MASPHID: A Model to Assist Screen Reader Users for Detecting Phishing Sites Using Aural and Visual Similarity Measures

Gunikhan Sonowal
Department of Computer Science
School of Engineering & Technology
Pondicherry University
gunikhan.sonowal@gmail.com

K. S. Kuppusamy
Department of Computer Science
School of Engineering & Technology
Pondicherry University
kskuppu@gmail.com

ABSTRACT

Phishing is one of the major issues in cyber security. In phishing, attackers steal sensitive information from users by impersonation of legitimate websites. This information captured by phisher is used for variety of scenarios such as buying goods using online transaction illegally or sometime may sell the collected user data to illegal sources. Till date, various detection techniques are proposed by different researchers but still phishing detection remains a challenging problem. While phishing remains to be a threat for all users, persons with visual impairments fall under the soft target category, as they primarily depend on the non-visual web access mode. The persons with visual impairments solely depends on the audio generated by the screen readers to identify and comprehend a web page. This weak-link shall be harnessed by attackers in creating impersonate sites that produces same audio output but are visually different. This paper proposes a model titled "MASPHID" (**M**odel for **A**ssisting **S**creenreader users to **Phi**shing **D**etection) to assist persons with visual impairments in detecting phishing sites which are aurally similar but visually dissimilar. The proposed technique is designed in such a manner that phishing detection shall be carried out without burdening the users with technical details. This model works against zero-day phishing attack and evaluate high accuracy.

CCS Concepts

•Security and privacy → Phishing; •Human-centered computing → Accessibility systems and tools;

Keywords

Phishing, Persons with Visual Impairments, Anti-phishing, Cyber-crime

1. INTRODUCTION

As the number of internet users are gradually increasing, cyber-crime is also upsurging day by day. Phishing is one of the major cyber-crimes where phisher would build a fake website which is similar to a legitimate site and make the users fall into the trap. This leads to various issues such as loss of revenue, loss of reputation for a brand name and also make people to lose their confidence on web based transactions. Anti-phishing work group which is a non profitable entity has reported on May 2016 that there are 123,555 unique phishing incidents are detected [2]. As recorded by WHO, there are 285 million people are estimated to be visually impaired worldwide, among them 39 million are blind and 246 have low vision [19].

As we know that phishing starts with an email because it's an easy and scalable technique to communicate with a group of target users. Phisher sends millions of phishing emails to customers and the email that look and feel like has originated from a legitimate site. So end users easily fall prey in phish. Some anti-phishing tools are able to stop the phishing emails to some extent but still many phishing email crosses the anti-phishing barrier [20] and move into user's mailbox.

In this paper we are not focusing at stopping phishing email, but we are trying to test the target web page link to identify the whether the link is phished on or legitimate. As we mentioned above that we specifically focus on persons with Visual impairments. Persons with visual impairments use screen reader software to identify and navigate web pages. The major problem for persons with visual impairments is that they are not able to see the page, so phishers can target them to fall in trap.

The objective of this paper is to develop an anti-phishing model which would be able to assist persons with visual impairments in detecting phishing site which are aurally similar but visually different. With this point, the proposed method differs from other approaches which assumes that phishing site would be a visual replica of original site. However, if a phishing a group targets, persons with visual impairments, then satisfying this *visual replica* criteria is not mandatory as this group of users depends on the audio generated by the screen readers in identifying and comprehending a webpage.

This paper is structured as follows:

- We briefly explain the about the phishing issues in Section 2. We presents some of solutions which are approached by researchers against phishing sites and on the last paragraph of this section 2 we provide some

ICIA '16, August 25-26, 2016, Pondicherry, India
 ISBN 978-1-4503-4756-3/16/08…$15.00
DOI: http://dx.doi.org/10.1145/2980258.2980443

important attributes of our model.

- This paper basically work on two metrics; one is aural similarity metric and another is visual similarity metric which is described in Section 3. In this section we provide an algorithm that will show the overall description of our model and then illustrate the algorithm with figures.

- To verify the model we have a prototype implementation and the outcome of our model by taking 20 popular banking site's web pages are explained in Section 4.

- In Section 5, discussions about the proposed work are provided and the conclusion with some future directions are provided in Section 6.

