Journal of Engineering and Applied Sciences 13 (Special Issue 8): 6371-6377, 2018
ISSN: 1816-949X

Website Forgery/Falsification Detection System Using Link Information and Images

Ji-Ho Cho and Geuk Lee
Department of Computer Engineering, Hannam University, Dae-Jeon, Republic of Korea

Abstract: In this study, we propose a website forgery/falsification detecting system that uses link information and images to detect website forgery/falsification. The proposed system first checks the URL address of each financial site to see whether the URL address is a normal one when the user accesses the financial site. Second, the proposed system collects images and web page link information of the page and compares them with normal web page information which is saved in previous to detect whether the site is abnormal. Third, the proposed system repeats the second phase while following the sub web pages linked in the page. As the number of sub page comparisons increases, the accuracy of detection is also increased.

Key words: Link information, website forgery/falsification detecting, feature matching, Levenshtein distance ORB algorithm, proposed system

INTRODUCTION

Use italic for emphasizing a word or phrase. Do not use boldface typing or capital letters except for section headings (cf. remarks on section headings). Use a laser printer, not a matrix dot printer users of financial services using online has increased continuously for convenient. In South Korea, based on the end of September 2016, the number of registered internet banking customers (including mobile banking) is 120.72 million people. The number of daily usage of internet banking (including mobile banking) is average 87.9 million. This is increased by 8.6% compared to the previous quarter (Bank of Korea, 2016).

Internet banking and mobile banking services enable handling various transactions such as deposit and loan inquiries and money transfers conveniently whenever and wherever with minimum spatial restriction. The increasing rate of mobile banking service user is higher than internet banking services user. However, thus far, the number of customers using internet banking services is larger than the number of mobile banking service customers. Hackers make a pharming site similar to the actual internet banking site. Hackers use this pharming site, so as to induce a user to enter the personal information that can be used for internet financial transactions. This process is also similar to the actual banking transaction process. The user has difficulty to distinguish the forgery or falsification website with normal one. Many institutions such as the National Cyber Security Center of the National Intelligence Service and the Computer Emergency Response Team Coordination Center of Korea Internet and Security Agency are detecting and responding to forged/falsified sites. However, these institutions have limitation to protect and respond to all personal cyber threats. A that makes individual users can identify and detect forgery/falsification site is necessary.

In this study, we propose a website forgery/falsification detecting system that uses link information and images to detect website forgery/falsification. The proposed system first checks the URL address of each financial site when the user accesses the financial site. Next, the proposed system collects images and web page link information (hyperlink information) of the page and compares them with normal web page information. Using the collected and compared information, the proposed system identifies forged/falsified sites.

Literature review

ORB algorithm: The ORB (Oriented Fast and Rotated BRIEF) algorithm was developed by Open CV (Open Computer Vision) Labs. The ORB algorithm is an image feature points detection algorithm proposed based on the FAST (Features from Accelerated Segment Test) (Rosten and Drummond, 2006) algorithm and the BRIEF (Binary Robust In dependent Elementary Features) (Calonder *et al*., 2010) algorithm. The ORB algorithm is implemented as shown in Fig. 1 (Calonder *et al*., 2010; Rublee *et al*., 2011; Anonymous, 2014).

Corresponding Author: Ji-Ho Cho, Geuk Lee, Department of Computer Engineering, Hannam University, Dae-Jeon, Republic of Korea

