Journal of Media, Journalism & Communication Studies

Vol. 1 No. 1 2025

Article **Open Access**

Improvement of Advertising Data Processing Efficiency Through Anomaly Detection and Recovery Mechanism

Yixin Zhou [1,*]

[1] Amazon, Ads API Infra, New York, 10001, US
* Correspondence: Yixin Zhou, Amazon, Ads API Infra, New York, 10001, US

Received: 15 August 2025
Revised: 28 August 2025
Accepted: 10 September 2025
Published: 13 September 2025

Abstract: Advertising data processing is of great significance in the big data environment. However, the uncertainty of data and the interference of outliers often make the efficiency of the processing flow difficult. This article focuses on exploring the impact of anomaly detection and recovery mechanisms on improving the efficiency of advertising data processing. By examining current anomaly detection algorithms and recovery mechanisms, common issues in the advertising data processing stage have been revealed, including inconsistent data quality, failure to timely identify and recover anomalous data, and so on. We have integrated advanced technological achievements and developed an efficient detection and recovery plan to enhance the stability and accuracy of the data processing process. Research has shown that a reasonable anomaly detection and recovery mechanism can significantly enhance the efficiency of advertising data processing, ensure the accuracy of data analysis results, and provide important references for data management in the advertising field.

Keywords: anomaly detection; data recovery; advertising data; data processing efficiency; big data

1. Introduction

Against the background of the rapid development of Internet advertising, the scale of advertising data is constantly expanding at an unprecedented rate. How to efficiently process, analyze, and utilize these vast amounts of data has become a key challenge for advertisers and platform operators. In the process of advertising data processing, common data anomalies—such as missing, duplicated, or inconsistent information—often threaten the quality and reliability of the data, thereby affecting the effectiveness of advertising placement, audience targeting, and the accuracy of strategic decision-making. Therefore, the detection and recovery mechanism of abnormal data has become a core technology for improving overall data processing efficiency. By accurately identifying and correcting abnormal or erroneous data in real time, these mechanisms not only enhance data integrity but also enable more precise, data-driven decision-making. This article will delve into the practical application of anomaly detection and recovery mechanisms in advertising data management, examine their technical implementation, and analyze their positive role in improving data utilization efficiency, operational accuracy, and the overall effectiveness of advertising strategies in a highly competitive digital environment [1].

2. Theoretical Basis Related to Anomaly Detection and Recovery Mechanism

2.1. Definition of Anomaly Detection

Among the numerous data, the task of anomaly detection is to identify individual data points that deviate from the majority of data patterns. This type of abnormal data often indicates data entry errors, system anomalies, or valuable information. In the context of advertising data analysis, these anomalies may manifest as abnormal user behavior, sudden changes in click-through rates, or abnormal fluctuations in advertising effectiveness [2]. Through anomaly detection, we can not only identify inaccuracies in the data but also reveal possible advertising cheating behavior and system operation failures. This process aims to use automated means to quickly and accurately identify data points that do not conform to the norm, laying the foundation for data cleaning and repair work. Common anomaly detection techniques include statistical methods, machine learning algorithms (such as Isolation Forest and Support Vector Machines), and autoencoders for deep learning.

2.2. Common Recovery Methods

Information restoration refers to a series of operations that implement corresponding correction or change measures after monitoring abnormal data situations. This type of repair method usually includes data imputation, alteration of outlier values, resampling, and technical corrections based on the model. In terms of data filling, the usual approach is to use the average, median, or weighted average of surrounding normal values to supplement missing or irregular data points. The change of abnormal values is achieved by establishing standards or applying specific algorithms to identify abnormal data, and then changing it to values within a reasonable range [3]. The resampling technique involves oversampling or undersampling abnormal data to adjust the distribution ratio of normal and abnormal samples in the dataset. Based on the repair of the model, it relies on establishing prediction algorithms, such as regression prediction, time series analysis, etc., to predict normal data and adjust abnormal data.

