COMPUTERS & SECURITY 84 (2019) 256–275

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/cose

Computers & Security

A keyword-based combination approach for detecting phishing webpages

Yan Ding[a], Nurbol Luktarhan[a,*], Keqin Li[b], Wushour Slamu[a]

[a] College of Information Science and Engineering, Xinjiang University, Urumqi, China
[b] Department of Computer Science, State University of New York, New Paltz, New York, USA

ARTICLE INFO

Article history:
Received 16 January 2018
Accepted 9 November 2018
Available online 23 March 2019

Keywords:
Heuristic rule
Machine learning
Phishing
Search engine
URL obfuscation techniques

ABSTRACT

In this paper, the Search & Heuristic Rule & Logistic Regression (SHLR) combination detection method is proposed for detecting the obfuscation techniques commonly used by phishing websites and improving the filtering efficiency of legitimate webpages. The method is composed of three steps. First, the title tag content of the webpage is input as search keywords to the Baidu search engine, and the webpage is considered legal if the webpage domain matches the domain name of any of the top-10 search results; otherwise, further evaluation is performed. Second, if the webpage cannot be identified as legal, then the webpage is further examined to determine whether it is a phishing page based on the heuristic rules defined by the character features. The first two steps can quickly filter webpages to meet the needs of real-time detection. Finally, a logistic regression classifier is used to assess the remaining pages to enhance the adaptability and accuracy of the detection method. The experimental results show that the SHLR can filter 61.9% of legitimate webpages and identify 22.9% of phishing webpages based on uniform/universal resource locator (URL) lexical information. The accuracy of the SHLR is 98.9%; thus, its phishing detection performance is high.

1. Introduction

1.1. Phishing

In 2017, the "Double 11" shopping carnival experienced a record high number of sales that amounted to hundreds of billions of Chinese e-commerce transactions. The continuous development of e-commerce has played a vital role in promoting the world economy and satisfying the needs of the population. However, during this period, phishing webpages and other online fraud behavior have also shown rapid growth trends. Phishing webpages, a form of cyberattack, seriously threaten the credibility and financial security of Internet users, and the annual economic losses caused by phishing websites amount to several hundred million dollars.

Phishing attacks are cybercrimes that steal user privacy data via social engineering or technical methods. Attackers use e-mail or text messages to send false security warnings or prize information to trick users into clicking on the phishing webpage and submitting critical personal information. Phishing attacks can lead to a series of serious problems for users, such as stolen bank account information, which can cause huge financial losses. Moreover, the use of identity information to substantiate false information will negatively affect the users credibility. Phishing webpages and their uniform/universal resource locators (URLs) are the primary sources used by phishing attackers to steal private data

* Corresponding author.
E-mail addresses: dingyan_mail@163.com (Y. Ding), nurbol@xju.edu.cn (N. Luktarhan), lik@newpaltz.com (K. Li), wushour@xju.edu.cn (W. Slamu).
https://doi.org/10.1016/j.cose.2019.03.018
0167-4048/

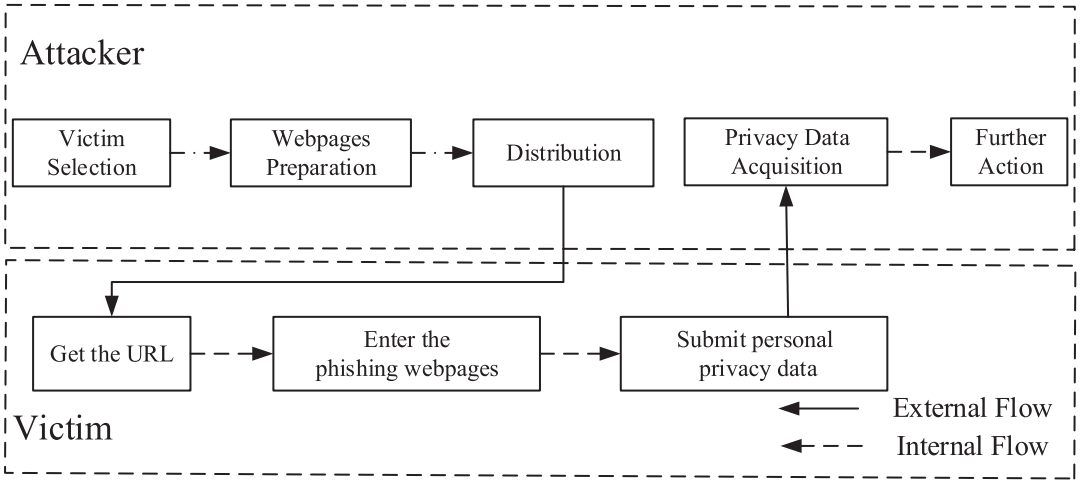

Fig. 1 – Phishing attack flow chart.

from users. According to the 2016 China Internet Security Report (Qihoo360, 2017) released by the 360 Internet Security Center, 1,199,000 new phishing websites were intercepted in 2016 compared with 2015, representing an increase of 25.5% (1,569,000). Over 225 phishing websites are created on average every hour. The scale of phishing webpages and their URLs shows a continuously growing trend.

With the advent of the big data era, phishing attacks have become more sophisticated. To increase the probability of attack success, attackers collect and analyze various data information that users leave on the Internet, such as their name, gender and contact information, and even basic information of the users' family members and social relations. As a result, phishing e-mails appear to be authentic and trustworthy, making users more likely to fall into the trap of phishing when they receive messages that appear relevant. According to Symantec's Internet Security Threat Report (Symantec, 2017), business email compromise (BEC) scams using spear phishing tactics attack more than 400 companies a day and have stolen more than $3 billion over three years. According to the Anti-Phishing Working Group (APWG) statistical report (APWG, 2017), a total of 255,065 phishing attacks against different enterprises or entities were detected globally in 2016, representing an increase of 10% from 2015. The effective detection of phishing webpages is a key measure to maintain Internet security.

1.2. *Phishing categories*

As shown in Fig. 1, phishing attackers need to perform preparatory work before conducting phishing attacks, including selecting subjects and methods of dissemination and creating fake webpages. A reasonable classification of phishing attacks can aid the adoption of effective identification measures (Aleroud and Zhou, 2017). Therefore, this paper classifies phishing attacks based on key processes used in the attack. First, phishing attacks can be divided into the following three types according to the target of the attack.

1. General phishing attacks are attacks that are not targeted to specific subjects. The attacker needs only to send the phishing URL to the user's terminal through various means and does not set a targeted scam plan. Over time, phishing attacks have become easy to conduct because of the evolution of the cybercrime market. An attacker can buy and sell tools for phishing kits, and these kits are usually priced at only $2–$10. Moreover, buyers do not need specialized expertise to execute the attack. Because obvious flaws are exposed with general phishing attacks, the probability of success is relatively low.
2. Spear phishing attacks are attacks on a specific class of subjects. Because such users are often the key information holders of a business or a group, the attack relies on well-planned social scams. An attacker sends false messages to the targets. If the attack does not work within a certain time frame, the attacker will launch other types of exploits against targets with the same background. However, in recent years, attackers have begun to change tactics to avoid exposure after many failures. The number and sophistication of spear phishing attacks have increased substantially compared with those of general phishing attacks. In 2015, 34.9% of all spear phishing e-mails were sent to financial enterprises, and financial industry enterprises had at least an 8.7% likelihood of experiencing at least one attack throughout the year (Symantec, 2016).
3. Whale phishing attacks are phishing attacks against corporate CEOs and political leaders. Attacker vectors use infected e-mails, collect and analyze users' online and offline information, and exploit various cyber vulnerabilities to develop a detailed attack strategy. Whale phishing has similarities to spear phishing, with both targeting specific groups of people; however, the losses caused by whale phishing are huge and catastrophic. For example, in 2016, Walter Stephan, the CEO of the Austrian aircraft parts maker FACC caused a $47 million loss because of a phishing e-mail.

Building a phishing webpage is a key part of a phishing attack, and how to disseminate URLs to clients is one issue

that attackers must consider. The methods of disseminating phishing webpages can be divided into three categories.

1. E-mail. Although instant messaging technology is gaining popularity among corporate and personal user groups, e-mail still dominates the field of digital communications. E-mail represents a simple and affordable dissemination method to spread phishing pages. In 2016, one out of every 2596 e-mails was a phishing attempt (Symantec, 2017).
2. Short Message Service (SMS). In 2016, 360 mobile guards blocked 17.35 billion spam messages across the China, of which fraud messages ranked second and accounted for 2.8% of the total spam messages (Qihoo360, 2017). However, because of the development of mobile Internet technology, users have become reliant on mobile devices, such as mobile phones; therefore, SMS represents a major method of delivering phishing URLs.
3. Others. Voice phishing is an attack technique that uses telephone systems or voice over IP (VoIP) systems. Users may receive a phone message, e-mail, or text message encouraging them to call a number. During the conversation, attackers will direct the user to provide detailed personal private data, such as credit card numbers and birthdays, all of which are used to implement phishing attacks. The attacker also uses the vulnerabilities in legitimate websites to hijack a site, change the legal link or intercept user communication data.

Attackers are not limited to using the forms of attack and routes of transmission considered in this paper. For example, a new phishing attack technology is the domain shadow, where an attacker steals a webmaster's domain account and creates tens of thousands of subdomains. Then, the attacker uses the subdomains to point to malicious sites or upload malicious code to the domain name server. Although new phishing attacks will continually be developed, any new phishing attack will require phishing webpages and URLs. Therefore, developing a method for effectively filtering and detecting these attacks will directly affect the reliability and safety of the entire Internet.

1.3. *Challenge*

At present, phishing attacks remain a serious security issue because they do not rely solely on vulnerabilities in terminals or communication protocols but also on psychological knowledge. Phishing is a form of attack that relies on users' mindsets. In recent years, many phishing detection methods have been proposed because of the tremendous losses accrued from phishing attacks. Detection methods are used to extract the relevant features of webpages to determine the nature of the site, and the commonly used features of these methods can be categorized as follows.

1. URL based.
   1) A URL is a standard Internet resource address. The standard naming convention of a URL is as follows: protocol://username:password@subdomain.domain_label.top_level_domain:port/path/filename.file?parameter. A URL can be divided into four parts, the domain, path, file, and parameters. Many unrelated words are included in phishing URLs; thus, phishing URLs present certain differences in their characters with respect to legitimate URLs (Mcgrath and Gupta, 2008). For example, Ma et al. (2009, 2011) extracted a number of features, including the length of the URL string, the number of domain labels, and the number of segmented characters to detect phishing webpages.
   2) Domain. The domain name is a unique identification that consists of two or more words separated by a period. The domain can be divided into three levels based on location, with the word at the far right of the domain name representing the top-level domain, which includes country names, such as.cn representing China's top-level domain, and top-level international domain names, such as.com (for business) and.org (for non-profit organizations). A number of new top-level domains are available, such as.red. The second-level domain name refers to the domain name formed by the top-level domain and a word or a phrase. Additional words or phrases included after the second-level domain name form the subdomain name. For example, in the URL http://www.jd.com/search, the domain name is http://www.jd.com, the path is /search, the top-level domain is .com, the second-level domain label is jd, and the second-level domain is jd.com. Vocabulary that can clearly indicate the identity of the URL is considered the identification name of the URL, such as jd. Domain abuse, such as http://www.jd.com.cn.org.net, which is clearly not consistent with URL naming conventions, is a commonly used tactic in URL obfuscation technology. For example, Yadav et al. (2010) calculated the similarity between URL labels and phishing words to detect phishing pages.
2. Host based.
   1) Whois. Whois is the transport protocol used to query the domain's IP and owner information. Such information can describe the registrant of the webpage, registration time and other registration information. If an attacker uses the same e-mail address or name to register a URL, we can determine whether the webpage is a phishing webpage (Sahoo et al., 2017). For example, Chu et al. (2013) detected malware pages by the domain name age.
   2) Domain Name System (DNS). DNS is a distributed database of mappings between domain names and IP addresses. This type of information describes all the IP addresses under the domain name as well as the A, MX and PTR information of the IP. To reduce the latency caused by detection, Seifert et al. (2008) detected malicious webpages by analyzing the relationships among the DNS, web servers and webpages. In addition, Holz et al. (2008) proposed the DNS Fluxiness feature for the use of proxy networks and malicious webpages that frequently change the DNS. Chiba et al. (2016) used DNS logs to detect malicious webpages. Bilge et al. (2011) proposed the EXPOSURE system, which uses passive DNS analysis techniques to detect malicious domains.

