International Journal of Computer Engineering in Research Trends
Multidisciplinary, Open Access, Peer-Reviewed and fully refereed
Research Paper Volume-8, Issue-12, 2021 Regular Edition E-ISSN: 2349-7084

Detecting Phishing Websites Using Natural Language Processing

Sherif Kamel Hussein[1], Aboubaker Wahaballah[2], Amal Alosaimi[3]

[1] *Associate Professor- Department of Communications and Computer Engineering, October University for Modern Sciences and Arts, Giza, Egypt*
Head of Computer Department – Arab East Colleges – Riyadh- KSA

[2] *Assistant Professor- Arab east college, Riyadh, Kingdom of Saudi Arabia*

[3]*Master of computer science, Arab east college, Riyadh, kingdom of Saudi Arabia*

**Corresponding Author:skhussein@arabeast.edu.sa*

Available online at: http://www.ijcert.org

Received: 24/12/2021, Revised: 31/12/2021, Accepted: 08/01/2022, Published: 14/01/2022

Abstract: - Phishing is one of the most cyber attacking tools. It targets both users and organizations. Several solutions have been proposed for detecting and preventing phishing websites, emails and SMSs. However, more research works are required to improve the phishing detection techniques such as improving the detection scalability and reducing false positive and negative alerts. This paper proposes a website phishing detection system based on natural language processing (NLP) features such as statements, words, and characters frequency. The proposed system first enables any user to find out if a specific website is phishing or not and, second, provides a search engine that 24/7 searches for the phishing websites and informs the system administrator (or publishes alerts online) about that. The system is evaluated in terms of its scalability and accuracy. The system accuracy here relies on the number of false-positive, false negative, true positive, and true negative alerts.

Keywords: Attacks, Natural language processing (NLP), Phishing, Scalability, Website, Information and communication technologies (ICTs).

---

1. Introduction

The increased use of information and communication technologies (ICTs) techniques has made our life better and easy. Nevertheless, it has yielded to new security challenges. In most countries, such security challenges, also called Cybersecurity crimes, can just be computer misuse or even cybercrimes according to cyberspace laws. Unfortunately, there is no simple solution for fighting cybercrimes, instead, several solutions have to be used in different stages and all these solutions must be managed and monitored in a well-defined process. This is because security is a process, not a product. Cybercrimes can be categorized into two types:

- Cybercrimes that target information and communication technologies (ICTs).
- Cybercrimes that use information and communication technologies (ICTs).

The former type targets the ICTs techniques and systems through several types of attacks such as malware (viruses, worms, Trojans, rootkit, spyware, adware, etc.), denial of service (DoS), distributed denial of service (DDoS), phishing, botnet, social engineering, spam emails, and so on.

 https://doi.org/10.22362/ijcert/2021/v8/i12/v8i1204

Again, this kind of attack targets the computing equipment and systems (such as servers and websites) for different purposes, for example, getting unauthorized access, altering data, or shutting down systems and services.

The second type of cybercrimes, which uses ICTs, can target humans and societies such as computer fraud, cyber-terrorism, and cyber extortion. In other words, the second type of crime is traditional or well-known crimes but now they are committed with ICTs (as a crime tool). In any way, they are considered cybercrimes since the crime scene is cyberspace. This research paper falls into the first type of cybercrimes, especially phishing.

To be more specific, this research paper focuses on phishing detection techniques using natural language processing (NLP). In fact, according to Verma *et al.*, [1] the NLP techniques are used in three different phishing detections, which are our website content, email content, and URLs phishing detections. The NLP techniques rely on natural language features, such as words and letters, to detect phishing in all directions. Another kind of phishing detection is based on email or website features such as color, font size, headers, titles, and so on. The goal, in the end, is to find out any matching URL, website, or email that is designed or prepared as a phishing attack.

