

Tamper Detection Based on Webpage Change Features

Hanji Shen^{ab}, Jun Li^a, Wei Wan^a, Chun Long^a, Jingdeng Zhou^a, Yuhao Fu^a, Xiaofan Song^a

a Computer Network Information Center, Chinese Academy of Sciences, China

b University of Chinese Academy of Sciences, China

shenhanji@cnic.cn, jlee@cstnet.cn, wanwei@cnic.cn, longchun@cnic.cn, zhoujingdeng@cnic.cn, fuyuhao@cnic.cn, songxiaofan@cnic.cn

Abstract—With the rapid development of the Network, the webpage tampering has become a problem that can not be ignored. In October 2019, for example, the number of websites tampered with in China reached more than 20,000, with the largest share coming from Beijing, Shandong and Guangdong respectively [1]. This shows that webpage tampering is no longer a mere simple problem, but requires greater attention. In order to find a better way to detect web page tampering, the website tamper-proof technology receives much attention in the area of web security. This paper* proposes a method of webpage tampering detection based on the webpage change features through analyzing the features of webpage changes and the illegal tampering purpose. Webpage changes will be determined before detecting. The detection model is decided by webpage change time, webpage change code amplitude, webpage change frequency, webpage change content location and webpage change content relationship. To be more specific, the detection process includes two training and detection stages, training phase and detection phase. In the training phase, the effective detection evidence only suitable for the webpage can be obtained through using the data set (multiple changes data in the webpage) to train model. In the detection phase, the detection model will detect the particular webpage according to the detection evidence, and then gives the detection result. If the result is misdeclaration, the detection evidence will be retrained. Furthermore, the simulation tampering experiment is designed to verify the feasibility of the detection method. And the optimal number of the experimental webpage changes firstly into the training phase is determined according to the accuracy and recall rate of the test results. After Verifying, the accuracy and recall of the test results were 98.32% and 99.12% respectively. The best number of changes was 55 and the risk value was 1.

Keywords—Webpage detection model, webpage tamper detection, web features

I. INTRODUCTION

The web as a new medium is accessed and viewed by hundreds of millions of people all the time.

However, network is a double-edged sword, it brings people opportunities and challenges, website tampering is one of the problems. Webpage tampering means that those web

pages accessed are not real or even wrong, which can bring about minor disruptions to browsing and serious damage to property. In this context, manipulated or tampered pictures are very common as part of informing articles, in today's misinformation crisis. [2].

Webpage tampering resulting in inaccurate or, in some cases, misleading information for visitors. At present, the method of tampering was more and more complex. Some webpages were modified directly by intrusion sites, and some webpages were inserted illegal html and JavaScript code through cross site Scripting or other vulnerabilities. The basic reason for the tampering webpage is the vulnerabilities in the websites. If the site is not vulnerable, the page will not be tampered with. But it is hard to ensure that every website has no vulnerability in the network. In other words, it is very difficult to prevent webpages from being tampered with. Therefore, the anti-tampering technology focus on tampered pages rapid detection.

In this paper, a method for tampering detection based on webpage change features is introduced. The method focuses on the page change itself, and it mainly performs tampering detection on a webpage by using a detection model based on some webpage change features. Its detection process includes training and detection phase. In the training phase, the detection model trains the data that changes many times in each webpage as the training set, then these sets will be used for the effective detection evidence, which is only suitable for the webpage. In the detection phase, the detection model will detect the particular webpage according to the detection evidence, and then the detection result can be obtained. If the result is misdeclaration, the detection evidence will be retrained.

The rest of this paper is organized as follows: The origin of the detection model is introduced in Section 2. In Section 3, the method for tampering detection based on webpage change features is proposed. And experiment analysis and conclusion are in Section 4 and Section 5 respectively.

II. DETECTION MODEL

A. Webpage Change Features

At present, the tamper detection method is used to detect whether a webpage is tampered with according to different

* Corresponding author: jlee@cstnet.cn.

