ISSN 2347 - 3983
Volume 8. No. 4, April 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter68842020.pdf
https://doi.org/10.30534/ijeter/2020/68842020

Enhanced detection of Phishing through pattern based recognition for securing human interaction through WEB pages

Mudunuri Bindu Naga Bhargavi[1], Chamarthi Venkata Pavan Kalyan[2], Prakash A[3], Dr, JKR Sastry[4]

[1]Koneru Lakshmaiah Education Foundation University, Vaddeswaram, bindumnb98@gmail.com
[2]Koneru Lakshmaiah Education Foundation University, Vaddeswaram ,chamarthi.vpk@gmail.com
[3]CMR Institute of Technology, Kandlakoya, Medchal, Hyderabad, prakash15aug@gmail.com
[4]Koneru Lakshmaiah Education Foundation University, Vaddeswaram,drsastry@kluniversity.in

ABSTRACT

Phasing the WEB sites is the most dangerous threat posed in proper functioning of the WEB sites especially those that are concerned with e-commerce sites and those sites that deal confidential information of the users such as details related to banking. Many attackers mimic the web pages of the application that collect confedentional information of the users. It has been a challenge to recognize the web pages instated by the attackers and then take corrective actions so that the users are exploited. Many have attempted to detect the Phasing pages based on visual similarity and the attackers mimic the WEB pages such that it quite complicated to detect the Phishing pages.

In this paper a method and an approach presented that is centered on conditional information collected through the WEB pages without much bothering about the Visual presentation of the content or the layout. A repository of the word phrases which are type identified are maintained that are related to confidential information. The confidential data related to a set of Phished pages obtained from Phishedpages.com are generated and stored in a database. Every time A web page is to be processed, confidential data refereed in the WEB page is extracted in consideration with the database of repository of confidential data phrases and the similarly of the same is detected based on Euclidian distance. The process of detection is done on the WEB server side with the client hinted with the possibility of Phishing.

Key words: Visual similarity of the WEB pages, WEB page Phishing, Confidential Information, Euclidian distance

1. INTRODUCTION

Phishing is increasingly being employed for getting hold of confidential information especially related to financialtransactions. Attackers Mimics the world famous web pages which acquire confidential information related the users for effecting especially the financial transactions. The attacking is done through luring the users to provide their confidential information through WEB pages which look alike. The attackers through emails, SMS messages and WEB popups to click the links to the WEB pages that collects confidential information that include credit card numbers, OTP numbers, Passwords, CVC numbers, Adhar card numbers, PAN numbers, Social security numbers etc..

Phishing attacks are evolving every day which can be generally categorised as Targeted phishing and Synchronised Phishing. WEB pages are customised suiting to a specific group through inclusion of photos, images, personal information to trick the users when it comes to targeted phishing. In the case of Synchronised attacking the user’s confidential information is collected and then used to exploit the users especially the financial transactions the attackers get the authentication for the specific system through use of confidential information that they gained out of attacking.

An attacker createsWEB pages similar to well-known web pages and collects important information related to the Victim users. Navigation to these mimic web pages is achieved through sending emails containing the web links to the mimic pages that will be clicked by the users. The WEB links to the Mimic web pages are also posted on the social sites that can be clicked by the users. Phishing attacks using the links to mimic web pages are heavily on the raise.

In literature many ways have been presented to detect the Phishing attacks which majorly include Analysis of URL[1], or determining the similarity in Visual appearance of the WEB pages[2].

In the Case of URL Analysis, the parameters contained in the URLS are analysed to find the similarity. Attackers to avoid detection, modifies the URL parameters to contain more unwanted parameters, thereby the similarly based on the contents of the URLS cannot be judged. Several recommendations [[3] [4], [5]] have been made to include parameters that include life time, time stamps; lock sign of SSL encrypted URLs. The attackers still recognise these features and remove the parameters from the URL as these parameters do not really affect the way the content is presented so that similarity of URL cannot be found.

In the case of visual similarity, the attackers maintain attacking WEB pages which are visually same as the original web pages. Search engines are used for determining the Similarity of the contents included into the WEB pages[[6],[7] based on snippets and frequency of those snippets. These kinds of similarity finding approaches have been defeated by the attackers who add some content which is invisible into the WEB pages fraudulently so that the content of attacking WEB page and the original web page differs.

