OPEN ACCESS

International
Journal
Of Modern Engineering Research (IJMER)

Trite Web Page Recognition by providing security and Measuring the Differences Using SIFT

Ayinavalli Ventaka Ramana [1], S Sowjanya [2]
**(Department of Information Technology, GMRIT/JNTUK, India*
*** (Department of Information Technology, GMRIT/JNTUK, India*

ABSTRACT: *Phishing is a form of online identity theft associated with social engineering and is a major threat to information security and personal privacy. Now-a-days, phishing plays a very important role in the network attacks. In this paper, we have measured the differences between the fake web pages and the legitimate web pages. For that we have proposed an anti-phishing method, by integrating text-based similarity and the visual based similarity for the identification of the fake web pages. We have implemented two algorithms: one is Scale invariant Featured Transform (SIFT) algorithm used to extract the signatures based on key-points from the screenshot of the web page. Second one is integration of MD5 and SHA used to generate hash value based on the content of the web page. Both of these values have to be stored in the database. In this paper, we have also providing security for the information of the user by using the RSA algorithm those are entered at the time of registration. When any web page is to be checked, then the two algorithms are to be applied to the web page to generate the signatures and those are compared with the values stored in the database. The results are very effective and accurate by using our proposed system.*

Keywords: *Phishing web pages, MD5&SHA, RSA, Phishing, SIFT.*

I. INTRODUCTION

The problem of plagiarism has recently increased because of the digital era of resources available on the World Wide Web. Phishing web pages are forged web pages which are generally used for phishing attacks. Plagiarism is an act of copying an idea from any possible source and presenting it without any citations to the origin. The act of illegal use can be done in many ways. Several researches have been done in the past decades which determine several techniques commonly practiced. Phishing typically begins with mass email which convinces the reader to visit the included website link. Typically the email looks legitimate and will include a company logo of the targeted site and a return address of the legitimate company which makes the email as a legitimate one at first glance. The phisher wants the lure to be as authentic as possible so that the victim will "bite". The phisher uses many techniques in order to convince the user or the reader of the email to visit the website. These techniques include updating certain information or avoiding account termination, to fill few details to win a lottery etc., many of these phishing techniques try to convince the reader that urgent in action is needed and prey upon emotion such as urgency, account interruption, or account termination. The main aim of the phisher is to steal the confidential information of the user. The phishing process can be done in 3 steps. The phisher creates a duplicate web page same as the original one with some small changes. They send continuous mails URL of the fake website to the user for clicking on that link. The users who click on that link are focused to the fake web page; in that the user's personal information may get stolen easily. Creating a phishing web site is very easy and those web sites are created within a very short time. By these phishing attacks, many of the organisations get lost their name and fame. Recognizing the trite webpage is not so easy because the phisher may create the fake webpage based on the content, image or layout. Many of the trite web pages have text or image similarity. In this paper, we have implemented a method for the detection of trite web pages. The administrator of the proposed system has to register the web pages those are original ones. User has a chance to check the web pages with the original web pages. In this paper, we have also provided security for the information of the user those are entered at the time of registration. For that, we have using the RSA algorithm for providing security. We have taken the screenshot of the webpage and stored it as an image in our system. And we have applied the algorithms for that webpage and those details are to be stored in the database. The MD5 algorithm has some disadvantages. So in this paper, we are using the combination of both MD5 and SHA to get accurate results. The server is trained to generate hash value for the web pages by using the integration of MD5 and SHA algorithms based on the content of the page. And the server is also generating key points based on the screenshot of the webpage. When any

suspected webpage is there, the system generates the hash value and key features for the current webpage. The recognition of the trite webpage is happened based on comparison of the both the key points and the hash value of the webpage with the details that are stored in the database for the original webpage.

II. Literature Survey

Phishing is a significant problem that tricks unsuspecting users into revealing private information involving fraudulent email and websites. This causes tremendous economic loss every year. Efforts to detect phish can be implemented at the phishing website level. To prevent phishing websites from reaching potential victims, traditional detection techniques such as Bayesian filter, blacklist, and rule based rankings can be applied. Generally speaking, research to detect phish at the website level falls into two categories: heuristic approaches, which use HTML or content signature to identify phish, and blacklist based methods.
Phishing is the most online fraudulent activity which is not new to the internet users.
Below are some categories of phishing attacks.

