Visualization of large-scale user-related feature data based on nonlinear dimensionality reduction method

Xiuzhuo Wei, Chunjie Wang, Bo Tang, Huinan Zhao*

Changchun Humanities and Sciences College, Changchuan,130117,China

*zhaohuinan1207@126.com

Abstract. Powerful tools and technical support are needed to extract meaningful information from massive data, understand the relationship between users and discover potential patterns. In this context, the visualization of large-scale user-related feature data has emerged as an important means to deal with the flood of data. The research direction of this paper is the visualization platform of large-scale user-related feature data based on nonlinear dimensionality reduction method, and the overall architecture of the visualization platform of large-scale user-related feature data is designed. Inspired by t-SNE algorithm, a new dimensionality reduction visualization method, namely Laplacian random neighbor distribution based on graph regularization, is proposed. This method aims to project large-scale user-related feature data into the visualization space, so that the visualization results can not only maintain the local neighbor structure of the original high-dimensional spatial data, but also maintain the overall structure of the data, and make the distribution of sample points in 2D space relatively loose. The research results show that the processing time of the test process is gradually reduced, which reflects the acceleration ratio of parallel processing. The improved t-SNE nonlinear dimensionality reduction model well preserves the category relationship between large-scale user-related feature data, and the data after dimensionality reduction is obvious.

Keywords:nonlinear dimensionality; user-related feature data; Visualization; t-SNE

1 Introduction

Visualization of large-scale user-related feature data is one of the most concerned fields in the information age. With the popularity of the Internet and the rapid development of digital technology, our lives are increasingly intertwined with the digital world, resulting in a large number of user data. These data include social media activities, online shopping, mobile application usage, geographical location information, etc., which contain valuable information and can be used to understand user behavior, predict trends, personalized recommendation, social network analysis and many other application fields. However, the scale and complexity of these data also bring unprecedented challenges to researchers and decision makers [1]. Powerful

K. Subramaniyam et al. (eds.), *Proceedings of the 3rd International Academic Conference on Blockchain, Information Technology and Smart Finance (ICBIS 2024)*, Atlantis Highlights in Computer Sciences 21,
https://doi.org/10.2991/978-94-6463-419-8_7

tools and technical support are needed to extract meaningful information from massive data, understand the relationship between users and discover potential patterns. In this context, the visualization of large-scale user-related feature data has emerged as an important means to deal with the flood of data.

The earliest concept of visualization refers to the study of visual representation of data, which mainly aims to express information clearly and effectively by means of graphical means. The modern concept of visualization is more extensive, which aims to study the visual presentation of large-scale information resources and help people understand and analyze data by using related technologies and methods of graphics and images [2]. Nowadays, visualization technology has become a basic tool to reveal the relationship between data in a data set and the hidden information behind it. Literature [3] puts forward a significant value measurement algorithm, which screens words and selects words with more information value to represent subject information. And in the visual presentation, the order of words is adjusted in a small range, and the related words are arranged together to enhance readability. Literature [4] introduced LDA (linear discriminant analysis) thematic model into the visualization and analysis of flow field. Define traces and features corresponding to articles and words respectively, regard traces as feature packages, and each feature represents a certain behavior of traces. Literature [5] applies the topic model to the analysis of mobile phone user data to extract crowd characteristics.

The challenge in this research field is not only to deal with huge data sets, but also to deal with different types of data, such as multi-source data, real-time data and multidimensional data, and to protect user privacy and data security. Through visualization, we can open a new perspective of data, reveal its value and promote the development of scientific research and commercial application. Based on the above research background, the research direction of this paper is a large-scale user-related feature data visualization platform based on nonlinear dimensionality reduction method, and a large-scale user-related feature data visualization platform with good interactivity and practicability is designed and implemented around descriptive statistical analysis, parallel coordinates and scatter matrix, data dimensionality reduction and visual clustering. The platform can provide user management, file management, data display, data dimension reduction, descriptive statistical analysis, data visualization and other functions.

2 Research method

2.1 Design of visualization platform for large-scale user-related feature data

Designing a large-scale visualization platform for user-related feature data needs to consider many aspects, including data processing, visualization technology, user interface and security. Identify data sources, including social media, mobile applications, online transactions, etc., and formulate appropriate data collection strategies. Deal with missing data, abnormal values and duplicate data, and standardize and transform the data. Choose a suitable database or distributed storage system to support the storage and retrieval of large-scale data [6-7].

