ISSN: 2633-4828 Vol. 5 No. S4 (July - Aug 2023)

International Journal of Applied Engineering & Technology

CLUSTERING PROCESS AND VISUALIZATION TECHNIQUES IN CYBER PROFILING: AN EXPLANATORY STUDY

Sakthidevi V[1] and Dr. Vishal Khatri[2]

[1] Research Scholar and [2]Research Supervisor, Department of Computer Science, Sikkim Professional University, Gangtok, (Sikkim)

ABSTRACT

The purpose of criminal profiling is not to assist in solving a case in which there is no evidence. Its primary purpose is to track down the perpetrator or perpetrators in question and make an effort to comprehend them. When a cybercriminal is found guilty of a crime, it is the obligation of the criminal profiler to collaborate with the investigators and review the produced criminal profile in order to determine whether or not there are any linkages to previous profiles that have been developed. Through the evaluation of recent trends, analysis of well-known incidents, and discussion of both mistakes and lessons learned, the purpose of this study was to highlight the considerable contributions that criminal profiling has made to the investigation of cybercrime. The CBR approach's similarity measure and clustering processing were utilized in this investigation to determine how closely related website defacement occurrences are to one another. Sanitization of the raw data that was retrieved from the resources of the hacked websites was accomplished through the use of data parsing and data cleaning procedures. According to the findings of this research, a substantial amount of real-world data was used in the data-driven hacker profiling study. This was accomplished by developing the case vector and selecting the critical elements for the case-based reasoning. A successful investigation into a computer crime must begin with the most fundamental and important stage, which is to profile the hacker who committed the crime using cluster analysis. The process of making decisions based on evidence and data should be the primary method used in order to locate the pertinent incident instances and a substantial amount of data on selected big events. In addition, in order to produce exceptionally valuable intelligence data, one must first reduce the amount of data and then analyze it.

Keywords: Clustering; Cyber Profiling; Cyber Crimes; Case – Based Reasoning (CBR).

INTRODUCTION

The efficacy of cybercriminal profiling has become a hot point of disagreement as cybercrime continues to grow in popularity. Anxieties about the usefulness of cybercriminal profiles have grown in tandem with the rise of online crime over the past decade. Literature review, case study research, and appraisal of possible issues have all lent support to criminal profiling as a crucial tool in cybercrime investigations (Lopez, 2022). Criminal profiling has helped law enforcement agencies identify trends among cybercriminals and distinguish between amateur and expert hackers. This strategy requires a rigorous scientific methodology and an encrypted connection to digital forensics, another investigation instrument (Yu, 2023). Through a literature review, an analysis of developing trends, and a discussion of mistakes and lessons gained, this study aimed to highlight the pioneering contributions of criminal profiling to cybercrime investigations. As more research is done to profile fraudsters, digital forensics may be incorporated into the process to help investigators extract more actionable information.

CBR, a method for solving complications by using relevant instances from the past. Even if it doesn't exactly match an earlier example, CBR can provide a partial answer to a new issue **(Quirion, & Chen, 2021).** CBR fulfils the criteria for a data-mining approach by categorising the supplied data's and forecasting the outcome. Because case studies are simple for people to grasp, CBR has been widely used in a variety of sectors, providing client technical assistance. Figure 1 depicts the overall process of CBR **(Martins and Neto, 2017).**

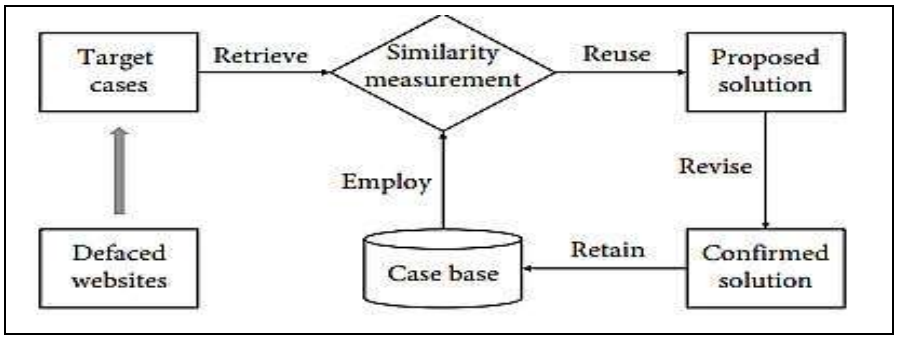

Figure 1: Cybercrime investigation processes leverage knowledge bases that are repeated over comparable fresh cases and stored for future use (Martins and Neto, 2017)

The previous literatures that are relevant to this study are discussed in greater detail in the next section.

