Hindawi
Security and Communication Networks
Volume 2019, Article ID 1901548, 21 pages
https://doi.org/10.1155/2019/1901548

Research Article

CBR-Based Decision Support Methodology for Cybercrime Investigation: Focused on the Data-Driven Website Defacement Analysis

Mee Lan Han, Byung Il Kwak, and Huy Kang Kim

Graduate School of Information Security, Korea University, Seoul, Republic of Korea

Correspondence should be addressed to Huy Kang Kim; cenda@korea.ac.kr

Received 25 March 2019; Revised 21 August 2019; Accepted 2 December 2019; Published 20 December 2019

Guest Editor: Jungwoo Ryoo

Criminal profiling is a useful technique to identify the most plausible suspects based on the evidence discovered at the crime scene. Similar to offline criminal profiling, in-depth profiling for cybercrime investigation is useful in analysing cyberattacks and for speculating on the identities of the criminals. Every cybercrime committed by the same hacker or hacking group has unique traits such as attack purpose, attack methods, and target. These unique traits are revealed in the evidence of cybercrime; in some cases, these unique traits are well hidden in the evidence such that it cannot be easily perceived. Therefore, a complete analysis of several factors concerning cybercrime can provide an investigator with concrete evidence to attribute the attacks and narrow down the scope of the criminal data and grasp the criminals in the end. We herein propose a decision support methodology based on the case-based reasoning (CBR) for cybercrime investigation. This study focuses on the massive data-driven analysis of website defacement. Our primary aim in this study is to demonstrate the practicality of the proposed methodology as a proof of concept. The assessment of website defacement was performed through the similarity measure and the clustering processing in the reasoning engine based on the CBR. Our results show that the proposed methodology that focuses on the investigation enables a better understanding and interpretation of website defacement and assists in inferring the hacker's behavioural traits from the available evidence concerning website defacement. The results of the case studies demonstrate that our proposed methodology is beneficial for understanding the behaviour and motivation of the hacker and that our proposed data-driven analytic methodology can be utilized as a decision support system for cybercrime investigation.

1. Introduction

Advanced persistent threat (APT) attacks, stealthily and continuously controlled by hackers or hacking groups targeting a specific entity, remain as a challenging threat, particularly to the companies or organizations that handle sensitive funding and information. When successful, APT attacks can have a catastrophic impact on critical infrastructures, such as banking, broadcasting system, and mass media sites. This impact is not speculative or theoretical—in fact, it is supported by various real-world incidents and actual attacks. For instance, in February 2016, a group of hackers stole $81 million from the Central Bank of Bangladesh through its account at the Federal Reserve Bank of New York through an APT attack which targeted constantly the SWIFT payment system for a year [1]. Furthermore, in May 2017, the WannaCry ransomware, another type of APT attack, spread due to the vulnerability in the Microsoft Server Message Block (SMB; the message format used to share folders and files and so on in Microsoft Windows OS). This attack caused catastrophic consequences, such a standstill and disruption of online work in hospitals, companies, and several government agencies. According to Symantec's 2017 annual report [2], the SWIFT case and the WannaCry ransomware case were perhaps launched by the Lazarus group that could be affiliated to the DarkSeoul (DS) case in 2013 and the Sony Pictures Entertainment (SPE) case in 2014. Symantec found that the hacking skills in the SWIFT case were very similar to those used by the Lazarus group, presumably one of the North

Korea's state-sponsored hacking group; the report also found that the malware of the WannaCry ransomware case was related to the one used by the Lazarus group [3]. In the Operation Blockbuster report released by Novetta in 2016, the Lazarus group was reported to hypothetically come in two basic classes—the features known as the wipers and the DDoS malware [4]. The noticeable features of these attacks underpin our interest in the Lazarus group's attack related to the DS case in 2013 and the SPE case in 2014.

On March 20, 2013, in the DS case, the DS's attack destroyed approximately 48,700 computerized and networked equipment items, such as PCs, servers, and network devices of major banks and TV broadcasters in South Korea. South Korea suffered a coordinated strike by a simple but very effective and destructive malware called Wiper A, B, and C [5]. In certain Windows OS environments, the wiper scripts attempted to remove any directories after attempting to overwrite each file with a specific string pattern (i.e., "HASTATI," "PRINCIPES," or "PR!NCPES") [6, 7]. In another incident initiated by the Lazarus group, the SPE was hacked by the self-named Guardians of Peace (#GOP) hacker group. Several malware analysis groups reported that the GOP attack was also related to the North Korean cyber army [8]. The malware used in this attack contained strings written using the Romanization of Korean words (i.e., Korean words were spelled using Latin letters following the English pronunciation). Of note, while the Korean language as spoken in North Korea and South Korea is linguistically identical, there are several important differences in terms of vowels and consonants, phonetic notation, and word spacing [9]. In the aforementioned case, the Romanized words captured in the malware were having various contemporary North Korean words.

From 2009 to 2017, along with the attacks mentioned above, the Lazarus group launched many other attacks (see Figure 1 for further details).

There have been numerous attempts in industry and academia to do hacker profiling and to handle attack incidents. These approaches can be categorized into the following three types: the human-centric analysis, malware-centric analysis, and case-centric analysis. The human-centric analysis approach focuses on hacker network analysis. Known hacker activities (e.g., message postings and discussion) on the hacker communities provide a clue to identify key actors by their reputation. In addition, it can classify the tendency of a hacker based on social networking methods [10, 11]. Unlike the human-centric analysis, the malware-centric approach primarily assumes that the same malware and its variants could be developed by the same or closely similar hacker groups. Among others, features such as API call sequence and control flow can be used to estimate the similarity between the newly detected malware and the known malware [12–14]. In fact, many previous studies on hacker profiling have primarily focused on using information derived from the analysis of the malware itself. While malware analysis could provide information about a malware's functionality and its similarity with the previously known malware family, tracing and analysing hacker information based on the malware centric could have the limitation where the core information can be circumvented. The last approach is the case-centric analysis, and our methodology falls into this category. Overall, only several proposals can be applied to the traditional investigation method, such as criminal profiling methods, to the cyber incident investigation; however, many systematic approaches are currently under development. From the viewpoint of the cyber intelligence analysis, the case-centric analysis has the advantage of making it possible to understand the purpose of attack campaigns; it is important to build profiles of attackers as with other methods of analysis. When performed successfully, such characterizations can facilitate estimating and predicting the attacker's next target in advance.

Based on this insight, the present study proposes the CBR-based decision support methodology for cybercrime investigation. In terms of data, website defacement attack cases occurring between 1998 and 2015 were retrieved from the public archival site zone-h.org (an archive of defaced websites, http://www.zone-h.org). After crawling web resources of the Hypertext Markup Language (HTML) type, data preprocessing for data parsing and data cleaning were performed to amend incomplete, improperly formatted, or duplicate data records. The case vector was designed to intuitively express defaced website cases collected from the public archival site. The reasoning engine will be able to start the major work only after completing the data preprocessing and the case vector design. The similarity measurement based on CBR was performed in them. The clustering algorithm was performed to group-abstracted crime cases into classes of similar cases. Based on the results concerning the DS and SPE cases, we evaluated the performance of the framework for cybercrime investigation by measuring the similarity and clustering algorithm. The results demonstrated that the proposed methodology can be used as a Decision Support System (DSS) to obtain meaningful information about the most similar past cases and related hacker groups.

The main contributions of the present study are summarized as follows:

(i) We present a CBR-based decision support methodology for cybercrime investigation. With the proposed cybercrime investigation scheme, security analysts can find past attack cases that are most similar to a given attack and thus obtain insights to uncover the networks of cybercrime related to website defacement. To deliver high-value intelligence, we adopt the data-driven analytic system.

(ii) We demonstrate the clustering processing and visualization. The clustering processing enables an investigator to efficiently explore large data and interpret the results. Furthermore, the visualization helps an investigator to intuitively recognize crime patterns.

(iii) We propose that it is possible to measure the similarity score and to perform the clustering algorithm by transforming unstructured data (i.e., web defacement cases) into calculable structured data.

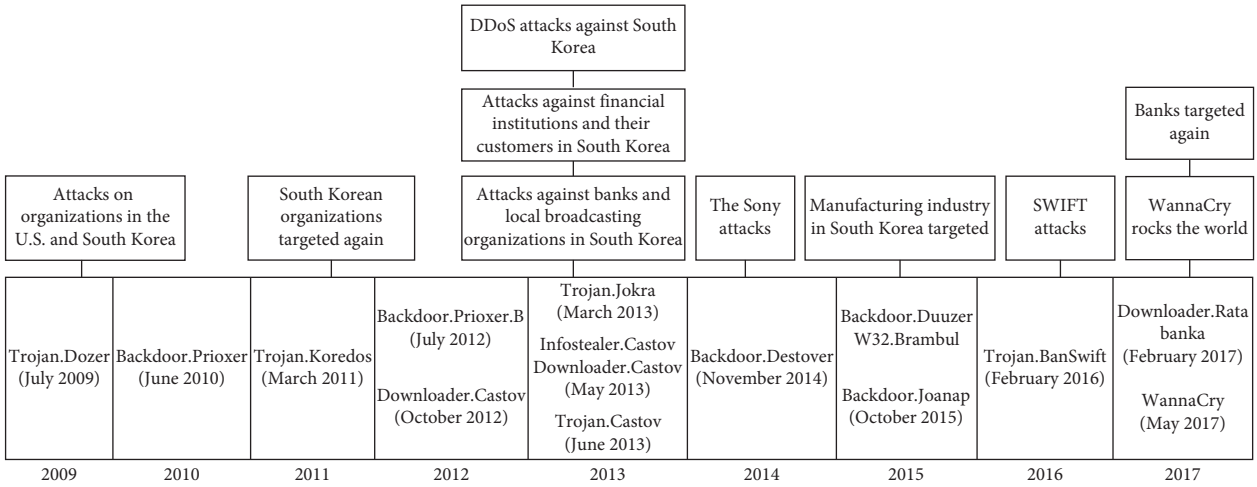

Figure 1: Timeline of the Lazarus group activities from 2009 to 2017.

(iv) We report case studies based on the real dataset gathered from the zone-h.org site to demonstrate the various aspects of our proposed algorithm. Finally, to foster further research, our dataset (the dataset for cybercrime investigation: focused on the data-driven website defacement analysis, http://ocslab.hksecurity.net/Datasets/web-hacking-profiling) is made publicly available [15, 16].

The rest of the paper is organized as follows. Section 2 provides a summary of the literature related to our work. The detailed methodology is described in Section 3. Section 4 reports the experimental results and analysis based on the case study. The limitations of our work and a discussion on the proposed approach are presented in Section 5. Finally, Section 6 concludes the paper and suggests directions of further research.

2. Related Work

In this section, we primarily highlight the previous studies closely related to the CBR and review two streams of literature on traditional criminal profiling and cybercrime profiling. We also elaborate the data mining-based cybercrime profiling pertaining to the following: (1) the CBR studies that help better understand our research context; (2) traditional criminal profiling and cybercrime profiling review that allow us to obtain an elusive criminal or a concealed clue; (3) data mining-based cybercrime profiling literature that can support and theoretically reinforce our methodology.

2.1. Case-Based Reasoning. CBR is a method that uses past experiences or cases to solve new problems. Even when the new problems are not exactly identical to the previous cases, CBR can suggest a partial solution to the new problems [17]. CBR can be categorized as a data-mining technique, as it can classify the given samples and predict the result for a new case. As case studies are intuitive and easily understood by humans, CBR has long been used in many fields, including customer technical support, medical case search, and legal case search. The general model of the four-step CBR process [18] is shown in Figure 2.