2. RELATED WORKS

Phishing detection is an active research area in which studies are being conducted with a spectrum of approaches to handle this major cyber security threat. [15, 9, 5]

Visual similarity of web pages has been considered as an important anti-phishing technique to identify phishing websites in comparison with legitimate sites, by studies. [18] Liu Wenyin, Guanglin Huang, Liu Xiaoyue , Zhang Min, Xiaotie Deng . Three similarity metrics are used to evaluate the visual similarities i.e. 1) Block level similarity, 2) layout similarity, 3) overall style similarity. DOMAntiPhish [14] is a novel approach to detect phishing sites by using layout similarity information. It is used to distinguish websites between phishing sites and legitimate. The DOM tree representation was used to calculate the layout similarity of two websites. If the two websites produce identical layout then they have same DOM-tree.

Yue Wang et. al [17] suggested a light-weight technique that is based on the observation of user's *whitelist* for home users . This Light-weight technique is a whitelist based approach that works on pattern matching method for effective protection. *Cantina* [21] is proposed by Jason Hong, Yue Zhang, Lorrie Cranor and it is content-based technique to detect phishing websites using TF-IDF information retrieval algorithm. CANTINA works as follows:

1. For Each term on given web page the TF-IDF score is computed.

2. The five terms with highest TF-IDF weights are used to generated a lexical signature.

3. Put that lexical signature to a search engine.

4. If it is legitimate websites then the domain name of top 30 search matches current web page domain name otherwise recognize as phishing site.

One problem of this approach is TF-IDF cannot count the hidden terms so that may arise problems in detecting phishing site. Some more features are included along with TF-IDF to overcome issues such as Domain age, Known Images, Suspicious URL, Suspicious Link, IP Address, Dotes in URL and Using Forms.

GoldPhish [7], When a user visit a new web page then an image is captured from the top of the page, optical character recognition(OCR) [12] is used to convert the image to text. After converting to text, these text is submitted to Google. The first four results are enough to identify because due to high PageRank. As it is known that the legitimate site always get displayed at top result in search engine. Sometimes logo is not converted correctly then the other text on the web page is searched in Google.

An Ideal Approach for Detection and Prevention of Phishing Attacks [16] are proposed to detect phishing site by combining both url based and webpage similarity based. The author also use LinkGuard Algorithm [4] to compare both actual url and visual url. Sadia Afroz and Rachel Greenstadt [1] proposed a new technique which depend on the whitelist i.e phishzoo. It is based on fuzzy hash technique i.e. distinguish content element i.e. HTML code, scripts, images.

Automated Individual White-List (AIWL)[3], a novel phishing detection technique where User submitted their credential to a Login User Interfaces (LUIs), AIWL will warn the user against possible attack if it is not Whitelist. AIWL can efficiently prevent against pharming attacks [1] means phishing without a lure i.e. malicious code is installed on a personal computer or server, misdirecting users to fraudulent Web sites without their knowledge or consent.

The authors in their paper(Detecting phishing web sites: A heuristic URL-based approach [13]) provided the structure of URL is as follows: $< protocol >$: // $< subdomain >$. $< primarydomain >$. $< TLD >$ / $< pathdomain >$. eg http://www.ebay.login.abc.net/login/web/index.html ,As per mention structure, there are six features : Protocol is http, Subdomain is ebay.login, Primarydomain is abc, TLD is net, Domain is abc.net, Pathdomain is login/web/index.html The author proposed a new algorithm to detect phishing site through weights computed through genetic algorithm [10]. The model has four phases to detect phishing site:

- Feature Extraction: It has 10 features - Page Rank, Alexa Rank, Age, DNS Records and Abnormal URL, Long URL, Prefix Suffix, Sub Domains, Http/Https and IP Address.

- Pre-processing: Categories into three type - Phishing, Suspicious and Legitimate classes are represented by -1, 0 and 1 respectively.

- Weight Adjustment: Classify the best weight to find website accurately. After that result

PhishBlock [8], it is based on both lookup and a support vector machine (SVM)[6] classifier. In addition, it checks features which is extracted from URL, text and linkage of websites. PhishBlock has three components to detect phishing site 1) Lookup System: In PhishBlock, three local lists is used i.e. Whitelist, Blacklist and doubtful. If the current URL is match in either of them then an alert is displayed to the user like safe, fraud and suspicious. If the URL not belong any of the lists then SVM test that URL. Moreover, PhishBlock also use PhishTank [2] to detect as a fraud or an unknown website, Escrow Fraud is used when it found unknown, otherwise add to the PhishBlock blacklist. In Escrow Fraud, if it is found unknown website then

