fu AY, Wenyin L, Deng X.

Deng et al. proposed an effective approach for detecting phishing Web pages, which employs the Earth Mover's Distance (EMD) to calculate the visual similarity of Web pages. The most important reason that Internet users could become phishing victims is that phishing Web pages always have high visual similarity with the real Web pages, such as visually similar block layouts, dominant colors, images, and fonts, etc. They follow the anti phishing strategy to obtain suspected Web pages, which are supposed to be collected from URLs in those e-mails containing keywords associated with protected Web pages. First convert them into normalized images and

then represent their image signatures with features composed of dominant color category and its corresponding centroid coordinate to calculate the visual similarity of two Web pages.
The linear programming algorithm for EMD is applied to visual similarity computation of the two signatures. An anti phishing system may be requested to protect many Web pages. A threshold is calculated for each protected Web page using supervised training. If the EMD-based visual similarity of a Web page exceeds the threshold of a protected Web page, they classified the Web page as a phishing one. Large-scale experiments have been carried out. Our final experiments show high classification precision, high phishing recall, and satisfactory time performance.

Issues:
- This method could not detect the phishing Web pages which are visually not similar to their attacking targets.
- This approach did not consider using textual features together with the visual features to overcome above problem.

2.9 Fighting phishing with discriminative key point features of web pages by Chen KT, Chen JY, Huang CR, Chen JY.
Phishing is a form of online identity theft associated with both social engineering and technical subterfuge. Specifically, phishers attempt to trick Internet users into revealing sensitive or private information, such as their bank account and credit card numbers. Chen et al. presented an effective image-based anti-phishing scheme based on discriminative key point features in web pages. They use an invariant content descriptor, the Contrast Context Histogram (CCH), to compute the similarity degree between suspicious pages and authentic pages.
First take a snapshot of a suspect webpage and treat it as an image in the remainder of the detection process. They use the Contrast Context Histogram (CCH) descriptors proposed to capture the invariant information around discriminative key points on the suspect page. The descriptors are then matched with those of the authentic pages of the protected domains, which are stored in a database compiled by users and authoritative organizations, such as the Anti-Phishing Working Group (APWG). The matching of CCH descriptors yields a similarity degree for a suspect page and an authentic page. Finally, use the similarity degree between two pages to determine whether the suspect page is a counterfeit or not. If the similarity degree between a phishing page and an authentic page is greater than a certain threshold, the suspect page is considered as a phishing page of the authentic page, and considered genuine if it is not a phishing page of any of the authentic pages in the database. The results show that the proposed scheme achieves high accuracy and low error rates.

Issues:
- This approach is susceptible to significant changes in the webpage's aspect ratio and colors used.

2.10 Phishing detection based associative classification data mining by Abdelhamid N, Ayesh A, Thabtah F.
Thabtah et al. investigated the problem of phishing detection using AC approach in data mining. They primarily test a developed AC algorithm called MCAC and compare it with other AC and rule induction algorithms on phishing data. The phishing data have been collected from the Phish tank archive, which is a free community site. In contrast, the legitimate websites were collected from yahoo directory. Thabtah et al showed that MCAC is able to extract rules representing correlations among website's features. These rules are then employed to guess the type of the website.
The novelty of MCAC is its ability not only to discover one class per rule, but rather a set of classes bringing up the classifier performance in regards to accuracy. This is unlike current AC algorithms that only generate a single class per rule. Thus, the new classes connected with the rules that have been revealed by MCAC correspond to new knowledge missed by the majority of the existing AC algorithms. This method outperformed the considered methods on detecting phishing with respect to accuracy rate. Further, the label-weight and any-label results of MCAC are better than those of the MMAC for the same phishing data. More importantly, MCAC was able to produce multi-label rules from the phishing data generating rules associated with a new class called ''Suspicious'' that was not originally in the training data set.

Issues:
- The rule used in this is based on human experience rather than intelligent data mining technique.
- This approach did not consider content based features to increase the collected features.

The survey focuses on discerning the significant features that discriminate between legitimate and phishing URLs. The foremost objective of this system is to identify an URL that is provided as input as a phished URL or not. This method consists of two phases URL search phase and the feature extraction phase. In the URL search phase, once the user accesses/requests an URL, a search is carried out to check whether the given URL is in the

repository of legitimate URLs. If a match is found in the repository, then the URL is considered to be a legitimate URL. In the Feature extraction phase, heuristics are defined to extract 14 features from the URL and are subjected to association rule mining to determine the legitimate and phished URL. These features are then subjected to associative rule mining, apriori and predictive apriori. The rules obtained are interpreted to emphasize the features that are more prevalent in phishing URLs. The main issues are User defined threshold, Frequent updation is needed in domain knowledge, Very low accuracy

III. SYSTEM ARCHITECTURE

The Construction of the binary matrix with the user provided threshold is difficult.Use of Apriori algorithm in binary Matrix Conversion may lead to wrong decision of phishing URL detection.

