Jeeva and Rajsingh *Hum. Cent. Comput. Inf. Sci.* (2016) 6:10
DOI 10.1186/s13673-016-0064-3

Human-centric Computing and Information Sciences

RESEARCH Open Access

Intelligent phishing url detection using association rule mining

S. Carolin Jeeva[1*] and Elijah Blessing Rajsingh[2]

*Correspondence:
caroljeeva@gmail.com
[1] Department of Computer Applications, Karunya University, Coimbatore, India
Full list of author information is available at the end of the article

Abstract

Phishing is an online criminal act that occurs when a malicious webpage impersonates as legitimate webpage so as to acquire sensitive information from the user. Phishing attack continues to pose a serious risk for web users and annoying threat within the field of electronic commerce. This paper focuses on discerning the significant features that discriminate between legitimate and phishing URLs. 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. Analyzing the knowledge accessible on phishing URL and considering confidence as an indicator, the features like transport layer security, unavailability of the top level domain in the URL and keyword within the path portion of the URL were found to be sensible indicators for phishing URL. In addition to this number of slashes in the URL, dot in the host portion of the URL and length of the URL are also the key factors for phishing URL.

Keywords: Phishing, Web security, Association rule mining

Background

Phishing is a malicious website that impersonates as a legitimate one to get sensitive data like credit card number or bank account password. A phisher uses social engineering and technical deception to fetch private information from the web user. The phishing web pages generally have alike page layouts, blocks and fonts to mimic legitimate web pages in an endeavor to influence web users to obtain personal details such as username and password. Over the last few years, online banking has become very popular as more financial institutions have begun to offer free online services. With the increase in online theft, financial crimes have changed from direct attacks to indirect attack. Phishing [1] is a quickly growing type of fraud and is taken into account as one of the foremost dangerous threats within the web which cause folks to mislay guarantee [2] in on-line transactions. It is relatively a current web crime as compared with virus, hacking and remains an ominous threat to client and business round the world.

According to the RSA's online fraud report [3], the year 2013 has been confirmed to be a record year where many phishing attacks have been launched globally. Additional1y, RSA estimates that over USD $5.9 billion was lost by global organizations due to phishing attacks at the same period. The Internet Security Threat Report 2014 [4] reports that cybercrimes are prevailing and damaging threats from cybercriminals still emerge over

businesses and customers. According to RSA monthly fraud report January 2014, the [5] big data analytics and broader intelligence will lead to faster detection resulting in lower financial losses. Data mining techniques are used to extract helpful information by analyzing the past information then predicting the future incidents.

In this paper, the rules are generated using association rule mining to detect phishing URL. The remaining section in the papers is organised as follows: The outline of literature survey is shown in second section. The system architecture is illustrated in third section. The features that are generated from the URL are discussed in fourth section. Fifth section explains the methodology used in detecting phishing URL. Sixth section presents an association rule mining technique to discover the rules concerning phishing URL and in the last section conclusions are presented.

Related work

Phishing is a major danger to web users. The fast growth and progress of phishing techniques create an enormous challenge in web security. Zhang et al. [6] proposed CANTINA, a completely unique HTML content method for identifying phishing websites. It inspects the source code of a webpage and makes use of TF-IDF to find the utmost ranking keywords. The keywords obtained are given as input to google search engine and examined whether the domain name of the URL matches with N top search result and is considered as legitimate. This approach fully relies on google search engine. CANTINA+ proposed by Xiang et al. [7] is an upgraded version of CANTINA, in which new features are included to achieve better results. In particular, the authors include the HTML Document Object Model, third party and google search engines with machine learning technique to identify phishing web pages.

Huang et al. [8] proposed SVM based technique to detect phishing URL. The features used are structural, lexical and brand names that exist in the URL. Liebana-Cabanillas et al. [9] proposed completely different technique to search out the variables that are most often utilized in financial institutions so as to predict the trust among electronic banking. Yuancheng et al. [10] proposed semi supervised based method for detection of phishing web page. The features of the web image and DOM properties are considered. Transductive Support Vector Machine is applied to detect and classify phishing web pages. Islam et al. [11] proposed filtering phishing email with the message content and header using multi-tier classification model.

Chen et al. [12] have proposed a hybrid approach that mixes extraction of key phrase, textual, financial data to ascertain the vicious of phishing attack using supervised classification strategies. Nishanth et al. [13] have proposed a method in which the structured style of the financial data are mined using machine learning algorithms. Liu et al. [14] have proposed the visual approach to identify phishing web pages. The similarity between the pages is assessed by block, layout and overall style. Medvet et al. [15] also adopted the visual similarity between webpages to calculate the similarity among a legitimate site and the suspected phishing website. The features used to verify page similarity are text piece, their style, images and the overall appearance of the webpage.

Antony et al. [16] have proposed a technique that uses EMD to decide the resemblance of webpage. In this methodology, the webpages are converted into images and the features like color and coordinate are used to generate signatures. The distance of

the webpage image signature is computed using EMD. The authors use a trained EMD threshold value to differentiate the legitimate and phishing webpages. Lam et al. [17] have proposed an image based approach for detecting phishing webpages. The authentic and the suspected pages are transformed into black and white image. The size and location of each blob are recorded and compared. The matched pair is selected by comparing the block pair with maximum similarity degree. The classifier categorizes the page based on the similarity score. Chen et al. [18] have proposed an approach that uses CCH to figure out the resemblance degree between fake and legitimate page.

Pshark is an approach proposed by Shah et al. [19] to detect and eliminate the identified phishing web page from host server. The WHOIS database is used to retrieve the information about the page. A notification is sent to the host server intimating that a phishing page resides in that server. He et al. [20] adopted heuristic method to categorize the legitimacy of the web page. The heuristic used in this system are based on the term identity set of a webpage. Aburrous et al. [21] have proposed fuzzy data mining technique to identify the phishing website. Zhang et al. [22] adopted domain feature enhanced classification model for the detection of Chinese phishing e- business websites.

Zhang et al. [10] proposed CANTINA, a completely unique HTML content method for identifying phishing websites. It inspects the source code of a webpage and makes use of TF-IDF to find the utmost ranking keywords. The keywords obtained are given as input to google search engine and examined whether the domain name of the URL matches with N top search result and is considered as legitimate. This approach fully relies on google search engine. CANTINA+ proposed by Xiang et al. [7] is an upgraded version of CANTINA, in which new features are included to achieve better results. In particular, the authors include the HTML Document Object Model, third party and google search engines with machine learning technique to identify phishing web pages. However, both the approaches rely on google search engine and the contents are downloaded from the webpages. But in our work the features related to URL is considered and thus downloading the content of the webpage is avoided. Moreover the system prediction is not exclusively based on querying search engine result.