The four phases are as follows:

(i) Retrieve: given a new website defacement case, relevant cases are retrieved from the knowledge base to solve the case at hand

(ii) Reuse: solutions from previous website defacement cases are mapped for reuse

(iii) Revise: on mapping and testing previous solutions to the target case, the solutions are revised to consider the changes in the cases

(iv) Retain: after the solution has been successfully adapted to the problem, a meaningful experience is stored as a new case in the knowledge base

CBR starts with a given set of cases for training, forms generalizations of the given examples, and subsequently identifies the commonalities between the retrieved case and the target case. When applied to the website defacement case composed of descriptive and nominal data, it can effectively determine the commonality from the crawled hacking cases and quickly search the nearest related case. Furthermore, CBR can be used to search the most similar cases and retrieve past solutions from the latest response cases. CBR facilitates security administrators to make better decisions. For example, Kim et al. proposed the DSS for an incident response based on CBR [19].

CBR has been extensively used in several areas, such as management for product development, medicine, and in engineering applications [20–22]. In addition, several CBR approaches were available for cyber incidents profiling. For instance, Kim et al. proposed an intelligent system that can measure the similarity between the past and new attacks. In their work, the author(s) demonstrated such capability in uncovering zero-day attacks using the string similarity analysis of the captured packet-level data [23]. Horsman et al. proposed the CBR-FT framework which is a method for collecting and reusing past digital forensic investigation information to highlight likely evidential areas on a suspect

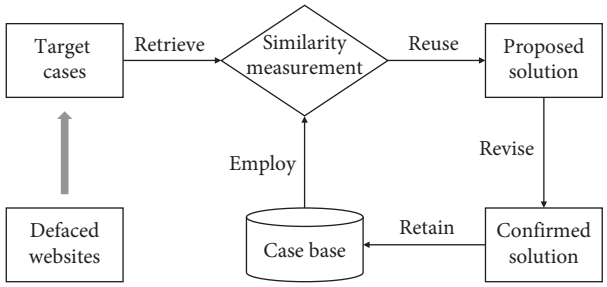

Figure 2: CBR process used in cybercrime investigation, employing knowledge base that is reused over similar new cases and retained for later use.

operating system. It enables an investigator to help quickly and precisely decide where to search for evidence [24].

2.2. Traditional Criminal Profiling and Cybercrime Profiling. Profiling is used in various sectors of the society to investigate a criminal's mentality. Criminal profiling is a profiling technique for criminal investigation based on the psychological and behavioural patterns of a criminal [25, 26]. The criminal aspects and crime factors can be identified through the evidences and insights of the psychological and behavioural bias [27]. In the field of criminology, the widely used profiling technique is called the Modus Operandi (MO). It is used to describe a suspect's behaviour and evidence elements in crime. That is, it means how a suspect commits their crimes. The Modus Operandi changes based on the offender's criminal conduct and interaction with the surrounding such as time, date, and location of crime. Moreover, it evolves based on how the offender reaches his/her victim [28, 29].

Based on only the traditional criminal profiling techniques and empirical knowledge, it is difficult for a cybercrime investigator to reduce the error of the investigative process and to untangle the complexity of a cybercrime. However, if the investigator is provided with sufficient information and detailed analysis data to understand the unclear motivation and the elusive pattern related to the cybercrime, they can infer the reason(s) of the crime at stake and produce both general and specific outlines of the criminal [26]. The cybercrime network and characteristics can be important indicators to differentiate between key figures in the cybercrime organizations and those of passing interest. In addition, their activity periods and message content patterns of the participants in an illegal community can support the investigator to carefully identify and scrutinize the key figure in the cybercrime network [30, 31]. By automating cybercrime profiling and data-mining methods of analysis through a cross-analysis of various behavioural patterns, we can anticipate potential criminal activities and identify new profiles that pose serious threats to the community. Furthermore, data-mining methods such as entity extraction, clustering/classification technique, and social network analysis make it possible to efficiently explore large data. Network visualization enables an investigator to intuitively recognize the crime pattern [32–34].

In general, the accuracy of CBR depends on the quality of the collected data, and the overall accuracy is difficult to evaluate [35]. Although the effectiveness of data-driven investigation can decrease owing to the dynamic and fast-evolving crime patterns, understanding the hidden correlations and latent behaviour in such data using large data analytic techniques is another promising direction in research. Accordingly, many law enforcement agencies have been adopting future crime prediction systems based on the statistics about weather, cleanliness, location, demographic distribution, education level, and wealth-level information. Based on the crime prevention through environmental design (CPTED) theory [36], many pieces of data correlated with the crime are collected and analysed to estimate the crime probability. However, while many data-driven approaches to support traditional criminal profiling are available, only several research efforts have focused on cybercrime profiling.

2.3. Data-Driven Cybercrime Profiling. In addition to the traditional criminal profiling for offline crime investigations, various profiling techniques have been developed; in these techniques, it is assumed that cybercriminals also show similar behavioural and psychological characteristics. Owing to the recent advances in data-mining and machine-learning algorithms, many studies regarding criminal pattern detection, classification, and clustering have emerged. The methods used in these studies include, among others, entity extraction, clustering, association rule mining, deviation detection, and classification of social network analysis. A combination of the traditional method and a newer method enables the pattern identification from both structured and unstructured data. For instance, entity extraction is used to understand concealed patterns in the data such as texts, images, and audio data. Furthermore, clustering is used to group objects into classes with similar characteristics [37, 38]. In addition, unsupervised methods, such as the self-organizing map (SOM), are used to support the results of the traditional criminal profiling [39]. In cases where the criminal and the related cases are known, supervised learning is applied [40]. However, although many advances have occurred in big data analytics and machine learning, these approaches are limited in supporting real-time processing, as they require high computing power to handle a large volume of training data. In fact, the large volume of crime data is a considerable challenge for the investigator in terms of gaining the appropriate understanding of a complicated relationship or in terms of a timely response. However, despite the limitations of this approach, data mining yields valid, useful, and appropriate results. By data preprocessing such as data cleaning, data integration, and data transformation, it intends to reduce noisy data, as well as incomplete and inconsistent data. It helps to uncover and conceptualize the concealed or latent crime patterns. By improving the efficiency of crime data understanding and reducing errors in the results afforded by the data-mining method, the investigator can perform reasoning, timely judgment, and quick problem solving [41].

CBR is also used to provide the reasoning power to search similar previous cases [25, 42]. However, biased or imperfect collected data deteriorate the quality of the decision support provided by CBR. Therefore, in many cases, setting the weight of the selected features is based on empirical knowledge, which can be subsequently used to enable the detection and analysis of crime patterns from the temporal crime activity data. Using clustering and classification techniques, as well as speculative models for searching similar crime cases in the past, investigators can easily extract useful information from the unstructured textual dataset [43]. Hence, investigators must collect and continuously update the comprehensive crime data.

Clustering is the task of determining a similar group in the data. Clustering includes supervised learning types. Zulfadhilah et al. compared four types of clustering algorithms: K-means, hierarchical clustering, SOM, and Expectation Maximization algorithm (EM clustering)—based on their performances. They concluded that the K-means algorithm and the EM algorithm are better than the hierarchical clustering algorithm. In general, partitioning algorithms such as the K-means and EM algorithm are highly recommended for use in large-size data [44]. In summary, the clustering algorithm can facilitate the investigator in detecting crimes patterns and accelerate crime solving. The weighting scheme for attributes can handle the limitations of the clustering techniques [45].

3. Methodology

In this section, we present the detailed scheme of decision support methodology for cybercrime investigation with the focus on the website defacement cases. A conceptual framework and its process are illustrated in Figure 3. The scheme is proceeded by the following three steps: data preprocessing, case vector design, and reasoning engine. First, we provide a brief outline of the dataset and describe the merits of the website defacement data. Also, we summarize the preprocessing for data parsing and cleaning regarding the collected data type. Next, we designed the case vector and chose the significant features to apply the reasoning performance. Finally, the reasoning engine has various functionalities, and it is intended for the grouping (clustering) of cases based on their similarity.

3.1. Preprocessing. As part of the proposed analytical framework, we have developed a crawler to automatically collect 212,093 website defacement cases from the zone-h.org site. Many website defacement cases are being daily recorded in the archive page of the zone-h.org site. Each case registered in the archive page provides information (i.e., IP address, Domain, Date, OS, Notifier, and Web server) of the same format through each mirror page. First of all, the crawler collects all public information relevant to each case. Thereafter, on accessing the domain site, it saves data in the raw format of the HTML source. After crawling the web resources of raw data, the data preprocessing is performed to amend incomplete, improperly formatted, or duplicate data records. More specifically, there are various tag attributes in the HTML source. Encoding and Font data are extracted through the <charset> and <font-style> tag of the HTML elements set between <head> and </head> tag in the HTML source. Also, image, sound file, and the linked site are extracted through the <font-family>, <img>, and <href> tag of them set between <body> and </body> tag in the HTML source. The web resources as original raw data were parsed and cleaned depending on the relevant case vector (see Figure 4). After cleaning the data, some significant data fields were selectively stored in the system's case database.

The selected data fields were related to the information about the website defacement date, related IP address, target domain, target system OS, and web server version; these aspects have proven to be useful for cyberattack investigations [46]. Specifically, the encoding method and the font whom the HTML source contains were necessary to speculate on the attacker's regional information. For example, if messages remaining in a defaced website are written in ISO/IEC 8859 encoding, we can subsequently infer that the hackers' language is German, Spanish, or Swedish. Furthermore, depending on whether all the messages are written in the same encoding method, the used special characters such as β or $\tilde{n}$ or $\mathring{a}$ can be used as a clue for guessing attacker's origin. In general, encodings from Windows-1250 to Windows-1258 are used in the central European languages, as well as in Turkish, Baltic languages, and Vietnamese. By contrast, GB encoding is used in Chinese, HKSCS encoding is used in Taiwanese, and EUC-KR or ISO-2022-KR encoding is used in Korean [47]. In addition to the font and encoding information, the text, image, audio, and video found in the messages are also necessary parameters for the case identification.

3.2. Case Vector Design. We designed the case vector in two types concerning the similarity measure and clustering processing. The case vector is summarized in Table 1. The features of various aspects such as the font, web server, thanks-to, notifier (hackers or hacking groups), as well as the features such as the encoding, IP address, domain, attack date, and OS, were extractable from the public archival site zone-h.org. Generally, more diverse features can be a significant factor for investigating relationships and associations among hackers or hacking groups and the scale and the density/intensity of the hacker community. However, such a premise has some shortcomings. The importance or the weight of all features may be different depending on the criterion. Also, if all features are important, machine-learning algorithms such as clustering or classification are difficult to perform in reality because of the high computational cost for analysing. Despite having similar meanings, some of the features can be reperformed unnecessarily. To this end, the dimensionality reduction and the feature selection were performed in the present study paper. After a thorough review by security experts, the significant features were selected for the case vector of website defacement cases. The detailed explanation of the dimensionality reduction and the feature selection is as follows.

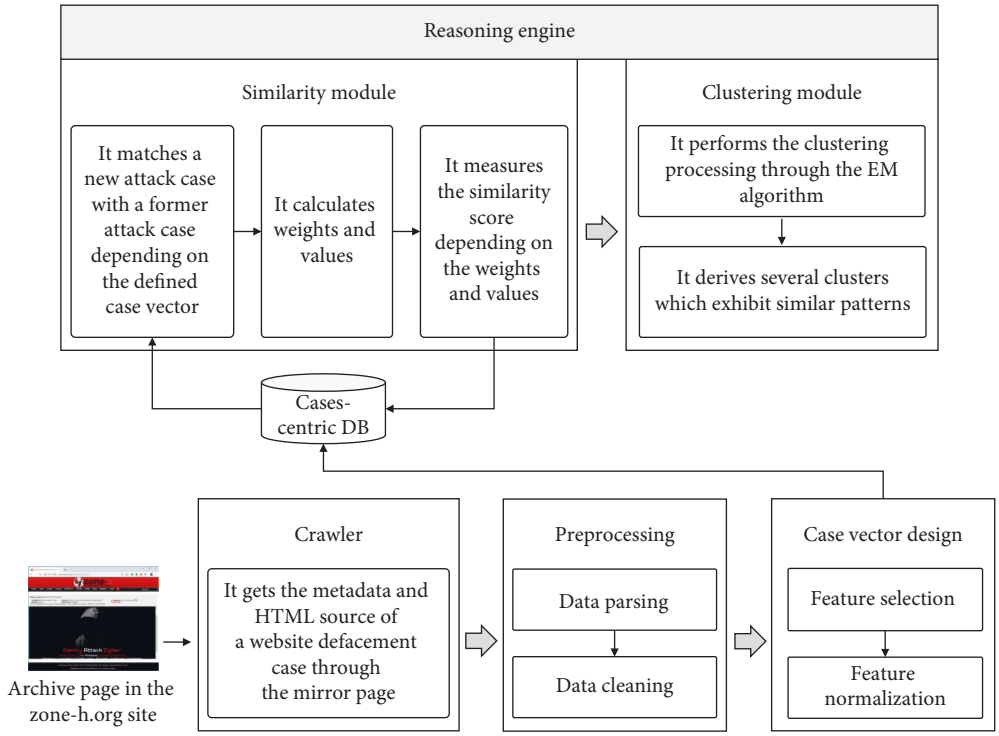

Figure 3: Proposed analytical framework for the data-driven website defacement cases.

