Redundant data high-efficiency compression based on distributed parallel algorithm

Jianhu Gong

Accepted: 11 August 2021 / Published online: 7 September 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract
Redundancy occurs when multiple consumers of the network attempt to access the similar content online. Many state-of-the-art studies have been proposed to remove redundancy from the network and to enhance the network efficiency of Internet applications. Still there is a necessity to explore more data redundancy elimination (DRE) techniques, especially for defense-based networks to enhance the data transmission speed and to compress the irrelevant data optimally for reducing the network latency. In order to improve the optimal storage capacity of redundant data over serial hybrid network cascade database, a high-efficiency redundancy elimination method based on the distributed parallel algorithm is proposed in this paper. The distributed storage structure model for handling redundant data of serial mixed network cascade database is designed and tested in simulated environment. The extraction of features of redundant data is performed by using distributed hybrid feature mining technique. The dimensionality reduction of redundant data is also attained by devising the feature transformation method to remove the unwanted features of data. The two benchmarked techniques have been selected for comparative study and to evaluate the performance of the proposed method. The simulation results show that the proposed DRE method can significantly reduce the redundancy from the network traffic. The bandwidth utilization is improved, the duplicate data are also compressed optimally and it is proved that the proposed DRE method is viable for large networks to eliminate the redundant data.

Keywords Distributed parallel algorithm · Protocol-independent redundancy elimination (PIDE) · Data redundancy elimination (DRE) · High performance · Compression

1 Introduction
Redundancy in a network takes place when multiple users attempt to use the Internet or attempts to access the similar content online (Lei et al. 2012). This allows repeated transfers of the similar content or data on the Internet and networking applications. Many research studies have been proposed by the academicians and researchers to remove the redundancies from the network traffic and to improve the network efficiency for Internet applications. An elimination of traffic redundancy is an effective solution for reducing the bandwidth costs, for reducing the network latency and for improving the efficiency of the networks (Anand et al. 2009a). Numerous studies are presented in the literature by the authors where diverse techniques have been devised for eliminating the redundancies from the network traffics. In military- and defense-based networks where network bandwidth is very imperious parameter for judging the health of the network to transmit the crucial information speedily, it is imperative to consider some smarter mechanisms that can eliminate the redundancy in network traffic efficiently and can optimize the bandwidth utilization.

Protocol-independent redundancy elimination (PIDE) is one of the popular methods used to eliminate redundancies from the network traffics (Abualigah et al. 2020). Numerous TRE solutions have been suggested so far, together with the middlebox-based solutions (Johansson et al. 2015). The middleboxes are placed at any of the wide-area-
network link, and end-to-end solution is devised at the client as well as at the server side. Redundancy exclusion middleboxes are used to enhance the bandwidth of network, for improving the access to the network links and for balancing the link loads in the areas with the smaller networking coverage. PIDE technique can be employed on multiple network routers, which are designed to support redundancy elimination at IP layer. PIDE assists in eliminating redundancy in multiple end-to-end applications (Halepovic et al. 2012a). In defense, the private cloud servers play major role for providing data and computational services to fulfill the requirements of defense-based applications. In order to optimize bandwidth cost and to reduce the data transfer time from cloud servers to application destinations, DRE (data redundancy elimination) techniques are to be refined. The quick delivery of information in the defense-based networks is a significant factor for judging the network efficiency and network security; therefore, smarter DRE techniques are required for the faster transmission of data and for improving the strength of the network. The delivery of the important information should not be delayed due to latencies in the network, redundancies in the traffic and over-consumption of the bandwidth.

Content delivery network (CDN) traffic is packed into streams; CDNs are distributed geographically with the help of proxy servers and data centers to provide quality services to the end users. For maintaining high levels of QoS (quality of services) with the increasing traffic volumes, it is indeed a challenging job to distribute the network services spatially and to control the data redundancy over the networks. New mechanisms are to be explored for eliminating data redundancies from the CDNs. The biggest challenge is the cost for operating CDNs. The cost depends on the volume of multimedia traffic exchanged over the CDNs. By analyzing the redundancy on CDN networks with the help of network data types, this paper explores the data redundancy elimination techniques for different network data types. In this paper, an attempt is made to research on the redundancy elimination novel technique for content-based networks, which can certainly improve the efficiency of networks for transmitting data smoothly by enhancing the bandwidth capabilities of the network. The mainstream data redundancy elimination technique is incorporated into the serial hybrid network infrastructure on the basis of distributed parallel algorithm.