Huang et al. [8] proposed SVM based technique to detect phishing URL. The features used are structural, lexical and brand names that exist in the URL. However, more features related to URL are considered in the proposed work. Neda et al. [23] proposed rule based classification algorithm to detect phishing URL. However the rule used in this is based on human experience rather than intelligent data mining technique. In the approach proposed by Han et al. [24] the system warns the user, when the user submits the username and password for the first time, although the current website is a legitimate website. This is because the information about the legitimate website is not maintained. This login problem is eliminated in our system as a repository of white list is effectively maintained.

System architecture for prediction of phishing URL

Figure 1 shows the system architecture for detecting phishing URL. The foremost objective of this system is to identify an URL that is provided as input as a phished URL or not. The proposed method consists of two phases (1) URL search phase and (2) feature

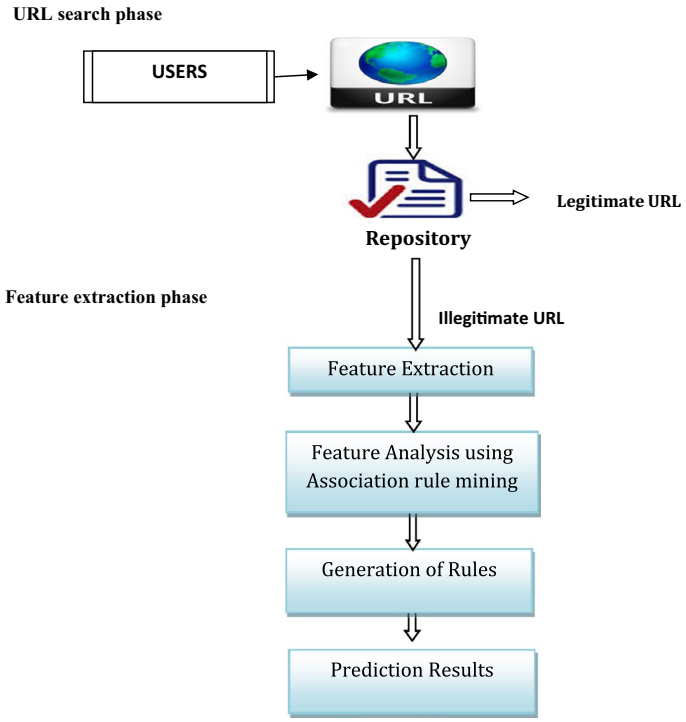

Fig. 1 System architecture

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. Otherwise the URL is not a legitimate URL and it undergoes the next phase. The main reason for carrying out the search phase before the feature extraction phase is it reduces the unnecessary computation during the feature extraction phase and improves the overall response time of the system. In the Feature extraction phase, we have defined heuristics to extract 14 features from the URL and are subjected to association rule mining to determine the legitimate and phished URL.

Feature extraction

The proposed work 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 (http://www.phishtank.com) 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.

Heuristic 1: length of the host URL

URL is a formatted text string utilized by internet users to recognize a network resource on the Internet. URL string consists of three elements such as network protocol, host name and path. For a given URL, the host name is extracted and host name length is examined. For the input data set (1200 phishing URLs and 200 legitimate URLs), domain

name length is analyzed for phishing and legitimate URLs. The distribution of the domain name length for phished URL is plotted in Fig. 2 and the average length of the domain name (ļ) in phishing URL is found to be greater than 25 characters. The distribution of the domain name length for legitimate URL is plotted in Fig. 3 and the average length of the domain name (ļ) in legitimate URL is found to be 20 characters.

Therefore, the heuristic is defined as

$$H_1 = \begin{cases} if\ length(host) > ļ \rightarrow Phishing\ URL \\ else, Legitimate\ URL \end{cases}$$

Heuristics 2: number of slashes in URL

The phishers try to trick web users by mimicking the doubtful URL look legitimate. One such technique used in scamming is the addition of slashes in URL. The present study, therefore, considers the number of slashes in URLs as a feature of identification of phishing and examines the number of slashes (μ) in legitimate and phishing URLs. For the input data set (1200 phishing URLs and 200 legitimate URLs), the number of slash is

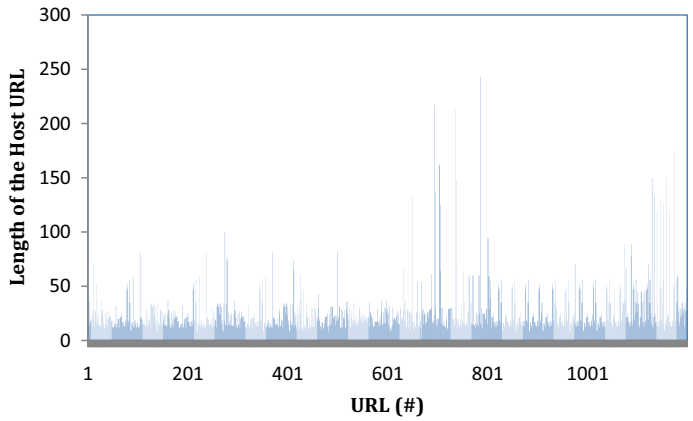

Fig. 2 URL domain length of phishing URL

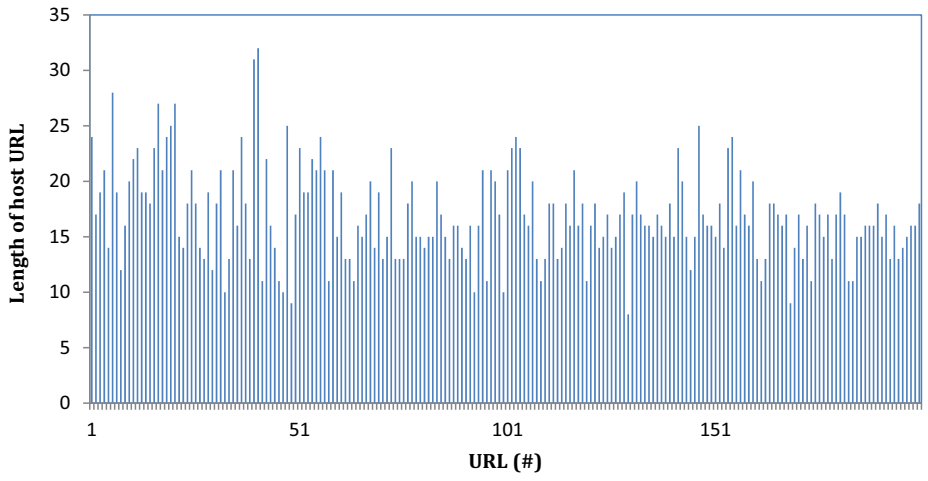

Fig. 3 URL domain length of legitimate URL

analyzed for phishing and legitimate URLs. The distribution of number of slashes in the phishing URLs and legitimate URLs are analyzed and are plotted in Figs. 4 and 5. The result shows that the average number of slashes (μ) in phishing URL is found to be greater than or equal to five and the average slash (μ) in legitimate URL is found to be three.

