Design of Distributed Multidimensional Big Data Classification System Based on Differential Equation

Pei-ying Wang(✉)

Tianhe College of Guangdong Polytechnical Normal University, Guangzhou 510540, China
wangpeiying258@sina.com

Abstract. In today's more distributed and disorderly network environment, how to organize this information simply and effectively, so that users can quickly obtain potentially valuable data is a common problem in all fields. The commonly used classification systems are based on genetic algorithms and orthogonal decomposition. These two types of systems have high memory usage and low classification accuracy. Aiming at the above problems, a distributed multidimensional big data classification system based on differential equations is designed. The system design is mainly divided into three parts: the first design system overall framework; the second design system hardware, including multidimensional data integration module, central processing module, storage module, result output and display module; third, designing multidimensional big data according to differential equation Classification software main program. The results show that compared with the big data classification system based on genetic algorithm and the big data classification system based on orthogonal decomposition, the classification accuracy of distributed multidimensional big data classification system based on differential equation is improved by 8.75% and 6.75%, and the system memory occupancy rate is improved. Reduce by 35% and 12%.

Keywords: Differential equation · Distributed multidimensional big data · Classification system

1 Introduction

In recent years, with the gradual development and widespread application of computers and the Internet, the amount of data in the Internet has gradually increased, but the rich data resources have made users face greater challenges. The large amount of data scattered and disorder has greatly increased people's The difficulty of using network information. Therefore, it is necessary to design a big data classification system to help users quickly and efficiently obtain the required information in a large amount of network data [1]. Big data is the concept of demand driven. With the popularization of database system and the expansion of Internet services, the data available to enterprises or individuals is expanding, and the existing technology is difficult to meet the data analysis needs in the era of big data. Therefore, we need to explore new theories and

S. Liu and L. Xia (Eds.): ADHIP 2020, LNICST 347, pp. 389–398, 2021.
https://doi.org/10.1007/978-3-030-67871-5_35

methods to support the application of big data. Although 4 V attributes of big data have been widely discussed, most of them describe the representation of big data, so it is difficult to abstract a unified data format. Therefore, it is necessary to find out the technical features that can be used for data formatting.

At present, there are many network big data classification systems, and relevant scholars have achieved good results. For the application requirements of big data with the main technical characteristics of distribution and mobility, reference [2] takes distributed data flow as the data expression carrier, and then designs the corresponding big data classification model and mining operator. At the same time, to solve the key problems of big data classification mining, the algorithm corresponding to the key steps is constructed, which proves the rationality of the micro cluster merging technology and the sample data reconstruction method in theory. Experiments show that the proposed classification model and algorithm of big data based on distributed data flow can not only greatly reduce the communication cost between network nodes, but also improve the global mining accuracy by about 10% on average (compared with the existing typical algorithm DS means). Although the time cost is slightly higher than DS means, the difference between them is very small under different data capacity tests, and the time climbing trend is similar. However, the storage module of the system does not use hierarchical structure, and the system memory occupation rate is high. KNN algorithm is a kind of big data classification algorithm with simple idea and easy implementation, but when the training set is large and there are many characteristic attributes, its efficiency is low and its time cost is large. To solve this problem, reference [2] proposes an improved KNN classification algorithm based on fuzzy C-means, which introduces the fuzzy c-means theory on the basis of the traditional KNN classification algorithm. In order to reduce the number of training sets, the sub cluster is used to replace all the sample sets of the sub cluster. Thus, the workload of KNN classification process is reduced and the classification efficiency is improved. KNN algorithm is better applied to data mining. The theoretical analysis and experimental results show that the algorithm can effectively improve the classification efficiency of the algorithm in the face of large data, and meet the needs of data processing, but the system's big data classification accuracy is low.

Aiming at the shortcomings of the above systems, a distributed multidimensional big data classification system based on differential equations is designed. The system design is mainly divided into three parts: system framework design, system hardware design and system software design. Finally, compared with the big data classification system based on genetic algorithm and the big data classification system based on orthogonal decomposition, the classification accuracy of distributed multidimensional big data classification system based on differential equation is improved, and the system memory occupancy is reduced. It can be seen that the performance of this system is better.

2 Big Data Classification System Based on Differential Equation

Classification systems have always played a very important role in the field of life and engineering. Speech recognition, handwriting recognition, identity recognition, etc. are all areas of classification system discussion. Because of its wide application value, the design and application of classification systems have always been valued [2].

In the past, organizing and organizing a large collection of original documents by manual means is not only time-consuming and laborious, but the effect may not be ideal. By directly filtering and classifying the data through the computer and submitting the parts that the user really needs to the user, the user can be freed from the cumbersome data processing work. Differentiating different types of data more quickly, systemizing a large amount of disordered data, greatly improving the utilization of information. Through the automatic data classification system, it can help users to organize and obtain information well, which is of great significance in improving the speed and accuracy of information retrieval, and has important research value [3].

