IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁒ Peer-reviewed & Refereed journal ⁒ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13461

Uncovering Threats: Data Mining Techniques for Cyber Security

Abhishek Guru[1], Anumolu Vasista Gopal[2], Sai Spandana Bandarupalli[3],

Nanduri Siva Sankar[4], Kakani Rama Rao[5]

Asst Prof, Department of CSE, KL University, Andhra Pradesh, India[1]

CSE, KL University, Andhra Pradesh, India[2-5]

Abstract: To keep an eye out for criminal actions like theft, data modification, and system disruption on one or more computers, we develop an intrusion detection framework. Digital attacks that are dynamic and complex are difficult for traditional intrusion detection systems to detect. But utilizing reliable methods, such various kinds of artificial intelligence, can raise detection rates, lower false alarm rates, and provide affordable solutions. Particularly in data mining, continuous pattern analysis, categorization, aggregation, and real-time data processing are made possible. This research study offers a targeted analysis of the literature on enhanced intrusion detection techniques using data mining and artificial intelligence.[2] In order to provide an analysis, synthesis, and succinct overview of their contents, we identify pertinent publications based on the volume of citations or emerging trends. We also emphasize data's crucial significance in data mining and artificial intelligence.[4]

Keywords: Intrusion detection framework, Artificial Intelligence, data mining, cyber security, cyber resilience

I. INTRODUCTION

Organizations are experiencing an increase in the frequency of cyberattacks, which calls for increased cyber resilience strategies that take both material and nonmaterial repercussions into account. In this context, it is essential for firms to anticipate disruptions, unexpected needs, and possibilities. To create strategic visions and anticipatory intelligence, foresight entails assembling important players and knowledge sources. Resilience engineering requires this capability to be established and managed. Analysts typically manage foresight efforts in businesses or government organizations by quickly discovering, assessing, mitigating, and documenting vulnerabilities and cyberattacks. Although there are diverse views on predicting, recent study by Schatz and Bashroush demonstrates that security experts' forecasts have consistently been right. Scaling up forecasting is difficult, though, because cyber dangers are growing quickly, and information is moving more quickly than analysts can process it. We'll explore relevant work in foresight, introduce the Horizon Scanner tool as a proof of concept, give a qualitative analysis of the findings, and offer some conclusions in our article.[7] A forward-thinking technique called foresight brings together powerful change agents and knowledge sources to produce strategic visions and anticipatory insights. It extends beyond keeping tabs on existing patterns and giving decision-makers useful information about upcoming developments. Understanding future risks and vulnerabilities is essential for developing sustainable strategies in many industries, including cybersecurity, given the field's constant evolution.[8] The need for real-time network anomaly detection methods has increased due to the prevalence of unwanted network activity.

The purpose of intrusion detection systems (IDS) is to identify intruders, detect attacks, collect evidence from unauthorized activity, and react immediately to anomalous circumstances that recognizes departures from the predefined typical characteristics as potential assaults. The data source treatment further divides intrusion detection approaches into host-based and network-based IDS, which examine audit data gathered by operating systems and track online network flow, respectively [10]. Knowledge discovery is the process of obtaining important, formerly undiscovered information from data. Low polynomial complexity in both space and time is a requirement for effective algorithms for this use. The knowledge that is culled from the data should provide original insights. There are two basic strategies: the first focuses on user-guided data exploration, while the second involves machine learning and statistical analysis for pattern finding. EXPLORA, KDW, and Spotlight are notable systems in the first category. Systems like Nielsen Opportunity Explorer and IMACS are common in the second group.[11].

II. LITERATURE REVIEW

Data mining plays a crucial role in fortifying cybersecurity defenses. By sifting through vast amounts of security data, data mining techniques can unearth hidden patterns and anomalies that might signal potential threats. This literature study delves into the application of data mining for cybersecurity, exploring:

DOI: 10.17148/IJARCCE.2024.13461

The Value of Data Mining in Cybersecurity: Articles like "Cybersecurity using Data Mining Techniques" discuss how data mining helps identify vulnerabilities, detect intrusions, and understand attacker behaviour. It can even expose previously unknown threats (zero-day attacks).