Some other approaches [8] [9] [10]suggested considerationof Images of the WEB pages to find the similarity. These mechanisms have been found to be complicated as is it involves too much of translation between the Images and html code. It has become more difficult to render the page. Even the differences that exists among the browsers did not much help is rendering the images

The appearance of a WEB page as such is dependent on the content and the layout used for rendering the content[11]. CSS (cascading style sheet) deals with the both the issues of Content format and the Layout used for displaying the content. While some CSS style rules are significant, others are insignificant in rendering the content. The attackers can make changes to the insignificant aspects of the style sheets as they do not affect the visual display thereby detection of similarity of the WEB pages is not possible.

An assessment of the Impact of the elements contained in CSS over the WEB page is first analysed, and the elements and the related CSS rules are shortlisted and the same are used to find the Similarity of the WEB pages based on visual similarity [12]. The attackers can pre-process a WEB page and concert the same to a WEB page that does not include the CSS bot still produce the same Visualisation of the WEB page.

Sometimes the Users do not bother about the visualisation similarity as the most of the reputed and most wanted WEB sites are changed most frequently making it difficult to develop WEB pages which are similar in visualisation. Similarity in the confidential information used is more important than the similarity in the Visualisation of the WEB site.

In this paper a method is presented that focus on parsing the vital data elements that are related to confidential information and maintain the same in a database and then check whether any WEB page is Phishing based on the similarity that is computed based on the Euclidian distance. The WEB pages available at phishtank.com are processed to find the confidential data and then a database is created using the confidential data. Every time a new WEB page is encountered the confidential information is extracted and the Euclidian distance is evaluated based on which the similarity of the WEB page is determined. The Apache WEB server is extended to call the Program that determines the Similarity finding the program and based on the result a specific type of WEB page is displayed through a browser.

2. PROBLEM DEFINITION

The main problem is to find the WEB page similarly based on the confidential information and not on the visual presentation of the content as the users these days ignore the Visual presentation of the WEB pages due to the fact that the vendors themselves keep changing the WEB pages time and again.

3. RELATED WORK

Many methods for detection of Phishing of the WEB pages proposed in the literature that include Black white List based Detection of Phishing, URL based detection, and content based detection, and based on many methods that use machine learning and Artificial Intelligence. Content based detection is more focused on visual similarly of the WEB pages

WEB page are developed using some features that include text, style, images, videos a and Layouts A phishing detection scheme is proposed by Eric et al. [13]that considers the comparison of the features of the Phished WEB pages with the features of a WEB page under investigation.

An algorithmic complexity theory is used by Chen et al. [14] for determining the similarity of the WEB Pages which has been found to be complex for implementation.

CANTINA [15] has used different types of heuristics combined with a new heuristics calledInverse document frequency to determine the similarity of the WEB pages based on the content.

Many have proposed use of search engines for finding the similarity among web pages. The techniques proposed by them failed as the users can change the contents of the web pages so that similarity among the Web pages cannot be found, thereby achieving their objective of luring the users to disclose their confidential data.

Some of presentations in finding the similarity among the WEB pages include comparing the images of the WEB pages. But these approaches have been proved to be impracticable for implementation. Optical character recognition (OCR) is employed to convert the images into text and then text is used for page ranking using Google Page ranking algorithm. The domains of the top ranked sites are fetched using search engines and then the domain of the top ranked WEB sites are compared with the suspected WEB page. The approach is called Goldfish [16].

Document object tree is constructed out of content of WEB pages form which textual clues are extracted and the abnormalities existing in the textual presentations are identified which is used to detect the suspicious web pages. The Suspicious WEB pages normally exhibit too many abnormalities within the content of the WEB pages[17].

A given web page is divided into a set of spatial rectangular blocks and finding similarity between the special blocks of the suspected web page and the related actual web page through special characteristics algorithms have been proposed by Zhang et al. [18]. They have used R-tree indexing algorithm for querying the existence of similar web pages within library that stores the special features of the WEB pages.

The Web pages are stored in files. The similarity of the WEB pages can be determined through matching the files in which the content of the WEB pages are stored.The WEB pages that match the pre-stored Fished WEB pages can be filtered through file Matching as presented by Wardman et al. [19].

H. Zang et al., [20] proposed a A novel framework based on Bayesian approach for detecting Phishing based on WEB page content using which the Web pages that tries to Mimic the original Web page can be determined. They have considered 2classifiers that include an Image classifier, and an algorithm to fuse both the classifiers. They have used Bayesian Model for estimating matching threshold. They have used the model to determine the WEB pages that fall in the category of Fished WEB pages.