Supported by National Natural Science Foundation of China under grant No.61672490.

features before and after the webpage change. However, in fact, the features that webpage changes themselves can be used for detection. Webpage changes can usually be divided into two categories, one is caused by updates, the other is caused by illegal tampering. The update of the webpage is seems that a new website vision publishing, so the webpage content is different from the previous one. But illegal tampering means that the act of hackers maliciously modifying the inner pages of a website in certain ways and is a form of cyber attack. It is true that web tampering is a shallow form of attack compared to other attacks, but the lack of website security has made it a common means of attacking various websites. [3]. Because of the particular webpage function, when the webpage is updating, these webpage functions should be used normally. Therefore, the webpage features changes is regular, and a limited part of the webpage content can be edited in the update time. But the changes caused by illegal tampering are irregular, any part of webpages can be modified at any time, and then the content that attackers modified will appear. Therefore, the difference between the two changes can be fully utilized as webpage change features. Considering the detection speed of each webpage, the webpage change features selected mainly include webpage change time, webpage change code amplitude, webpage change frequency, webpage change content location, and webpage change content relationship.

B. Detection Model

A detection model is designed according to the webpage change features. It can classify webpage changes by webpage change features so that it can detect whether a changed webpage tampered. The main structure of this model is shown in Figure 1. The model will sequentially judge the change time of the input information, the change code amplitude of the input information, the change frequency of the input information, the change content location of the input information, and the change content relationship of input information to determine whether a webpage tampered.

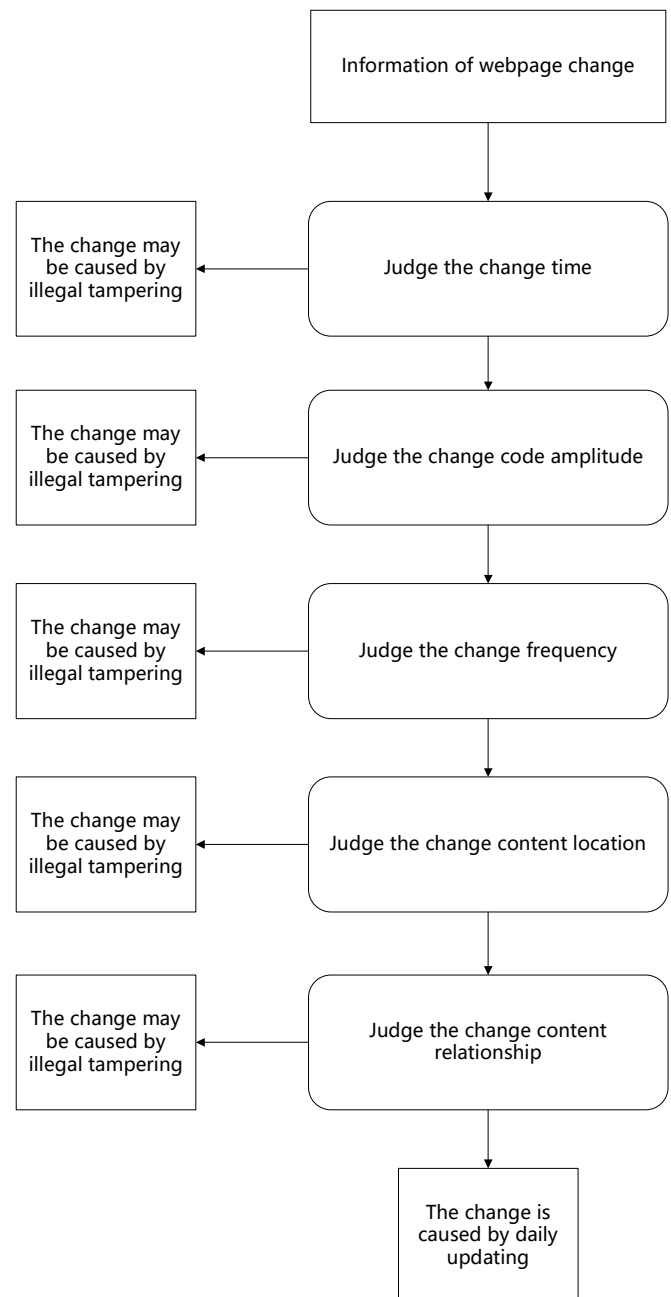


Figure 1. The structure of the detection model

III. TAMPER DETECTION METHOD

The structure of tampering detection based on webpage change features is shown in Figure 2. It contains six modules, which are the crawler module, the judgement module, the data processing module, the training module, the detection module, and the alarm module[4][5]. During the detection process, there are two phases, the training phase in the black box and the detection phase in the red box.