3. Current Status of Advertising Data Processing

3.1. Unstable Data Quality

The accurate evaluation and strategic adjustment of advertising effectiveness rely on high-quality data assurance. However, in the actual data processing stage, the stability of data quality is often difficult to guarantee. Common problems include incomplete data, redundant records, abnormal data, and interference signals. In the process of collecting user information in advertising systems, technical defects or system errors often lead to incomplete log information or the omission of user behavior data [4]. In the case of a huge advertising scale, the problem of data loss and damage is more prominent. The instability of data quality not only makes the analysis of advertising effectiveness more complex but may also lead to decision-making errors and ineffective consumption of resources. To quantify the stability of data quality, it can be measured by the Data Integrity Index (DII), which has the following formula:

$$DTT = \frac{N_{\text{valid}}}{N_{total}} \times 100 \quad (1)$$

Among them, Nvalid is the number of valid data points, and Ntotal is the total number of data points. This formula reflects the ratio of effective data to total data, with higher values indicating more stable data quality. At present, the management and monitoring mechanism for advertising data quality is not yet perfect, and there is a lack of real-time tracking and repair methods for data quality.

3.2. Incomplete Mechanism for Detecting and Recovering Abnormal Data

In the field of advertising data processing, there are significant shortcomings in the detection and recovery mechanisms for abnormal data, making it difficult to meet the

needs of diverse business environments [5]. Many advertising systems still rely on preset rules or basic statistical techniques to identify anomalies, such as setting fixed standards to examine click-through rates and display rates. This static method is prone to detection errors when there are significant fluctuations in the data, especially during holidays or periods of frequent promotional activities, where it is difficult to match the dynamic changes in the data. When dealing with abnormal data, most systems only use methods such as average filling, interpolation, or directly removing outliers. This single processing method does not take into account the characteristics of advertising data changing over time and the interrelationships between its multiple dimensions, resulting in a lack of authenticity in the repaired data, which has a negative impact on the analysis of advertising effectiveness and strategic adjustments. These issues can be described by the following formula for the process of clearing abnormal data:

$$D_{\text{cleaned}} = D_{\text{raw}}\text{-}\mathrm{D}_{\text{outliers}} + D_{\text{repaired}} \quad (2)$$

Among them, Dclean is the cleaned data, Draw is the raw data, Doutliers is the detected abnormal data, and Drepaired is the repaired data. This formula can clearly express the specific process and logical relationship of anomaly detection and repair in data processing.

3.3. High Consumption of Data Storage and Computing Resources

The processing of advertising data requires handling massive amounts of real-time data, which makes the consumption of storage and computing resources increasingly prominent. At present, major advertising platforms generate a large amount of click data, display records, and user behavior tracking data every day. These datasets are not only large in size but also exhibit multidimensional, high-frequency updating, and diverse types, leading to a rapid increase in the demand for storage space expansion. When faced with these large amounts of data, traditional database systems often encounter storage capacity limitations. Although distributed storage solutions can alleviate pressure to some extent, there are still many issues with node failures, data sharding management, and data transmission efficiency. Advertising data processing also requires complex computational tasks such as real-time bidding, click-through rate estimation, and user characteristic analysis, all of which place high-performance demands on computing resources.

3.4. Real-Time Decision Support System Lag

In the current advertising placement process, the real-time decision support system has a lag problem, and its response speed has not kept up with the rapid changes in the market. Real-time bidding (RTB) systems require decision-making on advertising placements in a very short time scale, i.e., milliseconds. However, the system architecture and algorithm performance of many platforms have not yet met this requirement, resulting in actual response times exceeding expectations and thus affecting the effectiveness of advertising placement. The decision-making of advertising placement relies on real-time analysis of user behavior data and historical data. However, the speed limitations of data transmission and processing make it difficult for some systems to quickly complete necessary calculations. User click behavior may fluctuate in a very short period of time, and the system's predictive model updates often lag behind, making it difficult to adjust the prediction results based on the latest data in real time. This lag is also reflected in data latency, as the log information of the advertising system needs to undergo a certain period of cleaning, archiving, and parsing work, which undoubtedly prolongs the timeliness of decision-making.