AUTHORS AND YEAR	METHODOLOGY	FINDINGS
Ovcharenko et al., (2020)	To study, using the profiling strategy, the hallmarks of a defendant who commits wrongdoing involving illicit drug trade on the internet; to analyze, using the describing procedure, characteristic discover circumstances that arise when committing crimes involving drug trafficking in people via their Internet; and to make complete ideas for reducing these kinds of acts of violence, taking into account both the particulars of the perpetrator.	This portrait should include socio-demographic and psychological characteristics of the offender's category, motivation, victimology, and evidence investigation.
Custers (2021)	Big statistics analyses to detect patterns in enormous data sets are fast growing, as is the portability and usability of information for cybercrime study. This brings up several data-driven research opportunities, including profiling and prediction.	However, emphasizing quantitative, data-driven research may provide methodological, practical, and ethical challenges. This study focused on privacy and profiling issues in big data research and criminality.
Oatley (2022)	Investigated the influence that recent AI-related technological big graph and information processing innovations have significantly impacted important research issues related to crime analytics.	Facts executives, law, and society must overcome AI-driven prescriptive law enforcement, massive amounts of data to stay legal, conflicting selections bias by using data analytics and big data in profiles and estimating illicit activity, anticipating crime risk and incidents, and maintaining AI tools.
Sunardi et al., (2023)	The objectives of this research are to: (1) create a profile of an internet scam customer; and (2) contrast and analyze the approaches normally employed in data mining activities for categorization based on recall, exactness, recollection, error, and efficiency.	According to the results, the Decision Trees Model along with the Nave Bayes Models both perform somewhat better than the random forests Approach. The reliability associated with the Random Forest (RF) approach is 76.8%, although that of the Naive Bayes, Naive Bayes, and decision trees approaches is 77.3%.

Table 1: Literature review

Even while many systematic techniques are being developed, only a few concepts, like criminal profiling, have historically been directly applicable to cyber event investigations. This study's primary objective is to conduct research on case-centric analysis, which calls for attacker profiles to expose the reasons for attack operations based on cyber intelligence analysis. These descriptions can estimate and foretell the future targets of hacker groups.

METHODOLOGY

The decision-support system for cybercrime investigation will be discussed in this section, with an emphasis on cases of website defacement. A conceptual framework and its process are shown in Figure 2. According to the approach, the three processes of data preparation, case vector design, and reasoning engine implementation are completed. At the outset, we'll quickly summarize the dataset and go through some of the advantages of the website defacement data. In accordance with the type of data that was gathered, we will also summarize the preparation for data cleaning and parsing. The important qualities will then be employed to generate the case vector, and reasoning performance will be taken into account. Last but not least, the reasoning engine is designed to classify situations into groups depending on how similar they are and has a number of features.