Consider using 3D visualization or virtual reality technology to enhance the user experience. Allows users to customize visualization parameters and settings to meet the needs of different users. Optimize the platform to adapt to various screen sizes and device types. Ensure the security of data during transmission and storage, and adopt encryption standards. Optimize data processing and visualization algorithms to cope with large-scale data. Allows users to export visualization results for further analysis. Provide social media sharing function to promote information dissemination and cooperation. Record user activities and system performance data for monitoring and troubleshooting [8]. Consider deploying the platform on the cloud to improve availability and flexibility. Ensure the stability and security of the platform, fix vulnerabilities and improve functions.

The visualization platform of large-scale user-related feature data is realized in the form of web site, based on Django service framework [9]. Using MySQL database and file storage, the front end includes data display plug-in, JS visualization module and JS calculation module. The overall architecture of the large-scale user-related feature data visualization platform is shown in Figure 1.

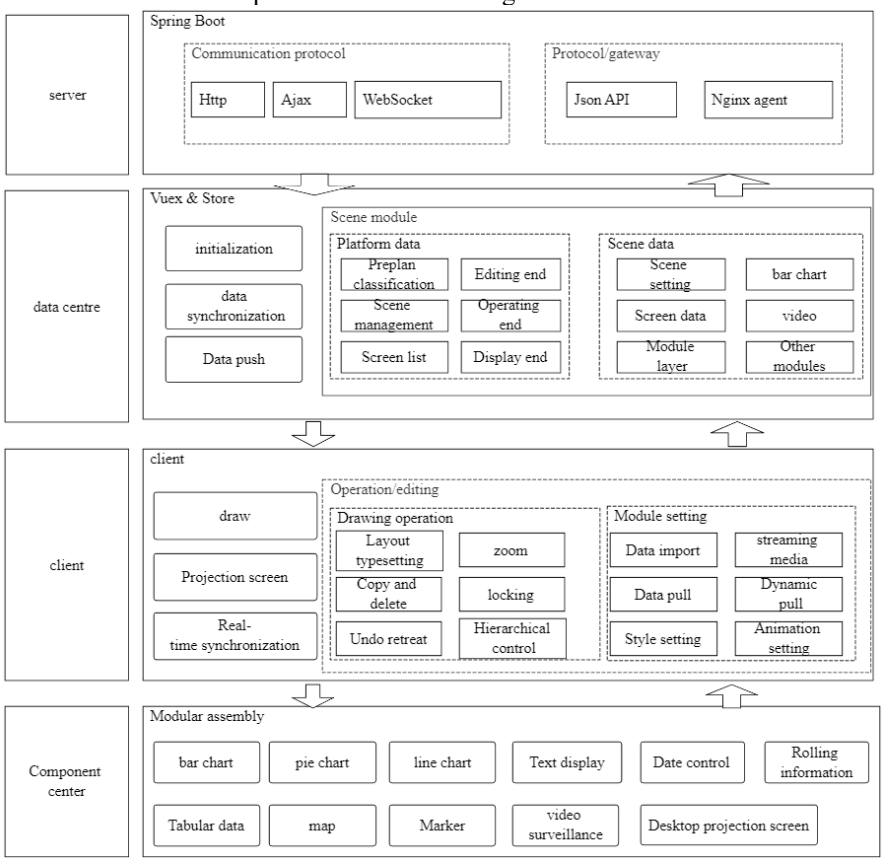

Fig. 1. Platform architecture

The server of data visualization platform is the core component supporting data processing, visualization generation and data distribution. The data processing engine extracts and calculates features from the original data to support different types of visualization. Visual generation uses chart library and visualization tools to transform the processed data into graphical presentation. API and integration allow other applications and systems to integrate with the platform for data acquisition and visualization. Web services support data exchange and visual sharing based on HTTP/REST.

The data center of data visualization platform is an important part of the whole system, which undertakes the key responsibility of storing, processing and managing data. Collect data from various data sources (databases, APIs, log files, sensors, etc.). Clear invalid data, remove duplicate values, and normalize and transform data. Data center is the foundation of data visualization platform, which needs high stability, availability and security to ensure the smooth visualization and analysis of data and protect the privacy and security of user data.

The client of data visualization platform is the interface for users to interact with the platform, browse and use visualization data. The client is usually a variety of applications or Web browsers, and its main task is to provide a user-friendly interface so that users can interact and explore data in an intuitive way. Visualization area is the main area used to present charts, graphs, maps or other visualization elements. Visual template selection allows users to select different types of visual templates, such as histogram, line chart, heat map, etc.