A. Y Liu et al., [22] proposed a method for detecting Phished WEB pages through computation of Earth Mover distance through which the similarity of the WEB pages can be measured. They proposed to convert the WEB pages to low resolution images and the signature of the images is represented through color and coordinates features. They have used EMD (Earth Movers Distance) for calculating the distance between the Signatures of the Images
M. Hara et al., [22] have presented that storing the details of the Phished Web pages in a database and then finding the whether a web page is suspected based on the comparison of phished WEB pages and the suspected web pages would be erroneous especially when similar WEB sites exists. They have proposed a method that uses CSS and Images to find similarity between WEB pages.

W. Liu et al., [23] have proposed a System called Site Watcher that runs on an email server and monitors keywords and URLs. The system finds whether a WEB page being navigated is suspicious by comparing the suspected WEB page with the WEB pages that are already phished and stored in database.

Rosiello A. P. E. et al., [24] have presented a system called “AntiPhish” that prevents the user furnishing sensitive information through suspected WEB pages. This system however fails if the user is non cooperative. They have presented at another improved method called DOMAntiPhish that considers the layout similarity of different web pages to determine the suspected WEB page.

S. Afroz et al., [25] have presented an approach called PhishZoo that considers the profiles of the Trusted WEB sites appearance. If any of the WEB sites does have a profile like the profiles of trusted web sites, then the WEB site is considered as suspected.

Y. Zhou et al., [26] have proposed a method of detecting phishing based on visual similarly of the WEB pages both on local and global visual features of the WEB page Images. They have combined the Local and Global Image features to determine the similarity of the suspected web pages with the local and global features of the images of the WEB pages.

Ankhith Kumar et al., [27] have provided comprehensive analysis of the approaches used for Phishing the WEB pages, the way the Phished WEB pages exploits the confidential information of the users. They have presentedcomparison of methods that consider the Visual similarities of the WEB pages.

Many contributions have been made in computing the quality of the WEB sites [28][29][30][31][32][33][34][35][36][37][38][39][40][41][42][43] but the quality of a WEB site form the perspective of the inability to Phish the web pages needs to be further investigated.

4. INVESTIGATIONS AND FINDINGS

Visual similarity of a Web site is dependent on the content and layout used for the development of a WEB page. Some WEB developers use DOM object models to provide the syntax and semantics used for the development of the WEB sites. CSS style sheets are used to define the content and the rules to be employed for display of the content. The kind of data accepted, the format of the data etc are defined within the style sheets. Some web sites also are developed without either using the style sheets or document object model.

The content of a WEB site can be changed without effecting the Visualization of a web page. Attackers can include some hidden content. WEB pages can also be changed by the original developers from time to time making it difficult for the attackers to mimic the WEB pages with the sole aim of deceiving the users and get hold of the confidential information of the user. Users do not pay much attention to the Visual display of the WEB site.

The main concern of the attacker is to get hold of the confidential information. An approach is presented in these papers that do not consider CSS or DOM for finding the similarity of the WEB pages. The method proposed focuses only on the confidential information so as to judge whether an active WEB page is a Phishing WEB page.

The proposed solution is implemented on the server side for detecting whether an active WEB page is suspicious WEB page and then informs the user accordingly. The Plug in engine on the server side is changed so that the program developed for detecting the WEB pages that aims to do phishing is done. The system is done to carry the normal processing if the active WEB page is not suspicious. If the WEB page is found to be suspicious, the user is informed accordingly and no processing is done with the current invoking of the URL. The processing flow is shown in Figure 1.

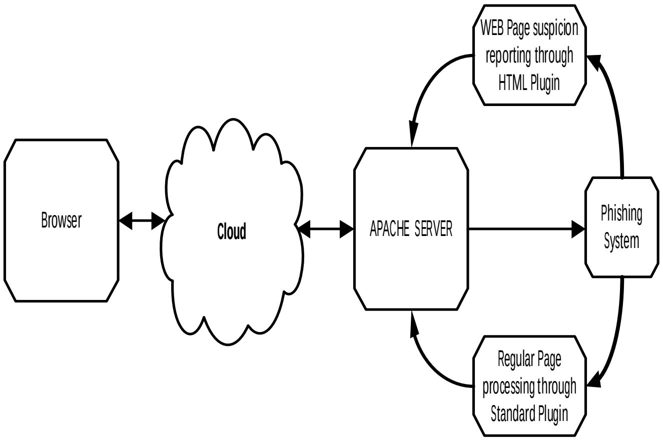
WEB Page suspicion reporting through HTML Plugin
Browser
Cloud
APACHE SERVER
Phishing System
Regular Page processing through Standard Plugin

Figure 1: Phishing detection on the WEB server side

User makes a request for WEB page and Web server hands over the request to the Phasing system. The phishing system checks whether the WEB page is suspicious. If the WEB page is found suspicious, an alert message is sent to the Client through HTML Plugin. If the WEB page is not found suspicious the same is handed over to the regular plugin to carry with further processing.