Therefore, this paper proposed a website phishing detection system based on NLP techniques. The proposed system consists of three modules, which are phishing search engine (PSE), phishing search website (PSW), and phishing search add-on (PSA). The PSE is a 24/7 search engine that finds the phishing website and alerts the admin or online about that phishing. It will be hosted on a server in the cloud so its 24/7 availability is ensured. While the PSW is a webpage that contains a search box. The user can put a URL to check it. The webpage will send this URL to the PSW and after a few time receives the result which will be displayed on the webpage as a report. Finally, the PSA is a component that is implemented as a browser add-on. When the user clicks on it, the system submits the opened URL to the PSW to find out if it is a phishing website or not.

The motivation behind this research is to provide a search engine for detecting website phishing and, as a result, contribute to the cybercrimes fighting efforts by alerting about the current phishing websites. The user can use the proposed system to know if a visited URL is phishing or not before responding to it. Organizations can use this system to mainly find out if abnormal phishing pages target their website pages (especially login pages).

The proposed system is implemented and evaluated. The evaluation focuses on two main criteria namely scalability and accuracy. The scalability is measured based on the number of submitted phishing detection requests. The system accuracy is measured using four main sub-criteria, which are the number of false positives, false negatives, true positive, and true negative.

This paper is structured as follows: Section II provides an overview of the phishing attack. Section III reviews and evaluates the related works followed, in Section IV, by introducing the proposed system architecture. The system implementation is discussed in Section V, followed, by Section VI, by discussing the system evaluation result. Finally, Section VII concludes the study and highlights the most critical future works.

2. Phishing Overview

According to Patil and Devale [2], phishing is "the manner of deception of an organization's customer to communicate with their confidential information in an unacceptable behavior". The goal of phishing is to target sensitive information (such as ID and password) or maybe systems using a fake URL leading to a fake website. The hacker here tries to let the user (victim) open a malicious link or website and, then, enter his sensitive data.

Historically, phishing started in 1996 by targeting American Online (AOL) [3]. The word "Phishing" comes from "fishing". According to Mohammad *et al.* [4], phishing strategies can be classified into three types, which are:

- Mimicking phishing attack: Using the mimicking attack, the attackers pull their victims to disclose their private information by email mimicking an official email from a well-known company.
- Forward phishing attack: The phishers/attackers get unauthorized access to the user's or victims' information by forwarding the user to a fake website. So, it is called a forward attack)
- Pop-up phishing attacks. The phisher uses a pop-up website or frame (mostly automatically opened without the user request).

There are several vectors for phishing and such vectors are used as a medium for launching the phishing attack. Mostly, the vector is used for sending a fake URL. A vector can be considered here as a phishing tool. These vectors are [5]: Emails messages; Instant messages; Short message service (SMS); Websites URLs; Social networks; Internet fax, eFax, or online fax.

Mohammad *et al.* [4] also present the lifecycle of phishing websites. As shown in Figure 1, the phishing lifecycle consists of three steps namely planning, collection,

and fraud. During the planning step, the phisher identifies who is the target (victim such as the user or even the whole organization), what is the target (information e.g., passwords), and how to target (a technique used such as a fake URL). In the collection step, the attacker or phisher targets the victim by leading him to reveal his access credentials, or, at this stage; the phisher can make unauthorized access on behalf of the victim. Finally, in the fraud step, the phisher uses the obtained unauthorized access to gain something such as private data access, financial fraud, and so on. Figure 1 shows the phishing steps.

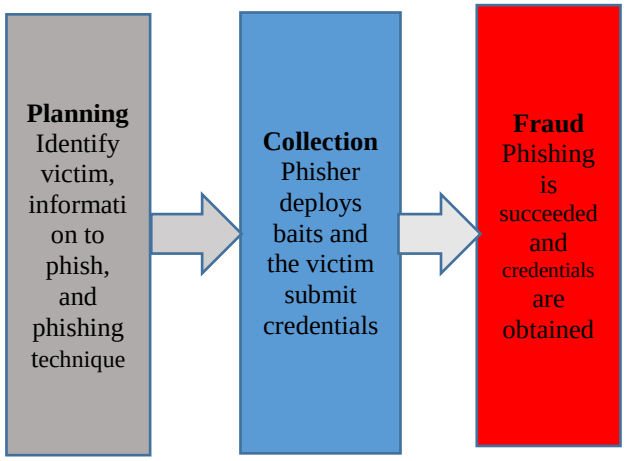

Figure (1): The phishing steps

It is clear that organizations, even private or public, have to fight phishing attacks using different directions. There are three phishing fighting directions, which are [6]:

- User awareness: Establish a strong cybersecurity awareness program that covers phishing attacks. As a result, the organization's staff can discover phishing URLs and websites. By the way, the awareness program must include other cyber security tips that also support fighting phishing, for example, awareness on email, browsing, and social network security.
- Phishing prevention: Using security solutions to provide a security layer that prevents phishing attacks. For example, using two-factor authentication instead of relying on the password only, blocking the known malicious websites, filtering unknowing emails by considering them as spam emails, monitoring the visited website by the organization's staff.
- Phishing detection: Working on detecting any phishing website or even URL using several techniques such as search engines, blacklists, machine learning, natural language processing (NLP), and so on. It is a good recommendation if the organization determine its critical URLs (in its website) and search if someone has fabricated such URLs to phish the organization's staff and the clients as well.

3. Related Works

In the literature, several phishing detection techniques are proposed for fighting the phisher, these techniques are whitelist-based, blacklist-based, content-based, visual-similarity-based, and URL-based techniques. The following sub-sections discuss these techniques in more detail.

A. Whitelist-based Technique

This technique relies on detecting phishing websites by comparing the URLs of both the fake website and a trusted website. Therefore, the solution must maintain a list of all trusted websites' URLs. Kang and D. Lee [7] proposed a solution called *Phishing Guard* which compares the URLs of the accessed website with a large number of well-known trusted websites, which seems impossible as the number of trusted websites is increased over time and in each minute. Another work was presented by Gao *et al.*, [8]. This work presents a solution called *Automated Individual White-List (AIWL)*. The AIWL manages a list of popular login pages. These login pages are automatically added to the list because the AIWL contains a search engine that continues the search for login pages. If the user tries to submit his access credentials to non-trusted URLs, the AIWL checks this URL and alerts the user if it is a phishing web page. It is clear that these solutions provide a fast detection mechanism but with a large number of false-positive alerts. This is because no such solution can define all trusted login pages in the world. The user will receive a large number of notifications whenever he signs in and most of these notifications/alerts are false-positive alerts.

B. Blacklist-based Technique

A black list of well-known phishing websites/pages is defined. The solutions use this approach as a third-party search engine (such as Google or Bing search engines) to find out all phishing websites. There are several proposed research works, based on this technique, propounded by Sharifi *et al.*, [9] and Prakash *et al.*, [10]. Nevertheless, subsequently, these proposed solutions rely on 3rd party search services (such as Google) for finding fake websites and comparing them with one another. As a result, their performance is getting worse while trying to find more phishing websites. Secondly, these solutions cannot detect what is called a *zero-hour phishing* attack. This is because phishing websites are changing over time and in each hour a large number of phishing websites are launched. On the other hand, the attacker uses a phishing website for a very short

time (one day or less) then changes to a new phishing website to target the same organization or another.

C. Content-Based Technique

This technique is based on text available on phishing webpages and emails. The content found in a website or email is analyzed, for instance, by listing all words and then searching over the internet (using a third-party search engine) to find out any website that may contain the same words. The phishing webpage always uses the same words and text used by the trusted website. Ardi and Heidemann [11] proposed a system called *AuntieTuna* which is a web browser plugin. The proposed plugin indexes all words in the visited website and searches for similar websites on the web. The hashing function is used while comparing the phishing website with the trusted website so this method provides a zero false positive but, at the same time, it will not be able to detect most phishing websites.

