IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES
ISSN PRINT 2319 1775 Online 2320 7876
Research Paper **© 2012 IJFANS. All Rights Reserved** UGC CARE Listed (Group -I) Journal Volume 11, S.Iss 6, 2022

DETECTION OF PHISHING URL BASED ON TEXT FEATURE EXTRACTION

[1]Shalini Gupta, [2]Shubha Jain, [3]Vijatashwa Awasthi, [4]Amit Sabharwal, [5]Mohit Kumar, [6]Ajay Gaur, [7]Waseed Mohamad Khan

Department of Computer Applications, Axis Institute of Higher Education, Kanpur, Uttar Pradesh, India

Abstract

Phishing poses a significant threat to users' personal information by tricking them into divulging sensitive data through deceptive websites. This study aims to develop a robust system to safeguard users against such attacks.

Our approach focuses on analyzing the content and URLs of web pages to identify potential phishing patterns. The process begins with parsing web pages to extract plain text terms and URLs. These elements are then assessed using advanced techniques like TF-IDF and a custom URL weighting system to determine their relevance in detecting phishing indicators. Furthermore, we employ a search engine lookup to correlate key terms, aiding in the identification of URLs likely to target victims. Additionally, WHOIS lookup is utilized to compare website registration details, ensuring accurate classification of websites as phishing or legitimate.By integrating these methodologies, our system aims to proactively mitigate phishing threats, thereby enhancing user security in the online environment.

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

I. INTRODUCTION

Phishing is a deceptive practice where attackers create fake replicas of legitimate websites to trick users into disclosing sensitive personal information [1]. It is often referred to as brand spoofing and differs from hacking. Phishing attempts involve redirecting users to fraudulent webpages designed to capture usernames, passwords, and credit card details.

The prevalence of phishing attacks is driven by their potential for substantial financial gain with minimal effort. Victims are manipulated into entering their private information, which is then exploited by the phisher to gain unauthorized access to their accounts.Various techniques are employed to deceive users, including replicating content from official websites, using email addresses that mimic legitimate ones, and deploying deceptive advertisements that redirect users to phishing pages.

Spear phishing stands out as a highly targeted and dangerous phishing technique, where attackers exploit personalized information such as your name and email address. Despite its name, spear phishing is not a game—it's a deceptive tactic orchestrated by cybercriminals who aim to steal sensitive information like credit card numbers, bank account details, and passwords stored on your computer.To shield users from falling prey to these sophisticated scams, a proactive system approach is proposed for detecting phishing webpages. This method involves extracting plain text and URLs from webpages, identifying domain names, and applying TF-IDF (Term Frequency-Inverse Document Frequency) weight calculations combined with search engine queries.During plain text extraction, the system discerns meaningful relationships between words within webpage content. URLs are extracted by

ISSN PRINT 2319 1775 Online 2320 7876
Research Paper © 2012 IJFANS. All Rights Reserved UGC CARE Listed (Group -I) Journal Volume 11, S.Iss 6, 2022

analyzing metadata such as titles, descriptions, and keywords. TF-IDF evaluates the importance of words based on their frequency and relevance within the webpage, utilizing a prioritized list of high-frequency terms. These terms are then used to query search engines for the top 30 related URLs, which are subsequently scrutinized using WHOIS lookup to verify domain ownership and discern potential phishing attempts.This systematic approach aims to fortify user defenses by proactively identifying and mitigating spear phishing threats through advanced analysis of webpage content and URLs.

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.

ISSN PRINT 2319 1775 Online 2320 7876
Research Paper © 2012 IJFANS. All Rights Reserved UGC CARE Listed (Group -I) Journal Volume 11, S.Iss 6, 2022

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 propose a phishing website detection system designed to classify websites as either phishing or legitimate. The use of term weighting for phishing pattern detection aims to minimize false positives and false negatives (see Fig. 3.1). The system's primary objectives are as follows:

- **Term and URL Extraction**: Extract terms and URLs from web pages using a DOM parser.
- **Term Identification**: Utilize TF-IDF and URL weighting schemes to identify key terms, such as brand names.
- **Brand Name Search**: Perform searches for brand names using a search engine API.
- **Victim Website Identification**: Identify the legitimate websites targeted by detected phishing sites.

The next section provides a detailed explanation of the proposed system architecture, demonstrating how it achieves these objectives.

System architecture

3.1.1 Webpage Parsing:

1. Goal:
   - Extract terms and URLs from web pages.
2. Procedure:
   - Utilize an HTML parser to build a Document Object Model (DOM).
   - The DOM is a tree-structured model representing the content of HTML or XML documents.
3. Document Object Model (DOM):
   - The DOM is a W3C standard for handling and manipulating HTML and XML content in memory.
   - It offers a set of objects and methods to simplify programming tasks.
   - Although initially created for HTML, the DOM now supports XML as well.
   - The DOM allows the complete reconstruction of a web page from its tree structure.
4. Anchor Tag Parsing:

ISSN PRINT 2319 1775 Online 2320 7876
Research Paper © 2012 IJFANS. All Rights Reserved UGC CARE Listed (Group -I) Journal Volume 11, S.Iss 6, 2022

- Extract URLs by parsing anchor tags.
- The typical syntax of an anchor tag is ....

