applied sciences

MDPI

Article

A Study on the Measuring Methods of Website Security Risk Rate

Yong-Joon Lee

Department of Hacking Security, Far East University (FEU), 76-32 Daehak-gil, Gamgok-myeon, Eumseong-gun 27601, Republic of Korea; 2020032@kdu.ac.kr

Abstract: Traditionally, website security risks are measured using static analysis based on patterns and dynamic analysis by accessing websites with user devices. Recently, similarity hash-based website security risk analysis and machine learning-based website security risk analysis methods have been proposed. In this study, we propose a technique to measure website risk by collecting public information on the Internet. Publicly available DNS information, IP information, and website reputation information were used to measure security risk. Website reputation information includes global traffic rankings, malware distribution history, and HTTP access status. In this study, we collected public information on a total of 2000 websites, including 1000 legitimate domains and 1000 malicious domains, to assess their security risk. We evaluated 11 categories of public information collected by the Korea Internet & Security Agency, an international domain registrar. Through this study, public information about websites can be collected and used to measure website security risk.

Keywords: website security risk; website security measurements; OSINT (open source intelligence); verifying DNS information; verifying IP information

check for updates

Citation: Lee, Y.-J. A Study on the Measuring Methods of Website Security Risk Rate. *Appl. Sci.* **2024**, *14*, 42. https://doi.org/10.3390/app14010042

Academic Editors: Bhanu Shrestha and Seongsoo Cho

Received: 4 August 2023
Revised: 11 September 2023
Accepted: 14 September 2023
Published: 20 December 2023

Copyright: © 2023 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Recently, cases of personal information infringement accidents and DDoS attacks that steal personal information after infecting users' PCs by distributing malicious codes to Internet users are increasing. Hackers have been exploiting security vulnerabilities such as cross-site scripts, SQL injections, and security configuration errors to steal personal information.

Among the various channels for distributing malicious code, the number of cases of distributing malicious code using vulnerabilities in websites and user PCs continues to increase. The homepage, which can infect a user's PC with malicious code, is divided into a distribution site that directly hides the malicious code and a hidden stopover so that it can be automatically connected to the distribution site. After opening a malicious code distribution site, the attacker hacks the website of the destination and inserts the distribution site URL to infect the malicious code by inducing the visitor to access the destination to the distribution site without knowing. Detection of stopovers is becoming increasingly difficult because they mainly target websites with high user visits, such as portals, blogs, and bulletin boards, and obfuscate and conceal malicious code within the stopover hacked by the attacker [1]. As a result, there is an increasing demand for measuring the security threat of websites. Static and dynamic analysis of website security risks is generally used to analyze website security risks [2]. Similar hash-based website security risk analysis and machine learning-based website security measurement methods have recently been proposed [3]. Additionally, website security measurement technology mainly aims at inspection speed, zero-day malicious code analysis possibility, and accuracy of analysis results.

In this study, a method was proposed to collect and measure information disclosed on the Internet to measure the risk of a website. DNS information, IP information, and website history information are required to measure the risk of a website. The website

history information can be checked for global traffic rankings, malicious code distribution history, and HTTP access status. In addition, public information collection on a website can be used to measure the security risk of a website. Existing cybersecurity intelligence analysis utilizes commercial services from antivirus companies. However, in this study, we conducted a study to measure cyberthreats by collecting only OSINT information for public purposes.

2. Related Research

The main method of measuring website security threats is reputation research based on the records of previous security incidents' logs. In addition, it is classified into dynamic checks that check malware by accessing the user's terminal environment and static checks that check malware detection patterns.

The main goals of website security measurement technology are inspection speed, zero-day malicious code analysis possibility, and accuracy of analysis results.

2.1. Static Analysis of Website Security Risks

There is a static analysis method for source code to check malicious code hidden on the website. Since the malicious code hidden on the website is obfuscated, it is decoded and checked [4]. As shown in Figure 1, the inspection method converts malicious codes previously distributed on the website into signatures and registers them as detection patterns. Then, the website to be analyzed is collected by crawling, and the source is checked on the website with the registered signature. The characteristic of this analysis is that it has a faster inspection time than other analysis methods, so it is efficient in detecting a large number of inspection targets in a short time [5].