2.1 Spoofing e-mails and web sites

Phishing attacks fall into several categories. The earliest form of phishing attacks were e mail based and they date back to the mid 90's. These attacks involved spoofed e-mails1 that was sent to users where attackers tried to persuade the victims to send back their passwords and account information. Although such attacks may be successful today, the success rate from the point of view of the attackers is lower because many users have learned not to send sensitive information via e-mail.

2.2 Clone Phishing

It is a type of an attack where a legitimate previously delivered email containing an attachment o-r link has had its content and recipient address (es) taken and used to create a cloned email. The attachment or link within the email is replaced with a malicious version and then sent from an email address spoofed to appear to come from the original sender. It may claim to be a re-send of the original or an updated version to the original. This technique could be used to pivot (indirectly) from a previously infected machine and gain a foothold on another machine, by exploiting the social trust associated with the inferred connection due to both parties receiving the original email.

2.3 Spear Phishing

It is a technique where specific victim is targeted. The information about the victim is known prior to the attack and the email is sent from the source known by the victim. Due to the nature of the trust on receiving email, this kind of attack has high probability to be successful. An example would be receiving an email from friend, colleague or financial institutions which prompt victim to provide the credentials.

2.4 Detection Based on Text

Phishing Web page detection is similar to duplicated or phishing document detection in some extent and the document similarity evaluation techniques can be used for the Anti Phishing task. Earlier research works on duplicated document detection approaches focus on plain text documents and use pure text features. Those are

- Collection statistics
- Syntactic analysis
- Displaying structure
- Vector space model

We proposed an effective hashing algorithm by the integration of MD5 & SHA in order to generate secure hash for the web page text which is unique for each and every web page. The dissimilarities can be easily detected using this hashing algorithm because plagiarized web pages cannot fingerprint the original web page hash signature.

2.5 Detection Based on Visual Similarity:

The visual similarity between two Web pages is evaluated in three metrics: block level similarity, layout similarity, and overall style similarity, which are based on the matching of the salient block regions. In our approach we take the web page snapshot and rescale the snapshot so that it can match the requirements of the SIFT algorithm in order to generate key points. Our visual similarity method depends on these key factors to assess the dissimilarities of two web pages.

III. Proposed System

The proposed system finding a solution for the phishing problem whether the webpage is fake or original. Our proposed system differs from the previous solutions for the phishing. In this, we have implemented the algorithm by combining both the textual based and the visual based methods to get efficient and accurate results for the recognition of phishing webpages. We have implemented three algorithms in this proposed system, in which the first algorithm is sift (scale invariant feature transform) algorithm and this algorithm is used to generate stable key points for every webpage based on the screenshot of the webpage. These key points are unique for each and every webpage. By using these visual similarity measures we didn't say that the webpage is fake or original accurately. For this reason, we have implemented another algorithm that is the integration of both sha&md5 algorithms and this is used to generate a hash value for the webpage based on the content of the webpage. By using these two values we recognize whether the webpage is original or fake. When there is any suspected page, then key points and the hash values are compared with the database. If both are similar then the current webpage is a legitimate webpage otherwise the current webpage is a trite webpage. Along with that, we have implemented an rsa algorithm for providing security for the passwords. When any intruder hacks the password of the administrator then he has a chance to store the trite webpages in the database. When any one wants to check the webpage and it matches with the fake webpage. It gives the result as legitimate webpage. If these signatures are similar then the current web page is a decriminalized web page or the web page is considered as a plagiarized web page.

IV. Methodology

We engaged the following steps for the recognition of trite webpages:

1. We have to store the webpages that are required to protect against phishing attacks for each and every web page

2.1 A hash value is generated based on the textual pattern of the web page using MD5 & SHA hashing algorithms to find the integrity of the webpage.

2.2 A screenshot for the web page is taken and store it as an image. For this, we have applied SIFT algorithm in order to generate featured key-points.

3. We have also stored the user details and admin details in the database and we have provided security for that data by using the security algorithm called RSA algorithm.
4. Store the signature and the trained image screenshot of every web page that is registered at step 1 in the database.
5. When a new web page is required to be checked, perform step 2.
6. These newly generated web page signature and key points are compared with its corresponding signature and clustered key-points from the database
7. If the verified web page is similar with that of web page that is stored in the database then the current web page is considered as legitimate web page and if they vary then the web page is considered as plagiarized web page.