The overall processing undertaken for detecting the Phishing of a WEB page is shown in Figure 2.

The method described in Figure 2 involves three strategies which include development of a repository of Confidential data through word phrases, conversion of WEB pages provided in Phishtank.com into a databased with reference to repository of confidential data, Checking whether an active WEB page is suspicious which also involves generation of confidential data with reference to the repository of confidential data and then use the data to find the similarity by comparing the data stored in a Database which has the data of about 7000 web pages which are hosted at Phishtank.com. Finally, the actual WEB processing is allowed if the current active WEB page is not suspicious or alerts the user if the Active WEB page suspicious

Generation of Confidential data repository

The confidential data repository is constructed as shown in Table 1. Different keywords are used and the connectivity between the Keywords is achieved through Soundex value and character representation of the data is considered while other types of data representations can also be presented. This repository is updated while parsing the WEB pages for checking the suspicious nature of the web page.

Generation of database for Phished WEB pages

Phishtank.com hosted more than 7000 pages that have been attacked through Phishing. These pages have been down loaded and the pages are subjected to a Parser which is written using the following algorithm. Table 2 shows the database layout using which the Phish tank pages stored. Each record in the Phish tank database is stored with primary key as WEB page serial+URL of the WEB page + Field Label

For Each WEB page

```
    Generate a Running WEB page Number
    Record the URL of the WEB page
    Determine data Labels either through Style sheets,
    or DOM or by Scanning the form Fields
    For each of the selected Label
        Compute the Soundex value
        Determine the Confidential data labels that
        closely match the Soundex value of the
        selected Label from the WEB page and the
        Soundex value of the confidential data
        Select the data type of the selected
        Confidential Data Labels
        Write the data Label. Soundex Value and
        data type of the filed joined with WEB
        page number
    Next Selected Label

Next WEB Page
```

Verifying the Active WEB page to find if it is Phished page

Generation of the confidential data related to the active Page and checking the similarity with reference to the confidential data stored in the database related to Phish Tank pages is carried using the following Algorithm

Algorithm

```
    Determine data Labels either through Style sheets,
    or DOM or by Scanning the form Fields
    For each of the selected Label

        Compute the Soundex value
        Determine the Confidential data labels that
        closely match the Soundex value of the
        selected Label from the WEB page and the
        Soundex value of the confidential data
        Select the data type of the selected
        Confidential Data Labels
        Write the data Label. Soundex Value and
        data type of the filed joined with WEB
        page number

    Next Selected Label
    Arrange the labels in the ascending order of the
    Soundex value of the Labels
```

```
For each of the WEB page stored in the Phish Tank Page

    Query the Labels in the ascending order of the Soundex values of the Labels
    Check whether the Number of variables same as the number of variables
    If the numbers of variables are not same, report that the page is non suspicious
    If the number of variables are same

        For each of the Label in the Active page, fetch the corresponding label in the phish tank page and compute the Euclidian distance and some the distance over total distance

    Next page

    If the Euclidian distance is non zero declare the Active page is non suspicious else declare the WEB page as Suspicious

Next WEB page
```

Calculating the Euclidian distance

Let A1, A2, A3, A4 and A5 are the Labels of the Active WEB page
Let P1, P2, P3, P4 and P5 are the Labels of the Phish Tank Pages
Euclidian Distance D=

$$\sqrt{comp(A1-P1)^2 + comp(A2-P2)^2 + comp(A3-P3)^2 + comp(A4-P4)^2 + comp(A5-P5)^2}$$

The value of the Expression Comp {Ai-Pi) is either zero or 1 or -1 depending on the type of variables involved in the Comparison

If D = 0 then the WEB pages are similar and therefore the active WEB page is identified as Suspicious. The WEB page is considered as non-suspicious for all the other values of the D.

The method is applied considering **100** suspicious pages and **100** Non suspicious pages and **7000** Phish Tank pages. All the 100 suspicious pages are identified as suspicious and **2 of the 100** Non Suspicious pages have been identified as semi suspicious pages.

5. CONCLUSION

Phishing is the most dangerous attacking system that is aimed at getting confidential information of the users and uses the same to exploit the financials of the legitimate users. The attacking is done through mimicking most of the important web pages making it necessary to detecting such kind of mimicking and takes corrective actions so that users are protected

Many approaches have been presented in the literature especially using the Visual similarity and most of the approaches have failed due to the reason that the attackers can change the content of the web sites without changing the visual representation or the layout of the WEB pages and also most of the businesses keep changing the content and the layout of their web pages.