As Internet traffic is growing at a rapid rate; defense-oriented applications require more delicate networks, which are capable to handle redundancies at the network level, and can provide QoS to the users. The redundancy in Internet traffic arises due to highly skewed popularity distribution of the content available at Internet. As a result, there are many repeated transfers of the similar content; both in client–server- and in P2P-based networking applications. From a logical perspective, repeated transfers of similar data were prone to over-consumption of network resources. In real-world scenario, redundant traffic can be a particularly serious problem for limited-bandwidth Internet access links. Redundant traffic can also be an issue economically, if Internet providers or users are charged based on the traffic volumes sent and received between peering points, i.e., usage-based billing, then the usage cost also goes high for the end users. Redundant traffic has to be controlled in a step-by-step approach for optimal usage of the network resources and to reduce the operational costs of the networks. Hence, elimination of redundant traffic from the networks motivated us for the research work presented in this paper.

Understanding the basic issues regarding network redundancy for improving the protocol design and ensuring network efficiency is need of the hour. Feature selection algorithms filter out redundant data features to reduce the dimensions of network data. Feature extraction is classified into three types, which are wrapping model, filtering model and embedded model. In the filter model, the features are assessed without predictor and use ranking gathering mechanism. A filter method has low computational cost and can also handle large datasets. The wrapping model uses classifiers to assess the key features subsets, whereas embedded model works by linear classifier (Pacella 2018). It is considered to have lower computation complexity than other models. The data fusion algorithm is used to extract features of data in our proposed work. It combines time and spatial correlation to extract relevant features of the data; removes redundant as well as erroneous data; and retains relevant information to form a set of multi-angle data. Data fusion also reduces energy consumption by reducing the consumption of communication bandwidth and energy. By filtering the redundant data, data transmission can be accelerated and the overall network efficiency also improves. Feature reduction is also performed on the extracted data, which is a common method to be used for data dimensionality reduction in the networks after extraction of features. It reduces the data redundancy by deleting unnecessary features of the data. In our work, we have made use of a distributed hybrid feature mining algorithm for feature extraction of redundant data from the serial networks, and then, feature reduction mechanism is devised to remove the irrelevant features. Finally, the output data are compressed optimally to improve the overall efficiency of the network.

This paper proposes a novel DRE mechanism for serial hybrid network cascade database (SHNCD) based on distributed parallel algorithm to handle the redundant data. The contributions of the paper are stated below:
A distributed storage structure model is designed for handling redundant data of serial hybrid network cascade database.

A distributed hybrid feature mining algorithm is used for extraction of associated features of redundant data.

Dimensionality reduction of redundant data is also performed using data fusion algorithm in distributed environment.

A mechanism for high data compression is also devised at the end that improves the optimal storage capacity of redundant data in serial hybrid network cascade database.

A simulation experiment is carried out to demonstrate the superior performance of the proposed method for high-performance compression capability and storage capability of the serial hybrid network cascade database. The experimental results demonstrate the superior performance of the proposed DRE method over the benchmarked approaches.

The rest of the paper is organized into six parts: In Sect. 2, the literature is reviewed. The proposed DRE mechanism is described in Sect. 3. Section 4 describes the experimental results and analysis on results. The conclusions of the paper and future research directions are presented in Sect. 6.

2 Related work

Several research papers have been published by the researchers on the various aspects of data redundancy in the area of networking. Data redundancy approaches are categorized into two types: object-level redundancy and packet-level redundancy. Traditional data redundant elimination is performed at the application layer such as data compression that removes the redundant data within the caching objects, and peer-to-peer caches, which are installed to serve the frequent requests. However, the application-layer caching is not efficient in eliminating data redundancy from the network traffic (Rekha and Raju 2019). Data redundancy elimination methods based on the personalized web objects are not detected by the traditional web caching platforms like Squid (Zhu et al. 2020). The protocol-independent technique is also proposed in the literature to find the content similarity for reducing the storage overhead (Priya and Enoch 2018a). The author presented a model where simulation of data localization is performed under redundant environment by employing the redundancy suppression technique for efficient network traffic analysis (Wang and Wang 2016). The authors have presented a trace-driven study of PIDE mechanisms. The finding states that 75 to 90% bandwidth of middlebox is consumed due to redundant byte-strings received from the client side traffic. They have suggested an end-to-end elimination solution for data redundancy to save bandwidth of the middlebox (Anand et al. 2009b).

