DETECTION OF PHISHING URL BASED ON TEXT FEATURE EXTRACTION

Abhinav Jagtap, Mahesh Adhav, Mayur Kadhane, Harshad Phalak
B.E.- Computer Engineering, Dr. D. Y. Patil School of Engineering, Pune, Maharashtra, India.

***Abstract*- Phishing is a technique of gaining personal information of user's from various websites. Sometimes it redirects the user to phished webpage to gain information of user like username, password, account and credit card details etc. Our main ambition here is to design system to provide safeguard to users against phishing attacks. Our working mainly focuses on use of terms and URL's from web page to detect possible phishing patterns from web pages of phishing websites. Process initiates with parsing of web page to extract plain text terms and URL's. Further detected terms are fed to TF-IDF and URL weighting system to identify importance of each detected term. Later search engine lookup is carried out for most important terms which help to detect possible victim URLs for given input website. Finally WHOIS lookup is used to compare registration details of websites to correctly categorize website as phishing or legitimate**

***Index Terms*- Search engine, WHOIS, DOM (Document Object Model), TF-IDF (Term Frequency Inverse/Document Frequency)**

I. INTRODUCTION

The phishing webpage is a replica of other sites that look like legitimate one [1]. Phishing is another crime than the hacking. Phishing also called as brand spoofing. It is a technique of taking personal information of user's from various websites. Sometimes it redirects the user to phish webpage to gain information of user like username, password, account and credit card details etc. Phishing attacks are becoming common now a days because of large financial gain with very little efforts. In phishing the victims have asked to enter their information Such as account numbers and password. The information you have entered is taken by phisher and gain access to that account. No of tricks to fool user duplicate content from official websites, duplicate email addresses looking like legitimate one, advertise of home page that redirect webpage to phish one.

The more dangerous phishing technique is called spear phishing. The spear phisher means he knows little bit about you, your name, email address.No it's not a sport, it is a scam and you are a target. Spear phishing is an email phishing that can be done by an individual that you know. But it isn't. It's from the same criminal hackers who want your credit card and bank account no, passwords, and the financial information on your pc. Spear phishing is a targeted email scam with the purpose of obtaining unauthorized access to sensitive data.

To protect user from being phished proposed a system approach for detecting a phish WebPages using a Plain text and URL's extraction, Domain name extraction, TF-IDF weight calculations with search engine lookup. In plain text extraction it extracts a text from web page by identifying relation between words in the text. Extract URL's using title, description, keywords and Meta data from a webpage. TF-IDF is a term frequency inverse/document frequency. TF-IDF is used for occurrence of words in webpage by making matrix of it. Shroff word frequency list are used for weight calculation. Then highest frequency word are selected and given them to search engine *for searching top 30 related urls and then comparing in WHO-IS lookup by domain name owner to detect whether a webpage is phish or not.*

II. RELATED WORK

In this section survey of various methodologies used for detecting phishing websites are provided. Even though different techniques are available (e.g. user's browser based dynamic security, predefined rules for web page creation by Website Company, visual and DOM tree similarity based approach and comparing URLs with blacklisted sites) our main focus is to incorporate use of text mining technique for classifying website as legitimate and fake. Thus following review illustrates application and observation of only text mining based website phishing detection techniques (Refer Table 2.0).

No	Paper Details	Methodology	Observation
1.	CANTINA: A Content-Based Approach to Detecting Phishing Web Sites. (*WWW 2007*)	❑ Traditional TF-IDF approach is used to identify more important words from testing web pages. ❑ Detected top 5 words with highest weight age used to retrieve search engine results. ❑ If domain name match found with top-n search engine results websites, website considered as legitimate otherwise phishing.	• Simple approach to detect phishing websites. • To get better results from traditional TF-IDF approach it is required that web page contains large no. of words.
2.	Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs (*ACM 2009*)	❑ Only lexical features of website URL and details of host are considered for phishing website detection. ❑ Classifiers are used for binary classification (e.g. phishing or benign) of websites.	• Less time required and applicable for checking any type of URL. • Website contents are ignored thus no way to detect whose contents are copied by website phisher.
3.	Teaching Johnny Not to Fall for Phish (*ACM 2010*)	❑ Phishing awareness and learning tool is modeled by author. ❑ Incorporates anti phishing technique against fraudulent email and websites. ❑ User's behavior tracked e.g. Whether or not user accesses the suspicious mail.	• Improves the self detection of phishing mails and websites by providing learning techniques. • It is difficult to force users to read instructions.
4	Phish Net: Predictive Blacklisting to Detect Phishing Attacks (IEEE 2010)	❑ Technique for generating suspicious phishing URLs from already available URL blacklist and approximate URL matching suggested ❑ Incorporates mechanism for checking validity of generated URL's and matching content of suspicious URL websites with possible victim websites	• Easy approach to predict URL's which can be generated by phishers

