Research on Web Application Vulnerability Scanning System based on Fingerprint Feature

Hao He [a], Lulu Chen and Wenpu Guo
Xi'an High-Tech research Institute, Xi'an 710025, China
[a]460827753@qq.com

Keywords: Web Application; Fingerprint Feature; Wappalyzer; KMN Algorithm; Vulnerability Scanning.

Abstract. In view of the existing network vulnerability scanning system does not distinguish between the characteristics of web applications, but in accordance with a single strategy to scan, resulting in scanning efficiency and accuracy is low, the scan results are not comprehensive and so on. In this paper, we propose a web application vulnerability scanning system based on fingerprint feature, in which Wappalyzer technology is used to extract the fingerprint feature of web application. Then, we use the Euclidean distance KMN algorithm to match the fingerprint data in the fingerprint database. Finally, through the comparative analysis of the experiments, the efficiency of the system is improved by 3% and the precision is improved by 6% in the process of vulnerability scanning for 1000 websites. The results show that the performance of the system is improved by 6%, which proves the superiority of the system.

1. Introduction

The concept of network vulnerabilities was first proposed by the US Department of Defense and Harvard University's military network communications project, when the US Department of Defense in order to ensure the safety of building networks [2], put forward the establishment of a dedicated network software testing tools, And then the United States Harvard University and Stanford University, under the joint research, launched a series of network vulnerability scanning tools and network vulnerabilities for scanning application script library [3], the vulnerability of the network security vulnerability, , These tools and libraries in the late development of the entire field of network security has played a very important role, such as wapiti and sqlmp are based on the prototype of these technologies developed. To the current vulnerability scanning in the network, there have been a lot of available tools and technologies, such as Discuz, wordPress, etc. These tools can help webmasters quickly build the relevant vulnerability detection environment [4], to achieve the loopholes in the site. But these vulnerability scanning technology is basically the use of network crawler network data collection, and then scan the results of the implementation of the final strategy will show the output of this model, and with the rapid development of Internet applications, network technology and application of the difference between the more , The vulnerability of a single strategy model [5], has been very difficult to fully detect web application vulnerabilities, even in many applications using new technology, these vulnerability scanning tools are no longer adapt, which is no longer adapt to this Scanning tools need to be able to identify the specific application types, and according to the type of application constantly changing scanning detection strategy to achieve more efficient and accurate scanning, a comprehensive detection of web applications vulnerability information [6].

To solve this problem, this paper proposes combining web fingerprinting technology with traditional web application vulnerability scanning technology to construct web application fingerprint data by using web fingerprinting technology and then matching with existing web application fingerprint database, Quickly find the type of web application, and then according to different web applications to take different scanning strategies for vulnerability scanning to improve web application vulnerability scanning efficiency and accuracy.

2. Fingerprint feature-based web application vulnerability scanning system design

2.1 Design Scheme of Web Application Vulnerability Scanning System

According to the analysis of the existing web application vulnerability scanning tool system, the current majority of web application vulnerability scanning system use a single strategy rather than base on the characteristics of web applications to take a specific way to detect web application vulnerabilities, which led to in the current diversified web application environment, the system can not be efficient and comprehensive for web application scan. Based on the existing web vulnerability scanning system, this paper introduces the web application fingerprint identification and matching technology. Before the web application vulnerabilities scanning, the scanning system uses the remote method to obtain the site fingerprint characteristics, and the use of web fingerprint feature matching technology to achieve the type of site application identification,.Then according to the identification of the type of application, using the corresponding vulnerability plug-in the application of the site scanning, the vulnerability plug-ins can be specific The application of different types of strategies to carry out vulnerability scanning, to avoid the blind use of a single strategy mode, resulting in low scanning efficiency, inaccurate and incomplete scanning and other issues, the specific system design shown in Figure 1 below:

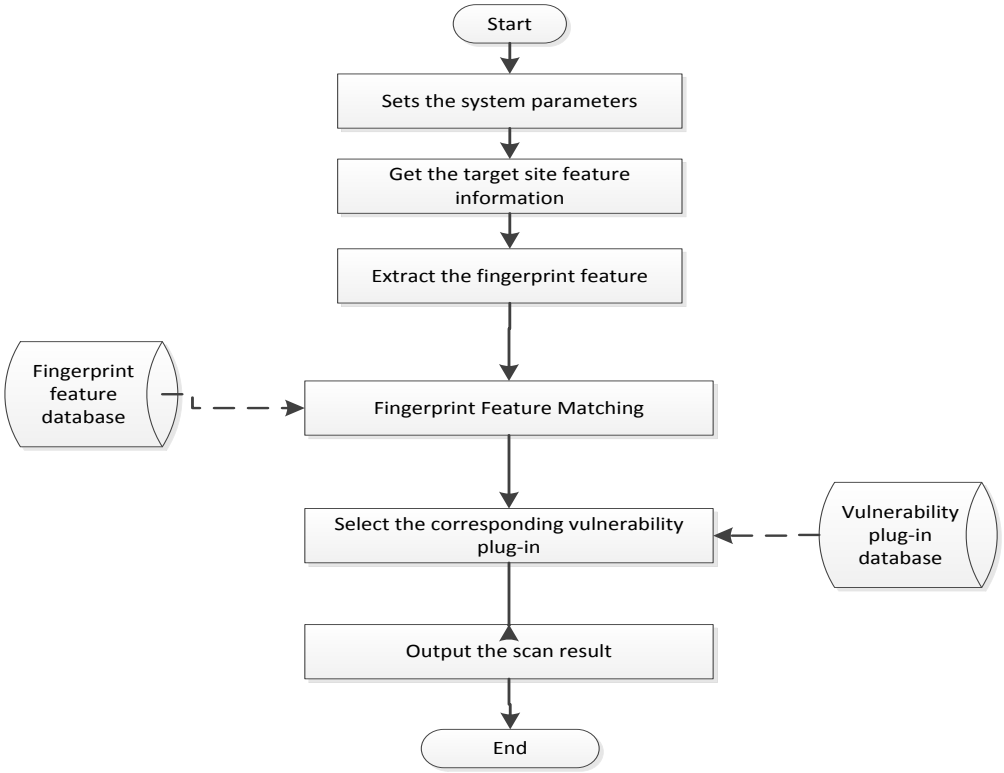

Figure 1. Vulnerability scanning system design based on fingerprint characteristics of the web application

In the above implementation, the configuration file is first formed in the system according to the parameters set by the user, and then the system starts to acquire the characteristic information of the target site according to the configuration information, including the keywords in the web application page, the special files Name, the hash value of the static file, the file tag tree eigenvalues, and then from the target site to obtain these features, use wappalyzer provide API processing to generate fingerprint information. With the fingerprint information of the target site, you can read from the local fingerprint feature database to match the relevant fingerprint information to find the corresponding fingerprint information corresponding to the application type, where the matching algorithm used is KMN adjacent Matching algorithm to find the most similar characteristics of the fingerprint corresponding to the type of web application, the system can be based on the type of application, the application for the theft of vulnerability scanning plug-in, targeted scanning to get a higher scan Efficiency and accuracy, to achieve a more comprehensive scan test.

2.2 web application fingerprint acquisition method design

In this paper, based on fingerprint characteristics of the web application vulnerability scanning system, fingerprint access method is one of the core of this design, how to obtain can be used to identify web application types of accurate fingerprint information, and to ensure that the matching process In the design process, the author selects the keyword, special file name, hash value of static file, and the characteristics of the file label tree in the web application page of the site through the synthetical analysis in the process of design. In this paper, the key of fingerprint acquisition method is designed, Value as the basic data of the fingerprint feature, and then use the setDataPrame () and startHashPro () functions provided by the Wappalyzer to finish the configuration of the fingerprint algorithm and generate the fingerprint data. The generated fingerprint eigenvalues can be used as the application fingerprint Which is used to match the fingerprint feature value in the fingerprint database and realize the identification of the application type. The specific implementation flow is shown in Figure 2 below:

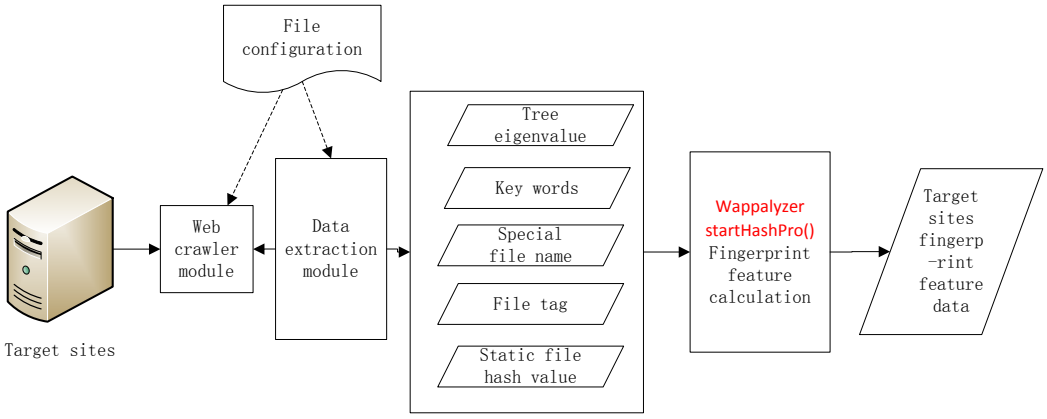

Figure 2. Web application fingerprint feature acquisition process

In the above method, the web crawler used in this paper is an open-source GooSeeker web crawler, which can automatically search the content information of the site according to the defined configuration file. In the search information, this paper mainly extracts the key of the site according to the pre-configured keywords. And then search the site of the tree characteristics of information and file labels, the hash of static files, and finally get the site's keywords, special file name, static file hash value, the file tag tree feature value as a fingerprint feature .The data of the target site can be calculated by calling the setDataPrame () and startHashPro () functions provided by the wappalyzer directly after the data are acquired, which can be used for the subsequent fingerprint feature matching. Specific implementation process and the core code is as follows:

First write the fingerprint rules json file is as follows:

Moretanjiti.json "apps": {

"Discuz!":{

"website": "www.discuz.net/forum.php",

"cats": [1],

"meta": { "generator": "Discuz"},

"headers": {"Set-Cookie": "_lastact.*_sid|_sid.*_lastact|_sid.*smile|smile.*_sid"},

"url": "/uc_server[/$]|uc_client[/$]",

"html": "Powered by (?:Discuz!|<a href=\"http://www\\.discuz\\.net/\"|UCenter)",

……

} }

Then use setDataPrame() and startHashPro()

SetDataPrame(words, flab, cms,cdn,sf, moretanjiti.json)

Start HashPro ()

2.3 Web application fingerprint feature matching algorithm

After extracting the fingerprint data from the target host to generate the fingerprint feature, we need to further determine the application type on the web server, which needs to match the fingerprint

database of the local web application fingerprint to find the most similar fingerprint data , And calculate the fingerprint similarity, if the similarity is greater than the preset threshold parameter, the fingerprint is considered as the corresponding application type in the fingerprint database, otherwise the fingerprint data does not exist in the fingerprint library, and the user needs to select and configure The fingerprint data into the fingerprint database, this time the user needs from the target server to obtain fingerprint-related data for manual analysis, and ultimately determine the use of the scanning program, after joining the user can perform the vulnerability scanning program on the target server scan , The realization of the process shown in Figure 3 below:

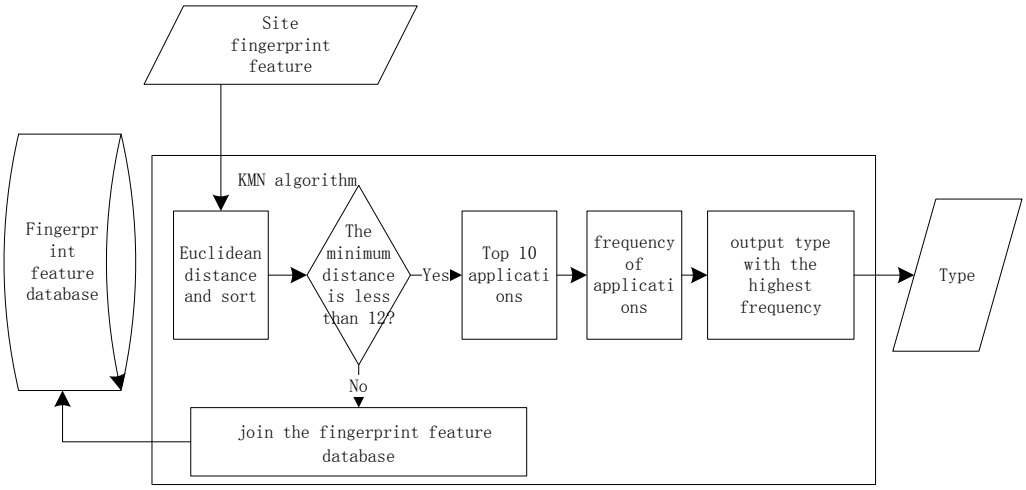

Figure 3. Web application fingerprint feature matching implementation process

In the above process, the core idea is to use KMN proximity algorithm to realize the application of fingerprint feature information to determine the type of application in the fingerprint feature database, there are different types of applications, fingerprint features, KMN algorithm module from the database read fingerprint characteristics and And then calculate the distance between the fingerprint feature of the target site and the fingerprint characteristics of the relevant application. After the calculation, the 12-bit threshold is used to discard the label and the corresponding fingerprint characteristic data, and then extract the first 10 items. The most frequently used application tag is output as the application type, and the corresponding application type of the fingerprint can be determined by the tag, and the KNN algorithm can be quickly matched to the corresponding fingerprint information from the database, and the fingerprint information of the small difference can not be matched.We will start a new configuration process, that the database does not have the application of fingerprint features, by configuring the new application information will be added to the fingerprint characteristics of the site to the fingerprints of the site, Feature database, and the establishment of the database with the mapping procedures, the next scan can automatically match the time to call the corresponding scan program.

In the process of matching KMN algorithm, the mathematical model of Euclidean distance in which the eigenvalues are calculated is as follows:

$$D = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \tag{1}$$

The eigenvalues of the network applications obtained in this paper are a four-dimensional eigenvector constructed with keywords, special file names, hash values of static files, and eigenvalues of the file tag tree (words, flab, cms, sf). According to the above- Formula, which is calculated as follows:

$$D = \sqrt{(words_1 - words_0)^2 + (\mathrm{flab}_1 - \mathrm{flab}_0)^2 + (\mathrm{cms}_1 - \mathrm{cms}_0)^2 + +(\mathrm{sf}_2 - \mathrm{sf}_0)^2} \tag{2}$$