Key Data Mining Techniques: Several research papers explore various data mining techniques used in cybersecurity. "A SURVEY OF DATA MINING TECHNIQUES FOR CYBER SECURITY" offers a comprehensive overview of techniques like:

Classification: Categorizes data points into predefined classes, useful for identifying malware or intrusions.
Clustering: Groups similar data points together, aiding in anomaly detection.
Association Rule Learning: Discovers relationships between data elements, helping predict suspicious activities.
Regression Analysis: Identifies how changes in one variable relate to others, potentially uncovering unusual patterns.

Applications of Data Mining in Cybersecurity: "Using Data Mining Techniques in Cybersecurity Solutions" [3] explores how data mining tackles various cybersecurity challenges:

Malware Detection: Data mining algorithms can analyse file behaviour and network traffic to identify malicious software.

Intrusion Detection: By examining system logs and network activity, data mining can spot suspicious attempts to access a system.

Fraud Detection: Data mining helps uncover fraudulent activities like financial scams or unauthorized access attempts.

Threat Intelligence Gathering: Data mining large datasets from various sources helps build a comprehensive picture of the threat landscape.

Insider Threat Detection: Analysing user activity patterns can help identify suspicious behaviours that might indicate insider threats.

III. Ease of Use

A. Related Work of Data Mining for Cyber Security

In order to create strategic visions and anticipatory intelligence, foresight is a method that looks forward while integrating information sources and change agents. It is essential in quickly developing domains like cybersecurity since it not only identifies current trends but also informs policymakers about upcoming developments. Traditional foresight depends on qualitative expert driven methods, which analyses literature and involve expert consultation through workshops, interviews, and surveys. Its use is made of techniques including horizon scanning, future-oriented technology analysis (FTA), and science and technology road mapping. Using the Delphi process, stakeholders can come to an agreement. Commercial technologies for predicting the future include the Gartner Hype Cycle, Trend Watching, and Technology Radar, with Intrada digitizing expert input. Although frequently time consuming and reliant on the opinions of experts, foresight encourages collaboration and perspective modification. Emerging methods for improving foresight by studying enormous amounts of data include data mining and information retrieval. [] Information is gathered from various websites by online applications like Google Trends, Alltop, Trending Reddit, and Bozsum. Europe Media Monitor and TIM are examples of tools created by the European Commission's Competence Centre on Text Mining and Analysis. Future words are visualized using the ITONICS tool Scout using a variety of data sources. With the help of the Horizon Scanner tool, you the use of crawling, scraping, indexing, trend analysis, and visualizations in search. It permits searching for particular phrases, unlike many other programs.[5] Our ICT infrastructure is seriously threatened by ransomware, which is steadily becoming a preferred technique for thieves. Even while the idea of leveraging encryption in Denial of Service (DoS) assaults has been around for a while, the emergence of currencies like Bitcoin has given attackers new options to demand ransom payments in exchange for access to user data.

Ransomware attacks have also been successfully avoided by technical countermeasures such limiting end-user capabilities and confirming program trustworthiness when accessing crypto libraries. Many ransomware detection programs use registry and disk events to spot malicious activity. The majority of the 1,359 ransomware instances studied used equivalent APIs and produced comparable filesystem activity logs. Accurate ransomware detection was accomplished with Bayesian Network models using filesystem and registry events as characteristics. With great accuracy, the ransomware classification system UNVEIL separated ransomware from other malware. A cloud-based detection technology called Cloud RPS identified ransomware based on unusual behaviours including quick file transcoding. EldeRan, which emphasizes the value of prompt detection, used links between OS events to find ransomware within seconds of execution. recognizing a family of ransomware threats. [11]

Anomaly Detection: To find unusual patterns or behaviours, data mining algorithms carefully monitor network traffic, system records, and user behaviour. These anomalies could be signs of unauthorized access or security lapses.

Data mining is essential for detecting intrusion attempts through continuous system log monitoring and the identification of questionable activity. It excels at differentiating between safe and harmful network traffic, which lowers false alarms.

Pattern Recognition: The use of data mining tools is very effective in identifying the tactics and patterns used by hackers. Organizations can take proactive steps to guard against well-known attack vectors by identifying these trends and taking note of them.

Behavioural Analysis: The construction of user behavior profiles through data mining is possible. The occurrence of deviations from these predetermined profiles sets up alerts, which demand additional study.