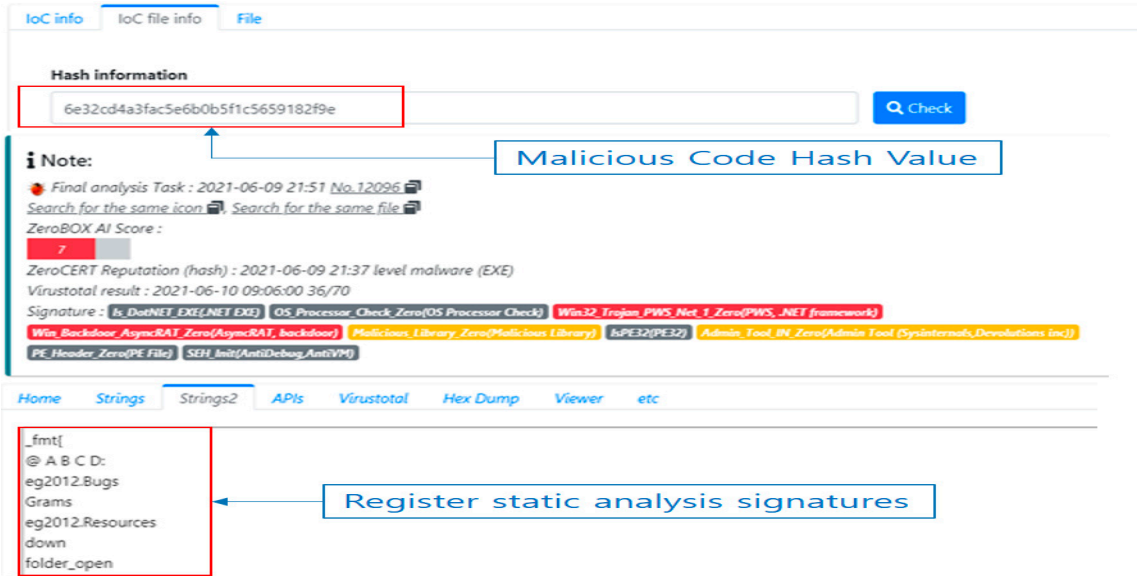


Figure 1. Register website security threat static analysis pattern.

2.2. Dynamic Analysis of Website Security Risk

Dynamic analysis is a method of analyzing behavior on a PC by executing malicious code and analyzing it on the same virtual machine as the user's PC. For detection, one must access the website and download malicious code [6]. This detects abnormal behavior such as creating additional files, changing the registry, and attempting to connect to a network on a virtual PC. As shown in (Figure 2), it is possible to collect malicious code downloaded from the website and information on command control servers (C & C, command and control). Dynamic analysis has the advantage of analyzing zero-day malicious codes that are difficult to detect with static analysis. However, the inspection speed takes a long time [7]. "Accept: */*" accommodates all types of files on the communication.

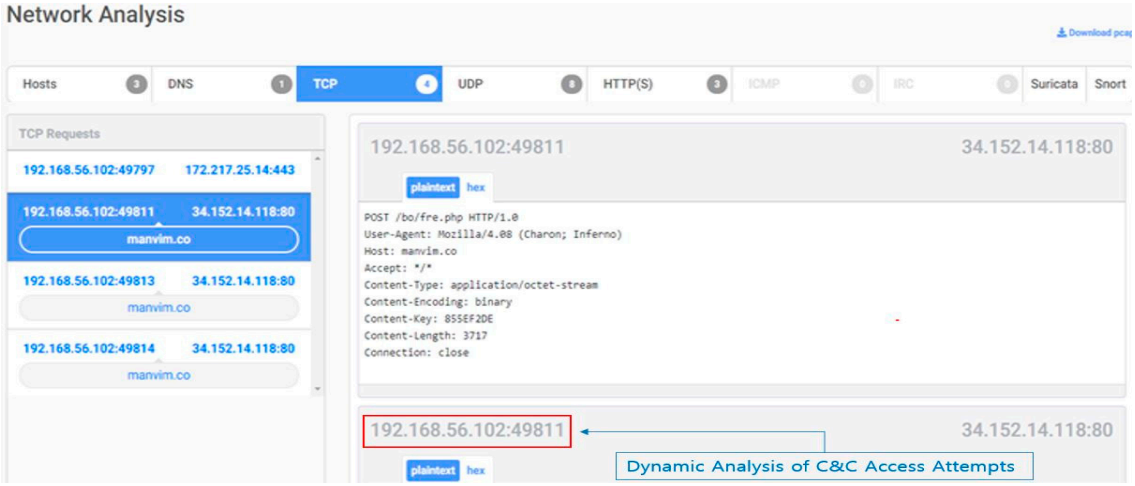


Figure 2. Dynamic analysis of security threats on website.

2.3. Analysis of Website Security Risk Based on Similarity Hash

A similarity hash method may be applied between hidden malicious codes of a website. As shown in (Figure 3), generating hash codes for malicious codes with similar hash values can compare the degree of similarity of files by deriving similarity values such as Euclidean distance measurement [8]. Similarity hash functions are applied to many malicious codes, and the K-NN method is used to classify whether they are malicious or not. K-NN classifies malicious code types classified as K into the closest type by applying similarity hash codes [9]. Blue dots are classified as normal files and red dots are classified as malicious files.



Figure 3. Security threat analysis of websites based on similarity hash.

2.4. Machine Learning (ML)-Based Website Security Threat Analysis

Attackers are evolving by continuously seeking ways to avoid detection techniques because they are aware of the website's security threat detection system Security experts have limitations in defending evolving attackers to evade existing detection systems [10]. Accordingly, to detect security threats, many websites are responding through machine learning (ML).