Based on the above principles to achieve the core code is as follows：

KMN algorithm

```
map<int, double> kmins;
// Look for the last 10 samples from the fingerprint in the database
for(int i = 0; i < nSamples; i++)
{
double dist = getDistance(data, DataSet[i]);
if(kmins.size() <10)
kmins.insert(map<int, double>::value_type(i, dist));
else
{
// Point to the current position in kmins that corresponds to the largest element in the distance
map<int, double>::iterator max = kmins.begin();
for(map<int, double>::iterator it = kmins.begin(); it != kmins.end(); it++)
…….
if(dist < max->second)
……
}
}
// The number of these 10 samples in the classification
map<double, int> votes;
for(map<int, double>::iterator it = kmins.begin(); it != kmins.end(); it++)
{
double tmp = DataSet[it->first][nFeatures - 1];
map<double, int>::iterator voteIt = votes.find(tmp);
if(voteIt != votes.end())
voteIt->second++;
else
votes.insert(map<double, int>::value_type(tmp, 1));
}

// Calculate the class to which the 10 samples belong
map<double, int>::iterator maxVote = votes.begin();
for(map<double, int>::iterator it = votes.begin(); it != votes.end(); it++)
{
if(it->second > maxVote->second)
maxVote = it;
}
data[nFeatures - 1] = maxVote->first;
return maxVote->first;
}
```

3. Experimental Simulation Analysis

In order to verify the correctness and feasibility of the web application vulnerability scanning system based on fingerprint feature, this paper uses open source wappalyzer platform and Stanford University's public web application vulnerability scanning plug-in library to complete the system building, and carries on the experimental analysis, In the experiment, we use the fingerprint extraction algorithm and fingerprint matching algorithm provided by the wappalyzer platform to get the fingerprint data from the target site, then get the fingerprint feature of the target site, and then match, according to the matching results from Stanford University web The application vulnerability

scanning program library called the corresponding web application vulnerability plug-in to perform vulnerability scanning, and finally the formation of each web site vulnerability report, the higher education institutions in China website as an example, select 1000 educational sites for analysis, and at the same time The average number of web vulnerabilities and the average time spent in performing the vulnerability scanning analysis are shown in Table 1. Table 1 shows the average number of web vulnerabilities in each site, and the average time spent in performing vulnerability scanning analysis.

Table 1. Experimental results

Statistical parameters	Discuz	WordPress	This article system
Average number of vulnerabilities (unit)	12.3	9.4	13.1
Vulnerability variance	2.4	3.6	1.87
Time-averaged (s)	11.23	14.98	10.87
Time Variance	3.1	4.3	2.9

It can be seen from the above results that the web application vulnerability scanning system based on fingerprint features has great advantages in exploiting the average number of vulnerabilities and exploiting the average exploit time compared with the traditional Discuz and WordPress platform, in which the number of vulnerabilities The average increase of 6.5% and 39%, with time than discuz and wordPress platform, respectively, increased by 3% and 29%, from the number of loopholes and the use of variance calculation point of view, the system design is more stable and better stability.

4. Conclusion

Ensuring the security of network application is an important direction in the future development of network security, this paper combines the traditional network vulnerability scanning technology and web fingerprinting technology to realize the design of vulnerability scanning method with different mining strategy for the specified web application. The experimental results show that the system designed in this paper has a great improvement in efficiency and precision, and it is very important to improve the efficiency and accuracy of web application exploit.

References

[1] Wang Chendong, Guo Yuobo, Huang Wei.Non-invasive network security scanning technology research [J]. Information security and communication security, 2016,09: 67-72 +76.
[2] Chen Yingcong, Chen Guangqing, Chen Zhiming, Wan Neng. Research on Binary Code Analysis Vulnerability Scanning Method for Smart Grid SDN [J] .Security of Information Network, 2016, 07: 35-39.
[3] Bai Yuanyuan.For VMware vulnerability detection model design and implementation [D]. Beijing Jiaotong University, 2016.
[4] Zhao Xing. Web vulnerability mining and security research [D]. North University, 2016.
[5] Huang Haijun, Wang Yang, Li Qiang, Yu Xiang.Vulnerability Assessment of Command Information System Based on Vulnerability Scanning [A]. Chinese Society of Command and Control. [C] And Control Society: 2016: 4.
[6] Yao Bingying. Fingerprint identification technology in WEB cloud storage security certification application [D]. Guangdong University of Technology, 2014.
[7] Xu Jingbo, Chen Taowei. Dual protection of Web pages by digital fingerprints [J]. Journal of Donghua University (Natural Science Edition), 2006,01: 116-119.