Ability to recognize the confidential data referred to in the WEB pages is the key. Confidential data referred to in different web pages having different URLS leads to suspicion of the page. It is rather a challenge to figure out the kind, type and sequence of the Confidential data being refereed in the WEB pages as styles used to refer to the confidential data could differ a lot..

REFERENCES

1 C. Inc. (Aug. 2016). Couldmark Toolbar. [Online]. Available: http://www.cloudmark.com/desktop/ie-toolbar

2 P. Likarish, E. Jung, D. Dunbar, T. E. Hansen, and J. P. Hourcade, ‘‘B-APT: Bayesian anti-phishing Toolbar,’’ in Proc. IEEE Int. Conf. Commun. (ICC), May 2008, pp. 1745–1749. https://doi.org/10.1109/ICC.2008.335

3 I. Fette, N. Sadeh, and A. Tomasic, ‘‘Learning to detect phishing emails,’’ in Proc. Int. World Wide Web Conf. (WWW), May 2007, pp. 649–656. https://doi.org/10.21236/ADA456046

4 iTrustPage. Accessed on Nov. 1, 2013. [Online]. Available: http://www.cs.toronto.edu/ronda/itrustpage/

5 T. Ronda, S. Saroiu, and A. Wolman, ‘‘itrustpage: A user-assisted anti- phishing tool,’’ in Proc. Eurosys, Apr. 2008, pp. 261–272. https://doi.org/10.1145/1357010.1352620

6 A. Nourian, S. Ishtiaq, and M. Maheswaran, ‘‘CASTLE: A social frame- work for collaborative anti-phishing databases,’’in Proc. ACM Trans. Inter- net Technol., 2009, pp. 1–10.

7 Y. Zhang, J. I. Hong, and L. F. Cranor, ‘‘CANTINA: A content-based approach to detecting phishing Web sites,’’ in Proc. Int. World Wide Web Conf. (WWW), May 2007, pp. 639–648.

8 Y. Cao, W. Han, and Y. Le, ‘‘Anti-phishing based on automated individual white-list,’’ in Proc. 4th ACM Workshop Digit. Identity Manage, 2008, pp. 51–60.

9 X. Deng, G. Huang, and A. Y. Fu, ‘‘An anti-phishing strategy based on visual similarity assessment,’’ Internet Compute., vol. 10, no. 2, pp. 58–65, 2006. https://doi.org/10.1109/MIC.2006.23

10 W. Liu and X. Deng, ‘‘Detecting phishing Web pages with visual similarity assessment based on earth mover’s distance,’’ IEEE Trans. Depend. Sec. Comput., vol. 3, no. 4, pp. 301–311, Apr. 2006.

11 J. Mao, P. Li, T. Li, T. Wei, and Z. Liang, ‘‘BaitAlarm: Detecting phishing sites using similarity in fundamental visual features,’’ in Proc. 5th Int. Conf. Intell. Netw. Collaborative Syst. (INCoS), 2013, pp. 790–795.

12 Jian Mao, Wenqian Tian, Pei Li, Tao Wei, Zhenkai Liang, “Phishing-Alarm: Robust and Efficient Phishing Detection via Page Component Similarity”, IEEE Access (Volume: 5), pp. 17020-17030, August 23. 2017

13 E. Medvet, E. Kirda, and C. Kruegel, ‘‘Visual-similarity-based phishing detection,’’ in Proc. SecureComm, Sep. 2008, p. 22. https://doi.org/10.1145/1460877.1460905

14 T.-C. Chen, S. Dick, and J. Miller, ‘‘Detecting visually similar Web pages: Application to phishing detection,’’ ACM Trans. Internet Technol., vol. 10, no. 2, pp. 1–38, May 2010.

15 A. Nourian, S. Ishtiaq, and M. Maheswaran, ‘‘CASTLE: A social frame- work for collaborative anti-phishing databases,’’in Proc. ACM Trans. Inter- net Technol., 2009, pp. 1–10.

16 M. Dunlop, S. Groat, and D. Shelly, ‘‘GoldPhish: Using images for content-based phishing analysis,’’ in Proc. 5th Int. Conf. Internet Monitor. Protection (ICIMP), May 2010, pp. 123–128. https://doi.org/10.1109/ICIMP.2010.24

17 Y. Pan and X. Ding, ‘‘Anomaly based Web phishing page detection,’’ in Proc. 22nd Annu. Comput. Security Appl. Conf., 2006, pp. 381–392.

