A Review of the Application of Big Data Technology in Computer Network Information Security Management

Xiaofei Fang

Greentown Technology Industry Service Group Co., Ltd., Hangzhou 310000, Zhejiang, China

Abstract: ***This paper focuses on the application value of big data technology in network information security management, reviewing its core role in transforming security defense paradigms. Leveraging capabilities in massive data processing, high-speed analysis, and intelligent modeling, big data technology offers a new path for building proactive, intelligent, and collaborative network security defense systems. By systematically integrating multi-source heterogeneous security data and employing intelligent analysis methods, it significantly enhances the breadth and depth of threat situational awareness, enabling precise identification, real-time early warning, and efficient handling of security risks.***

Keywords: Big data technology; Computer network information security; Security management.

1. INTRODUCTION

With its core ability to process data characterized by Volume, Variety, Velocity, and Value, big data technology enables efficient collection, integration, storage, and near-real-time analysis of heterogeneous, multi-source cybersecurity data. By integrating intelligent analysis methods such as machine learning and data mining, big data platforms can uncover latent threat patterns, identify anomalous user behavior, detect advanced attack signatures, and predict security posture evolution, thereby supporting the construction of an adaptive, intelligent, and global proactive defense system. More importantly, its distributed architecture provides elastic scalability and disaster recovery capabilities, allowing security platforms to continuously adapt to ultra-large-scale data and ever-changing attack scenarios. Therefore, in-depth exploration of innovative application mechanisms of big data technology in network information security management holds significant theoretical and practical importance for building next-generation intelligent security defense systems. Xu [1] proposed an innovative graph convolutional network (GCN)-based approach for optimizing healthcare facility design to enhance sustainability. The financial sector has seen parallel developments, with Jiang et al. [2] creating an advanced investment advisory system using deep neural networks for personalized financial guidance, while Yang and Duan [5] contributed to risk management through knowledge graph construction for the US stock market. Network optimization technologies advanced through Tu's [3] Log2Learn system for intelligent real-time log analysis. Environmental health research by Ma et al. [4] provided crucial insights into the relationship between metal exposure in maternal and cord blood and fetal liver function. Computer vision technologies have made notable progress, particularly in Lu et al.'s [6] DeepSPG framework for low-light image enhancement using multimodal learning. Logistics automation benefited from Luo et al.'s [7] novel path planning algorithm that integrates transformer networks with GCNs for intelligent logistics robots. Human resources technology was significantly advanced by Li et al. [8], who optimized resume-job matching through a combination of generative pretrained transformers and hierarchical graph neural networks. In medical diagnostics, Wang et al. [9] developed CPLOYO, an advanced pulmonary nodule detection model employing multi-scale feature fusion. Cross-cultural AI studies by Shan et al. [10] offered valuable insights into large language model applications across different cultural contexts. E-commerce and supply chain management saw substantial AI integration, with Chew et al. [11] developing an AI-powered system for accounting data integration and financial risk assessment, complemented by Saunders et al.'s [12] exploration of AI-driven solutions for supply chain efficiency enhancement. Computer vision applications continued to advance through Guo et al.'s [13] improvements to vehicle detection using enhanced YOLOv8 networks, and Jin et al. [14] achieved breakthroughs in object detection and pose estimation using hybrid task cascade networks.

2. THEORETICAL FOUNDATIONS OF BIG DATA TECHNOLOGY

2.1 Concepts and Characteristics of Big Data Technology

Big-data technology refers to the technical system that effectively collects, stores, manages, analyzes, and mines

data objects that are broad in origin, massive in scale, and complex in structure, thereby extracting valuable information. The data it processes usually reaches the petabyte (PB) or even exabyte (EB) level, far exceeding the capacity limits of traditional database software tools. Moreover, these data sources are extremely diverse, spanning IoT devices, Internet applications, social-media platforms, various sensors, and enterprise business systems. The data types are also exceptionally rich, including not only structured relational data but also large volumes of semi-structured data (e.g., log files, XML/JSON) and unstructured data (e.g., text, images, audio, video). With its distinctive characteristics, big-data technology demonstrates significant application advantages and value across different domains.

Volume. With the widespread adoption and deepening of IoT and Internet applications, the speed and scale of data generation are exploding. Relevant statistics show that the amount of new data created worldwide each day has reached the trillions of bytes: social-media platforms publish and upload billions of messages, images, and videos daily; e-commerce systems continuously generate and accumulate massive transaction records and user-behavior trails; IoT devices are constantly producing large volumes of sensor data. Faced with such an enormous scale, traditional data-processing technologies face severe challenges in storage, computation, and analytical efficiency.