4.1 Scale Invariant Feature Transform Algorithm

Scale Invariant Feature Transform is an algorithm which was published by David Lowe in 1999, this algorithm is implemented for extracting invariant features from images that can be used to perform reliable matching between different views of an image. This algorithm also provides accurate recognition from suspected web pages by matching individual features to a database of features from known web pages.

The main purpose of this algorithm is to derive and describe key points which are robust to scale, rotation and change in illumination. We analyzed the algorithm in six steps

(A) Scale-space extrema detection: In order to create a scale space, the image (in our case the web page screen shot) is taken and generates progressively blurred out images. Then the original image should be resized to half of its original size and generate blurred out image again and this process should be repeated until necessary octaves and blur levels (scales) are generated. The generation of progressively blurred images was done using Gaussian blur operator.

Mathematically "blurring" for this algorithm is referred to as convolution of Gaussian operator and the image

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y)$$

Symbols

L is a blurred image
G is a Gaussian blur operator
x , y are location coordinates
σ is the scale parameter

* Is the convolution operator for x,y coordinates, this applies Gaussian blur G on image I

$$G(x,y,\sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}$$

The above represented equation is to calculate the Gaussian blur operator.

(B) LOG Approximation: For Laplacian of Gaussian (LOG) we need to take the image(web page screenshot) and blur it a little and then we need to calculate the second order derivatives on it(or "Laplacian"). This LOG locates edges and corners on the image which are good for finding key-points. As the second order derivatives are sensitive to noise they are intensive computationally, we calculated the difference between two consecutive scales i.e., Difference of Gaussian (DOG). Similarly all the consecutive pairs are considered to perform DOG operation.

(C) Finding Key points: Finding the key points includes two parts Locate maxima/minima in DOG images-here the trained system iterated through each pixel and check all its neighbors, this process is checked for current image along with the Images that are above and below it. And the points that are checked are the "appropriate" maxima and minima

Find subpixel maxima/minima-using the available pixel data, subpixel values are generated using the method Taylor expansion of image around the approximate key-point which is mathematically represented as follows

$$D(x) = D + \frac{\partial D^T}{\partial x} x + \frac{1}{2} x^T \frac{\partial^2 D}{\partial x^2} x$$

On solving the above equation subpixel key point locations will be generated, this increases the chances of matching and robustness of the algorithm.

(D) Get rid of bad key points: Previous step produces lot of key points but some of them lie on edge or they don't have enough contrast. In both the cases they are not useful as features. Harris Corner Detector [23] technique is implemented in this algorithm to remove edge features. And the intensities of the key-points are also calculated in order to verify the contrast. To calculate the intensity we again used the Taylor expansion for the sub-pixel key points to get the intensity value of the subpixel location. In general the image around the key point can be

A flat region: both the gradients will be small

An edge: the perpendicular gradient will be big and the other will be small.

A corner region: both the coordinates are big.

Depending on their location coordinates also there is a scope to eliminate the unnecessary key-points by calculating threshold maxima/minimal values.SIFT Detector Parameters is Threshold to maintain every image threshold value. For every given input it detects the edges. To measure the edges we have parameter known as EdgeThreshold.

(E)Orientation assignment: Now we got legitimate key points which are tested as stable. And we already know the scale invariance that is the scale at which the key point is detected. Gradient magnitude and orientation is calculated with the below formulae

$$m(x,y) = \sqrt{(L(x+1,y) - L(x-1,y))^2 + (L(x,y+1) - L(x,y-1))^2}$$

$$\theta(x,y) = tan^{-1}\left((L(x,y+1) - L(x,y-1))/(L(x+1,y) - L(x-1,y))\right)$$

All the pixels around the key-point will be calculated using these formulae to generate magnitude and orientation. Most prominent orientation(s) is figured out and is assigned to the key-point. The size of the "orientation collection region" around the key-point depends on its scale. After calculating orientation a histogram is generated. In this histogram, the 360 degrees of orientation are broken into 36 bins (each 10 degrees). Let's say the gradient direction at a certain point (in the "orientation collection region") is 17.589 degrees, then it will go into the 10 to19 degree bin.

Generate SIFT Features: Now a unique fingerprint is generated for every keypoint. The orientation histograms summarize the contents over 44 regions with 8 orientation bins which can allocate 128 element features for each key-point. Correspondence of feature points is generated by the ratio of distance for the descriptor vector from the closest neighbor to the distance of second closest.