18 W. Zhang, H. Lu, B. Xu, and H. Yang, ‘‘Web phishing detection based on page spatial layout similarity,’’ *Informatica*, vol. 37, no. 3, pp. 231–244, 2013.

19 B. Wardman, T. Stallings, G. Warner, and A. Skjellum, ‘‘High-performance content-based phishing attack detection,’’ in *Proc. eCrime Res. Summit*, San Diego, CA, USA, Nov. 2011, pp. 1–9. https://doi.org/10.1109/eCrime.2011.6151977

20 H. Zhang, G. Liu, T. W. S. Chow, and W. Liu, “Textual and visual content-based anti-phishing: a Bayesian approach,” IEEE Transactions on Neural Networks, vol. 22, no.10, pp.1532–1546, 2011

21 Y. Fu, W. Liu, 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 https://doi.org/10.1109/TDSC.2006.50

22 M. Hara, A. Yamada, and Y. Miyake, “Visual similarity-based phishing detection without victim site information, ”in Proceedings of the IEEE Symposium on Computational Intelligence in Cyber Security (CICS ’09), pp. 30–36, IEEE, Nashville, USA, April 2009.

23 Liu, X. Deng, G. Huang, and A. Y. Fu, “An anti-phishing strategy based on visual similarity assessment,” IEEE Internet Computing, vol. 10, no. 2, pp. 58–65, 2006 https://doi.org/10.1109/MIC.2006.23

24 P. E. Rosiello, E. Kirda, C. Kruegel, and F. Ferrandi, “A layout-similarity-based approach for detecting phishing pages,” in Proceedings of the 3rd International Conference on Security and Privacy in Communications Networks and the Workshops (SecureComm’07), pp.454–463,September2007

25 S. Afroz and R. Greenstadt, “Phish Zoo: detecting phishing websites by looking at them,” in Proceedings of the 5th Annual IEEE International Conference on Semantic Computing (ICSC’11), pp. 368–375, Palo Alto, California, USA, September 2011

26 Zhou, Y. Zhang, J. Xiao, Y. Wang and W. Lin, "Visual Similarity Based Anti-phishing with the Combination of Local and Global Features," *2014 IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications*, Beijing, 2014, pp. 189-196.

27 Ankit Kumar Jain and B. B. Gupta, Phishing Detection: Analysis of Visual Similarity Based Approaches**,** Hindawi. Security and Communication Networks, Volume 2017, Article ID 5421046, 20 pages https://doi.org/10.1155/2017/5421046

28 Jammalamadaka, S.B. Duvvuri,B.K.K. Jammalamadaka, K.R.S. Priyanka, J.H., Automating WEB interface about user behavior, Advances in Intelligent Systems and Computing, 815, pp. 91-102, https://doi.org/10.1007/978-981-13-1580-0_9

29 Sastry, J.K.R. Chittibomma, C.S. Alla, T.M.R, Enhancing the performance of search engines based heap-based data file and hash-based indexing file, International Journal of Engineeringand Technology(UAE), 7, pp. 372-375, 2018 https://doi.org/10.14419/ijet.v7i2.7.10722

30 Sastry, J.K.R. Sri Harsha Vamsi, M. Srinivas, R. Yeshwanth, G., Optimizing performance of search engines based on user behavior, International Journal of Engineering and Technology (UAE), 7, pp. 359- 362, 2018 https://doi.org/10.14419/ijet.v7i2.7.10715

31 Sastry, J.K.R. Jyothsna Sai Sree, M.Mani Dedeepya, T. Kamesh, D.B.K., On selection of a user interface dynamically for displaying data mined results, ARPN Journal of Engineering and Applied Sciences, 12(11), pp. 3561-3572, 2017,

32 Kamesh D.B.K., SasiBhanu J., SastryJ.K.R, An architectural approach for assessing the quality of web sites, ARPNJournal of Engineering and Applied Sciences, 13(15), pp,4503-451, 2018

33 Sastry, J.K.R. Sreenidhi, N. Sasidhar, K, Quantifying quality ofWEB-site based on usability, International Journal of Engineering and Technology(UAE), 7, pp. 320-322, 2018 https://doi.org/10.14419/ijet.v7i2.7.10606

34 VenkataRaghavarao, Y. Sasidhar, K. Sastry, J.K.R., Chandra Prakash, V. Quantifying quality of WEB sites based on content, International Journal of Engineering and Technology(UAE), 7(2), pp. 138-141, 2018 https://doi.org/10.14419/ijet.v7i2.7.10280

