Received: 16 Jul. 2023 | Revised: 28 Jul. 2023 | Accepted: 14 Aug. 2023 | Published online: 29 Aug. 2023

RESEARCH ARTICLE

Journal of Global Humanities and Social Sciences
2023,Vol. 4(4)171-175
DOI : 10.61360/BoniGHSS232014160804

Big Data-Driven Threat Intelligence Analysis and Early Warning Model Construction

Peng Zhang*,[1]

[1]*Guangdong Mechanical & Electrical Polyitechinc，China*

Abstract: With the development of big data technology, its application in various fields is becoming increasingly widespread, especially in the field of threat intelligence analysis and early warning. By using the processing power of big data and data mining technology, patterns and laws of threats can be discovered from the massive and multifaceted data, thus realizing effective early warning of threats. However, big data-driven threat intelligence analysis and early warning model construction is a challenging process, which involves various aspects such as understanding and processing of big data, methods, and tools for threat intelligence analysis, design and implementation of early warning models, and validation and evaluation of early warning models. This paper aims to comprehensively describe this process to provide guidance and reference for big data-driven threat intelligence analysis and early warning.

Keywords: big data-driven; threat intelligence; early warning model

Introduction:

Big data is a phenomenon that describes data sets whose volume exceeds the processing power of traditional database applications and therefore requires new processing methods to capture value, discover insights, and make decisions, a phenomenon that stems from the increasing amount of machine-generated data, social media-generated content, and machine learning algorithms that demand data, resulting in massive, complex, and frequently changing data sets. As a result, big data is often characterized by the "4V" model: volume, velocity, diversity, and veracity. Volume represents the size of the data, and for Big Data, we are usually talking about data measured in terabytes or petabytes; velocity describes the speed of data flow or generation, and for many Big Data applications, such as social media monitoring or high-frequency transactions, the speed of data flow is particularly important; diversity refers to the type of data, including structured, semi-structured, and unstructured data, and Big Data environment usually contains various types of data; authenticity represents the quality and accuracy of data, which is crucial for data-based decision making.

1. Trends in Big Data Technology

The development trend of Big Data technology shows extensive and profound changes. On the one hand, the sources and types of data are expanding rapidly, which makes the acquisition, organization, and understanding of data more complex. Due to the rapid growth of data generated by the Internet of Things, social media, mobile devices, etc., Big Data now includes not only traditional structured data, but also unstructured data such as pictures, videos, and audio. Therefore, how to process and utilize these unstructured data to extract useful information and knowledge has become a major development trend of big data technology. On the other hand, big data technology is also developing toward improving processing speed and efficiency. Technologies such as distributed computing, parallel processing, and in-memory computing are changing the way people deal with big data, making it possible to analyze and make decisions in real-time on large-scale, complex

Corresponding Author: Peng Zhang
Guangdong Mechanical & Electrical Polyitechinc，China
Email:282765761@qq.com

data sets. Especially with the development of artificial intelligence and machine learning, automated data processing and intelligent data analysis techniques are gaining attention to find patterns in massive amounts of data, generate insights, and be able to learn and optimize themselves to process and use big data more efficiently. Big data technologies are also moving to the cloud, where data storage, processing, and analysis can be performed in the cloud, greatly reducing hardware investment and maintenance costs and increasing the flexibility and scalability of data processing, and cloud service providers are offering richer and more specialized big data services to meet the needs of different industries and fields. In addition, as the volume of data increases and data usage becomes more widespread, issues such as data leakage, data misuse, and data discrimination are becoming increasingly prominent. Therefore, how to protect personal information and corporate confidentiality and avoid data misuse in the big data environment has become an urgent issue. To this end, many new techniques and methods, such as differential privacy, homomorphic encryption, and blockchain, are being developed and applied to achieve better security and privacy protection in the use of big data (Zhang, & Ren, 2022).

2. Basics of threat intelligence analysis

Threat intelligence analysis is a specialized technique and process that aims to better understand and prevent potential risks that could impact individuals, organizations, or countries by collecting, identifying, evaluating, and interpreting information about various security threats, including attacks from networks, malware, fraud, and any other behavior that could threaten security, operations, or interests [2]. At the core of threat intelligence analysis is a data-driven approach to identifying and predicting threats, and analysts need to gather information from a variety of sources, both public (e.g., news reports, public web forums) and private (e.g., dedicated threat intelligence services, closed web forums, or the dark web), and even internal data collected from their network environment, including various forms such as text, image, audio, and network traffic data, etc. After collecting enough data, threat intelligence analysts need to use a variety of tools and techniques to process, clean, organize, and analyze this data to identify possible threat patterns, behaviors, and trends. Threat intelligence analysis must be able to not only identify existing threats but also predict possible future threats, requiring a deep understanding of threat trends, attacker intent and behavior, and the systems or resources that may be affected. In this way, threat intelligence analysis can provide decision-makers with detailed information and recommendations on how to prevent and respond to security threats and better protect their assets and interests (Chen, 2022).

