Research on Non-structural Big Data Analysis Algorithm of Communication Network Based on Feature Fusion

Xin Meng*
Wuhan Qingchuan University, Hubei, Wuhan, 430000, China

*Corresponding author: baobwfe94533@163.com

Abstract. Nowadays, along with the fast development of information technology, the scale of data is increasing in an exponential way, and the worth of big data has been paid increasingly interested. At present, there are two key problems in the category of big data: how to delegate big data as a integrate model and what efficiently reduce dimensions of big data. Although the processing technology of various mold of data has been studied, the efficiency of parallel query has not been optimized deeply. Especially, it is more difficult to analyze the information of online users with widely dispersed information. A communication network non-structural big data analysis algorithm based on semantic relevance feature fusion is proposed, which reconstructs the high-dimensional phase space of the distributed information flow of cloud storage big data, extracts the semantic relevance dimension feature quantity of big data from the reconstructed phase space, and conducts adaptive learning training with the extracted feature quantity as the test set. Knowable from the simulation experiment that the algorithm effectively reduces the generation of redundant messages and improves the network running environment.

Keywords: Feature fusion; non-structural; Big data; feature extraction

1. Introduction
In non-structural P2P networks, object location will seriously consume network resources and restrict network scalability. Therefore, search algorithm is an important research topic. A basic search algorithm is flooding, which has short response time and high search success rate; But a big number of redundant request messages will be generative [1]. At the same time, because most network users frequently enter and leave the system, and will constantly update their favorite lists, the search algorithm should be able to adapt to dynamic environmental changes and rapid topological changes. Non-structural data may not have explicit connection with each other, and even can't be stored as relational data.

In the step of managing big data, big data administer consist of data model building, data storage, data cleaning, data application, etc. In the above four processes, how to build a concise data model is the best fundamental contact [2-3]. Big data fusion model designed to combine or succeed structured, semi-structured and non-structural data at various presentative horizontal, so as to obtain a unified a model that can fuse a large number of high dimensional data together. However, for the data with high dimension and large volume, the traditional data model can't deal with such complicated data [4]. This
paper proposes an non-structural big data analysis algorithm for correspond networks in view of semantic relevance feature fusion. Reconstruct the phase space of big data, extract the semantic correlation dimension features in the reconstructed phase space, and conduct adaptive learning training. Finally, the performance is verified by simulation experiments. The superiority of this method is demonstrated.

2. Non-Structural Data
The characteristic of non-structural data is that the data structure has no specific form, which can not be stored directly by relational database, but can only be stored by different types of files. Such as office documents, notepad files, jpg, tif picture files, log files, etc. [5].

non-structural data communication usually needs professional software to browse, and in terms of storage, it can only be converted into specific fields for storage, which has nothing to do with it. It is not conducive to adding, deleting, modifying and checking data in the future. The expression relationship of this database is not clear, so users can't directly get the information they want to provide from the data table. XML [6], a new data technology derived from SGML, can realize the mutual conversion between heterogeneous data. XML brings the technology of transforming non-structural data into structured data, which brings great convenience to the system. On the one hand, it can meet the requirements of different systems for their own file formats, on the other hand, it can achieve the unified distribution and management of data.

3. Implementation of non-structural big data analysis algorithm in communication network

3.1 Feature extraction and information fusion of semantic correlation dimension
Based on the high-dimensional phase space reconstruction of the distributed information flow of big data in cloud storage, this paper proposes an non-structural big data analysis algorithm based on semantic relevance feature fusion in communication network, and carries out feature extraction and mining optimize project of big data. The semantic correlation dimension feature amount of big data is pick up from the reconstructed phase space, and the limited data set of big data information flow in phase space is given as [7-8]:

$$X = [s_1, s_2, \ldots, s_K] = (x_n, x_{n-1}, \ldots, x_{n-(m-1)})$$ (1)