The detection system using ML learns the executable file for the PE (portable executable) structure in the form of EXE and DLL [11]. The PE structure is the file structure of the executable downloaded from the website, which can also be checked for similarities to malware. Additionally, it learns nonexecutable (Non-PE) structures such as standard document files and scripts. As shown in (Figure 4), this analysis checks the similarity of files hidden on the website to be inspected for maliciousness. For example, when checking the website IP registered on Google Virustotal, it can be seen that more than 30 phishing domains are registered [12].

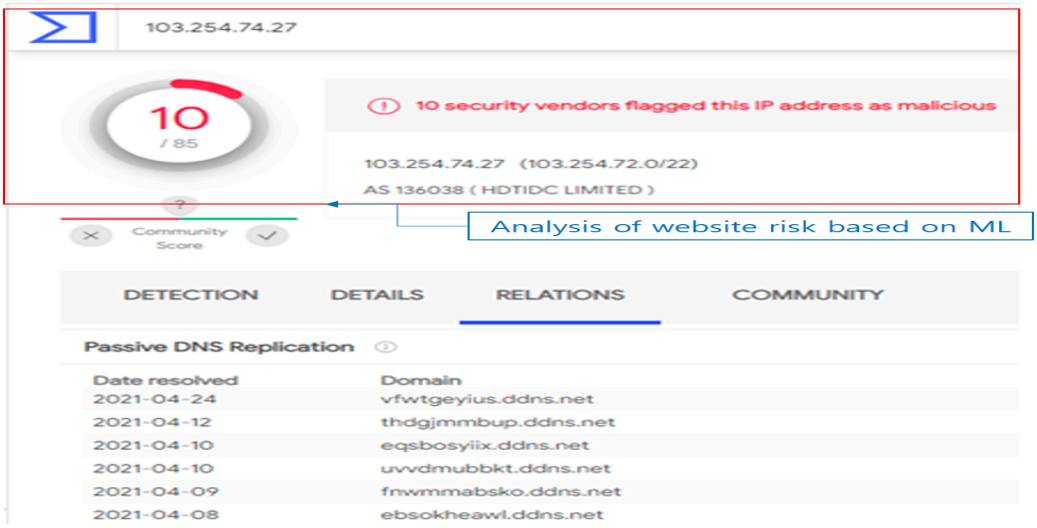


Figure 4. Checking website security risk using ML.

In the same way, most of the methods for measuring security threats on a website are building a DB by crawling website source files and collecting executable files distributed on the website, which are presented in (Table 1). Items that measure website risk were analyzed as inspection method, inspection technology, inspection speed, zero-day detection, and Accuracy. The inspection method refers to the website malicious code detection method, and the inspection technology refers to the information collection method. Inspection speed is the detection speed, zero-day detection is the detection of unknown malicious code, and Accuracy refers to the accuracy. The static analysis method registers malicious code patterns and then collects and checks website sources through web crawling. It has the advantage of being fast because it checks patterns. However, it has the disadvantage of being unable to detect nonregistered zero-day malware. In addition, the accuracy is very good with registered pattern searches. Dynamic analysis connects the website to the virtualized PC environment to determine whether it is dangerous through abnormal behavior analysis. The inspection technology uses a virtualized PC, which takes some time for risk behavior analysis. It can detect even zero-day malicious codes through behavioral analysis and has excellent accuracy. Similarity hash analysis technology configures a dataset for each type of malicious code to determine similarity hash values and similarity to files collected from the website. The inspection technology is a similarity hash technology for malicious codes, and the inspection speed is high. Zero-day detection is also possible, but the accuracy is moderate. ML learning analysis constructs malicious code and standard file datasets to predict the risk of files collected on the website after performing ML learning. The inspection technology is ML learning for malicious and standard code, and the inspection speed is high. Zero-day detection is possible, and accuracy is excellent, depending on the amount of learning.

Table 1. Comparison of analysis techniques on website security risk.

Comparison Item	Static Analysis	Dynamic Analysis	Similarity Hash Analysis	ML Analysis
Inspection method	Check website sources by registering malicious code patterns	Access websites from virtualized PCs to analyze risk behavior	Check the similarity of malicious code datasets for hash values and files collected on the website	ML of malicious code and healthy files to predict risk to files collected on the website
Inspection technology	Web crawling	Virtualized PC	Malicious code dataset similarity hash	Malicious code, regular code ML
Inspection speed	High speed	Low speed	High speed	High speed
Zero-day detection	Unable to check	Inspection available	Inspection available	Inspection available
Accuracy	Good	Good	Middle	Good

3. Suggested Method of Website Risk Information Collection

This study proposes collecting information disclosed online to measure website risk. In order to measure the risk of a website, DNS information, IP information, and website history information are required. The website history information may be based on global traffic ranking, malicious code distribution history, and HTTP access status. This study proposed a three-step plan to collect the published website risk information.