In (Kaur and Singh 2016), authors have presented a study based on packet payloads to examine the usefulness and proficiency of packet-level PIDE methods. Various array of redundancy detection algorithms have been compared in terms of their usefulness and proficiency. Novel approaches are proposed in this study for removing the redundancies. From their experimental result analysis, it is also found that redundant data matches in traces follow a special distribution, which results in lower compression by increasing the data cache size. Paper (Halepovic et al. 2012b) describes a study on numerous parameters that affect the practical efficacy of PIDE techniques. They have used strategies to improve the current TRE protocol-independent procedures on the basis of their experimental work. In (Zhang and Ansari 2014), the authors have conducted a study on different potential techniques involved in traffic redundancy elimination such as fingerprinting, cache management and chunk matching mechanisms. They have offered modifications on current redundancy elimination systems for overall performance improvement. In (Ditto: A system for opportunistic caching in multi-hop wireless networks. 2008), application-independent caching is proposed to enhance the quantity of data transfers. Web applications are modified by introducing data markers that assist in elimination of traffic redundancy.

In (Priya and Enoch 2018b), redundant traffic elimination (RTE) is used for finding and removing duplicate chunks of data by byte-by-byte scanning of the network layer. It is used for finding out identical content, which works on heterogeneous traffic. In (Failed 2018), REDA based on pattern generation approach is proposed. It is particular to the sensed data. The redundant data are collected from the sensor nodes from the similar cluster. The results of REDA prove that energy consumption is hoarded up to 44% without using data aggregation techniques. In (Failed 2019), MLDAT approach is proposed that produces useful data for all large networks by eliminating the unwanted data. It prioritizes the data based on the sending query to the destined nodes. The aggregator nodes forward the valuable data to the sink node to reduce the latency, and for better bandwidth utilization. The authors present the data fusion techniques for eliminating the redundancy in WSNs. The information is extracted with respect to redundant data to provide relevant data in energy-efficient way. Data accuracy is also preserved.

Many existing works have contributed for solving the problem of data redundancy over the networks. Some techniques extract the features of data and then minimize the redundancy; a few techniques use feature reduction
techniques to reduce dimensionality of data; and many techniques are based on compression of data, but there is no distributed and parallel technique, which can combine all the redundancy techniques and obtain redundancy free data by reducing large volume of data into manageable chunks. There is still a scope to explore new techniques that can enhance the bandwidth capability, reduce the network latency and enhance the overall efficiency of the network. Hence, we are proposing a novel technique for redundancy elimination in this paper to minimize the redundancy optimally, and the proposed DRE method uses a combination of all the redundancy mechanisms and is deployed in distributed and parallel fashion to yield better output.

3 Proposed method

The proposed DRE model works in four steps.

Step 1. The data are collected from serial hybrid network cascade database for ascertaining the redundancy, and fuzzy logic-based approach is used for automatic allocation of location for database. By using fuzzy logic, database of redundant data is obtained.

Step 2. The next step is to extract the features of the redundant data. The DRE module collects the information of the data block corresponding to each data fingerprint at runtime; the recent data block utilization rate and the contribution of data block are used as factors for ranking of the redundant data. Once the features of the redundant data are extracted irrespective of types of data and type of network, then the feature reduction module is activated.

Step 3. The feature reduction module removes the unwanted features to reduce the burden of redundant data and to promote the flow of relevant data through the network. A distributed parallel algorithm is used, which works in parallel for feature extraction, feature reduction and data compression to remove redundancy from data.

Step 4. Finally, the data without redundant features are available; then, high compression algorithm is devised to compress the data, to optimize the bandwidth over the network channels and to improve the overall efficiency of the network.

3.1 Data collection

Data redundancy elimination technology is evolved due to realization of the need of elimination of the duplicate data over the networks and to safeguard the network traffic (Klapa et al. 2017). Data blocks with high redundancy rate are analyzed from TCP byte stream and are used as redundancy dictionaries to replace duplicate data with smaller data fingerprints in order to eliminate the redundant part in network traffic. The redundant data block selection algorithm used by the data compression engine is based on data package level. The redundancy mode of the data package is analyzed to further decide upon the data compression, which is independent of the protocols used in the heterogeneous networks.