Therefore, the heuristic is defined as

$$H_2 = \begin{cases} if\ (Slash\ in\ URL) \geq \mu \rightarrow Phishing\ URL \\ else,\ Legitimate\ URL \end{cases}$$

Heuristics 3: dots in host name of the URL

This heuristics verifies the presence of dots in the host name of the URL. However phishing URL usually have many dots to make users believe that they are genuine page. For the input data set (1200 phishing URLs and 200 legitimate URLs), the number of dot in hostname is analyzed for phishing and legitimate URLs. The presence of dots (α) in both phishing and legitimate URLs are analyzed and are plotted in Figs. 6 and 7. The

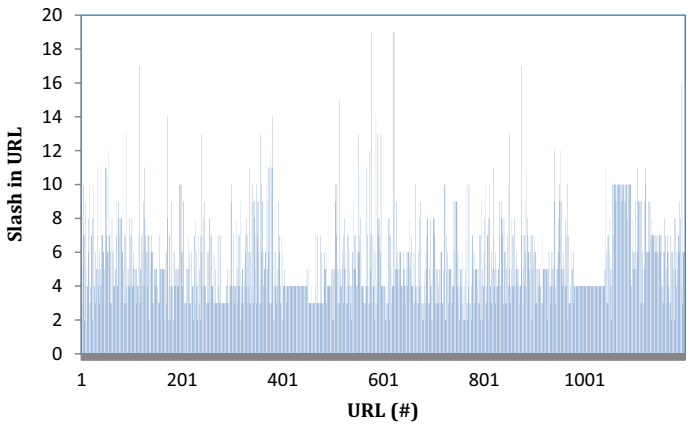

Fig. 4 Number of slashes in phishing URL

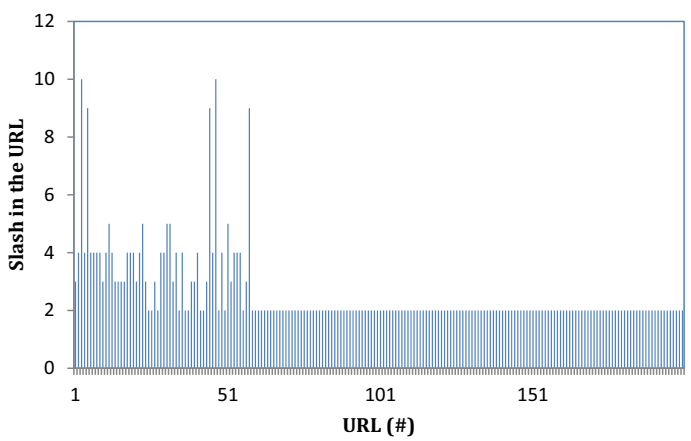

Fig. 5 Number of slashes in legitimate URL

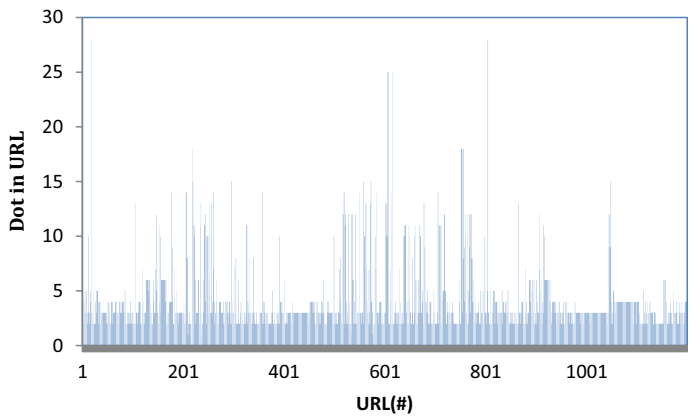

Fig. 6 Number of dots in phishing URL

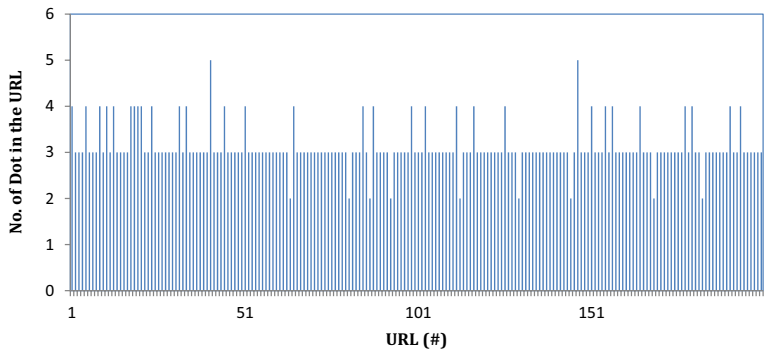

Fig. 7 Number of dots in legitimate URL

result shows that the suspicious URL has more than four dots and the number of dots (α) in legitimate URL is almost three.

Therefore, the heuristic is defined as

$$H_3 = \begin{cases} if\ \left(Dots\ in\ host\ name\ of\ URL\right) > \alpha\ \rightarrow\ Phishing\ URL \\ else,\ Legitimate\ URL \end{cases}$$

Heuristics 4: number of terms in the host name of the URL

Domain names are used to ascertain a unique identity. The host name component is extracted from the URL. The terms in the host name component are tokenized. If the number of terms in the host name portion of the URL is (∂) then the URL is legitimate otherwise phishing. For the input data set (1200 phishing URLs and 200 legitimate URLs), the number of terms in the host name of the URL is analyzed for phishing and legitimate URLs. The distribution of number of terms in the host name for phishing URL is plotted in Fig. 8. The number of terms in the host name (∂) in the phishing URL is found to be greater than four. The distribution of number of terms in the host name for legitimate URL is plotted in Fig. 9. The average number of terms in the host name (∂) in the legitimate URL is found to be four.

