Scientific Journal of Intelligent Systems Research Volume 7 Issue 8, 2025
ISSN: 2664-9640

Research on Intelligent Analysis and Visualization Technology of Big Data for Massive Heterogeneous Data Processing

Shuyu Zhang

Donghua University, Shanghai, China

Abstract

The research puts forward solutions from three levels: in the aspect of heterogeneous data fusion, a semantic alignment model based on dynamic Knowledge Graph (KG) is constructed, and semantic conflicts are solved through multimodal ontology modeling and dynamic entity alignment mechanism (fusion of structure, semantics and temporal similarity); On the level of intelligent analysis, a collaborative hybrid computing framework based on improved Lambda architecture is designed, which combines real-time processing of sliding window and deep mining of batch data lakes, and realizes the optimization of results through dynamic weight fusion. In the visual interaction, a three-tier engine integrating progressive rendering and predictive loading is developed. By using quadtree index, dynamic level of detail (LOD) and LSTM behavior prediction, the interaction delay at 1080P resolution is controlled within 100ms. The experimental results show that the cross-modal link accuracy rate of this technical scheme is 92.7%, the semantic conflict resolution rate is 91.2%, the batch processing delay is stable at 210±15ms, the throughput is increased by 18.75%, and the visual interaction delay is significantly reduced, which achieves a performance breakthrough in theory and application.

Keywords

Heterogeneous Data; Intelligent Analysis; Visual Interaction; Big Data; Knowledge Graph.

1. Introduction

With the deep integration of 5G communication, Internet of Things (IoT) and AI technology, the global data volume is growing exponentially. More than 80% of the data exists in unstructured or semi-structured forms such as text, images, videos and sensor logs, showing obvious multi-source heterogeneity characteristics [1-2]. For example, in a smart city system, traffic camera video streams, social media texts, weather sensor values and other types of data need to be processed at the same time; Cross-modal information such as equipment vibration signals, production logs and quality inspection images should be integrated in the industrial Internet scene [3]. The significant differences in format, semantics and time series characteristics of this kind of data lead to the double challenges of semantic gap and computational complexity in the traditional analysis framework based on a single data source.

At present, there are core contradictions in big data processing technology at three levels: data fusion, intelligent analysis and visual interaction. In terms of data fusion, traditional ETL (Extract-Transform-Load) tools are difficult to deal with the semantic conflict of dynamic heterogeneous data sources, such as the information loss caused by the difference in the coding of the same disease in medical electronic medical record system [4]; In intelligent analysis, the independent use of batch processing and stream processing framework limits the collaborative optimization of real-time and historical data analysis, and increases the error rate in some forecasting tasks [5]; On the visual interaction level, the existing tools either can't support the dynamic exploration of large-scale data, or face the problems of rendering delay and equipment

ISSN: 2664-9640

compatibility, which leads to poor performance in scenes that need rapid response, such as the long response time for identifying abnormal trading patterns in financial risk control.

In this study, a dynamic semantic alignment model based on Knowledge Graph (KG) is proposed to solve the semantic conflict of heterogeneous data, and a hybrid computing framework of stream and batch collaboration is designed to realize collaborative optimization of analysis and mining, and a visualization engine integrating progressive rendering and predictive loading is developed to control the interaction delay within 100ms at the resolution of 1080P, so as to achieve overall performance breakthrough in theory, method and system.

2. Core Technical Method

2.1. Heterogeneous Data Fusion

In order to solve the semantic conflict of heterogeneous data sources, this study proposes a semantic alignment model based on dynamic KG, and constructs a multimodal ontology model as the core of the fusion architecture. By defining the core classes such as devices, events and geographical locations and their attribute relationships, the model establishes cross-modal data mapping rules, realizes the unified modeling of different types of data such as sensors and images, and provides a structural basis for subsequent semantic integration.

On the basis of ontology modeling, the system introduces dynamic entity alignment mechanism, and carries out entity matching by integrating the similarity of structure, semantics and time series [6]. Weighted similarity function is used to measure attribute matching degree, cross-modal semantic embedding similarity and time sequence consistency respectively, in which each weight parameter can be initialized according to the application scenario and support subsequent dynamic adjustment to improve alignment accuracy. This method enhances the alignment reliability through time-space constraints, such as reducing the matching weight of entities across time periods by using Gaussian attenuation function with time offset, and capturing deep semantic association by combining with cross-modal BERT model, thus effectively coping with the challenges brought by diverse data sources, heterogeneous formats and dynamic evolution, and providing robust semantic alignment capability for data fusion in complex scenes [7].