[1]http://windowsitpro.com/networking/security-update-phishing-and-pharming-june-22-2005

[2]$http://www.phishtank.com/$

pass to google otherwise add to PhishBlock blacklist. 2) Classifier System: This is client-side tools based on features based approach extract from website content or information of domain registration. 3) Fishblock Checks : In this check basically focus on URL Features and content features. The authors proposed a novel technique to detect phishing site on base of user device[11]. They noticed that as mobile device increasing, so popular website are build both device different ways. But phishers don't make website for two devices at the same time. E.g. the desktop website of facebook is www.facebook.com and mobile website is m.facebook.com but both the domains are genuine domain.

MASPID is hybrid-base phishing detection technique and it works using both list based and image based comparison. This model is developed specifically for persons with visual impairments. The following are the important attributes of the proposed MASPHID model:

1. This model incorporates both aural and visual similarity comparison.
2. The User are not required to do any complex things. Simply run the MASPID tool and get the result about the *url* whether it is phished or not.
3. It is a *zero day* detection model
4. The result of the MASPHID model is provided to the user through audio alerts which are better suited for persons with visual impairments.

3. THE MODEL OVERVIEW

The section explores the proposed MASPHID approach for detecting phishing. First, we provide an algorithm then we explain briefly about the algorithm along with figures.

3.1 MASPHID Algorithm

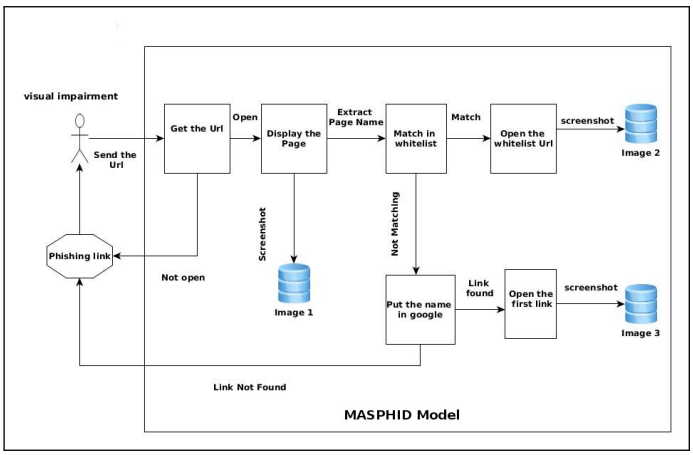

Figure 1: MASPHID MODEL

MASPHID receives the *url* as input, then it opens the page of the url. If the page is failed to open then it is informed to the user and the algorithm is terminated. and if it is open then take a screenshot of the page and save the image for future comparison.

Next, MASPHID prompts the user to enter the descriptive text for url, in text field. After that the MASPHID checks

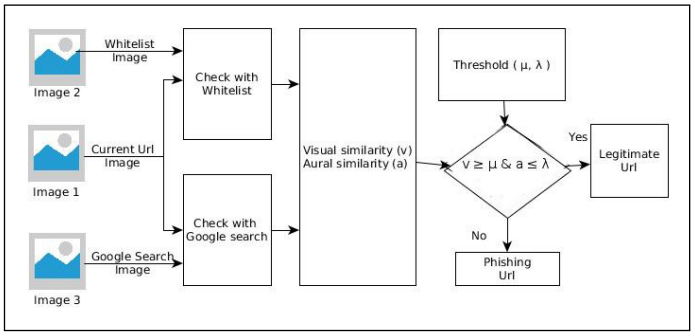

Figure 2: MASPHID MODEL

the url in whitelist. If it is matched with whitelist then extract the genuine url. After that take a screenshot of the genuine page. If the name of page is not match with whitelist, The MASPID put the name in search engine and open the first link and take a screen shot. If the search engine is not returning the relevant page, then it is confirmed that it is phishing site and MASPHID send the information to user.

MASPHID Compare both the screenshot images using root means square [3] and evaluate the similarity percentage and fixed one threshold value to check the page whether it is phishing or legitimate. If visual similarity and aural similarity are compared with threshold values (μ, λ). Based on the result of the comparison, the phishing or legitimate nature of the site is informed to the user.