Considering the equivalent semantic mapping of two different attributes, the objective function of big data semantic relevance feature fusion is constructed by using the method of relevance dimension feature extraction:

$$X_p(u) = \begin{cases} \frac{1 - \frac{1}{2} \frac{1}{2} \alpha + \frac{1}{2} \alpha - \cos \alpha}{2\pi} e^{\frac{1}{2} \alpha + \frac{1}{2} \alpha - \cos \alpha} \int_{-\infty}^{+\infty} x(t)e^{\frac{1}{2} \alpha - \cos \alpha} dt, \alpha \neq n\pi \\ x(u), \alpha = 2n\pi \\ x(-u), \alpha = (2n \pm 1)\pi \end{cases}$$ (2)

In which: $x(t)$ is the ontology segment containing the attribute set of mining target in massive big data information, $P$ is the measurement distance between trajectory $x_k$ and $V_i$ in reconstructed phase space, which is deliver by euclidean distance as follows:

$$p = \|x_k - V_i\|^2$$ (3)

And satisfy

$$\sum_{i=1}^{\varsigma} \mu_{ik} = 1, k = 1, 2, \ldots, n$$ (4)

Optimize and solve the objective function of semantic relevance feature fusion to obtain the optimal
solution:

\[ F_j = \sum_{k=1}^{n} X_{kj}, Q_j = \sum_{k=1}^{n} \left( X_{kj} \right)^2 \]  

(5)

The multi-source fusion method of semantic relevance features is used for anti-interference processing [8], from which the semantic relevance dimension characteristic amount of big data is pick up, and the iterative formula of feature extraction is obtained:

\[ x_i(k+1) = (1 - \omega)x_i(k) + \left( b_i - \sum_{j=1}^{m} a_i x_j(k) - \sum_{j=m+1}^{n} a_i x_j(k) \right) \]

\[ i = 1, 2, \ldots, n; k = 1, 2, \ldots, n \]  

(6)

The extracted feature quantity is used as the test set to carry out adaptive learning training, and the limited data set \( X \) in each idle time slice is divided into \( c \) class, where \( 1 < c < n \), and the adaptive training weight is \( \omega = \left( \omega_1, \omega_2, \ldots, \omega_c \right)^T \), \( \omega_j \in [0, 1] \). Through adaptive training, improve the stability and convergence of big data mining process.

3.2 Analysis of restriction algorithm

It is assumed that each node has \( k \) child nodes. Calculate how many redundant messages can be reduced by using the above method. Any node connected with other nodes in the same layer or with non-child nodes in the lower layer will generate redundant request messages [9]. If a node is connected with a sibling node in the same layer, redundant request messages will not be generated by using the above method; If a node is connected with the child node of the lower sibling node and also connected with this sibling node, the above method will not generate redundant request information.

It is assumed that a condition is satisfied: if a node is connected with a child node of a lower-level sibling node, then it is connected with the sibling node. A node in the \( m \) layer now has an extra edge, and there are \( k^m + k^{(m+1)} \) possibilities for this extra edge to connect with the same layer node or the lower layer node; There are \( k + k^2 \) possibilities that satisfy the conditions of the algorithm and do not generate redundant request messages. The improved ratio is \( \left( k + k^2 \right) \left( k^m + k^{(m+1)} \right) \). The total improvement probability of the first three layers is \( 3 * \left( k + k^2 + k^3 \right) * 100\% \).

It can be seen that when \( m \) and \( k \) increase, the efficiency of improvement also decreases. Although the graph and the conditions are assumed, the real graph is the deformation of this graph. The efficiency of improvement also changes with the change of graphics. However, no matter what pattern, the improvement ratio decreases with the increase of \( m \) and \( k \). At the same time, it can be found that this method can play a certain role in the first three layers, and the request message has been transmitted to the fourth layer.

