Algorithm for Detection of Phishing URL and Web Scraping

Parth Shah and P. Beaulah Soundarabai
Department of Computer Science, CHRIST (Deemed to be University), Bengaluru, India

Abstract: Social Networking Sites (SNS) gives administrations to group of people or to individuals to exchange information and data. SNS contains tremendous measure of individual data such as user/client profiles, user/client relations and user/client activities. Usually, the data which are stored or collected by the SNS are personal in nature and it usually consists of very sensitive yet valuable information. Once, the information comes out this might lead to various cases of harassment or even monetary misfortune and identity theft. So, securing client protection in SNS is fundamental. Numerous danger warning and assaults have been found in SNS. This study shows a security danger display for all the URL which will be entered, type by the user/client. It is grouped under 3 sections: safe, malicious and phishing. Once, the websites are classified they are stored in the form of dataset. Also, the websites which are classified as phishing also goes through the process of web scraping. The content of the entire website is processed and if any inappropriate word, sentences or phases are detected, the website is described as an inappropriate site for viewing. Database of the inappropriate sites will be created through our system model and will be made available to public for social usage.

Key words: Social networking sites, web scrapper, security, identity, websites, inappropriate

INTRODUCTION

Social Networking Sites (SNS) (NaliniPriya and Asswini, 2015) are also known as Online Informal Communities (OIC). Over the past few years, it has been noticed that the number of users getting connected to the social networking sites has increased to a huge extent. Social networking has taken its prime position when it comes to friendly interactions as well as professional communications. It is the process of extending one’s business with help of contacts by making associates through various social networking sites or online informal communities. Social networking sites usually, generates a list of people with whom the user wants to create a relationship or not. The social networking sites are known as social networking websites also.

The word social networking site (Ahmad, 2011) was first coined or began in the year 1997 and it was followed by many other sites. The different types of social networking websites are: Facebook, LinkedIn, Twitter, Myspace, etc. Over the time it has been noticed that almost 85% of the world population are present in various social networking sites. According to a statistical analysis, Facebook alone has over 2 billion users during the end of 2017. Some of the examples of the social networking sites are:

Facebook: Facebook is an online social media site as well as service was developed by Mark Zuckerberg along with few of his friends. Facebook was deployed globally on 2004, February 4. Report taken from the year 2010 shows that there are more than 500 million users currently active on Facebook.

LinkedIn: LinkedIn is a social networking platform where various researchers, companies share information related to researches made by their department and look for suitable employees for their company. It is a place where one can share their views and ask questions where renowned person might come forward to answer it.

Instagram: Instagram is a photo and video sharing social platform where people mainly post about their weekend getaways, parties and other happening things occurring in their lives. It is now currently owned by Facebook Inc. It was launched in 2010 October and later it was acquired by Facebook on April 2012. Statistics shows that Instagram had a stronghold of about 800 million users at the end of 2017.

Social network security: The term security in the social networking sites (Donald Zulu et al., 2014) usually, refers to the state of being tension free or danger free. The social networking sites face high risks with respect to security threats. With the increasing number of online social networking users these sites not only bring the families, friends and the businesses together; the intruders who want to get the data for a wrong intension are also attracted to the social networking sites. For a newcomer to the internet field, social networking sites are a
phishing attack with very less failure rates. The algorithm was deployed in a Windows XP system. The algorithm is based on end-host system and the name given to the algorithm is “link Guard algorithm”. The algorithm is used to find out whether a website is legitimate or a phishing one. It uses the current domain name and matches with the domain present in the white list. There are 2 main objectives. They are: domain and IP matching one and if the domain doesn’t matches with the white list then hyperlinks features are extracted to decide whether the website is a legitimate or a phishing one.

Naresh et al. (2013) have proposed another anti phishing algorithm by using the various characteristics of hyperlinks which were used in phishing attacks. The algorithm is based on end-host system and the name given to the algorithm is “link Guard algorithm”. The algorithm was deployed in a Windows XP system. The algorithm was able to detect both known and unknown types of phishing attack with very less failure rates.

Basnet et al. (2014) has done a study on the various features or techniques to detect phishing URL’s. A lot of discussion was done on the various search engine methods. The parameters which were taken for the study are those which are available or applicable features of the URL. An analysis was done between the various different kind of machine learning techniques which are used for detecting a website URL whether it is phished or not with respect to the efficiency of their respective proposed model. Various different kind of features such as lexical features, reputation based features, search engine based features, keyword based features were used.

Damodaram (2016) did an extensive study on phishing. This study mainly focusses on what phishing really is what are the different types of phishing tools how phishing causes a threat to the e-commerce company how phishing is planned what are the necessary steps that needs to be taken and also provides some suggestion or measures on how to reduce the chances of phishing.