Full View | Item View | Script Output |

	Num	Date	Domain	gTLD	ccTLD	IP_Octet1	IP_Octet2	IP_Octet3	IP_Octet4	OS	Encoding	Notifier
1	1	37364.93354	gocomp	com		209	197	254	55	linux		Mr.X
2	2	37267.52771	outfitters	com						linux	utf-8	null
3	3	41558.821	0.static.wix	com		216	139	213	107	linux		HueHueHue
4	4	39820.82729	1.soodoo	com		59	36	101	41	window		Xiaoyu
5	5	41871.3169	132		ir	5	61	24	52	linux		Adr-Elite404
6	6	39981.98212	001bolanhui	net	cn	58	215	85	94	window	WestEurope	federal-atack.org
7	7	37116.06927	001hq	com						unix	utf-8	Kebracho
8	8	39981.98212	001zhanxiaohui	com		58	215	85	94	window	WestEurope	federal-atack.org
9	9	41562.53906	2.huangshi	gov	cn	221	233	125	180	linux		Iranian_Dark_Coders_Team
10	10	41562.53903	4.huangshi	gov	cn	221	233	125	180	linux		Iranian_Dark_Coders_Team
11	11	41562.53892	5.huangshi	gov	cn	221	233	125	180	linux		Iranian_Dark_Coders_Team
12	12	40166.94075	7.hypnotic		hu	217	116	47	129	linux		K-E-H
13	13	41562.47785	1.taranefa1	com		69	43	161	134	linux		Hidden Pain
14	14	39991.80419	0101tech	com		98	130	83	69	linux	WestEurope	TechnicaL
15	15	39993.10155	010fsw	com		61	156	39	67	window		federal-atack.org
16	16	39982.84795	115318878	com		216	46	178	77	linux	Arabic	TheWayEnd

Figure 4: Sorted dataset through the preprocessing.

In the Windows operating system, if a specific font is not designated as the tag inside the HTML code, such as the <font-family> property, the characters on a website page may appear as broken. In particular, some of the fonts among the Chinese characters' cultural area depend on the character encoding (e.g., font-family: Gulim, MingLiU, and STHeiti) [48]. Similar to the encoding feature, although this characteristic may be the key evidence to uncover a correlation between the victim and the attacker, it is extremely rare in each of the collected website defacement cases. Therefore, it is not suitable as a case vector for cybercrime investigation. Meanwhile, in the case of a web server, it provides HTML, CSS, JavaScript, etc. when a client requests a web page using the web server. While the Apache and IIS web servers are primarily used in the Windows environment, the LiteSpeed web server is primarily used in the Linux environment and the Enterprise web server is primarily used in the UNIX environment. Therefore, the web server is selectively dependent on the OS environment. As with the font feature described before, since the web server feature could not be found in the collected website defacement cases, it was not suitable as a case vector for cybercrime investigation. Finally, although the case vector concerning thanks-to and notifier can be used to analyse a hidden network between the hackers and hacker groups, the analysis of a network among hackers and hacking groups through them should be addressed in future research.

As a result, we defined the case vector by dividing into two types, i.e., a version for the similarity measure and a version for the clustering processing. As the features of the case vector, the encoding, IP address, domain (i.e., service name, gTLD, and ccTLD), attack date, and OS were used in the similarity measure. However, the encoding, gTLD, ccTLD, and OS were used in the clustering processing. The encoding is a case vector that provides decisive clues related to the attacker's region information. In the case of the IP

Table 1: Case vector design, highlighting two groups of features.

Case vector	Used in process		Description
	S	C	
Encoding	O	O	It is used to represent the different types of language information on the computer. It determines the usable characters and the methods to express them. The feature was normalized based on MS Windows and the ISO character set
IP address	O	N/A	A unique number that allows devices on the network to identify and communicate with each other
Domain			
Service name	O	N/A	The service name is individually made with a different name depending on the service categories such as gTLD or ccTLD
gTLD	O	O	The gTLD feature was normalized depending on the element having the same meaning (e.g., .go, .gob, and .gobr feature were normalized into .gov)
ccTLD	O	O	The ccTLD is a unique code assigned to the domain name that represents the country, specific region, or an international organization The ccTLD normalized by the continent is used in the clustering process, and the original ccTLD is used in the similarity process
Date	O	N/A	The attack date performed by the hacker or the hacking group
OS	O	O	A part of a computer system that manages all hardware and software (e.g., Windows, Linux, and UNIX)

S, similarity measure; C, clustering processing.

address and domain, it gives clues related to the victim's location and position. Furthermore, the attack date gives clues to the relation between the attacker and the victim. The detailed explanation of key features is provided in Table 1.

The normalization result of various feature elements stored in the raw form of the HTML source is presented in Figure 5. In the case of encoding, ISO series and MS Windows series are applied by normalizing depending on the encoding used in each region or country. In the case of gTLD, it was applied by normalizing depending on the groups or organizations with similar characteristics. In the case of ccTLD, it was applied by normalizing depending on each continent. Although the compression and normalization of features enable making the analysis, such as clustering processing and similarity measure, simple and clear, on the contrary, it may also bring about the loss of information in the original data or make it more difficult to analyse in detail.

3.3. Reasoning Engine. In the reasoning process, the reasoning engine first performs a similarity search based on CBR. Discrete similarity scores are defined to calculate the distance of nominal data (e.g., IP address and domain). Algorithm 1 shows how the similarity module operates by comparing a retrieved website defacement case and all cases in the cases-centric DB on a case-by-case basis. Subsequently, the reasoning engine evaluates the similarity score between the given new attack case vector and vectors of other attack cases. Next, the reasoning engine performs clustering to group-abstracted crime cases into classes of similar crime cases. In crime investigation, a cluster grouped as similar crime case subsets helps to infer crime patterns and speeds up the process of solving a crime due to a better understanding of a complicated relationship or in terms of a timely response. In the present study, we implemented the reasoning engine consisting of two processing entities: the similarity measure processing and the clustering algorithm processing (see below for further details).

3.3.1. Similarity Measure. As the similarity measure based on the CBR algorithm, we proposed the similarity algorithm operated by comparing a retrieved website defacement case and all cases in the cases-centric DB. To begin with, if one of the retrieved cases (RC: a new case) is given and there are "n" cases in the cases-centric DB (TCs: all cases in the cases-centric DB), a comparison between RC and TCs are conducted as "n" times. We defined the extent of similarity between RC and TCs as a numeral value from "0" to "1," where "0" means that RC and TC are unrelated and "1" means that RC and TC are identical. Similarity score ($0 < S < 1$) specifies the extent of similarity between RC and TC. If the similarity score is much closer to "1," RC and TC are more analogous to each other. In the event of multiple case vectors, similarity can be expressed as a weighted sum of case vectors:

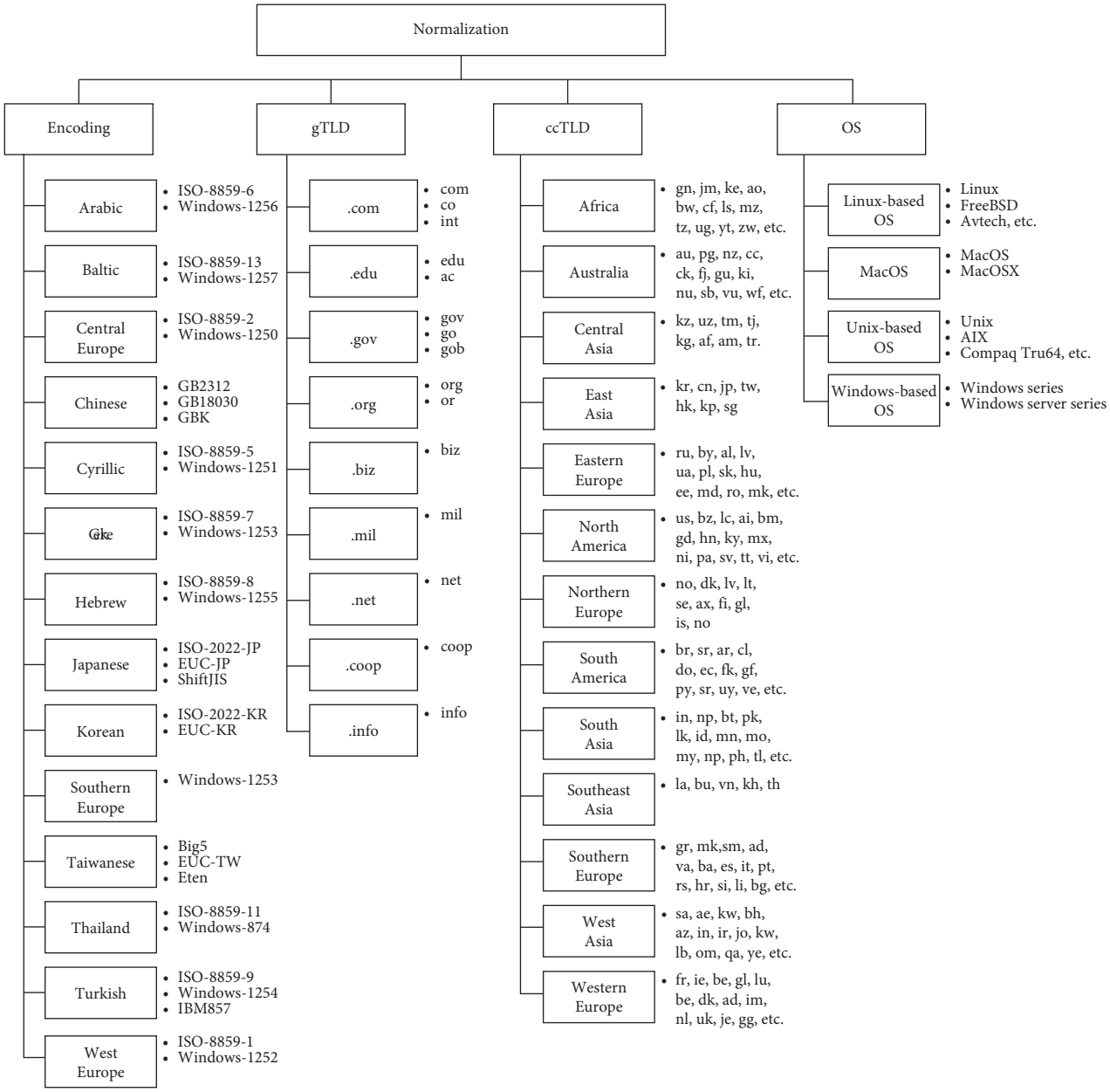

Figure 5: Normalization of each feature elements.

$$\text{Similarity score} = \sum_{i=1}^{\text{cv}} \left[\text{distance}\left(\text{RC}_{\text{cv}}, \text{TC}_{\text{cv}}\right) \times \text{weight}_{\text{cv}}\right],$$

$$\text{cv: case vector (i.e., encoding, IP address, domain, date, and OS).} \quad (1)$$