3.3 Unified representation framework of big data

The conventional way to multiply two tensors to get a new tensor, without considering the inner data connection. In the step of expressing data with tensors, not only the new tensor should be gain, but also the characteristics of the data in each tensor ought to retain unchanged. In this paper, structured, semi-structured and non-structural data are quantized firstly, and then multiple tensors are merged by tensor merging algorithm of semi-tensor product, which ensures the stability of data internal structure while merging tensors.

- Sub-sheet quantization model of multi-source heterogeneous data

Firstly, non-structural data, semi structured data and structured data are described by tensor according to their different data characteristics. Finally, each kind of data is expressed as a corresponding sub-tensor model, and the specific representation process of sub-tensor quantization
model is explained by using three kinds of data: ordinary video data, semi-structured text data and database table.

1) Sub-sheet quantization delegate method of non-structural video data

The major characteristics of videotape data are time framework, width and height of framework picture and color. Hence, a fourth-order tensor is able to applied to delegate videotape data in MP4 format, and the formula is as follows [10]:

\[ T_{\text{video}} = I_f \times I_w \times I_h \times I_c \]  

In which:
- \( I_f \) — Frame;
- \( I_w \) — Video width;
- \( I_h \) — Video height;
- \( I_c \) — RGB colors.

Sub-sheet quantization representation of semi-structured text data

The data base of semi-structured data is a focus of nodes, each node is a leaf node or an internal node, and each semi-structured data has a hierarchical structure, can be broken down into a tree structure. Fig. 1 is semi-structured data representing a movie and its stars. Among them, the elements are represented by circles, and the connecting lines between circles represent the relationship between elements. The tree structure diagram of semi-structured data is composed of root nodes, internal nodes and leaf nodes. Semi-structured data can be represented by triples, including tags, types and values of objects. In this paper, semi-structured data is delegate as a third-order sub-tensor, and the formula is as follows [11]:

\[ X \in R^{I_{en} \times I_{ec} \times I_{er}} \]  

In which:
- \( I_{er} \) — Identifies the rows of the matrix;
- \( I_{ec} \) — Identifies the columns of the matrix;
- \( I_{en} \) — The ASCII encoding of the element.

![Fig.1 Semi-structured data of a movie and its stars](image)

3) Representation method of structured data sub-sheet quantization model

Structured data is simply a database. Such as financial system, medical database, one-card data and other core databases. Organization data is data logically deliver and realized by two-dimensional table organization. Primarily cross over the dependent database to manage and store. Genre database is widely used to manage structured data. In a ordinary database table, a space is often delegate by numbers or characters, which is able to expressed as a matrix.
Tensor space fusion method based on semi-tensor product

In fact, the multiplication of tensor and tensor is to multiply some matrices in each tensor to get a new tensor. In traditional matrix multiplication, the number of rows in the front array and the back array of the matrix must be the same before the matrix multiplication can be meaningful. However, in the fusion process of tensors and tensors, the dimension of a some order of the tensor involved in the operation is different from that of another tensor. The following introduces a new multiplication between matrices, that is, the semi-tensor product between matrices. Semi-tensor product of matrices means that even if the number of front and back arrays of two matrices is different, the matrices can be multiplied meaningfully. Moreover, the generalized matrix multiplication can keep the main characteristics of the original matrix unchanged.

Set \( A \in M_{nm}, B \in M_{pq} \). The minimum common multiple \( t \) of \( n \) and \( p \) can be obtained, which is written as \( t = \text{lcm}(n, p) \), and the half-tensor product of \( A \) and \( B \) is defined as:

\[
A \propto B = \left( A \otimes I_{t/n} \right) \left( B \otimes I_{t/p} \right)
\]

In which \( \otimes \) is Kronecker product of matrix.

On the basis of ensuring that the original matrix multiplication rules are not destroyed, the semi-tensor product of matrix and matrix has good characteristics such as concise definition and convenient calculation. Cross over the represent in the former part, we have transformed non-structural data, semi-structured data and structured data into sub-tensor models with lower order. In this part, all sub-tensor ways with lower order are melting by means of semi-tensor product to obtain a unified tensor model with higher order.