4. EXPERIMENTAL RESULTS

We have collected 20 popular banking site's web pages to test our technique. On Table 1, Experimental Results along with barchart (Figure 8) comparing visual and aural similarity measures. The *fake url* collection is amalgamated for validating the proposed approach. To develop the MASPHID prototype tool, we have used *Python* language and as we are developing an online application, we have used *django* [4] and *Mozilla Firefox* browser to run the program.

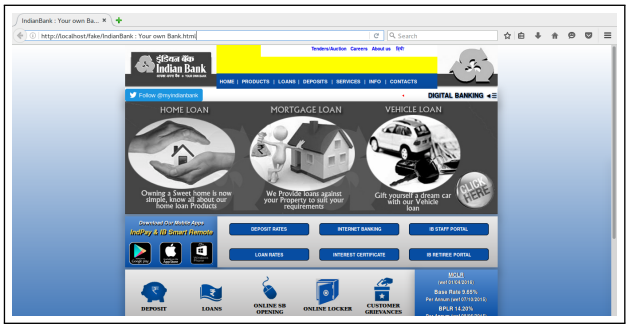

Figure 3: An Altered Sample Page with High Aural and Reduced Visual Similarity

We have provided some screenshots to illustrate the workflow of our technique. When user visit any page, if he/she wants to test the legitimacy (Sample shown in Figure 3 is an altered page for testing with high aural similarity and

[3]http://mathworld.wolfram.com/Root-Mean-Square.html
[4]https://www.djangoproject.com/

Algorithm 1 MASPHID Algorithm

procedure PHISHINGCHECK($link$)
  if $fail \leftarrow Open(link)$ **then**
    Terminate
  else
    $page \leftarrow Open(link)$
    $image_1 \leftarrow Screenshot(page)$
  $url\_name \leftarrow ExtractName(link)$
  $ImageValue \leftarrow Null$
  if $url\_name \in whitelist$ **then**
    $whitelist\_url \leftarrow Extract(whitelist)$
    $page \leftarrow Open(whitelist\_url)$
    $image_2 \leftarrow Screenshot(page)$
    $ImageValue \leftarrow SimilarityCheck(Image_1, Image_2)$
  else
    if $Result \leftarrow Search\_engine(url\_name)$ **then**
      $first\_link \leftarrow ExtractFirstLink(Result)$
      $page \leftarrow Open(frist\_link)$
      $image_3 \leftarrow Screenshot(page)$
      $ImageValue \leftarrow SimilarityCheck(Image_1, Image_3)$
    else
      $Phishing \leftarrow No\ link\ found$
      Terminate
  $auralSim \leftarrow ComputeAuralSim(link, page)$
  if $(ImageValue \geq \mu) \wedge (auralSim \leq \lambda)$ **then**
    return Legitimate Site
  else
    return Phishing site

reduced visual similarity) then he/she run the MASPHID. MASPHID displays the first page url to screen and read the url through screen reader.

If the user know the name of the url then he/she type the name of the url on text field as shown Figure 5.

Then MASPHID searches in whitelist, search engine and finally MASPHID informs the user about the site whether phishing as shown Figure 6 or legitimate as shown Figure 7.

5. DISCUSSION

Phishing is an evolving issue with new dimensions. Researchers are continuously developing new ideas to reduce phishing. Nevertheless day by day phishing crime is increasing. The proposed MASPHID model is an initial version of the approach which shall be strengthened further with additional modules with machine learning based classification etc. There are some operational limitations of the proposed technique, one is dependency on Mozilla Firefox Browser. When user visit any site then their URL is saved in *recovery file* from which MASPHID extracts the URL. But the problem shall arise when the path of recovery file is changed.

6. CONCLUSIONS AND FUTURE WORK

In this paper a novel phishing detection tool entitled MASPHID was presented. The objective of the model is to provide phishing detection for screen reader users with aural and visual similarity measures.

The MASPHID model also incorporates Whitelist technique to detect phishing site. The proposed model is tailored

Figure 4: Original Indian Bank page

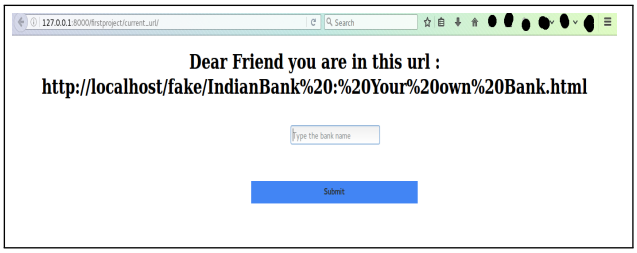

Figure 5: MASPHID - Step I

for detecting phishing sites which are targeted towards persons with visual impairments. The non requirement of visual replica condition makes the persons with visual impairments, a easy target for phishing. This paper has proposed an early-stage solution for detecting phishing sites with high aural similarity and reduced visual similarity. These similarity measures shall be further enhanced with the incorporation of machine learning techniques.