There are various approaches to set the weight of the case vector, such as the heuristic method, logistic regression analysis, and attribute weighting methods. Furthermore, these weight values need to be periodically updated to be applied to the study of recent attack trends. However, for the initial setting, it is difficult to set the exact numerical value for each weight values in accordance with the case vector. In our experiment, we set the impact and the weight of the case vector as high, medium, and low according to their importance so that to concretely categorize the attacker and the victim. Above all, since encoding makes it possible to infer the static located information of the attacker, we defined encoding as high-quality information. IP address and domain were defined as medium-quality information. These case vectors enable the identification and specification of the victim. Finally, the targeted date and OS were defined as low-quality information. To measure clustering and similarity, all values of the case vector mixed as numbers and letters were normalized to have a value from 0 to 1. Obviously, since these values can be subjective, in order to prevent this subjective bias, these values should be acquired and thoroughly reviewed by several experts. This technique can be easily applied using expert knowledge of investigation experts and is easy to understand from researchers' viewpoint. The quantitative method for setting and

Input: $TCs(Tested\_DB)$/∗ The Tested_DB indicates the cases-centric DB ∗/
RC (Retrieved_Case) $\longleftarrow \{Encoding_{RC}, IP_{RC}, Domain_{RC}, Date_{RC}, OS_{RC}\}$/∗ RC means one of the retrieved cases. ∗/
W (Weight) $\longleftarrow \{Encoding_{W}, IP_{W}, Domain_{W}, Date_{W}, OS_{W}\}$
Output: *Similarity_score*
TC$\{Encoding_{TC}, IP_{TC}, Domain_{TC}, Date_{TC}, OS_{TC}\} \longleftarrow$ TCs
While *RC in TCs* **do**
  if $Encoding_{RC} == Encoding_{TC}$ **then**
    $Encoding\_similarity\_value \longleftarrow 1.0$
  else
    $Encoding\_similarity\_value \longleftarrow 0.0$
  end
  IP_{RC} $\{Octet\_A_{RC}, Octet\_B_{RC}, Octet\_C_{RC}, Octet\_D_{RC}\}$, IP_{TC} $\{Octet\_A_{TC}, Octet\_B_{TC}, Octet\_C_{TC}, Octet\_D_{TC}\}$
  if $(Octet\_A_{RC} == Octet\_A_{TC}) \parallel (Octet\_B_{RC} == Octet\_B_{TC}) \parallel (Octet\_C_{RC} == Octet\_C_{TC}) \parallel (Octet\_D_{RC} == Octet\_D_{TC})$ **then**
    $IP\_similarity\_value \longleftarrow 1.0$
  else if $(Octet\_A_{RC} == Octet\_A_{TC}) \parallel (Octet\_B_{RC} == Octet\_B_{TC}) \parallel (Octet\_C_{RC} == Octet\_C_{TC})$ **then**
    $IP\_similarity\_value \longleftarrow 0.75$
  else if $(Octet\_A_{RC} == Octet\_A_{TC}) \parallel (Octet\_B_{RC} == Octet\_B_{TC})$ **then**
    $IP\_similarity\_value \longleftarrow 0.5$
  else if $(Octet\_A_{RC} == Octet\_A_{TC})$ **then**
    $IP\_similarity\_value \longleftarrow 0.25$
  else
    $IP\_similarity\_value \longleftarrow 0.0$
  end
    $Domain_{RC}$ $\{ServiceName_{RC}, gTLD_{RC}, ccTLD_{RC}\}$, $Domain_{TC}$ $\{ServiceName_{TC}, gTLD_{TC}, ccTLD_{TC}\}$
  if *an identical domain* **then**
    $Domain\_similarity\_value \longleftarrow 1.0$
  else if $(ServiceName_{RC} == ServiceName_{TC}) \parallel (gTLD_{RC} == gTLD_{TC}) \parallel (ccTLD_{RC} == ccTLD_{TC})$ **then**
    $Domain\ similarity\_value \longleftarrow 0.8$
  else if $(gTLD_{RC} == gTLD_{TC}) \parallel (ccTLD_{RC} == ccTLD_{TC})$ **then**
    $Domain\_similarity\_value \longleftarrow 0.3$
  else if $(ServiceName_{RC} == ServiceName_{TC})$ **then**
    $Domain\_similarity\_value \longleftarrow 0.1$
  else if $(ccTLD_{RC} == ccTLD_{TC})$ **then**
    $Domain\_similarity\_value \longleftarrow 0.1$
  else if $(gTLD_{RC} == gTLD_{TC})$ **then**
    $Domain\_similarity\_value \longleftarrow 0.1$
  else
    $Domain\_similarity\_value \longleftarrow 0.0$
  end
    $Date\_variance \longleftarrow |Date_{RC} - Date_{TC}|$/∗ It converts a date format year, month and day (i.e., yyyy-mm-dd) into a day calculated with numeric. ∗/
  if $0 \leq Date\_variance \leq 365$ **then**
    $Date\_similarity\_value \longleftarrow 1.0$
  else if $365 < Date\_variance \leq 1095$ **then**
    $Date\_similarity\_value \longleftarrow 0.75$
  else if $1095 < Date\_variance \leq 1825$ **then**
    $Date\_similarity\_value \longleftarrow 0.5$
  else if $1825 < Date\_variance \leq 2555$ **then**
    $Date\_similarity\_value \longleftarrow 0.25$
  else if $2555 < Date\_variance$ **then**
    $Date\_similarity\_value \longleftarrow 0.0$
  end
  if $OS_{RC} == OS_{TC}$ **then**
    $OS\_similarity\_value \longleftarrow 1.0$
  else
    $OS\_similarity\_value \longleftarrow 0.0$
  end
  $Similarity\_score \longleftarrow (Encoding\_similarity\_value \times Encoding_{W}) + (IP\_similarity\_value \times IP_{W}) + (Domain\_similarity\_value \times Domain_{W}) + (Date\_similarity\_value \times Date_{W}) + (OS\_similarity\_value \times OS_{W})$
  return *Similarity score between RC and TC*
end while

Algorithm 1: Similarity measure module.

updating the weight value is an issue worth addressing in further research. In the present study, we set the weight values for the case vector including the encoding, IP address, domain, attack date, and OS (see Table 2).

Some case vectors' distance cannot be directly estimated, as they have mixed numerical and nominal data (such as IP address range and domain name). For this reason, to calculate the distance between the nominal data, we defined the discrete similarity measure. The similarity of IP addresses was calculated by measuring the similarity among the same octet of two given IP addresses. The IP address space is composed of a number combination of four octets separated by ".". In the present study, we compared if octets from the 1st octet to the 4th octet of RC and TC were identical. Subsequently, a similarity value was assigned to the IP address vector. We suggested the discrete similarity value between two IP addresses as visible in Table 2. The proposed approach is advantageous in that it enables the distance calculation between the IP addresses efficiently:

(i) IP address of RC: zzz: yyy: xxx: www

(ii) IP address of TC: zzz: yyy: xxx: www

Meanwhile, the similarity between domains is calculated according to their domain properties. The domain is composed of the gTLD, ccTLD, and service name. The gTLD refers to a generic top-level domain in the domain rule. For instance, .com and .co are used for commercial companies or organizations, .org and .or are used for nonprofit organizations. .go and .gov are used for government and state agencies. Besides, ccTLD refers to a country code top-level domain in the domain rule and means a unique sign that represents a specific region, such as .kr, .cn, .br, and .uk. DNS makes change in the IP address into a unique Domain Name which is easy to remember because it consists of a combination of an alphabet letter and a number. Among the Domain Name, the service name is built corresponding with the characteristics of the groups, organizations, or corporations that the gTLD is intending and pursuing. The service name has diverse and different names depending on the categories of the gTLD, such as educational institutions, commercial enterprises, military organizations, nonprofit organizations, and government and state agencies. Unlike other case vectors, we set the rule for estimating the similarity of the domain, as depicted in Table 2.

Furthermore, we defined the attack date similarity. Similar to the offline criminal investigation case, if the time of a crime occurrence is near, we can analyse the cases as a similar crime with a cross-analysis of the target, area, and the criminals' patterns. The similarity value depends on the period difference between a new case and existing cases. As visible in Table 2, the similarity value is described according to the date gap of two cases that occurred on different dates. In summary, according to the similarity degree of a variation range of a section, the similarity values of the attack IP address, domain, and attack date were set to the similarity value between 0 and 1.

3.3.2. Clustering Processing. Merely sorting the data and visually analysing them render it difficult for an investigator to infer the correlations and similarity among the potential features of incidents. Hence, an advanced tool that would capture the complex underlying structures and data properties is required. Accordingly, in the present study, we conducted the clustering process using the EM algorithm based on the probability of the individual data attributes. This algorithm does not restrict the number of clusters in the parameters but automatically generates a number of valid clusters by cross-validation. Thereafter, the algorithm determines the probability that some data items existed in the cluster by maximizing the correlation and dependence among the objects. We applied practically the EM algorithm to 80,948 data items having the information of encoding, gTLD, ccTLD, and OS from 212,093 data for clustering. The character encoding was normalized by a group of congenial cover code units (ISO-8859, MS Windows character set, GB, and EUC series). We excluded the Unicode because it is too general, which accounts for the majority of the collected encoding data for clustering. In the case of the service name, even if we can find out similar combinations of alphabet letters or numbers, it is not easy to find commonality or relevance between them. Therefore, it is not suitable for being used as the similarity measure of the reasoning engine. Consequently, characteristics and metadata concerning the 12 clusters were obtained (see Table 3). These clustering results are also visualized and stored in the database (see Figure 6).

The donut charts include the different features from outside to inside (in order), with the corresponding share of each feature value separated by a different colour code within this same circle. Each cluster consists of four circles, and the circle represents, from the outside to the inside, the encoding, gTLD, ccTLD, and OS. The percentage in Table 3 represents how many cases one cluster contains among all of website defacement cases collected from the zone-h.org site. The representative hacker represents a notable hacker or hacking group among the members of them in each cluster. As described in Figure 6, clusters of similar patterns were found in the clusters. The most conspicuously similar clusters were 4 and 7, which had the feature of using Arabic and Chinese, a feature of the attack against an industrial organization whose headquarters are located in Western Europe. The cases in Clusters 4 and 7 accounted for 41.29 percent among all of website defacement cases collected from the zone-h.org site. The results of the clustering process contribute to the concretization of the similarity between the new and existing cases. A large number of new cases have flowed in the database, and then, if the clustering process is performed with the dataset, a clustering result may take on a different pattern, of course.

4. Application

4.1. Experimental Results and Analysis. Considering that the assumption that the attackers tend to use similar or unique attack methods is not always valid, and it is difficult to evaluate the accuracy of the similarity mechanism. As time progresses, attackers' hacking skills advance, and in addition, the attack plan, campaign purpose, and target groups can change depending on the situation. Therefore, in the present

Table 2: Value and the weight for the similarity score by the case vector. All of the values of the similarity score are normalized to 0 or 1.

Case vector	Weight	Impact	The similarity measure between a new case and existing cases	Value
Encoding	0.5	High	—	0 or 1
IP address	0.2	Medium	If the same (e.g., 143.248.1.6 and 143.248.1.6)	1
			If the 1st, 2nd, and 3rd octet are matched (e.g., 143.248.1.6 and 143.248.1.8)	0.75
			If the 1st and 2nd octet are matched (e.g., 143.248.1.6 and 143.248.4.4)	0.5
			Only the 1st octet is matched (e.g., 143.248.1.6 and 143.13.2.4)	0.25
			No common octet (e.g., 143.248.1.6 and 163.13.2.5)	0
Domain	0.15	Medium	An identical domain	1
			Service name is matched, and one of the gTLD and ccTLD is matched	0.8
			gTLD and ccTLD is matched	0.3
			Service name is matched	0.1
			ccTLD is matched	0.1
			gTLD is matched	0.1
			Nonidentical domain	0
Date	0.1	Low	Period of about 6 months back and forth (1 year)	1
			Period of about 18 months back and forth (3 years)	0.75
			Period of about 30 months back and forth (5 years)	0.5
			Period of about 42 months back and forth (7 years)	0.25
			Over period of about 42 months (over 7 years)	0
OS	0.05	Low	—	0 or 1

Table 3: Characteristics and metadata of several different clusters derived from the clustering processing.