2.2. Intelligent Analysis Engine

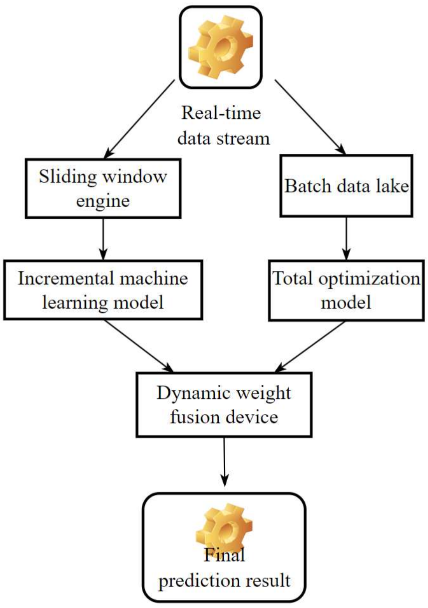

Figure 1. Flow-batch collaborative hybrid computing framework

ISSN: 2664-9640

Build a flow batch collaborative hybrid computing framework based on the improved Lambda architecture, giving consideration to real-time and accuracy (as shown in Figure 1). In this framework, real-time data streams are connected to the sliding window engine for low-latency processing, and at the same time, all the data are stored in the batch data lake for deep mining, which respectively drive the incremental learning model and the total optimization model. Finally, the results are integrated by the dynamic weight fusion device, and the advantages of stream processing and batch processing are complementary [8].

In the aspect of streaming computing, the framework adopts the sliding window aggregation mechanism with time attenuation to extract the features of the data stream $S=\{x_1,x_2,\cdots,x_n\}$ in the window:

$$F_w(S)=\sum_{i=1}^{n}\phi(x_i)w(|t_c-t_i|) \tag{1}$$

$\phi(\cdot)$ is the characteristic projection function as the time attenuation weight, which ensures that recent events get higher attention and improves the response ability of the model to dynamic changes. $w(\Delta t)$ is the time attenuation weight, t_c is the current timestamp, and t_i is the event occurrence time.

In order to ensure the long-term accuracy of the model, the system introduces batch results to periodically correct the flow model, realizes parameter fusion by Formula (2), and corrects the drift of the flow model by using high-precision total model, thus continuously optimizing the prediction performance while maintaining low delay, and forming an intelligent analysis engine with closed-loop optimization.

$$\theta_{stream}=\theta_{stream}+\eta(\theta_{batch}-\theta_{stream}) \tag{2}$$

Where $\eta=0.05$ is the fusion rate of model parameters.

2.3. Visual Interactive System

In order to solve the problem of high delay and interaction jam in PB-level data visualization, a visual interactive system based on progressive rendering is designed, and the core is to build an efficient and intelligent three-layer rendering pipeline (as shown in Figure 2). The system optimizes the whole process from data organization, preloading strategy to graphics rendering, realizes the smooth interactive experience of large-scale complex data at 1080P resolution, ensures that the end-to-end response time is controlled within 100ms, and meets the strict requirements of real-time decision-making scenarios.

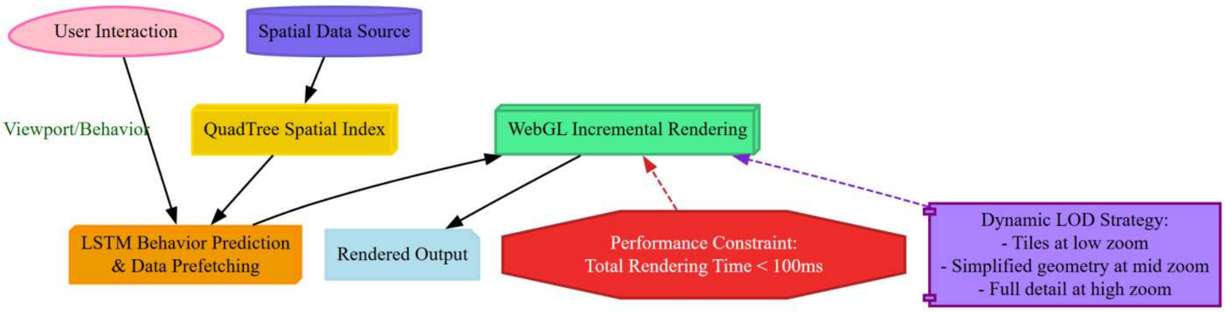

Figure 2. Visual interactive system

Firstly, the system manages the massive geographic or topological data in a structured way through the spatial division technology, and uses QuadTree index to slice the data by regions, thus realizing rapid positioning and local updating. When users zoom or pan, the system only loads the data tiles within the viewport, which significantly reduces the data volume of a single rendering. Combined with the dynamic level of detail (LOD) mechanism, the system automatically selects the rendering accuracy according to the zoom level of the current view: light tile is displayed at low zoom level, simplified geometry is used at medium zoom level, and the complete detail model is loaded at high zoom level, which ensures the visual quality and maximizes the performance efficiency [9].