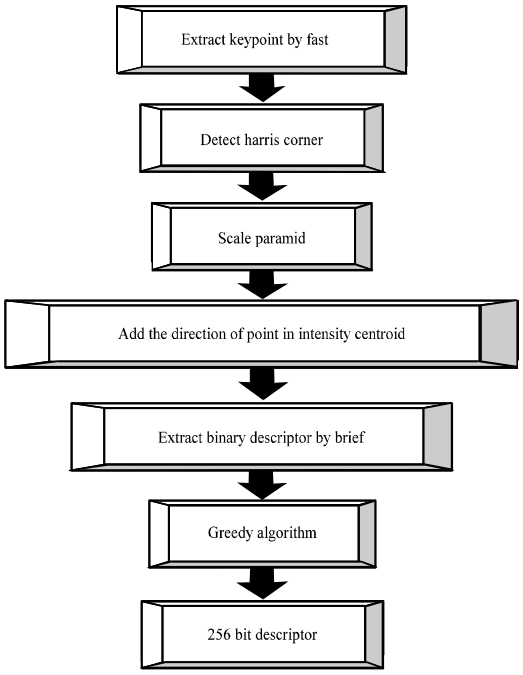

Fig. 1: ORB algorithm

Levenshtein distance algorithm: The name Levenshtein distance algorithm was derived from the name of Russian scientist Vladimir Levenshtein. The Levenshtein distance algorithm is also called edit distance algorithm. The Levenshtein distance algorithm is an algorithm designed to measure the similarity of two character strings. The Levenshtein distance algorithm is used for spell checking, speech recognition and plagiarism checking. The Levenshtein distance algorithm compares two character strings using a two-dimensional array and performs insertion, deletion and changes in each character string to obtain a minimum edit distance value. The accumulated values of minimum edit distances obtained from individual parts using the Levenshtein distance algorithm is final edit distance value of the two character strings. The final edit distance value is used as a measure of similarity (Yu *et al*., 2015).

Table 1 shows the process of the Levenshtein distance algorithm which compares two strings 'ALLIGATOR' with 'ELEVATOR' to obtain the edit distance. When 'ALLIGATOR' and 'ELEVATOR' are compared, if the values are equal, the value above the left diagonal line should be brought. When 'ALLIGATOR' and 'ELEVATOR' are compared, if the values are different, the smallest value among the values on the top, on the left and above the left diagonal line plus 1 should be brought. When the operation of the Levenshtein distance algorithm has been completed, the value at the right bottom becomes the edit distance of the two character strings.

Table 1: Edit distance value between string 'ALLIGATOR' and 'ELEVATOR'

	String A									
Edit distance	–	A	L	L	I	G	A	T	O	R
String B										
-	0	1	2	3	4	5	6	7	8	9
E	1	1	2	3	4	5	6	7	8	9
L	2	2	1	2	3	4	5	6	7	8
E	3	3	2	2	3	4	5	6	7	8
V	4	4	3	3	3	4	5	6	7	8
A	5	4	4	4	4	4	4	5	6	7
T	6	5	5	5	5	5	5	4	5	6
O	7	6	6	6	6	6	6	5	4	5
R	8	7	7	7	7	7	7	6	5	4

MATERIALS AND METHODS

Forged/falsified website detecting system: In this study, we propose a website forgery/falsification detecting system using link information and images. When the user accesses a financial site, the proposed detecting system is activated to check the site. The detecting system collects, compares and analyzes the information on the financial site accessed by the user and the information on pre-saved normal website information. The system informs the user whether the website is forged/falsified or not.

Figure 2 is a block diagram of the system presented in the study to detect forged/falsified websites. The system compares the image and link information collected from the top page of the current website with the image and link information collected from the top page of the normal website. We define this process called 'one phase'. If the result of one phase is determined to be abnormal, the result will be informed to the user and the operation will be terminated. If the result is determined to be normal, the system moves to the next linked sub page and repeats the process to collect and compare image and link information for the sub page to determine the site is normal or not. Phases can be repeated up to the nth depth (bottom) of the detection tree as shown in Fig. 2.

Operation of the detecting system: To detect a forged/falsified website, the detecting system repeats the collection, comparison and analysis process n times according to link information. As shown in Fig. 3, the detecting system implements and repeats five steps (1. URL comparison, 2. image and link information collection, 3. image analysis, 4.link information analysis, 5. forgery/falsification discrimination). Steps from collection to discrimination is called 'one phase' as described in previous part.