The component center of data visualization platform is a key part for managing, configuring and integrating various core components. The center allows administrators and developers to centrally manage and coordinate the various components of the data visualization platform. Component center is the backbone of data visualization platform. By centralized management of various components, the platform is more flexible, extensible and easy to maintain. This helps to improve the usability of the platform, reduce management costs and promote the rapid deployment of new functions.

2.2 Realization of nonlinear dimension reduction method

Nonlinear visualization technology is a method for presenting and analyzing complex and nonlinear data structures. These technologies can help people find patterns, relationships and structures in large-scale user-related feature data, and are usually used in data mining, machine learning, statistical analysis and visualization [10-11].

t-SNE(t-Distributed Stochastic Neighbor Embedding): t-SNE is a nonlinear dimensionality reduction technique for mapping large-scale user-related feature data to low-dimensional space. It performs well in preserving the similarity between data points, especially suitable for clustering and category separation tasks.

UMAP(Uniform Manifold Approximation and Projection): UMAP is another dimensionality reduction technology, which realizes dimensionality reduction by preserving local topology on data manifold. UMAP performs well in data visualization and clustering with high speed.

NLPCA(Nonlinear Principal Component Analysis): Similar to linear PCA, NLPCA tries to find a low-dimensional representation to preserve the variance of data to the greatest extent. The difference is that NLPCA can capture nonlinear structures.

Self-encoder: Self-encoder is a neural network structure, which can learn to map data to a low-dimensional representation and then restore data from the low-dimensional representation. This is very useful for learning the nonlinear feature representation of data.

MDS(Multidimensional Scaling): MDS technology can reduce the dimension of data by minimizing the distance between large-scale user-related feature data and low-dimensional representation, and can be used for nonlinear data.

KPCA(Kernel PCA): Similar to standard PCA, but by using kernel techniques to map data to high-dimensional space, so as to work on nonlinear data.

TD(Tensor Decomposition): TD technology is used for dimensionality reduction and analysis of multidimensional data, and it is suitable for dealing with nonlinear data with complex structure.

The choice of these nonlinear visualization techniques depends on the nature of the data and the analysis goal. They help researchers and data scientists to better understand data, identify patterns, carry out feature engineering and visualize data, so as to better guide decision-making and discover information hidden in data.

Among the existing dimensionality reduction visualization methods, t-SNE has been widely used to visualize high-dimensional nonlinear data in different fields. Although the boundaries between clusters are clear in the visualization results of t-SNE, from the visualization point of view, the samples within the class overlap too much, which is not conducive to further interactive analysis of the visualization results [12]. Therefore, inspired by t-SNE algorithm, this paper combines the idea of manifold regularization with graph theory, and proposes a new visualization method of dimensionality reduction, namely Laplacian random neighbor distribution based on graph regularization.

This method aims to project large-scale user-related feature data into the visualization space, so that the visualization results can not only maintain the local neighbor structure of the original high-dimensional spatial data, but also maintain the overall structure of the data, and make the distribution of sample points in 2D space relatively loose.

t-SNE algorithm constructs a probability distribution between data points in high-dimensional space, which will make similar data points correspond to higher probability and dissimilar data points correspond to lower probability. The calculation formula is shown in the following formula:

$$p_{j|i} = \frac{\exp\left(-\left\|x_i - x_j\right\|^2 / \left(2\sigma_i^2\right)\right)}{\sum_{k \neq i} \exp\left(-\left\|x_i - x_j\right\|^2 / \left(2\sigma_i^2\right)\right)} \tag{1}$$

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n} \tag{2}$$

Where σ_i is the variance corresponding to the Gaussian distribution with x_i in the middle. Because the density of sample points in the data set is different, different sample points i correspond to different σ_i. The denser the data points are, the smaller the corresponding σ_i is. In the area where data points are sparsely distributed, the corresponding σ_i is larger. σ_i value is calculated by bisecting the rope.

Manifold regularization is a machine learning technology for processing large-scale user-related feature data and nonlinear data. Its idea stems from the manifold hypothesis, that is, data are usually distributed on low-dimensional manifolds, rather than evenly distributed in high-dimensional spaces. The goal of manifold regularization is to better capture the structure and characteristics of data by reducing dimensions or mapping data to a lower dimensional space.

We adopt a symmetrical KL divergence, and at the same time, we introduce the influence of p_{ij} on q_{ij}, which can make the points with great differences between sample points in 2D map farther away. On the basis of symmetric divergence, manifold regularization is introduced, and the Laplacian matrix is constructed as the penalty term of the objective function, so that similar sample points have a larger probability distribution in 2D space, thus better maintaining the geometric structure between samples.