In order to realize high-performance compression of serial hybrid network cascade database (SHNCD), redundant data are eliminated on the basis of distributed parallel algorithm in this paper. A distributed storage structure is designed for handling redundant data, and then, a distributed hybrid feature mining method is deployed to collect redundant data for SHNCD. An automatic allocation of location for database of redundant data is determined by combining the multi-feature transformation method and a high-performance compression model. Using the control technology, the fuzzy correlation set of the serial hybrid network cascade database of redundant data is obtained to satisfy \( A \cap B \subseteq V \) and \( A \cap B = \emptyset \). 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 \( X = \{ x_1, x_2, \ldots, x_n \} \), cross feature items of multilevel serial network contact points are calculated. Ontological models A and B sets 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 proposed to eliminate the data redundancy at the packet level. In order to eliminate the data at the packet level, there is a need to analyze the bytes. Initially, winnowing algorithm is applied for similarity detection; it divides the substrings by selecting the window value and then selects the data fingerprint as the document feature; it then compares the document feature to detect the similarity. The motive of compression algorithm is to select data blocks with high redundancy from the network traffic at packets level to eliminate the redundant information (Verma and Singh xxxx), and then, the data are analyzed for the entire network. The focus of this paper is to adopt different redundancy elimination strategies for data received from different media in the network to achieve effective way for eliminating redundant data from the diverse network traffic.

According to data compression quality disturbance, the transmission data set of serial networks is obtained as given in Eqs. 1 and 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. The quantitative recursive analysis method is implemented to adjust the output of the redundant data of the serial hybrid network cascade database, and for the adaptive output control of the redundant data of the serial hybrid network.

The adaptive output is formulated as shown in Eq. 3.

\[
x_n = a_0 + \sum_{j=1}^{M_{d}} a_j x_{n-j} + \sum_{j=0}^{M_{a}} b_j y_{n-j}
\]

where \( a_0 \) is the sampling amplitude of distributed communication information in the serial network and \( x_{n-j} \) is the scalar time series of network-carrying capacity. A fuzzy clustering model is established as shown in Eq. 4, to obtain the characteristic acquisition output \( Z \) of the serial hybrid network cascade database redundant data obeying Gaussian distribution with parameter \( \beta_d \).

\[
\beta_d = (MPDist - d + 1)/MPDist, \quad d \in [2, MPDist]
\]

where \( adj(a, c) \) denote the correlation probability distribution of redundant data of SHNCD, and the optimized storage structure expression of redundant data of SHNCD is obtained as shown in Eq. 5.

\[
p(y|x, \theta) = \sum_{k=1}^{K} \pi_k p_k(y|x_k, \sum_k)
\]

According to the oscillation-type fluctuation of redundant data in the cascade database of the serial hybrid network, the output is shown in Eq. 6.

\[
C_T(f)Y_T(f) = C_T(f)\sum_n x(f - \frac{n}{f}) \cdot e^{j2\pi(f - \frac{n}{f})t_0}
\]

\[
= C_T(f)X(f) e^{j2\pi t_0}
\]

According to the above-mentioned steps, a multi-parameter information fusion model of serial hybrid network cascade database redundant data is established, fuzzy clustering is performed according to the data acquisition results and optimal mining of SHNCD redundant data is proposed.

### 3.2 Data feature extraction

Before the network transmission, the 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 to compress the data during the transmission. Since both ends of the data transmission maintain a common redundancy dictionary, as soon as the data reach at the receiving end, it is replaced by the original data from the redundancy dictionary and achieves the effect of decompression. In the application of CDN service environment, when analyzing all the undifferentiated network data, the block selection strategy of the data redundancy elimination (DRE) strategy is selected in the DRE compression engine.

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. If different algorithm window values are used to generate redundant dictionaries for different types of traffic, the data redundancy elimination effect of different types of network data will be the most useful (Procedia Computer Science 2018). Because of the end-to-end data redundancy elimination, redundant dictionary is used to record the corresponding relationship between redundant data block and the 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. 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 an updation is required. Some data fingerprints with high contribution rate need to be replaced when required, so it is necessary to rank the compression effect of data fingerprints in redundant dictionary. According to the ranking, the fingerprint of the data is determined, which is required to be replaced.

The DRE module collects the information of the data block corresponding to each data fingerprint at runtime; 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 used as factors for ranking of the data. The replacement algorithm uses the collected information to determine the replacement object. For the multimedia 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, judgment logic is required to determine what type of data is currently processed, so as to select corresponding redundancy dictionaries for search.
and replacement. For proving the viability of the proposed DRE mechanism, the research study is performed on TCP connection-oriented network as it provides secured connectivity.