3. Content based.
   1) HTML. The HTML contains all the resources of the page and often includes the length of the document, the number of words and the label attributes as detection features. Hou et al. (2010) extracted the HTML features, including the number of words in each line, the number of asymmetric labels and the number of invisible labels. Choi et al. (2011) used the number of page links and Iframe tags as feature categories to achieve better detection results. Zhang et al. (2010) detected webpages by calculating the similarity of the DOM-tree structure between webpages.
   2) Code. Malicious code written in JavaScript can be executed without the user's permission. Hou et al. (2010) sorted 154 JavaScript functions for phishing webpage detection. Choi et al. (2009) used n-gram, entropy, and word size to detect JavaScript obfuscation.
   3) Visual. Webpages often contain a large number of visual elements, such as logos, page layouts and colors. The focus of detection methods based on visual features is calculating the similarity between webpages and protected webpages. Chang et al. (2013) and Kang et al. (2015) used webpage logos to identities the real identity of webpages. Chen et al. (2010) determined whether a page was a phishing website via the webpage layout. Additionally, Dunlop et al. (2010) saved the whole image as an image and used optical character recognition (OCR) technology to extract the text content of the webpage as a search keyword and used the search results to determine the nature of the webpage.

Compared with other URL-based and host-based feature types, web content-based feature types contain more information. When the webpage type cannot be determined according to the URL-based and host-based features, extracting the content features can achieve better detection effects. Because of the popularity of various mobile apps, the use of social networks to spread phishing pages has become a common method for attackers. Liu et al. (2012) evaluated the reputation of certain users by analyzing their social connections and then determined whether the pages they visited were phishing based on the reputations of these users.

Phishing attacks are evolving rapidly and present new challenges to current phishing detection methods. First, URL obfuscation is one common escape tactic used by phishing attackers (Lin et al., 2015). URL obfuscation refers to technical methods of including junk characters in the URL, changing the encoding method, using the IP address instead of the domain name, randomly generating the domain name, etc (Sha et al., 2016). Various URL obfuscation techniques are used to change and eliminate the original features of phishing URLs and evade the detection of anti-phishing tools. Second, phishing attackers add irrelevant words to webpage text to hide the real webpage keywords, and these attacks are challenging for detection methods based on web content. Third, phishing attackers generate webpages that are visually similar to real pages, and increased time and computational resources are required to calculate the similarity between webpages. In the field of phishing webpage detection, the use of such detection methods leads to greater delays and is not conducive to real-time detection needs. Finally, detection methods based on the user profile are often based on the browser. The browser records and analyzes the user's behavior data locally, which can produce good results for fixed users. However, false positives can occur when different users are using the same computer.

Based on an analysis of collected phishing webpages, a combination approach based on keywords, which is called the SHLR (Search & Heuristic Rule & Logistic Regression), is proposed. This method can effectively filter legal webpages and detect phishing webpages that adopt escape technology. The method is composed of three parts: legitimate webpage recognition based on the search engine; phishing webpage recognition based on a heuristic rule; and webpage identification based on a logistic regression classifier.

The SHLR can be divided into the following three phases.

1. The title tag content of the webpage is used as keyword inputs for the Baidu search engine. If the original URL is included in the designated search result set, the webpage may be recognized as a legal webpage.
2. If the webpage cannot be identified as legal, whether the webpage is a phishing page is determined based on heuristic rules defined by the characteristics of the phishing URL.
3. A logistic regression classifier is used to determine the nature of the remaining pages to enhance the adaptability and accuracy of the detection methods.

The main contributions of this paper are as follows.

1. To detect escaping technology, which inserts many unrelated words in the phishing webpage, we use the title tag content of the webpage as keywords and filter legal webpages quickly with the help of the search engine.
2. Seven heuristic rules are proposed for detecting URL obfuscation technology, which can quickly identify phishing websites.
3. A combination approach based on keywords that can achieve good detection performance and meet the needs of real-time detection is proposed.

The remainder of this paper is organized as follows: Section 2 briefly introduces the available phishing detection methods; Section 3 introduces the SHLR in detail; Section 4 describes the experimental schemes and evaluates the performance of the SHLR; Section 5 presents a discussion of the SHLR and future issues in phishing detection; and Section 6 presents our conclusions.

2. Related work

As mentioned above, effectively detecting phishing is a key factor for maintaining Internet security. A variety of methods have been proposed for phishing detection, and they can be divided into the following four categories.

1. Education. One key measure of combating phishing is user education (Mohammad et al., 2015). Alsharnouby et al. (2015) found that the gaze time on Chrome browser

elements is correlated with an increased ability to detect phishing. Parsons et al. (2015) found that the knowledge gained by users completing a phishing study may have improved their diligence and vigilance in detecting phishing. These studies show that anti-fraud education can help users avoid phishing attacks.

2. URL Blacklist/Whitelist. Blacklisted and whitelisted URLs refer to confirmed phishing and legal website domain names and IP addresses, respectively. When a user visits a webpage, the browser's plugin will warn the user to block or allow the linked action based on the blacklist or whitelist. Blacklist/whitelist technology remains the most widely used detection technology and includes the Google Safe Browsing blacklist developed and maintained by Whittaker et al. (2010). However, in recent years, attackers have registered domain names using low-cost methods, such as automatic domain name generation, and selected parts of the sites as phishing links. Updating and maintaining such a large list of sites in real-time with blacklist technology is difficult. According to the survey (Aleroud and Zhou, 2017), 93% of phishing websites are not included in the blacklist. The method often requires manual inspection, honeypot technology, machine learning, and other techniques to efficiently identify phishing webpages, and the detection performance depends on the size of the list. Moreover, the method can only identify whether a page occurs on the list and has a high rate of false negatives.
3. Heuristic Rule. This method relies on expert systems, data mining and machine learning algorithms to build a heuristic rule base to detect phishing attempts. Moghimi and Varjani (2016) extracted the heuristic rules underlying the support vector machine (SVM) algorithm model to detect phishing. Gastellier-Prevost et al. (2011) defined 20 heuristic rules based on the characteristics of phishing URLs and webpages. Hadi et al. (2016) proposed a new associative classification algorithm called the fast associative classification algorithm (FACA) to detect phishing websites. Tan et al. (2016) proposed a phishing detection method based on differences between the target and actual identities of a webpage. Abdelhamid et al. (2014) developed an associative classification method called the multi-label classifier based on associative classification (MCAC) to detect phishing. Prakash et al. (2010) proposed five heuristic rules to enumerate simple combinations of known phishing sites to identify new phishing URLs. In contrast to the blacklist method, the heuristic rule method can recognize freshly created phishing websites in real-time (Miyamoto et al., 2009; Mohammad et al., 2015). However, this method relies on building a heuristic rule base, and difficulties associated with updating the rules are observed. Moreover, this method has higher false positive rates.
4. Machine Learning. This method regards phishing webpage recognition as a classification or clustering problem. Li et al. (2016) proposed an approach based on the minimum enclosing ball support vector machine (BSVM) to achieve high speed and high accuracy for detecting phishing websites. Hu et al. (2016) proposed a method consisting of five single classifiers and four ensemble classifiers to detect malicious domains. Lin et al. (2013) generated two filtering models by using lexical features, descriptive features and combined the models with an online learning algorithm that reduced the number of benign webpages by 75%. Gowtham and Krishnamurthi (2014) proposed an approach that first confirms whether the URL is in the user's whitelist list and whether the webpage has a login form, and then extracts 15 features to detect phishing webpages. Compared with the blacklist-based and rule-based detection methods, this method has better generalization ability. However, the accuracy depends heavily on the feature vector. Therefore, developing a method for extracting more effective features is one of the main research topics in this field. During the process of converting a webpage into a feature vector, especially with high-dimensional vectors, machine learning-based detection methods introduce a large delay, which affects the real-time detection of phishing webpages.

Current detection methods present certain limitations. First, user training must based on the user's age, gender, academic qualifications and other information, and the effect cannot be guaranteed. Moreover, even when participants are well educated and computer literate, they may not detect phishing attempts (Blythe et al., 2011). Second, no valid detection method has been proposed for phishing webpages that use URL obfuscation techniques. Third, detection methods based on a heuristic rules database only have the advantage of detection speed and cannot identify new types of phishing webpages; thus, these methods have poor adaptability and a high false positive rate. Finally, machine learning-based methods perform well only in detection and require more computational resources to extract features and perform model training. Moreover, these methods do not meet the needs of real-time detection. We have the following observations to address these issues.

1. Users often do not remember the domain name of a web page. Instead, they enter the keyword of the webpage in the search engine and click the feedback link to enter the relevant webpage. The search engine queries the resources in cyberspace according to the keywords provided by the user, and returns the search results according to the keywords contained in the network resources and its specific website ranking strategy. Simultaneously, each search engine also follows the laws and regulations of different countries and regions to establish search strategies and search results (Yang et al., 2017). For example, most search engines will automatically block online resources involving phishing, child pornography, infringement of intellectual property rights, and user privacy to comply with ethical principles and local legal requirements. Therefore, we can use the search engine's own search strategy and its technical strength to efficiently identify website identity information and reduce unnecessary feature extraction processes. However, phishing webpages add a large number of unrelated words to improve their relevance to search keywords, improve their website rankings, and defraud users. Thus, the ability of search engines to identify legal and phishing pages is not consistent. Therefore, we propose a method of

filtering legitimate webpages based on search engines to reduce the delay in detecting legitimate webpages.
2. Heuristic rules are an effective way to discover abnormal information based on practical experience. Phishing URLs often use character obfuscation techniques to confuse users. Therefore, the identification of such phishing webpages can be achieved by detecting typical abnormal behaviors. Instead of using a search engine to search for resources in the entire cyberspace, heuristic rules can quickly detect anomalous behavior in a limited search space. We statistically analyze 2776 phishing URLs obtained in PhishTank and design a rule base for phishing URLs using confusing techniques based on the results of the statistical analysis to further improve the ability of real-time detection. However, the escape methods of phishing URLs are various. A complete set of screening rules to effectively distinguish between phishing webpages and legitimate webpages is difficult to construct. Therefore, we propose a phishing web filtering method based on heuristic rules as the second detection phase, and we identify targeted parts of phishing webpages to meet the needs of real-time detection.
3. A large number and variety of phishing pages are impossible to detect by relying solely on search engines or using only a limited rule base. Therefore, we hope to perform a third detection for webpages that are not effectively identified in the first two detection processes to improve the accuracy and adaptability of the detection. The characteristics of machine learning classifiers exactly meet our needs. Therefore, we propose a phishing webpage detection method based on machine learning. After comparing the detection effects of the current popular machine learning classifiers in the field of phishing detection, this paper chooses the logistic regression classifier with L1 norm regularization as the third-phase detection method.