3.1. Technique of Website Risk Information Collection

A method for collecting public information for measuring the risk of malicious code is presented as follows. As shown in (Figure 5), risk measurement information for the website can be collected through the website DNS information after inquiring Whois in step 1. Then, in step 2, website IP information can be collected through Whois inquiry. The third step is to collect website history information, which can collect Amazon Alexa inquiries, Google Virustotal, and HTTP response codes.

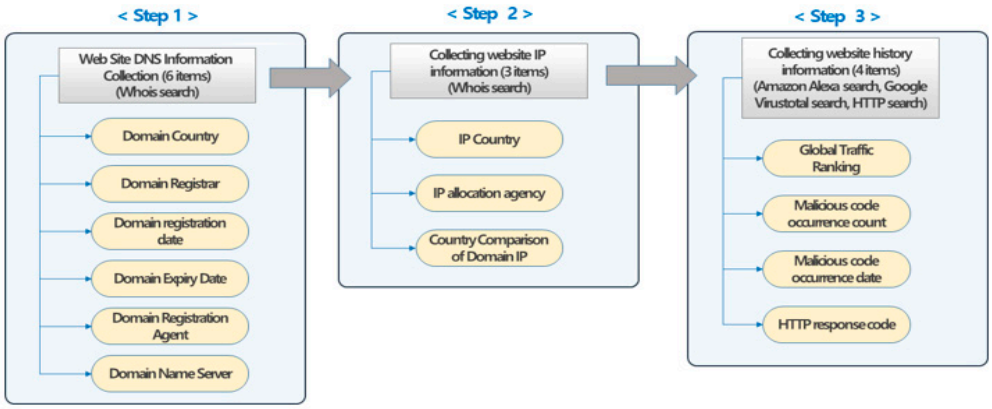


Figure 5. Collection technique of website risk information.

3.2. Technique Analysis of Website Risk Information Collection

Table 2 shows that the website risk information collection can collect 13 items. Through DNS information, it can check the risk of the website according to the country, registrant, registration date, end date, and registered agent information. It is essential whether the IP country, IP allocation agency, DNS country, and IP country are the same through IP information. Among the website history information, if the global track pick ranking is low, the security risk is high and the number of malicious code occurrences increases. In addition, the more recent the occurrence date, the higher the risk. If the HTTP response code is not accessible to the website, the risk can be assumed to increase.

Table 2. Collection information for website security risk measurement.

No.	Contents of Collection	Information of Collection	Methods of Collection
1	Domain country	- Registrant country	Whois search
2	Domain registrar	- Registrant name - Registrant organization - Registrant	
3	Domain registration date	- Creation date - Registered date - Registered on	
4	Domain expiry date	- Expiry date - Expiration date - Expiration date	

Table 2. *Cont.*

No.	Contents of Collection	Information of Collection	Methods of Collection
5	Domain registration agent	- Authorized agency - Registrar - Organization	Whois search
6	Domain name server	- Name server - Host name	
7	IP country	- Country	Whois search
8	IP allocation agency	- Organization - Netname	
9	Country comparison	- Country comparison of domain and IP	
10	Global traffic rank	- Global traffic rank search	Amazon Alexa search
11	Number of occurrences	- Malicious code occurrence count	Google Virustotal search
12	Time of occurrence	- Malicious code occurrence date	
13	HTTP response code	- Checking website access status	HTTP response code search

3.3. How to Collect Website Risk Disclosure Information

3.3.1. Website DNS Information Collection

Information for determining the risk of a website can collect essential information through DNS. As shown in Figure 6, the country, registrar, registration date, end date, registration agent, and name server information is collected through Whois queries. This is to be able to check in advance whether the website is dangerous through basic information. Collecting DNS public information on the website through Whois queries is possible through a string for each keyword when inquiring Whois. For example, to extract a domain country, it can extract it using "Registrant Country". "Reglstrant Country:(.*?)\n/I" means a search for a pre-registration area for a website domain.

```
<Example Code>
$domain = "malware.me";
$whois = LookupDomainName($domain); //whois Search function
$result = NULL;
if($whois) {
        if(preg_match_all("/Registrant Country:(.*?)\n/i", $whois, $tmp)) {
                if($tmp[1][0]) {
                        $result = trim($tmp[1][0]);
                }
        }
}
echo "Result : $result";

<Execution Results>
Result : KR
```

Figure 6. Collecting domain country information with Whois DNS search.

3.3.2. Collecting Website IP Information

Information for determining the risk of a website may collect basic information through IP. As shown in (Figure 7), country, allocation agency, and country comparison information through Whois queries were collected. Collecting public information about the website IP through Whois queries is possible. Collecting public information on the website IP through Whois queries is possible through a string for each keyword when inquiring Whois. For

example, IP country extraction can be extracted using "country". "/Country:(.*?) \n/i" means a search for the entire country where the website is located.

```
<Example Code>
$ip = "1.209.199.227";
$whois = LookupIPAddress($ip); //whois Search function
$result = NULL;
if($whois) {
        if(preg_match_all("/country:(.*?)\n/i", $whois, $tmp)) {
                if($tmp[1][0]) {
                        $result = trim($tmp[1][0]);
                }
        }
}
echo "Result : $result";

<Execution Results>
Result : KR
```

Figure 7. IP country information collection via Whois DNS search.