3. Application of big data in threat intelligence analysis

Big data plays a crucial role in threat intelligence analysis, data is the basis of threat intelligence analysis, and in the modern network environment, security threats come from a variety of sources and forms, involving a large amount of network traffic, system logs, user behavior, social media content, and other data. These data are not only huge in volume but also complex in format, including structured tabular data, semi-structured weblogs, unstructured text, images, etc. Therefore, how to effectively gather threat intelligence from various sources and types of data has become an important issue. In this regard, big data technologies, such as distributed data collection and storage, web crawlers, and data lakes, can collect and manage data more effectively. In the data processing phase, data needs to be cleaned, preprocessed, and organized to improve the quality and usability of the data, as the collected data may have noise, errors, redundancy, and missing values (Hu, 2022). Big data processing tools and methods, such as ETL, data cleaning, and data integration, can process and optimize large amounts of data. The data analysis phase requires the discovery of useful threat patterns and behaviors from the data to identify and predict security threats and involves complex data analysis, pattern recognition, and predictive modeling tasks that become more complex and difficult in the face of large-scale data. However, big data analytics techniques, such as parallel computing, in-memory computing, and distributed computing, as well as machine learning and artificial intelligence, allow for effective data analysis in big data environments. Examples include the use of distributed computing frameworks to process large-scale network traffic data, machine learning algorithms to identify complex threat patterns, deep learning techniques to process unstructured text and image data, and predictive modeling techniques to predict possible future threats. The result interpretation stage is to transform complex data and analysis results into easily understandable and trustable information for decision-makers and the public (Shen, et al.,2022), and big data technologies such as data visualization, analysis interpretation, and result validation can help us to present complex data and analysis results in the form of charts, images, and stories so that decision-makers and the public can understand and trust the results of threat intelligence analysis.

4. Construction of an early warning model

4.1Definition of early warning model

The early warning model is a method or tool that collects and analyzes data to predict and warn early about possible threats or dangers, which can identify potential risks as early as possible and give organizations or individuals time to take measures to prevent or mitigate the impact of risks. Early warning models can be used in a wide range of fields, including but not limited to risk management in financial markets, disease prevention in public health, disaster prevention and control in environmental science, and threat intelligence analysis in cyber security. Early warning models can be classified according to the technology they rely on, the type of data they process, the type of problem they solve, and the form of warning they generate (Liu, & Zhou, 2022), and the technology-based classification can be divided into statistical-based early warning models, machine-learning-based early warning models, and artificial intelligence-based early warning models. Statistical-based models rely on traditional statistical theories and methods, such as time series analysis, regression analysis, etc.; machine learning-based models use sophisticated machine learning algorithms, such as support vector machines, decision trees, neural networks, etc., to automatically learn and extract threat features and patterns from data; and artificial intelligence-based models try to mimic the human way of thinking, such as expert systems, knowledge graphs, and deep learning, etc., to perform more in-depth and comprehensive threat warning.

4.2 Steps of building early warning models based on big data

The problem definition is the starting point of the big data-based early warning model construction, i.e., to clearly define what the problem is to be solved, what type of threat to be warned of, what is the goal of the warning, what is the temporal and spatial scope of the warning, and what are the desired warning effect and criteria (Sun, 2022). After that, various data related to the early warning problem are collected, including historical and real-time data, structured and unstructured data, internal and external data, subjective and objective data, etc. The quality, completeness, diversity, and real-time data will directly affect the effectiveness and accuracy of the early warning model. Next, various data processing methods and tools are used to clean, transform, integrate, and de-select data to improve the quality, consistency, compatibility, and usability of data, and various data analysis methods and tools are used to explore, describe, mine and visualize data to understand the distribution, correlation, patterns, and trends of data, etc. Once the above phases are completed, the model building can be performed to select or design suitable early warning models, such as statistical-based models, machine learning-based models, artificial intelligence-based models, etc., as well as to determine model inputs and outputs, parameters and variables, structures and functions, evaluation and optimization, etc. Similarly, model training is the process of building early warning models based on big data, using training data to train early warning models, such as using supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning, etc. Through model training, early warning models can learn and extract features and patterns of threats from data, so that they can provide early warning of threats. Model application and model updating are then the goals of big data-based early warning model construction, applying the trained and validated early warning models to actual early warning scenarios (Xu, et al., 2022), for example, using real-time warning, batch warning, and dynamic warning. The early warning models are also updated regularly or irregularly based on new data and feedback, e.g., using online learning, migratory learning, incremental learning, etc., so that the early warning models continuously provide effective and accurate warnings to prevent and respond to various threats.