Based on the above discussions, we propose a simple and effective method of phishing detection. First, the SHLR filters certain legal webpages with the help of the Baidu search engine. Second, a rule-based detection method is adopted for certain obfuscation techniques to avoid the feature extraction employed by phishing webpages and meet the needs of real-time detection. Finally, we use a logistic regression classifier and extraction features from DNS, Whois, similarity with phishing vocabulary, lexical feature and HTML to determine the nature of the remaining pages to improve the adaptability and accuracy of the detection method. The SHLR combines the advantages of a search engine, heuristic rule method and machine learning method. When real-time detection is satisfied, the SHLR reduces the false positive rate caused by the lack of rules.

3. Method architecture

Fig. 2 shows the SHLR architecture; the specific steps are defined as follows.

Step 1. Enter the search engine-based detection phase.
Step 2. Use the content of the web's title tag as the search keywords in the Baidu search engine. The webpage is legal if the webpage domain matches the domain name of any of the top-10 search results. If the domain is absent from the ten search results, then the procedure continues to Step 3.
Step 3. Enter the heuristic rule-based detection phase.
Step 4. Check the URL against the rules in the rule base; if the URL matches, the webpage is classified as a phishing webpage. Otherwise, the procedure continues to Step 5.
Step 5. Enter the logistic regression (LR) classifier-based detection phase.
Step 6. Extract features from the URL's DNS, Whois, similarity with phishing vocabulary, lexical feature and HTML.
Step 7. Use a logistic regression classifier to assess the webpage.

Next, we introduce the SHLR in detail based on the detection process.

3.1. *Search engine-based detection phase*

The traditional search engine-based phishing detection method uses the frequency and inverse document frequency (TF-IDF) algorithm to extract the keywords of the webpage or to extract the logo image as the search keyword and then treats a website as suspicious if its domain is absent from the top-*N* search results. Those methods were originally used in the CANTINA (Zhang et al., 2007). The keyword-retrieval component utilizes information algorithms to exploit the power of search engines and does not require training data and prior knowledge (Xiang and Hong, 2009). The phishing pages studied in this paper are phishing sites that mimic legitimate pages and steal users private data. As a result, the keywords for these phishing pages are often not related to their domain name. However, the drawbacks of the traditional methods are obvious. First, attackers can add irrelevant words to the phishing webpage to hide the real keywords. Second, complete reliance on the browser to detect phishing pages can cause a considerable number of false positives. For example, certain newly registered webpages may be miscategorized as phishing pages because of their page rankings.

In response to these problems, we propose the following improvements.

1. We regard the title tag content of the webpage's HTML as the search keywords of the webpage. Many legitimate webpages include their own identification name in the title tag. Search engines obtain mainly the title tag content to verify the relevance of search keywords and webpages; therefore, the title tag content often includes the core identification name of the webpage. Moreover, use of the title tag content as the search word can also reduce the latency caused by a keyword extraction algorithm.
2. Search engines are used to identify only legitimate webpages. Search engine strategies can help us to quickly identify legitimate pages related to the keyword, although they do not effectively recognize phishing pages. Therefore, we use the search engine to detect only whether the webpage

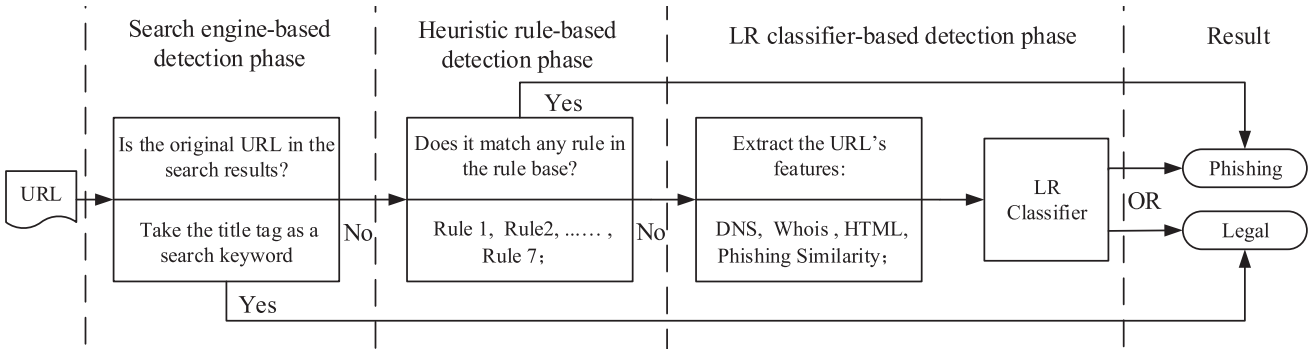

Fig. 2 – SHLR architecture.

is a legal webpage, and the subsequent detection phase determines whether the webpage is a phishing page. With the help of search engines, we can reduce the detection time of legitimate webpages and improve the real-time detection ability.

Our experiments have shown that the proposed detection phase can filter 61.9% of legitimate webpages and presents an error rate of only 0.04%, which means that the method exhibits a good ability to recognize authentic websites.

3.2. *Heuristic rule-based detection phase*

Pattern matching through strings can detect URL obfuscation techniques, such as URL naming standard violations and hidden phishing target words. If the detection phase can directly determine whether the URL is a phishing website, the feature extraction process required for classifier detection can be omitted to meet the requirements of real-time detection. Three modes of splitting the URL are detailed below, and different segmentation methods will produce different URL vocabulary lists.

Mode 1. Split via non-alphanumeric and non-numeric characters not including the underscore '_' and hyphen '-' characters.
Mode 2. Split via '/' and '\'.
Mode 3. Split via non-alphanumeric and non-numeric characters including the underscore '_' and hyphen '-' characters.

In legal URLs, the identification name is used as the second- or third-level domain name label, for example, in http://www.taobao.com, taobao is the identification name of Taobao. The phishing target word refers to the identifying name that the phishing URL is attempting to disguise (Ramanathan and Wechsler, 2013). The phishing URL adds the phishing target word to their domain, path, or file part to mimic the legal website (Ramanathan and Wechsler, 2013). Attackers also use mixed-case, alphanumeric and misspelled words as a disguise, such as pa.ypal and taobao123.com. We have chosen the names of companies in the fields of e-commerce, mail communication, online banking, and APP to be included in the collection according to the identification names of famous global enterprises or groups (Apple, Alibaba, Google, and Gmail). As shown in Table 1, we establish a word set containing 90 phishing target words. By detecting the phishing target word in the URL, we can determine whether it is a phishing URL and identity the subject that the phishing webpage is attempting to camouflage, allowing for more effective countermeasures.

Because of the special nature of the identification name, a legitimate webpage will not insert another company's identification name in their URL. One URL obfuscation technique is the addition of a phishing target word in the path section, which results in multiple identification names appearing in one URL. Table 2 shows the PhishTank and Yahoo dataset URL totals and the proportion of various types of exception information. As shown in Table 2, 6.9% of phishing URLs have multiple identification names, such as http://www.frrtrt2.com/Dropbox/id. This link is a phishing URL masquerading as Dropbox. The path portion of the URL contains the phishing target word dropbox but does not match the domain label frrtrt2. We propose Rules 1 and 2 based on this phenomenon.

Rule 1. The URL is segmented by Mode 1. If a phishing target word is included in the path and the word is different from the second-level or third-level domain label, the URL is considered to be a phishing website.
Rule 2. The URL is segmented by Mode 1. If two or more different identification names are included in the domain or path part, the URL is considered to be a phishing website.

The method of using numbers, delimiters, and concatenated strings to forge an identification name is a common obfuscation technique used by attackers. As shown in Table 1, this confusion technique is used at a frequency of 7.1%. Rule 3 can effectively detect obfuscation with mixed numbers. Thus, we propose Rule 4, which is a URL word reorganization method to detect more complicated obfuscation technology. For example, Fig. 3 shows a flow chart of the URL word list reorganization. In the domain name section, the standard top-level domain terms do not participate in the reorganization of words. Moreover, for a name with less than 4 characters, Rule 4 does not apply.

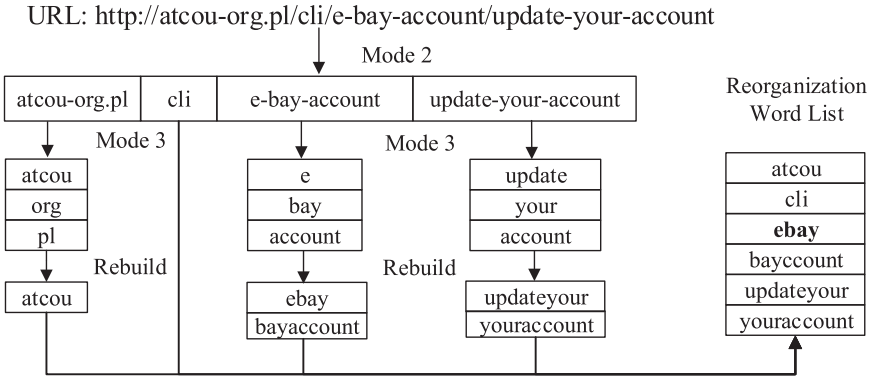

Fig. 3 – URL word list reorganization.

Table 1 – Phishing target word list.

amazon, ayol, banco, alibaba, gmail, google, adobe, metrobank, facebook, epay, lottomatica, whatsapp, cartasi, wellsfargo, americangreetings, absabank, usaa, usbank, alliancebank, blockchain, itau, pkopolishbank, iconic, hotmail, paypal, cariparmacreditagricole, mastercard, allegro, blizzard, huawei, britishtelecom, alliedbanklimited, standardbankltd, cooperativebank, linkedin, pncbank, hmrc, southafricanrevenueservice, microsoft, steam, firstnationalbank, asbbanklimited, aetnahealthplans, hsbcgroup, deutschebank, netflix, bankofamericacorporation, yahoo, vodafone, barclaysbankplc, capitalone, cielo, discovercard, gucci, twitter, westpac, dropbox, royalbankofscotland, santanderuk, natwestbank, lloydsbank, apple, bancaintesa, nationalaustraliabank, royalbankofcanada, groupon, bancodebrasil, bradesco, ramweb, keybank, walmart, tamfidelidade, westernunion, deltaairlines, accurint, capitecbank, suncorp, bankofamerica. americanexpress, internalrevenueservice, sulakecorporation, citibank, runescape, halifax, visa, amarillionationalbank, cibc, posteitaliane.

Table 2 – URL exception information statistics.

Exception	Data source	
	PhishTank	Yahoo
Multiple identification name (%)	6.9	0.02
Hidden phishing target words (%)	7.1	0.08
Violation of naming standards (%)	4.1	0.03
IP as a domain name (%)	0.68	0
The total number of URLs	2,776	5,883

Rule 3. The URL is segmented by Mode 1, and no phishing target word is found. However, if the phishing target word is found after the number in the URL is omitted, then the site is considered to be a phishing website.

Rule 4. The URL is segmented by Mode 1, and no phishing target word is found. The URL is then divided into sections by Mode 2, and the sections are segmented by Mode 3. In each section, words in the consecutive index positions are concatenated to form new words, and a new URL word list is generated. If a phishing target word is included in the word list or a word containing the substring of the other identification name is included, the site is considered to be a phishing website.

The direct use of an IP address as a domain name often means that the link points to a private page, such as http://124.248.212.46/caixa/, which is not the URL for a legal webpage. We propose Rule 5 accordingly.

Rule 5. An IP address is used as the domain name. This site is considered to be a phishing site.

In the URL standard naming convention, the first part is the protocol type, such as ftp or https. To confuse the user, the non-protocol part of the phishing URL will write the protocol field, such as http://www.oilchangeasheville.com/https:/www2.antander.com.br, in an attempt to pretend to be Santander's website. Thus, we propose Rule 6 based on these situations. Top-level domains should appear only in the domain name section of the URL and not in other sections. Therefore, we propose Rule 7.

Rule 6. The URL includes a protocol field written in the non-protocol part of the URL. This site is considered to be a phishing website.