Cluster number	Ratio (%)	Description	Representative hacker (group)
0	7.84	The group uses Central European languages. They principally attacked against the profit organization and Linux-based OS in Western Europe.	JaMaYcKa, Super2li
1	8.16	The group uses Arabic and Cyrillic. They principally attacked against the organization that manages the network and Linux-based and Unix-based OS. Their attack region is distributed throughout Southern Europe, South America, Eastern Europe, and Southeast Asia.	BI0S
2	10.36	The group uses Central European languages. They principally attacked against the organization that manages the network and nonprofit organizations in Western Europe.	JaMaYcKa
3	9.33	The group uses Central European languages. They principally attacked against the profit organization and Windows-based OS in Western Europe.	1923Turk
4	25.36	The group uses Arabic and Chinese. They principally attacked against the profit organization and Windows-based OS in Western Europe.	EL_MuHaMMeD, federal-atack.org
5	1.73	The group uses Central European languages. They principally attacked against the profit organization and Unix-based OS in Southern Europe and Eastern Europe.	d3b~X, SuSKuN
6	5.24	The group uses Central European languages. They principally attacked against the profit organization, the educational institution, the government and state agencies, and also Windows-based OS in East Asia.	1923Turk

Table 3: Continued.

Cluster number	Ratio (%)	Description	Representative hacker (group)
7	15.93	The group uses Arabic, Chinese, and Turkish. They principally attacked against the profit organization and Linux-based OS in Western Europe.	Rya, iskorpitx
8	9.11	The group uses Central European languages. They principally attacked against the profit organization and Windows-based OS in Western Europe.	1923Turk
9	3.63	The group uses Central European languages. They principally attacked against the profit organization and Linux-based OS in South America and Eastern Europe.	Hmei7
10	1.39	The group uses Central European languages. They principally attacked against Windows-based OS in South America and Southeast Asia. Their attack target is mostly the educational institution and the government and state agencies.	BHS, F4keLive
11	1.92	The group uses Arabic and Central European languages. They principally attacked against the profit organization and Windows-based OS in Southern Europe.	EL_MuHaMMeD, linuXploit_cre

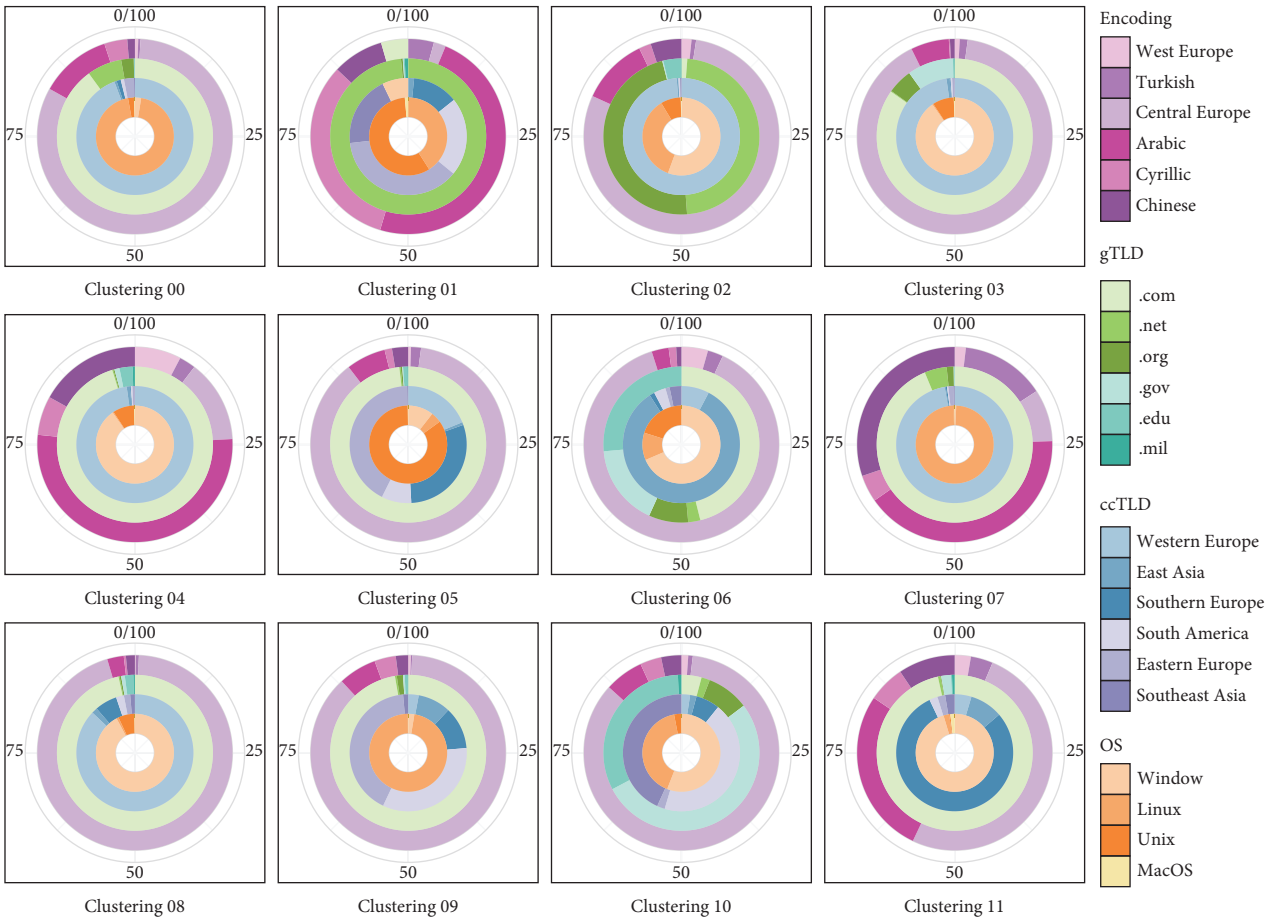

Figure 6: Visualization of the 12 different clusters (00 through 11) in our data annotated with various features: encoding, gTLD, ccTLD, and OS and their corresponding share (legend on the right side).

study, rather than evaluating the accuracy of the similarity mechanism, we tested the overall performance of the proposed methodology with the ratio of correctly identified hackers. The developed testing procedures unfolded in the following four steps and are depicted in detail in Figure 7: where "K" presents all hackers within the database.

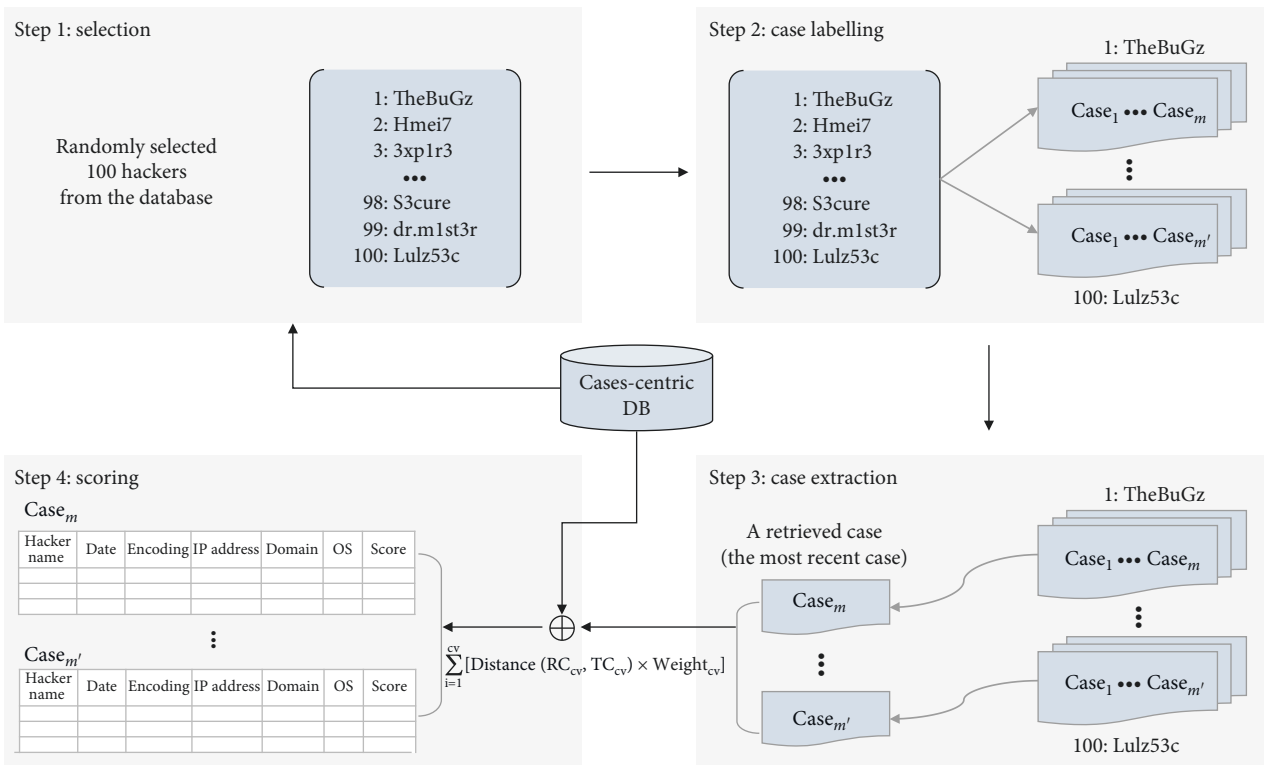

Figure 7: The developed testing procedures from step 1 to step 4.

$$R_k = \frac{\text{Count}\left(\text{Cases}_k^m\right)}{\text{Count}\left(\text{Cases}_k^{\text{all}}\right)}, \tag{3}$$

where "m" means the past cases which are within the defined scope concerning a randomly selected hacker "k."

(i) Step 1, selection: the measurement objects, i.e., 100 hackers were randomly selected from the database.

(ii) Step 2, case labelling: we retrieved all previous attack cases conducted by the randomly selected 100 hackers in Step 1 and then subsequently labelled all previous attack cases by each hacker name.

(iii) Step 3, case extraction: we selected the most recent case among the cases labelled in Step 2 as an input value. The similarity score was then estimated by comparing the most recent case (i.e., RC—one of the retrieved cases) with all other cases in the database (i.e., TCs—all cases in the cases-centric DB).

(iv) Step 4, scoring: similarity score was sorted depending on the value and the weight for the similarity score by the case vector (see Table 2), in the descending order. Whenever the similarity value was 0, it was not displayed on the scoring list of Step 4. The feasibility of the proposed methodology was evaluated based on how many past cases of a hacker there were in the N scope at the scoring list of Step 4; that is, regarding the ratio of the attack cases by each hacker, we checked whether the cases were included at the top N scope (N scope: from the top 1 percent to the top 30 percent):

$$N_{\text{Scope}} = \frac{\text{Count}\left(\text{Cases}_K^{\text{Scope}}\right)}{\text{Count}\left(\text{Cases}_K^{\text{all}}\right)} \times 100, \tag{2}$$

First, we randomly picked 100 hackers from the collected dataset (i.e., cases-centric DB); thereafter, we retrieved and extracted all past attack cases for each hacker. The extracted past cases were labelled with the hacker's name. Figure 8 depicts the number of website defacement attack cases in the past for each hacker. In Steps 3 and 4, similarity between a retrieved case (i.e., the most recent case) and all other stored website defacement cases were measured.