Predictive Analysis: Data mining can foresee potential security threats or vulnerabilities by utilizing previous data and trend analysis, allowing companies to be proactive in putting preventive measures into place.

Vulnerability Assessment: Data mining helps firms prioritize and address important security problems by methodically evaluating and scanning software and system vulnerabilities.

Phishing Detection: Phishing emails and websites can be recognized using data mining techniques. To do this, suspicious trends are found by carefully examining email content, URLs, and user interactions.

Malware Detection: The early detection and elimination of harmful software is made easier by data mining's contribution to the discovery of malware signatures and behaviours.

The term "User and Entity Behaviour Analytics" (UEBA) refers to the use of data mining tools to examine user and entity activity within a network in order to identify insider threats or compromised accounts.

Forensic Analysis: Post-security incidents, data mining plays a pivotal role in forensic analysis by reconstructing events, pinpointing attack vectors, and tracing the source of the breach.

Security Information and Event Management (SIEM): Data mining seamlessly integrates with SIEM solutions, aggregating, correlating, and analysing security event data from diverse sources to offer a comprehensive view of an organization's security posture.

Fraud Detection: Beyond traditional cybersecurity, data mining techniques are extensively used in financial institutions to spot fraudulent transactions and activities, enhancing security in the financial sector.

Data mining techniques significantly enhance organizations' capabilities to identify, respond to, and prevent cybersecurity threats and incidents, solidifying their role as indispensable tools in the ever-evolving landscape of cybersecurity.

B. Horizon Scanner Tool

The Horizon Scanner tool is designed to assist analysts in recognizing new developments and technology trends in the area of cyber operations, covering both defensive and offensive aspects from a military perspective. These themes cover issues that are relevant to companies in terms of weaknesses, advancements, and potential threats. Several essential features for the Horizon Scanner tool were discovered during the initial requirements-gathering session, which involved about 20 cyber professionals. First and foremost, experts underlined the need to recognize new technical advancements and trends over time. Topics or trends with a noticeable increase in mentions or publications are particularly intriguing. While focusing only on the total number of publications clarifies well-known issues, it falls short in revealing new ones. Time graphs showing the most prominent phrases are used. [6]

The main objective of the tool is to offer perceptions and situational awareness about what is occurring or anticipated to occur in a specific field. A Horizon Scanner Tool's primary attributes and capabilities often include:

C. Overview of the Horizon Scanner Tool Architecture

The Horizon Scanner tool and its modules' inputs and outputs are shown in the figure. It shows the data collection procedure on the right side. Through the user interface, users can submit documents, and everyday internet sources are used. The information is automatically gathered (crawled) and processed (scraped). Entity extraction finds pertinent single, double, or triple word pairings. After that, semantic models (word2vec), which are involved in query expansion, are trained using the cleaned-up data. Additionally, this information is indexed in a text database. Section [specified section] has a thorough overview of the crawling, scraping, and indexing procedure. [5]

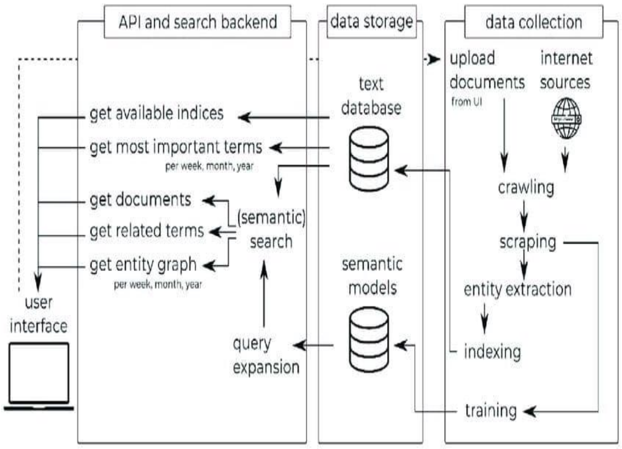

Fig 1. Overview of the Horizon Scanner Tool

Using trend analysis, one can better understand how technologies and ideas develop within a certain domain by looking at data trends and patterns across time.

Tracking Mentions and Publications: This feature keeps track of mentions and publications pertaining to particular subjects or trends, enabling the discovery of new themes and the significance of such themes.