Rule 7. The URL includes TLDs in the non-domain part. This type of site is considered to be a phishing website.

By assessing the nature of the URL through the rule base, we can avoid the process of feature extraction in certain phishing webpages. Moreover, using the LR classifier to perform a third detection avoids the false negatives caused by the size of the rule base.

3.3. *LR classifier-based detection phase*

We use a logistic regression classifier and extract more features from DNS, Whois, similarity with phishing vocabulary, lexical feature and HTML to determine the nature of the remaining pages to improve the adaptability and accuracy of the

Table 3 – High-frequency DNS statistics.

Data source	PhishTank		Yahoo	
High frequency DNS	domaincontrol.com	369	awsdnsc.com/net/org/co.uk	1775
	websitewelcome.com	97	ns.cloudfare.com	393
	hostgator.com	82	domaincontrol.com	317
	ns.cloudfare.com	62	dreamhost.com	219
	ezyreg.com	35	worldnic.com	180
Total of DNSs	2,823		8,819	

detection method. Next, we detail the constructed feature set and the feature extraction method.

3.3.1. *DNS doubt*

Mature DNS providers often take quick action on phishing webpages for security and commercial purposes. Compared with mature DNS providers, free or cheap DNS providers are the preferred choice for attackers. Therefore, this paper proposes a new detection feature, DNS doubt, which is inversely proportional to DNS credibility. We first count the DNS used by the URLs in the PhishTank (2776 phishing URLs) and Yahoo (5883 legal URLs) datasets to initialize the frequency of each DNS used on phishing and legitimate webpages. DNS doubt is calculated as follows:

$$F_d = \frac{P_{dns}}{Q_{dns}} \tag{1}$$

where P_{dns} is the frequency at which the DNS is used in the PhishTank URL dataset and Q_{dns} is the frequency at which the DNS is used in the Yahoo URL dataset. For example, as shown in Table 3, $P_{domaincontrol.com}$ is 0.13 and $Q_{domaincontrol.com}$ is 0.04. Therefore, the DNS doubt of domaincontrol.com is 3.25. For Q_{dns}, if the value is 0, then the URL cannot be properly resolved; subsequently, the DNS doubt is calculated as follows:

$$F_d = \frac{P_{dns\_max}}{Q_{dns\_min}} \tag{2}$$

where P_{dns_max} is the maximum frequency of the DNS used in the PhishTank URL dataset and Q_{dns_min} is the minimum frequency of the DNS used in the Yahoo URL dataset. Thus, P_{dns_max} is 0.13 (369/2823) and Q_{dns_min} is 0.0001 (1/8819). In the actual situation, this paper sets the DNS doubt of unrecorded DNS to zero. After the assessment is complete, the DNS will be included in the category according to the actual situation, and the DNS doubt of the DNS will be recalculated.

In addition to DNS doubt, this paper extracts a total of 7 features, including the number of IPs, IPs in the blacklist, RETRY values, TTL values, REFRESH values and EXPIRE values. The URL's DNS feature vector is defined as follows:

$$V_D = < F_{0\_1}, F_{ttl}, F_{retry}, F_{expire}, F_{refresh}, F_{ip\_num}, F_d > \tag{3}$$

3.3.2. *HTML*

1. Meta. The meta tag is a key tag in the HTML head area. This tag provides the most basic information about a document, including the document character set, language, author, keyword and webpage level. The rational use of descriptions and keywords in meta tag attributes is a crucial factor for search engines. For example, $< meta\ http-equiv =" refresh"\tilde{}content = "\ldots" >$ indicates that this page is a redirected page link. The contentâs text refers to the jump link. Link redirection is commonly used by attackers. Therefore, we calculate the number of $< meta\ http-equiv = "refresh"\tilde{}content = "\ldots" >$ type tags in the webpage, and we determine whether the original URL and the jump link are under the same domain. The feature extraction formula is as follows:

$$F_{H_{meta\_1}} = \begin{cases} 0, & \text{if } length(S_{meta}) = 0, \\ 1, & \text{if } length(S_{meta}) > 0. \end{cases} \tag{4}$$

$$F_{H_{meta\_2}} = \begin{cases} 0, & \text{if } length(S_{meta\_domain} \cap S_{domain}) = 0, \\ 1, & \text{if } length(S_{meta\_domain} \cap S_{domain}) > 0. \end{cases} \tag{5}$$

where S_{meta} is the set of $< meta\ http-equiv =" refresh"\ content = "\ldots" >$ tags, S_{meta_domain} represents the word set of the redirect link domain, and S_{domain} is the word set of the original domain.
2. Title. As mentioned previously, title tags tend to have the most important keywords of a webpage. However, certain webpages may not be directly identified as legal webpages based on the title. Therefore, we use Mode 3 to cut the content of the title tag and the original domain to obtain the intersection of the two sets. The feature extraction formula is as follows:

$$F_{H_{title}} = \begin{cases} 0, & \text{if } length(S_{title\_domain} \cap S_{domain}) = 0, \\ 1, & \text{if } length(S_{title\_domain} \cap S_{domain}) > 0. \end{cases} \tag{6}$$

where S_{title_domain} represents the word set of the title content.
3. A & Link. The $< link >$ tag defines the relationship between the document and external resources. The $< a >$ tag defines a hyperlink from one page to another. Extracting the properties of two types of tags plays an important role in determining the nature of webpages. For example, links within legitimate webpages generally point to their own domains, whereas phishing links often insert resources under other domains for the purpose of improving page rankings. We extract the href attribute of those tags as the detection feature.
   (1) HTTPS. Hypertext Transfer Protocol over Secure Socket Layer is a security-oriented HTTP protocol. HTTPS can help users confirm the identity of webpages and is widely used. Therefore, the use of HTTPS can, in a certain sense, improve the credibility of webpages. Therefore, we extract the use percentage of the HTTPS

protocol in the two types of tags in a webpage as one of the detection features. The extraction formula used for this feature is as follows:

$$F_{a_{https}} = \frac{Count(a\_tag\_https)}{Count(a\_tag)} \tag{7}$$

$$F_{link_{https}} = \frac{Count(link\_tag\_https)}{Count(link\_tag)} \tag{8}$$

where $Count(a\_tag)$ and $Count(link\_tag)$ represent the number of $<a>$ and $<link>$ tags in the webpage, respectively, and $Count(a\_tag\_https)$ and $Count(link\_tag\_https)$ indicate the number of links using the HTTPS protocol in the $<a>$ and $<link>$ tags, respectively.

(2) When the href = "#", the attribute field appears in the tag; clicking on the corresponding link will not jump to that link and will return the user to the current page. Users can be deceived by adding such tags in the phishing webpage. The extraction formula for this feature is as follows:

$$F_{a_{hash\_tag}} = \frac{Count(a\_hash\_tag)}{Count(a\_tag)} \tag{9}$$

$$F_{link_{hash\_tag}} = \frac{Count(link\_hash\_tag)}{Count(link\_tag)} \tag{10}$$

where $Count(a\_hash\_tag)$ and $Count(link\_hash\_tag)$ represent the number of $< a\ href = "\#" >$ and $< link\ href = "\#" >$ tags in the webpage, respectively.

(3) The attribute href = javascript:void(0) indicates a dead link, which means that clicking on the link does nothing. Standard does not recommend this method of use. The extraction formula of this feature is as follows:

$$F_{a_{javascript}} = \frac{Count(a\_javascript)}{Count(a\_tag)} \tag{11}$$

$$F_{link_{javascript}} = \frac{Count(link\_javascript)}{Count(link\_tag)} \tag{12}$$

where $Count(a\_javascript)$ and $Count(link\_javascript)$ represent the number of $< a\ href = javascript : void(0) >$ and $< link\ href = javascript : void(0) >$ tags in the webpage, respectively.

4) Legitimate webpages tend to point to web resources under their own domain names, whereas attackers try to increase the number of external links to improve the phishing webpage's search ranking and reputation. The extraction formula for this feature is as follows:

$$F_{a_{diff\_domain}} = \frac{Count(a\_diff\_domain)}{Count(a\_tag)} \tag{13}$$

$$F_{link_{diff\_domain}} = \frac{Count(link\_diff\_domain)}{Count(link\_tag)} \tag{14}$$

where $Count(a\_diff\_domain)$ and $Count(link\_diff\_domain)$ represent the number of links in the $<a>$ and $<link>$ tags that are not of the same domain name as the original URL, respectively.

(5) Adding top-level domains is a commonly used obfuscation technique. Therefore, we extract the average dot number of the related tag's links as one of the detection features. The extraction formula of this feature is as follows:

$$F_{a_{domain\_dot}} = \frac{Sum(a\_domain\_dot)}{Count(a\_tag)} \tag{15}$$

$$F_{link_{domain\_dot}} = \frac{Sum(link\_domain\_dot)}{Count(link\_tag)} \tag{16}$$

where $Sum(a\_domain\_dot)$ and $Sum(link\_domain\_dot)$ represent the sum of the domain's dot number for the $<a>$ and $<link>$ tags in the webpages, respectively.

(6) Some unusual symbols such as "@" are also added to the phishing URL. Therefore, we extract the number of links containing "@" in the relevant tag links as one of the detection features. The feature extraction formula is as follows:

$$F_{a_{at}} = \frac{Sum(a\_tag\_at)}{Count(a\_tag)} \tag{17}$$

$$F_{link_{at}} = \frac{Sum(link\_tag\_at)}{Count(link\_tag)} \tag{18}$$

where $Sum(a\_tag\_at)$ and $Sum(link\_tag\_at)$ represent the sum of the URLs containing "@" for the $<a>$ and $<link>$ tags in the webpages, respectively.

(7) The links in phishing webpages generally have hidden phishing target words. Therefore, the number of links using this obfuscation technique is extracted as one of the detection features. The feature extraction formula are as follows:

$$F_{a_{phish}} = \frac{Sum(a\_tag\_phish)}{Count(a\_tag)} \tag{19}$$

$$F_{link_{phish}} = \frac{Sum(link\_tag\_phish)}{Count(link\_tag)} \tag{20}$$

where $Sum(a\_tag\_phish)$ and $Sum(link\_tag\_phish)$ represent the sum of the obfuscation URLs for the $<a>$ and $<link>$ tags in the webpages, respectively.

Ultimately, we extract 17 features of the webpages HTML and define the feature vector of the HTML of the URL as follows:

$$V_H = < F_{H_{meta\_1}}, F_{H_{meta\_2}}, \ldots, F_{a_{domain\_dot}}, F_{link_{domain\_dot}} > \tag{21}$$

3.3.3. Similarities with phishing vocabulary

Attackers use time or other information as a seed and randomly generate URLs according to certain rules (Lin et al., 2015). This method is less costly than more complex methods and works well. Because of the inability to obtain unregistered phishing URLs and their vocabulary, the blacklist and bag-of-words methods cannot effectively detect such URLs. Yadav et al. (2010) concluded that the Jaccard similarity is the most suitable method for phishing detection via comparison with the K-L distance and edit distance. According to the research

of Yadav et al. (2010), we add the domain label and the vocabulary of alphanumeric characters in phishing URLs to the phishing vocabulary library. Ultimately, we extract a total of 3675 phishing words from 2776 phishing URLs (PhishTank).