III. PROPOSED SYSTEM

We are proposing phishing website detection system which can categorize website as either phishing or legitimate. Preferred use of term weighting based phishing pattern detection reduces the false consideration of phishing website as legitimate one and vice versa (Refer Fig.3.1). Following are the main objectives of the system:

- ➢ Extract terms and URLs from web page using DOM parser.
- ➢ Identify important terms (brand name) using TF-IDF and URL weighting scheme.
- ➢ Search results for brand name using search engine API.
- ➢ Identify victim website for detected phishing website.

In following section in detail explanation of proposed system architecture is given which helps to achieve mentioned objectives.

System architecture

3.1.1 Webpage parsing: This phase parses the collected web pages, one page at a time, to get terms and URLs from respective web page (Refer Fig.3.1). HTML parser is used to create a Document Object Model (DOM). The Document Object Model (www.w3.org/DOM) is a standard for making and controlling inmemory representations of HTML (and XML) content. DOM is presented through tree style

structure which is transformable and can be used to reproduce a complete page [3]. *DOM (Document Object Model):*The DOM is an interface for working with the structure of XML & HTML documents. A project of the W3C, the DOM was designed to provide a set of objects and methods to make life simpler for programmers. (Technically, the DOM predates XML; the early DOM work was meant for HTML documents.).Ideally, you should be able to write a program that uses one DOM-compliant parser to process an XML document, and then switch to another DOM-compliant parser without changing your code. When you parse an XML document with a DOM parser, you get a hierarchical data structure (a DOM tree) that represents everything the parser found in the XML document. You can then use functions of the DOM to manipulate the tree. You can search for things in the tree, move branches around, add new branches, or delete parts of the tree.

Web page parsing phase is divided into three sub modules. Functionality of each module described in following section:

3.1.1.1 Plain text extraction*:* In this we extract the HTML code of the web pages which lies in the source code. Files that contain markup or other meta-data are generally considered plain-text, as long as the entirety remains in directly human-readable form(as in HTML, XML, and so on). In our work we are not considering all the words from the source of web page as plain text. Rather only textual content from specific tags is considered as plain text. List of those tags is as follows:

<meta>...</meta>
<title>...</title>
<body>...</body>.

3.1.1.2 URL Extraction: URL Extractor can be used to extract URLs. Generally URLs are placed inside the anchor tag where href property denotes the URL path. URL can be relative path or absolute path. In our work we are focusing on extracting absolute URLs. For the same purpose, anchor tag is parsed. Anchor tag has following syntax structure:

a. *<a href = > ...</a>*
b. *<a href = > ...</a>*
c. *<a href = > ...</a>*

3.1.1.3 Domain name Extraction*:* To get the information about domain of input web page is extracted which is later useful for identifying web page is legitimate or not. Generally domain information proceeded by hostname & '.' Symbol and followed by '/' & remaining URL path. Thus regular expression technique is used to extract domain of web page accurately.

3.1.2 Shroff's word Frequency: The shroff word frequency is used to assign the weight to the words that have been obtained from the web page parser. After the generation the weight is calculated and mostly important keywords are searched for that (Refer Fig.3.1).

TF-IDF [Term Frequency-Inverse Document Frequency]: The TF-IDF weight is a weight utilized as a part of data recovery and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus [2].

TF: Term Frequency, which measures how as often as possible a term occurs in a record. Since every record is different in length, it is possible that a term would appear much more time in long records than shorter one. Thus, the term frequency is often divided by the record length as a way of normalization:

TF (t) = (Number of times term t appears in a record) / (Total number of terms in the record).

IDF*:* Inverse Document Frequency, which measures how important a term is *in the record*. While computing TF, all terms are considered equally important. However it is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance [2]. Thus we need to weight down the frequent terms while scale up the rare ones, by computing the following:

IDF (t) = log_e (Total number of documents / Number of documents with term t in it)

3.1.3 *URL Based Weight:* In this mainly URL's are weighted according to the occurrences in the document of the source code (Refer Fig.3.1).