Let \( A \in R_{l_1 \times l_2 \times l_3} \), \( B \in R_{l_1 \times l_2 \times l_3} \), tensor spread multiplication follows the following function:

\[
f = \begin{cases} 
A \propto B \to C, C \in R_{l_1 \times l_2 \times l_3} \\
A \propto B = \left( A \otimes I_{t/n} \right) \left( B \otimes I_{t/p} \right)
\end{cases}
\]

In the above formula, the operator \( \propto \) satisfies the associative law. That is, \( (A \propto B) \propto C = A \propto (B \propto C) \), the order of tensor is able to elongate to the extant tensor way through semi-tensor product operation in different directions. Structured data, semi-structured data and non-structural data can be expressed as low-order sub-tensors first, and then merged into high-order tensor space by tensor extension operator in view of semi-tensor product, thus realizing the unified representation of big data. The following is a summary initiate to the semi-tensor product of any two tensors, taking the semi-tensor product of two graphs as an example.

Data includes structured data \( d_u \), semi-structured data \( d_{semi} \) and non-structural data \( d_s \). As all types of heterogeneous data need to be processed, a integrate tensor way is able to deliver as:

\[
f : (d_u \propto d_{semi} \propto d_s) \to T_u \propto T_{semi} \propto T_s
\]

In formula (11), the independent variables respectively delegate structured data, semi-structured data and non-structural data. These various form of data are represented in the low-order sub-tensor space, and then melting by tensor spread operator in view of semi-tensor product and consistently represented as high-order tensor, which corresponds to a unite variable data structure in the computer system.

4. Experimental analysis

Simulate the establishment of a P2P network composed of 12 000 nodes. Each node in the network is randomly connected with other nodes, and the degree of each node is randomly selected between 2 and 10. Use any node as a source point to send out request information. The files to be searched are randomly placed on 10 nodes. Search files with the proposed selective flooding algorithm and flooding algorithm respectively, and compare the two algorithms. The two algorithms are compared by changing different parameters. On the whole seven layers, compare the improvement ratio of the selection.
algorithm suggest in this paper in the number of search messages compared with the flooding algorithm: that is, the quantity of information generated by the selection algorithm is 100% compared with the quantity of information generated by the flooding algorithm.

![Fig.2 Improve the change of ratio with average degree](image1)

Fig.2 Improve the change of ratio with average degree

It can get from fig. 2 that in the first four layers, along with the raise of the average degree of nodes, the improvement ratio of the flood limiting algorithm is also increasing. This is mainly because under the condition that the quantity of nodes in the network is unchanged, the average degree of a single node increases, which is conducive to the formation of the condition of not generating redundant request information. Similarly, the number of nodes can be changed, and other parameters remain unchanged.

![Fig.3 Change of improvement ratio with the total number of nodes (abscissa unit thousand)](image2)

Fig.3 Change of improvement ratio with the total number of nodes (abscissa unit thousand)

It can be seen from fig. 3 that with the increase of the number of nodes, the improvement ratio also decreases correspondingly. This phenomenon can be understood in this way because it is not conducive to meeting the conditions of not generating redundant messages; Or is understood to be due to the increase of $k$.

From the experimental data, it can be seen that compared with flooding, the improved ratio of the limited flooding algorithm is 7.6%. This efficiency is not very high, but when there are not many redundant messages generated in the first four layers, the redundant messages generated after reduction are acceptable. On the whole seven layers, the selection algorithm and flooding algorithm mentioned in this paper are used to compare the amount of search messages generated when searching files. It can be measured that the improvement ratio of the selection algorithm is 11.5%.