Therefore, the heuristic is defined as

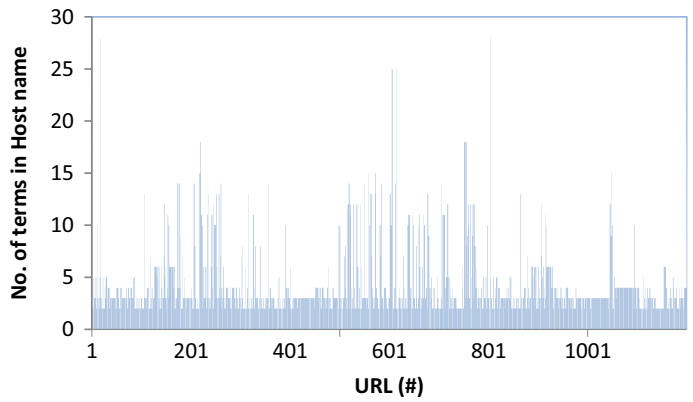

Fig. 8 Number of terms in the host name of phishing URL

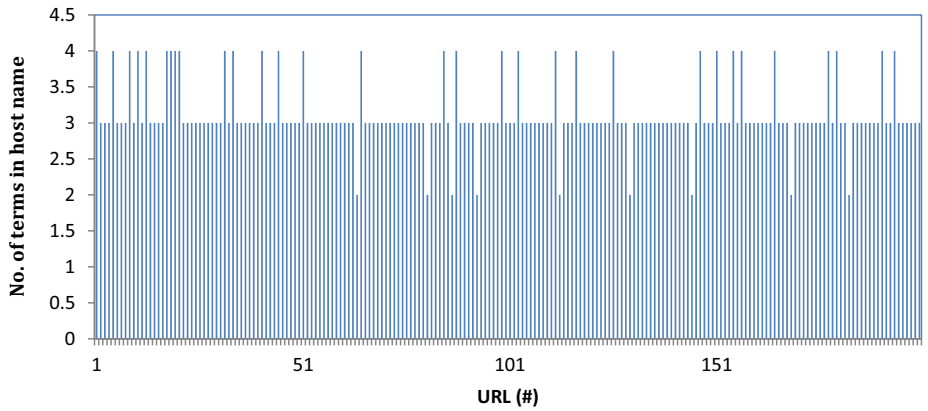

Fig. 9 Number of terms in the host name of legitimate URL

$$H_4 = \begin{cases} if\ Host(terms) > \partial \rightarrow Phishing\ URL \\ else,\ Legitimate\ URL \end{cases}$$

Heuristics 5: special characters

The URL is unique in the cyberspace. The identity of the legitimate website is obtained from the host name of the URL. The hostname in the URL of the legitimate and phished dataset is investigated for understanding the presence of special characters in both the data sets. While examining, it was found that 77.75 % of phished URLs are with special characters.

Therefore, the heuristic is defined as

$$H_5 = \begin{cases} if\ (Special\ characters\ in\ Host) \rightarrow Phishing\ URL \\ else,\ Legitimate\ URL \end{cases}$$

Heuristic 6: IP address

In general, the legitimate websites are addressed by their domain name. In the dataset, the hostname in the URL is examined for the presence of IP address. It was found that 9.4 % of phished URLs contained IP address.

Therefore, the heuristic is defined as

$$H_6 = \begin{cases} if\ (IP\ address\ in\ URL) \rightarrow Phishing\ URL \\ else,\ Legitimate\ URL \end{cases}$$

Heuristics 7: unicode in URL

Unicode provides a unique number for every character. On analyzing the input data set (1200 phishing URLs and 200 legitimate URLs), it is scrutinized that most of the phishing URLs are coded with unicode characters. The presence of Unicode in the host name of the URL indicates that the URL is a phished URL. It was found that 65.16 % of phished URLs contained unicode characters

Therefore, the heuristic is defined as

$$H_7 = \begin{cases} if\ (Unicode\ in\ URL) \rightarrow Phishing\ URL \\ else,\ Legitimate\ URL \end{cases}$$

Heuristics 8: transport layer security

The URL is broken down into host component and a path component. The URL uses Transport Layer Security to determine whether the URL is protected. The presence of HTTPs protocol is required when delicate information is transferred across network. Therefore the existence of Transport Layer Security is examined for the input URL. On analyzing the phishtank dataset, 99.16 % URLs were found without transport layer security.

Therefore, the heuristic is defined as

$$H_8 = \begin{cases} if\ (URL\ is\ http) \rightarrow Phishing\ URL \\ else,\ Legitimate\ URL \end{cases}$$

Heuristics 9: Subdomain

The securityweek network, [25] reports that subdomain leads to increase uptime for phishing attack. Fraudsters, scam users by adding sub domains to make the link look legitimate. Adding subdomain to the URL makes the cyber space user believe that the URL belongs to the authentic website. Hence, the number of subdomain present in the phishing URLs is analyzed and 64 % of phished URLs are found to be with subdomain.

Therefore, the heuristic is defined as

$$H_9 = \begin{cases} if\ (Subdomain\ in\ URL) \rightarrow Phishing\ URL \\ else,\ Legitimate\ URL \end{cases}$$

Heuristics 10: certain keyword in the URL

While investigating the URLs of the phished and legitimate dataset, it was found that certain keywords occur frequently in the phished URL. Therefore (TF-IDF) method is used to obtain the term frequency identity set from the path portion of the URLs and repeated keywords were examined. It was found that the Phishing URL contains many frequently repeated keywords such as suspend, confirm, PayPal etc. The frequency of the keyword is calculated for the entire URLs. If the most commonly used keywords are present in the path portion of the URL, then the URL is a phished URL. On analyzing the phishtank dataset, 91.08 % of URL has certain keywords in the path portion of the URL. Table 1 shows the algorithm of TF-IDF.

Therefore, the heuristic is defined as

Table 1 TF-IDF algorithm

Algorithm: Term Frequency Inverse Document Frequency

Input: Path portion of the URLs

Output: Term Identity set

1. Extract the path portion of the URLs
2. Text Preprocessing
   a. Lowercase and tokenize the text into terms
   b. Ignore short terms and stop words
3. The term frequency $Tf_{i,j}$ value of term t_i in the URL d_j is

$$Tf_{i,j} = \sqrt{\frac{n_{ij}}{\sum_k n_{k,j}}}$$

4. The inverse document frequency idf of term t is

$$\mathrm{idf}_i = \ln\left(\frac{|D|}{|\{d_j : t_i \epsilon d_i\}+1|}\right) + 1$$

The Tf-idf weight of the term t_i in the URL d_j is

$$Tf - idf = Tf_{i,j} * idf_i$$

5. Return the top highest Tf-idf terms

$$H_{10} = \begin{cases} if\ (Keyword\ in\ the\ path\ portion\ of\ URL) \rightarrow Phishing\ URL \\ else,\ Legitimate\ URL \end{cases}$$

Heuristics 11: top level domain

The host name of the URL includes the top level domain, secondary level domain and the domain. The top level domain part of the URL is checked for existence. If the top level domain is not available in the host name of the URL then it is a phished URL otherwise the URL is a legitimate URL. Hence, the existence of the top level domain in the phishing URLs is analyzed and 66.5 % of phished URLs are found without top level domain.