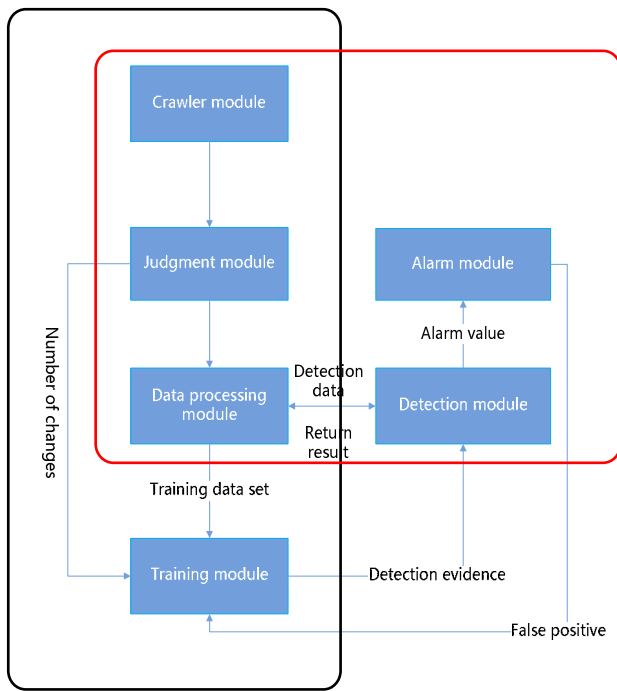


Figure 2. The structure of tampering detection

A. Training Phase

The crawler module gains the webpage information and transports it to the judgment module. Then the judgment module will detect the webpage changes according to the hash value of the webpage source code, the hash value of all the images in webpage, the hash value of the JavaScript file in the webpage, and the hash value of the CSS file in the webpage. And transport data to the data processing module after detecting the webpage change. The data processing module will calculate the webpage change time, the webpage change code amplitude value, the webpage change frequency value, the webpage change content location value, and the webpage change content relationship value based on the data. The specific process is as follows:

- Get the current time as the webpage change time.
- Get the difference between the number of lines before and after the change as the page change code amplitude value.
- Get the difference between the change time of the webpage and the last change time of the webpage as the webpage change frequency value.
- Detect changes of source code, images, JavaScript files, and CSS files in the webpage. Take 1 for change, 0 for no change. Compose a four-digit binary number, and convert it to a decimal number. The decimal number is the webpage change content location value.
- Detect the change between source code and JavaScript files, source code, and CSS files, image and picture URL, text content and the rest of the URL at the same time. Take 1 for yes, 0 for no. Compose a four-digit binary number, and convert it to a decimal number. The decimal number is the webpage change content relationship value.

Start the training module when the number of changes in the judgement module reaches a certain value. At this point,

the data processing module will transport the training data set to the training module. And it will be called to train the model described in Chapter 2. The training process is simple. It uses a supervised machine learning method. During training, a range is set in advance for each change feature, and then the detection model uses the range to detect the data set and confirm the detection result. Then the training module can narrow this range and detect again until a minimum range or value that guarantees all test results are correct. The training module will transport the range or value as to detection evidence to the detection module after training is finished [6].

B. Detection Phase

In the detection phase, the functions of the crawler module, the judgment module, and the data processing module are the same as in the training phase. But in the phase, the data processing module will transport data that will be detected to the detection module [7] [8]. And the detection model, wherein the detection module will use detection evidence to detect it. In addition to the detection model, an alarm is also provided in the detection module. If the result of one of the detectors of the detection module is “may be caused by illegally tampering”, the alarm will be incremented by one. The detection module will transport alarm value to an alarm module. The alarm module will decide whether to alarm according to a risk value. Finally, if the result is a false positive, the alarm module will return the result to the training module and make it restart.