A statistical analysis showed that the probabilities of having the same two tuples, three tuples and four tuples between any two URL strings were 95.7%, 75.8% and 33.6%, respectively (Huang et al., 2014). To reduce the error rate and time complexity, this paper calculates the similarity between legitimate URLs and phishing URLs by calculating the Jaccard similarity between triplets. When the number of identical triplets between strings is at least 50% of the number of triplets contained in the target string, then the similarity between the two URLs is calculated; otherwise, the Jaccard similarity between the two URLs is considered to be zero. To reduce the computational complexity of the operation, a word search is first performed. When the domain label of the URL is observed in the phishing vocabulary, then the similarity between the URL and the phishing vocabulary is set to one. The Jaccard similarity is calculated as follows:

$$JM = \frac{\|A \cap B\|}{\|A \cup B\|} \tag{22}$$

where $\|\ldots\|$ represents the number of elements in the collection, A and B represent the 3-gram set of a string. The similarity F_{sim} between the URL and the phishing vocabulary is calculated as follows:

$$JM_j = \frac{1}{n}\sum_{i=1}^{n}\frac{\|A_j \cap B_i\|}{\|A_j \cup B_i\|} \tag{23}$$

$$F_{sim} = \frac{1}{k}\sum_{j=1}^{k} JM_j \tag{24}$$

where JM_j is the Jaccard similarity between the jth word in the URLs word list and the phishing vocabulary, A_j is the 3-gram set of the jth word of the URL, B_i is the 3-gram set of the ith word of the phishing vocabulary, n is the total number of words in the phishing vocabulary, and k is the number of elements in the word set formed by splitting the URL by Mode 3.

The traditional method first builds a phishing vocabulary library composed of the phishing URL domain label and then calculates the average similarity between the URL domain name label and the phishing vocabulary library. Finally, the preset threshold is used to determine whether the website is a phishing website. Compared with the traditional Jaccard-based random domain name recognition method, this paper adds the source item of the phishing vocabulary and does not use the threshold as the basis for judging the URL. Instead, the Jaccard similarity is considered a feature value. Therefore, the Jaccard similarity feature vector is defined as follows:

$$V_S =< F_{sim} > \tag{25}$$

3.3.4. *Lexical features*

1. Information entropy. The probability of randomly generating different URL characters is usually equivalent. To identify these randomly generated phishing URLs, we introduce the information entropy of the URL alphabetic characters and numeric characters as detection features. The information entropy is calculated as follows:

$$H = \sum_{i=0}^{n} p_{x_i} \ln p_{x_i} \tag{26}$$

where p_{x_i} is the frequency with which the alpha or numeric character x_i appears in the URL.

2. Confused string. Mixed-case and alphanumeric strings are both common obfuscation techniques. First, the URL is segmented by Mode 3 to form a word list. Then, we count the number of uppercase letters that appear as inner characters of a word and the number of such words. We also count the number of alphanumeric words as one of the detection features.
3. Other features. In this paper, the URL is divided into four parts: domain, path, file and parameter. The length of each part, the length of the longest word of each part, the number of "-" or "_" characters and the number of dots are extracted. In addition, we extract a total of 37 features of a URL as shown in Table 4. The URL's lexical feature vector is defined as follows:

$$V_C =< F_1, F_2, \ldots, F_{37} > \tag{27}$$

3.3.5. *Whois features*

Based on a scenario in which an attacker uses the same information to register phishing webpages, we extract the information of the registration, update, and expiration date of the webpage, whether it is a private registration, whether the IP is locked, and whether the registration authority, registrant, subnet, country or region code, and autonomous system number of the URL are blacklisted. Ultimately, we extract 10 features of the webpageâs Whois and define the feature vector as follows:

$$V_W =< F_1, F_2, \ldots, F_9, F_{10} > \tag{28}$$

3.3.6. *LR classifier*

In this paper, five feature types are selected to define the feature vector of the URL. The feature vector of a URL is defined as follows:

$$V_{url} =< V_C, V_D, V_H, V_W, V_S > \tag{29}$$

where V_{url} is extracted from the undetected URLs in the first two detection phases. Then, the LR classifier is used for further testing to determine the nature of the webpage. For a given sample and tag $< V_{url}^{(i)}, y^{(i)} >, i = 1, \ldots, n, y^{(i)} \in \{-1, 1\}$ represents the label of the URL, and $V_{url}^{(i)} \in \mathbb{R}^n$ represents a feature vector of the URL. $\mathbb{R}^n$ refers to the n-dimensional real number space.

The logistic regression algorithm proposed by Cox (1958) is a common algorithm for solving classification problems. The decision function of the LR classifier is:

Table 4 – Lexical feature set.

Part	Feature	Amount
URL	URL length, longest word length, longest word length/URL length, longest word length - average length of the URLs words, mixed-case word number, uppercase letters, alphanumeric word number, letters number/URL length, dot number, the number of other separators, alphabetic entropy, numeric entropy, character entropy, singular characters(\, ^, @, etc.) number, % number, % number/URL length.	17
Domain	domain length, domain length/URL length, word number, longest word length, "-" and "_" number, port number, dot number.	7
Path	path deep, path length, path length/URL length, dot number, longest word length, path length/domain length.	6
File	file length, "-" and "_" amount, dot number.	3
Parameter	parameter length, variables number, longest variable length, "-" and "_" number.	4

$$h_w(V_{url}) = \frac{1}{1+\exp\left(w^T V_{url}\right)} \tag{30}$$

The probability model for the LR classifier to make a decision is:

$$P(y|V_{url};w) = h_w(V_{url})^{1\{y=1\}}\left(1-h_w(V_{url})\right)^{1\{y=-1\}} \tag{31}$$

$$l(w) = \sum_{i=1}^{n}\Big\{1\{y^{(i)}=1\}\ln\left(h_w(V_{url}^{(i)})\right) + 1\{y^{(i)}=-1\} \ln\left(1-h_w(V_{url}^{(i)})\right)\Big\} \tag{32}$$

where w is the feature weight vector that must be learned. $1\{\dots\}$ is an indicator function that takes a value of 1 if its argument is true, and 0 otherwise (i.e. $1\{True\}=1$, $1\{False\}=0$). We maximize the likelihood $l(w)$ of obtaining w. Then, we insert w into Eq. (30) and obtain the logistic regression classifier.

Regularization is an effective way to address overfitting problems, so we select regularized logistic regression (Li, 2012). We use the L1-regularized logistic regression to solve the following optimization problem to obtain the LR classifier (Fan et al., 2008):

$$\min_{w,b}\|w\| + C\sum_{i=1}^{n}\ln\left(\exp\left(-y^{(i)}(V_{url}^{(i)}w+b)\right)+1\right) \tag{33}$$

We can also use the L2-regularized logistic regression to solve the following optimization problem to obtain the LR classifier (Fan et al., 2008):

$$\min_{w,b}\frac{1}{2}w^T w + C\sum_{i=1}^{n}\ln\left(\exp\left(-y^{(i)}(V_{url}^{(i)}w+b)\right)+1\right) \tag{34}$$

where C is the penalty coefficient, b is the bias term and w is a weight vector. Note that C is a preset constant. The difference between Eqs. (33) and (34) is that the former can produce sparsity. Sparsity can transform many items in w to zero, which is conducive to the calculations.

By comparing the detection results of three supervised classification algorithms, i.e., logistic regression (LR with L1 norm regularization and LR with L2 norm regularization), SVM and Naive Bayes (NB), on the test dataset, we conclude that the LR with L1 norm regularization is more suitable for phishing detection. See Section 4 for additional details. Therefore, the SHLR uses the LR classifier.

Table 5 – Experimental data group.

Data group	1		2		3	
Data source	PhishTank	Yahoo	PhishTank	Yahoo	URLB	DMOZ
URL number	226	488	2776	5883	2216	9509

4. Experiments

The experimental environment of this paper is an Ubuntu 14.04 system with 32GB RAM. We use a total of four datasets: the PhishTank blacklist (2776 unique domain URLs) maintained by Internet users, Yahoo datasets (5883 unique domain URLs), URLBlacklist (2216 unique domain URLs) collected by SquidGuard, and DMOZ (9509 unique domain URLs), the world's largest directory community. PhishTank and URLBLacklist are phishing URL sets, and Yahoo and DMOZ are legal URL sets.

4.1. *Experimental setting*

The experiment is designed as four phases according to the detection process. Moreover, different data groups are selected based on the purpose of the experiment. Table 5 shows the experimental data groups. Table 6 shows the data and purposes of experiments 1-4. A detailed description of each experiment is provided as follows:

1. Experiment 1 is used to verify the effectiveness of the search engine-based detection phase. The main purposes of the experiment are as follows:
   (1) Use data group 1 to test the effectiveness of the proposed webpage keyword extraction methods. In this experiment, TF-IDF in CANTINA is the comparison object of the keyword extraction algorithm. CANTINA uses the TF-IDF algorithm to extract the keyword of the webpage as the search keyword and then treats a website as suspicious if its domain is absent from the top-N search

Table 6 – Experimental setting.

Experimental sequence	Data group	Experimental purpose
1	1,2	Test the search engine-based detection phase
2	2	Test the heuristic rule-based detection phase
3	2	Test the classifier-based detection phase
4	2,3	Test the effectiveness of the SHLR

results. However, we regard the title tag content of the webpage's HTML as the search keyword, and use the search engine to identify only legitimate webpages.

(2) Use data group 2 to select the search engine and the value of the top-*N*.

2. Experiment 2 is used to verify the effectiveness of the heuristic rule-based detection phase. The main purposes of the experiment are as follows:
   1) Use data group 2 to test the ability of each rule to identify phishing webpages.
   2) Use data group 2 to test the impact of reducing the phishing target word.
3. Experiment 3 is used to select the appropriate machine learning classification algorithm and verify the effectiveness of the LR classifier-based detection phase. This experiment uses data group 2. The main purposes of the experiment are as follows:
   (1) Compare the detection abilities of LR with L1 norm regularization, LR with L2 norm regularization, SVM, and NB using our constructed feature set. Furthermore, the classifier used in this detection phase is selected according to the experimental results.
   (2) Test the impact of DNS doubt on the accuracy, precision and recall of the LR classifier detection phase.
4. Experiment 4 is used to comprehensively analyze our SHLR detection methods. In addition, the experiment selects three representative detection methods for comparison: PhishDef (Le et al., 2010), BeyondB (Ma et al., 2009) and BigData (Lin et al., 2013). BigData extracts only the lexical features from the URL. BeyondB and PhishDef extract the lexical and host-based features from the URL, although the features that PhishDef extracts are more exhaustive. All three detection methods use the bag-of-words model. The main purposes of the experiment are as follows:
   (1) Use data group 2 to compare and analyze the effectiveness of the detection using the three methods individually as well as their effectiveness in different combinations.
   (2) Use data group 2 to compare SHLR with other detection methods in terms of accuracy, recall, F-Score, precision, average feature extraction time for a single URL, average recognition time for a single URL, PRC curve. Then, we use data group 3 to test the generalization ability of the SHLR.

4.2. *Experimental results*

4.2.1. *Experiment 1*

CANTINA extracts the top-*M* words from the webpage content ranked by the TF-IDF metric and searches those words and the webpage domain keywords in the search engine. The webpage is legal if the webpage domain matches the domain name of any of the top-*N* results. Otherwise, the webpage is regarded as a phishing page. In contrast with CANTINA, we use the webpage's title tag content as the keyword search. The webpage is legal if the webpage domain matches the domain name of any of the top-*N* results. Otherwise, further analysis is required. Table 7 shows the number of legitimate webpages identified by different detection methods and search engines. In this experiment, we set $N = 10$. As shown in Table 7, with the Baidu search engine, the SHLR can filter 63.9% of the legitimate webpages with an error rate of 0.14%, whereas CANTINA can filter only 28.1% of the legitimate webpages. The experiments show that the SHLR has a better detection ability for the identification of legitimate webpages.

Different search engines produce different results. Fig. 4 shows the effects of setting different *N* values for two search engines. As shown in Table 8 and Fig. 4, to address the needs of real-time detection, we choose the Baidu search engine and set $N = 10$.