3.3.3. Collection of Website History Information

Websites with many Internet users invest a lot in security infrastructure to protect users at the corporate level. Therefore, the ranking of website user traffic becomes basic information that can determine whether there is a security risk. As shown in Figure 8, global website ranking information can be collected through Amazon's Alexa Traffic Rank service. As shown in (Figure 8), the Amazon Alexa Rank public information does not identify real-time security threats, but it does show similarities to websites that have been the focus of security incidents. Additionally, you can collect global website ranking information. "#56,586" means the website's global ranking is 56,586.

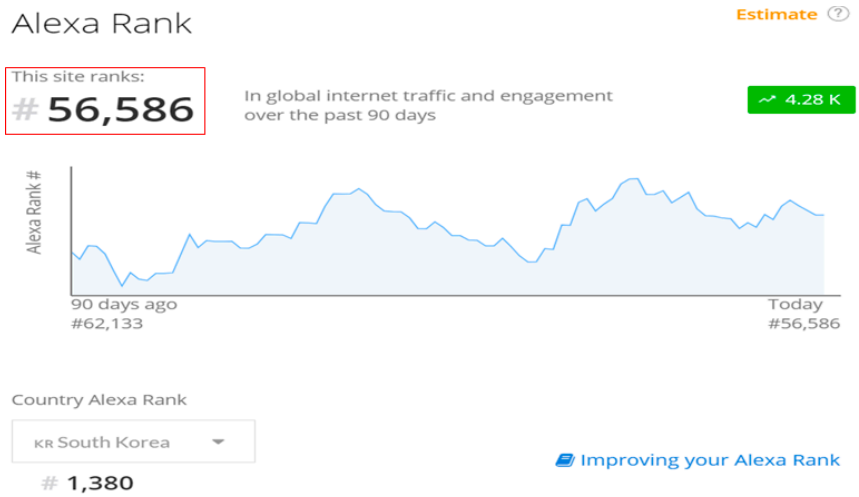


Figure 8. Alexa Traffic Rank search results.

As shown in Figure 9, global traffic ranking information was collected through the Alexa Traffic Rank API.

```
<Example Code>
$domain = "malware.me";
$rank_alexa = array();
$alexa = new Get_Alexa_Ranking();
$rank_alexa = $alexa->get_rank($domain);
print_r($rank_alexa);

<Execution Results>
Array
(
[0] => 56586 // In global internet traffic and engagement over the past 90 days
[1] => KR
[2] => 1380 // Country Alexa Rank
)
```

Figure 9. Gathering global traffic ranking information through Amazon.

Alexa Traffic Rank API Search

In order to measure the security risk of a website, it is essential to have a history of malicious code distribution on the website. In this study, malicious code distribution history information provided by Google Virustotal was used. As shown in (Figure 10), information on the number and timing of malicious code distribution on the website can be collected by linking with the Google Virustotal API.

```
<Example Code>
$domain = "papaya.gotdns.ch";
$result = array();
$result = vtChk($domain, "domain"); // Virustotal API
print_r($result);

<Execution Results>
Array
(
[detected_urls] => Array
(
[0] => Array
(
[date] => 2021-06-08 23:03:17
[url] => http://papaya.gotdns.ch/pawpaw/afo.docx
[positives] => 15
[total] => 88
)
```

Figure 10. Collection of malicious website history information.

Virustotal API

Attackers often open a website quickly to spread malicious code and then delete the website. Alternatively, collecting status information on the website is essential because it may be operated only when malicious code is distributed. As shown in (Figure 11), HTTP

status codes are a way to check the risk of a website's availability in real time. For example, a status code of 200 is normal, while a 404 status code is unavailable and does not provide availability. If the HTTP status code is 0, the risk of the website increases as it does not operate the web server.

```
<Example Code>
$domain = "amazno.co.jp.jwet.xyz";
$httpCode = NULL;
$ch = curl_init();
curl_setopt($ch, CURLOPT_URL, $domain);
curl_setopt($ch, CURLOPT_HEADER, true);
curl_setopt($ch, CURLOPT_RETURNTRANSFER, 1);
curl_setopt($ch, CURLOPT_SSL_VERIFYPEER, false);
curl_setopt($ch, CURLOPT_NOBODY, true);
curl_setopt($ch, CURLOPT_TIMEOUT, 3);
curl_setopt($ch, CURLOPT_CONNECTTIMEOUT, 3);
$response = curl_exec($ch);
$httpCode = curl_getinfo($ch, CURLINFO_HTTP_CODE);
curl_close($ch);
echo "Reult: $httpCode";

<Execution Results>
0
```

Figure 11. Gathering website HTTP status code information.

