Redundant Data High Efficiency Compression Based on Distributed Parallel Algorithm

Jianhu Gong (gongjhlw@163.com)
Guangdong Peizheng College

Research Article

Keywords: Distributed, Parallel algorithm, Redundant data, High performance, Compression

Posted Date: April 20th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-410158/v1

Redundant Data High Efficiency Compression Based on Distributed Parallel Algorithm

Jianhu Gong*
School of Data and Computer Science, Guangdong Peizheng College, Guangzhou, 510830, China
gscilw@163.com

Abstract: In order to improve the optimal storage capacity of redundant data in serial hybrid network cascade database, a high efficiency compression algorithm for redundant data in serial hybrid network cascade database based on distributed parallel algorithm is proposed. The distributed storage structure model of redundant data of serial mixed network cascade database is constructed, the association feature extraction of redundant data of serial mixed network cascade database is carried out by using distributed hybrid feature mining method, the dimension reduction of redundant data of serial mixed network cascade database is carried out by combining with feature transformation method, the automatic location allocation of redundant data of serial mixed network cascade database is carried out by using high-order spectrum decomposition method, and the high-efficiency energy compression output model of redundant data of serial mixed network cascade database is constructed. The simulation results show that this method has good losslessness for redundant data compression of serial mixed network cascaded database and good fidelity of data output.
Keywords: Distributed; Parallel algorithm; Redundant data; High performance; Compression

1. Introduction

With the development of the intelligent serial network transmission technology, the remote cascaded database load transmission is realized through the serial network, and the automatic control ability of the serial network output is improved. In the intelligent serial network design and maintenance, it is necessary to combine the operation and maintenance management technology to carry out the redundant data storage and optimal transmission design of the serial hybrid network cascaded database in the intelligent serial network, to establish the optimal storage and transmission model of the redundant data of the cascaded serial network cascaded database, and to improve the remote communication and adaptive control ability of the cascaded database. Through the high efficiency compression design of redundant data of serial mixed network cascaded database (Wei et al. 2017), the channel equalization design of cascaded database telecontrol communication is carried out, which improves the distributed storage performance of redundant data of serial mixed network cascaded database, reduces the storage overhead of redundant data of serial mixed network cascaded database, and improves the throughput performance of cascaded database telecontrol (Razavian et al. 2016).

At present, the research on the elimination of network data redundancy is mostly carried out in a stand-alone environment. The analysis of the data redundancy mode in the network requires a large amount of original network data as the source data. It obviously takes a lot of time to analyze in a stand-alone environment, with the help of efficient computing in distributed computing environment, the time of data analysis can be reduced. By analyzing the redundancy elimination of CDN network data by network data types, this paper explores a redundancy mode analysis method for different network data types, so as to promote the research on the redundancy elimination of CDN network data and provide an analysis idea for improving the efficiency of network data redundancy elimination, And design an analysis system that can analyze different types of data redundancy in network data, and provide reference for the research of data redundancy elimination of network data from the design and

implementation of research tools, How to choose the block with high redundancy rate is the key to achieve better effect of data redundancy elimination. It is obviously unrealistic to calculate the fingerprint of each packet, because a large fingerprint table will be generated. At present, the mainstream data redundancy elimination algorithm is improved on the basis of winnowing algorithm. The data fingerprint generation mechanism of winnowing algorithm is: use the sliding window to divide the data area, the size of the window can be defined by itself, select the most special hash value in the hash value of the window as the fingerprint, the selection mechanism not only ensures that enough and representative fingerprint can be selected, And it doesn't make the fingerprint bigger.

Compression of serial hybrid network cascade database redundant data is the key to realize optimal storage of communication data. In traditional methods, the compression methods of serial hybrid network cascade database redundant data mainly include K-means clustering high-performance compression method, mutual information entropy extraction method and high-performance compression algorithm of serial hybrid network cascade database redundant data fused by block area chain information. A large data information processing model of serial hybrid network cascade database redundant data is established, and the optimal clustering and high-performance compression of serial hybrid network cascade database redundant data are carried out by adopting an associated data information fusion method (Zhang et al. 2014). A high-performance compression algorithm of serial hybrid network cascade database redundant data based on a multi-view clustering model is proposed in document (Guo et al. 2016), and a phase space distribution structure model of serial hybrid network cascade database redundant data is constructed. The multi-view clustering analysis method is adopted to realize high-performance compression of redundant data in serial hybrid network cascade database. However, the feature clustering of data compression by this method is not good and the ability of discrimination and recognition is not strong. Document proposes a high-performance compression algorithm for serial hybrid network cascade database redundancy data based on regional chain information fusion (Xiong et al. 2018). The method of association data information fusion is adopted to compress the serial hybrid network cascade database redundancy data, but the method has high computational complexity and poor real-time compression (Dai et al. 2016; Ma et al. 2016; Zhou and Xu 2018).

In view of the above problems, this paper proposes a high-performance compression algorithm for serial hybrid network cascade database redundancy data based on distributed parallel algorithm. Firstly, a distributed storage structure model of serial hybrid network cascade database redundant data is constructed, automatic location allocation in the storage process of serial hybrid network cascade database redundant data is carried out, and feature dimensionality reduction of serial hybrid network cascade database redundant data is realized by combining a feature transformation method, so that redundant data can be compressed with high efficiency. Finally, a simulation experiment is carried out to demonstrate the superior performance of the method in improving the high-performance compression capability and storage capability of the serial hybrid network cascade database redundant data.

2. Basic Definitions

Data redundancy elimination technology is to eliminate duplicate data in network data and save network traffic. Dre technology is saved at both ends of data transmission. Data blocks with high redundancy rate analyzed from TCP byte stream data are used as redundancy dictionaries to replace duplicate data with smaller data fingerprints, so as to eliminate redundant part in network data. The redundant data block selection algorithm used by the data compression engine is a data redundancy

elimination algorithm oriented to the data package level. The redundancy mode of the data package is analyzed, which is independent of the protocol.

In order to realize high-performance compression of serial hybrid network cascade database redundant data based on distributed parallel algorithm, a distributed storage structure model of serial hybrid network cascade database redundant data is constructed, a distributed hybrid feature mining method is adopted to collect serial hybrid network cascade database redundant data, an automatic location allocation of serial hybrid network cascade database redundant data is realized by combining a multi-feature transformation method (Wu et al. 2020), and a high-performance compression model is established. Using the control technology, the fuzzy correlation set of the serial hybrid network cascade database redundant data is obtained to satisfy $A \subset V$, $B \subset V$ and $A \text{ I } B = \varphi$. According to the transmission characteristics of the serial network, node allocation and fuzzy control of the serial hybrid network cascade database redundant data are carried out. Assuming the serial hybrid network cascade database redundant data set $X = \{x_1, x_2, \text{L}, x_n\}$, cross feature items of multilevel serial network contact points are calculated. Ontological models A and B are used for statistical feature analysis and adaptive collection of redundant data of cascade databases of serial hybrid networks.