As shown in Table 8, the search engine-based detection phase can directly determine 61.9% of legitimate webpages, and the error rate is only 0.04%. Based on PhishTank, the URL http://i.try8.info/taobao/try was mistakenly identified as a legitimate webpage. However, the URL is actually the website of Hangzhou Zhuanbao Technology Co. Ltd, and evidence that it is a phishing site has not been obtained. In addition, 1.7% of legitimate webpages did not have a title tag and could not be directly identified as legitimate webpages.

4.2.2. *Experiment 2*

A URL that cannot be directly identified as a legal webpage enters the second phase of the SHLR detection method. Table 9 shows the number of phishing webpages identified by different rules, and the results indicate that 22.9% of the phishing webpages can be detected with an error rate of the SHLR heuristic rule detection phase of only 0.36%. A total of 188 URLs were found to hide the phishing target words, and 143 phishing URLs did not comply with the naming conventions. Rule 1 determines that a webpage falsely categorized as legal is actually a phishing site because it is a magazine website that includes "walmart" as part of the URL file name, and the rule base identifies an anomaly of multiple identification names. Two ULRs are detected by Rule 7 because the URLs violate the naming convention. The remaining false detections are found by Rule 4 because three URLs use google and microsoft keywords despite not being the official URLs of google or microsoft. The other URLs are misjudged because of the use of the high-frequency word "apple".

Table 10 shows the effect of reducing the phishing target words on the detection effect. In this process, we use data group 2 and randomly select part of the phishing target words to participate in the detection. Table 10 shows that the number of phishing target words directly affects the heuristic rule detection results. In Table 11, we list information about

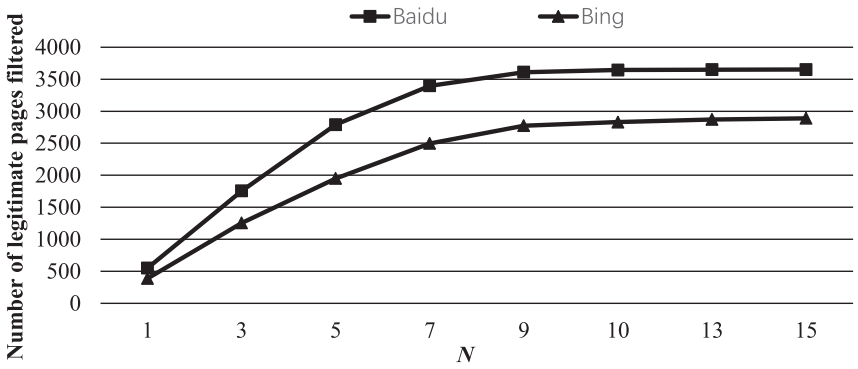

Fig. 4 – The effects of different N and search engine on detection.

Table 7 – Webpage recognition effect on data group 2.

Search engine	Baidu		Bing	
Detection method	CANTINA	SHLR	CANTINA	SHLR
Yahoo	137 (28.1%)	312 (63.9%)	135 (27.7%)	350 (71.7%)
PhishTank	0 (0%)	1 (0.14%)	0 (0%)	0 (0%)

Table 8 – The effect of search engine-based detection phase.

Data group 2	PhishTank	Yahoo
Unique domain URL number	2776	5883
The number (percentage) of URLs without title tag	482 (17.4%)	99 (1.7%)
The number (percentage) of URLs judged to be legal	1 (0.04%)	3,644 (61.9%)

high-frequency phishing target words. Dropbox, google, paypal, yahoo, alibaba, facebook and apple are popular subjects being forged. Therefore, we can improve the ability to filter phishing webpages by adding phishing target words.

4.2.3. *Experiment 3*

As shown in Table 9, because of the limited size of the rule base, the heuristic rule-based detection phase has a high false negative rate. Therefore, the SHLR introduces a classifier based on machine learning to improve the accuracy and adaptability of the detection methods. As shown in Tables 8 and 9, the number of unrecognized URLs in the first two stages is 4369, of which 2138 are URLs of PhishTank and 2231 are URLs of yahoo. For this part of the URL, the machine learning classifier is used for the third detection. We perform ten 10-fold cross-validations to assess the effects of different classifiers. As shown in Table 12, among the four classifiers, LR with L1 norm regularization is more suitable for phishing URL detection. Therefore, the SHLR uses LR with L1 norm regularization. In the following, LR refers to logistic regression with L1 norm regularization.

Table 13 shows the impact of DNS doubt on the accuracy, precision and recall of the LR classifier detection phase, indicating that DNS doubt has a good effect on phishing detection. Table 14 shows the effect of setting different C values on the classification accuracy, precision and recall of the LR. Tables 8, 9 and 12 show that the SHLR detection method achieves an

Table 9 – The effect of heuristic rule-based detection phase.

Data group 2	PhishTank	Yahoo
Unique domain URL amount	2776	5883
The number of URLs filtered by the search engine	1	3644
The number of remaining URLs	2775	2239
The number of URLs filtered by the Rule 1	229	1
The number of URLs filtered by the Rule 2	50	0
The number of URLs filtered by the Rule 3	3	0
The number of URLs filtered by the Rule 4	188	5
The number of URLs filtered by the Rule 5	24	0
The number of URLs filtered by the Rule 6	24	0
The number of URLs filtered by the Rule 7	119	2
The number(percentage) of URLs judged to be phishing	637 (22.9%)	8 (0.36%)

Table 10 – The impact of phishing target words amount.

Detection phase	Heuristic rule					
Data group	2					
Phishing target word amount	0	10	30	50	70	90
Filtered URLs amount of the Yahoo dataset	217	238	272	336	562	673
Filtered URLs amount of the PhishTank dataset	3	3	3	5	6	8

Table 11 – The High-frequency phishing target word.

Data group 2	PhishTank	
	dropbox	99
High	google	63
Frequency	paypal	32
Phishing	yahoo	31
Target	alibaba	31
Word	facebook	21
	apple	15
Total of phishing target word	90	

Table 12 – The comparison of the four classifier.

Classifier	Accuracy (%)	Recall (%)	F-Score (%)	Precision (%)
Bayes	92.2	92.1	92.1	91.1
LR (L1)	98.1	98	98	97.8
LR (L2)	97	97	97	97.2
SVM	78.9	72	70.5	64.2

Table 13 – The impact of DNS doubt on accuracy and precision.

DNS feature	Accuracy (%)	Precision (%)	Recall (%)
Ignore DNS	98	97.1	97
Binary	98.1	97.2	96.1
DNS doubt	98.1	97.8	98

Table 14 – The effect of different C on detection.

Sequence	C	Accuracy (%)	Precision (%)	Recall (%)
1	1	97.8	97.7	97.8
2	2	97.9	97.6	97.9
3	3	98.0	97.8	98
4	4	98.1	97.8	98
5	5	98.1	97.8	98
6	10	98.1	97.8	98

accuracy of 98.9%, a recall rate of 99.1%, an F-Score of 99% and a precision of 98.9%.

4.2.4. *Experiment 4*

For a better analysis of the SHLR detection method, we compare and analyze the detection methods using the three detection methods alone and in different combinations. Fig. 5 shows the flow of other detection methods. As shown in Fig. 5, the S method uses only the Baidu search engine to determine the nature of the webpage. That is, if the domain name of the original URL is in the top-10 search results, the webpage is considered to be a legitimate webpage; otherwise it is considered to be a phishing webpage. The H method uses the only rule base proposed by us to detect phishing webpages; that is, webpages matching the rules are phishing webpages and are otherwise legitimate webpages. The LR method refers to judging the webpage using only LR based on the feature set constructed in this paper. S+H, H+LR, H+S+LR and other methods are combinations of the above detection methods. As shown in Table 15, SHLR performed well on each evaluation indicator. The analysis of the SHLR detection methods and other detection processes will be presented in detail in Section 5.

Table 16 shows a comprehensive comparison of the four detection methods: BeyondB, PhishDef, BigData and SHLR. Compared with the traditional detection methods, the SHLR improved the accuracy by 0% –3.2%, the recall rate by 0.2%–3.4%, the F-Score value by 0.1%–3.3% and the precision by 0.01%–3.3%. In addition, the time consumed by SHLR is greatly reduced. Based on the comprehensive recognition time and test performance, the SHLR performs better than the other three detection methods. Fig. 6 shows the PRC comparison of the four detection methods, which indicates that the SHLR has the largest area under the curve (AUC).

An experiment was conducted to test the adaptability of the SHLR. In this experiment, the initial DNS and malicious IP list use the data provided by data group 2. Table 17 compares the detection effects of the four detection methods on data group 3, and the SHLR has better adaptability than the other methods.

5. Analysis and discussion

5.1. *Analysis*

Based on the results in Fig. 5 and Table 15, the ability of the three detection methods to independently detect phishing webpages is unequal, but each method has its own merits. First, a benefit is achieved from the search strategy of the search engine, and the search engine-based phishing detection has a high recall rate (99.9%). However, the search engine's own search strategy causes many legitimate pages with low ranks to be misjudged as phishing pages. As a result, the accuracy and precision of this method are low. Second, the experiment validates that the rule base constructed in this paper can achieve good detection of URL obfuscation technology. The precision of the heuristic rule-based phishing

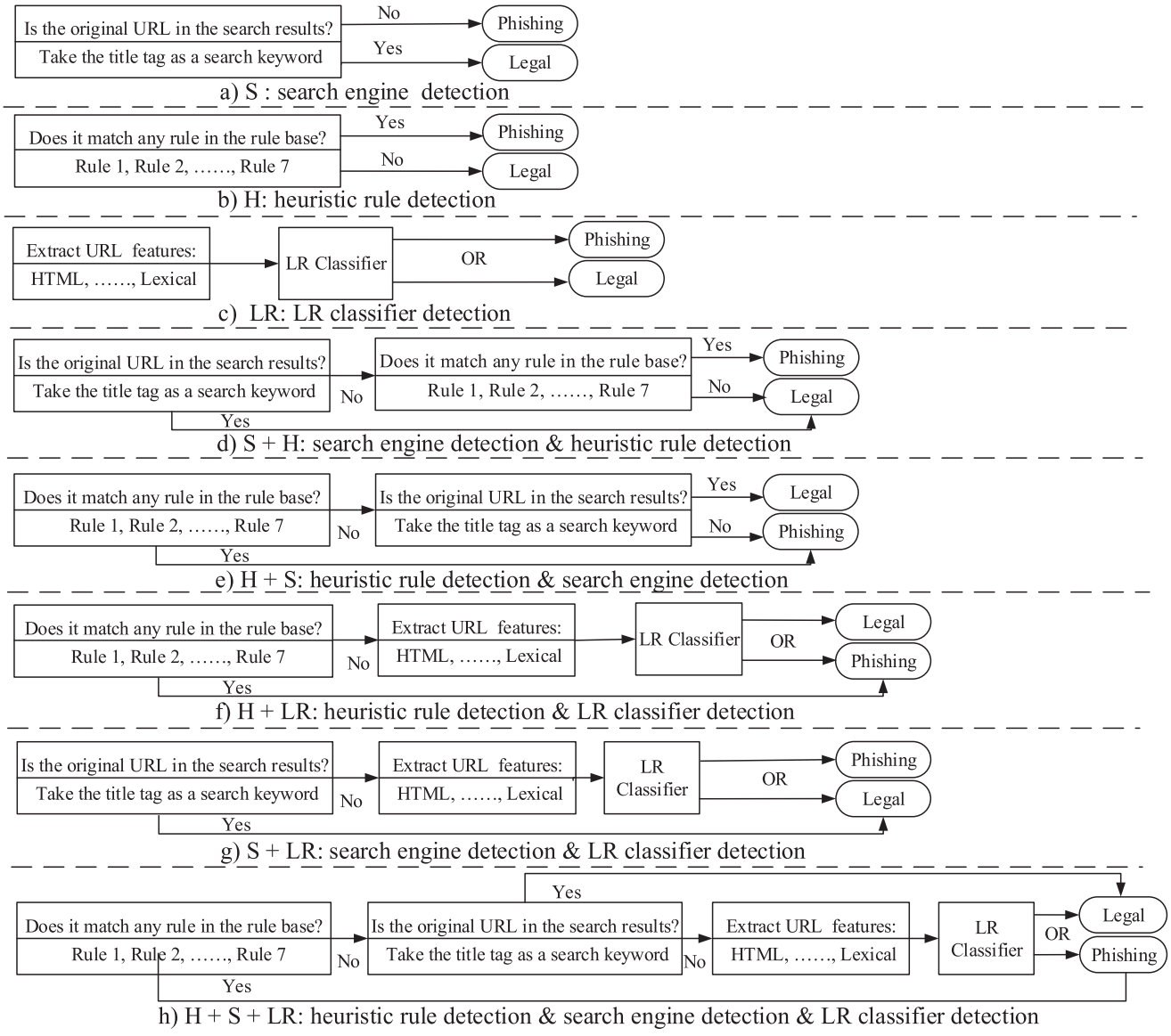

Fig. 5 – Process of different methods.