Therefore, the heuristic is defined as

$$H_{11} = \begin{cases} if\ (Top\ Level\ Domain\ does\ not\ exists) \rightarrow Phishing\ URL \\ else,\ Legitimate\ URL \end{cases}$$

Heuristics 12: number of dots in the path of the URL

The path portion is retrieved from the URL. The path portion of the URL is examined for the occurrence of dots. The number of dots in the path portion of the URL is analyzed for both URLs. The distribution of the dots in the path of the URL for phishing URL is plotted in Fig. 10. The number of dots (Ω) in the path of the URL is found to be greater than two. The distribution of dots in the path portion of the URL for legitimate URL is plotted in Fig. 11. The number of dots (Ω) in the path of the URL is found to be less than two.

Therefore, the heuristic is defined as

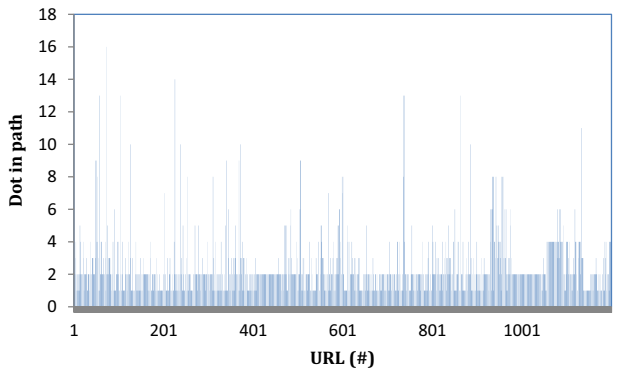

Fig. 10 Number of dots in the path of phishing URL

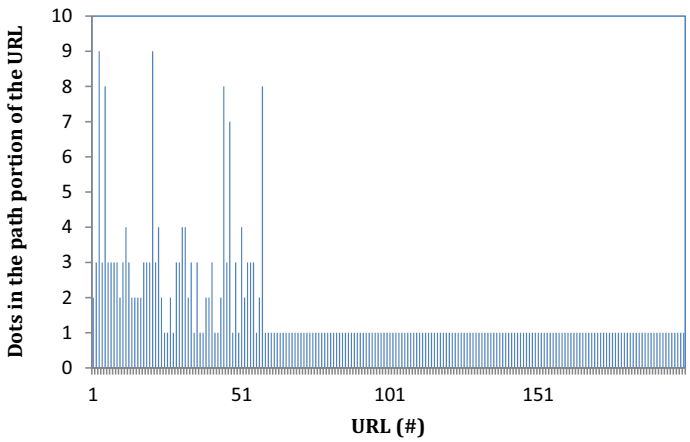

Fig. 11 Number of dots in path of legitimate URL

$$H_{12} = \begin{cases} if\ path(dot) > \Omega \rightarrow Phishing\ URL \\ else,\ Legitimate\ URL \end{cases}$$

Heuristics 13: hyphen in the host name of the URL

The results confirm that the presence of more than one hyphen in the host name of the URL may indicate that it is a phishing URL. Legitimate URLs are found to have at the most one hyphen in majority of cases. Figures 12 and 13 show the distribution of number of hyphen in the host part of phishing and legitimate URLs. The number of hyphen (λ) in the host part of the URL is found to be greater than 1 for phishing URLs. The number of hyphen (λ) in the host part of the URL is found to be 1 for legitimate URLs. http://www.merchant-credit-card-account.net/PeyPol/profile.php is a phishing URL in which the hyphen available in the host portion of the URL is found to be three.

Therefore, the heuristic is defined as

$$H_{13} = \begin{cases} if\ hyphen(hostname) > \lambda \rightarrow Phishing\ URL \\ else,\ Legitimate\ URL \end{cases}$$

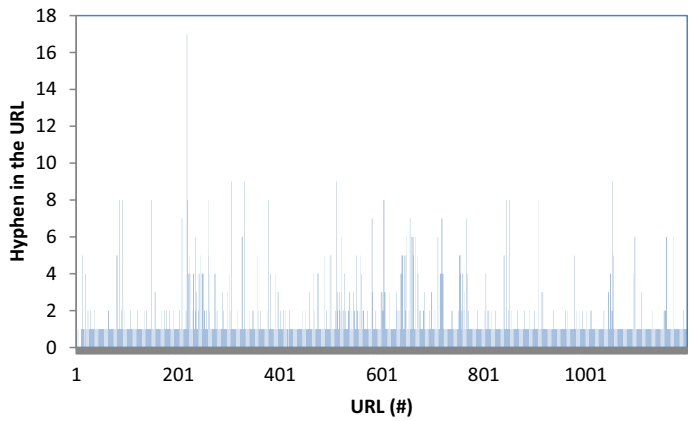

Fig. 12 Phishing hyphen in the URL

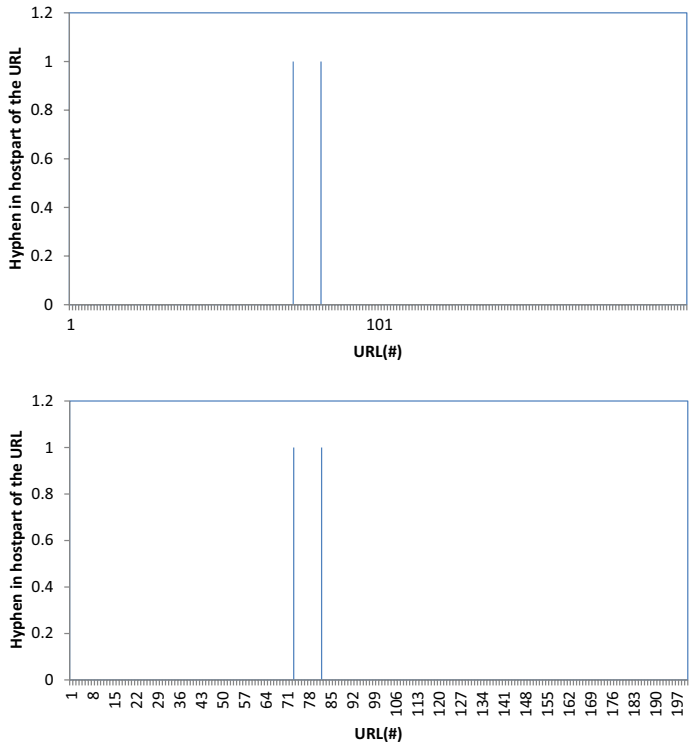

Fig. 13 Legitimate hyphen in the host name

Heuristics 14: URL length

The URL is checked for the occurrence of phishing attack. Long URLs are used by the phisher to hide the doubtful part of the URL. If the length of the URL is greater than β then the URL is identified as phished URL otherwise legitimate URL. Figures 14 and 15 show the distribution of the length of the URL for both phished and legitimate URLs. The average length of the legitimate URL (β) is found to be 40. The average length of the phishing URL (β) is found to be greater than 75.