The protocol independent data redundancy elimination is to ignore the transmission protocol used by the upper layer and eliminate the network data redundancy from the packet level. To eliminate the redundancy from the packet level, we need to analyze the bytes. At first, winnowing algorithm is an efficient algorithm applied to document similarity detection. It divides the substring by selecting the window value, then selects the data fingerprint as the document feature, and compares the document feature to detect the similarity. The feature of compression algorithm is to select data blocks with high redundancy rate from network data packets so as to eliminate the redundant part of network data better, and the data analyzed are all network data without difference. The direction of this paper is to adopt different redundancy elimination strategies for data of different media information in network data, so as to achieve better effect of eliminating redundant data.

When $G(A)$ and $G(B)$ is satisfied, communication channel equalization is carried out in combination with voltage fluctuation, and bearing capacity prediction method of cascade database communication networks is combined. According to data compression quality disturbance, the transmission data set of serial networks is obtained as follows:

$$y(t) = \sqrt{k}\, x(kt), k > 0 \tag{1}$$

$$W_y(t, v) = W_x(kt, v / k) \tag{2}$$

Where, k represents the disturbance frequency of the serial network, v represents the sampling period of redundant data in the cascade database of the intelligent serial hybrid network, W_x is a function of the source load fluctuation of the serial network (Ye et al. 2015). The quantitative recursive analysis method is adopted to adjust the output balance of the redundant data of the serial hybrid network cascade database, and the output adaptive control of the redundant data of the serial hybrid network cascade database is realized. The output is as follows:

$$x_n = a_0 + \sum_{i=1}^{M_{AR}} a_i x_{n-i} + \sum_{j=0}^{M_{MA}} b_j \eta_{n-j} \tag{3}$$

Where, a_0 is the sampling amplitude of distributed communication information in the serial network and x_{n-i} is the scalar time series of network carrying capacity. A fuzzy clustering model is established

to obtain the characteristic acquisition output Z of the serial hybrid network cascade database redundant data obeying Gaussian distribution with parameter β_d , wherein:

$$\beta_d = (MPDist - d + 1)/MPDist,\ \ d \in [2, MPDist] \tag{4}$$

Where, $adj(a,c)$ represents the correlation probability distribution of redundant data of serial hybrid network cascade database, and the optimized storage structure expression of redundant data of serial hybrid network cascade database is obtained as follows:

$$p(y \mid \alpha, \theta) = \sum_{k=1}^{K} \alpha_k p_k(y \mid \mu_k, \sum_k) \tag{5}$$

According to the oscillation type fluctuation of redundant data in the cascade database of the serial hybrid network, the output is:

$$C_{T'}(f)Y_{T'}(f) = C_{T'}(f)\sum_n x\left(f - \frac{n}{T'}\right)e^{j2\pi\left(f-\frac{n}{T'}\right)\tau_0} = C_{T'}(f)X(f)e^{j2\pi f\tau_0} \tag{6}$$

According to the above analysis, a multi-parameter information fusion and feature extraction model of serial hybrid network cascade database redundant data is established, fuzzy clustering is carried out according to the data acquisition results, and optimal mining of serial hybrid network cascade database redundant data is carried out (Hu et al. 2013).

3. Data Feature Extraction

Before the network transmission, the part with higher repetition frequency is added to the redundancy dictionary, and then the high-frequency repetition field is replaced by the corresponding smaller fingerprint in the redundancy dictionary according to the redundancy dictionary, so as to further compress the data during the transmission. Since both ends of the data transmission maintain a common redundancy dictionary, after the data reaches the receiving end, it is replaced with the original data from the redundancy dictionary, Achieve the effect of decompression. In the application of CDN service environment, when analyzing all the undifferentiated network data, the data redundancy elimination effect is related to the block selection strategy of the data redundancy elimination algorithm selected in the DRE compression engine. How to select the block with high redundancy rate as the redundancy dictionary is a hot spot of current research. However, in the analysis of the network data packet captured by CDN server, it is found that the data redundancy elimination effect of the DRE compression engine is different for different types of network data, so that the best data redundancy elimination effect of each type of data will inevitably increase the overall data redundancy elimination effect. If different algorithm window values are used to generate redundant dictionaries for different types of data, the data redundancy elimination effect of different types of network data will be the best.

Because of the end-to-end data redundancy elimination, redundant dictionary is used to record the corresponding relationship between redundant data block and data fingerprint. The generation of redundant dictionary is based on the analysis of the content of the data package, selecting the blocks with high redundancy, using hash function to calculate the data fingerprint of these blocks with high redundancy, storing them in the redundant dictionary, and using the smaller data fingerprint to replace the original data block in data transmission. When there are consistent redundant dictionaries at both ends of data transmission, the synchronization of redundant dictionaries can be updated by timing.

The data fingerprints in the redundant dictionary are not always the same, but need to replace some expired data fingerprints to store new data fingerprints when they need to be updated. The most classical replacement algorithm is the least recently unused algorithm (LRU). Using LRU algorithm

directly and simply can achieve the purpose of replacement. However, this situation will make some data fingerprints with high contribution rate and need to be replaced in the future, so it is necessary to rank the compression effect of data fingerprints in redundant dictionary, According to the ranking, the fingerprint of the data to be replaced is determined.

The DRE program collects the use information of the data block corresponding to each data fingerprint at runtime, and the recent data block utilization rate (i.e. the frequency of data block occurrence) and the contribution of data block (i.e. how much data is compressed) are the basis for ranking. The replacement algorithm uses the collected information to determine the replacement object. For the multi-media data redundancy mode analysis system, it is necessary to generate corresponding redundancy dictionaries for different data types. When using the multi redundancy dictionary mode to replace redundant data blocks and data fingerprints, a judgment logic is needed to determine what type of data is currently processed, so as to select corresponding redundancy dictionaries for search and replacement.

The information of obtaining TCP flow starts from the recovery of TCP connection. The recovery of TCP connection is based on the analysis of the captured network data packet, and the judgment of whether it is TCP load from the protocol field of IP packet. Then, the connection establishment and disconnection are completed according to the syn, fin and rst of TCP packet header. The key information of connection establishment is the source IP address, destination IP address, source port number and destination port number (27). The specific TCP connection simulation recovery process is as follows:

A. From the analysis of the captured packets, the common server port numbers of HTTP protocol are 80 and 8080, and the destination port number or source port number of the packets are used to determine whether the packets are request packets or HTTP request packets or HTTP response packets. The HTTP response packet contains the content type field information that needs to be extracted.

B. The source IP address, destination IP address, source port number and destination port number in the IP package are used as the identification of a TCP flow. As the key value of the TCP flow information storage, a map can be created based on the key value. The value part of the map is used to store other information related to the TCP flow.

C. After that, the TCP packet header protocol field is detected. If there are syn and ACK, the packet is the response of the server during the establishment of the connection by the three-way handshake of the TCP connection. At this time, the TCP connection information should be added to the map data structure storing the TCP flow information.

D. If there are fin and ACK in the header protocol field of the TCP packet, the packet is the response of the server in the process of four times of wave disconnection of the TCP connection. At this time, it indicates that the information of the TCP connection has been recorded, and the next TCP connection detection can be started.

E. If there is RST in the header protocol field of the TCP packet, it means that the current TCP connection is abnormal, and the connection needs to be reestablished. At this time, the current processed TCP connection information should be deleted from the map data structure.