Table 15 – The comparison of different detection methods or processes.

Method	Accuracy (%)	Recall (%)	F-Score (%)	Precision (%)	Feature extraction time $\mu s/url$	Recognition time $\mu s/url$
S	74.1	99.9	85.1	55.3	240.9	10.7
H	74.9	22.9	35.1	95.2	0	2.9
LR	98.6	98.6	98.6	97.8	581.2	38
S + H	76	22.9	35.2	99.2	240.9	12.4
H + S	71	99.4	86.4	57.7	225.7	12.8
H + LR	98	98	98	96.8	536.3	37.9
S + LR	91.8	97.6	94.6	90.7	577.5	32.7
H + S + LR	98.4	97.1	97.7	89	497.81	30.8
SHLR	98.9	99.1	99	98.9	521.6	29.1

Table 16 – The comparison of four detection methods.

Method	Feature vector latitude	Accuracy (%)	Recall (%)	F-Score (%)	Precision (%)	Feature extraction time $\mu s/url$	Recognition time $\mu s/url$
BeyondB	6982	95.7	95.7	95.7	94.9	6651.5	60.1
PhishDef	6995	98.9	98.9	98.9	97.4	6761.6	59.2
BigData	6014	97.1	97.1	97.1	96.3	6829.1	72.1
SHLR	78	98.9	99.1	99	98.9	521.6	29.1

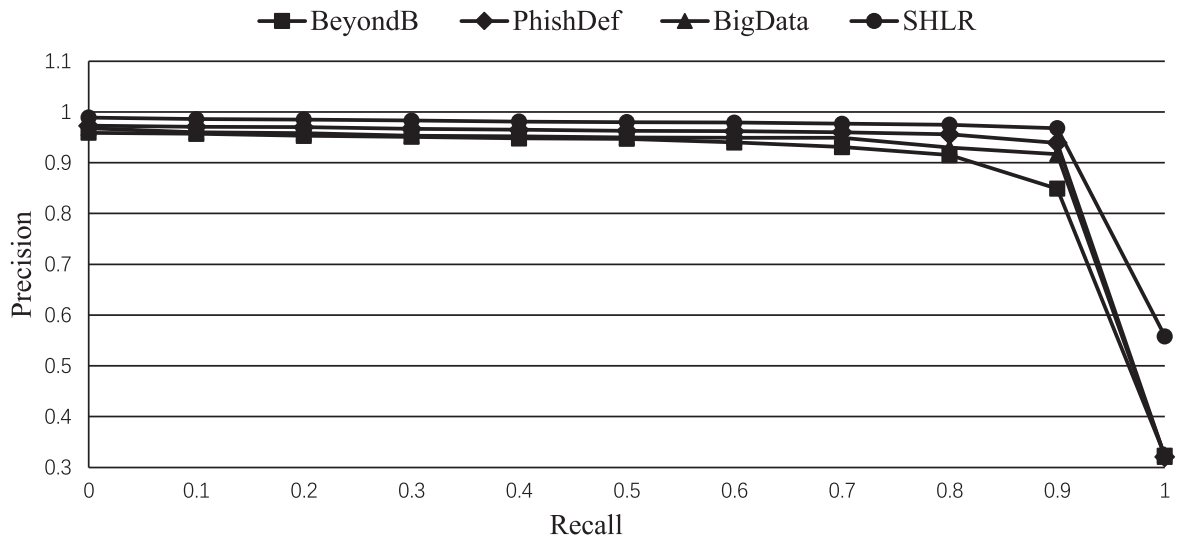

Fig. 6 – The comparison of the precision and recall curve.

Table 17 – The comparison of the detection effects on data group 3.

Method	Accuracy (%)	Recall (%)	F-Score (%)	Precision (%)
BeyondB	94.2	94.2	93.1	90.5
PhishDef	96.9	97	96.9	94.1
BigData	97.4	97.4	97.4	94.5
SHLR	98	98.5	98.2	94.8

detection reaches 95.2%. Moreover, since the rule-based detection method does not need to extract features and perform efficient matching of strings, it requires the shortest detection time. However, the problem with this method is clear. Limited by the size of the rule base, the detection method is powerless for other types of phishing webpages. Therefore, this method has a high false negative rate and low accuracy. Finally, the LR classifier-based detection is outstanding in terms of accuracy, recall, F-score and precision. However, the method consumes considerable time to extract features and has lowest detection efficiency, which is not conducive to real-time detection. We consider combining the mothods to benefit from each of their advantages.

The first strategy is to combine two methods. Table 15 shows that for the S + H and H + S methods, the recall rate of S + H is higher, and the precision of H + S is higher. The combination of H and S causes S + H to focus more on detecting phishing webpages and H + S to focus more on detecting legitimate webpages. If H is placed after S, because the rules in H can only detect phishing webpages related to abnormal information, other types of webpages are identified as legitimate webpages. Therefore, the S + H detection method causes many phishing pages to be unreported. If S is placed after H, the search engine is more inclined to detect legitimate webpages with high rankings, and for webpages not in the top-10 are identified as phishing pages. Therefore, the H + S detection method will result in misjudgment of many legitimate webpages. Thus, the accuracy of the above two detection methods is not high. However, the combination of S and LR or H and LR greatly improves the performance on the four indicators, but with an increase in the time required for detection. Therefore, we need to comprehensively consider the factors of detection time and detection effect to achieve an appropriate combination of H, S and LR.

The second strategy is to combine all three methods. In addition to the extraction time, Table 15 shows that SHLR is better than H+S+LR in terms of various evaluation indicators. One problems occurs in the heuristic rule-based detection phase because many legitimate URLs are also more or less in violation of the URL naming standards. Simultaneously, for some navigational pages, the identification names of other webpages are also placed in the path of the URL. In addition, some high-frequency words such as apple appears in the URLs of many legitimate webpages. These behaviors are

considered to be uses of URL obfuscation techniques, so the detection method determines that the webpage is a phishing page. Therefore, the search engine is first used to filter out legitimate webpages that may have abnormal conditions, which can help to reduce the false positive and improve the detection.

Based on the above analysis, this paper combines the advantages of the three detection methods and proposes the SHLR phishing webpage detection method. The advantages of this method are the ability of search engines to quickly identify legitimate webpages, the ability of heuristic rules to quickly identify phishing, and the adaptability of the LR. The experiments show that the detection method can effectively reduce the number of detections of legitimate webpages, and effectively detect phishing webpages using URL obfuscation technology. In contrast to traditional detection methods, the SHLR does not use the bag-of-words model; therefore, the feature vector latitude of the SHLR is much lower than that of the other methods, which reduces the time complexity and the required computational resources. By ignoring network communication delays, the SHLR greatly reduces the detection time and requires only 521.6 microseconds to extract features and 29.1 microseconds to judge the URL. The experiments show that the SHLR is more suitable for real-time detection than the other three detection methods. Meanwhile, benefiting from the detection feature set and the detection process we constructed, the SHLR is also significantly better than the traditional detection methods.

As an actual application, we can develop a browser plugin based on the SHLR. When the user clicks on a link, the plugin is triggered to determine the nature of the webpage. If the webpage is determined to be a phishing webpage, a warning message is displayed at a prominent location in the browser. According to the experimental results, the average delay of determining the nature of a webpage is only 550.7 microseconds, which satisfies the needs of real-time detection. Because the plugin is concurrent with user activity, it does not affect the page loading speed.

5.2. *Discussion*

In this section, we briefly analyze the problems we encountered during the experiment in the hope of further enhancing the detection capability of the method in the future.

In China, people are gradually becoming accustomed to paying online, and millions of dollars are being turned around online every day. This phenomenon is convenient for people's lives, but it is also a huge economic temptation for cyber attackers. Although various network security methods have been widely adopted, and good prevention and detection effects have been achieved, the types of attack methods change daily, with no steadfast security strategy. In the foreseeable future, phishing attacks will continue to seriously threaten the credibility and property safety of Internet users.

First, many browsers trigger danger notifications when visiting a link that has an IP as its domain or even block the webpage. Because phishing pages that use an IP as their domain name are easily identified, attackers must use domain name servers to conceal the phishing website. If the webpages on the side of the DNS can be identified and timely feedback is provided, then the financial losses caused by phishing will be largely avoided. The proposed DNS doubt parameter represents the expectation that malicious webpages, such as phishing attacks, can be effectively filtered on the DNS side in the future. Therefore, a set of DNS reputation evaluation standards can be used as an important reference for detecting various cyberattacks in the future.

Second, e-mail is still the main means of communication, and accurately identifying phishing e-mails is also an important topic in phishing detection. Phishing e-mail detection should pay more attention to the detection accuracy. The SHLR detection method can automatically extract and identify the nature of the URLs in the e-mail. However, to improve the detection, we should expand the size of the rule base and the detection feature set according to the characteristics of the phishing e-mail. For example, the header and text information of the e-mail are used as detection features. In fact, phishing e-mails are more dependent on the user's psychology, and their content usually appears to be authentic and trustworthy. Therefore, we can use psychological knowledge to improve the accuracy of the phishing e-mail detection.

Third, spear phishing attacks have gradually become a major trend in phishing attacks. Such attacks are long-lasting and are not easy to identify. Therefore, effectively detecting such well-planned phishing attacks should be an important direction for future studies. To better respond to spear phishing attacks, better detection methods for new phishing attacks must be continuously provided. However, users should be encouraged to learn about phishing scams to avoid falling into their traps because no security strategy is consistently effective, and phishing detection requires the joint efforts of all involved parties.

Finally, we believe that no phishing attack detection method is always applicable. On the one hand, researchers largely carry the burden of detecting various new types of online frauds in a timely manner and providing effective solutions. On the other hand, identifying common cyber frauds should be a basic skill of Internet users obtained through general training.

6. Conclusion

We propose a simple and effective phishing detection method to identify the obfuscation techniques commonly used by phishing websites and meet the needs of real-time phishing detection. This method can effectively filter legal webpages and detect phishing webpages by adopting escape technology. First, to implement escaping technology that identifies many unrelated words in the phishing webpage, we use the title tag content of the webpage as the webpage keywords and filter legal webpages quickly with the help of the Baidu search engine. Second, a rule-based detection method is adopted for certain obfuscation techniques to avoid the feature extraction of phishing webpages and meet the needs of real-time detection. Third, the SHLR combines the advantages of a search engine-based method, heuristic rule-based method and machine learning-based method. While satisfying

real-time detection, the SHLR method reduces the false positive rate caused by the lack of rules and improves the adaptability and accuracy of the method.