Domain Name Extraction:

1. Purpose:
- Retrieve the domain information from a web page to verify its legitimacy.
2. Technique:
- Employ regular expressions to accurately extract the domain name.
- The domain is typically found after the hostname and a period (.), followed by a slash (/) and the rest of the URL path.

Shroff's Word Frequency:

When processing XML documents, the Document Object Model (DOM) is used to create a tree structure representing the XML data. This allows efficient navigation and manipulation of the document.

Key Points:

1. DOM Parser:
- Parses the XML document to construct a DOM tree, providing a hierarchical data structure.
- Enables operations like searching, adding, moving, or deleting elements within the document.
2. Switching Parsers:
- You can switch to any DOM-compliant parser without modifying your code.
- As long as the new parser adheres to the DOM standard, your existing code will work without issues.
3. Web Page Parsing Phases:
- Parsing: Converts web page content into a DOM tree.
- Manipulation: Utilizes DOM methods to change the structure or content of the document.
- Rendering: Updates and displays the modified content on the web page.

3.1.1.1 Plain Text Extraction:

This process involves extracting the HTML code from web pages as it appears in the source code. Typically, files with markup or metadata are considered plaintext if they are directly human-readable (such as HTML or XML). However, in our approach, we do not treat all text from the web page source as plaintext. Instead, we extract textual content only from specific tags. These tags include:

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

3.1.1.2 URL Extraction URL Extraction:

- Function: Extract URLs from anchor (<a>) tags.
- Type: Focus on extracting absolute URLs.

Keyword Weighting:

- Function: Assign weights to keywords found on the web page.
- Process: Calculate keyword weights to identify significant terms (refer to Fig. 3.1).

Research Paper © 2012 IJFANS. All Rights Reserved

TF-IDF [Term Frequency-Inverse Document
TF-IDF Weight:
TF-IDF Weight:
Purpose: A statistical measure used in data retrieval and text mining to assess the significance of a word within a document relative to a larger corpus.
TF (Term Frequency):

- **Definition:** Indicates how frequently a term appears in a document.
- **Normalization:** To adjust for document length, TF is often normalized:
  TF(t)=Number of times term t appears in a documentTotal number of terms in the docum ent\text{TF}(t) = \frac{\text{Number of times term } t \text{ appears in a document}}{\text{Total number of terms in the document}}TF(t)=Total number of terms in the documentNumber of times term t appear s in a document

IDF (Inverse Document Frequency):

- **Definition:** Evaluates the importance of a term across the entire corpus.
- **Purpose:** Decreases the weight of frequently occurring terms and increases the weight of infrequent terms:
  IDF(t)=log⁡e(Total number of documentsNumber of documents containing term t)\text {IDF}(t) = \log_e \left(\frac{\text{Total number of documents}}{\text{Number of documents containing term } t}\right)IDF(t)=loge (Number of documents containing term tTotal number of documents)

3.1.3 URL-Based Weight:

- Objective: Assign weights to URLs based on their frequency of occurrence in the document's source code (see Fig. 3.1).

Search Engine Lookup:

- Objective: Use Google's search engine API to gather information related to keywords.
- Steps:

1. Select Keywords: Identify three keywords based on Shroff's word frequency analysis.
2. Perform Search: Query Google with these keywords to obtain the top 30 search results.
3. Extract Domains: Extract domain names from the search result URLs for subsequent URL verification.

Ex. URL for getting result from Google search engine API:
https://www.googleapis.com/customsearch/v
1?key=<API Key>&cx=<Google Custom Search Engine Control Panel Setting ID>& q=<Keywords to search>

3.1.4 URL Verification:

- Objective: Authenticate domain ownership using the WHOIS API for domains found in web pages and search engine results (see Fig. 3.1).

WHOIS Lookup:

- Description: Provides information about domain ownership for domains registered with organizations like VeriSign or Network Solutions .
- Steps:

ISSN PRINT 2319 1775 Online 2320 7876
Research Paper © 2012 IJFANS. All Rights Reserved UGC CARE Listed (Group -I) Journal Volume 11, S.Iss 6, 2022

1. Input Domains: Enter domain names extracted from both web pages and search engine results.
2. Evaluate Legitimacy: Compare the legitimate domain (Dl) with the queried domain (Dq) to assess the risk of phishing.