IV. EXPERIMENT AND ANALYSIS

A. Simulation Tampering Experiment

To verify the above-mentioned webpage tampering detection method, established a library where the data collected from a large number of real tampering cases in China Science Technology Network (CSTNET), and based on the library designed a simulation tamper experiment. Except that it is not illegal, the tampering behaviour in the experiment had all the characteristics of the real tampering cases. The experimental results were based on the accuracy and recall rate of the detection module in 100 tests. And in the process of the experiment, only the number of changes in the judgement module and the risk value of the detection module would affect the detection result.

B. Set the Number of Change

The detection results of different number of changes are shown in Figure 3 When the risk value is 1.

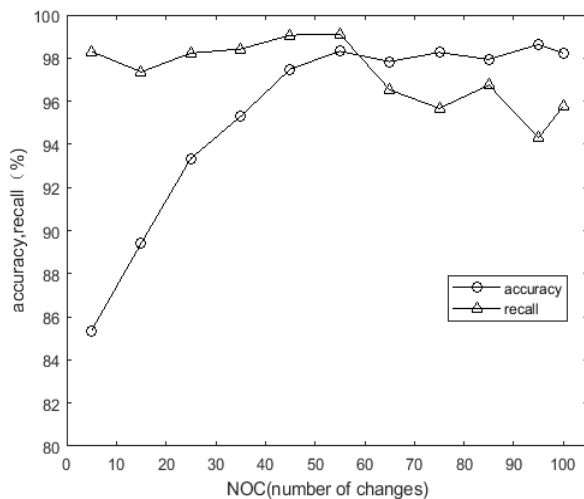


Figure 3

In the figure, the number of change is indicated by NOC. The figure shows the change in accuracy and recall rate for different number of changes. Accuracy and recall rate gradually rise when $1 \leq \text{NOC} \leq 55$. And the accuracy rate is stable and the recall rate is gradually decreasing when $55 < \text{NOC} \leq 100$. Therefore, the best number of changes is 55.

C. Select the Risk Value

Risk value selection range is 1 to 5. The detection results of different risk value are shown in Figure 4 When the number of change is 55.

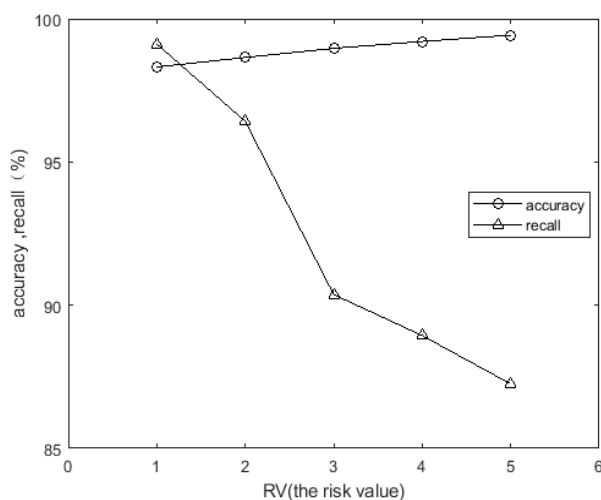


Figure 4

In the figure, the risk value is indicated by RV. The figure shows the change in accuracy and recall rate for different risk value. The accuracy gradually increasing and the recall rate gradually decreasing when $\text{RV} > 1$. Taking into account the detection time, so select 1 as the risk value finally.

V. CONCLUSION

This paper has proposed a method of tampering detection base on webpage change features. It only tampers detection of

changed webpages. It can get the webpage information through the webpage URL that inputted. Then judge whether the webpage changes according to this information. And calculate the relevant value after judging the webpage is changed. Next, gain the detection evidence by using the historical value of the webpage changes to training the model that was built in advance. Finally, tampering detection of the webpage based on the detection evidence and alarm confirmation. If the result of the detection is misdeclaration, the detection evidence will be retrained. The method can get the detection evidence through a training detection model for each webpage that will be detected. And experimentally verified that the accuracy and recall of the detection results were 98.32% and 99.12% respectively when the number of changes was 55 and the risk value was 1.