Variety. The broad range of data sources directly determines the complexity and diversity of data types. In addition to conventional structured data, semi-structured and unstructured data are becoming increasingly prominent in big data, covering everything from documents, emails, and web content to multimedia materials and even complex graph data. This diversity greatly enriches the information dimensions contained in the data, but it also raises the technical difficulty of data integration, efficient processing, and in-depth analysis, requiring specialized technical solutions.

Velocity. Emphasizes the high-speed generation, real-time flow, and rapid processing needs of data. In today's ubiquitous connected environment, data is produced at a continuously explosive growth rate: global financial trading systems process millions of transaction logs per second; industrial IoT sensors stream device-status data at millisecond intervals; social-media platforms are flooded with massive user interactions and content updates every minute. This extreme timeliness demands near-real-time responsiveness from data-processing systems, making traditional batch-processing architectures inadequate for scenarios highly sensitive to latency, such as real-time intrusion detection, fraud-transaction interception, and dynamic risk scoring.

2.2 Advantages of Big-Data Technology in the Field of Network Security

At the data collection layer: big-data technologies can efficiently aggregate network data from multiple sources, including user-behavior information, network-traffic data, and device and system security logs. These data span every layer and phase of the network and originate from all kinds of devices, applications, and systems. Leveraging advanced distributed collection technologies (e.g., Flume, Kafka), big-data technologies ensure efficient, complete, and comprehensive data gathering even in extremely large-scale scenarios.

At the data storage layer, big-data technologies generally adopt distributed storage architectures (e.g., HDFS, HBase, Amazon S3) that excel at meeting the storage demands of massive data while markedly improving system reliability and horizontal scalability. Based on distributed file systems, data are scattered across multiple nodes, and redundancy mechanisms such as replication or erasure coding safeguard data security while allowing flexible capacity expansion to accommodate continuous data growth.

In data analysis and processing, big-data technologies demonstrate formidable power, enabling high-speed processing and analysis of vast data sets to rapidly identify potential security threats. By leveraging parallel and distributed computing frameworks (e.g., MapReduce, Spark, Flink), data-processing speeds are multiplied. Integrating advanced algorithms from data mining, machine learning, and deep learning, big-data technologies can unearth hidden patterns and regularities, thereby detecting novel or advanced persistent threats (APTs) that traditional security monitoring methods often miss.

In predictive analytics and visual insights, it can synthesize historical and real-time data streams to effectively forecast the probability and evolution of cybersecurity incidents, providing a basis for proactive defense deployment. Meanwhile, big-data technologies can render complex cybersecurity data intuitively through charts, dashboards, heat maps, and other visual forms, greatly aiding security managers in grasping the overall situation and performing in-depth analysis. Via interactive visual interfaces, security personnel can quickly understand the overall health of network security and pinpoint potential risks.

3. APPLICATION MECHANISMS OF BIG-DATA TECHNOLOGY IN COMPUTER NETWORK INFORMATION SECURITY MANAGEMENT

3.1 Data Collection

In computer network information security management, efficient and comprehensive data collection is the primary link in applying big-data technology, focusing on capturing relevant data in real time and in full from diverse network data sources.

Network traffic data is one of the most fundamental and critical monitoring objects. By deploying dedicated network traffic monitoring tools (such as NetFlow, sFlow, packet sniffers, etc.), the contents of packets traversing network links can be captured and analyzed in real time. These tools can deeply parse key elements such as source/destination IP addresses, port numbers, protocol types, and even payload characteristics, thereby providing administrators with a global insight into the state of data flows across the entire network, potential anomalies, and the nature of the traffic. In enterprise networks, continuous traffic collection and analysis are of great significance for rapidly identifying abnormal traffic patterns.

User behavior data records users' operational trajectories and activity characteristics within the network, serving as a crucial basis for establishing security baselines and identifying insider threats. It meticulously logs details such as login times, geographic locations, accessed resources, operation frequencies, and command execution histories. Through cumulative analysis of this data, fine-grained user behavior profiles and dynamic baseline models can be constructed, laying a solid foundation for detecting suspicious activities that deviate from normal patterns.