35 Bhanu, J.S. Kamesh, D.B.K. Sastry, J.K.R., Assessing completeness of a WEB site from Quality Perspective, International Journal of Electrical and Computer Engineering, 9(6), pp. 5596-5603,2019 https://doi.org/10.11591/ijece.v9i6.pp5596-5603

36 B. Vishnu Priya, Dr. JKRSastry, Computing Quality of Structureof a Web-Site, International Journal of Advanced Trends in Computer Science and Engineering, Volume 8, No.5, pp. 2142-2148, 2019 https://doi.org/10.30534/ijatcse/2019/44852019

37 B. Vishnu Priya, Dr. JKRSastry, Assessment of Website Quality based on Appearance, International Journal of Emerging Trends in Engineering Research, Volume 7, issue 10, pp. 360 – 375, 2019 https://doi.org/10.30534/ijeter/2019/017102019

38 SastryJKR, Talluri SL, A framework for assessing the quality of the web site, PONTE, 73(6),pp.20-34, 2017 https://doi.org/10.21506/j.ponte.2017.6.2

39 Sujatha, M.M., PrasadaRao, P.V.R.D.Sastry, J.K.R., Metrics for assessing the quality of WEB sites, International Journal of, Innovative Technologyand Exploring Engineering 8(8), pp. 1710-1714, 2019

40 B. Vishnu Priya, Dr. JKRSastry, Framework for Assessing the Quality of Multimedia objects hosted on a WEB site International Journal of Emerging Trends in Engineering Research, Volume 7, Issue 11.

41 Gorantla Sahana*, ST. Mary Manasa, Dr. JKRSastry, Dr. V Chandra Prakash, Evaluating the quality of Navigation designed for a WEB site, International Journal of Engineering & Technology, 7 (2.7) (2018) 1004-1007

42 Naga PrudhviKolla 1 *, Dr. JKRSastry 1, Dr. V Chandra Prakash 1, Siva Krishna Onteru 1, Yeshwanth , Surya Pinninti, Assessing quality of web sites based on multimedia content, International Journal of Engineering & Technology, 7 (2.7) (2018) 1040-1044

43 V. Sai Virajitha , Dr. JKRSastry, Dr. V Chandra Prakash , P. Srija, M.Varun, Structure based assessment of quality of WEB sites, International Journal of Engineering & Technology, 7 (2.7) (2018) 980-983

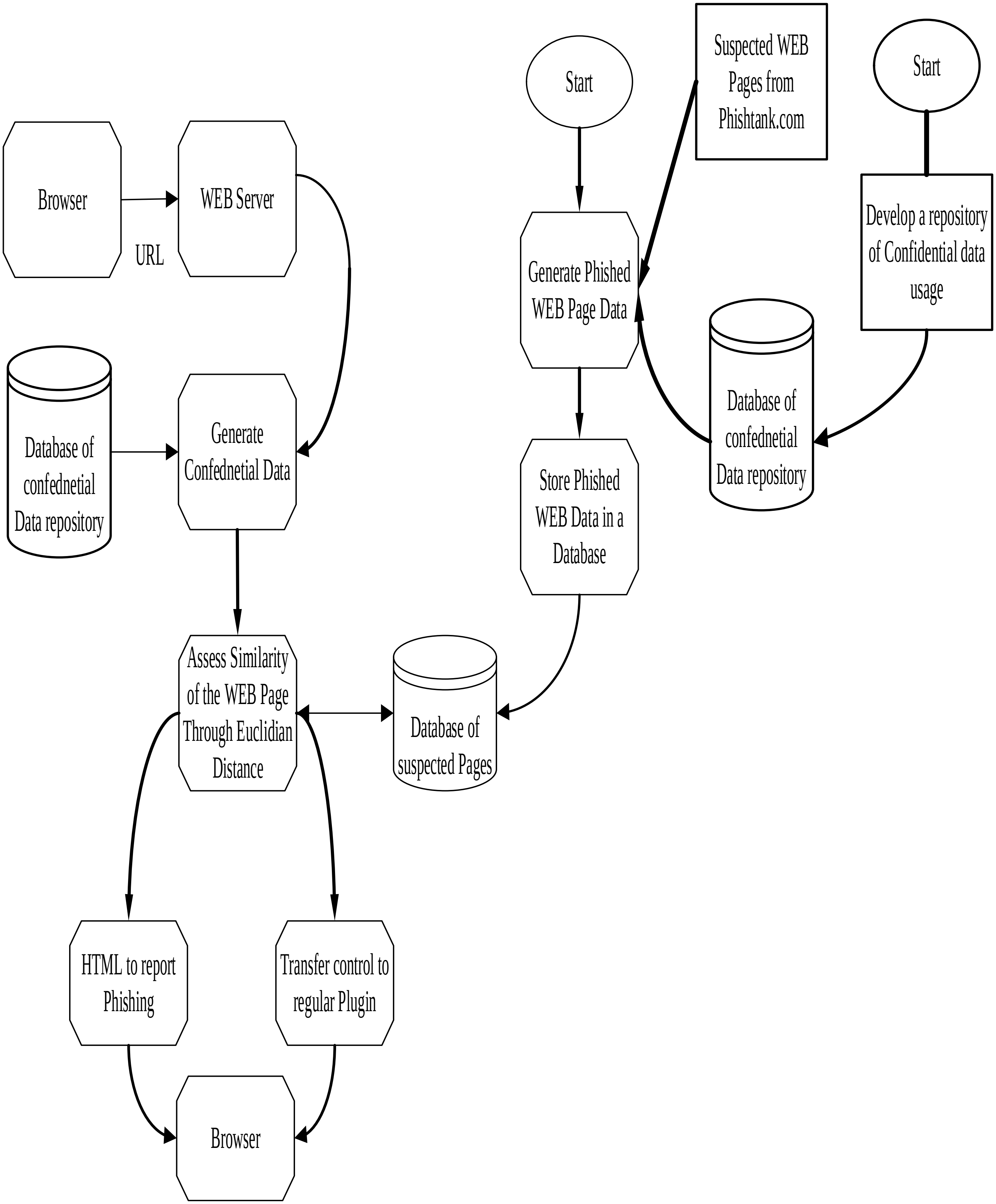
Start
Suspected WEB Pages from Phishtank.com
Start
Browser
URL
WEB Server
Generate Phished WEB Page Data
Develop a repository of Confidential data usage
Database of confednetial Data repository
Generate Confednetial Data
Database of confednetial Data repository
Store Phished WEB Data in a Database
Assess Similarity of the WEB Page Through Euclidian Distance
Database of suspected Pages
HTML to report Phishing
Transfer control to regular Plugin
Browser