The method of distributed hybrid feature mining is adopted to extract the associated features of the redundant data of the serial hybrid network cascade database, and the method of feature transformation is combined to reduce the dimension of the redundant data of the serial hybrid network cascade database (Huang et al. 2013), and the feature transformation quantization learning function of the remote communication number of the cascade database is as follows:

$$\begin{cases} \min \sum_{1\le i\le K}\sum_{e\subseteq k(e)} \dfrac{f(e(i))}{C(e,i)} \\ 0\le f(e,i)\le C(e,i) \\ F=const \\ \sum_{1\le i\le K, e\subseteq k(e)} \dfrac{f(e(i))}{C(e,i)} + \sum_{e\subseteq k(e)} \dfrac{f(e^{'}(i))}{C(e^{'},i)} \le k(v) \end{cases} \tag{7}$$

Combined with the random feature transformation compression method, the feature reconstruction is carried out, and the binary coding of redundant data in the serial hybrid network cascade database is carried out. The low-frequency flicker distribution is obtained as follows:

$$\begin{aligned} Computition(n_j) &= (E_{elec}+E_{DF})l\delta + E_{Tx(l,d_j)} \\ &= (E_{elec}+E_{DF})l\delta + lE_{elec} + l\varepsilon_{fs}d_j{}^2 \\ &= [(E_{elec}+E_{DF})\delta + E_{elec} + \varepsilon_{fs}d_j{}^2]l \end{aligned} \tag{8}$$

Due to the random distribution of load and power consumption of cascade databases, spectral feature extraction of redundant data in cascade databases of serial hybrid networks is carried out in combination with fuzzy update rules, and the zero frequency feature distribution threshold of data compression is obtained:

$$\eta_k^w(\omega) = \mathrm{E}\left(T_k^w \middle| T_k^w > \xi_k^w(\omega)\right), k\in R_w, w\in W \tag{9}$$

Where: $\xi_k^w(\omega)$ is a multi-queue fuzzy scheduling function for redundant data of cascade database in serial hybrid network, which can be expressed as:

$$\xi_k^w(\omega) = \min\left\{\xi \middle| \Pr\left(T_k^w \le \xi\right) \ge \omega\right\} = \mathrm{E}\left(T_k^w\right) + \gamma_k^w(\omega)\ k\in R_w, w\in W \tag{10}$$

Extract a high-order spectrum statistical characteristic quantity of redundant data of a serial hybrid network cascade database, and obtaining a statistical characteristic distribution index set of redundant data of that serial hybrid network cascade database in a neighborhood space (t, f) as follow:

$$f(x) = \begin{cases} f(x), x\in Levf \\ a, x\in Levf \end{cases} \tag{11}$$

Calculating the fitness function with high performance compression, according to the distribution of frequent itemsets (Zheng and You 2013), the information fusion scheduling model of serial hybrid network cascade database redundant data is obtained as $\mathrm{E}\left(T_k^w - \xi_k^w(\omega) \middle| T_k^w \ge \xi_k^w(\omega)\right)$, and the global optimal solution of serial hybrid network cascade database redundant data can be expressed as:

$$s_h^w = \mathrm{E}\left[\min_{k\in R_w}\left\{\mathrm{H}_{h,k}^w\right\} \middle| \boldsymbol{\eta}^w\right] = -\frac{1}{\theta}\ln\sum_{k\in R_w}\exp\left(-\theta\eta_{h,k}^w(\omega)\right), w\in W, h\in H \tag{12}$$

According to the above analysis, a high-order spectrum statistical feature extraction model for redundant data of serial hybrid network cascade database is constructed, and a high-order spectrum decomposition method is adopted for automatic location allocation in the storage process of redundant data of serial hybrid network cascade database to improve the data compression and clustering capabilities (Han et al. 2019). Data is mostly transmitted between CDN servers and between CDN servers and mobile terminals. There are consistent redundant dictionaries at both ends of data transmission. The generation of redundant dictionaries is based on the analysis of the content of data packets, selecting data blocks with higher redundancy rate to calculate data fingerprints, and establishing the corresponding relationship between redundant data blocks and data fingerprints. Then, before data transmission, the redundant data block in the redundant dictionary is used to compress the data to be transmitted. After the data receiving end, the receiving end reconstructs the original data packet according to the redundant dictionary, thus saving the network resources consumed by the fingerprint table attached to each data transmission.

First of all, divide the source data by day to form a data package every day, which is more intuitive

when analyzing the chart. There are ready-made data package splitting tools in this part of work. Then, in the case of single redundant dictionary mode with default algorithm window value, analyze and count the network data in redundant mode, record the redundancy elimination efficiency of various network data types, and form "date network data type data" The table of "size before compression - size after data compression" forms pie chart of "proportion before compression of data types of various networks" and "proportion after compression of data types of various networks". The network data type with the largest difference between the proportion before data compression and the proportion after data compression is the network data type with the best compression effect in the current situation with this method, the most suitable algorithm window value of each network data type is obtained. Then, after generating redundant dictionaries of different types of network data, the efficiency of redundancy elimination is calculated by analyzing the pattern of redundant dictionaries. Through the comparison of data redundancy elimination efficiency between single redundancy dictionary mode and multi redundancy dictionary mode, the redundancy modes of different network data types are analyzed. Then, every five bytes are divided, the byte with the largest byte value is selected as the starting point, and four consecutive bytes are selected as the data block to calculate the data fingerprint. The fingerprint set of all the collected data blocks is used to generate redundant dictionaries, which is the process of dictionary generation. After generating the dictionary, when eliminating the redundancy of the transmission data, we still analyze the data according to this method. After calculating the data fingerprint, we will find out whether the redundancy dictionary exists in the dictionary (because the dictionary in the application is not updated in real time, but updated regularly, and the data to be transmitted will not be the set of data that generated the dictionary). If it exists in the dictionary, we will use the relative the fingerprint with smaller volume should replace the original data segment to achieve the purpose of redundancy elimination. If it does not exist in the dictionary, the redundant dictionary should be updated according to the replacement strategy of the dictionary. After the data arrives at the receiving end, according to the received data fingerprint, the original transmission data can be reconstructed by searching the corresponding original data in the redundant dictionary, which is a complete redundant elimination process.