4. Experiment

This study conducted experiments on 1000 regular domains and 1000 domains classified as malicious code dissemination or phishing websites in 2021. As a result, information related to 2000 website security risks was collected and analyzed. This paper presents the results of an experiment to measure the risk of a website using only OSINT information for public interest purposes.

4.1. Results of Website Domain Security Risk Measurement

To collect public information, we configured an experimental environment by connecting DNS and IP public websites and linked APIs for reputation research. The difference between the information collected for 1000 regular domains and 1000 malicious domains is as follows. As shown in (Table 3), among the unknown items in the collection of domain information were the domain registration date (regular 12%, malicious 66%), the domain termination date (regular 11%, malicious 66%), the domain registration agent (regular 10%, malicious 65%), and the domain name server (10%, malicious 65%). It was confirmed that the four collection items are information that needs to be collected for domain security risk. However, domain countries (64% regular, 75% malicious) and domain registrants (62% regular, 77% malicious) were found to have low discrimination among the unknown items.

Table 3. Results of website domain information collection.

No.	Contents of Collection	Extraction Keyword	
		Regular Website	Malicious Website
1	Domain country	- Unknown (64%) - TOP5: ○ KR (309) ○ US (39) ○ CA (6) ○ JP (2) ○ IS (1)	- Unknown (75%) - TOP5 ○ IS (70) ○ US (66) ○ CN (17) ○ FR (16) ○ KR (14)
2	Domain registration authority	- Unknown (62%) - TOP5 ○ Whois Domain Admin (8) ○ Wikimedia Foundation Inc (8) ○ Redacted for Privacy (7) ○ Organization (3) ○ eBay Korea LLC (3)	- Unknown (77%) - TOP5 ○ REDACTED FOR PRIVACY (77) ○ Withheld for privacy ehf (53) ○ See PrivacyGuardian.org (9) ○ Data protected (5) ○ Not applicable (4)
3	Domain registration date	- Unknown (12%) - TOP5 ○ 1997.09.15 (18) ○ 1996.08.10 (10) ○ 2001.01.13 (8) ○ 1995.01.18 (5) ○ 1999.06.22 (5)	- Unknown (66%) - TOP5 ○ 2001.06.28 (17) ○ 2013.04.12 (15) ○ 2021.01.22 (2) ○ 2020.05.11 (2) ○ 2001.08.10 (2)
4	Domain end date	- Unknown (11%) - TOP5 ○ 2021.10.15 (37) ○ 2022.10.15 (23) ○ 2028.09.14 (18) ○ 2023.10.15 (13) ○ 2023.01.13 (8)	- Unknown 66% - TOP5 ○ 2022.06.28 (17) ○ 2029.04.12 (15) ○ 2022.01.22 (8) ○ 2022.01.09 (5) ○ 2021.12.28 (4)
5	Domain registration agent	- Unknown (10%) - TOP5 ○ Inames Co., Ltd (Seongnam-si, Republic of Korea) (167) ○ MarkMonitor Inc (Meridian, ID, USA) (120) ○ Gabia, Inc (Seoul, Republic of Korea) (114) ○ Whois Corp (Singapore, Singapore) (95) ○ CSC Corporate Domains, Inc (Wilmington, DE, USA) (55)	- Unknown (65%) - TOP5 ○ NameCheap, Inc (Phoenix, AZ, USA) (40) ○ Namecheap (Phoenix, AZ, USA) (40) ○ NameSilo, LLC (Phoenix, AZ, USA) (23) ○ TLDS L.L.C. (Jacksonville, FL, USA) SRSPlus (22) ○ Gandi SAS (Paris, France) (17)
6	Domain name server	- Unknown (10%) - TOP5 ○ ns.gcc.go.kr (43) ○ NS1.GOOGLE.COM (38) ○ ns.gabia.co.kr (17) ○ ns.kornet.net (14) ○ PRMNS.LG.CO.KR (12)	- Unknown (65%) - TOP5 ○ DNS1.REGISTRAR-SERV ○ ERS.COM (23) ○ NF1.NO-IP.COM (23) ○ DNSPOD.COM (22) ○ DNS1.NAMECHEAPHOST ○ ING.COM (19) ○ NS1.DUCKDNS.ORG (15)

4.2. The Security Risk Measurement Results of the Website IP

The security risk measurement results of the website IP are as follows. As shown in Table 4, among the unknown items in the collection of IP information were the results of IP countries (regular 12%, malicious 48%), IP allocation agencies (regular 12%, malicious 48%), and DNS and IP countries comparison (regular 4%, malicious 48%).

Table 4. The information collection results of website IP.