Growth Analysis: The tool evaluates the growth in mentions or publications concerning particular topics. Rapidly growing trends may indicate emerging technologies or areas of interest.

Entity Graphs: Some tools offer visual representations of relationships between different knowledge areas, innovations, and trends. This aids in connecting various topics and comprehending their interconnections.

Customizable Alerts: Users can configure alerts for specific keywords, topics, or trends. When significant developments occur in line with the chosen criteria, the tool can promptly notify users.

Data Sources: Horizon Scanner Tools draw data from diverse sources, including websites, research papers, patents, news articles, and social media. This diversity enhances the tool's data coverage.

Visualization: These tools often incorporate visualization capabilities, making it easier for analysts to interpret data trends and relationships through user-friendly displays.

Customization: Users can tailor the tool to meet their specific needs and interests, selecting which topics or trends to track and receive insights on.

Cross-Domain Analysis: Certain Horizon Scanner Tools enable the analysis of trends and innovations across different domains or sectors, offering a broader perspective.

Time Series Data: The inclusion of time series data allows users to observe the evolution of specific trends over time.

The Horizon Scanner Tool proves valuable for various purposes, including strategic planning, research and development, and cybersecurity. In the context of cybersecurity, for example, this tool assists analysts in staying informed about emerging threats, vulnerabilities, and defensive technologies within the cyber operations domain, facilitating proactive planning and threat mitigation.

DOI: 10.17148/IJARCCE.2024.13461

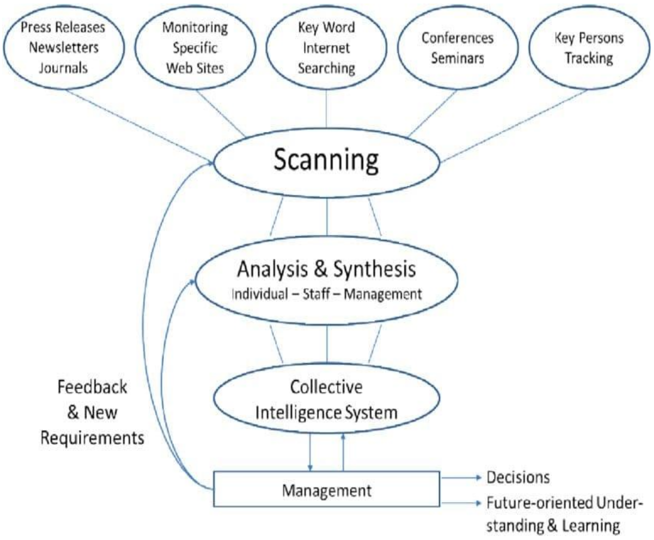

Fig 2. Conceptual model of a horizon scanning system

The architecture of the Horizon Scanner Tool is designed to facilitate the identification and analysis of emerging trends and technological developments within a specific domain.

User Interface: The tool starts with a user interface that allows users to interact with and configure the tool according to their specific needs.

Data Collection: The tool collects data from various sources. This includes two main data collection methods Document Upload: Users can upload documents or data sources relevant to the domain of interest. These documents may include reports, research papers, patents, and more. Internet Access: The tool also has the capability to access and retrieve data from internet sources. It can crawl websites, and access news articles, research publications, and other online content.

Data Processing: Crawling and Scraping: Data collected from internet sources undergoes crawling and scraping processes. Crawling involves visiting websites and collecting data, while scraping extracts specific information from web pages.

Entity Extraction: Relevant entities, which can be one-, two-, or three-word combinations, are extracted from the processed data. These entities could be keywords, topics, or specific terms related to the domain.

Semantic Models (Word2Vec): The tool uses semantic models like Word2Vec to train on the extracted entities. These models help in understanding the relationships and associations between different terms and concepts in the data.

Indexing: The processed and enriched data is stored in a text database or index. This index makes it easier to search and retrieve information during analysis.

Query Expansion: The semantic models trained earlier are utilized in query expansion. This means that when users search for specific terms or trends, the tool can provide expanded and related search results based on the semantic understanding of the data.