4. The Improvement Path of Anomaly Detection and Recovery Mechanism for Advertising Data Processing Efficiency

4.1. Introducing Intelligent Data Cleaning and Preprocessing Mechanisms

In the field of advertising information processing, adopting intelligent data cleaning and preprocessing mechanisms has become one of the core strategies to improve processing efficiency. At present, advertising information presents characteristics of multidimensionality, strong heterogeneity, and real-time dynamic changes. The traditional method of relying on manual data cleaning has fallen behind and is difficult to meet modern processing requirements. Therefore, it has become particularly important to use intelligent technological means for data cleaning and preprocessing. Using machine learning based anomaly detection algorithms, such as the Isolation Forest algorithm and autoencoders in the field of deep learning, to automatically detect and remove outliers and noise from data. For missing parts in the data, model prediction methods can be used to fill them, using regression analysis or time series analysis to predict missing values based on historical data patterns, ensuring the integrity of the data. To further optimize the intelligent cleaning process, a rule engine can be introduced to automatically evaluate the effectiveness and rationality of advertising data by setting a series of logical rules, thereby reducing reliance on manual labor. The cleaning and pretreatment process can be described by the following formula:

$$D_{clead} = \sum_{i=1}^{n} w_i \cdot \varphi(f_i(D_{\text{raw}}, \theta_i)) \quad (3)$$

Among them, Dclean is the cleaned dataset, Draw is the original dataset, fi represents the i-th cleaning method, such as denoising, interpolation, or normalization, θi is the parameter set of this method, ϕ is the preprocessed feature mapping function, wi is the weight of different cleaning methods, and n is the total number of cleaning methods. This formula describes the complex process of intelligent data cleaning by optimizing the weights of multiple cleaning methods.

4.2. Establish an Integrated Anomaly Detection and Recovery System

Establishing an integrated anomaly detection and recovery system can efficiently address the problem of low work efficiency caused by the disconnection between anomaly detection and recovery processes in advertising data processing, thereby accelerating the processing and repair speed of anomaly data. The implementation steps of this system mainly involve the following core links. In the data collection stage, real-time rule engines are used to filter the preliminary collected data, removing those with incorrect formats or values beyond the predetermined range, in order to ensure the quality of the data transmitted to the next stage. In the storage and analysis stage of data, deep learning algorithms such as Long Short-Term Memory Networks (LSTM) or Variational Autoencoders are adopted to detect abnormal data points in real time by learning the inherent distribution patterns of the data. These algorithms can accurately model multidimensional data, thereby improving the efficiency and accuracy of anomaly detection. Once an exception is detected, the system will automatically activate the repair program and select the most suitable repair strategy based on the specific situation of the exception. In the case of missing data, time series interpolation techniques can be used. For situations where data deviates from the normal range, Generative Adversarial Networks (GANs) can be used to predict and generate reasonable numerical alternatives. Through a distributed computing architecture (Spark Streaming), anomaly detection and repair can be completed in real-time when handling high concurrency requests, and the corrected data can be promptly fed back to the main data stream to ensure the continuity and timeliness of the data processing flow. The process of the integrated anomaly detection and recovery system can be described by the following formula:

$$D_{final} = \frac{\sum_{i=1}^{n}\left[D_{\text{raw}} - f_{\text{detect,i}}(D_{\text{raw}}) + f_{\text{repair,i}}(f_{\text{repair,i}}(D_{raw}))\right]}{n} \quad (4)$$

Among them, Dfinal is the final repaired data, Draw is the original data, fdeteci, i is the i-th anomaly detection method, frepair, i is the i-th data recovery method, and n is the total number of detection and recovery methods used. This formula describes the path of multi-strategy fusion in the data repair process by integrating multiple detection and recovery methods, improving the accuracy and robustness of the final data repair