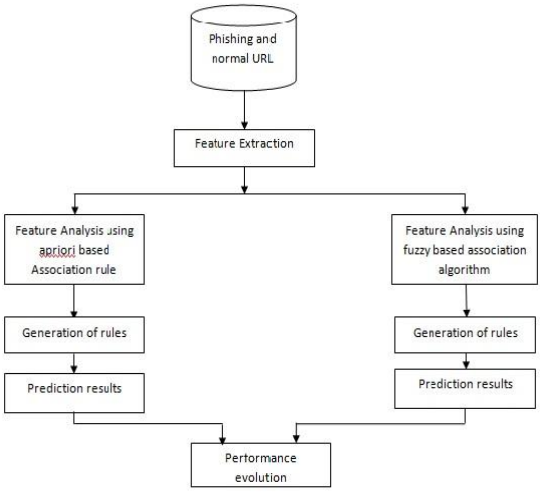


Figure 3.1: System Architecture

PROPOSED WORK

The Fuzzy logic based Association rule mining algorithm classifies the legitimate and phishing URLs based on the features that are converted to membership values. The extracted features are converted to fuzzy membership values as "Low",' Medium' and "High. Then the association rule mining algorithm is applied to generate the rules to detect the phishing URLs as shown in Figure 3.1. Features are extracted from the URL, and then apriori algorithm is used to generate binary matrix method. Association rule mining is to find out association rules that satisfy the predefined minimum support and confidence from a given database. From the confidence values rules are generated. The rules are used to detect whether the given URL is phishing or not.

Advantages

- Greater computation speed and memory usage.
- Reduces the Space and Time complexity of the overall system.

Feature Extraction

It focuses on identifying the relevant features that differentiate phishing websites from legitimate websites and then subjecting them to association rule mining. In order to identify the relevant features, certain statistical investigations and analysis were carried out on the phish tank and legitimate dataset. Based on the heuristics, fourteen features were defined and are subjected to association rule mining to effectively determine the legitimate and phished URL. The features are length of the host URL, number of slashes in URL, dots in host name of the URL, number of terms in the host name of the URL, special characters, IP address, Unicode in URL, transport layer security, Sub domain, certain keyword in the URL, top level domain, number of dots in the path of the URL, hyphen in the host name of the URL and URL length.

The heuristic is defined as

$$H = \begin{cases} if(Featurename) > averagefeatureval \\ else, LegitimateURL \end{cases}$$

… (3.1)

Input : Legitimate or Phishing URL.
Output : Different Features extracted from the input URL.

Apriori Algorithm Based Association Rule Mining

Data mining is the method that tries to get patterns in massive information sets. The overall objective of data mining method is to extract information from data set and remodel it into comprehensible structure. The objective of the association rule mining is used to discover associations among items in a set, by mining knowledge from the database. Support and confidence techniques are used to assess the association rules. Support is the proportion of transactions wherever the rule holds. Confidence is the conditional probability of C with reference to A or, in different words, the relative cardinality of C with reference to A.

Apriori is an important algorithm for mining frequent itemsets. It uses past information of frequent item set properties. To select fascinating rules from the set of possible rules, constraints on numerous measures of significance and interest are used. Support and confidence are the measures of rule that replicate the quality and certainty of a rule. Predictive apriori exploits the accuracy of association rules, on hidden data. Apriori grades the rules with respect to confidence alone but predictive apriori deliberates the confidence and support together in ranking the rules. The support and confidence is joined in a single measure called accuracy.

$$Supp\,(X) = \frac{Numberof\ \times Xappears}{N} = P(X)$$

..(3.2)

$$Supp\,(XY) = \frac{Numberof\ \times X \wedge Yappearstogether}{N} = P(X \cap Y)$$

..(3.3)

.. (3.4)

{S **.. (3.5)**

The process of identifying the type of a URL is generated using association rules in which the different heuristics are utilized to acquire unknown knowledge. It is used to ascertain the URL type when a user accesses it. The experiments have been performed with apriori and predictive apriori rule generation algorithms. The experiment is done to discover the rules based on phishing URLs.

Rule Extracted From Apriori

Association rules play a major role in finding interesting patterns. Association rules deliver information within the kind of "if–then" statements. The rules are computed from the information and are probabilistic in nature. All the attributes selected in the data set are binary attributes and only the phishing URLs are mined using the apriori algorithm for identifying the recurring patterns. The strong rule generated by the apriori with 100 % confidence alone has been considered for further analysis and the other rules are discarded.

Input: Extracted features

Output: Support and confident values for each feature.