Specifically, we checked whether the result (i.e., the sorted hacker's past cases with a high similarity score) stemming from the similarity measurement was included at the top N scope. This process was meant to check, based on the similarity score, how many past attack cases of randomly picked 100 hackers were included in the defined top N scope. To this end, we divided the top N scope into eight criterion factors from the top 1 percent to the top 30 percent and the ratio R all the past attack cases for each hacker into six criterion factors from 50 percent to 100 percent (i.e., at 10 percent intervals). As illustrated in equations (2) and (3), the N scope and the ratio R were categorized as ratios according to the defined measure rule. More specifically, the criterion of the top N scope, i.e., "top N percent" was based on the result derived from the similarity measurement. Attack cases were sorted in order of high similarity score, and therefore, the cases were within the range of top N scope (see Figure 9). Also, in the case of the hacking case ratio of a randomly

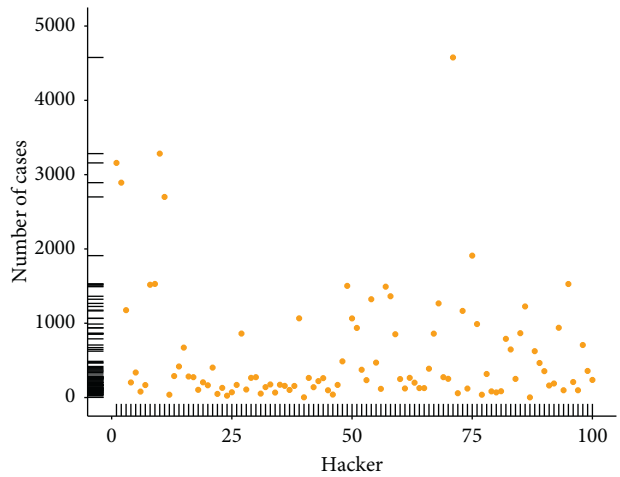

Figure 8: The number of website defacement attack cases in the past of each hacker.

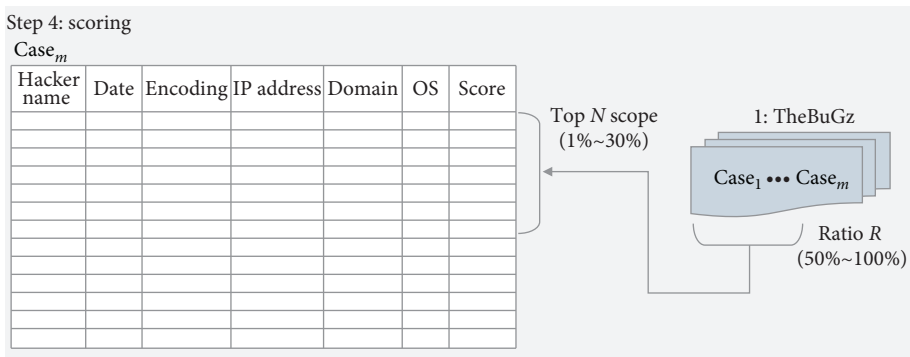

Figure 9: Scoring step on the top N scope and the ratio R.

selected hacker, some parts of the past attack cases (i.e., ratio R) concerning a hacker were within the defined N scope (see Figure 9).

Figure 10 shows the number of an identified hacker for a retrieved case (i.e., the most recent case) among all hacking cases of each hacker. The X-axis in Figure 10 shows the criterion of the top N scope including the eight criterion factors (%) and of the ratio R including the six criterion factors (%). The Y-axis presents the number of an identified hacker in the top N scope among the randomly selected 100 hackers in Step 1. As can be seen in Figure 10, the higher the ratio R and the narrower the N scope, the lower the number of an identified hacker in the top N scope among the randomly selected 100 hackers. On the other hand, the lower the ratio R and the wider the N scope, the higher the number of identified hackers in the top N scope among the randomly selected 100 hackers. Consequently, even if hacking cases were caused by the same hacker, as the hacker or hacking group which only attacked the same or similar objects were rare, it is impossible to draw results with a high similarity score for all cases of a hacker. Nevertheless, the results demonstrated that the proposed CBR-based decision support methodology can successfully reduce the number of hackers and their cases and suggest potential top N percent candidates among hundreds of thousands of cases.

Therefore, an investigator should consider the availability and flexibility of data with respect to the data selection criteria for the similarity measurement. As mentioned above, when a new attack occurs, they can limit the search range of the data and determine the direction of the criminal investigation. With such the reduction in the number of candidate-related cases, the outcomes of our similarity mechanism are highly valuable in terms of reducing the investigation time to determine the potential suspect of a given hacking incident.

4.2. Case Study. As mentioned above, the accuracy of the CBR depends on the quality of the collected data, and the overall accuracy is difficult to evaluate. Nevertheless, although the data are insufficient to evaluate the proposed methodology, the DS and SPE cases include the ground-truth data with specific information related to the hacker or hacking groups. Based on the public ground-truth data of the DS and SPE cases, we found the most similar top three hackers or hacking groups to them and noticed their characteristic, by the proposed similarity measure and the clustering processing.

The hackers of the DS cyberattack defaced the groupware homepage of LG U$^+$, the 3rd largest telecommunication company in South Korea and the English version of the

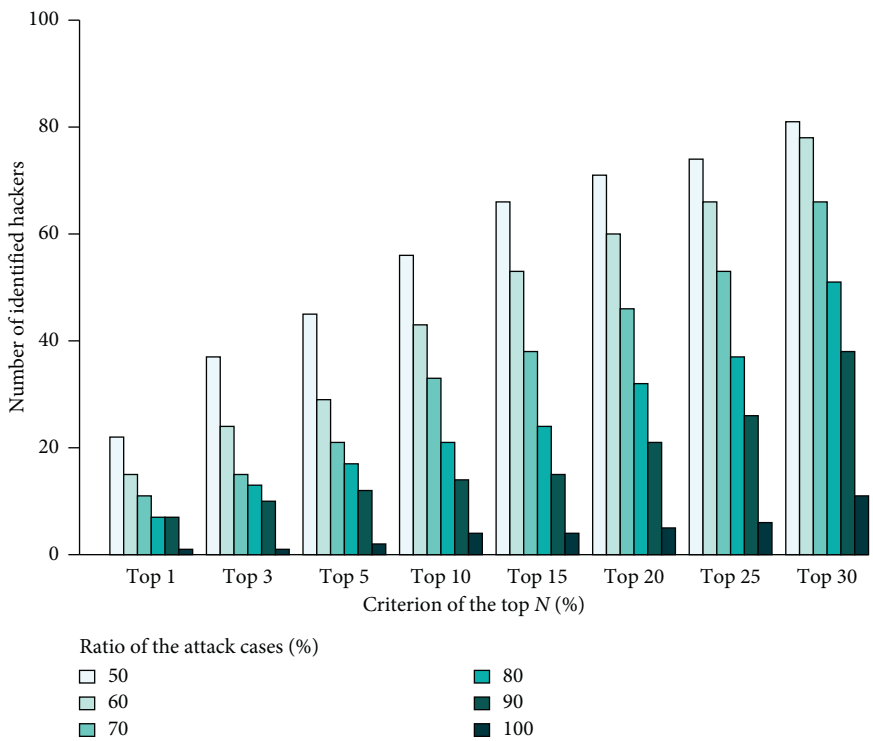

Figure 10: The number of identified hackers in the top N scope among the randomly selected 100 hackers.

Korean Broadcasting System (KBS) homepage. They left unique images and many messages on the defaced websites. The three Calaveras image (i.e., skull image) used in the LG U$^+$'s defaced website appeared on many European websites. The character encoding set of the message was the Western European language system. Based on these insights, we could infer that the hackers' background is European. "HASTATI" was the word written on the KBS homepage, meaning the forefront line of the Roman troops, hinting that the DS cyberattack could be a starting point; rather than a transient attack, it was a persistent one. Even if we excluded other images and messages, as well as other features from the similarity processes due to the unanticipated loss or absence of data, one could establish the similarity and intent of the attackers with reasonable confidence. However, given the sufficiently large hacker profiling source, such abundant data could support and enhance the accuracy of inference. Figure 11 shows the screenshots of the defaced websites at that time.

In the SPE case, similarly to the DS case, some images and messages were left on the computers of SPE. Regarding colour, skulls image, and misspellings, the images Figure 11(c) used in the SPE cases took on the characteristics similar to those of the images Figure 11(b) used in the DS cases. As shown in Figure 11, the colour schemes in green and red and the visual similarities seen in skull image are other crucial elements for crime tracing. In both the DS and SPE cases, the phrase such as "this is the beginning" and "your data" were commonly found in the messages. However, given the intentional hacking nature of forging or hiding their identity, motivation, and location, some experts say that these characteristics are not the conclusive proof that Sony has been attacked by the same hacker [49–51].

For the evaluation of the results of the case study, we first measured the similarity between the new website defacement cases (i.e., the DS and SPE cases) and the collected existing cases in the database. This approach coheres with the CBR process used in cybercrime investigation (see Figure 2). Two new website defacement cases, the DS and the SPE were applied as RC, and the similarity score for each of these two cases was computed using the similarity measure (see equation (1)) proposed in Section 3.3.1. Provided that, because the DS and SPE cases do the function of the target cases as an input value, we considered a direct comparison between the DS and SPE cases for the similarity score was not appropriate [52].

The similarity measure mentioned in the previous paragraph is based on the metadata released by an analysis report of the DS and SPE real cases. We summarized further the characteristics and metadata associated with them in Table 4. The similarity score was derived through comparison between the presented metadata of the DS and SPE cases and all cases in the cases-centric DB. We gave the most similar top three cases among the result of the similarity score (see the right side in Table). Notifier Hmei7 and d3b_X are among the cases that belonged to Clusters 0 and 8, which were the two clusters that exhibited identical characteristics. It can thus be understood that they used the encoding system pertinent to Central European languages based on the Latin language system and typically launched attacks against a profit organization located in Western Europe. Notifier oaddah, M@TRiX, and EL_MuHaMMeD were all classified

(a) (b) (c)

Figure 11: A snippet of website defacement cases by a comparison of examples of the DS and SPE: the defaced LG U$^+$ groupware homepage (a) and KBS homepage (b) in the DS case and the defaced website in SPE case (c).

Table 4: Further characteristics and metadata associated with the DS and SPE cases.

	Retrieved case	Tested cases		
	Case name	Notifier		
	DarkSeoul (DS)	Hmei7	d3b_X	StifLer
Encoding	**Windows-1252**	Windows-1252	Windows-1252	ISO-8859-9
IP address	**203.248.195.178**	203.86.238.68	203.124.37.66	77.92.108.3
Domain	gyunggi.onnet21.com	http://www.garycheng.com	health.ajk.gov.pk	yapikimyasallari.com.tr
Date	**20 Mar 2013**	6 Feb 2014	4 Feb 2014	8 Jun 2013
OS	**Windows**	Windows	Windows	Windows
Similarity	—	0.690	0.675	0.665
Cluster	—	0	8	4
	Retrieved case	Tested cases		
	Case name	Notifier		
	Sony pictures Entertainment (SPE)	Oaddah	M@TRiX	EL_MuHaMMeD
Encoding	**EUC-KR, EUC-CN**	GB2312	GB2312	GB2312
IP address	**203.131.222.102**	203.124.15.55	208.29.19.8	208.116.45.34
Domain	http://www.sonypicturesstockfootage.com	http://www.hzkcgg.com	dax.digitalrom.com	digitalairstrip.net
Date	**24 Nov 2014**	14 Jun 2012	16 Dec 2002	18 June 2009
OS	**Windows**	Windows	Windows	Windows
Similarity	—	0.615	0.615	0.600
Cluster	—	7	7	7

The metadata are arranged according to the defined case vector, corresponding with the DS and SPE cases on the left side (shown in part in boldface type).

as the same cluster (Cluster 7), where the hackers of Cluster 7 used the encoding system pertinent to Arabic and Chinese languages and typically attacked against the profit organization located in Western Europe.

Next, to ensure the objectivity of the similarity score based on the case study by the DS and SPE, we computed the similarity score of any randomly selected pair from the whole case. Figure 12(a) shows the distribution of the similarity score of the randomly selected cases. We took the distribution of the similarity score using the central limit theorem, which describes the average distribution of random samples extracted from a finite population. The distribution shows that the calculation of the similarity score of the randomly selected two website defacement cases was repeatedly performed for 10,000 times. The similarity scores of any randomly selected pair of cases were typically distributed around 0.3. This result (Figure 12(a)) substantiates that the similarity scores are not low, even if the similarity scores of the DS and SPE cases (Figure 12(b)) do not appear numerically high. Figure 12(b) shows the similarity scores of the DS and SPE cases. The top score of the similarity was 0.69 in the DS case, and all measured cases concentrated around the similarity score (X-axis) of 0.0 to 0.15 and of 0.5 to 0.6. In the SPE case, the top score of the similarity was 0.615, and all measured cases concentrated around the similarity score (X-axis) of 0.0 to 0.2.