Search Engine lookup: In this lookup, Google search engine API used to search keyword related information. In this phase the three keywords are selected from the shroff's word frequency given to the search engine and the result is generated in the form of top 30 results related to the keyword. After that extract domain name of that related URLs is extracted and is given to next phase of the URL verification.

Ex. URL for getting result from Google search engine API:

https://www.googleapis.com/customsearch/v1?key=<API Key>&cx=<Google Custom Search Engine Control Panel Setting ID>& q=<Keywords to search>

3.1.4 URL Verification*:* In This phase we use the WHOIS API for verifying owner information of processed web page domain & mostly detected domain in search engine results (Refer Fig.3.1).

WHOIS lookup: WHOIS provides information about owner of any second-level domain name who has registered it with VeriSign (or with Network Solutions, which was acquired by VeriSign) [2]. Network Solutions was originally the only Internet registrar of the com, net, and org domain names) and many domain names are still registered with VeriSign. In this, the extracted domain name of webpage and extracted domain name of related URL's are given to WHOIS lookup. And also in this phase the comparison to justify us whether the URL is legitimate or phished. This is a step of comparing legitimate domain name is Dl and query of domain name is Dq.

Ex. URL for retrieving domain information from API:

http://www.whoisxmlapi.com/whoisserver/WhoisService?domainName=<domain name>&username=xxxxx& password=xxxx

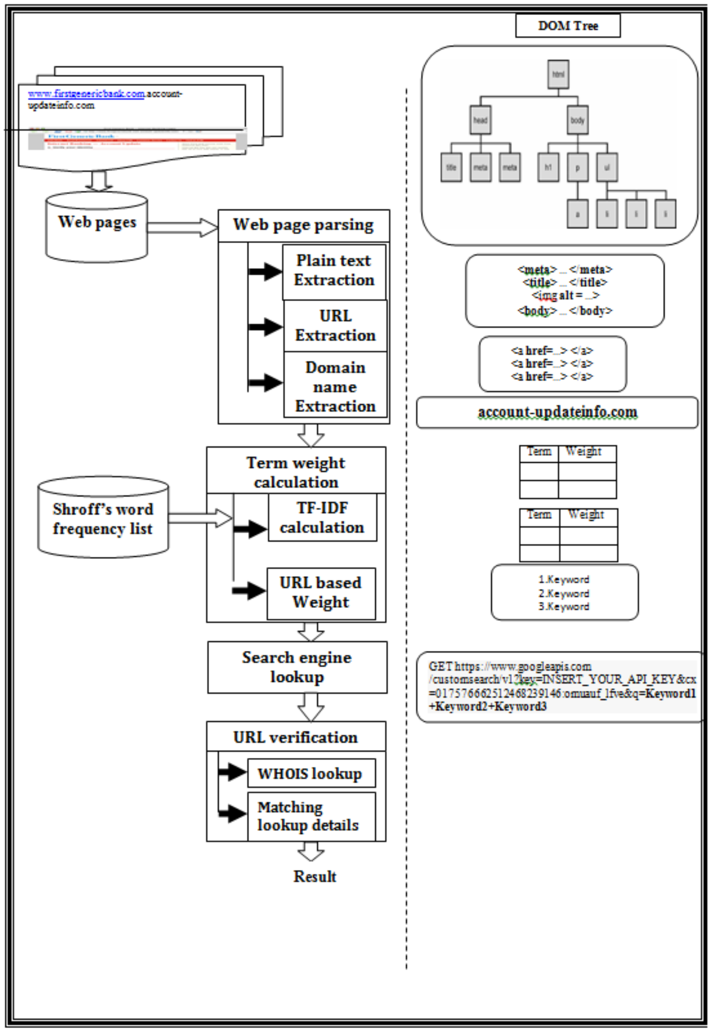

Fig 3.1 Detection of phishing webpage URL

IV. ALGORITHM:

1. Randomly select web page from collection of legitimate & phished web pages.
2. Parse web page to extract textual contents from tags (e.g. meta, anchor, title & body) and tag attributes (e.g. alt).
3. For the words in extracted textual content perform:
    a. TF/IDF weight calculation using revised formula (based on Shroff's word frequency list).
    b. Term weight modification based on word occurrence in URL.

Finally select three words with highest weight.