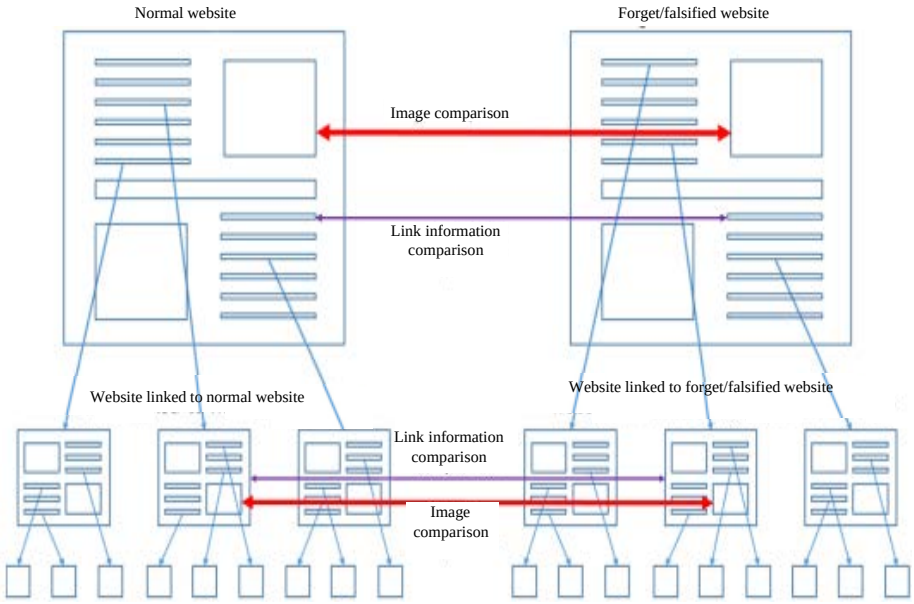

Fig. 2: Tree sructure of the detecting system

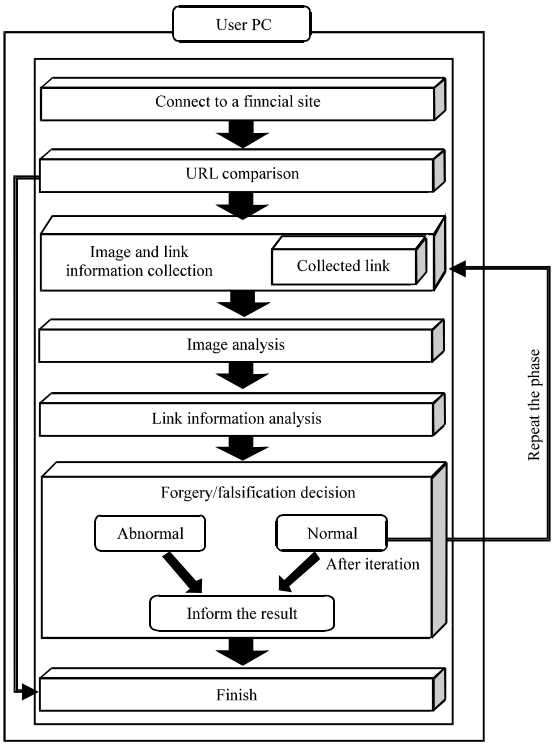

Fig. 3: Operation of the detecting system; a) Comparision result 1and b) Comparision result 2

RESULTS AND DICUSSION

URL comparison: The system compares the URL of the page currently accessed by the user with the normal URL addresses that have been already collected. When the URL addresses are same, the system progresses to the next step (step to collect web page internal images and link data). When results of comparison of URL addresses are not same, the detection process is terminated.The system uses the Levenshtein distance algorithm in this URL comparision step.

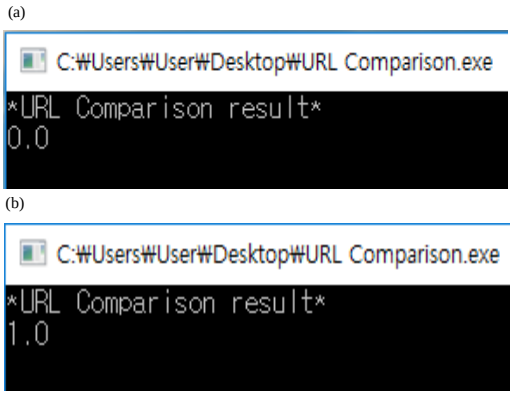

Fig. 4: Results of URL comparison using the Levenshtein distance algorithm

Figure 4 shows the resultant value obtained by comparing the URL address of the currently accessed page with the URL addresses collected in advance using the Levenshtein distance algorithm. Figure 4a shows a result value is 0.0. It means the two URL character strings are different from each other even only in a part. Figure 4b shows the result value is 1.0. It means the two character strings are accurately same.