The information of TCP flow starts from the recovery of TCP connection. The recovery of TCP connection is based on the analysis of the collected network data packet and from the protocol field of IP packet. Then, the connection establishment and disconnection are completed according to the SYN (synchronization bit), FIN (finish bit) and RST (reset bit) of TCP packet header. The key information is available in the source IP address, destination IP address, source port number and destination port number (TCP port 20 and 27). The specific TCP connection simulation recovery process is as follows:

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

B. The source and destination IP addresses’ source and destination port numbers in the IP package are consumed 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 SYC and ACK, the packet is the response of the server during the formation of the connection by the 3-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 FIC 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 related features of the redundant data of the SHNCD, and the method of feature transformation is combined to reduce the dimension of the redundant data of the SHNCD, and the feature transformation quantization learning function of the remote communication number of the cascade database is expressed in Eq. 7.

\[
\min \sum_{1 \leq i \leq K, e \subseteq k(e)} \frac{f(e(i))}{C(e,i)} \\
0 \leq f(e,i) \leq C(e,i) \\
F = \text{const} \\
\sum_{1 \leq i \leq K, e \subseteq k(e)} \frac{f(e(i))}{C(e,i)} + \sum_{e \subseteq k(e)} \frac{f'(e(i))}{C'(e,i)} \leq k(v)
\]

(7)

Combined with the random feature transformation compression method, the feature extraction is performed, and the binary coding of redundant data in SHNCD is performed. The low-frequency flicker distribution is obtained using Eq. 8.

\[
\text{Computation}(n_i) = (E_{elec} + E_{DF})\delta + E_{TSL(d_i)} \\
= (E_{elec} + E_{DF})\delta + lE_{elec} + l_s d_f^2 \\
= [(E_{elec} + E_{DF})\delta + E_{elec} + l_s d_f^2] l
\]

(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 performed in combination with fuzzy update rules as shown in Eq. 9, and the zero-frequency feature distribution threshold is obtained:

\[
\gamma_k^w(\omega) = E(T_k^w | T_k^w > \xi_k^w(\omega)), \ k \in R_w, w \in W
\]

(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 shown in Eq. 10.

\[
\xi_k^w(\omega) = \min \{\xi | \text{Pr}(T_k^w \leq \xi) \geq \omega\} = E(T_k^w) + \gamma_k^w(\omega) \\
\]

(10)

Extract a high-order spectrum statistical characteristic quantity of redundant data of a SHNCD and obtaining a statistical characteristic distribution index set of redundant data of SHNCD in a neighborhood space \((t,f)\) is expressed in Eq. 11.

\[
f(x) = \begin{cases} (x), x \in \text{Levf} \\ a, x \in \text{Levf} \end{cases}
\]

(11)

Calculating the cost function with high-performance compression, as per the distribution of frequent itemsset, the information fusion model of serial hybrid network cascade database redundant data is obtained as \(E(T_k^w - \xi_k^w(\omega) | T_k^w \geq \xi_k^w(\omega))\), and the global optimal solution of serial hybrid network cascade database redundant data can be expressed as shown in Eq. 12.

\[
s_h^w = E \left[ \min_{k \in R_w} \{H_k^w | \eta^w\} \right] = -\frac{1}{\theta} \ln \sum_{k \in R_w} \exp \left(-\theta \eta_h^w(\omega)\right), \ w \in W, h \in H
\]

(12)
A high-order spectrum feature extraction method for redundant data of SHNCD is established, 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 grouping capabilities. Data are mostly transmitted between network servers and device terminals. There are consistent redundant dictionaries at both ends of data transmission. The generation of redundant dictionaries depends on the analysis of the content of data packets, selection of data blocks with higher redundancy rate to calculate data fingerprints, and establishment of the corresponding relationship between redundant data blocks and data fingerprints. Then, before data transmission, the redundant data block in the redundant dictionary is utilized 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.

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, the data are again analyzed. After calculating the data fingerprint, it is determined whether the redundancy dictionary exists in the dictionary. If it exists in the dictionary, the fingerprint used 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 arrive 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. After collection of redundant data and extraction of features of redundant data by applying the distributed hybrid feature mining method to extract the association features of the database of redundant data, the next step is the reduction of irrelevant features that are extracted and added in the dictionary.

### 3.3 Feature transformation dimension reduction

The distributed hybrid feature mining method is adopted in the previous subsection to extract the relevant features of the database of redundant data, and then, the high-performance compression processing is performed to reduce the transformation dimensionality based on the distributed parallel algorithm in this subsection. The feature extraction model of fuzzy association rules for redundant data of cascade databases is designed, and the automatic location allocation in the storage process of redundant data is carried out by adopting a high-order spectral decomposition method, and the feature transformation dimensionality reduction processing is performed by high compression of data.

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 (Mohapatra et al. 2020). 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 of the data, the data size is counted after compression by the DRE engine. The intermediate results of the analysis of data by the DRE engine are shown in Fig. 1.