Figure 13 shows the distribution of the similarity score for randomly selected 100 hackers mentioned in Section 4.1. To know the mean value of the similarity score for each hacker case, we calculated the similarity score from the hacker's own past cases. Cases used for the similarity score means not all cases in the cases-centric DB but just the past cases conducted by the hacker in the cases-centric DB. The mean value of the similarity scores in the hackers is 0.5233. The similarity scores of the tested cases in Table 4 is above the mean value. Thus, the similarity scores for each hacker adequately underpin the similarity scores from the TCs in DS and SPE.

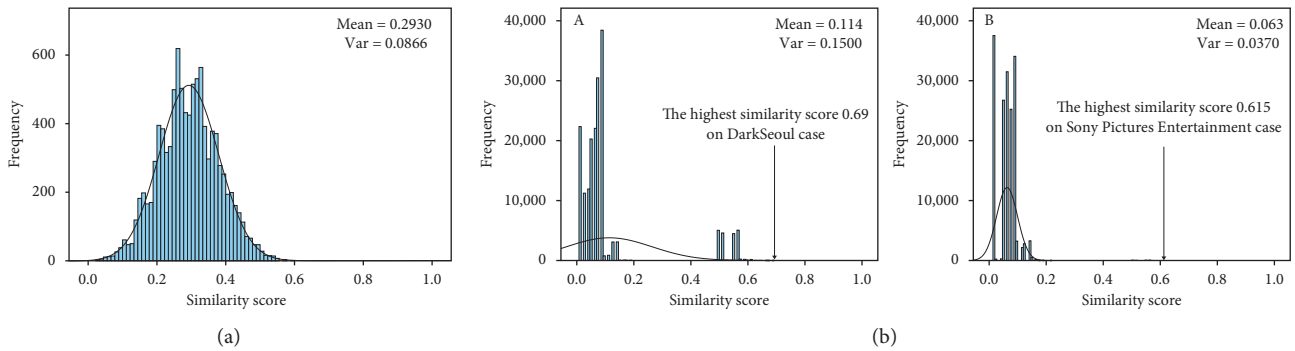

Figure 12: (a) Probability distribution of the similarity score for any pair of randomly selected cases; (b) distribution of the similarity value between the collected website defacement cases with the DS case (A) and the distribution of the similarity value between the collected website defacement cases with the SPE case (B). The similarity was calculated between each studied case and all other cases in our system.

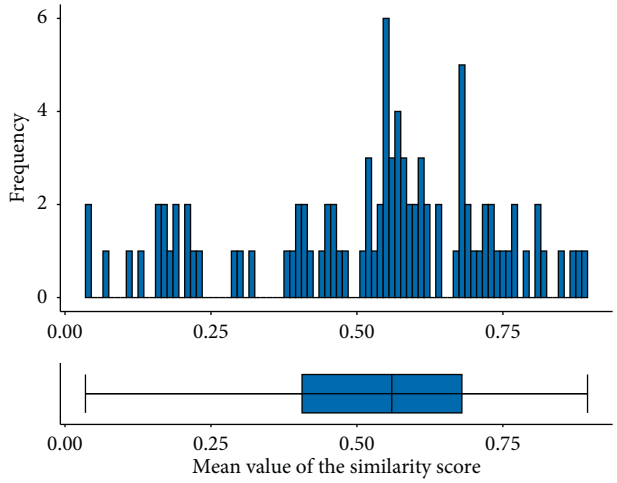

Figure 13: Distribution of the similarity score for randomly selected 100 hackers.

4.3. Follow-Up Investigation. A case study is a research method involving an in-depth and detailed investigation of a subject of study, as well as its related contextual methodology. Hence, we conducted follow-up investigations of the most similar top three hackers, as mentioned above in Table 4. According to the results, specifically, over 93 percent of the hacker's attacks were similar to the DS case that occurred in 2013 and 2014. Their major targets were .com domain sites, and they targeted primarily Germany, Italy, New Zealand, Russia, Turkey, Taiwan, and South Korea (see Table 5). Two hackers (i.e., Hmei7 and d3b_X) primarily attacked government agencies. Interestingly, 20 percent of the attacks by the hackers named d3b_X targeted South Korea. In the SPE incident, the similar hacker's attacks occurred throughout the period from 2002 to 2014. The hackers named M@TRiX and EL_MuHaMMeD intensively executed such attacks in 2003 and 2009. Their major targets were .com (or .co) and .org domain sites, and they targeted primarily Brazil, Canada, Denmark, France, Greece, Hong Kong, and Italy (see Table 5). Two hackers (i.e., M@TRiX and EL_MuHaMMeD) primarily attacked commercial agencies and additionally attacked the public and network agencies. As shown in Figure 14, to describe the follow-up investigation more discernibly and to focus on the attack flow, we used an alluvial diagram, which is a type of Sankey diagram developed to represent changes in a network structure over time [53]. It shows the investigation of the top three hackers with website defacement cases most similar to the DS case and SPE case. The case vectors were based on the attack year, ccTLD, and gTLD. The thickness of the attack flow in this figure means the degree of attack. This network visualization method could support an investigator to understand the flow and core of the crime clearly, by listing the multidimensional evidence that is complicatedly entangled or hidden, such that it does not look presentable.

5. Limitations and Discussion

The CBR algorithm has the disadvantage that the performance evaluation may be degraded if the property describing the case is inappropriate. Therefore, in order to obtain more accurate results, cross-data analysis with other various data sources should be considered. For example, cybercrime statistics data from law enforcement agencies, threat intelligence data from malware analysis groups, and vulnerability databases could be useful resources to

Table 5: Follow-up investigation on the top three hackers with website defacement cases most similar to the DS case and SPE case. The case vector value means the hacker's attack rate.

Domain	DS case			SPE case		
	Hmei7	d3b_X	StifLer	Oaddah	M@TRiX	EL_MuHaMMeD
Com	78.32	85.81	100.00	100.00	86.27	82.98
Edu	1.62	0.96	—	—	1.76	1.91
Net	3.40	3.20	—	—	5.46	5.74
Gov	12.16	6.51	—	—	1.06	—
Year	Hmei7	d3b_X	StifLer	Oaddah	M@TRiX	EL_MuHaMMeD
2002	—	—	—	—	10.74	—
2003	—	—	—	—	89.08	—
2006	—	—	—	—	—	—
2007	0.09	—	—	—	0.18	—
2008	—	—	—	—	—	—
2009	3.15	—	—	—	—	99.57
2010	0.09	—	—	—	—	—
2011	0.34	—	—	—	—	—
2012	3.40	—	—	100.00	—	—
2013	34.86	39.17	100.00	—	—	—
2014	58.08	59.77	—	—	—	0.43
2015	—	1.07	—	—	—	—

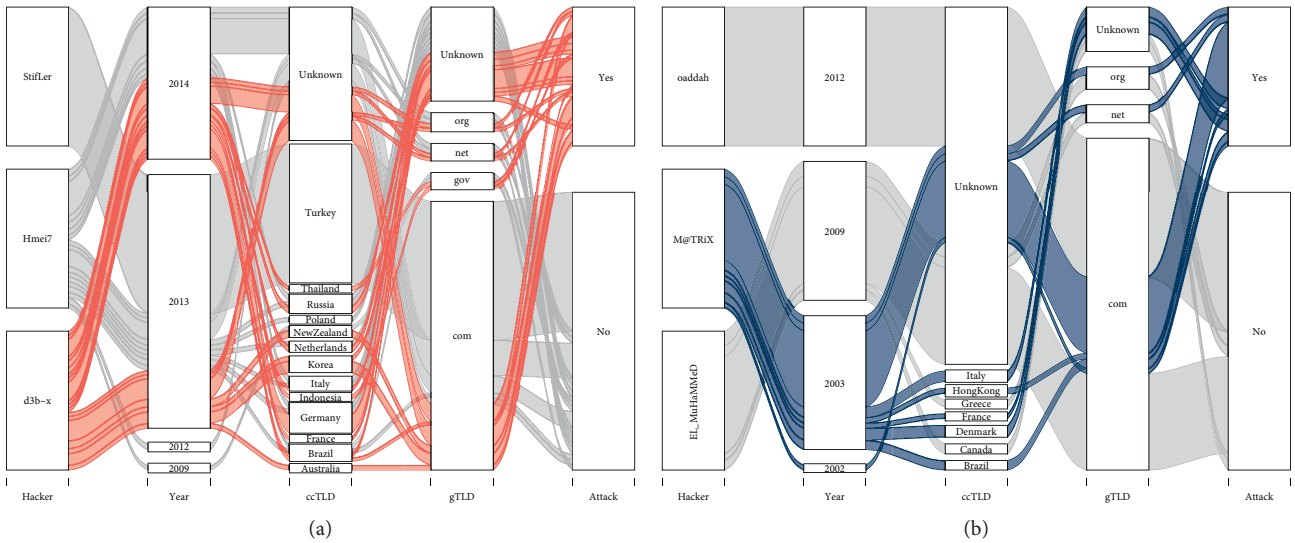

Figure 14: Follow-up investigation on the top three hackers with website defacement cases that are most similar to the DS case (a) and SPE case (b).

improve the accuracy and usability of our proposed methodology. However, at the time of writing the present paper, we did not have access to open and public data concerning cybercrime.

For that reason, we tried to demonstrate the practicability of the proposed methodology as a proof of concept. Therefore, we focused on the dataset of the zone-h.org that includes a large number of website defacement cases. Although the zone-h.org provides an extensive dataset on the past incident events, not all incidents can be included in our study. Therefore, if a hacker penetrated some target organizations by APT attacks and performed stealthy activities, such hacking activities would not be reported in the dataset of the zone-h.org, and the proposed methodology would not be able to detect similar cases with reasonable confidence.

6. Conclusion and Future Work

In this study, the similarity of website defacement cases was assessed through the similarity measure and the clustering processing using the CBR as a methodology. The collected raw data of the defaced web sites' resources was sanitized via data parsing and data cleaning process. Also, based on the large size of real dataset, data-driven analysis for the hacker profiling is achieved. To this end, the case vector was designed, and the significant features were chosen for applying to the case-based reasoning. For a successful cybercrime investigation, hacker profiling via clustering analysis is the most basic and important process; in order to find out the relevant incident cases and significant data on some prime incidents, data-driven

and evidence-driven decision making should be the critical process. Also, reducing the amount of data and time to be analysed are important factors to deliver the high value of intelligence data.

Although the obtained results appear to be sound and meaningful, it is difficult to evaluate the accuracy of the results unless the attacker is captured. Naturally, the ground-truth data with specific information about the involved hacking groups for verification are rare (i.e., no adversary claimed that the two attacks were the result of their actions). However, it is noteworthy that our methodology provides a meaningful insight into the confidential and undercover network of cybercrime as well, especially when there is a lack of information. Also, the proposed methodology contributes to facilitate the analysis and reducing the time required for searching for possible suspects of cybercrime. We believe that the proposed system is meaningful for further exploration and correlation of various website defacement cases.

As mentioned in Discussion and Limitations, a cross-data analysis with other various data sources should be reviewed. Said differently, the use of additional online or offline information acquired by human intelligence (HUMINT) or different types of signal intelligence (SIGINT) and sources may also help to reason composition requirements of crime and reduce the category of investigation. Furthermore, the proposed methodology can be expanded into incident information for compatibility and information exchangeability with other cyberthreat intelligence system as the Structured Threat Information eXpression (STIX) and Trusted Automated eXchange of Indicator Information (TAXII), which are key strategic elements of the information-sharing system [54].

There are features such as the particular messages (i.e., thanks-to, notifier, nationality, religion, and anniversary) or image and mp3 file in the web resources which are gathered from the zone-h.org site. Although these features are limited to only a small number of hackers of the web resources, in future research, we will try to study a close-knit network among them, such as the hub hacking group, key player, and followers. Furthermore, we also plan to more definitely classify and systemize the hackers' intents using text mining and mood detection techniques. The findings of this prospective study will contribute meaningful insights to trace hackers' behavioural patterns and to estimate their primary purpose and intent.