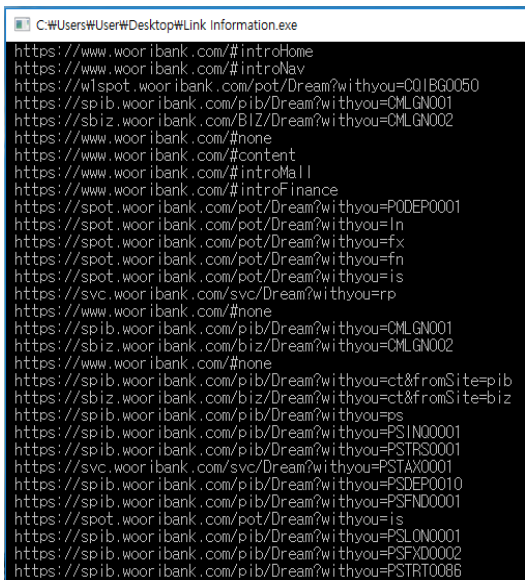

Fig. 5: Result of link information collection (captured screen: a) Image 1 and b) Image 2

Image and link information collection: In this step, images and link information are collected. In the image collection part, images existing inside the accessed page are captured. The captured images are used in the image analysis step. The link information collection step uses jsoup, a JAVA library for HTML work, to crawl the page currently accessed by the user. The system collects link information of the web page using the crawler.

Figure 5 shows the results of link information collection by crawling the main page of a certain financial site using jsoup. The number of links collected from the financial site used in the experiment was larger than 250 and Fig. 5 shows a part of them.

Image analysis: Open CV 7 is used for image recognition and image retrieval and it provides many image processing algorithms. We use ORB algorithm in Open CV for the feature matching. The ORB algorithm determines whether the images collected in the image collection step are matching with original images. The ORB algorithm finds key points of individual images and matching operation. As for the results of the matching operation, the time spent for the image matching, the number of matching features and whether the images are same with original ones are shown.

Figure 6 shows captured images of the full screen of the main page of a certain South Korea financial site. The parts marked with red squares in a) and b) of Fig. 6 look similar but are different. Figure 7 shows the enlarged images of the red square parts in Fig. 6.

Figure 8 shows the results of image matching for Fig. 6 and 7. In Fig. 8a is the results of matching between a and b in Fig. 6.

Figure 6 which corresponds to a in Fig. 8 is decided to similar images as the matching time was of 1017 msec and the number of matching features was 430. However, although a and b in Fig. 6 look similar but are different images. The feature matching value of Fig. 6 indicates that the images are similar despite that they are different in detail. Therefore, the images match with each other cannot be determined based on only the results shown in Fig. 6. the partial image matching 1 in Fig. 8b shows the result of matching between a and b in Fig. 7. In Fig. 8b, the partial image matching1 is decided to different images as the matching time was 19 msec and the number of matching features is 0.

Although, accurate decision is not possible with the results of analysis of the full image as shown in Fig. 8a, sophisticated decision is possible through partial image analysis as shown in Fig. 8b) (Ji-Yong *et al*., 2016).

Link information analysis: In the link information analysis step, the information collected in the link information collection stage are compared and analyzed with the normal web page link information using the Levenshtein distance algorithm. The results of using the Levenshtein distance algorithm have values in a range of 0.0-1.0. The value becomes 1.0 when the comparison and analysis results of the link information are exactly same. When the comparison and analysis results of the link values are other than 1.0, it means they are different.