Message Digest 5 (MD5) Algorithms:

Message Digest (MD5) algorithm was developed by Ron Rivest, which is a secure version of his previous work MD4. MD5 is the most widely used secured algorithm which takes input as a message of arbitrary length and produces 128 bit output for the inputted message. The message processing for MD5 involves following steps

(A) Padding: the message is padded with single 1 and followed by 0's to ensure that the message length plus 64 is divisible by 512. Padding is must even if the length of the message is congruent to 448 modulo 512

(B) Appending length: the result of step 1 is attached with 64 bit binary representation of original length. This should be a multiple of 512 bits. Equivalently, this message has a length that is an exact multiple of 16(32-bit) words. Let S[0 ---- N-1] denote the words of resulting message, where Nis a multiple of 16.

(C) Initialize MD Buffer: In order to compute the message digest a four word buffer (A,B,C,D) is used, where each A,B,C,D is a 32-bit register. These registers are initialized with the following Hex values, which initiates the lower-order bytes first.

Word A : 01 23 45 67
Word B : 89 ab cd ef
Word C : fe dc ba 98
Word D : 76 54 32 10

(D)Process message in 16-Word Block: four auxiliary functions are defined three 32-bit words as input and produce one 32-bit word as output.

F(X, Y, Z) = XY v not(X) Z
G(X, Y, Z) = XZ v Y not(Z)
H(X, Y, Z) = X xor Y xor Z
I(X, Y, Z) = Y xor (X v not(Z))

If X,Y and Z bits are independent and unbiased, then each bit of F(X,Y,Z), G(X,Y,Z), H(X,Y,Z) and I(X,Y,Z) will be independent and unbiased.

(E)Output: the message digest is produced as output is A,B,C,D. this output is generated by initiating with the lower-order byte of A and by concluding with the higher-order byte of D.

The main logic used to implement this algorithm is to check the integrity of the web page. When a web page is needed to verify whether it is legitimate or fake, MD5 generates the signature for that web page and is compared with its corresponding signature stored in the database. If both the hashes are similar then the web page is considered as original or the integrity of the web page is considered as dishonest.

THE RSA ALGORITHM

The RSA algorithm is used for both public key encryption and digital signatures. It is the most widely used public key encryption algorithm. The basis of the security of the RSA algorithm is that it is mathematically infeasible to factor sufficiently large integers. The RSA algorithm is believed to be secure if its keys have a length of at least 1024-bits.

RSA Algorithm

1. Choose two very large random prime integers:
   p and q
2. Compute n and $\varphi(n)$:
   $n = pq$ and $\varphi(n) = (p-1)(q-1)$
3. Choose an integer e, $1 < e < \varphi(n)$ such that:
   $\gcd(e, \varphi(n)) = 1$(where gcd means greatest common denominator)
4. Compute d, $1 < d < \varphi(n)$ such that:
   $ed \equiv 1 \pmod{\varphi(n)}$

- the public key is (n, e) and the private key is (n, d)
- the values of p, q and $\varphi(n)$ are private
- e is the public or encryption exponent
- d is the private or decryption exponent

Encryption

The cipher text C is found by the equation 'C = M^e mod n' where M is the original message.

Decryption

The message M can be found form the cipher text C by the equation 'M = C^d mod n'.

V. Results

In proposed system we have maintained two types of datasets, first one is the trained dataset and it contains the database of the original webpages or legitimate webpages. The second one is the test database and it contains the database of the trite webpages. To test the performance of our proposed system, we collected some web pages by using different keywords on different categories i.e., banking, mail, social networking. When any webpage is to be checked then the values are to be checked with both of the trained and test databases. If it matches with the trained database values then the webpage is considered as a legitimate one. Otherwise it matches with the test dataset then the current webpage is considered as a trite one.

Implementing SIFT to extract Image Signatures

We considered SIFT for measuring the dissimilarities between the web page screenshots. To compute Scale Space, SIFT scale space parameters are SigmaN, Sigma0, O is for measuring Number of Octaves, S for Number of Levels, Omin for First Octave, in addition to these Smin, Smax are used to calculate threshold values. To generate key points, maxima/minima and threshold values we use SIFT Detector Parameters which are Threshold to maintain every image threshold value. For every given input it detects the edges. To measure the edges we have parameter known as Edge Threshold. For orientation assignment and histogram generation, there are SIFT Descriptor parameters NBP: Number of Spacial Bins, NBO Number of Spacial Orient Bins.