Since the intermediate result is in the form of key value, map reduction 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. After the intermediate result collection is completed, the map reduce script is executed 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, and put them into the output stream. The 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, one can use the map reduce program after collecting the intermediate results of each node every day or in a fixed interval of time. In this way, the implementation of handling 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 are 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 window size to generate different redundant dictionaries to find the best window size 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 expressed in Eqs. 13, 14 and 15.

$$q^w = E(Q^w) = \sum_{k \in R_w} f_k^w, \ w \in W$$ \hspace{1cm} (13)

$$v_a = E(V_a) = \sum_{w \in W} \sum_{k \in R_w} \delta_{w,k} f_k^w \cdot a \in A$$ \hspace{1cm} (14)

$$f_k^w \geq 0, \ k \in R_w, \ w \in W$$ \hspace{1cm} (15)

The high-performance compressed binary structure model of redundant data is obtained as expressed in Eq. 16.

$$\min_x \frac{1}{2} \left\{ \sum_{i=1}^{l} \sum_{j=1}^{l} y_i z_j x_i x_j K(x_i, x_j) - \sum_{j=1}^{l} z_j \right\}$$

s.t. $$\sum_{j=1}^{l} y_j z_j = 0$$

$$0 \leq z_j \leq u(x_j) C \quad j = 1, 2, ..., l$$ \hspace{1cm} (16)

In the reconstructed phase space of the redundant data distribution of the SHNCD, the unbalanced degree fusion method is adopted for dimensionality reduction of the redundant data, and the quantized characteristic distribution set of the redundant data of the serial hybrid network cascade database is defined as

$$D_i = \{S_{ij}(t), T_{ij}(t), U_{ij}(t)\}$$

wherein $S_{ij}(t)$ represents the adaptive weight of the redundant data of the serial hybrid network cascade database, $T_{ij}(t)$ represents the fused grouping characteristic of distribution set of the redundant data that belong to the serial hybrid network cascade database, $U_{ij}(t)$ represents a similarity (correlation) model, the redundant data of the serial hybrid network cascade database are subjected to block-wise high-performance compression and the output is represented in Eq. 19.

$$S_{ij}(t) = \frac{p_{ij}(t) - sp_{ij}(t)}{p_{ij}(t)}$$ \hspace{1cm} (17)

wherein $T_{ij}(t)$ represents the fuzzy feature distribution set of redundant data for serial hybrid network cascade database, as expressed in Eq. 18.

$$T_{ij}(t) = \frac{[p_{ij}(t) - \Delta p(t)]}{p_{ij}(t)}$$ \hspace{1cm} (18)

According to the above analysis, the high-order spectral feature quantity of the redundant data of the SHNCD is extracted, and the classification processing of the redundant data of the SHNCD is carried out by using the methods of statistical information mining and fuzzy information clustering analysis; the feature dimension reduction of the redundant data of the SHNCD is realized by using the method of feature transformation, so that the redundant data can be efficiently compressed as shown in Fig. 2.

3.4 Redundant data compression for TCP-oriented traffic

The processing of TCP packets is distributed as the TCP packets are truncated when they are very lengthy; only the first packet will have the HTTP protocol header, which
leads to the problem of correct identification of packets of a TCP stream in an accurate order. The existing TCP packet reassembly is to ensure the correct sorting of TCP packets according to the transmission sequence numbers of the TCP packet header. However, for a large number of network packets, sorting analysis by serial number requires many resources, and it eventually reduces the efficiency of the analysis. 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. To correctly identify each data type and quantitatively count the data traffic size are two important aspects of TCP-oriented traffic. The connection methods of HTTP protocol include non-persistent connection and persistent connection. For the non-persistent connection mode, the specific establishment steps are: (i) 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. (ii) 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 is used to communicate from the establishment of TCP connection to the closure of TCP connection, in which the order of sending data should be sequential and it should not continue to send the data of the second request until the data of the first request are not sent. 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 the TCP packets, the content-type field in the HTTP protocol header is extracted, and then, the content-type information of the current TCP stream can be updated. Before the next content-type updation, 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. In this way, a variable recording the content-type value can be used to identify the data type in multiple HTTP transfers.

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 window-size set during the TCP handshaking process.

B. Adjust window size to generate different redundant dictionaries to find the best window size for network data such as text type. The process uses—w to specify different window size, single redundancy dictionary mode, and statistics the efficiency of data redundancy elimination through the results comparisons to find the most appropriate window size. 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.

E. In the reconstructed phase of the redundant data distribution of the SHNCD, the unbalanced degree fusion method is adopted for high-performance compression, and the quantized feature distribution set of the redundant data of the serial hybrid network cascade database.