Figure 9 shows the result screen shot in which the link information collected in the link information collection step is compared and analyzed with normal one. In Fig. 9, the result value of a was 1.0, indicating that the link information is the same. In Fig. 9, the result value of b was 0.8701803051317615 indicating that the link information is slightly different.

Forgery/falsification decision: In the forgery/falsification decision step, the system uses the result values obtained in the image analysis step and the link information analysis step. The system decides whether the web page is a forged/falsified page or a normal page. If the result values obtained in the image analysis step and the link information analysis step are abnormal, the page will be classified into forged/falsified pages and the detection process will be terminated. If the result values are normal, the system will move to the sub page of the

Fig. 6: Two full screen images of Korea Woori Bank website: a) Image 1and b) Image 2

Fig. 7: Two partial images of Korea Woori Bank website (red square parts in Fig. 7): a) Image 1 and b) Image 2

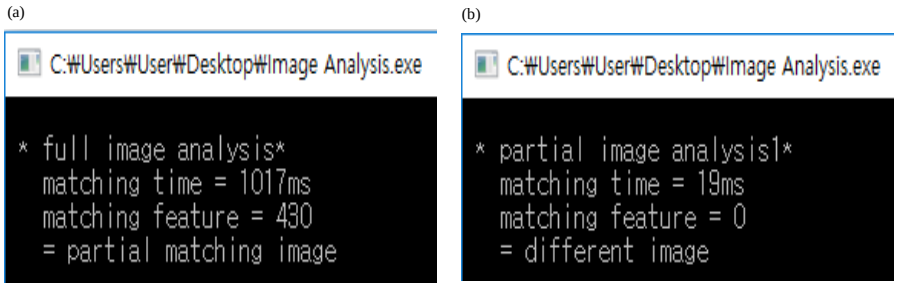

Fig. 8: Result of image analysis of Fig. 7 and 8 (captured screen): a) Result of analysis between Fig. 7a, b and b) Result of analysis between Fig. 8a, b

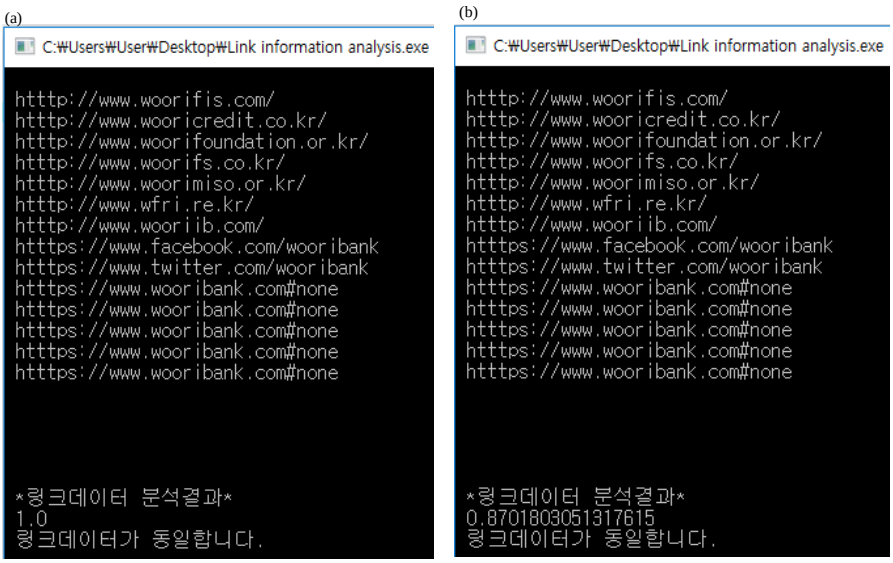

Fig. 9: Link information analysis (captured screen): a) Analysis result of same site and b) Analysis result of similar site