Therefore, the heuristic is defined as

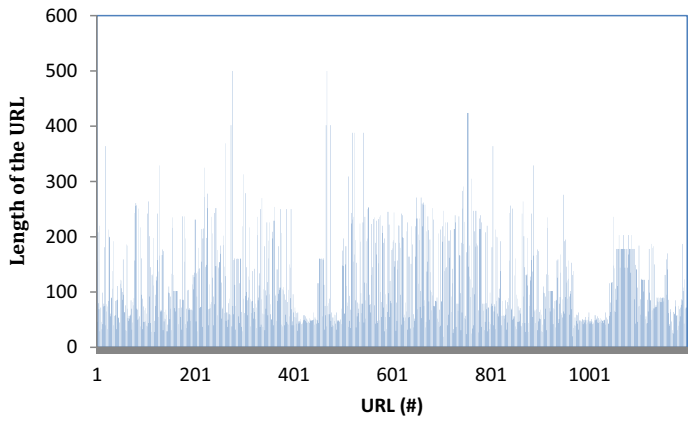

Fig. 14 Length of phishing URL

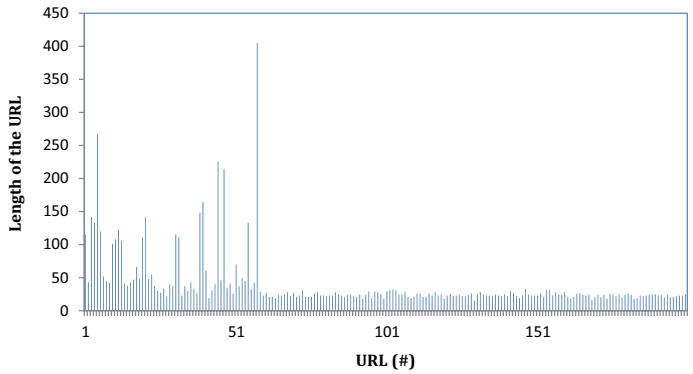

Fig. 15 Length of legitimate URL

$$H_{14} = \begin{cases} if\ \left(length\ of\ URL\right) > \beta \rightarrow Phishing\ URL \\ else,\ Legitimate\ URL \end{cases}$$

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 essential knowledge from the database. The algorithm was proposed by Agrawal et al. [26]. 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. This algorithm uses past information of frequent itemset 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 is an algorithm proposed by [27], which 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.

$$Supp(X) = \frac{Number\ of\ times\ X\ appears}{N} = P(X) \tag{1}$$

$$Supp(XY) = \frac{Number\ of\ times\ X\ and\ Y\ appears\ together}{N} = P(X \cap Y) \tag{2}$$

$$Conf(X \rightarrow Y) = \frac{Supp(XY)}{Supp(X)} = \frac{P(X \cap Y)}{P(X)} \tag{3}$$

$$\{Support,\ Confidence\} = Accuracy \tag{4}$$

Association rule mining to detect phishing URL

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. This rule is used to ascertain the URL type when a user accesses it. We have recognized different heuristics extracted from the URLs and collected over 1400 URLs from several sources. Legitimate URLs is acquired from five sources and is shown in Table 2 and we collected 1200 phishing URLs from phishtank database (http://www.phishtank.com). The feature extraction is implemented in PHP. The experiments were performed using WEKA. WEKA a data mining tool incorporates collection of machine learning algorithms. 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. Detail of this experiment is provided in the following subsections.

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. Association rule mining is used to explore the hidden relationship between the attributes. In the proposed work, association rule mining has been used for detecting the frequently occurring features in phishing URLs. 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.

Table 2 Legitimate data source

Source	Link
Yahoo most visited sites	http://dir.yahoo.com/Business_and_Economy
Most visited sites google's top 1000	http://www.google.com/adplanner/static/top1000
Alexa's top targeted sites	http://www.alexa.com/topsites
Netcraft's most visited sites	http://toolbar.netcraft.com/stats/topsites
Millersmile's top targeted sites	http://www.millersmiles.co.uk

The result of the analysis is shown in Table 3. The results show that most of phishing URL are without transport layer security. Other features indicating a phishing URL include subdomain in the URL and certain keyword in the path portion of the URL. Similarly the special characters in the host name of the URL, top level domain does not exist in the URL and Unicode in the URL were all possible features that are available in the phishing URL. To conclude the investigation features such as transport layer security, unavailability of the top level domain and keyword in the path portion of the URL frequently occur together in phishing URL.

Rule extracted from predictive apriori

The detailed study and analysis of phishing URL were carried out and the results indicate that in predictive apriori algorithm few different rules are generated apart from apriori. The rules generated by predictive apriori are based on accuracy. The result obtained from predictive apriori is shown in Table 4. The strong rule generated by the predictive apriori with accuracy level above 99 % has been considered for further analysis and the other rules are discarded. An investigation on the itemsets reveals that the features like transport layer security, unavailability of the top level domain in the URL and keyword within the path portion of the URL were found to be sensible indicators for phishing URL. In addition to this number of slashes in the URL, dot in the host portion of the URL and length of the URL are also the key factors for phishing URL. Other features such as special characters in the URL, Unicode in the URL, length of the URL is greater than 75 and more than four dots in the host name of the URL were also found to be significant features of phishing URL.