Figure 2 : Overall process flow within Phishing detection system

Table 1: Repository of confidential data related word Phrases

Confidential Data Element	Soundex Value	Alternative Contextual usage	Soundex Value	Data Type	Maximum Data Size
Password	121	Passwd	127	Char	20
		Wordpass	789	Char	20
		Password	121	Char	20
		Secret data	534	Char	20
		OTP	847	Char	20
		Pass***	130	Char	20
User ID	987	User name	960	Char	20
		Name of the User	544	Char	20
		ID of the User	322	Char	20
		User ID	967	Char	20
Date of Birth	254	DDMMYY	430	Char	12
		MMDDYY	579	Char	12
		DDMMYYYY	423	Char	12
		MMDDYYYY	590	Char	12
		YYMMDD	980	Char	12
		YYYYMMDD	987	Char	12
		Birth Date	234	Char	12
		YYYYDDMM	999	Char	12
Type of Payment	677	Payment Mode	787	Char	12
		Netbank	678	Char	12
		Credit card	324	Char	12
		Debit card	454	Char	12
		Visa	876	Char	12
		Master	565	Char	12
Card Number	232	Account Number	252	Char	20
		Number	666	Char	20
CVX Number	232	3 Digit data	123	Char	3
		Secret data Behind the card	675	Char	3
Name on the card	767	Card Name	232	Char	20
		Customer name on the card	333	Char	20
Captcha	234	Captcha	343	Char	8
		Cap Cha	343	Char	8
		Robot Identification	675	Char	8

Table 2: Database Layout of the Phish tank pages

WEB page Number	URL of the WEB page	Field LABEL	Field Soundex Value	Field Type	Field Size
1	WWW. XYZ.com/ html/Home.html	Password	121	Char	20
1	WWW. XYZ.com/ html/Home.html	User ID	987	Char	20
2	WWW. XYZ.com/ html/Home.html	Password	121	Char	20
2	WWW. XYZ.com/ html/Home.html	User ID	987	Char	20
2	WWW. XYZ.com/ html/Home.html	Date of Birth	254	Date	12
2	WWW. XYZ.com/ html/Home.html	Type of Payment	677	Char	20
2	WWW. XYZ.com/ html/Home.html	Card Number	232	Char	20
2	WWW. XYZ.com/ html/Home.html	CVX number	232	Char	3
2	WWW. XYZ.com/ html/Home.html	Name on the Card	767	Char	20