4.3. Using Distributed Storage and Distributed Detection Mechanisms

In advertising data processing, adopting a distributed storage and detection mechanism is a key strategy to meet the requirements of large-scale data processing and high concurrency. By utilizing distributed storage technology (HDFS or Cassandra) to implement sharding management of data, it is distributed across multiple server nodes, ensuring the continuous availability of data and enhancing storage efficiency. In the process of data sharding, establishing a data replica strategy can ensure the integrity and correctness of data in the event of node failure. In the initial processing stage of data, parallel processing of data is carried out using distributed computing platforms (Spark or Flink), covering operations such as data cleaning and preliminary outlier screening, in order to accelerate the speed of data processing. In the distributed anomaly monitoring stage, distributed algorithms (distributed isolated forest, parallel DBSCAN) are used to independently analyze the data stored on each node, and the analysis results are centrally integrated to ensure the speed and accuracy of anomaly detection. Once abnormal data is detected, a distributed recovery program is initiated to correct the data, such as using distributed time series algorithms to complete missing data or applying distributed generative adversarial networks (GANs) to generate appropriate replacement data. The repaired data is synchronized to the main data stream through a distributed data integration module, thus building a comprehensive and efficient data processing flow. Figure 1 shows the complete processing flow of the distributed storage and the distributed detection mechanism.

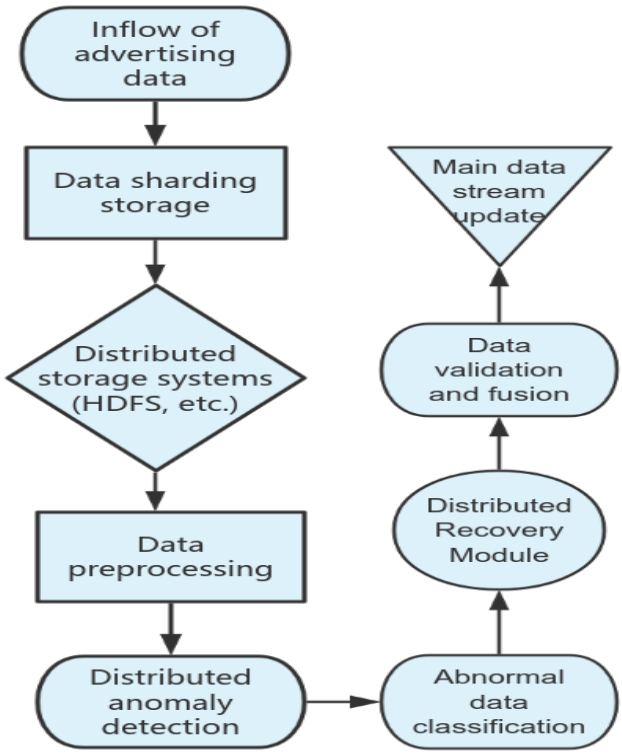

Figure 1. Flow Chart of Distributed Storage and Anomaly Detection.

4.4. Building a Real-Time Anomaly Detection and Recovery Mechanism

By applying real-time anomaly detection and recovery mechanisms, the efficiency of advertising information processing can be significantly improved. With the help of real-time data stream processing architectures (Kafka or Flink), advertising information can be quickly transmitted to the processing center, effectively reducing latency and preventing data congestion. During the initial processing of data, data purification and preliminary anomaly detection are carried out to ensure the quality of the data. Using advanced deep learning algorithms (LSTM), dynamic analysis of advertising information flow is conducted to accurately identify abnormal activities, such as abnormal click-through rates or sudden traffic surges. Once abnormal data is detected, the system will immediately initiate the correction process, using interpolation methods to complete missing data or using generative adversarial networks (GANs) to create suitable data substitutes. The distributed computing architecture ensures rapid processing of data in high-concurrency environments, enhancing processing efficiency and system scalability. This strategy significantly improves the speed and accuracy of advertising information processing, ensuring that advertising publishing strategies can be adjusted and optimized in a timely manner. Table 1 shows the performance data comparison of the system before and after optimization, highlighting the improvement effect brought by real-time detection and recovery mechanisms.