REFERENCES

- [1] National Computer Network Emergency Response Technical Processing Coordination Centre, October 2019 Cybersecurity Monitoring Data, 2019.
- [2] Anastasia N. Katsaounidou, et al, News authentication and tampered images: evaluating the photo-truth impact through image verification algorithms, 2020:6-12.
- [3] Liu Jiang. Design and Implementation of Webpage tampering monitoring System, 2018.3.
- [4] Lai Qingnan, Chen Shiyang, Ma Hao, Zhang Bei. A Method of Batch Webpage Tampering Detection Based on Machine Learning[J]. Journal of Huazhong University of Science and Technology(Natural Science Edition), 2016, 44(11): 16-20.
- [5] Wang wenyi, Design and Implementaton of Network Tamper Detecton System, Southwest Jiaotong University, 2019.08.
- [6] Li keyan, Liu xiaodong, Web Tampering Detection Dystem Based on Feature Recognition, Electronic Design Engineering, 2020.9.
- [7] Wei xu, Cheng weiqing, The Training Module Will Transport The Range Or Value As To Detection Evidence to the Detection Module After Training is Finished, Journal of Nanjing University of Posts and Telecommunications (Natural Science Edition), 2019.4
- [8] Mukherjee A, Dey G, Dey M, et al. Web-based intelligent EEG signal authentication and tamper detection system for secure telemonitoring[M]//Brain-Computer Interfaces. Springer, Cham, 2015: 295-312.



Hanji Shen is currently a full engineer in Computer Network Information Center, Chinese Academy of Sciences, China. He received M.S. degree in Engineering in the field of Computer Technology from University of Chinese Academy of Sciences, and he is currently studying for a Ph.D. in University of Chinese Academy of Sciences, majoring in Computer Software and Theory. He has been an IEEE member since 2014. His research interest is Cyber Security.



Jun Li is currently a researcher, deputy chief engineer and doctoral supervisor in Computer Network Information Center, Chinese Academy of Sciences, China. He received Ph.D. degree in Computer System Structure from University of Chinese Academy of Sciences. He is one of the earliest experts engaged in computer network technology research in China. His research interests include Network Architecture and Cyber Security.



Wei Wan is currently a full engineer in Computer Network Information Center, Chinese Academy of Sciences, China. He received Ph.D. degree in Computer Software and Theory from University of Chinese Academy of Sciences. He has presided over and participated in many national projects. Moreover, he developed Investigation of state division in botnet detection model, Botnet detecting method based on activity similarity and so on. His research interests include Information Security, Cyber Security, Cyber Risks and Web Vulnerabilities.



Chun Long is currently a full engineer in Computer Network Information Center, Chinese Academy of Sciences, China. He received Ph.D. degree in Computer Software and Theory from University of Chinese Academy of Sciences. He developed Compound Attack Prediction Method based on the Attack Graph and Multi-source Security Event Fusion method based on EA-DS Evidence Theory. His research interests include Information Security, Cyber Security, Cyber Risks and Web vulnerabilities.



Jingdeng Zhou is currently a full engineer in Computer Network Information Center, Chinese Academy of Sciences, China. He received B.S. degree in Network Engineering from Jinggangshan University. His research interests include Network Security Guarantee Technology and Cyber Space Security.



Yuhao Fu is currently a full engineer in Computer Network Information Center, Chinese Academy of Sciences, China. He received B.S. degree in Computer Science and Technology from Chongqing University of Post and Telecommunications. His research interests include Network Security Guarantee Technology and Cyber Security.



Xiaofan Song is currently a full engineer in Computer Network Information Center, Chinese Academy of Sciences, China. She received M.S. degree in Cyber Security from University of Southampton. Her research interests include Information Security, Cyber Security, and Cyber Risks.