Che *et al*., [12] proposed a content-based email phishing detection approach based on the semantic web and fuzzy control concepts. The meaning of the phishing email is analyzed and compared with other trusted emails after searching for it in a database. So, this approach focuses on the meaning of phishing emails, not on the exact contents.

Peng *et al*., [13] proposed an email phishing detection method based on the natural language processing (NLP) concept. This method considers an email as a phishing email if it inquires sensitive information such as ID and password, credit cards information, name, address, phone number, and email address. The Peng *et al*., [13] solution has the worst result in terms of accuracy.

Another use of the natural language processing (NLP) concept is by Egozi and Verma [14]. This method extracts the language features from the email message such as the number of words and so on. Then, these features are compared with other emails to detect if this email is a phishing email or not. Nevertheless, there are no publically available emails to compare with.

D. Visual-Similarity-URL-Based Technique

This technique is based on the visual feature of the website. It is well-known that a phishing website will use the same visual features. The technique was first proposed by Wenyin *et al*., [15] based on the three visual features, which are website overall-style, website block style, website layout. Later on, in 2006, Wenyin *et al*., [16] suggested another approach based on the *Earth Mover Distance (EMD)*, in which the signature of the images in both websites (phishing and trusted websites) are compared together. The EMD technique has a good performance as well as a good accuracy result. It has 89% true positives and only 0.71 false positives.

E. URL-Based Technique

The fast technique is URL-based phishing detection, which compares the URL of the phishing website with the URL of the trusted website. This approach is very fast because the compared data is small (URL only) and there is no need for comparing the page content.

4. The Proposed System Architecture

The proposed system will have three components as shown in Fig 2. These components are:

- Phishing Search Website (PSW): This is a webpage that contains a search box. The user can put a URL to check it. The webpage will send this URL to the PSE and after a few time receives the result which will be displayed on the webpage as a report.
- Phishing Search Add-on (PSA): This component will be implemented as a browser add-on. When the user clicks on it, the system submits the opened URL to the PSW to find out if it is a phishing website or not.
- Phishing Search Engine (PSE): This component will 24/7 search for the phishing website and alert the admin or online about that phishing. It will be hosted on a server in the cloud so its 24/7 availability will be ensured.

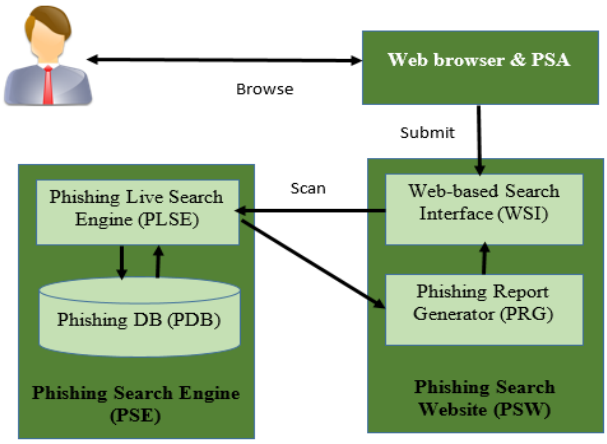

Figure (2): The general architecture of the proposed system

Interactions between the proposed system components are presented in Fig 2 as under:

- The user submits his request to the phishing search website (PSW). The request is just a webpage URL that may be a phishing website. The request can be submitted by visiting the PSW website and inserting the URL in the search textbox or by clicking on the PSA installed on the user web browser as an add-on.
- The PSW forwards the request to the phishing search engine (PSE).
- The PSE starts by finding this URL if it is already known as a phishing website by searching inside the PDB. If not, it starts searching on the web using Google search engine to find out any similar webpage with a minute difference on the URL of the page.
- The PSE returns the result to the phishing report generator (PRG) which will give the result a scaling to indicate the matching level between the submitted URL and the trusted website.
- The PSW shows the report to the user.

The following sub-section will discuss each component in more detail.