4)Fuzzy Based Association Rule Mining

Discovering Fuzzy Sets

The traditional way to discover the fuzzy sets needed for a certain data set is to consult a domain expert who will define the sets and their membership functions. This requires access to domain knowledge which can be difficult or expensive to acquire. In order to make an automatic discovery of fuzzy sets possible, an approach has been developed which generates fuzzy sets automatically by clustering. The whole process of automatically discovering fuzzy sets can be subdivided into four steps:

1. Transform the database to make clustering possible (the value of all the attributes has to be positive integer)
2. Find the clusters of the transformed database using a clustering method.
3. For each quantitative attribute, fuzzy sets are constructed using the medoids.
4. Generate the associated membership functions.

Generate membership functions (triangular function):

$$f_{ij}\left(x: min_j, a\frac{k}{2}j, max_j\right) = \begin{cases} 1, if\, a\frac{k}{2}j - x \\ \frac{x - min_j}{a\frac{k}{2}j - min_j}, if\, min_j \le x \le a\frac{k}{2}j \\ \frac{max_j - x}{max_j - a\left(\frac{k}{2}\right)j}, if\, a\left(\frac{k}{2}\right)j \le x \le max_j \\ 0, otherwise \end{cases} f_{ij}\left(x: min_j, a\frac{k}{2}j, max_j\right)$$

$$= \begin{cases} 1, if\, a\frac{k}{2}j - x \\ \frac{x - min_j}{a\frac{k}{2}j - min_j}, if\, min_j \le x \le a\frac{k}{2}j \\ \frac{max_j - x}{max_j - a\left(\frac{k}{2}\right)j}, if\, a\left(\frac{k}{2}\right)j \le x \le max_j \\ 0, otherwise \end{cases} \quad ..(3.6)$$

The algorithm first searches the database and returns the complete set containing all attributes of the database. In a second step, a transformed fuzzy database is created from the original one. The user has to define the sets to which the items in the original database will be mapped. The frequent item sets F will be created from the candidate item sets C. New candidates are being generated from the old ones in a subsequent step. Ck is generated from C_{k-1} as described for the Apriori algorithm in step 1. The following pruning step deletes all item sets of C_k if any of its subsets does not appear in C_{k-1}.

Input: Support and confident values

Output: Fuzzy based values.

The Fuzzy value classification is computed based on apriori rule which are generated from the features extracted. The next chapter deals with the implementation details and the functions of each module.

PERFORMANCE ANALYSIS

In this section, the performance evaluation of this work is done to prove the performance improvement over the proposed methodology than the existing system in terms of time, memory usage and quality of generated rules with maximizing the fitness function.

The proposed application concentrates on the fuzzy value classification based on the apriori rule applied.

IV. CONCLUSION

The features are extracted from the URL and the apriori algorithm is used to generate the binary matrix. Fuzzy based association rules are generated based on confident values are generated from the URLs. After the Feature Extraction and Apriori algorithm based Association Rule Mining, the Fuzzy based Association Rule Mining where the support and confidence values are compared to all the features extracted. Based on the confident value the association rules are applied and the rules are generated. After the association rule applied, the performance evolution is done for the association rule applied and binary matrix generated values. The upcoming developments that could be expected is Fuzzy based association rule mining.

REFERENCES

[1]. Abdelhamid N, Ayesh A, Thabtah F (2014) Phishing detection based associative classification data mining. Science-Direct 41:5948–5959.

[2]. Chen KT, Chen JY, Huang CR, Chen JY (2009) Fighting phishing with discriminative key point features of webpages. IEEE Internet Comput 13:56–63.

[3]. Chen X, Bose I, Leung ACM, Guo C (2011) Assessing the severity of phishing attacks: a hybrid data mining approach. Expert Syst Appl 50:662–672.

[4]. Fu AY, Wenyin L, Deng X (2006) Detecting phishing web pages with visual similarity assessment based on earth mover's distance. IEEE Trans Dependable Secure Comput 3(4):301 321.

[5]. Islam R, Abawajy J (2013) A multi-tier phishing detection and filtering approach. J Netw Comput Appl 36:324–335.

[6]. Li Y, Xiao R, Feng J, Zhao L (2013) A semi-supervised learning approach for detection of phishing webpages. Optik 124:6027–6033.

[7]. Nishanth KJ, Ravi V, Ankaiah N, Bose I (2012) Soft computing based imputation and hybrid data and text mining: the case of predicting the severity of phishing alerts. Expert Syst Appl 39:10583–10589.

[8]. Medvet E, Kirda E, Kruegel C (2008) Visual-similarity-based phishing detection. SecureComm. In: Proceedings of the 4th international conference on Security and privacy in communication netowrks. pp 22–25.

[9]. Xiang G, Hong J, Rose CP, Cranor L (2011) CANTINA+: a feature-rich machine learning framework for detecting phishing web sites. ACM Trans Inf Syst Secur 14:21.

[10]. Zhang Y, Hong JI, Cranor LF (2007) CANTINA: a content based approach to detecting phishing web sites. In: Proceedings of the 16th international conference on world wide web, Banff, p 639–648.