4. Optimization of High Performance Compression Algorithm for Redundant Data

4.1. Feature transformation dimension reduction

On the basis of the above-mentioned construction of the distributed storage structure model of the serial hybrid network cascade database redundant data and the adoption of the distributed hybrid feature mining method to extract the association features of the serial hybrid network cascade database redundant data, the high-performance compression processing of the serial hybrid network cascade database redundant data is carried out. This paper proposes a high-performance compression algorithm of the serial hybrid network cascade database redundant data based on the distributed parallel algorithm (Yu et al. 2015; Li 2019). The feature extraction model of fuzzy association rules for redundant data of cascade databases in serial hybrid networks is constructed, and the automatic location allocation in the storage process of redundant data of cascade databases in serial hybrid networks is carried out by adopting a high-order spectral decomposition method. The energy overhead of communication transmission of cascade databases:

$$R_2 = \{X_{d+1}, X_{d+2}, \cdots X_{d+m}\}^T \tag{13}$$

When $R_2^T R_2 = \{X_{d+1}, X_{d+2}, \cdots X_{d+m}\}\{X_{d+1}, X_{d+2}, \cdots X_{d+m}\}^T$, the window function $V = [V_1, V_2, \cdots, V_m] \in R^{m \times m}$ for state space reorganization of redundant data in cascade database of serial

hybrid network takes the minimum value. When $V \in R^{m \times m}$, $VV^T = I_M$ have minimum values, the method of statistical information mining and fuzzy information clustering analysis is used to classify the redundant data of the serial hybrid network cascade database, and the state information fusion of the redundant data of the serial hybrid network cascade database is carried out. According to the application scenario or voltage level, the method of minimum utility threshold scheduling is adopted to extract the statistical characteristic quantity of the redundant data of the serial hybrid network cascade database, and the fuzzy membership function of the redundant data of the serial hybrid network cascade database is obtained as follows:

$$y(t) = \frac{1}{\pi} P \int \frac{x(\tau)}{t-\tau} d\tau = x(t) * \frac{1}{\pi t} \tag{14}$$

In the above formula, P is the time domain distribution set of redundant data of the serial hybrid network cascade database, $x(t)$ is the sampling delay of the DC data center, and based on the transient voltage characteristic analysis method, the impact load of redundant data of the serial hybrid network cascade database is obtained, thus a high-performance compressed output model of redundant data of the serial hybrid network cascade database is constructed, and feature transformation dimensionality reduction processing is realized (Wang and Wang 2016; Xiao 2021).

4.2. Redundant data can be efficiently compressed and output

When processing scattered TCP packets, because TCP packets will be truncated when they are too long, only the first packet will have the HTTP protocol header, which leads to the problem of how to correctly identify the packets of a TCP stream in order. The existing TCP packet reassembly is to ensure the correct sorting of TCP packets according to the transmission sequence number of the TCP packet header, However, for a large number of network packets that need to be processed, sorting analysis by serial number needs a lot of resources and reduces the analysis efficiency. Even if the problem of TCP packet order is not considered, in the same TCP connection, it is possible to initiate multiple HTTP requests and responses, so it is possible to send multiple types of data. How to correctly identify each data type and quantitatively count the data traffic size is an important problem. The connection methods of HTTP protocol include non persistent connection and persistent connection.

For the non persistent connection mode, the specific establishment steps are: the HTTP client initializes a TCP connection with the host of the HTTP server, and then the client sends a request message to the HTTP server. The HTTP server responds to the request and transmits the data. After the object is transmitted, it will notify the http client to close the TCP connection, and the HTTP client will establish a TCP connection again when it requests again. In this case, it is necessary to establish and close a TCP connection once to transmit a data object. In the process of communication between HTTP client and HTTP server, the TCP connection has been used to communicate from the establishment of TCP connection to the closure of TCP connection, in which the order of sending data is sequential, It will not continue to send the data of the second request when the data of the first request has not been sent, so even if only the first TCP data segment has the packet header of HTTP protocol, the subsequent data is the network data of this type until the next TCP packet with the packet header of HTTP protocol is encountered.

There is a strict order between the non persistent connection and persistent connection modes of HTTP when the server responds to different requests from the client. Therefore, when processing TCP packets, whenever the content type field in the HTTP protocol header is extracted, the content type information of the current TCP stream can be updated. Before the next content type update, All arrived packets are

calculated according to the current content type. When a packet containing the content type field is encountered, it indicates a new HTTP request response. Just update the value of the content type field. In this way, a variable recording the content type value can be used to identify the data type in multiple HTTP transfers. For the characteristics of the network data types of different domain names, if the domain name image data accounts for a large proportion, then the domain name can improve the redundancy elimination efficiency of the image data by using the dictionary generated by the algorithm window value suitable for the image data when sending the data. Because the image data accounts for a large proportion of the total data, This may improve the overall efficiency of redundancy elimination (Liu et al. 2021; Zhu et al. 2020).

The workflow in different redundant dictionary modes includes the following points:

A. Calculate the overall redundancy elimination efficiency of three different domain name network data. This process uses the default value of windowsize, single redundant dictionary mode. Because the input data file is pcap file divided by day, shell script is used to ensure the continuous automatic operation and statistics of the import record file of dancing fruit.

B. Adjust windowsize to generate different redundant dictionaries to find the best windowsize for network data such as text type. The process uses - w to specify different windowsize, single redundancy dictionary mode, and statistics the efficiency of data redundancy elimination, through the results comparison to find the most appropriate windowsize. Shell script is also used to ensure continuous automatic operation and import the statistical results into the record file.

C. After the best window size value of each network data type is obtained from process 2, the window value is used to generate different network data type dictionaries of daily data, which are stored in files named by day, and different redundant dictionaries of different network data types are named by different data type names.

D. Calculate the overall redundancy elimination efficiency of multi redundancy dictionary mode. This process is mainly for the use of multi redundant dictionary. When analyzing data, the - M parameter is added to represent the multi redundant dictionary mode. In the data analysis stage, the judgment branch of selecting redundant dictionary is added, and the corresponding redundant dictionary is selected according to the content type value to replace the redundant data block and data fingerprint. Shell script is also used to ensure the continuous automatic operation and import the statistical results into the record file.

The statistics of redundant data elimination efficiency adopts map structure as information storage structure. The current TCP connection is identified by the source IP address, destination TP address, source port number and destination port number of data packets. As the key of map, the value of map stores data size before redundancy elimination and data size after redundancy elimination in the way of user-defined structure. This advantage is expandable Good expansibility. If you need to add statistics fields or information in the future, just add definition fields in the structure. After determining the network data type of the current TCP connection, it is necessary to make statistics on the data size of the incoming packets as the data size before compression by the DRE engine. After analysis, the data size after redundancy elimination is counted as the data size after compression by the DRE engine, forming the intermediate results of the analysis by the DRE engine, as Figure 1 shows:

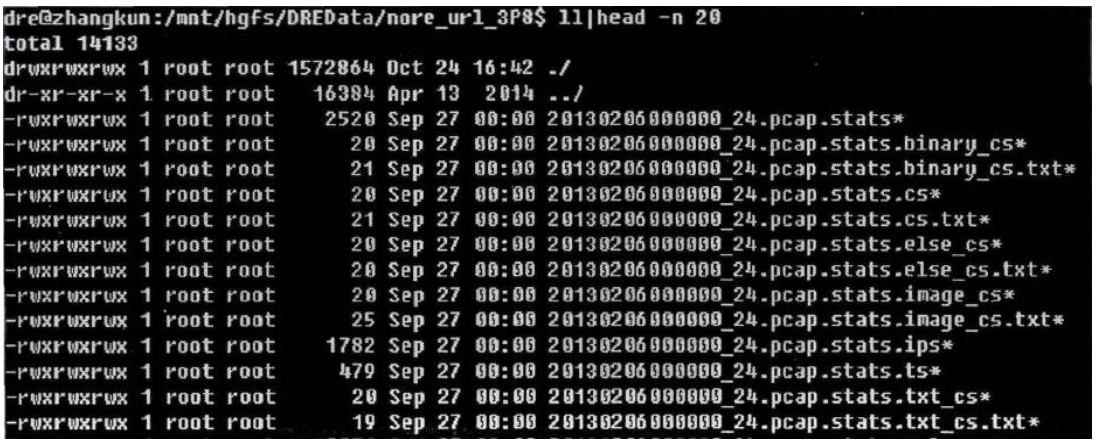

Fig. 1 Intermediate results of engine analysis

Because the intermediate result is in the form of key value, map reduce is more suitable for statistical processing. Here, the streaming stream in the Hadoop environment is used to process the intermediate result files in the HDFS file system. The PHP program and the map reduce program are used. After the intermediate result collection is completed, the map reduce script is run for all day data statistics. Map reduce program is written to open the currently processed file with fopen function of PHP, send the file content into the processing part of map in the form of stream, divide the content of current line with "\ T", record the values of the second and third fields, put them into the output stream, reduce program reads all the second and third fields extracted from map from the output stream, and then makes statistics of the results. When using the system in the actual CDN node server, you can use the map reduce program that crontab runs regularly, run the map reduce program after collecting the intermediate results of each node every day or in a period of time for the statistics of that day, or select some intermediate results as the input of the map reduce program for statistics according to the situation. In this way, the implementation of putting the intermediate results of each dre engine analysis into the HDFS file system is similar to the current mature log processing system, which is flexible and convenient for data mining and system expansion.

The data is divided into the following parts according to the workflow:

A. Calculate the overall redundancy elimination efficiency of three different domain name network data.

B. Adjust windowsize to generate different redundant dictionaries to find the best avindowsize for network data such as text type.

C. Calculate the overall redundancy elimination efficiency of multi redundancy dictionary mode.

A high-performance compression model of serial hybrid network cascade database redundant data is constructed by adopting a feature transformation method, deviation and steady-state adjustment are carried out on the serial hybrid network cascade database redundant data, and the output of the feature transformation is as follows:

$$q^w = \mathrm{E}\left(Q^w\right) = \sum_{k \in R_w} f_k^w \,,\, w \in W \tag{15}$$

$$v_a = \mathrm{E}\left(V_a\right) = \sum_{w \in W} \sum_{k \in R_w} \delta_{ak}^w f_k^w \,,\, a \in A \tag{16}$$

$$f_k^w \geq 0 \,,\, k \in R_w \,,\, w \in W \tag{17}$$

By decomposing Fourier transform into superposition of harmonics, the high-performance compressed binary structure model of redundant data is obtained as follows:

$$\left.\begin{aligned} &\min_{\alpha} \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j K(x_i, x_j) - \sum_{j=1}^{l} \alpha_j \\ &s.t. \quad \sum_{j=1}^{l} y_j \alpha_j = 0 \\ &\qquad 0 \le \alpha_j \le u(x_j)C, \qquad j = 1, 2, ..., l \end{aligned}\right\} \tag{18}$$

In the reconstructed phase space of the redundant data distribution of the serial hybrid network cascade database, the unbalanced degree fusion method is adopted for high-performance compression, and the quantized characteristic distribution set of the redundant data of the serial hybrid network cascade database is defined as D, $D = \left\{S_{i,j}(t), T_{i,j}(t), U_{i,j}(t)\right\}$, wherein $S_{i,j}(t)$ represents the adaptive weight of the redundant data of the serial hybrid network cascade database, $T_{i,j}(t)$ represents the fused clustering characteristic distribution set of the redundant data of the serial hybrid network cascade database, $U_{i,j}(t)$ represents a similarity (correlation) model, and the redundant data of the serial hybrid network cascade database are subjected to block-wise high-performance compression, and the output is:

$$S_{i,j}(t) = \frac{p_{i,j}(t) - sp_{i,j}(t)}{p_{i,j}(t)} \tag{19}$$

Wherein, $T_{i,j}(t)$ represents the fuzzy feature distribution set of redundant data compression for serial hybrid network cascade database, and the calculation expression is:

$$T_{i,j}(t) = \frac{\left|p_{i,j}(t) - \Delta p(t)\right|}{p_{i,j}(t)} \tag{20}$$

According to the above analysis, the high-order spectral feature quantity of the redundant data of the serial hybrid network cascade database is extracted, and the classification processing of the redundant data of the serial hybrid network cascade database is carried out by using the methods of statistical information mining and fuzzy information clustering analysis, and the feature dimension reduction of the redundant data of the serial hybrid network cascade database is realized by using the method of feature transformation, so that the redundant data can be efficiently compressed, and the realization flow is shown in figure 2.

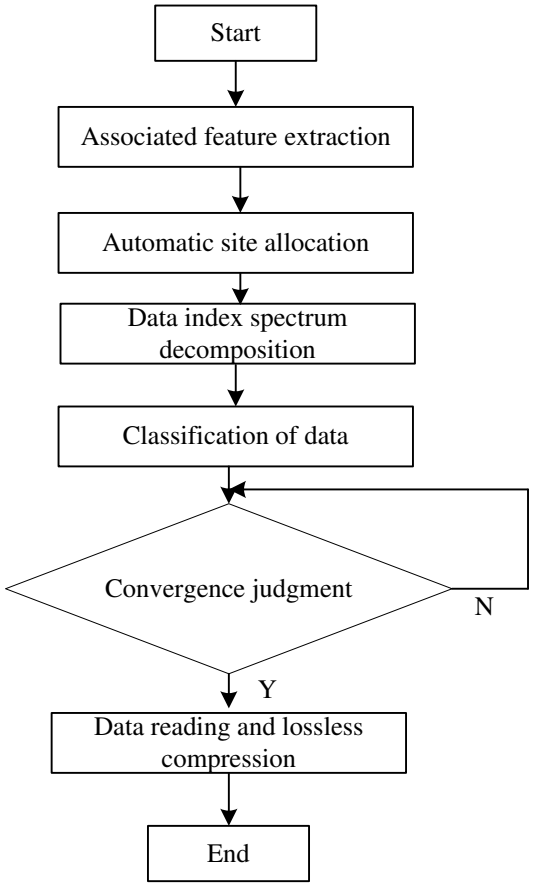

Fig. 2 Flow of high efficiency compression of redundant data

5. Simulation Experiment and Analysis of Results

In order to test that application performance of the method in realize high-performance compression of the redundant data of the serial hybrid network cascade database, a simulation experiment is carried out by using Matlab. The sampling sample length of the redundant data of the serial hybrid network cascade database is 2000, the code bit sequence length of the data is 120, the spatial embedding dimension of the redundant data of the serial hybrid network cascade database is set to 6, and the data load deviation of the serial hybrid network is 0.25. The deviation limit of the redundant data is 1.48, and the adjacent load is 10dB. according to the above simulation parameter settings, high-performance compression of the redundant data of the serial hybrid network cascade database is performed to obtain the original redundant data of the serial hybrid network cascade database as shown in figure 3.