No.	Contents of Collection	Extraction Keyword	
		Regular Website	**Malicious Website**
7	IP country	- Unknown 12% - TOP5 ○ KR (387) ○ US (341) ○ AU (42) ○ CN (31) ○ DE (9)	- Unknown 48% - TOP5 ○ US (230) ○ AU (53) ○ RU (25) ○ NL (21) ○ GB (21)
8	IP allocation agency	- Unknown 12% - TOP5 ○ IRT-KRNIC-KR (389) ○ IRT-APNIC-AP (100) ○ Google LLC (Mountain View, CA, USA) (82) ○ Cloudflare, Inc (San Francisco, CA, USA) (62) ○ RIPE Network Centre (Amsterdam, The Netherlands) (48)	- Unknown 48% - TOP5 ○ IRT-APNIC-AP (153) ○ RIPE Network Centre (123) ○ Cloudflare, Inc (San Francisco, CA, USA) (53) ○ Internet Assigned Authority (13) ○ Google LLC (12)
9	Country comparison	- Unknown 4% - false 67% - true 29%	- Unknown 45% - false 53% - true 2%

4.3. The Security Risk Measurement Results of Website Reputation Inquiry

As shown in Table 5, the security risk measurement results of the website reputation inquiry are as follows. Among the unknown items in the collection of reputation information were the global traffic rank (13% regular, 89% malicious), the number of malicious codes (96%, 1% malicious), the date of occurrence of malicious codes (96%, 1% malicious), and HTTP response code (11% regular, 59% malicious). In particular, when the number of malicious codes and the date of occurrence were usual, it was confirmed that the higher the unknown number, the lower the risk.

Table 5. The information collection results of website reputation inquiry.

No.	Contents of Collection	Extraction Keyword	
		Regular Website	**Malicious Website**
10	Global traffic rank	- Unknown (13%) - TOP5 ○ 1 (18) ○ 13 (8) ○ 233 (5) ○ 11 (4) ○ 572 (3)	- Unknown (89%) - TOP5 ○ 369,628 (9) ○ 2645 (4) ○ 6942 (3) ○ 2,149,797 (3) ○ 6722 (3)
11	Malicious code occurrence count	- Unknown (96%) - TOP5 ○ 1 (23) ○ 2 (4) ○ 460 (1) ○ 29 (1) ○ 7 (1)	- Unknown (1%) - TOP5 ○ 1 (825) ○ 2 (76) ○ 3 (38) ○ 4 (17) ○ 5 (9)
12	Malicious code occurrence date	- Unknown (96%) - TOP5 ○ 2021.03.28 (1) ○ 2021.03.27 (1) ○ 2020.01.13 (1) ○ 2018.09.20 (1) ○ 2018.07.27 (1)	- Unknown (1%) - TOP5 ○ 2021.01.13 (6) ○ 2021.01.11 (6) ○ 2021.01.18 (5) ○ 2021.01.11 (5) ○ 2021.01.14 (5)
13	HTTP response code	- Unknown (11%) - TOP5 ○ 301 (488) ○ 302 (198) ○ 200 (126) ○ 400 (19) ○ 403 (17)	- Unknown (59%) - TOP5 ○ 200 (165) ○ 301 (134) ○ 302 (48) ○ 403 (35) ○ 404 (12)

4.4. Expected Effect

As shown in (Table 6), public information collection is used to measure website security risk. Domain countries and domain registrants were not adopted as items that were difficult to distinguish between normal and malicious. Domain registration date, domain termination date, domain registration agent, domain name server, IP country, IP allocation agency, DNS-IP country comparison, global traffic rank, and HTTP response code showed that the higher the unknown ratio, the higher the risk. It was confirmed that the risk of malicious code occurrence and malicious code occurrence date decreased as the unknown ratio decreased. As a result of this study, 11 items were found to increase the accuracy of risk measurement in public information on the website. We presented experimental results showing that it is possible to collect basic information about publicly available websites and check them for maliciousness. 'o' is an important item and '×' is a non-important item. It indicates that '↑' there are many applicable items, '↓' there are few applicable items, and '-' corresponding items are meaningless.

Table 6. Results of public information utilization experiment for measuring website security risk.

Search Contents	Website Security Risk Utilization of Public Information					
	Contents of Collection	Unknown		Significant	Unknown Rate	Risk
		Regular	Malicious			
Domainin formation	• Domain country	64%	75%	×	-	-
	• Domain registrar	62%	77%	×	-	-
	• Domain registration date	12%	66%	○	↑	↑
	• Domain expiry date	11%	66%	○	↑	↑
	• Domain registration agent	10%	66%	○	↑	↑
	• Domain name server	10%	66%	○	↑	↑
IP information	• IP country	12%	48%	○	↑	↑
	• IP allocation agency	12%	48%	○	↑	↑
	• Country comparison of DNS and IP	4%	45%	○	↑	↑
Reputation information	• Global traffic rank	13%	89%	○	↑	↑
	• Malicious code occurrence count	96%	1%	○	↑	↓
	• Malicious code occurrence date	96%	1%	○	↑	↓
	• HTTP response code	11%	59%	○	↑	↑

5. Conclusions

Recently, malicious code distribution using vulnerabilities in websites accessed by many users has been increasing. Accordingly, the importance of measuring the security risk of websites is increasing. Static and dynamic analysis of website security risks is generally used to analyze website security risks.