Tensor model may delegate the customary features of non-structural data, semi-structured data and structured data, and know the merge and delegate of multisource different data in high-order and high-dimensional tensor space. We discuss the application of big data fusion model in smart bus. The tensor model of bus big data is composed of video data, weather data and GPS data, in which video data features are composed of many aspects, such as images, key frames, colors and so on. When JSON describes video data, it can extract key frame pictures of video data, and then systematically describe the basic components of video such as roughness, contrast, color and texture. When weather data is
The data has been converted into XML document format. The basic attributes of weather data include city name, date, weather condition, etc. The attribute values are extracted and set, and then the attributes are described by JSON.

Fig.4 Comparison of efficiency between this model and traditional model

The data set selected in the experiment is the bus big data set, which contains structured semi-structured and non-structural data. As displayed in Fig. 4, the data set increases as the serial number increases. Experiments show that the model is as shown in Figure 4, and it is lightweight and can support complex retrieval. Through the analytical of the unified way in view of tensor, the data of getting on and off passengers in the future can be predicted by the passenger flow in a period of time.

5. Conclusion
In this paper, the problem of correspond network non-structural big data analysis algorithm is studied, and a communication network non-structural big data research algorithm in view of semantic relevance feature fusion is proposed, which reconstructs the information flow of big data in phase space, calculates the average mutual information and semantic relevance dimension features of big data in the reconstructed phase space, and extracts the relevance features of big data. Any P2P network can take a node as the vertex and turn it into a pyramid shape. On the first layers of the graph, nodes can limit the generation of redundant messages through the information sent by neighboring nodes. When the message reaches Layer 4, quite a lot of source nodes have already been generated, and there are relatively few redundant messages generated at this time. According to the test of the algorithm in this project, compared with the traditional data processing methods, the proposed algorithm has obvious improvement in data query and analysis ability. Especially in terms of processing efficiency, the proposed algorithm is obviously superior to traditional data processing methods.

References
[1] Du Z. Energy analysis of Internet of things data mining algorithm for smart green communication networks. Computer Communications, 2020, 152:223-231.
[2] Saura J R, BR Herráez,Reyes-Menendez A. Comparing a traditional approach for financial Brand Communication Analysis with a Big Data Analytics technique. IEEE Access, no. 99, pp. 1-1, 2019.
[3] Bu, Fanyu. An intelligent efficient scheduling algorithm for big data in communication systems. International Journal of Communication Systems, vol. 31, no. 16, pp. e3465.1-e3465.9, 2018
[4] Feijoo-Martinez J R,Munoz-Romero S,Soguero-Ruiz C, et al. Event Analysis on Power Communication Networks with Big Data for Maintenance Forms. IEEE Access, 2018, PP:1-1.
[5] Wu P,Lu Z,Zhou Q, et al. Bigdata logs analysis based on seq2seq networks for cognitive Internet of Things. Future Generation Computer Systems, vol. 90, no. JAN., pp. 477-488, 2018.
[6] Zhou N,Zhan X,Lin S, et al. Information diffusion on communication networks based on Big Data analysis. The Electronic Library, vol. 35, no. 4, pp. 00-00, 2017.
[7] Lee J. Optimal power allocating for correlated data fusion in decentralized WSNs using algorithms based on swarm intelligence. Wireless Networks, vol. 23, no. 5, pp. 1-13, 2017.
[8] Yuan J,Holtz C, T Smith. Autism spectrum disorder detection from semi-structured and non-structural medical data. EURASIP Journal on Bioinformatics and Systems Biology, vol. 2017, no. 1,
[9] Rikos N, Linardakis M, Rovithis M, et al. Features of recording practices and communication during nursing handover: a cluster analysis. Revista da Escola de Enfermagem da USP, 2018, 52.

[10] An Qiangqiang, Li Zhaoxing, Zhang Feng, et al. non-structural Big Data Analysis Algorithm for Communication Network Based on Machine Learning. Electronic Design Engineering, vol. 026, no. 014, pp. 53-56, 2018.

[11] Mi Jie, Liu Daohua. Big data mining method based on semantic relevance feature fusion. Journal of Xinyang Normal University no. Natural Science Edition), vol. 32, no. 01, pp. 147-151, 2019.