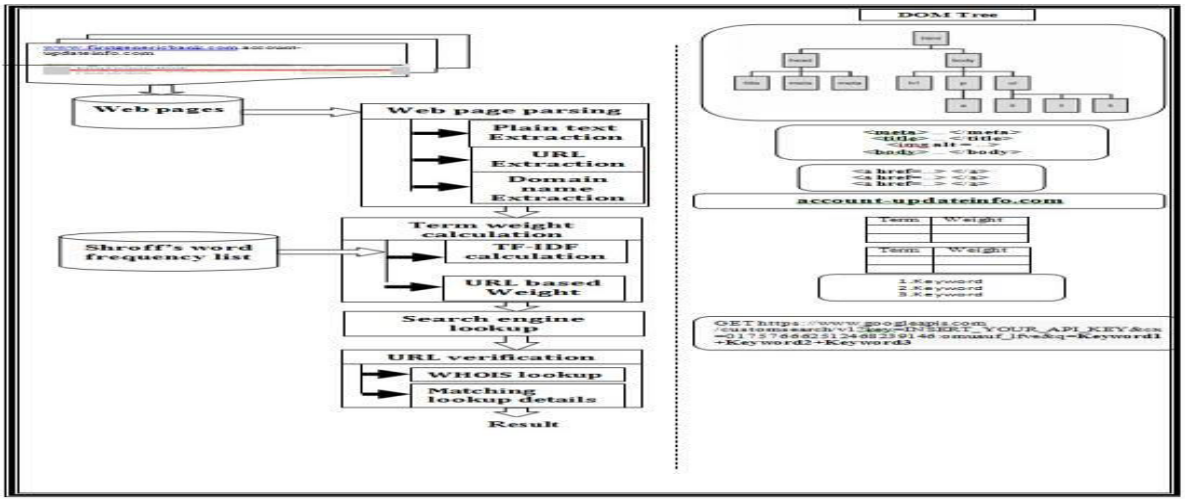

Fig 3. Phishing Webpage URL Detection

Algorithm:

1. Select Web Page:
   - Choose a web page randomly from a dataset containing both legitimate and phishing pages.
2. Extract Content:
   - Retrieve textual content from the following tags: meta, anchor, title, and body, as well as from attributes such as alt.
3. Calculate Weights:
   - TF/IDF: Compute term weights using the updated formula based on Shroff's word frequency.
   - Adjust for URLs: Modify term weights based on their occurrence in URLs. Identify the top three most weighted words.
4. Submit Search Query:
   - Use the top three words to query a search engine API and gather the top 30 results.
5. Analyze Domains:
   - Identify the most frequently occurring domain among the search results.
6. Compare Domains:
   - Use the WHOIS API to compare the domain information from the web page with the domain from the search results.
7. Classify Page:
   - Mark the web page as legitimate if the domain information matches; otherwise, classify it as a phishing page.

V. RESULT

For our experimentation, we use a dataset compiled by Choon Lin Tan et al., which includes both legitimate and phishing web pages. Phishing pages are sourced from Phish Tank, a

Research Paper © 2012 IJFANS. All Rights Reserved

repository that tracks phishing sites and provides justification for their classification. Legitimate pages are collected manually. The accuracy of the results is assessed as follows:

- **IS Result:** The outcome provided by the implemented system for each processed web page.
- **PT Result:** The phishing status of the web page as recorded by Phish Tank.

	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

2 APPLICATIONS

The proposed system, when used as a standalone application, aids data mining researchers in uncovering patterns related to phishing websites. If developed into a browser plug-in, it could serve end users by automatically analyzing websites they visit, such as banking and social networking sites, to detect and alert them to potential phishing threats.

3 FUTURE SCOPE:

In future developments, the community detection system will emphasize improving phishing website detection by incorporating multiple search engines. This strategy is intended to reduce the potential for biased results that may arise from using only one search engine.

4 CONCLUSION

III. The phishing website detection system boosts accuracy by using both term and URL weights to detect potential phishing threats. Traditional methods like TF-IDF often struggle due to limited textual data and the requirement to analyze a large number of web pages. The proposed system addresses these challenges by integrating URL weights and Shroff's word list, thereby streamlining the detection process and enhancing overall efficiency.

IV. Evaluation Criteria:

- **Detection of Genuine Phishing Sites:** Evaluates how well the system identifies authentic phishing websites.
- **Identification of Actual Victims:** Determines the accuracy of the system in recognizing real victims of phishing attempts.

ACKNOWLEDGEMENT

We are deeply thankful to our internal guide, Prof. Nitin Hambir, for his unwavering support and guidance throughout this project. His thoughtful suggestions and assistance have been invaluable, and we are truly appreciative of his help.

We also wish to acknowledge Prof. S. S. Das, Head of the Computer Engineering Department at DYPSOE, Lohegaon, Pune, and our Project Coordinator, Prof. J. L. Chaudhari, for their essential support, guidance, and motivation during the course of the project. Their contributions have been greatly appreciated.

ISSN PRINT 2319 1775 Online 2320 7876
Research Paper **© 2012 IJFANS. All Rights Reserved**, UGC CARE Listed (Group -I) Journal Volume 11, S.Iss 6, 2022

REFERENCES:

[1] Choon Lin Tan, Kang Leng ***Chiew***, ***Han*** Nah ***Size***, "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/ge ***neral understand phishing***

[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/10.1145/1242572.12426 59*