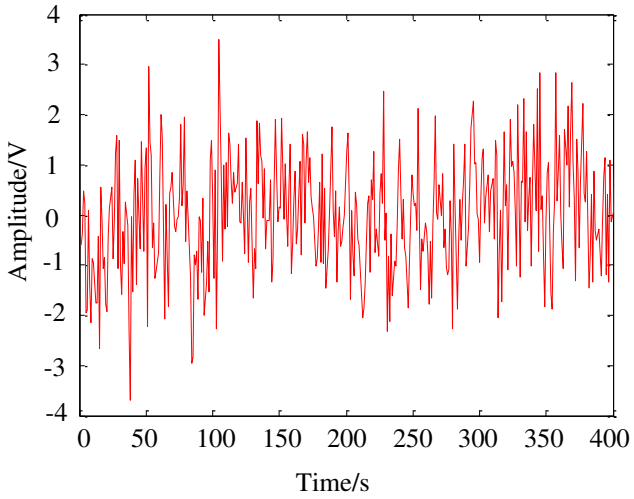

Fig. 3 Redundant data from the original serial hybrid network cascade database

Taking the data in figure 3 as the research object, a high-performance compression output model of serial hybrid network cascade database redundant data is constructed. statistical information mining and fuzzy information clustering analysis are used to classify the serial hybrid network cascade database redundant data. feature transformation is used to achieve high-performance compression of the data, and the compressed output is shown in figure 4.

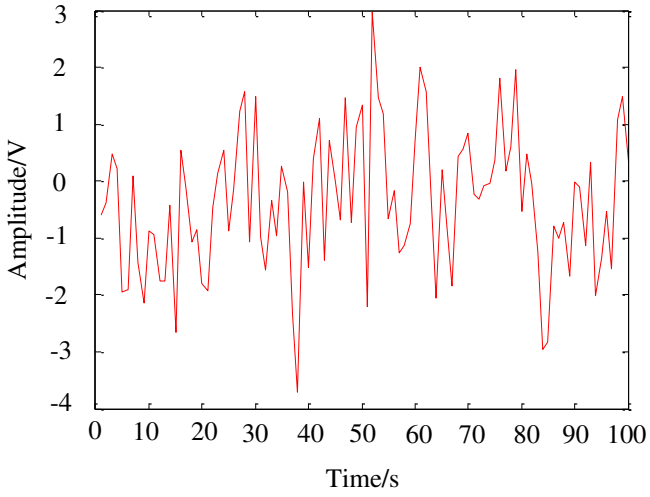

Fig. 4 Data compression output

Analysis of fig. 4 shows that the data compression can be effectively realized by using this method, and the storage cost of data is reduced. testing the fidelity of high-performance compression output of redundant data of serial hybrid network cascade database by different methods shows that the comparison results are shown in table 1. Analysis of table 1 shows that the fidelity of redundant data compression of serial hybrid network cascade database by using this method is higher and the lossless is better.

Table 1. Comparison of redundant data compression performance of cascaded databases in serial hybrid networks.

Number of iterations	This method	Reference (Guo et al. 2016)	Reference (Xiong et al. 2018)

100	0.845	0.843	0.734
200	0.966	0.857	0.878
300	0.987	0.888	0.904
400	0.993	0.912	0.917

6. Conclusions

In this paper, an optimized storage and transmission model of serial hybrid network cascade database redundant data is established to improve the remote communication transmission and adaptive control capability of cascade database. This paper proposes a high-performance compression algorithm of serial hybrid network cascade database redundant data based on distributed parallel algorithm.

The method comprises the following steps of: extracting high-order spectrum statistical characteristic quantities of redundant data of a serial hybrid network cascade database; adopting a high-order spectrum decomposition method to carry out automatic location allocation in the storage process of redundant data of the serial hybrid network cascade database; constructing a high-efficiency compression model of redundant data of the serial hybrid network cascade database; and adopting a characteristic transformation method to realize feature dimensionality reduction of redundant data of the serial hybrid network cascade database and realize high-efficiency compression of redundant data. The research shows that the method in this paper has good lossless performance and high fidelity of communication output when compressing redundant data in serial hybrid network cascade database.As a new research hotspot, the efficient storage of data has been paid more and more attention by academia and industry. Its ultimate purpose is to alleviate the rapid growth of storage system space demand, reduce the actual storage space occupied by data, simplify storage management, maximize the use of existing resources and reduce costs.

The application of big data leads to the increasing capacity demand of storage system. Considering the different performance and cost of different storage devices, and the local characteristics of time and space of data access, in order to save cost and reduce energy consumption, reduce the actual data storage, and store the appropriate data in the right place, hierarchical storage is often used in large data centers The mode adopted. Most of the existing granularity of hierarchical storage is data migration at the level of subvolume or file, which makes a large number of redundant data in the storage system. With the rapid development and application of SSD, its read-write performance is close to that of DRAM, and its energy consumption is far lower than that of mechanical HDD. This characteristic of SSDs will change the traditional hierarchical access mode of data that needs to go through memory and then to hard disk, bring about the change of the current storage system structure with mechanical hard disk as the main body, and also make it possible to realize the data migration strategy based on variable length block level, which can greatly improve the data De duplication rate and reduce the physical storage capacity required by data. Therefore, in practical use, how to use the advantages of SSD as much as possible, and how to introduce the capacity reduction technology based on variable length block granularity into the hierarchical storage mode, will be a big challenge for big data storage management, and also one of the future research directions.

Ethical approval: This article does not contain any studies with human participants performed by any of the authors.

Conflict of Interest: The authors declare no conflict of interest.