4 Experimental results and analysis

In order to test the application performance of the proposed DRE method in SHNCD, a simulation experiment is performed by using MATLAB. The sampling sample length of the redundant data of the serial hybrid network cascade database is 2000 tuples, the code bit sequence length of the data is 120, the spatial embedding dimension of the redundant data 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 10 dB. According to the above simulation parameter settings, proposed data redundancy elimination method is applied along with the conventional method and state-of-the-art methods for comparative study on the redundant data of the SHNCD.

It can be depicted from Fig. 3 that for 100% redundancy, the proposed DRE method provides occupancy of bandwidth near to zero; the method is capable to avoid redundant data and to utilize the bandwidth optimally. The methods considered for comparative study are temporal correlation perceptual data de-redundancy method (TCPDD) and hybrid algorithm for redundancy and concurrency control (HARCR), which also yield good output, but the proposed DRE method outperforms these methods. In the conventional method, the data aggregation and redundancy elimination happen in a sequential manner; therefore, the occupancy of bandwidth is quite high, whereas the distributed methods utilize bandwidth optimally by eliminating the redundant traffic.

It can be depicted from Fig. 4 that the proposed DRE method provides better utilization of the bandwidth as compared to other methods; the method is proficient to reduce the redundancy of data and to enhance the utilization of the channel bandwidth optimally. The methods considered for comparative study are also capable to enhance the utilization of the bandwidth, but as soon as the data increase over the network channels, their performance deteriorates with respect to the bandwidth utilization. Since bandwidth is the most important issue for the smooth flow of network traffic, the redundancy techniques should be proficient enough to optimize the bandwidth utilization.

Looking at the data in Fig. 5a as the research object, the redundant data available at serial hybrid network cascade database is presented. Feature transformation is used to achieve high-performance compression of the data, and the compressed output is shown in Fig. 5b. It is also observed that the data compression can be effectively comprehended by using the proposed DRE method, and the storage cost of data will also be reduced eventually. Testing of the fidelity of high-performance compression output of redundant data by different methods is shown in Table 1 and Fig. 6, which reveals that the fidelity of redundant data compression is higher as compared to the state-of-the-art methods.

The application of big data leads to the increasing capacity demand of storage system. In order to save cost, to reduce energy consumption, to minimize the actual data storage, and to store the appropriate data in the right place, hierarchical storage is often used in large data centers. The proposed DRE method not only provides redundancy elimination but also compresses the data for effective utilization of the bandwidth.

Fig. 3 Average bandwidth occupancy achieved by the proposed and other methods
5 Conclusions

In this paper, an optimized storage and transmission model of serial hybrid network cascade database for handling of redundant data is proposed to improve the remote communication transmission and adaptive redundancy control capability of cascade database. This paper proposes a high-performance compression algorithm to minimize the effect of redundant data based on the distributed parallel algorithm. The method comprises four stages: it begins with the extraction of high-order spectrum statistical characteristic of redundant data of a serial hybrid network cascade database, then the high-order spectrum decomposition method for automatic location allocation in the storage process of redundant data is devised, and the third phase reduces the irrelevant features of data, the fourth phase performs high-efficiency compression to reduce the redundancy optimally. The proposed DRE method basically carries a characteristic transformation method to realize feature dimensionality reduction of redundant data of the serial hybrid network cascade database and

---

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

| Number of iterations | Proposed DRE mechanism (Zhu et al. 2020) | TCPDD method (Zhu et al. 2020) | HARCR method (Klapez et al. 2017) |
|----------------------|------------------------------------------|-------------------------------|-----------------------------------|
| 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                             |

---

**Fig. 4** Average bandwidth utilization by the proposed DRE and other methods

**Fig. 5** (a) Redundant data from the original SHNCD, (b) data compression output by proposed DRE
comprehend high-efficiency compression of redundant data. The research outcome shows that the proposed DRE method has high fidelity of communication output when compressing redundant data in SHNCD; it optimizes the bandwidth utilization and bandwidth occupancy. The ultimate purpose is to alleviate the rapid growth of storage system space demand, to reduce the actual storage space occupied by data, and to reduce the data management costs. In future work, we will also attempt to calculate the cost-effectiveness from the proposed DRE method by setting a cost function through the fuzzy logic-based model. Therefore, practically, an attempt will be made to introduce the capacity reduction technology based on variable length block granularity into the hierarchical storage mode for big data storage management that will optimize the bandwidth usage and also reduce the cost of data handling.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Abualigah L, Aldulaimi A, Shinwan Al, Mohammad & Shehab, Mohammad. (2020) A Proposed Hybrid Feature Selection Method for Data Mining Tasks. International Journal of Science and Applied Information Technology. https://doi.org/10.30534/ijssat/2019/218620198

Anand A, Muthukrishnan C, Akella A, Ramjee R (2009a) Redundancy in network traffic: findings and implications. ACM SIGMETRICS Perform Evaluation Rev. https://doi.org/10.1145/1555349.1555355

Ashok Anand, Chitra Muthukrishnan, Aditya Akella, and Ramachandran Ramjee. (2009) Redundancy in network traffic: findings and implications. In Proc. of SIGMETRICS, 2009.