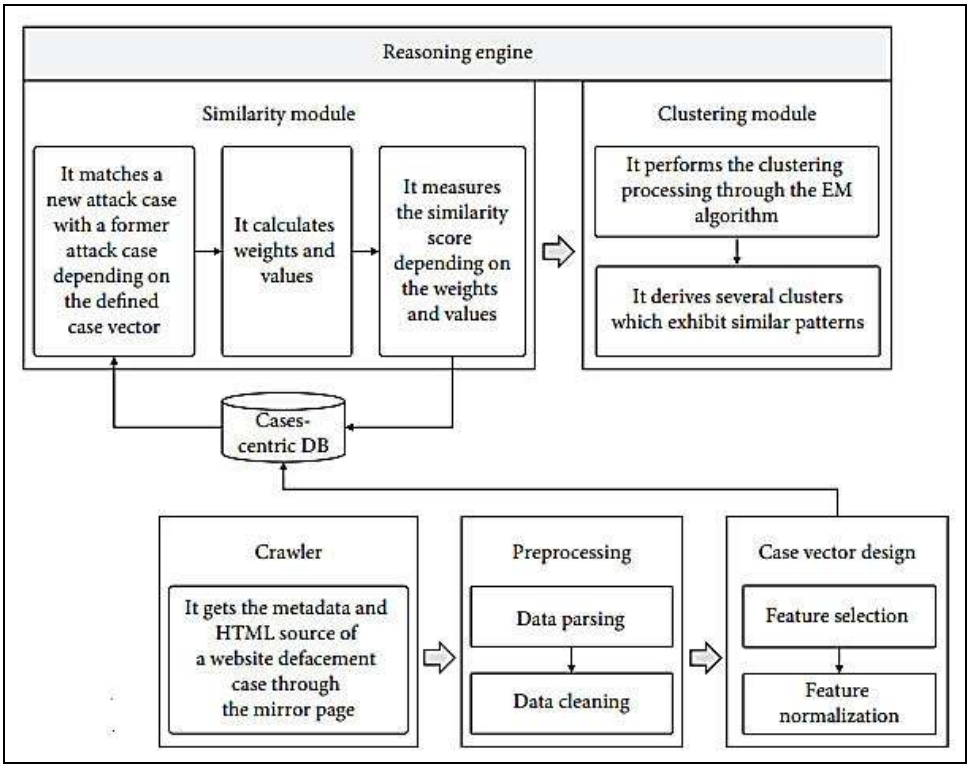

Figure 2: Conceptual structure of website defacement cases based on data

Since attackers may not always employ similar or distinctive attack strategies, it is difficult to evaluate the accuracy of the similarity mechanism. With time, hackers become more proficient, and depending on the situation, the attack tactic, campaign goal, and target audiences may also change. As a result, rather than examining the accuracy of the similarity mechanism in the current experiment, we evaluated the overall performance of the recommended technique using the ratio of successfully discovered hackers. The four stages of the testing procedures that were developed are as follows, where "k" stands for all of the database's hackers.

$$R_k = \frac{\text{Count}(\text{Cases}_k^m)}{\text{Count}(\text{Cases}_k^{all})}, \quad \text{---------- (1)}$$

where “m” refers to the historical instances involving a certain hacker “k” that fall inside the stated parameters.

1. **Selection:** 100 hackers, who serve as the measuring objects, will be chosen at random from the database.
2. **Case labelling:** All prior attacks carried out by the 100 hackers who were chosen at random in Step 1 will be retrieved, and all prior attacks will then be labelled with the names of the 100 hackers.
3. **Case extraction:** we will choose the most recent case as an input value. The most recent case was then compared to every other case in the database to determine the similarity score.
4. **Scoring:** The similarity score will be ordered in decreasing order based on the value and weight that the case vector supplied to it. It was not displayed on the Step 4 score list when the similarity value was 0.

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

RESULTS AND DISCUSSIONS

The CBR algorithm has the drawback that the performance evaluation could get worse if the attribute that describes the situation is unsuitable. This is a potential problem with the algorithm. Therefore, in order to acquire results that are more accurate, one should take into consideration conducting a cross-data analysis using a variety of other data sources. For instance, the statistics on cybercrime that are collected by law enforcement agencies, the threat information that is gathered by the risk inventories that these organisations maintain as well as malware research organisations may all be useful sources for enhancing the precision and usefulness of our proposed method. nevertheless, at the time the current paper was being edited, researchers had no access to information on cybercriminals that is free and open for the general public.

The particular objective of this investigation was to ascertain whether or not the similarity measurement's output—specifically, the ranked hacker's prior instances with a substantial resemblance score—was contained at the very top of N scope. The goal of this process was to count the number of prior attacks carried out by 100 random cybercriminals that fell outside the top N scope, based on similarities score. To achieve this, the strongest N scope was divided into twelve requirements factors with ranges of top just one percent to top twenty percent, and a percentage R to each hacker's prior attack occasions was divided into six requirement aspects with ranges of fifty percent of the total to 100 percent (i.e., at 10% intervals). According to the established determine rule, both the N subject matter and the value of R were categorized as ratios that are used, as shown by the illustrations in equations (1) and (2). To be more exact, the result that was generated from the similarity assessment was used as the basis for the criterion of the top N scope, which was referred to as the "top N percent." Since attack cases were arranged according to their high similarity scores, it should come as no surprise that these examples were inside the top N scope (see Figure 3 below).

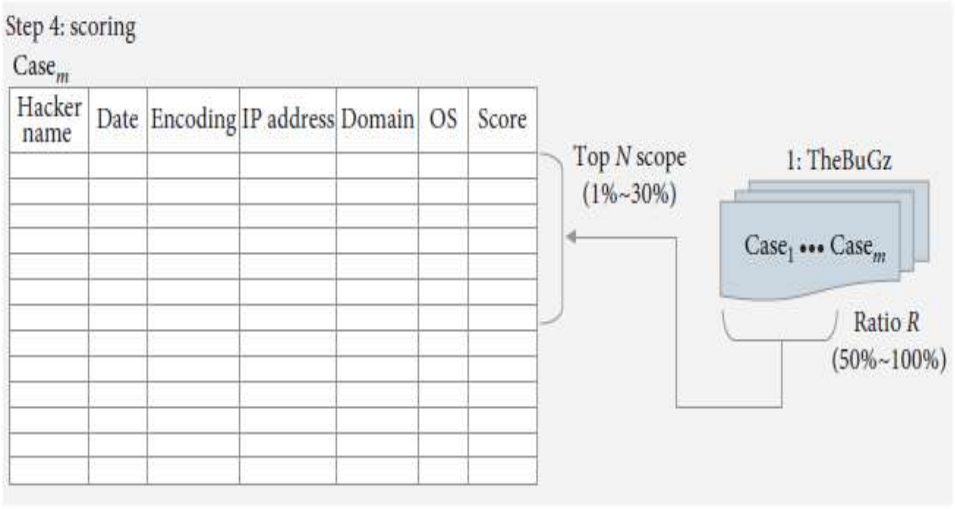

Figure 3: Scoring step on the top N scope and the ratio R.