The objective function C is shown in Formula (3).

$$C = \lambda \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} + (1-\lambda) \sum_i \sum_j q_{ij} \log \frac{q_{ij}}{p_{ij}} + \beta \pi^T L \pi \tag{3}$$

Where λ is a symmetric KL divergence parameter with a value of 0.5 and β is a manifold regularization parameter with a value range of [0,0.01]. L is Laplace matrix.

The iterative method of objective function is the same as t-SNE, and the descent speed and time are optimized by momentum and learning rate self-adaptation, so as to prevent the descent process from falling into local minimum instead of global minimum.

$$y^{(t)} = y^{(t-1)} + \eta \frac{\delta C}{\delta y_i} + \alpha(t)\left(y^{(t-1)} - y^{(t-2)}\right) \tag{4}$$

Where η is the learning rate, the initial value of the learning rate in this paper is set to 100, and α represents the impulse. In this paper, α is selected when the number of iterations is $t < 200$, $\alpha(t) = 0.7$, and when $t \geq 200$, $\alpha(t) = 0.4$.

3 Result analysis

In order to verify the usability of the visualization platform of large-scale user-related feature data based on nonlinear dimensionality reduction algorithm, the system is mainly tested by data sets. The whole experiment process is carried out on MATLAB 2015a, and the main configuration of the computer is Intel Core TM i5-4690 CPU 3. 50 GHz and 16 GB DDR 3.

In the model training, the MNIST data set is used, which includes 0 ~ 9 gray-scale pictures of ten handwritten numbers, and the scale of each image is 28 *28 =784. During the model training, 60,000 samples are used for training and 10,000 samples are used for testing.

In this section, 1-10 slave nodes are selected to experiment on the above MNIST data set. As can be seen from Figure 2, with the increase of slave nodes, the processing time of the test process gradually decreases, thus reflecting the acceleration ratio of parallel processing.

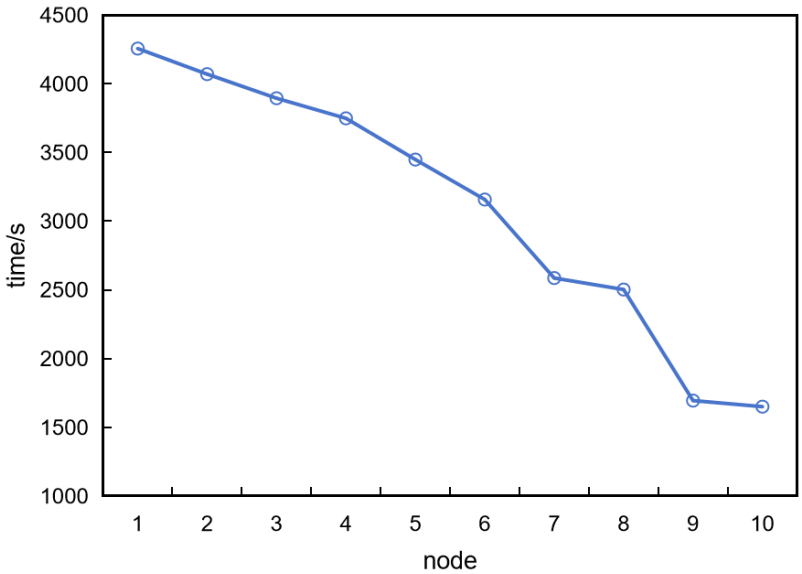

Fig. 2. Performance of the algorithm in different number of nodes

In order to verify the superiority of the improved t-SNE nonlinear dimensionality reduction model selected in this paper, the experiment in this section mainly compares the visualization results of directly reducing the original image from 784 dimensions to 2D by different dimensionality reduction methods. The classical PCA method is used for comparative experiments, and the experimental results are shown in Figure 3.