References

Dai YY, Li CF, Xu H (2016) Density spatial clustering algorithm with initial point optimization and parameter self-adaption. Comput Eng 42(1):203-209.
Guo H, Liu H, Wu C (2016) Logistic discrimination based on G-mean and F-measure for imbalanced problem. J Int Fuzzy Syst 31(3):1155-1166.
Han JH, Yuan JX, Wei X, Lu Y (2019) Pedestrian visual positioning algorithm for underground roadway based on deep learning. J Comput Appl 39(3):688-694.
Hu LR, Wu JG, Wang L (2013) Application and method for linear projective non-negative matrix factorization. Comput Sci 40(10):269-273.
Huang XJ, You RY, Zhou CJ (2013) Study on optical properties of equivalent film constructed of metal nanoparticle arrays. J Optoelectron·Laser 24(7):1434-1438.
Li LR (2019) Simulation research on intelligent verification of massive user information in wireless networks. Comput Sim 36(5):341-344.
Liu J, Zhao LY, Luo XY, Zou D (2021) Cyclic redundancy check method of serial communication data stream based on ZigBee. Comput Sim 38(1):226-230
Ma CL, Shan H, Ma T (2016) Improved density peaks based clustering algorithm with strategy choosing cluster center automatically. Comput Sci 43(7):255-258.
Razavian AS, Sullivan J, Carlsson S (2016) Visual instance retrieval with deep convolutional networks. ITE Trans Media Technol Appl 4(3):251-258.
Wang L, Wang F (2016) Simulation of efficient data localization in network database under redundant environment. Compu Sim 33(4):364-367.
Wei XS, Luo JH, Wu J (2017) Selective convolutional descriptor aggregation for fine-grained image retrieval. IEEE Trans Image Process 26(6):2868-2881.
Wu KH, Wu QR, Guan JY, Wei H, Wang B (2020) The Design and Implementation of a Real-Time Detection and Analysis System for ECG Signal. Comput Sci Appl 10(9):1631-1638.
Xiao F (2021) Research on feature detection method of high-dimensional discrete data in dual redundant networks. J Ningxia Norm Univ 42(1): 67-72.
Xiong H, Guo YQ, Zhu HH, Wang S (2018) Robust nonnegative matrix factorization on manifold via projected gradient method. Inform Contr 47(2):166-175.
Ye M, Qian Y, Zhou J (2015) Multitask sparse nonnegative matrix factorization for joint spectral-spatil hyperspectral imagery denoising. IEEE Trans Geosci Remote Sens 53(5):2621-2639.
Yu Z, Li L, Liu J (2015) Adaptive noise immune cluster ensemble using affinity propagation. IEEE Trans Knowl Data Eng 27(12):3176-3189.
Zhang Y, Fu P, Liu W (2014) Imbalanced data classification based on scaling kernel-based support vector machine. Neural Comput Appl 25(3/4):927-935.
Zheng YZ, You RY (2013) Study on segmented correlation in EEG based on principal component analysis. Chin J Biomed Eng 22(3):93-97.
Zhou SB, Xu WX (2018) A novel clustering algorithm based on relative density and decision graph. Control Decis 33(11):1921-1930.
Zhu RB, Li YL, Ding QA, Yu M (2020) Temporal correlation perceptual data de-redundancy algorithm based on maximum time threshold and adaptive step size. J S-Cent Univ Nationalities(Nat Sci Ed) 39(3):295-301.

Jianhu Gong, male, born in December 1967. He is an associate professor with a doctor's degree. He graduated from the computer technology major of Sun Yat-sen University. He is now a teacher in School of Data and Computer Science, Guangdong Peizheng College. His research interests are software engineering and big data technology. He has published many excellent academic articles.