Table 3 Rules extracted for phishing URL through apriori algorithm

Algorithm:	Apriori
Rules:	
Phishing URLs	
Rule 1: *if* {Transport Layer Security = http ∩ Keyword in the path portion of the URL = yes ∩ Top level domain = yes} => *class phishing*(*conf.*, 1).	
Rule 2: *if* { Number of slash in URL ≥ 5 ∩ Transport Layer Security = http ∩ Keyword in the path portion of the URL = yes} => *class phishing* (*conf.*, 1).	
Rule 3: *if* {Special characters = yes ∩ Transport Layer Security = http ∩ Number of terms in the host name of the URL > 4} => *class phishing* (conf., 1).	
Rule 4: *if* {Dot in the host URL > 4 ∩ Transport Layer Security = http ∩ Number of terms in the host name of the URL > 4} => *class phishing* (conf., 1).	
Rule 5: *if* {Number of slash in the URL ≥ 5 ∩ Dot in the host URL > 4 ∩ Length of the URL > 75} => *class phishing* (conf., 1).	
Rule 6: *if* {*Special Characters = yes ∩ Transport Layer Security = http ∩ Top level domain = yes* } = > *class phishing* (*conf.*, 1).	
Rule 7: *if* {*Dot in the host URL* > 4 ∩ *Transport Layer Security = http ∩ Keyword in the path portion of the URL = yes* } => *class phishing* (*conf.*, 1).	
Rule 8: *if* {*Special Characters = yes ∩ Transport Layer Security = http ∩ Keyword in the path portion of the URL = yes* } => *class phishing* (*conf.*, 1).	
Rule 9: *if* { *Dot in the host URL* > 4 ∩ *Keyword in the path portion of the URL = yes ∩ Top level domain = yes*} = > *class phishing* (*conf.*, 1).	

Table 4 Rules extracted for phishing URL through predictive apriori algorithm

Algorithm: Predictive Apriori

Rules:

Rule 1: *if* {*Number of slashes in URL* $\geq 5 \cap$ *Special characters* $=$ *yes* $\cap$ *Transport Layer Security* $=$ *http* $\cap$ *Keyword in the path portion of the URL* $=$ *yes* $\cap$ *Top Level Domain* $=$ *yes* $\cap$ *Length of the URL* > 75 } => *class phishing* (*acc.*, 0.995).

Rule 2: *if* {*Number of slashes in URL* $\geq 5 \cap$ *Unicode in URL* $=$ *yes* $\cap$ *Transport Layer Security* $=$ *yes* $\cap$ *Length of the URL* > 75} => *class phishing* (*acc.*, 0.995).

Rule 3: *if* {*Number of slashes in URL* $\geq 5 \cap$ *Dots in host name of the URL* $> 4 \cap$ *Unicode in URL* $=$ *yes* $\cap$ *Keyword in the path portion of the URL* $=$ *yes* $\cap$ *Top Level Domain* $=$ *yes*} $=$ $>$ *class phishing* (*acc.*, 0.995).

Rule 4: *if* {*Number of slashes in URL* $\geq 5 \cap$ *Special characters* $=$ *yes* $\cap$ *Unicode in URL* $=$ *yes* $\cap$ *Top Level Domain* $=$ *yes* $\cap$ *Length of the URL* > 75 } => *class phishing* (*acc.*, 0.995).

Rule 5: *if* {*Number of slashes in URL* $\geq 5 \cap$ *Dots in the hostname of the URL* $> 4 \cap$ *Special characters* $=$ *yes* $\cap$ *Unicode in URL* $=$ *yes* $\cap$ *Top Level Domain* $=$ *yes* } $=$ $>$ *class phishing* (*acc.*, 0.995).

Rule 6: *if* {*Number of slashes in URL* $\geq 5 \cap$ *Dots in the hostname of the URL* $> 4 \cap$ *Special characters* $=$ *yes* $\cap$ *Length of the URL* > 75 } => *class phishing* (*acc.*, 0.995).

Rule 7: *if* {*Special characters* $=$ *yes* $\cap$ *Subdomain* $=$ *yes* $\cap$ *Number of terms in the hostname of the URL* > 4 } $=$ $>$ *class phishing* (*acc.*, 0.995).

Rule 8: *if* {*Number of slashes in URL* $\geq 5 \cap$ *Dots in the hostname of the URL* $> 4 \cap$ *Transport Layer Security* $=$ *yes* $\cap$ *Keyword in the path portion of the URL* $=$ *yes* $\cap$ *Top Level Domain* $=$ *yes* $\cap$ *Length of the URL* > 75 } $=$ $>$ *class phishing* (*acc.*, 0.995).

Rule 9: *if* {*Unicode in URL* $=$ *yes* $\cap$ *Keyword in the path portion of the URL* $=$ *yes* $\cap$ *Number of terms in the hostname of the URL* $> 4 \cap$ *Length of the URL* > 75 } => *class phishing* (*acc.*, 0.995).

Overall the two algorithms generated different rules. Apriori algorithm is faster comparing to predictive apriori. Comparing the experimental results, it is seen that the features such as subdomain, URL without transport layer security and keyword in the path portion of the URL are found in most of the phishing URL and these are obtained by using apriori algorithm. However, the number of slashes in the URL is greater than or equal to five, top level domain does not exists in the URL, keyword in the path portion of the URL and special characters in the URL are found in most of the phishing URL and are obtained using predictive apriori algorithm. The algorithms are compared with the execution time for different number of instances. The execution time for both the algorithms is shown in Table 5. The graphical representation of the computational time for apriori and predictive apriori is shown in Fig. 16. It has been observed that the predictive apriori requires more time than the apriori to generate rules. Finally, it is inferred that when the number of instances is increased the time taken for both the algorithm is increased.

The best unique rules generated from apriori and predictive apriori for various size of the phishing input dataset is considered. The graphical representation of the number of rules mined for apriori and predictive apriori is shown in Fig. 17. Thirty-one unique rules are mined by predictive apriori when 100 % of the input dataset is used. When

Table 5 Execution time of algorithms

No. of instances	Apriori (ms)	Predictive apriori (ms)
500	0	10,958
1000	1	23,996
1500	1	39,002
2000	2	48,000
2500	2	109,982
3000	3	198,001