The percentage of occurrences a known hacker was in charge of a case that was recovered after being hacked (i.e., its most severe event) is displayed in Figure 4. This number is relative to the total number of times each hacker was responsible for a hacked case. Figure 4.4's X-axis depicts the criterion of the top N scope, which takes into account all eight of the criteria components, as well as the criterion of the ratio R, which takes into account all six of the criteria factors.

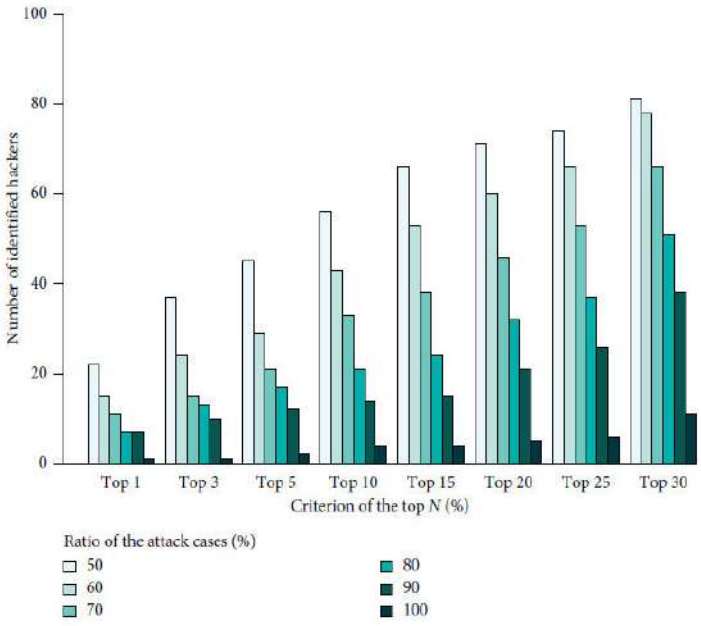

Figure 4: The percentage of detected cybercriminals in the most prominent N scope out among 100 hackers that selected at random.

CONCLUSION

As a result of this, an attempt was made to provide evidence of conception by showing that the suggested approach is feasible. Consequently, we focused on the collection of data supplied by zone-h.org, which includes a large number of various examples of webpage defacement. Despite the fact that zone-h.org offers a sizable dataset

on past incident occurrences, space restrictions prevent us from include each case in our study. Hence, the recommended method wasn't going to be ready to recognize akin instances at the degree of belief that will be accepted to the end user if the hacker gained access to a number of desired companies via APT harms and carried out enigmatic decisions.

REFERENCES

1. Quirion-Blais, O., & Chen, L. (2021). A case-based reasoning approach to solve the vehicle routing problem with time windows and drivers' experience. Omega, 102, 102340.
2. Martins, N. (2017). -*Cybercrime investigation processes leverage knowledge bases that are repeated over comparable fresh cases and stored for future use Computers & Security*, *78*, 398-428.
3. Ovcharenko, M. O., Tavolzhanskyi, O. V., Radchenko, T. M., Kulyk, K. D., & Smetanina, N. V. (2020). Combating illegal drugs trafficking using the internet by means of the profiling method. *Journal of Advanced Research in Law and Economics*, *11*(4 (50)), 1296-1304.
4. Custers, B. (2021). Profiling and predictions: challenges in cybercrime research datafication. *Researching Cybercrimes: Methodologies, Ethics, and Critical Approaches*, 63-79.
5. Oatley, G. C. (2022). Themes in data mining, big data, and crime analytics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *12*(2), e1432.
6. Sunardi, S., Fadlil, A., & Kusuma, N. M. P. (2023). Comparing Data Mining Classification for Online Fraud Victim Profile in Indonesia. *INTENSIF: Jurnal Ilmiah Penelitian dan Penerapan Teknologi Sistem Informasi*, *7*(1), 1-17.
7. Yu, S. (2023). Social media intelligence: AI applications for criminal investigation and national security. In *Handbook of Research on Artificial Intelligence Applications in Literary Works and social media* (pp. 152-170). IGI Global.
8. López, B. (2022). *Case-based reasoning: a concise introduction*. Springer Nature.