4.3 Validation and Evaluation of early warning models

Model validation is a test of the predictive performance of a model by using independent data sets, such as a test set or a validation set, to confirm whether the model can successfully learn generalized knowledge from the data and perform well on unseen data. During validation, attention is usually paid to whether the model is over- or under-fitted, i.e., whether the model can find general trends in the data and not just learn the noise in the training data, or whether the model is so simple that it cannot learn from the data effectively. To prevent overfitting, model validation is often performed using cross-validation methods, such as K-fold cross-validation, which can make better use of limited data and can yield more robust model validation results (Ren, 2022). Model evaluation, on the other hand, is a quantitative measure of model performance, which is performed by calculating various performance metrics, such as accuracy, recall, F1 score, the area under the ROC curve, etc. Accuracy tells the proportion of samples correctly predicted by the model to the total samples; recall reflects the ability of the model to detect real threats (Zhao , et al., 2021); and the F1 score is the summed average of accuracy and recall, which can be considered in accuracy and recalls while giving a single performance metric; and the ROC curve and AUC-ROC can give the overall performance of the model while considering different threshold settings. More attention is paid to recall in early warning models, with the hope of finding as many potential threats as possible, even if this may result in some false positives. Model optimization based on evaluation results, on the other hand, is based on model validation and evaluation to adjust and optimize the model to get better early warning performance, including adjusting the parameters of the model, choosing different features, changing the complexity of the model, or trying different models and algorithms so that the early warning model provides higher early warning performance and lower false alarm rate while meeting specific needs and constraints (Yang, 2020).

Conclusion

This paper has discussed in detail the whole process of big data-driven threat intelligence analysis and early warning model construction, including the understanding of big data, methods, and tools for threat intelligence analysis, steps of early warning model construction, and validation and evaluation of early warning models. Through in-depth understanding and effective application of these key aspects, big data technologies can be better utilized to analyze and alert threats and improve defense and response capabilities. However, big data-driven threat intelligence analysis and early warning model construction is a dynamic process that requires continuous learning of new knowledge, technologies, and methods to address changing and escalating threats. In addition, attention needs to be paid to ethical, legal, and social issues in the processing and application of big data to respect and protect the rights of individuals and society.

Conflict of Interest

The authors declare that they have no conflicts of interest to this work.

Acknowledgement

This research was funded by:

Guangzhou Science & Technology Plan Project (Grant No. 202102080243)

References

Zhang, Q., & Ren, X. Y. (2022). The innovation path of sludge co-disposal management driven by big data. *Renewable Resources and Circular Economy*, *15*(12), 41–44.

Chen, J. (2022). Research on human-machine optimization strategy for open source threat intelligence analysis in cyberspace. *Information Security and Communication Secrecy*, *2022*(07), 17–24.

Hu, D. (2022). Operational mechanism and development measures of collaborative innovation in hubei universities driven by big data. *Internet Weekly*, 2022*(24),*13-15.

Shen, C., Liu, L., Xu , T., & Xiumei, X. (2022). I. research on dynamic social governance driven by big data in government affairs. *Journal of Nanjing University of Posts and Telecommunications (Social Science Edition)*, *24*(06), 48–57.

Liu, L., & Zhou, X. (2022). Design of automatic user portrait generation model based on big data-driven. *Computer Simulation*, *39*(12), 259–263.

Sun, H. (2022). From mosaic theory to preventive rule: The principle of legal regulation of data-driven investigation. *Journal of Hunan Police Academy*, *34*(06), 14–24.

Xu, Z., Liang, D., & Pan, W. (2022). Research on agricultural pest monitoring technology using the internet of things and big data-driven. *Agricultural Equipment Technology*, *48*(06), 4–6, 18.

Ren, P. Y. (2022). *Threat intelligence analysis for tor network services*. Xi'an University of Electronic Science and Technology.

Zhao, N., Li, L., Liu , Q., & Ye , R. (2021). Threat intelligence analysis and management based on network open source intelligence. *Journal of Intelligence*, *40*(11), 16–22, 73.

Yang, L. (2020). *Analysis and research of dark network threat intelligence*. Qilu University of Technology.

How to Cite: Zhang, P.(2023). Big Data-Driven Threat Intelligence Analysis and Early Warning Model Construction. *Journal of Global Humanities and Social Sciences*, 04(04), 171-175. https://doi.org/10.61360/BoniGHSS232014160804.