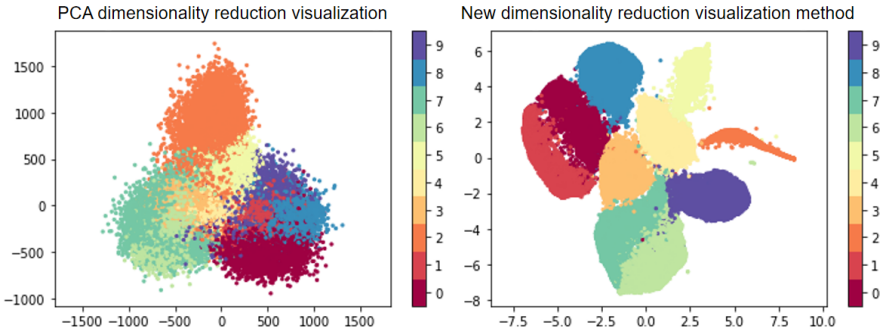

Fig. 3. Comparison of dimensionality reduction visualization results

It can be seen that PCA method has a great loss in direct dimensionality reduction of large-scale user-related feature data, resulting in that the original category of the image after dimensionality reduction of the original data is not separable. However, the improved t-SNE nonlinear dimensionality reduction model well preserves the category relationship between large-scale user-related feature data, and the data after dimensionality reduction is obvious. It shows that the loss caused by dimensionality reduction can be effectively reduced by selecting the improved t-SNE nonlinear dimensionality reduction model.

4 Conclusions

The research direction of this paper is a large-scale user-related feature data visualization platform based on nonlinear dimensionality reduction method, and a large-scale user-related feature data visualization platform with good interactivity and practicability is designed and implemented around descriptive statistical analysis, parallel coordinates and scatter matrix, data dimensionality reduction and visual clustering. The platform can provide user management, file management, data display, data dimension reduction, descriptive statistical analysis, data visualization and other functions. However, the improved t-SNE nonlinear dimensionality reduction model well preserves the category relationship between large-scale user-related feature data, and the data after dimensionality reduction is obvious. It shows that the loss caused by dimensionality reduction can be effectively reduced by selecting the improved t-SNE nonlinear dimensionality reduction model. Users can easily show the correlation between different dimensions of data by exchanging high-dimensional visualization charts.

References

1. Wang, J. , Liu, X. , & Shen, H. W. (2019). High-dimensional data analysis with subspace comparison using matrix visualization. Information Visualization, 18(1), 94-109.

2. Gerber, S. , & Potter, K. (2012). Data analysis with the morse-smale complex: the msr package for r. Journal of Statistical Software, 050(2), 1-22.
3. Vyrynen, E. , Kortelainen, J. , & Seppnen, T. (2013). Classifier-based learning of nonlinear feature manifold for visualization of emotional speech prosody. IEEE Transactions on Affective Computing, 4(1), 47-56.
4. Dorgo, G. , Kulcsar, T. , & Abonyi, J. (2021). Genetic programming-based symbolic regression for goal-oriented dimension reduction. Chemical Engineering Science(244), 244.
5. Li wenfa, Gongming, W. , Ke, L. , & Su, H. (2017). Similarity measurement method of high-dimensional data based on normalized net lattice subspace. chinese high technology letters, 23(2), 6.
6. Bi, X. , Li, B. , & Cang, Y. (2014). Visualization of high dimension multi-objective of rotation basis based on interactive decision making. Jisuanji Fuzhu Sheji Yu Tuxingxue Xuebao/Journal of Computer-Aided Design and Computer Graphics, 26(5), 739-746.
7. Yao, F. , Coquery, J. , & Kim-Anh Lê Cao. (2012). Independent principal component analysis for biologically meaningful dimension reduction of large biological data sets. Bmc Bioinformatics, 13(1), 24.
8. Khan, F. , Kari, D. , Karatepe, I. A. , & Kozat, S. S. (2016). Universal nonlinear regression on high dimensional data using adaptive hierarchical trees. IEEE Transactions on Big Data, 2(2), 175-188.
9. Li, Y. , Wang, G. , Chen, H. , Shi, L. , & Qin, L. (2013). An ant colony optimization based dimension reduction method for high-dimensional datasets. Journal of Bionic Engineering, 10(002), 231-241.
10. Khosla, K. , Jha, I. P. , Kumar, A. , & Kumar, V. (2020). Local-topology-based scaling for distance preserving dimension reduction method to improve classification of biomedical data-sets. Algorithms, 13(8), 192.
11. Ryu, J. , Kim, H. , Kim, R. M. , Kim, S. , Jo, J. , & Lee, S. , et al. (2021). Dimensionality reduction and unsupervised clustering for eels-si. Ultramicroscopy(231-), 231.
12. Zhai, Y. , Wang, N. , Zhang, L. , Hao, L. , & Hao, C. (2020). Automatic crop classification in northeastern china by improved nonlinear dimensionality reduction for satellite image time series. Remote Sensing, 12(17), 2726.