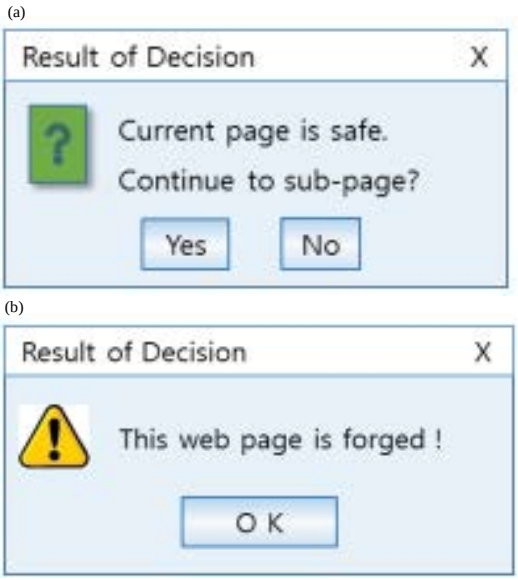

Fig. 10: Result of forgery/falsification decision; a) Normal page and b) Abnormal page

link information and performs phase again from the image and link information collection step in the sub page. This phase is repeated n times following the link information. As the number of sub page comparisons increases, the accuracy of detection is also increased.

Figure 10 shows the result screen shots after forgery/falsification decision. Figure 10a shows a case where the page is decided to a normal page. After the page is decided to a normal page, whether the detection process should be continued to sub-pages or not is asked. Figure 10b shows a case where the page is decided to an abnormal page. The detection process is terminated in this case.

CONCLUSION

Although, phishing and pharming attacks for financial sites increase, most of all users cannot distinguish forgery/falsification sites with normal ones. This study proposes a forgery/falsification detecting system that uses link information and images to detect forged/falsified website. When the user accesses a certain financial web site, the proposed system compare the URL address currently accessed by the user with URL addresses collected in advance. If the URLs match, the system performs the internal image and link information collection steps of web page. In the image collection step, images in the current page are collected. In the link information collection step, all link information connected to the web page are collected using jsoup. In the image analysis step, images are compared using feature matching method of ORB algorithm. In the link information analysis step, the current link information are compared with the normal web page link information using forgery/falsification decision stage, normal and non-normal data are discriminated using the results obtained in the image analysis step and the link information analysis step. If the result from the decision step is abnormal, the detection process is terminated. If the result is normal, the system moves to the sub page indicated by the link data to repeats the detection phase again from the image and link information collection step to

forgery/falsification decision step. In the proposed detecting system, the accuracy of forgery/falsification detection rate is improved as the detection phase is repeated in sub-pages.

ACKNOWLEDGEMENT

This research was supported by human resources exchange program in scientific technology through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and future Planning (NRF-2016H1D2A2916091).

REFERENCES

Anonymous, 2014. OpenCV documentation ORB. OpenCV, Massachusetts, USA. http://docs.opencv.org/3.0-beta/.

Bank of Korea, 2016. Present situation of use of domestic internet banking services in the 3rd quarter of 2016. NTT DATA, Tokyo, Japan. http://www.nttdata.com/global/en/investor/library/ results-briefing/pdf/2016/fy2015_pre_3q_01.pdf.

Calonder, M., V. Lepetit, C . Strecha and P. Fua, 2010. Brief: Binary Robust Independent Elementary Features. In: Computer Vision-ECCV, Daniilidis, K., P. Maragos and N. Paragios (Eds.). Springer, Berlin, Germany, ISBN:978-3-642-15560-4, pp: 778-792.

Ji-Yong, S., C. Ji-Ho, H. Lee, K. Jeong-Min and G. Lee, 2016. A study on website forgery/falsification detection techniques using images. J. Convergence Secur., 16: 81-87.

Rosten, E. and T. Drummond, 2006. Machine learning for high-speed corner detection. Eur. Conf. Comput. Vision, 1: 430-443.

Rublee, E., V. Rabaud, K. Konolige and B . Gary, 2011. ORB: An efficient alternative to SIFT or SURF. Proceedings of the IEEE International Conference on Computer Vision (ICCV), November 6-13, 2011, IEEE, Barcelona, Spain, ISBN:978-1-4577-1101-5, pp: 2564-2571.

Yu, L., Z. Yu and Y. Gong, 2015. An improved ORB algorithm of extracting and matching features. Intl. J. Signal sProcess. Image Process. Pattern Recognit., 8: 117-126.