In order to further shorten the data acquisition time, the system introduces a preloading mechanism based on user behavior prediction. The LSTM model is used to analyze the historical interaction sequence H_t (such as zooming, panning, clicking, etc.), predict the viewing area that users may visit next, calculate the prefetch probability $P(load_i|H_i)$, and load relevant data from the back end in advance.

$$P(load_i|H_i) = \text{softmax}(W_h LSTM(H_i) + b) \tag{3}$$

Finally, the system uses WebGL to call the parallel computing capability of GPU to achieve incremental rendering, and optimizes the data acquisition and GPU rendering time, and achieves a near real-time visual interactive experience with high resolution under the constraint of T_fetch+T_render < 100ms.

3. Experimental Design and Verification

3.1. Configuration of Experimental Environment

The hardware platform consists of three Dell R750 servers, each equipped with dual Xeon 6330 processors, 512GB RAM and two NVIDIA A40 GPU, providing powerful computing power. In terms of storage, Ceph distributed storage system is adopted, with a total capacity of 1.2PB and a throughput of 12GB/s, ensuring efficient data access and management. The test data set covers 8TB of industrial equipment data set (including vibration signals, infrared images and logs) and 5.6TB of smart city data set (including traffic video streams, weather sensor data and social media information).

3.2. Experimental Result

Table 1. Verification of heterogeneous data fusion effect

Training rounds	Cross-modal link accuracy	Semantic conflict resolution rate
1	78.3%	72.6%
3	87.1%	85.4%
5	92.7%	91.2%

The industrial equipment data set is used for five rounds of incremental training, and the cross-modal entity link effect is gradually optimized through the dynamic KG alignment model. After each training round, the cross-modal link accuracy and semantic conflict resolution rate are evaluated. The results show that with the increase of training rounds, the model performance continues to improve. The data in Table 1 shows that the initial round link accuracy rate is 78.3% and the semantic conflict resolution rate is 72.6%. In the third round, it rose to 87.1% and 85.4% respectively; After five rounds of training, the two indexes reach 92.7% and 91.2% respectively, which shows that the proposed semantic alignment model can effectively learn and fuse the structural and semantic information in multi-source heterogeneous data, and significantly improve the quality of data fusion.

As shown in Figure 3, the flow batch collaborative computing scheme proposed in this study significantly outperforms pure streaming and pure batch processing methods in terms of latency stability and throughput. Although the average delay of pure flow calculation is relatively low (186ms), there is a problem of state drift, which requires periodic resetting and affects continuity; Pure batch processing updates every two hours, with a latency peak of up to 14 minutes, which is difficult to meet real-time requirements; In contrast, this method stabilizes the processing delay at 210±15ms, balancing real-time performance and accuracy. At the same time, the system throughput reached 38000 events/second, an increase of 18.75% compared to the pure stream based 32000 events/second, verifying the efficiency and stability of the stream batch collaboration framework in high load scenarios.