A. Phishing Search Add-on (PSA)

The PSA is developed as a browser add-on and by using it the user can send any URL to the phishing search website (PSW). It is just a direct method for submitting suspicious and opened URLs. In other words, if the user wants to check if the currently accessed webpage is phishing or not, he just clicks on the PSA (which is shown as a button on the web browser toolbar) and the PSA will submit the URL to the PSA.

Therefore, the PSA will also not show any result regarding the phishing detection scanning. The browser will automatically open a new browser tab that will present the result of the scanning given by the PSW.

B. Phishing Search Website (PSW)

The PSW helps as the phishing searching interface that receives the scan requests and shows the scan result. The scan requests come to PSW by the user using two main methods, which are:

- The user visits the PSW and enters the suspected URL.
- The user submits the suspected URL through the PSA installed in his web browser.

However, the PSW will then show a report about the scan result. If the scanned URL is a phishing webpage, the result will show all indicators that lead to considering this webpage or URL as a phishing webpage or URL.

C. Phishing Search Engine (PSE)

The PSE is the main search engine that decides if the submitted URL is phishing or not. It detects the phishing webpage by comparing the texts found inside the scanned webpage with other web pages on the web after searching each statement and word on the web using the Google search API.

However, before using the text search technique, the PSE check also the following conditions on the scanned URL:

- If the URL is new, means created within the last two days. This is because the phishing webpages are deleted within some days and the old webpages are considered trusted.
- If the URL is not archived in google search. If archived the URL is trusted.
- If the URL did not start with https and/or has a secure SSL certificate. If not the webpage is considered phishing.
- If the URL is too long or not. The phishing URLs are almost long.

If all the above conditions are true, the PSE starts using the webpage's text by searching each statement on the web using the Google search APIs. The search result will be sent after to the PSW, which will show the scan result report.

5. System Implementation

For implementing the phishing search engine (PSE) module, the Java programming language is chosen and the NetBeans, which is a java integrated development environment (IDE), is used. Using the Java programming language helps in making the implemented PSE portable, which means it can be executed in any operating system such as Windows, Linux, MAC, etc. However, the proposed system is hosted in Linux Ubuntu operating system.

Google search APIs are application programming interfaces (APIs) provided by Google and they can be used by developers to search over the web. These APIs are already integrated with other Google services such as Search, Translate, Gmail, Google Maps, etc. Many websites and applications are now relying on Google Search APIs to find out some content on the web.

The Google search API is accessed by the PSE since the PSE needs to search over the web to find out if the scanned webpage or URL is phishing or not. This API is used for this purpose.

For making and managing tables by the PSE to list all found phishing URLs in a black list, MySQL is used. MySQL is well-known as the most popular open-source relational SQL database management system. It is widely used by developers in developing several Java and Python applications as well as web-based applications.

The Java Network API is used by the PSE module for finding several helpful pieces of information that help in deciding if the scanned URL or webpage is phishing or not. First, this information needs checking of the domain name system (DNS) used by the scanned URL or webpage is registered in Google search or not. Secondly, the finding of the DNS is new or not. Mostly the phishing webpages' DNS is new and its age is just a few days. Third, check if the URL is abnormal (long words, contains similar words with the secure URL, etc.). Finally, if the scanned URL starts with https or http.

The java REST API is used to let the modules (PSE, PSW, and PSA) communicate with each other. This communication is highly important to exchange important information between the suggested modules.

JavaScript is a programming language used for developing web pages efficiently. It is used by this research for developing the PSA. Only the Chrome browser will be supported during the implementation in PSA due to the limited time.

The Hypertext Markup Language (HTML) is used for developing the phishing search website (PSW) in which the user can enter a URL to scan if this URL is phishing or not. The PSW is also responsible for showing the scan result.

The Cascading Style Sheets (CCS) is used in this research project for describing how the PSW looks like. It helps in describing the colors, size, locations, images, etc. during developing the PSW using the HTML as discussed in the previous sub-section.