Bourgeois et al 2014	Generate information	Generate action	Cooperation & networking	
UK commons science & technology committee 2014	Support strategy development	Make policymaking resilient	Improve operational delivery	
Nicolini & bagni 2012	Build strategic vision and create a shared sense of commitment	Informing policymaking	Build networks	Develop capabilities including foresight culture
Peter Ho 2010	Identify emergent risks	Develop policy and new capabilities	Build global networks & partnerships	Develop policy and new capabilities

TABLE 1. Types of objectives in foresight projects

DOI: 10.17148/IJARCCE.2024.13461

IV. CONCLUSION

This article proposes the Horizon Scanner tool, which helps analysts search the web for new cybersecurity threats and vulnerabilities. The program collects data from web sources and PDFs using text mining and information retrieval algorithms, then stores, searches for, and displays this data using an entity graph, trend visualization, and key term summary. We evaluated the tool's ability to help analysts find hot issues in cybersecurity through an initial requirements session and user evaluation. It's important to note that the proof of concept was not speed optimized, and the tool's data volume and performance are inferior to those of commercial search engines. However, it was discovered to be useful for extracting important phrases over particular time frames, assisting in the detection of subtle threat signals.

REFERENCES

[1] Y. Ye, T. Li, D. Adjeroh, and S. S. Iyengar, “A Survey on Malware Detection Using Data Mining Techniques,”ACM Computing Surveys, vol. 50, no. 3, pp. 1–40, Jun. 2017, doi: 10.1145/3073559.

[2] Y. Yang, X. Li, Z. Yang, Q. Wei, N. Wang, and L. Wang, “The Application of Cyber Physical System for Thermal Power Plants: Data-Driven Modeling,” Energies, vol. 11, no. 4, p. 690, Mar. 2018, doi: 10.3390/en11040690.

[3] 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, doi: 10.1109/access.2019.2891694

[4] G. A. Afzali and S. Mohammadi, “Privacy-preserving big data mining: association rule hiding using fuzzy logic approach,” IET Information Security, vol. 12, no. 1, pp. 15– 24, Jan. 2018, doi: 10.1049/iet ifs.2015.0545

[5] [1] M. H. T. de Boer, B. J. Bakker, E. Boertjes, M. Wilmer, S. Raaijmakers, and R. van der Kleij, “Text Mining in Cybersecurity: Exploring Threats and Opportunities,” Multimodal Technologies and Interaction, vol. 3, no.3, p.62, Sep. 2019, doi: 10.3390/mti3030062.

[6] K. S. Manoj and P. S. Aithal, “Data Mining and Machine Learning Techniques for Cyber Security Intrusion Detection,” International Journal of Engineering and Advanced Technology, vol. 9, no. 3, pp. 4084–4090, Feb. 2020, doi: 10.35940/ijeat.c5979.029320

[7] S. Liu et al., “Model-Free Data Authentication for Cyber Security in Power Systems,” IEEE Transactions on Smart Grid, vol. 11, no. 5, pp. 4565–4568, Sep. 2020, doi: 10.1109/tsg.2020.2986704.

[8] Wu, Q.; Shao, Z. Network Anomaly Detection Using Time Series Analysis. In Proceedings of the Joint International Conference on Autonomic and Autonomous Systems and International Conference on Networking and Services (ICAS-ISNS’05), Papeete, French Polynesia, 23–28 October 2005; IEEE: Piscataway, NJ, USA, 2005; Volume 5, p. 42

[9] Miles, I.; Harper, J.C.; Georghiou, L.; Keenan, M.; Popper, R. The many faces of foresight. In The HPractice; Edward Elgar Publishing: Cheltenham

[10] Eldman, R.; Dagan, I. Knowledge Discovery in Textual Databases (KDT). In Proceedings of the First International Conference on Knowledge Discovery and Data Mining, Montreal, QC, Canada, 20–21 August 1995; IEEE: Piscataway, NJ, USA, 1995; Volume 95, pp. 112–117

[11] S. Homayoun, A. Dehghantanha, M. Ahmadzadeh, S. Hashemi, and R. Khayami, “Know Abnormal, Find Evil: Frequent Pattern Mining for Ransomware Threat Hunting and Intelligence,” IEEE Transactions on Emerging Topics in Computing, vol. 8, no. 2, pp. 341–351, Apr. 2020, doi:10.1109/tetc.2017.2756908.