Background study

Phishing: Phishing depicts bunches of various tricks, however, the sign of each is that it includes casualties surrendering individual data to con artists under an appearance of authenticity. These tricks use a type of advanced correspondence routed to a potential target for the most part by means of email, to bait casualties into trading off their gadget or personality. For instance, phishers may get in touch with you by email, professing to be your bank, an online life webpage or some other establishment you trust. They’ll regularly ask for that you click a connection which prompts a site they control, to “sign into” your record or they’ll request that you tap on a malevolent connection that will introduce malware on your machine.

Whom does phishing affect? The extend or reach of phishing can’t be exaggerated as it’s presently a threat tormenting each level of society. Basic email tricks and how they identify with phishing and keeping in mind that email is the default correspondences medium for phishing in all trustworthiness any computerized content based medium-like talk customers or content informing (regularly called SMiShing or SMS phishing) is helpful for phishing. Basically, on the off chance that researchers have an email address, mobile phone, web based life account or impart online in any capacity researchers are a potential focus for a phishing.

Issues with respect to research in phishing: The primary issue with respect to research in phishing are: there doesn’t exist a single package of phishing which can detect or identify each different kind of Phish. Spotting different types of phishing emails is quite a hard task. The basic advice which is given on different phishing package are to look for standard mistakes like grammatical mistake and checking for spelling mistakes. But the main issue with this advice is that there are people whose grammar and spelling may be good. They may not make such mistakes.

Reacting to messages and tapping on links is an indispensable piece of work. Endeavoring to stop the propensity for clicking isn’t just to a great degree troublesome, however is it what you need?
Web scraping: Web scraping can be defined as the automated technique or the process of sampling and analyzing of the data or particular information of a particular website. The law with respect to scratching web information is perplexing, fluffy and ready for change, however that is another issue. It isn’t so much, that nobody is considering or composing, about the morals in scratching but instead that both those scratching and those being scratched can’t concede to fundamental standards.

Research issues related with web scraping: Scratching whole HTML website pages is truly simple and scaling such a scraper isn’t troublesome either.

The primary choice is hard on the grounds that locales change constantly and keeping up hundreds or thousands of scrapers can get exceptionally tedious. The second alternative is hard in light of the fact that the engineer basically needs to make code that can consequently decide the structure of a site and the structure of each important page of that site. Doing the process of web scraping at 20% exactness isn’t that difficult. However, doing the same process at 99% precision is a genuine test. One that no single innovation can fathom. It is required to utilize a mix of numerous ways to deal with quality control the information that comes in and auto-amend when errors show up. Scaling a solitary scraper arrangement will end up harder when you are attempting to achieve dynamic substance, i.e. when you are trying to scrap inside a ordered list.

MATERIALS AND METHODS

The main objective of this research is to reduce the chances of phishing that is happening with the social networking sites by proposing an algorithm which will check the URL of the websites and after processing the URL it will determine whether the website is a malicious one or a legitimate one. The URL’s will be stored as a dataset for the algorithm. If the URL is present in the dataset then it is categorized as a malicious one. With the help of a web scrapper the website is searched for inappropriate contents in the web pages and the URL is added to the database of “abusive sites” if inappropriate content is found. The system will also block the malicious as well as abusive sites for ever from accessing by issuing an appropriate warning message. The database of “abusive sites” can be uploaded to internet servers for societal usage.

Model design: Web sites which are malicious or phishing largely supports various criminal activities which hinders the development of web services. So, different kinds of steps or measures have been implemented, so that, the users don’t make the same mistake of visiting that side again.

Researchers propose a systematic approach to assort the web sites into 3 main categories: safe, malicious and phishing. It is based on the mechanism that the system only checks the URL of the website without going in depth towards the contents of the website.

If a website is found to be phishing then with the help of ParseHub researchers will scrap the website in order to find any content whether it is appropriate for viewers or it has to be removed. The websites which will be found having inappropriate contents will be then added to a table or list which will be helpful for societal usage.

Design approach: URLs of the websites are separated into 3 classes:

Safe: Websites are classified as safe if the algorithm doesn’t detect any suspicious activities.

Malicious: Websites are classified as malicious if the algorithm detects suspicious activities such as fake online dating sites, various questionnaire’s and surveys.

Phishing: Website will be classified as phishing sites if the websites is found to be either a fake webpage asking for log in details or other important information such as credit card details.

Tools used

ParseHub: ParseHub is a powerful web scraping tool. It is used for extracting data from various webpages. It can be used to select the portion of the website which needs to be scrapped. Data which are available in the sites are only available for the view purpose. Web Scraping is a toll which help us to scrap the website and also with it allows us to store the data in various format such as CSV.