Security log data is a key carrier for recording system security events, typically generated by critical security devices such as firewalls, operating systems, intrusion detection/prevention systems (IDS/IPS), routers/switches, and antivirus gateways. It faithfully records various security-related events, including but not limited to permission change attempts, frequent login failures, critical system error alerts, and policy modification behaviors. These detailed log entries form a valuable evidence chain for tracing the root causes of security incidents, conducting diagnostic analysis, and post-incident forensics. Their comprehensiveness and accuracy are vital for understanding attack chains and assessing the scope of impact.

3.2 Data Integration

After extensive multi-source data collection, the complex data integration process becomes a key step to ensure analytical effectiveness. The core task of data integration is to cleanse data from different channels with varying structures and formats, ultimately providing a unified, accurate, and complete high-quality data foundation for subsequent security analysis. Due to inherent differences among data sources, the first step is data format conversion and standardization. This process aims to transform heterogeneous data into a consistent, easily processed format or data model, significantly improving the efficiency of subsequent data storage, retrieval, and analysis.

Data cleaning is decisive for ensuring data quality and the reliability of analytical results. Its core task is to detect and eliminate errors (such as format mistakes, missing fields, or logical contradictions), redundant copies, and various noise disturbances. Security log data is especially prone to containing large amounts of such dirty data, which may stem from misconfigurations, brief device failures, or inappropriate alert thresholds. The presence of these invalid or misleading pieces of information can severely interfere with the judgment of security analysis engines and even cause machine-learning models to misclassify. Therefore, professional data-cleaning tools or customized rule/script strategies (e.g., rule-based cleaning, fuzzy-matching deduplication) must be applied to rigorously deduplicate, correct, and filter out invalid information, ultimately ensuring data trustworthiness and the accuracy of analytical results.

3.3 Data Analysis and Mining

Predictive analytics occupies a central role in network security situational awareness. It relies on a comprehensive assessment of historical security-event data (including attack types, occurrence times, target systems, attack vectors, etc.) and current real-time network-state information (such as traffic fluctuations, abnormal connections, configuration changes, etc.), employing predictive algorithms like time-series analysis (ARIMA, LSTM, etc.) and

regression analysis to build mathematical models. These models can prospectively forecast the probability of future security events within a specific time window, potential attack targets, and the most likely attack types, providing security teams with a valuable early-warning window.

Association-rule mining is another highly valuable data-analysis technique that excels at uncovering hidden relationships among different network-security events. For example, when analyzing vast network-log data, association-rule mining algorithms (Apriori, FP-Growth, etc.) can reveal strong correlations between key event patterns such as "frequent login failures" and "abnormally rapid IP address switching for a specific account" and the conclusion that "the account is at high risk of compromise." Consequently, when the system detects multiple login attempts for an account from different geographic locations within a very short time, it can promptly infer—based on the above rules—that the account may have been hijacked and immediately trigger predefined security-response procedures, such as enforcing multi-factor authentication (MFA) or temporarily freezing account access. This proactive, rule-based analysis effectively reduces the risk of account theft and enhances overall security.

3.4 Security Prediction and Early Warning

Security forecasting and early warning are key initiatives for big-data-driven proactive defense in network information security management. Their essence lies in building high-precision security prediction models through data analysis and establishing a rapid early-warning mechanism that works in concert with them. This proactive capability enables the system to identify the signs and evolution trends of potential threats earlier—before they fully surface or inflict real damage—thereby providing strong support for targeted, pre-emptive security measures (such as configuration optimization, rule updates, and resource allocation) and ultimately reducing security risks and ensuring the continuous, stable operation of network information systems.

4. CONCLUSION

4.1 Reconstructing the Security Defense Line with Core Capabilities

Big-data technology, leveraging its ability to aggregate massive volumes of data (spanning traffic, logs, user behavior, IoT, OSINT, and other multidimensional sources), the elastic processing power of distributed architectures, and intelligent analytics engines (machine learning, association-rule mining, predictive models), has enabled a threefold leap in security management. First, from reactive response to proactive prediction: time-series analysis and deep-learning models anticipate attack trends. Second, from localized protection to global collaboration: multi-source data are fused to build a unified security view. Third, from rule dependence to intelligent decision-making: behavioral-baseline models and real-time anomaly detection enable dynamic defense.