6. Result and Discussion

The proposed system is evaluated using these two criteria namely efficiency and security scan result. The evaluation of the above main criteria is presented in the next two sub-sections.

A. System Efficiency

System efficiency is evaluated by measuring the cost, in terms of time, of each step discussed in section 4.3. Each step is evaluated three times with three different websites, then the average result is taken. Fig 3 shows the evaluation result of the system efficiency.

Based on the result shown in Fig 3, it is clear that the proposed system is efficient as it can detect any phishing website in less than 1.5 seconds. Whereas, most phishing websites can be detected in the first five steps. This means without searching on Google to find any similar websites. Therefore, in most cases, the cost is 0.5 seconds only after reducing the cost of Google search.

In addition, the proposed system did not have any bottleneck; means no step can take an unreasonable time. As a result, the proposed system by design can scale well with the increased number of concurrent requests. The only thing required is applying more resources (or servers) for hosting the phishing search engine (PSE).

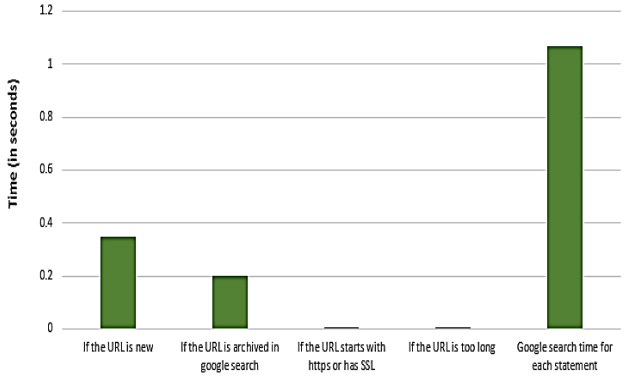

Figure (3): The result of the efficiency evaluation

B. Security Scan Result

The security of the proposed system is evaluated by measuring the number of false-positive, false negative, true positive, and true negative alerts. Table 1 shows the result of the security scan when scanning ten phishing web pages.

Table (1): The security scan result

No.	Sub-criteria	Result
1	Number false positive	0
2	Number false negative	0
3	Number true positive	0
4	Number true negative	10

The above result is when scanning 10 phishing webpages only and this is due to the limited available number of real phishing websites at the time of the system testing. All the web pages are discovered using the first four steps (discussed in Section 4.3) and before using the Google search engine. However, testing the proposed system with more phishing web pages may lead to some false positive results and this can be discovered while using the system with more websites over time.

7. Conclusion

This paper proposed a phishing detection system based on natural language processing (NLP) to detect phishing websites. The proposed system enables the user to scan a URL after submitting it directly to an 'add-on' installed in his web browser or via submitting the URL to a phishing search interface.

The proposed system is evaluated by measuring its efficiency and security scan result. As a result, the system performs well at an acceptable time. The security scanning result shows that the number of true positive is also ten. This means the number of false positive, false negative, and true negative alerts are zero.

More research work is required in the future to improve the proposed system e.g. evaluating the system with a large number of phishing websites over time, improving the detection of phishing websites using the media items (e.g., photos, videos, etc.) found on the webpages, and also improving the phishing black list to list more phishing websites resulting in improving the efficiency of the proposed project, and finally implementing the PSA to support others web browsers as it supports now only Google Chrome web browser.

References

[1] Verma, R., Shashidhar, N., & Hossain, N., "Detecting Phishing Emails the Natural Language Way",Computer Security–ESORICS 2012, 824-841.

[2] Patil, P.; Devale, P. "*A literature survey of phishing attack technique*", Int. J. Adv. Res. Comput. Commun. Eng. 2016, 5, 198–200. 17.

[3] Rakesh M. Verma and Nabil Hossain. "Semantic feature selection for text with application to phishing email detection", InProc. *16th International Conference on Information Security and Cryptology ICISC*, Revised Selected Papers, pages 455–468. Springer, 2013.