Website security risk measurement technology mainly aims at inspection speed, zero-day malicious code analysis possibility, and accuracy of analysis results.

The main method of measuring website security threats is reputation research based on the records of previous security incidents. It is classified into dynamic checks that check malware by accessing the user's terminal environment and static checks that check malware detection patterns.

This study proposed a method that can be collected through information disclosed on the Internet in measuring website risk. DNS, IP, and website history information are

required to measure website risk. Website history information can be checked for global traffic rankings, malicious code distribution history, and HTTP access status.

This study proposed a three-step plan to collect published website risk information. First, this study experimented with collecting and analyzing information related to 2000 website security risks on 1000 regular domains and 1000 domains that were distributed malicious code or classified as phishing websites in 2021. In order to measure the security risk of the website through the experimental results, 11 items for collecting public information with high accuracy and meaningful results are proposed. This study collected basic information on OSINT, a publicly available website, and checked whether it is malicious. Finally, this study confirmed the possibility of big data analysis of OSINT in the future.

Funding: This work was supported by the 2022 Far East University Research Grant (FEU2022S07).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of these data. Data was obtained from KISA and are available at KISA (www.kisa.or.kr, accessed on 13 September 2023) with the permission of KISA.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Shin, K.; Yoo, J.; Han, C.; Kim, K.M.; Kang, S.; Moon, M.; Lee, J. A study on building a cyber attack database using Open Source Intelligence (OSINT). *Converg. Secur. J.* **2019**, *19*, 113–133.
2. Kim, K.H.; Lee, D.I.; Shin, Y.T. Research on Cloud-Based on Web Application Malware Detection Methods. In *Advances in Computer Science and Ubiquitous Computing*; CUTE CSA 2017 2017; Lecture Notes in Electrical Engineering; Park, J., Loia, V., Yi, G., Sung, Y., Eds.; Springer: Singapore, 2018; Volume 474. [CrossRef]
3. Lee, Y.J.; Park, S.J.; Park, W.H. Military Information Leak Response Technology through OSINT Information Analysis Using SNSes. *Secur. Commun. Netw.* **2022**, *2022*, 9962029. [CrossRef]
4. Tan, G.; Zhang, P.; Liu, Q.; Liu, X.; Zhu, C.; Dou, F. Adaptive Malicious URL Detection: Learning in the Presence of Concept Drifts. In Proceedings of the 2018 17th IEEE International Conference on Trust, Security and Privacy in Computing and Communications/12th IEEE International Conference on Big Data Science and Engineering (TrustCom/BigDataSE), New York, NY, USA, 1–3 August 2018; pp. 737–743. [CrossRef]
5. Nandhini, K.; Balasubramaniam, R. Malicious Website Detection Using Probabilistic Data Structure Bloom Filter. In Proceedings of the 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 27–29 March 2019; pp. 311–316. [CrossRef]
6. Shibahara, T.; Takata, Y.; Akiyama, M.; Yagi, T.; Yada, T. Detecting Malicious Websites by Integrating Malicious, Benign, and Compromised Redirection Subgraph Similarities. In Proceedings of the 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), Turin, Italy, 4–8 July 2017; pp. 655–664. [CrossRef]
7. Mishra, H.; Karsh, R.K.; Pavani, K. Anomaly-Based Detection of System-Level Threats and Statistical Analysis. In *Smart Computing Paradigms: New Progresses and Challenges*; Springer: Singapore, 2019; Volume 2, pp. 271–279. [CrossRef]
8. Khan, N.; Abdullah, J.; Khan, A.S. Defending Malicious Script Attacks Using Machine Learning Classifiers. *Wirel. Commun. Mob. Comput.* **2017**, *2017*, 5360472. [CrossRef]
9. Husak, M.; Kaspar, J. Towards Predicting Cyber Attacks Using Information Exchange and Data Mining. In Proceedings of the 2018 14th International Wireless Communications Mobile Computing Conference (IWCMC), Limassol, Cyprus, 25–29 June 2018.
10. Singhal, S.; Chawla, U.; Shorey, R. Machine Learning & Concept Drift based Approach for Malicious Website Detection. In Proceedings of the 2020 International Conference on COMmunication Systems & NETworkS (COMSNETS), Bengaluru, India, 7–11 January 2020; pp. 582–585. [CrossRef]
11. Torres, J.M.; Comesaña, C.I.; García-Nieto, P.J. Machine learning techniques applied to cyber security. *Int. J. Mach. Learn. Cybern.* **2019**, *10*, 2823–2836. [CrossRef]
12. Liu, D.; Lee, J. CNN Based Malicious Website Detection by Invalidating Multiple Web Spams. *IEEE Access* **2020**, *8*, 97258–97266. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.