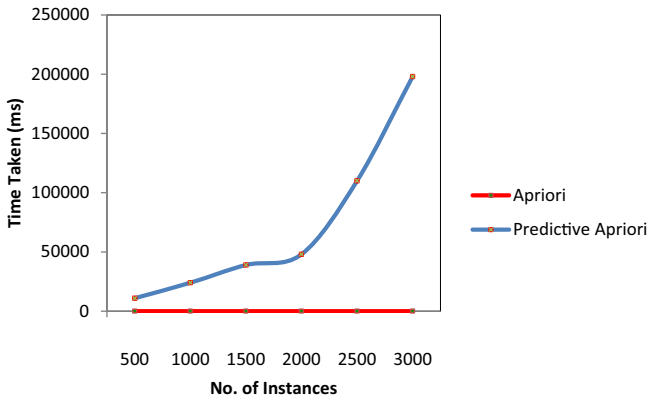

Fig. 16 Comparision of computational time of algorithms

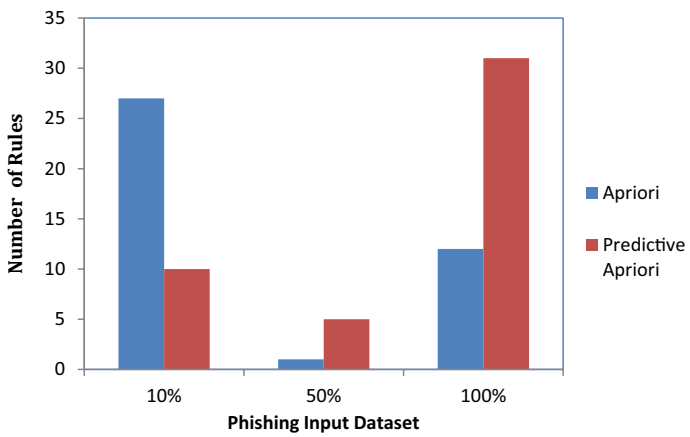

Fig. 17 Number of effective rules of algorithms in various instances

10 % of the input data set is mined using apriori algorithm 27 best unique rules are generated The time taken to generate the rules by predictive apriori is much higher than apriori. Apriori is able to mine the rules must faster. Although, predictive apriori mines unique rules the rules mined by apriori is considered for further process. The rule generated by the training dataset is checked for a sample input data set. 93 % of the test dataset contain URL without transport layer security, keyword in the path portion of the

URL and without top level domain. 90 % of the test data set contain number of slashes in URL is greater than or equal to five, without transport layer security and keyword in the path portion of the URL. 88 % of the test data set contains more than four dots in the host portion of the URL, without transport layer security and number of terms in the host portion of the URL is greater than four. About 79 % of the test data set contains number of slashes in the URL is greater than or equal to five, dots in the host URL is greater than four and length of the URL is greater than 75. Ninety percentage of the test data set contains dots in the host name of the URL, without transport layer security and keyword in the path portion of the URL. About 88 % of the test data set dots in the host URL is greater than four, keyword in the path portion of the URL and does not contain top level domain in the host portion of the URL. Overall 93 % of the phishing URL are detected using the rules obtained by apriori algorithm.

Conclusion

In this paper, the features of the URL are analyzed and are 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 results obtained from rule mining have highlighted the useful features available in the phished URL. Analyzing the information available on phishing URL and considering confidence as indicator, the features such as transport layer security, unavailability of the top level domain in the URL and keyword within the path portion of the URL were found to be sensible indicators for phishing URL. In addition to this number of slashes in the URL, dot in the host portion of the URL and length of the URL are also the key factors for phishing URL.

Authors' contributions
SCJ carried out the studies, and drafted the manuscript. EBR provided full guidance and revised the manuscript to high standards. Both authors read and approved the final manuscript.

Author details
[1] Department of Computer Applications, Karunya University, Coimbatore, India. [2] Director (Collaborations), Karunya University, Coimbatore, India.

Competing interests
The authors declare that they have no competing interests.

Received: 1 September 2015 Accepted: 2 May 2016
Published online: 10 July 2016

References

1. https://ers.trendmicro.com/guide/en_us/AG/AppA/Phish_Attack.htm. Accessed May 2015
2. http://www.consumeraffairs.com/news04/2005/gartner.html. Accessed June 2015
3. http://www.emc.com/collateral/fraud-report/rsa-online-fraud-report-012014.pdf. Accessed June 2015
4. http://www.symantec.com/content/en/us/enterprise/other_resources/bistr_main_report_v19_21291018.en-us.pdf. Accessed June 2015
5. RSA Anti-Fraud Command Center. RSA monthly online fraud report. (2014). http://www.emc.com/collateral/fraud-report/rsa-online-fraud-report-012014.pdf. Accessed July 2015
6. 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
7. 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
8. Huang H, Qian L, Wang Y (2012) A SVM based technique to detect phishing URLs. Int Technol J 11(7):921–925
9. Liébana-Cabanillas F, Nogueras R, Herrera LJ, Guillén A (2013) Analysing user trust in electronic banking using data mining methods. Expert Syst Appl 40:5439–5447
10. Li Y, Xiao R, Feng J, Zhao L (2013) A semi-supervised learning approach for detection of phishing webpages. Optik 124:6027–6033
11. Islam R, Abawajy J (2013) A multi-tier phishing detection and filtering approach. J Netw Comput Appl 36:324–335

12. 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
13. 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
14. Liu W, Deng X, Huang G, Fu AY (2006) An antiphishing strategy based on visual similarity assessment. IEEE Computer Society 1089-7801/06 IEEE, IEEE Internet Computing
15. 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
16. 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
17. Lam IF, Xiao WC, Wang SC, Chen KT (2009) Counteracting phishing page polymorphism: an image layout analysis approach. Springer-Verlag, Berlin
18. Chen KT, Chen JY, Huang CR, Chen JY (2009) Fighting phishing with discriminative keypoint features of webpages. IEEE Internet Comput 13:56–63
19. Shah R, Trevathan J, Read W, Ghodosi H (2009) A proactive approach to preventing phishing attacks using Pshark. In: Sixth international conference on information technology: new generations. IEEE, Las Vegas, pp 915–921
20. He M, Horng SJ, Fan P, Khan MK, Run RS, Lai JL et al (2011) An efficient phishing webpage detector. Expert Syst Appl 38(10):18–27
21. Aburrous M, Hossain MA, Dahal K, Thabtah F (2010) Intelligent phishing detection system for e-banking using fuzzy data mining. Expert Syst Appl 37(12):7913–7921
22. Zhang D, Yan Z, Jiang H, Kim T (2014) A domain-feature enhanced classification model for the detection of Chinese phishing e- business websites. Inf Manag 51:845–853
23. Abdelhamid N, Ayesh A, Thabtah F (2014) Phishing detection based associative classification data mining. Science-Direct 41:5948–5959
24. Han W, Cao Y, Bertino E, Yong J (2012) Using automated individual whitelist to protect web digital identities. Expert Syst Appl 39:11861–11869
25. http://www.securityweek.com/use-subdomains-leads-increased-uptime-phishing-attacks. Accessed June 2015
26. Agrawal R, Imielinski T, Swami A (1993) Mining association rules between sets of items in large databases. ACM-SIGMOD 22:207–216
27. Scheffer T (2001) Finding association rules that trade support optimally against confidence. In: Proceedings of the 5th European conference on principles and practice of knowledge discovery in databases (PKDD'01), Springer-Verlag, Freiburg