Python: Python is a language interpreted, object-oriented programming language which consist of dynamic semantics. Python is usually used as it provide very high level of data structures to the users and when it is joined with dynamic writing of code, it makes the rapid process of application development very attractive. It is very simple and easy to use and also it has a huge collection of built in library which can be used for various purposes for our application.

Tkinter: Tk has been a core part from the beginning. The graphical interface which is developed is done by using tkinter. It gives a powerful and stage autonomous windowing toolbox that is accessible to Python developers utilizing the tkinter bundle and its expansion, the tkinter.tix and the tkinter.ttk modules.

Google transparency report: A transparency report is an announcement issued all the time by an organization, unveiling an assortment of insights identified with
demands for client information, records or substance. Transparency reports by and large reveal how as often as possible and under what specialist governments have asked for or requested information or records over a specific time frame (Fig. 1). This type of corporate straightforwardness enables people in general to recognize what private data governments have accessed through court orders and court subpoenas, among different strategies. Some straightforwardness reports depict how frequently because of government activity or under copyright arrangements, content was evacuated. Google initially propelled a straightforwardness report in 2010 with Twitter following in 2012. Google transparency report was used to check whether the website is a malicious one or phishing one.

**PhishTank:** PhishTank is like a hostile environment for phishing site. PhishTank was designed in October, 2006 by business visionary David Ulevitch as a branch of OpenDNS. The company provides a phishing check framework which is network based. It also provides us with the facilities that if a user wants to upload a URL as Phish he can do so. On doing so, the URL will be then asked to be voted as a Phish or not a Phish by the registered users and on the basis of that it will be decided that whether the site is a phished one or not. In 2018, it was reported that PhishTank will be decommissioned before the finish of 2018.

**Dataset:** To check the efficiency of the proposed method/framework, researchers have taken the dataset of around 1000 websites which have been collected from two specific sources. They are: PhishTank and OpenPhish. The dataset consist of various URL which have been reported by the users as either malware or phishing. It consist of sites which are active as of now or the URL has been removed over the course of time. The dataset also consist of various authorized or reliable websites which includes various payment gateways, social media websites like Facebook, Twitter as well as various video blogs, blogs etc.

**Evaluation criteria:** Using the Tkinter, the GUI of the proposed system has been developed. A URL is usually taken from the dataset and then it is put on the rectangular
box. On clicking the submit button, given single URL, researchers extract its features and categorize them into 3 categories:

**Lexical features:** Lexical features are an important property of URL which is generally present in the textual form. These properties incorporate the length of the hostname, the length of the whole URL and also the quantity of dabs in the URL, twofold component for every token in the hostname (delimited by '.') and in the way URL (strings delimited by '?', '?', ':', '=' and '_'). This is too known as a "pack of-words".

**Site popularity features:** Over the time, it has been seen that the malicious or the phishing websites tend to be less popular than the original one. It is because the domain name and other components tend to be similar to the original one but it is not and so, it is ranked lower in the site popularity.

**Host-based features:** It has been noticed that the illegal or malicious or phishing sites tend to save their domain and data in areas or region which are less famous.

**Comparison with other framework:** After the feature extraction of the website is done it is then declared whether the particular website or the URL of the website is safe, malicious or phishing one. Once, it has detected as a phishing, safe or malicious the website is then compared with the PhishTank and Google transparency report. The output of all the three methods are taken into consideration.

**Web scraping:** If a webpage is reported to be phished then the website is again scrapped using the ParseHub tool and it looks for inappropriate words or content present in the webpage. If there is any content which might be inappropriate the webpage is declared not suitable for viewing and it is stored in a table which will act as a database which can be uploaded for various societal usage.

### RESULTS AND DISCUSSION

A number of experiments have been carried out to check the accuracy value of our proposed system. The proposed method is been compared with the PhishTank and Google transparency report.

**Total website taken for expriment purpose:** About 100 websites:

**Tools used:**
- Google transparency report
- PhishTank
- Proposed method

**Google transparency report:** This is the report which shows that for 100 websites 77 websites are safe, 15 are phishing, 1 of them have some pages of phishing and 7 of them are malicious according to Google transparency report (Fig. 2 and Table 1).