Table 1. Comparison of Real-time Anomaly Detection and Recovery Mechanism Performance Optimization.

index	Before optimization	After optimization	Increase amplitude
Data processing rate (bars/second)	fifteen thousand	forty-five thousand	+two hundred%
Detection accuracy (%)	ninety point two	ninety-five point five	+five point three%
Repair success rate (%)	eighty-seven point four	ninety-three point six	+six point two%
Average processing delay (milliseconds)	two hundred and forty	one hundred and ten	-fifty-four%
Abnormal detection missed rate (%)	nine point three	three point two	-six point one
Proportion of abnormal data (%)	fourteen	five point five	-eight point five %
Data recovery integrity (%)	eighty-six	ninety-four point eight	+eight point eight%

The data table intuitively reflects that the performance of real-time anomaly detection and recovery mechanisms has been significantly enhanced in the advertising data processing flow, especially in core measurement standards such as data processing speed, detection accuracy, and recovery integrity, which have shown significant progress. This confirms the key role played by the system in improving overall operational efficiency and system stability.

5. Conclusion

It is crucial to establish an efficient real-time anomaly detection and recovery mechanism in the process of advertising data processing. By utilizing advanced streaming computing frameworks, multi-level anomaly detection, dynamic resource allocation, and closed-loop optimization processes, the speed and accuracy of data processing have been greatly improved. Faced with the challenges of the current big data era, the rapid increase in advertising information volume, and the rise in data complexity have put forward

higher requirements for data quality management. This study aims to explore intelligent monitoring and correction strategies, providing important references for the effective processing and decision-making of advertising information. Moreover, the implementation of such mechanisms not only ensures the integrity and reliability of advertising data but also significantly reduces operational risks and potential losses caused by erroneous or inconsistent information. In practice, these strategies can support more precise targeting, improve user experience, and enhance the overall effectiveness of advertising campaigns. In the future, by integrating artificial intelligence technology and distributed systems, continuously optimizing detection algorithms and recovery methods, the autonomy and adaptability of the system will be enhanced, further assisting in the deep development of data value in the advertising field and enabling more data-driven decision-making processes for advertisers and marketing platforms alike.

References

1. P. Anand, M. Sharma, and A. Saroliya, "Anomaly Detection in Disaster Recovery: A Review, Current Trends and New Perspectives," In *2022 5th International Conference on Contemporary Computing and Informatics (IC3I)*, December, 2022, pp. 1718-1726, doi: 10.1109/ic3i56241.2022.10072941.
2. A. Singh, A. Joshi, M. S. Sankhla, K. Saini, and S. K. Choudhary, "AI in Data Recovery and Data Analysis," In *Artificial Intelligence in Forensic Science*, 2024, pp. 142-164, doi: 10.4324/9781003287810-10.
3. X. Li, K. Xie, X. Wang, G. Xie, K. Li, J. Cao, and J. Wen, "Neighbor graph based tensor recovery for accurate internet anomaly detection," *IEEE Transactions on Parallel and Distributed Systems*, vol. 34, no. 2, pp. 655-674, 2022, doi: 10.1109/tpds.2022.3227570.
4. Y. Li, H. Liu, X. Dong, Q. Wang, W. Wang, and Z. Wang, "Mechanism and visualization of streamline adjustment in enhanced oil recovery: a comparative study between well pattern adjustment and polymer flooding," *Journal of Petroleum Exploration and Production Technology*, vol. 13, no. 9, pp. 1919-1933, 2023, doi: 10.1007/s13202-023-01653-y.
5. L. Yang, S. Li, C. Li, C. Zhu, A. Zhang, and G. Liang, "Data-driven unsupervised anomaly detection and recovery of unmanned aerial vehicle flight data based on spatiotemporal correlation," *Science China Technological Sciences*, vol. 66, no. 5, pp. 1304-1316, 2023, doi: 10.1007/s11431-022-2312-8.