Figures

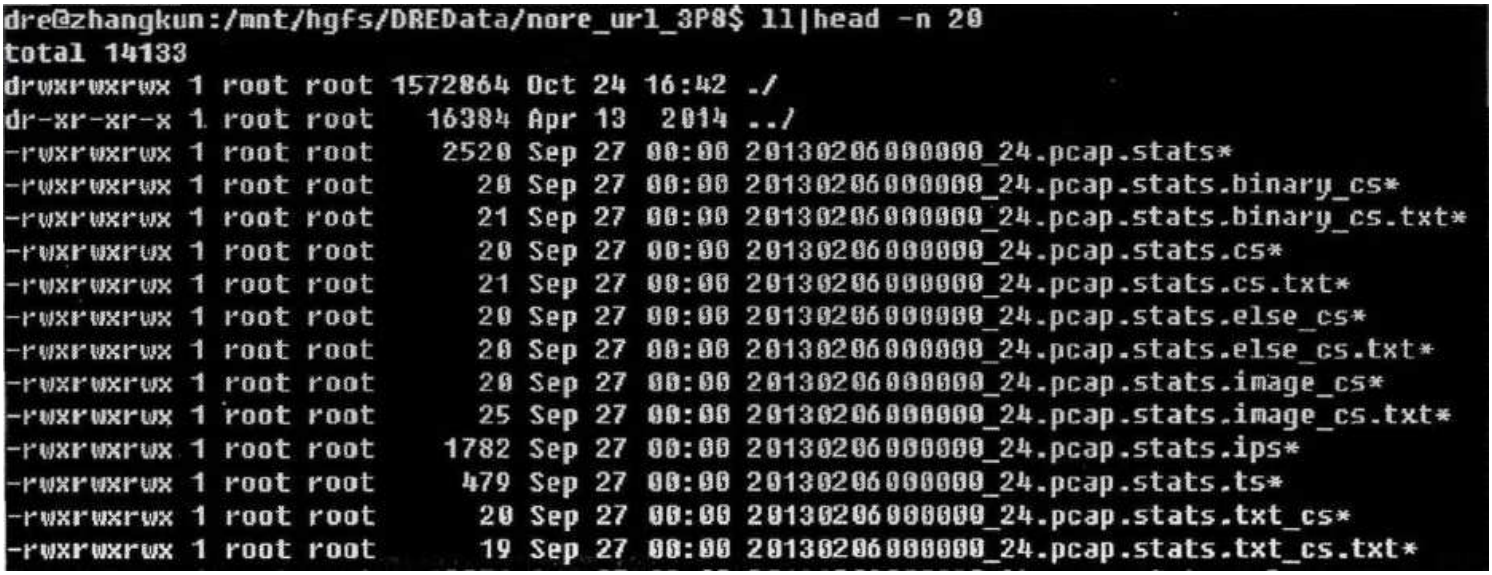

Figure 1

Intermediate results of engine analysis

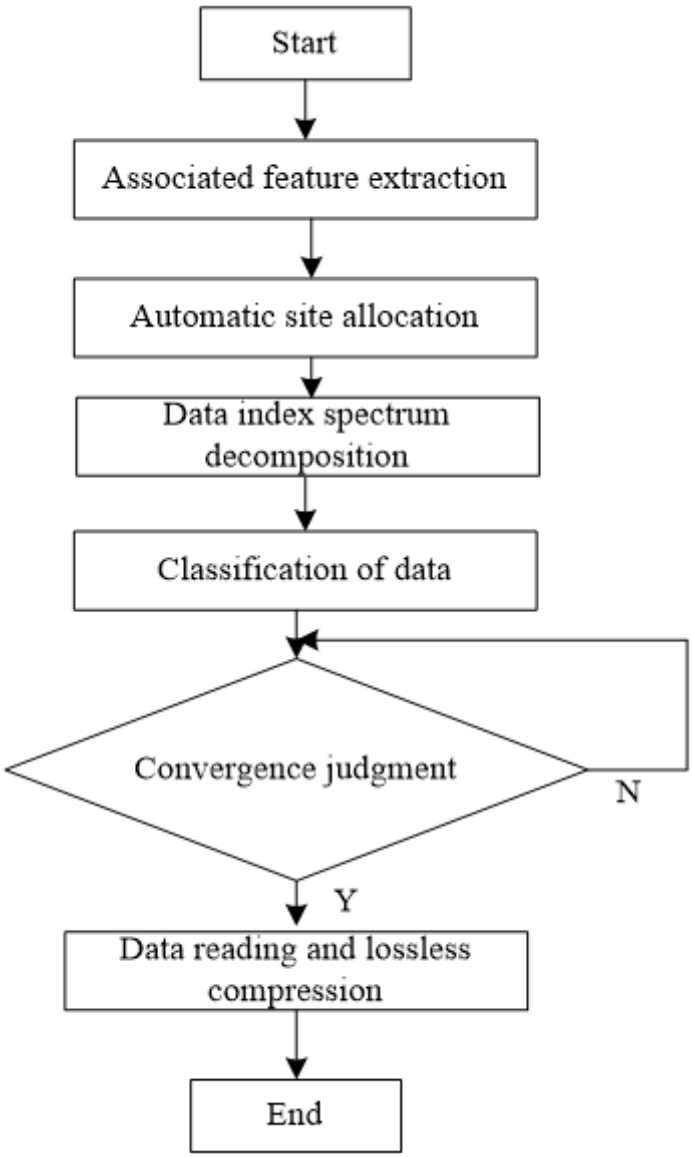

Figure 2

Flow of high efficiency compression of redundant data

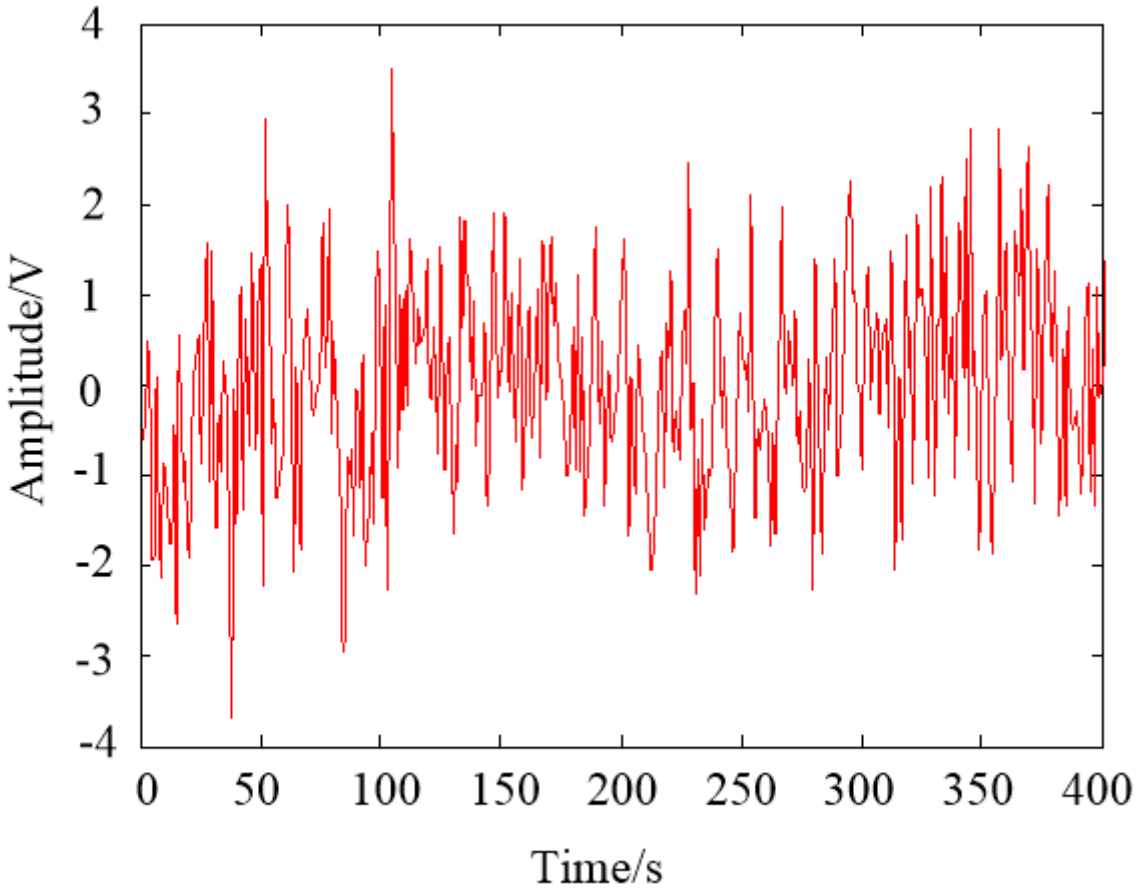

Figure 3

Redundant data from the original serial hybrid network cascade database

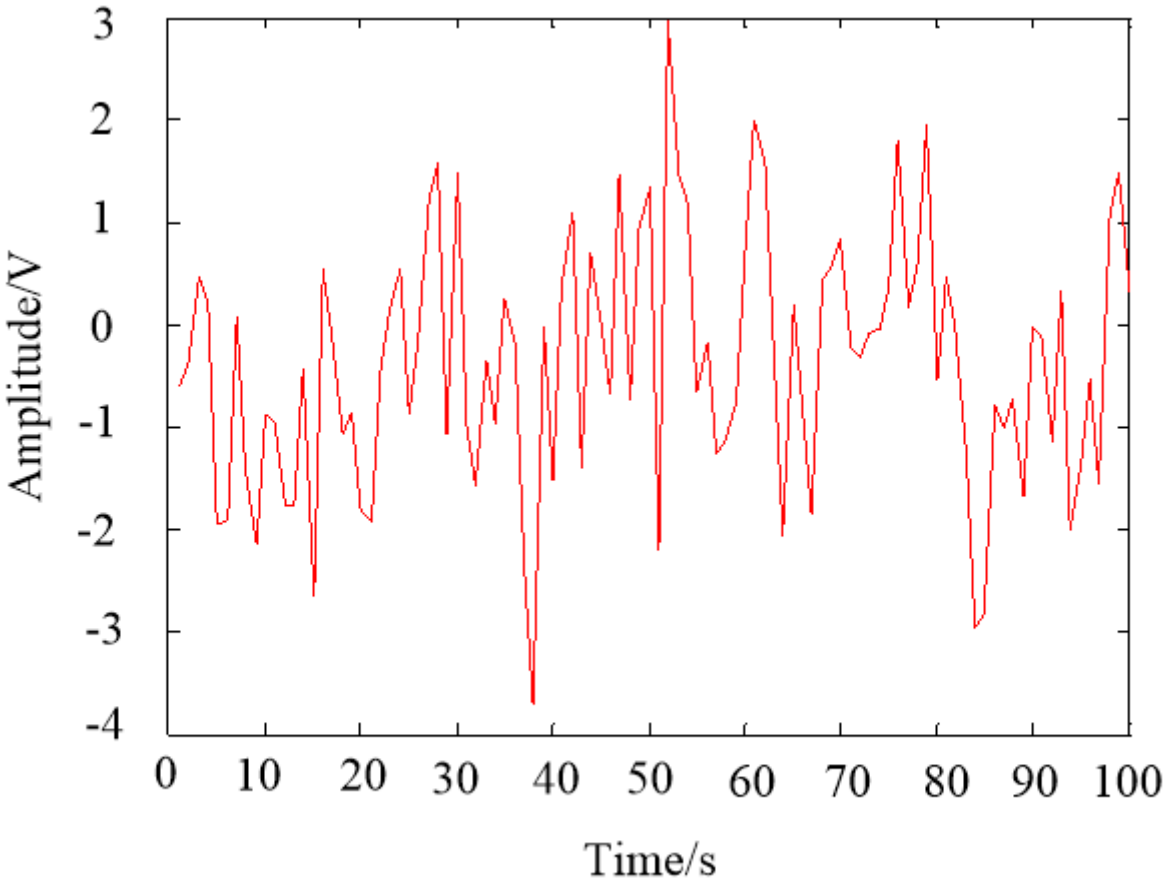

Figure 4

Data compression output