![Fig. 2: Chart representing Google transparency report](image)

<table>
<thead>
<tr>
<th>URL</th>
<th>Google transparency report</th>
<th>PhishTank</th>
<th>Phishing algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://goal.com">http://goal.com</a></td>
<td>Some pages are phishing</td>
<td>Safe</td>
<td>Safe</td>
</tr>
<tr>
<td><a href="http://google.com.bd">http://google.com.bd</a></td>
<td>Safe</td>
<td>Safe</td>
<td>Safe</td>
</tr>
<tr>
<td><a href="http://ndtv.com">http://ndtv.com</a></td>
<td>Safe</td>
<td>Safe</td>
<td>Safe</td>
</tr>
<tr>
<td><a href="http://huffingtonpost.com">http://huffingtonpost.com</a></td>
<td>Safe</td>
<td>Safe</td>
<td>Safe</td>
</tr>
<tr>
<td><a href="http://apple.com">http://apple.com</a></td>
<td>Safe</td>
<td>Safe</td>
<td>Safe</td>
</tr>
<tr>
<td><a href="http://cigarettes.com">http://cigarettes.com</a></td>
<td>Safe</td>
<td>Safe</td>
<td>Phishing</td>
</tr>
<tr>
<td><a href="http://buycheappillsonline.com">http://buycheappillsonline.com</a></td>
<td>Safe</td>
<td>Safe</td>
<td>Phishing</td>
</tr>
<tr>
<td><a href="http://rapidfatloss.com">http://rapidfatloss.com</a></td>
<td>Safe</td>
<td>Safe</td>
<td>Phishing</td>
</tr>
<tr>
<td><a href="http://ladamejeanne.fr">http://ladamejeanne.fr</a></td>
<td>Safe</td>
<td>Safe</td>
<td>Phishing</td>
</tr>
<tr>
<td><a href="https://christuniversity.in/">https://christuniversity.in/</a></td>
<td>Safe</td>
<td>Safe</td>
<td>Phishing</td>
</tr>
<tr>
<td><a href="http://avantpack.ru">http://avantpack.ru</a></td>
<td>Safe</td>
<td>Safe</td>
<td>Phishing</td>
</tr>
<tr>
<td><a href="http://protectionpage-110121a.esy.es/">http://protectionpage-110121a.esy.es/</a></td>
<td>Safe</td>
<td>Phishing</td>
<td>Phishing</td>
</tr>
<tr>
<td><a href="http://youtube.com/watch?v=oVBOvn-xrEY">http://youtube.com/watch?v=oVBOvn-xrEY</a></td>
<td>Safe</td>
<td>Safe</td>
<td>Safe</td>
</tr>
<tr>
<td><a href="http://www.natalcielo.890m.com/">http://www.natalcielo.890m.com/</a></td>
<td>Safe</td>
<td>Phishing</td>
<td>Phishing</td>
</tr>
<tr>
<td><a href="http://printforprofits.com">http://printforprofits.com</a></td>
<td>Safe</td>
<td>Safe</td>
<td>Phishing</td>
</tr>
</tbody>
</table>
PhishTank: According to the searching mechanism for detecting URL in PhishTank, 67 of the sites are safe, 1 of them is malicious and 32 of the sites have been detected as phishing (Fig. 3).

On carrying out experiments with our proposed method, it has been found that out of 100 websites which were taken for the experiments 38 of them are safe, 43 are of phishing category and 19 of them are malicious (Fig. 4).

Web scraping: From the URL which has been marked as phishing websites by our proposed method (Fig. 5 and Table 2), researchers do a web scraping of that particular website and if any inappropriate content is found then the website is classified as abusive sites and the link is stored in the form of table which can be used for societal usage.
Table 2: Classification on the basis of web scraping

<table>
<thead>
<tr>
<th>URL</th>
<th>Abusive sites</th>
<th>Abusive content</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="https://www.indy100.com/article/british-swear-words-ranked-ofcom-7340446">https://www.indy100.com/article/british-swear-words-ranked-ofcom-7340446</a></td>
<td>Yes</td>
<td>Arse, Cunt etc.</td>
</tr>
<tr>
<td><a href="http://goal.com">http://goal.com</a></td>
<td>No</td>
<td>-</td>
</tr>
<tr>
<td><a href="http://ndtv.com">http://ndtv.com</a></td>
<td>No</td>
<td>-</td>
</tr>
</tbody>
</table>

CONCLUSION

It has been found that the results which are produced by our phishing algorithm is much better as compared to the results with Google transparency report and PhishTank. It is because there are many sites in our dataset and URL which google or PhishTank are unable to detect or figure out whether they are phishing website or malicious website or safe website. Our algorithm is able to figure out information about each of the URL in our database whether it is a safe, malicious or phishing efficiently. The accuracy rate of the proposed method is much better than the Google or PhishTank. This study has mainly focused on the issues which are related to phishing and a method is proposed which classifies the websites into three categories: safe, phishing and malicious; a comparative analysis is also done with Google transparency report and PhishTank. In future, the contents of the entire website with its various linking pages will also be tested for the safe content. A database of such unsafe site would be generated and made available for the public.

REFERENCES