[4] R. M. Mohammad, F. Thabtah, L. McCluskey, "Tutorial and critical analysis of phishing websites methods," Computer Science Review, vol. 17, pp. 1-24, 2015.

[5] Kang-Leng Chiew, Kelvin S. C. Yong, Choon Lin Tan:" A survey of phishing attacks: Their types, vectors and technical approaches", Expert Syst. Appl,106: 1-20

[6] Rakesh Verma, Narasimha Shashidhar, and Nabil Hossain, "Detecting phishing emails the natural language way". European Symposium on Research in Computer Security, pages 824–841. Springer, 2012.

[7] J. Kang and D. Lee, "Advanced white list approach for preventing access to phishing sites,", Proc. International Conference on Convergence Information Technology (ICCIT 2007), pp.491-496, 2007.

[8] Y. Cao, W. Han, and Y. Le, — "Anti-phishing based on automated individual white-list", Proceedings of the 4th ACM workshop on Digital identity management. New York, NY, USA: ACM, 2008, pp. 51–60.

[9] M. Sharifi and S. H. Siadati, "A phishing sites blacklist generator," IEEE/ACS International Conference on Computer Systems and Applications, pp. 840-843, 2008.

[10] P. Prakash, M. Kumar, R. R. Kompella, and M. Gupta, "Phishnet: predictive blacklisting to detect phishing attacks,", Proc. IEEE INFOCOM, 2010, pp.1-5, 2010.

[11] Ardi C, Heidemann J, Auntietuna: "personalized content-based phishing detection", NDSS usable security workshop (USEC). https://doi.org/10.14722/usec.2016.23012

[12] Hongming Che, Qinyun Liu, Lin Zou, Hongji Yang, Dongdai Zhou, Feng Yu, "A Content-Based Phishing Email Detection Method", QRS Companion 2017: 415-422

[13] Peng, T., Harris, I. and Sawa, Y., "Detecting phishing attacks using natural language processing and machine learning", IEEE 12th International Conference on Semantic Computing (ICSC) (pp. 300-301), 2018.

[14] Egozi, G. and Verma, R., "Phishing Email Detection Using Robust NLP Techniques", *IEEE International Conference on Data Mining Workshops (ICDMW)* (pp. 7-12), November 2018.

[15] L. Wenyin, G. Huang, L. Xiao Yue, Z. Min, X. Deng, "Detection of phishing webpages based on visual similarity,", Special interest tracks and posters of the 14th

International Conference on World Wide Web, pp. 1060-1061, 2005.

[16] Y. Fu, L. Wenyin and X. Deng, "Detecting phishing web pages with visual similarity assessment based on earth mover's distance (EMD)," IEEE Transactions on Dependable and Secure Computing, vol. 3, no. 4, pp. 301-311, 2006.

Author Profile

[1] Sherif Kamel Hussein Hassan Ratib: Graduated from the faculty of engineering in 1989 Communications and Electronics Department, Helwan University. He received his Diploma, MSc, and Doctorate in Computer Science-2007, Major Information Technology and Networking. He has been working in many private and governmental universities inside and outside Egypt for almost 15 years. He shared in the development of many industrial courses. His research interest is GSM Based Control and Macro mobility based Mobile IP. Now he is working as an Associate Professor in Communications and Computer Engineering department at October University for Modern Sciences and Arts – Egypt. He is also visiting professor at Arab East College for graduate studies – KSA.

[2] Abubaker Wahaballa : Researcher at University of Electronic Science and Technology of China.He holds a Doctoral degree in Cyber Security. He is also a visiting professor at Arab East College, Saudi Arabia. His research interests include information technology communication, IT security, cryptography, and Steganography.

[3] Amal Alosaimi : Graduated from the faculty of computer science –Alemam University –KSA .She got master degree in computer science – information security track 2020 from Arab East Colleges – Riyadh – KSA . Her research interest – Information Security applications based Artificial Intelligence.