4.2 Challenges and Future Directions

Although big-data technologies have revolutionized cybersecurity, their practical deployment still faces significant bottlenecks. First, the real-time fusion of heterogeneous data from diverse devices and systems—such as traffic, logs, and sensor feeds—presents efficiency hurdles; the massive, heterogeneous data torrent is hard to integrate and analyze rapidly enough to counter instantaneous threats. Second, accurately extracting faint security-threat signals from extremely low-value-density data is highly challenging; critical attack clues are often drowned in overwhelming noise, leading to high false-positive and false-negative rates. Finally, AI-based analytic models themselves are exposed to security risks: attackers can craft sophisticated adversarial examples to deceive the models and render them ineffective, revealing latent vulnerabilities in existing defenses.

To address these challenges and unlock the full potential of big-data security, innovative technologies must be integrated to open new paths. Future efforts should focus on: (1) pushing analytic capabilities closer to data sources (edge computing) and exploring collaborative learning mechanisms (federated learning) to process data locally, filter out useless information, and drastically reduce transmission and analysis loads while preserving privacy; (2) deepening the mining of deep semantic relationships in data—e.g., by combining knowledge graphs with semantic analysis—which is crucial for complex sources such as dark-web intelligence and unstructured logs, so that hidden threats can be uncovered through their intrinsic associations. The ultimate goal is to build an "adaptive security architecture" that evolves dynamically and self-adjusts, whose core defense strategies and models continuously advance alongside evolving attack techniques, ensuring sustained effectiveness of protection.

REFERENCES

[1] Xu, Haoran. "Sustainability Enhancement in Healthcare Facility Design: Structural and Functional Optimization Based on GCN." (2025).

[2] Jiang, Gaozhe, et al. "Investment Advisory Robotics 2.0: Leveraging Deep Neural Networks for Personalized Financial Guidance." (2025).

[3] Tu, T. (2025). Log2Learn: Intelligent Log Analysis for Real-Time Network Optimization.

[4] Ma, Haowei, et al. "Maternal and cord blood levels of metals and fetal liver function." Environmental Pollution 363 (2024): 125305.

[5] Yang, Wei, and Jincan Duan. "Knowledge Graph Construction for the US Stock Market: A Statistical Learning and Risk Management Approach." Journal of Computer Technology and Applied Mathematics 2.1 (2025): 1-7.

[6] Lu, Jialang, et al. "DeepSPG: Exploring Deep Semantic Prior Guidance for Low-light Image Enhancement with Multimodal Learning." arXiv preprint arXiv:2504.19127 (2025).

[7] Luo, H., Wei, J., Zhao, S., Liang, A., Xu, Z., & Jiang, R. (2024). Intelligent logistics management robot path planning algorithm integrating transformer and gcn network. IECE Transactions on Internet of Things, 2(4), 95-112.

[8] Li, Huaxu, et al. "Enhancing Intelligent Recruitment With Generative Pretrained Transformer and Hierarchical Graph Neural Networks: Optimizing Resume-Job Matching With Deep Learning and Graph-Based Modeling." Journal of Organizational and End User Computing (JOEUC) 37.1 (2025): 1-24.

[9] Wang, Meng, et al. "CPLOYO: A pulmonary nodule detection model with multi-scale feature fusion and nonlinear feature learning." Alexandria Engineering Journal 122 (2025): 578-587.

[10] Shan, X., Xu, Y., Wang, Y., Lin, Y. S., & Bao, Y. (2024, June). Cross-Cultural Implications of Large Language Models: An Extended Comparative Analysis. In International Conference on Human-Computer Interaction (pp. 106-118). Cham: Springer Nature Switzerland.

[11] Chew, J., Shen, Z., Hu, K., Wang, Y., & Wang, Z. (2025). Artificial Intelligence Optimizes the Accounting Data Integration and Financial Risk Assessment Model of the E-commerce Platform. International Journal of Management Science Research, 8(2), 7-17.

[12] Saunders, E., Zhu, X., Wei, X., Mehta, R., Chew, J., & Wang, Z. (2025). The AI-Driven Smart Supply Chain: Pathways and Challenges to Enhancing Enterprise Operational Efficiency. Journal of Theory and Practice in Economics and Management, 2(2), 63–74. https://doi.org/10.5281/zenodo.15280568

[13] Guo, Haocheng, Yaqiong Zhang, Lieyang Chen, and Arfat Ahmad Khan. "Research on Vehicle Detection Based on Improved YOLOv8 Network." Applied and Computational Engineering 116 (2025): 161-167.

[14] Jin, Yuhui, Yaqiong Zhang, Zheyuan Xu, Wenqing Zhang, and Jingyu Xu. "Advanced object detection and pose estimation with hybrid task cascade and high-resolution networks." In 2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML), pp. 1293-1297. IEEE, 2024.