In future, this work shall be strengthened with machine learning based classification approaches. Till now the MASPHID work only on Mozilla Firefox and in future we will deployment it in all major browsers.

7. REFERENCES

[1] S. Afroz and R. Greenstadt. Phishzoo: Detecting phishing websites by looking at them. In *Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on*, pages 368–375, 2011.

[2] APWG. Anti-phishing work group. http://www.antiphishing.org/, Access on 1 june 2016.

[3] Y. Cao, W. Han, and Y. Le. Anti-phishing based on automated individual white-list. In *Proceedings of the 4th ACM Workshop on Digital Identity Management*, DIM '08, pages 51–60. ACM, 2008.

[4] J. Chen and C. Guo. Online detection and prevention of phishing attacks. In *2006 First International Conference on Communications and Networking in China*, pages 1–7, Oct 2006.

[5] K.-K. R. Choo. The cyber threat landscape: Challenges and future research directions. *Computers & Security*, 30(8):719–731, 2011.

[6] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297.

[7] M. Dunlop, S. Groat, and D. Shelly. Goldphish: Using images for content-based phishing analysis. In *Internet Monitoring and Protection (ICIMP), 2010 Fifth*

Figure 6: MASPHID Result - Phishing

Figure 7: MASPHID Result - Legitimate Site

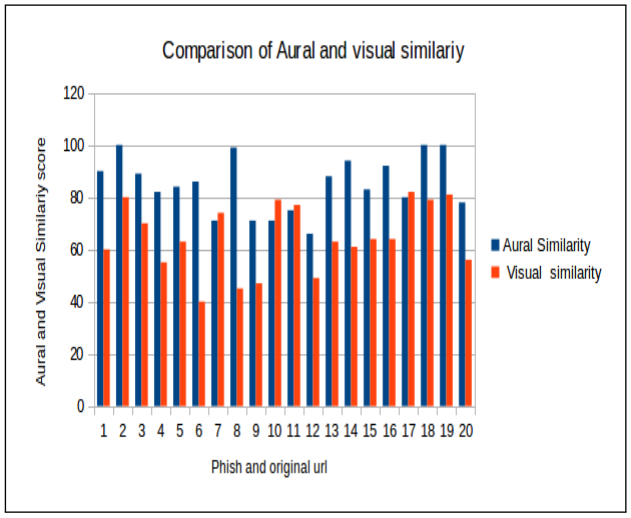

Figure 8: Barchart of Aural and Visual Similarity

International Conference on, pages 123–128, May 2010.

[8] H. M. A. Fahmy and S. A. Ghoneim. Phishblock: A hybrid anti-phishing tool. In *Communications, Computing and Control Applications (CCCA), 2011 International Conference on*, pages 1–5, March 2011.

[9] T. N. Jagatic, N. A. Johnson, M. Jakobsson, and F. Menczer. Social phishing. *Commun. ACM*, 50(10):94–100, Oct. 2007.

[10] S. Kaui and A. Kaur. Detection of phishing webpages using weights computed through genetic algorithm. IEEE, 2015.

[11] I.-C. Lin, Y.-L. Chi, H.-C. Chuang, and M.-S. Hwang. The novel features for phishing based on user device detection. *JOURNAL OF COMPUTERS*, 11(2):109–115, 2016.

[12] S. Mori, H. Nishida, and H. Yamada. *Optical Character Recognition*. 1st edition, 1999.

[13] L. A. T. Nguyen, B. L. To, H. K. Nguyen, and M. H. Nguyen. Detecting phishing web sites: A heuristic url-based approach. In *2013 International Conference on Advanced Technologies for Communications (ATC 2013)*, pages 597–602. IEEE, 2013.

[14] A. P. E. Rosiello, E. Kirda, . Kruegel, and F. Ferrandi. A layout-similarity-based approach for detecting phishing pages. In *Security and Privacy in Communications Networks and the Workshops, 2007. SecureComm 2007. Third International Conference on*, pages 454–463, Sept 2007.