Data Availability

The web-hacking dataset applied to our paper can be downloaded from the linked site below. http://ocslab.hksecurity.net/Datasets/web-hacking-profiling.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported under the framework of international cooperation program managed by the National Research Foundation of Korea (No. 2017K1A3A1A17092614).

References

[1] S. S. Response, "Swift attackers' malware linked to more financial attacks," 2016, https://www.symantec.com/connect/blogs/swift-attackers-malware-linked-more-financial-attacks.

[2] S. S. Response, "Wannacry: ransomware attacks show strong links to lazarus group," 2017, https://www.symantec.com/connect/blogs/wannacry-ransomware-attacks-show-strong-links-lazarus-group.

[3] K. lab, "Lazarus under the hood," 2018, https://media.kasperskycontenthub.com/wp-content/uploads/sites/43/2018/03/07180244/Lazarus_Under_The_Hood_PDF_final.pdf.

[4] Operation Blockbuster, "Destructive malware report," 2016, https://www.operationblockbuster.com/wp-content/uploads/2016/02/Operation-Blockbuster-Destructive-Malware-Report.pdf.

[5] D. Martin and SANS Institute InfoSec Reading Room, "Tracing the lineage of DarkSeoul," 2016, https://www.sans.org/reading-room/whitepapers/critical/tracing-lineage-darkseoul-36787.

[6] D. S. C. T. U. T. Intelligence, "Wiper malware threat analysis," 2013, https://www.secureworks.com/research/wiper-malware-analysis-attacking-korean-financial-sector.

[7] R. Sherstobitoff, M. L. Itai Liba, and O. O. T. C. James Walter, "Dissecting operation troy: cyberespionage in South Korea," 2013, https://www.mcafee.com/enterprise/en-us/assets/white-papers/wp-dissecting-operation-troy.pdf.

[8] N. Horton and A. DeSimone, "Sony's nightmare before christmas: the 2014 North Korean cyber attack on Sony and lessons for US government actions in cyberspace," 2018, https://www.jhuapl.edu/Content/documents/SonyNightmareBeforeChristmas.pdf.

[9] I. K. Lee and S. R. Ramsey, *The Korean Language*, State University of New York, Albany, NY, USA, 2000.

[10] V. Benjamin and H. Chen, "Securing cyberspace: identifying key actors in hacker communities," in *Proceedings of the 2012 IEEE International Conference on Intelligence and Security Informatics*, pp. 24–29, Arlington, VA, USA, June 2012.

[11] Y. Lu, X. Luo, M. Polgar et al., "Social network analysis of a criminal hacker community," *Journal of Computer Information Systems*, vol. 51, no. 2, pp. 31–41, 2010.

[12] J.-W. Jang, H. Kang, J. Woo, A. Mohaisen, and H. K. Kim, "Andro-autopsy: anti-malware system based on similarity matching of malware and malware creator-centric information," *Digital Investigation*, vol. 14, pp. 17–35, 2015.

[13] J. W. Jang and H. K. Kim, "Function-oriented mobile malware analysis as first aid," *Mobile Information Systems*, vol. 2016, Article ID 6707524, 11 pages, 2016.

[14] Y. Ki, E. Kim, and H. K. Kim, "A novel approach to detect malware based on api call sequence analysis," *International Journal of Distributed Sensor Networks*, vol. 11, no. 6, Article ID 659101, 2015.

[15] M. L. Han, H. C. Han, A. R. Kang et al., "Web-hacking dataset for the cyber criminal profiling," 2016, http://ocslab.hksecurity.net/Datasets/web-hacking-profiling.

[16] M. L. Han, H. C. Han, A. R. Kang, B. I. Kwak, A. Mohaisen, and H. K. Kim, "WAHP: web-hacking profiling using case-based reasoning," in *Proceedings of the 2016 IEEE Conference*

on Communications and Network Security (CNS), pp. 344-345, Philadelphia, PA, USA, October 2016.

[17] A. Aamodt and E. Plaza, "Case-based reasoning: foundational issues, methodological variations, and system approaches," *AI Communications*, vol. 7, no. 1, pp. 39–59, 1994.

[18] D. M. L. Martins and F. B. D. Lima Neto, "Hybrid intelligent decision support using a semiotic case-based reasoning and self-organizing maps," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, no. 99, pp. 1–8, 2017.

[19] H. K. Kim, K. H. Im, and S. C. Park, "DSS for computer security incident response applying CBR and collaborative response," *Expert Systems with Applications*, vol. 37, no. 1, pp. 852–870, 2010.

[20] J.-B. Lamy, B. Sekar, G. Guezennec, J. Bouaud, and B. Séroussi, "Explainable artificial intelligence for breast cancer: a visual case-based reasoning approach," *Artificial Intelligence in Medicine*, vol. 94, pp. 42–53, 2019.

[21] M. Relich and P. Pawlewski, "A case-based reasoning approach to cost estimation of new product development," *Neurocomputing*, vol. 272, pp. 40–45, 2018.

[22] E. R. Reyes, S. Negny, G. C. Robles et al., "Improvement of online adaptation knowledge acquisition and reuse in case-based reasoning: application to process engineering design," *Engineering Applications of Artificial Intelligence*, vol. 41, pp. 1–16, 2015.

[23] H. K. Kim, S.-K. Kim, and S.-H. Kim, "Decision support system for zero-day attack response," *Applied Mathematics and Information Sciences*, vol. 6, no. 1, pp. 221S–241S, 2012.

[24] G. Horsman, C. Laing, and P. Vickers, "A case-based reasoning method for locating evidence during digital forensic device triage," *Decision Support Systems*, vol. 61, pp. 69–78, 2014.

[25] G. Horsman, C. Laing, and P. Vickers, "A case based reasoning system for automated forensic examinations," in *Proceedings of the PGNET 2011 the 12th Annual Postgraduate Symposium on the Convergence of Telecommunications, Networking and Broadcasting*, pp. 26–31, Liverpool, UK, June 2011.

[26] Z. Yin, Y. Gao, and B. Chen, "On development of supplementary criminal analysis system based on cbr and ontology," in *Proceedings of the 2010 International Conference on Computer Application and System Modeling (ICCASM 2010)*, vol. 14, Taiyuan, China, October 2010.

[27] A. J. Pinizzotto and N. J. Finkel, "Criminal personality profiling: an outcome and process study," *Law and Human Behavior*, vol. 14, no. 3, pp. 215–233, 1990.

[28] P. Chen and J. Kurland, "Time, place, and modus operandi: a simple apriori algorithm experiment for crime pattern detection," in *Proceedings of the 2018 9th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pp. 1–3, Zakynthos, Greece, July 2018.

[29] C. J. R. Collie and K. Shalev Greene, "Examining modus operandi in stranger child abduction: a comparison of attempted and completed cases," *Journal of Investigative Psychology and Offender Profiling*, vol. 16, no. 2, pp. 91–109, 2019.

[30] V. Benjamin, B. Zhang, J. F. Nunamaker Jr., and H. Chen, "Examining hacker participation length in cybercriminal internet-relay-chat communities," *Journal of Management Information Systems*, vol. 33, no. 2, pp. 482–510, 2016.

[31] V. Benjamin and H. Chen, "Time-to-event modeling for predicting hacker IRC community participant trajectory," in *Proceedings of the 2014 IEEE Joint Intelligence and Security Informatics Conference*, pp. 25–32, The Hague, The Netherlands, September 2014.

[32] K. Veena and K. Meena, "Identification of cyber criminal by analysing the users profile," *International Journal of Network Security*, vol. 20, no. 4, pp. 738–745, 2018.

[33] F. Iqbal, B. C. M. Fung, M. Debbabi, R. Batool, and A. Marrington, "Wordnet-based criminal networks mining for cybercrime investigation," *IEEE Access*, vol. 7, pp. 22740–22755, 2019.

[34] N. Qazi and B. L. W. Wong, "An interactive human centered data science approach towards crime pattern analysis," *Information Processing & Management*, vol. 56, no. 6, p. 102066, 2019.

[35] N. Jain, P. Sharma, R. Anchan et al., "Computerized forensic approach using data mining techniques," in *Proceedings of the ACM Symposium on Women in Research 2016*, pp. 55–60, ACM, New York, NY, USA, 2016.

[36] P. M. Cozens, G. Saville, and D. Hillier, "Crime prevention through environmental design (cpted): a review and modern bibliography," *Property Management*, vol. 23, no. 5, pp. 328–356, 2005.

[37] H. Hassani, X. Huang, E. S. Silva, and M. Ghodsi, "A review of data mining applications in crime," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 9, no. 3, pp. 139–154, 2016.

[38] A. Sharma and S. Sharma, "An intelligent analysis of web crime data using data mining," *International Journal of Engineering and Innovative Technology (IJEIT)*, vol. 2, no. 3, 2012.

[39] S.-T. Li, S.-C. Kuo, and F.-C. Tsai, "An intelligent decision-support model using FSOM and rule extraction for crime prevention," *Expert Systems with Applications*, vol. 37, no. 10, pp. 7108–7119, 2010.

[40] Y.-H. Tseng, Z.-P. Ho, K.-S. Yang, and C.-C. Chen, "Mining term networks from text collections for crime investigation," *Expert Systems with Applications*, vol. 39, no. 11, pp. 10082–10090, 2012.

[41] A. Malathi and S. S. Baboo, "An enhanced algorithm to predict a future crime using data mining," *International Journal of Computer Applications*, vol. 21, no. 1, 2011.

[42] S. Kapetanakis, A. Filippoupolitis, G. Loukas et al., "Profiling cyber attackers using case-based reasoning," in *Proceedings of the 19th UK Workshop on Case-Based Reasoning (UKCBR 2014)*, Cambridge, UK, December 2014.

[43] R. Al-Zaidy, B. C. Fung, A. M. Youssef et al., "Mining criminal networks from unstructured text documents," *Digital Investigation*, vol. 8, no. 3-4, pp. 147–160, 2012.

[44] M. Zulfadhilah, Y. Prayudi, and I. Riadi, "Cyber profiling using log analysis and k-means clustering," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 7, pp. 430–435, 2016.

[45] S. V. Nath, "Crime pattern detection using data mining," in *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops*, pp. 41–44, Hong Kong, China, December 2006.

[46] ITP.net, "Syria, Egypt crises spur escalation of me cyber attacks," 2013, http://www.itp.net/594742-syria-egypt-crises-spur-escalation-of-me-cyber-attack.

[47] A. McEnery and R. Xiao, "Character encoding in corpus construction," in *Developing Linguistic Corpora: A Guide to Good Practice*, Oxbow Books Ltd., Oxford, UK, 2005.

[48] B. Bos, T. Çelik, I. Hickson et al., "Cascading style sheets level 2 revision 1 (CSS 2.1) specification," W3C Working Draft, 2005, http://www.w3.org/TR/CSS21/.

[49] W. Stuckey, "Massive sony breach sheds light on murky hacker universe," 2018, http://america.aljazeera.com/articles/2014/12/24/sony-hacker-universe.html.
[50] S. Gallagher, "Sony pictures malware tied to Seoul, "Shamoon" cyber-attacks," 2018, https://arstechnica.com/information-technology/2014/12/sony-pictures-malware-tied-to-seoul-shamoon-cyber-attacks/.
[51] J. Pagliery, "Sony hack: signs point to North Korea," 2018, https://money.cnn.com/2014/12/05/technology/security/sony-hack-north-korea-employee/index.html.
[52] K. Ketler, "Case-based reasoning: an introduction," *Expert Systems with Applications*, vol. 6, no. 1, pp. 3–8, 1993.
[53] M. Rosvall and C. T. Bergstrom, "Mapping change in large networks," *PLoS One*, vol. 5, no. 1, Article ID e8694, 2010.
[54] OASIS, "STIX/TAXII standards," 2017-2018, https://oasis-open.github.io/cti-documentation/.