Training is done with the default parameters SigmaN is 0.50000, SigmaO is 2.015874. Number of Octaves per iteration is 6. In each Octave Number of Levels are 3. Number of Spatial and Orientations are 4 and 8 respectively.

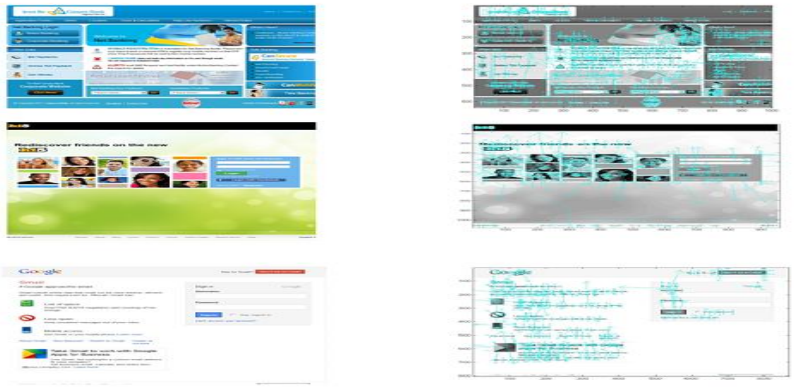

Figure1 implementing SIFT to generate key points on various web page snapshots

Implementing MD5&SHA to extract Text Signatures:

After extracting the HTML text between the HTML tags, a digital signature is generated using md5&SHA algorithms written in java code using the extracted text as input and generates a digital hash signature as its output.

The digital hashing signatures are generated in both the phases of our detection mechanism namely training and testing phases, these generated signatures are unique for each and every web page. If the web page generates unique signature then the webpage is stated as legitimate one else the web page is stated as obscure one.

WEBSITE	HASH SIGNATURE	REAL OR FAKE
AOL	20abb1dd5a1d02878251b59c1fc5b0db	Fake
D2JSP	16a480f056dd6f672fd1f24054381932	Fake
eBay	57ee5af0ce8a5b64438b7e8213576edd	Real
D2JSP	17cf453f0e8b010665354bfedb400ee3	Real
Gmail	4f0c64913da7adafa2132c833bc723e1	Fake
Gmail	b188c5a447427725e7d6b3c93b1f2b9c	Real
ICICI Bank	d771d64a3ec92556c61e61f78d5fe0fe	Real
Yahoo Mail	da8b3e36d408f0988c26a96a1563a271	Fake

Figure 2 represents a table showing a list of few real and fake web pages and their unique hashes from our database.

Combining Digital and image signatures:

Our aim is to combine both the image signatures and the digital signature in order to produce efficient and accurate results from our trained system. By combining the Image signature as well as digital signature we achieved the following results.

S.No	Image signature	Digital signature	Un-Identified
1	34/32	153/1	1
2	49/48	89/3	0
3	25/25	12/2	0
4	118/116	87/0	1
5	89/88	40/1	0
6	172/172	89/1	0
7	25/25	42/0	0

Table 1 Combining both the signatures to generate final output

VI. Conclusion and Future Work

In this paper, we offered a successful approach in order to recognize trite web pages by comparing visual similarities along with the textual similarities between a suspicious web page and the potential, legitimate target page. As my proposed system is purely trained server side system there is no burden on the client in order to justify whether the received web page is legitimate or not. We have implemented an RSA algorithm to provide security for the passwords. We considered a visual similarity based approach because the users are typically believed that they are visiting a legitimate page by the look-and-feel of a web site. But only visual similarity based approach might not derive efficient and accurate results. For this reason we integrate digital signature based approach along with visual similarity based anti phishing solution. We performed an experimental evaluation of our comparison technique to assess its accuracy and effectiveness in detecting trite web pages. We used a dataset containing real trite phishing pages with their corresponding target legitimate pages in our trained system and the results are satisfactory.

The future work can be extended by implementing a hybrid mechanism to detect plagiarized web pages using layout based detection mechanism along with the designed system which can produce better results than the proposed mechanism.