Dogar, Fahad & Phanishayee, Amar & Pucha, Himabindu & Ruwase, Olatunji & Andersen, David. (2008). Ditto: A system for opportunistic caching in multi-hop wireless networks. 279–290. https://doi.org/10.1145/1409944.1409977.

Fitzek FHP, Seeling P, Höschele T, Jacobfeuerborn B (2021) On the need of computing in future communication networks. Comput Commun Netw. https://doi.org/10.1016/b978-0-12-820488-7.00011-6

Halepovic E, Williamson C, Gadheri M (2012a) Enhancing redundant network traffic elimination”. Comput Netw 56(2):795–808

Halepovic E, Williamson C, Ghaderi M (2012b) Enhancing redundant network traffic elimination. Comput Netw 56:795–809

Kaur M, Singh J (2016) Data de-duplication approach based on hashing techniques for reducing time consumption over a cloud network. Int J Comput Appl 142:4–10

Lei Yu, K. Sapra, Huaying Shen and Lin Ye. (2012) Cooperative end-to-end traffic redundancy elimination for reducing cloud bandwidth cost, 2012 20th IEEE International Conference on Network Protocols (ICNP), 2012, pp. 1–10, doi: https://doi.org/10.1109/ICNP.2012.6459973.

Martin Klapacek, Carlo A. Grazia, Maurizio Casoni, A hybrid algorithm to combine redundancy and concurrency in virtual network resource pooling. 2017. DOI. https://doi.org/10.1102/it2.18

Mohapatra S., Bajpai N., Swarnkar T., Mishra M. (2020) Raw Data Redundancy Elimination on Cloud Database. In: Das A., Nayak J., Naik B., Dutta S., Pelusi D. (eds) Computational Intelligence in Pattern Recognition. Advances in Intelligent Systems and Computing, vol 1120. Springer, Singapore. https://doi.org/10.1007/978-981-15-2449-3_34

Neetu Verma, Dinesh Singh, Data Redundancy Implications in Wireless Sensor Networks.

Pacella M (2018) Unsupervised classification of multichannel profile data using PCA: an application to an emission control system. Comput Ind Eng 122:161–169. https://doi.org/10.1016/j.cie.2018.05.029

Priya D, Enoch S (2018b) The effect of packet redundancy elimination technique in sensor networks. J Comput Sci 14:740–746. https://doi.org/10.3844/jcssp.2018.740.746

Priya D, Enoch S (2018) The effect of packet redundancy elimination technique in sensor networks. J Comput Sci, pp.740–746. Procedia Computer Science, Volume 132, 2018, pp. 1210–1217, https://doi.org/10.1016/j.procs.2018.05.036.

R. Johansson, A. Horndahl and M. Rosell. (2015) A data association framework for general information fusion,” 2015 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI), pp. 233–239, doi: https://doi.org/10.1109/MFI.2015.7295814.

R. Nagesh, S. Raga and S. Mishra, ”Elimination of redundant data to enhance wireless sensor network performance using Multi level data aggregation technique,” 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2019, pp. 1–5, doi: https://doi.org/10.1109/ICCCNT45670.2019.8944647.

Rekha Ch, Raju S (2019) A survey on protocol independent technique for eliminating redundant network traffic and the power of prediction cloud bandwidth and operating cost. Int J Res 8:1065

S. Khriji, G. Vinas Raventos, I. Kammoun and O. Kanoun, ”Redundancy Elimination for Data Aggregation in Wireless Sensor Networks,” 2018 15th International Multi-Conference on Systems, Signals & Devices (SSD), 2018, pp. 28–33, doi: https://doi.org/10.1109/SSD.2018.8570459.

Wang L, Wang F (2016) Simulation of efficient data localization in network database under redundant environment. Compu Sim 33:364–367

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

Zhang Y, Ansari N (2014) On protocol-independent data redundancy elimination. Commun Surv Tutorials, IEEE 16:455–472
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 39:295–301

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.