[15] J. A. Schibrowsky, J. W. Peltier, and A. Nill. The state of internet marketing research. *European Journal of Marketing*, 41(7/8):722–733, 07 2007.

[16] S. Unnikrishnan, S. Surve, D. Bhoir, N. M. Shekokar, C. Shah, M. Mahajan, and S. Rachh. Proceedings of 4th international conference on advances in computing, communication and control (icac3'15) an ideal approach for detection and prevention of phishing attacks. *Procedia Computer Science*, 49:82 – 91, 2015.

[17] Y. Wang, R. Agrawal, and B. Y. Choi. Light weight anti-phishing with user whitelisting in a web browser. In *Region 5 Conference, 2008 IEEE*, pages 1–4, April 2008.

[18] L. Wenyin, G. Huang, L. Xiaoyue, Z. Min, and X. Deng. Detection of phishing webpages based on visual similarity. In *Special Interest Tracks and Posters of the 14th International Conference on World Wide Web*, WWW '05, pages 1060–1061, 2005.

[19] WHO. World health organization: Visual impairment and blindness. http://www.who.int/mediacentre/factsheets/fs282/en/, Updated August 2014.

[20] H. Z. Zeydan, A. Selamat, and M. Salleh. Survey of anti-phishing tools with detection capabilities. In *Biometrics and Security Technologies (ISBAST), 2014 International Symposium on*, pages 214–219, Aug 2014.

[21] 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*, WWW '07, pages 639–648. ACM, 2007.

Table 1: Experimental Result

Serial no	Real Url	Fake Url	Aural similarity	Visual Similarity
1	https://www.allahabadbank.in/english/Home.aspx	http://localhost/fake/Allahabad%20Bank-%20home.html	90	60
2	http://www.axisbank.com/	http://localhost/fake/Axis%20Bank.html	100	80
3	http://www.indian-bank.com/	http://localhost/IndianBank%20:%20Your%20own%20Bank.html	89	70
4	http://www.bankofbaroda.com/	http://localhost/fake/Bank%20of%20Baroda%20-%20India's%20International%20Bank.html	82	55
5	http://www.bankofindia.co.in/english/home.aspx	http://localhost/fake/Bank%20Of%20India%20-%20Home.html	84	63
6	http://www.bankofmaharashtra.in/	http://localhost/fake/Maharashtra%20Bank%20-%20Ek%20Parivaar%20Ek%20Bank.html	86	40
7	http://www.icicibank.com/	http://localhost/fake/Personal%20Banking,%20Online%20Banking%20Services%20-%20ICICI%20Bank.html	71	74
8	http://www.canarabank.com/English/Home.aspx	http://localhost/fake/Canara%20Bank.html	99	45
9	http://www.denabank.com/	http://localhost/fake/Welcome%20to%20Dena%20Bank.html	71	47
10	https://www.pnbindia.com/En/ui/Home.aspx	http://localhost/fake/Welcome%20to%20Punjab%20National%20Bank.html	71	79
11	http://www.sbi.co.in/	http://localhost/fake/State%20Bank%20of%20India.html	75	77
12	http://www.syndicatebank.in/	http://localhost/fake/Syndicate%20Bank-%20home.html	66	49
13	http://www.ucobank.com/	http://localhost/fake/UCO%20Bank,%20Global%20Indian%20Bank%20for%20Personal,%20Corporate,%20Rural%20Banking%20Services.html#.V1yEufk4-PQ	88	63
14	http://www.unitedbankofindia.com/English/HomePage.aspx	http://localhost/fake/United%20Bank%20of%20India%20-%20Home.html	94	61
15	http://www.vijayabank.com/	http://localhost/fake/Vijaya%20Bank.html	83	64
16	http://www.hdfcbank.com/	http://localhost/fake/HDFC%20Bank:%20Personal%20Banking%20Services.html	92	64
17	http://www.federal-bank.com/	http://localhost/fake/Federal%20Bank.html	80	82
18	https://www.americanexpress.com/india/	http://localhost/fake/American%20Express%20India%20|%20Log%20in%20|%20Credit%20Cards,%20Travel%20&%20Rewards.html	100	79
19	https://www.jpmorgan.com/country/IN/en/jpmorgan	http://localhost/fake/J.P.%20Morgan%20Home%20|%20J.P.%20Morgan.html	100	81
20	http://www.hsbc.co.in/1/2/homepage	http://localhost/fake/%20HSBC%20India.html	78	56