Conflict of Interest

None.

Acknowledgments

We would like to express our gratitude to four anonymous reviewers for their criticism and comments on improving the quality of the manuscript. This work was partially supported by the Natural Science Foundation of China (Grant nos: 61303231, 61433012, U1435215), the Major State Basic Research Development Program of China (Grant no: 2014CB340506) and the China Postdoctoral Science Foundation funded project (Grant no: 2018T110829).

REFERENCES

Abdelhamid N, Ayesh A, Thabtah F. Phishing detection based associative classification data mining. Expert Syst Appl 2014;41(13):5948–59.

Aleroud A, Zhou L. Phishing environments, techniques, and countermeasures: a survey. Comput Secur 2017;68:160–96.

Alsharnouby M, Alaca F, Chiasson S. Why phishing still works: User strategies for combating phishing attacks. Int J Human Comput Stud 2015;82:69–82.

APWG. In: Report, APWG Internet Policy Committee. APWG, Global phishing survey for 2016; 2017. http://docs.apwg.org/reports/APWGGlobal-Phishing-Report-2015-2016.pdf

Bilge L, Kirda E, Kruegel C, Balduzzi M. Exposure: finding malicious domains using passive DNS analysis. Proceedings of the network and distributed system security symposium, NDSS. California, USA: San Diego, 2011.

Blythe M, Petrie H, Clark JA. F for fake:four studies on how we fall for phish. In: Proceedings of the SIGCHI conference on human factors in computing systems; 2011. p. 3469–78.

Chang EH, Kang LC, Sze SN, Wei KT. Phishing detection via identification of website identity. In: Proceedings of the international conference on it convergence and security; 2013. p. 1–4.

Chen TC, Dick S, Miller J. Detecting visually similar web pages:application to phishing detection. ACM Trans Internet Technol 2010;10(2):1–38.

Chiba D, Yagi T, Akiyama M, Shibahara T, Yada T, Mori T, Goto S. Domainprofiler: discovering domain names abused in future. In: Proceedings of the IEEE/IFIP international conference on dependable systems and networks; 2016. p. 491–502.

Choi H, Zhu BB, Lee H. Detecting malicious web links and identifying their attack types. Proceedings of the Usenix conference on web application development, 2011. 11–11

Choi YH, Kim TG, Choi SJ, Lee CW. Automatic detection for javascript obfuscation attacks in web pages through string pattern analysis. Int J Secur Appl 2009;4(2):13–26.

Chu W, Zhu BB, Xue F, Guan X. Protect sensitive sites from phishing attacks using features extractable from inaccessible phishing urls. In: Proceedings of the IEEE international conference on communications; 2013. p. 1990–4.

Cox DR. The regression analysis of binary sequences. J R Stat Soc 1958;20(2):215–42.

Dunlop M, Groat S, Shelly D. Goldphish: Using images for content-based phishing analysis. In: Proceedings of the international conference on internet monitoring & protection; 2010. p. 123–8.

Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ. Liblinear: a library for large linear classification. J Mach Learn Res 2008;9(9):1871–4.

Gastellier-Prevost S, Granadillo GG, Laurent M. Decisive heuristics to differentiate legitimate from phishing sites. In: Proceedings of the network and information systems security; 2011. p. 1–9.

Gowtham R, Krishnamurthi I. A comprehensive and efficacious architecture for detecting phishing webpages. Comput Secur 2014;40(40):23–37.

Hadi W, Aburub F, Alhawari S. A new fast associative classification algorithm for detecting phishing websites. Appl Soft Comput 2016;48:729–34.

Holz T, Gorecki C, Rieck K, Freiling FC. Measuring and detecting fast-flux service networks. In: Proceedings of the network and distributed system security symposium, NDSS. California, USA: San Diego; 2008. p. 487–92.

Hou YT, Chang Y, Chen T, Laih CS, Chen CM. Malicious web content detection by machine learning. Expert Syst Appl 2010;37(1):55–60.

Hu Z, Chiong R, Pranata I, Susilo W, Bao Y. Identifying malicious web domains using machine learning techniques with online credibility and performance data. Evol Comput 2016:5186–94.

Huang D, Xu K, Pei J. Malicious URL detection by dynamically mining patterns without pre-defined elements. World Wide Web-Internet Web Inf Syst 2014;17(6):1375–94.

Kang LC, Chang EH, Sze SN, Wei KT. Utilisation of website logo for phishing detection. Comput Secur 2015;54:16–26.

Le A, Markopoulou A, Faloutsos M. Phishdef: Url names say it all. In: Proceedings of the IEEE INFOCOM; 2010. p. 191–5.

Li H. Statistical learning method. Beijing: Tsinghua University Press; 2012.

Li Y, Yang L, Ding J. A minimum enclosing ball-based support vector machine approach for detection of phishing websites. Optik - Int J Light Electron Opt 2016;127(1):345–51.

Lin HL, Li Y, Wang WP, Yue YL, Lin Z. Efficient segment pattern based method for malicious URL detection. J Commun 2015;36:141–8.

Lin MS, Chiu CY, Lee YJ, Pao HK. Malicious URL filtering a big data application. In: Proceedings of the IEEE international conference on big data; 2013. p. 589–96.

Liu X, Jia CF, Liu GY, Hu ZC, Wang D. Collaborative defending scheme against malicious web pages based on social trust. J Commun 2012;33(12):11–18.

Ma J, Saul LK, Savage S, Voelker GM. Beyond blacklists:learning to detect malicious web sites from suspicious urls. In: Proceedings of theACM SIGKDD international conference on knowledge discovery and data mining. France: Paris; 2009. p. 1245–54.

Ma J, Saul LK, Savage S, Voelker GM. Learning to detect malicious URLs. ACM Trans Intell Syst Technol 2011;2(3):1–24.

Mcgrath DK, Gupta M. Behind phishing: an examination of phisher modi operandi. Proceedings of the usenix workshop on large-scale exploits and emergent threats. CA, USA: San Francisco, 2008.

Miyamoto D, Hazeyama H, Kadobayashi Y. An evaluation of machine learning-based methods for detection of phishing sites. In: Proceedings of the advances in neuro-information processing, international conference, ICONIP. New Zealand: Auckland; 2009. p. 539–46. Revised Selected Papers

Moghimi M, Varjani AY. New rule-based phishing detection method. Expert Syst Appl 2016;53:231–42.

Mohammad RM, Thabtah F, Mccluskey L. Tutorial and critical analysis of phishing websites methods. Comput Sci Rev 2015;17(C):1–24.

Parsons K, Mccormac A, Pattinson M, Butavicius M, Jerram C. The design of phishing studies: challenges for researchers. Comput Secur 2015;52:194–206.
Prakash P, Kumar M, Kompella RR, Gupta M. Phishnet: predictive blacklisting to detect phishing attacks. In: Proceedings of the IEEE INFOCOM; 2010. p. 1–5.
Qihoo360. In: Report. 2016 china internet security report; 2017. http://zt.360.cn/1101061855.php?dtid=1101062514&did=490278985
Ramanathan V, Wechsler H. Phishing detection and impersonated entity discovery using conditional random field and latent Dirichlet allocation. Comput Secur 2013;34(34):123–39.
Sahoo D., Liu C., Hoi S.C.H.. Malicious url detection using machine learning: A survey. arXiv:1701.071792017.
Seifert C, Welch I, Komisarczuk P, Aval CU, Endicott-Popovsky B. Identification of malicious web pages through analysis of underlying dns and web server relationships. In: Proceedings of the IEEE conference on local computer networks, LCN; 2008. p. 935–41.
Sha HZ, Liu QY, Liu TW, Zhou Z, Guo L, Fang BX. Survey on malicious webpage detection research. Chin J Comput 2016;39(3):529–42.
Symantec. In: Report. Internet security threat report; 2016. https://www.symantec.com/content/dam/symantec/docs/reports/istr-21-2016-en.pdf
Symantec. In: Report. Internet security threat report; 2017. https://www.symantec.com/content/dam/symantec/docs/reports/istr-22-2017-en.pdf
Tan CL, Kang LC, Wong KS, Sze SN. Phishwho: phishing webpage detection via identity keywords extraction and target domain name finder. Decis Support Syst 2016;88:18–27.
Whittaker C, Ryner B, Nazif M. Large-scale automatic classification of phishing pages. Proceedings of the network and distributed system security symposium, NDSS. California, USA: San Diego, 2010.
Xiang G, Hong JI. A hybrid phish detection approach by identity discovery and keywords retrieval. In: Proceedings of the international conference on world wide web, WWW. Spain: Madrid; 2009. p. 571–80.
Yadav S, Reddy AKK, Reddy ALN, Ranjan S. Detecting algorithmically generated malicious domain names. In: Proceedings of the ACM SIGCOMM conference on internet measurement. Australia: Melbourne; 2010. p. 48–61.
Yang H, Ma X, Du K, Li Z, Duan H, Su X, Liu G, Geng Z, Wu J. How to learn Klingon without a dictionary: detection and measurement of black keywords used by the underground economy. In: Proceedings of the security and privacy, S & P. California, USA: San Jose; 2017. p. 751–69.
Zhang W, Zhou Y, Xu L, Xu B. A method of detecting phishing web pages based on hungarian matching algorithm. Chin J Comput 2010;33(10):1963–75.
Zhang Y, Hong JI, Cranor LF. Cantina:a content-based approach to detecting phishing web sites. In: Proceedings of the international conference on world wide web, WWW. Alberta, Canada: Banff; 2007. p. 639–48.

Yan Ding received his B.S. degree in Software Engineering from North University of China, Taiyuan, China in 2014. He is currently a graduate student at Xinjiang University and a student member of the China Computer Federation. His research interests include machine learning, artificial intelligence and network security.

Nurbol Luktarhan received his B.S., M.S. and Ph.D. degrees in Computer Science from Jilin University, Changchun, China in 2005, 2008 and 2010. From May 2015 to July 2016, he was a visiting scholar at Tsinghua University. He is currently an associate professor at Xinjiang University and the deputy director of the Network and Information Technology Center of Xinjiang University. His research interests include network security and data mining.

Dr. Keqin Li is a SUNY Distinguished Professor of computer science in the State University of New York. He is also a Distinguished Professor of Chinese National Recruitment Program of Global Experts (1000 Plan) at Hunan University, China. He was an Intellectual Ventures endowed visiting chair professor at the National Laboratory for Information Science and Technology, Tsinghua University, Beijing, China, during 2011–2014. His current research interests include parallel computing and high-performance computing, distributed computing, energy-efficient computing and communication, heterogeneous computing systems, cloud computing, big data computing, CPU-GPU hybrid and cooperative computing, multicore computing, storage and file systems, wireless communication networks, sensor networks, peer-to-peer file sharing systems, mobile computing, service computing, Internet of things and cyber-physical systems. He has published over 550 journal articles, book chapters, and refereed conference papers, and has received several best paper awards. He is currently serving or has served on the editorial boards of IEEE Transactions on Parallel and Distributed Systems, IEEE Transactions on Computers, IEEE Transactions on Cloud Computing, IEEE Transactions on Services Computing, and IEEE Transactions on Sustainable Computing. He is an IEEE Fellow.

Wushour Slamu is an academician of the Chinese Academy of Engineering. He has presided over 7 projects of The 863 Program, 5 projects of the National Natural Science Foundation of China and 30 projects of provincial and ministerial high-tech research programs including the National Development and Reform Commission. Additionally, he has published more than 120 papers and 9 books. He is currently a professor of College of Information Science and Engineering at Xinjiang University, the Vice chairman of the China Chinese Information Society, the Director of Xinjiang Multilingual Information Technology Key Laboratory, and Honorary Chairman of Xinjiang Union of IT. His research interests include natural language processing, machine learning, and artificial intelligence.