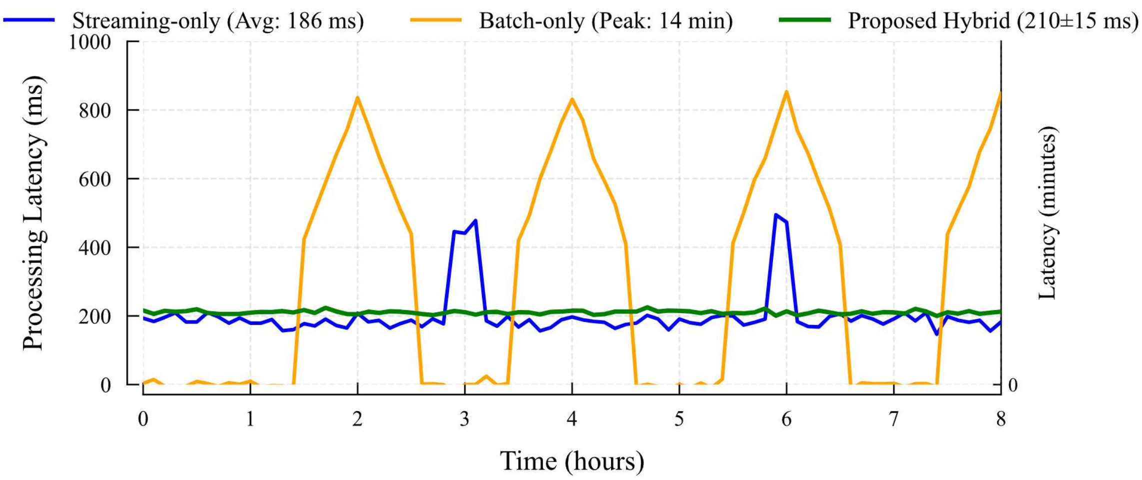

Figure 3. Efficiency comparison of flow batch collaborative computing

Table 2. Visual engine performance test

Interactive operation	Traditional scheme delay	Delay of this scheme
Global scaling	320ms	75ms
Hot spot area focusing	280ms	68ms
Cross-modal association query	410ms	**83ms**

Experiments are conducted to evaluate the performance of the proposed visualization engine when processing large-scale data. The test conditions include a resolution of 1920×1080@60Hz and a data set containing 1 million spatial entities. The results in Table 2 above show that in the global scaling operation, the delay of the traditional scheme is 320ms, while this scheme only needs 75ms; For the focus of hot spots, the traditional delay is 280ms, and this scheme is reduced to 68ms; For cross-modal correlation query, the traditional method takes 410ms, but this scheme greatly reduces the delay to 83 ms. These data fully demonstrate that the new engine can significantly shorten the interactive response time while ensuring high-resolution display, and greatly improve the user experience and operational efficiency.

4. Conclusion

The semantic alignment model based on dynamic KG effectively solves the semantic conflict problem of multi-source heterogeneous data. The experimental results show that after five rounds of incremental training, the cross-modal link accuracy and semantic conflict resolution rate of the model reach 92.7% and 91.2% respectively, which significantly improves the quality of data fusion. At the same time, this study designed a hybrid computing framework of batch-flow collaboration, which realized the complementary advantages of real-time data processing and historical data analysis. The processing delay was stably controlled at 210±15ms, and the throughput reached 38,000 events/second, which was 18.75% higher than that of pure streaming processing. In addition, the developed visualization engine can control the interactive delay of large-scale complex data within 100ms through progressive rendering technology, which greatly improves the user experience and operational efficiency. This study has achieved a breakthrough in overall performance in theory, method and system, and provided effective technical support for processing massive heterogeneous data.

References

[1] Anastasia Semykina.(2025).Estimating heterogeneous effects in static binary response panel data models.Econometric Reviews,44(8),1209-1233.

[2] Jiuyun Hu,Naichen Shi,Raed Al Kontar & Hao Yan.(2025).Personalized Tucker Decomposition: Modeling Commonality and Peculiarity on Tensor Data.Technometrics,67(3),409-425.

[3] Oluwatobi Bamigbade,Emonena Patrick Obrik Uloho,Faith Hauwa Oluwapamilerin Kolo,Akinde Michael Ogunmolu & Temilade Oluwatoyin Adesokan Imran.(2025).Data-Centric AI for Zero-Carbon Power Systems Security: A Framework for Learning with Noisy, Sparse, and Heterogeneous Data.Journal of Energy Research and Reviews,17(7),167-185.

[4] Haoming Wang & Wei Gao.(2025).Tackling Intertwined Data and Device Heterogeneities in Federated Learning with Unlimited Staleness..Proceedings of the ... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence,39(20),21080-21089.

[5] M. A. Poltavtseva & M. O. Kalinin.(2024).Access Control Analysis in Heterogeneous Big Data Management Systems.Programming and Computer Software,50(7),549-558.

[6] Chanmin Kim,Yisheng Li,Ting Xu & Zhongxing Liao.(2024).Bayesian Nonparametric Model for Heterogeneous Treatment Effects With Zero-Inflated Data..Statistics in medicine,43(30),5968-5982.

[7] Nathalia Hoffmann Guarda Aguzzoli,Ana Laura Bueno,Yağmur Halezeroğlu & Renan Rangel Bonamigo.(2024).Skin cancer after heart transplantation: a systematic review..Anais brasileiros de dermatologia,100(1),63-86.

[8] Xin Yang,Qi Jing Yan & Mi Xia Wu.(2024).Adaptive Distributed Inference for Multi-source Massive Heterogeneous Data.Acta Mathematica Sinica, English Series,40(11),2751-2770.

[9] Anwar AhmedKhan,ShamaSiddiqui & Muhammad ShoaibSiddiqui.(2024).Delay management for heterogeneous traffic in vehicular sensor networks using packet fragmentation of low priority data.IET Communications,18(20),1753-1764.