4. Provide highest word weights to search engine API and collect top 30 results.
5. Identify most occurred domain from collected search engine results.
6. Compare domain information of web page domain & detected domain (in previous phase) using WHOIS lookup API.
7. Mark web page as legitimate, if domain information match or as phished otherwise.

V. RESULT

Collection of web pages (legitimate & phished) by Choon Lin Tan et al. [1] is used for our experimentation. In mentioned dataset, phishing web pages was collected from Phish Tank e.g. repository which maintains phishing web pages along with justification for phishingness while legitimate web pages are manually collected. Accuracy of result is calculated on following basis:

IS Result: Implemented System result for processed web page

PT Result: Phish Tank status for processed web page.

	Positive	Negative
TRUE	1. IS Result: Phish PT Result: Phish 2. IS Result: Legitimate PT Result: -	IS Result: Phish PT Result: -
FALSE	1. IS Result: Phish PT Result: -	IS Result: Legitimate PT Result: Phish

VI. APPLICATIONS

Current implementation of proposed system (e.g. as standalone application) is helpful for data mining researchers to identify various data patterns used for phishing websites.

Implemented system can be useful for end users if implemented as "Plug-in" to the browsers where browser can take care of doing mentioned process for user recommended websites e.g. banking, social networking & so on.

VII. FUTURE SCOPE:

The future development of community detection system will be concentrated on improving phishing website detection system results by incorporating more than single search engine as it would reduce the chances of getting biased search engine results.

VIII. CONCLUSION

Implemented phishing website detection system considers term and URLs weights for identification of possible victims for phished web pages. Previous approaches rely on traditional term weighting approaches e.g. TF-IDF technique which does not provides accurate results due to lack of sufficient textual content and need of processing no. of web pages. Proposed system adopts URL and shroff's word list weights in weighting scheme which eliminates need of processing multiple web pages. Following parameters are going too considered while evaluating system:

(i) Identification of true phishing websites.
(ii) Identification of true victims of phishing website.

ACKNOWLEDGEMENT

We would like to take this opportunity to thank our internal guide **Prof. Nitin Hambir** for giving us all the help and guidance we needed. We are really grateful to him for his kind support. His valuable suggestions were very helpful.

We are also grateful to **Prof. S. S. Das**, Head of Computer Engineering Department, DYPSOE, Lohegaon, Pune and to our Project Co-coordinator **Prof. J. L. Chaudhari** for their indispensable support, suggestions and motivation during the entire course of the project.

REFERENCES:

[1] Choon Lin Tan, Kang Leng Chiewy, San Nah Sze, "Phishing Website Detection Using URL-Assisted

Brand Name Weighting System", *In IEEE International Symposium on Intelligent Signal Processing and*

Communication Systems (ISPACS), Pages 54-59, 2014

[2] (2014, June) Phishing guide part 1. PayPal Inc. [Online].
Available:
https://www.paypal.com/au/webapps/mpp/security/generalunderstandphishing

[3] *(2014, June) Phishing activity trends report, 2nd half / 2010.Anti-Phishing Working Group. [Online].Available: http://docs.apwg.org/reports/apwg report h2 2010.pdf*

[4] *Y. Zhang, J. I. Hong, and L. F. Cranor, "Cantina: A content-based approach to detecting phishing web sites," in Proceedings of the 16th International Conference on World Wide Web, ser. WWW '07.New York,NY,US: ACM, 2007, pp. 639–648. [Online]. Available:http://doi.acm.org/10.1145/1242572.1242659*

BIODATA AND CONTACT ADDRESSES OF AUTHORS

1. **ABHINAV JAGTAP**
B.E.- COMPUTER ENGINEERING,
DR. D. Y. PATIL SCHOOL OF ENGINEERING,
PUNE,MAHARASHTRA, INDIA.
7350889919

2. **HARSHAD PHALAK**
B.E.- COMPUTER ENGINEERING,
DR. D. Y. PATIL SCHOOL OF ENGINEERING,
PUNE, MAHARASHTRA, INDIA.
9403731367

3. **MAHESH ADHAV**
B.E.- COMPUTER ENGINEERING,
DR. D. Y. PATIL SCHOOL OF ENGINEERING,
PUNE, MAHARASHTRA,INDIA.
8055479404

4. **MAYUR KADHANE**
B.E.- COMPUTER ENGINEERING,
DR. D. Y. PATIL SCHOOL OF ENGINEERING,
PUNE, MAHARASHTRA, INDIA.
9967603930