REFERENCES

[1]. APWG. http://www.anti-phishing.org/.
[2]. Microsoft.SenderIDHomePage. http://www.microsoft.com/mscorp/safety/technologies/senderid/default.ms%px
[3]. Yahoo. AntiSpam Resource Center. http://antispam.yahoo.com/domainkeys.
[4]. Y.Zhang, J.I.Hong,L.F.Cranor, CANTINA: A Content-based approach to detecting phishing web sites, in: the international World Wide Web conference, www 2007, ACM Press, Banff, Alberta, Canada, 2007, pp. 639-648.
[5]. Microsoft Corporation. PhishingFilter: Help protect yourself from online scams. http://www.microsoft.com/protect/products/yourself/phishingfilt er.mspx
[6]. Mozilla Project. Firefox phishing and malware protection. http://www.mozilla.com/en-US/firefox/phishing-protection/.
[7]. R. Dhamija and J.D. Tygar, "The Battle Against Phishing: Dynamic Security Skins," Proc. Symp. Usable Privacy and Security, 2005.
[8]. Google, Inc. Google safe browsing for Firefox. http://www.google.com/tools/firefox/safebrowsing/.
[9]. stopbadware.org. Badware website clearinghouse. http://stopbadware.org/home/clearinghouse
[10]. Firefox add-on PhishTank SiteChecker, https://addons.mozilla.org/en-us/firefox/addon/phishtanksitechecker/
[11]. Firefox add-on for anti-phishing Firephish, https://addons.mozilla.org/en-US/firefox/addon/firephish-antiphishing-extens/
[12]. CallingID LinkAdvisor, www.callingid.com
[13]. SpoofGuard is a tool to help prevent a form of malicious attack, http://crypto.stanford.edu/SpoofGuard/
[14]. SpoofStick a simple browser extension for IE and Firefox, http://www.spoofstick.com/
[15]. Y. F. Anthony, W. Liu, X. Deng. "Detecting Phishing Web Pages with Visual Similarity Assessment Based on Earth Mover's Distance (EMD)". IEEE Transactions on Dependable and Secure Computing, October 2006, Volume 3 (4), 301-311.
[16]. Angelo P.E. Rosiello, Engin Kirada, Christopher Kruegel and Fabrizio Ferrandi. A Layout-Similarity Approach for detecting Phishing pages.
[17]. Eric Medvet, Engin Kirda,and Christopher Kruegal. Visual- Similarity-Based Phishing Detection
[18]. Yu Meng, Dr. Bernard Tiddeman, Implementing the Scale Invarient Feature Transform(SIFT) Method.
[19]. T.C. Hoad and J. Zobel, "Methods for Identifying Ver-sioned and Plagiarized Documents," J. Am. Soc. Infor-mation Science and Technology, vol. 54, no. 3, pp. 203-215, 2003.
[20]. David.G.Lowe, Distinctive Image Features from Scale-Invariant Keypoints.
[21]. Janaka Deepakumara, Howard M.Heys and R. Venkatesan, FPGA Implementation of MD5 Hash Algorithm.
[22]. Anti-Phishing Group of the City University of Hong Kong, http://antiphishing.cs.cityu.edu.hk, 2005.

[23]. W. Liu, X. Deng, G. Huang, and A.Y. Fu, "An Anti-Phishing Strategy Based on Visual Similarity Assess-ment," IEEE Internet Computing, vol. 10, no. 2, pp. 58-65, 2006.
[24]. W. Liu, G. Huang, X. Liu, M. Zhang, and X. Deng, "Detection of Phishing Web Pages Based on Visual Similar-ity," Proc. 14th Int'l World Wide Web Conf., pp. 1060-1061, 2005.
[25]. T. Nanno, S. Saito, and M. Okumura, "Structuring Web Pages Based on Repetition of Elements," Proc. Sev-enth Int'l Conf. Document Analysis and Recognition, 2003.
[26]. A. Emigh, "Online identity theft: Phishing technilogy, chokepoints and countermeasures, " Radix Labs, Tech. Rep., 2005, retrieved from Anti- Phishing Working Group: http://www.antiphishing.org/resources.html.
[27]. W. Liu, G. Huang, X. Liu, M. Zhang, and X. Deng, "Detection of Phishing Web Pages Based on Visual Similarity," Proc. 14th Int'l World Wide Web Conf., pp. 1060-1061, 2005.
[28]. PhishGuard.com. Protect Against Internet Phishing Scams http://www.phishguard.com/.