A differential equation is a mathematical equation used to describe the relationship between a class of functions and their derivatives. It is widely used and can solve many derivative-related problems. Many kinesiology and dynamics problems involving variability, such as the resistance of air to the falling motion of the velocity function, can be solved by differential equations. Differential equations are different from linear equations, quadratic equations, higher-order equations, exponential equations, logarithmic equations, trigonometric equations, and equations. It is not the process of finding the relationship between the known number and the unknown, the equation of the column, and the process of finding the solution of the equation. Rather, the process of finding one or several unknown functions that satisfy certain conditions is the most viable mathematical branch equation. This time, the differential equations are combined with the big data classification system to design a distributed multidimensional big data classification system based on differential equations.

2.1 Overall Framework of Big Data Classification System

Big data classification is not simply to find one or several fixed values. In the past, the staff needed to find the relationship between each data and establish a relationship function to complete the classification process. However, based on the Internet cloud computing, a distributed multidimensional big data classification model is established, and different data can be effectively classified by differential equation algorithm. The classified distributed multidimensional big data includes unstructured environmental data and semi-structured environmental data [4]. The structure of the big data classification system is shown in Fig. 1.

From the perspective of the data classification system, the system includes five modules: multidimensional data integration module, central processing module, storage module, result output and display module.

The distributed multidimensional big data classification system described above shows the basic laws followed by data changes. As long as the corresponding differential

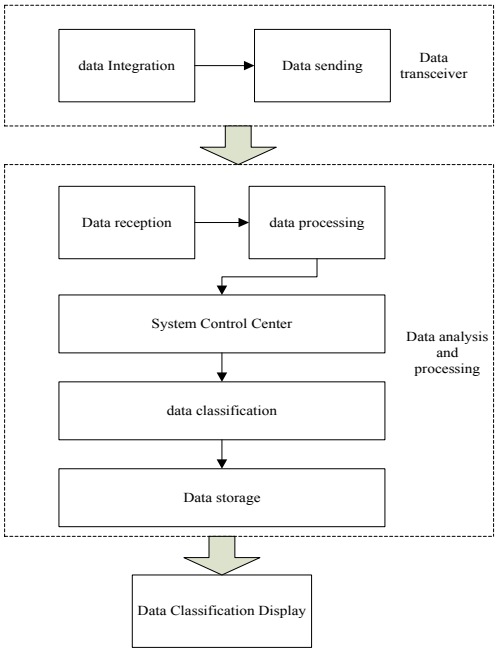

Fig. 1. Structure of big data classification system based on differential equation

equations are listed, the connections and differences between different data can be found and then classified.

2.2 Hardware Design of Distributed Multidimensional Big Data System

(1) Multidimensional data integration module

Distributed multidimensional big data in various forms, such as papers, bibliographies, conference records, journals, etc. These diverse data resources are often heterogeneous (structured, semi-structured, unstructured), so how to automatically migrate these massive, scattered, and heterogeneous data to a central station according to statistical criteria is data classification. The basis [5].

The multi-dimensional data integration mainly completes the operation through the CP2210 integrated chip CP2210, realizes big data acquisition, and then transmits the collected data to the central processor through the network interface. The specific process is as follows: The power supply sends 6 V voltage, which is transmitted to the voltage regulator of the MCU through the REGIN pin of the microcontroller, adjusts the voltage to the 4 V voltage required for the operation of the MCU, and sends the remaining 2 V voltage to the remaining components through the VDD pin. The MCU exchanges information with P3, P4 and other I/O pins. The signal obtained from the network passes through the signal adjuster, and the P25 pin of the single chip reaches

the A/D converter, and the A/D converter converts the signal into corresponding data, thereby completing the collection of the network data.

(2) Central processing module

The role of the master chip is to control the operation of the entire system, all the programs of the system need to be written on the master chip. There are currently four main control chips on the market, such as microcontroller, FPGA, ARM and DSP. Among the above four main control chips, the data processing capability and operation speed of the DSP are optimal, which is in line with the design goal of the system. Here, the TMS320DM642 chip in the C6000 series specially designed for audio processing is selected from TI. The main hardware included in the TMS320DM642 is program memory FLASH, power supply circuit, clock circuit, reset circuit and JTAG port [6].

(3) Storage module

The data storage part includes three parts: FLASH, SDRAM and CF card. FLASH memory has the function of electric erasing and writing in the system, and the information is not lost after power-off. It is used to save the system self-starting code and system program code. This system uses ATMEL's AT29LV020 FLASH chip, which is a NOR type FLASH chip. The total capacity is 256 KB and the data bus is 8 bits. When the EMIFA boot mode is selected by the DSP, the program is automatically loaded from the CE1 space after power-on, so the FLASH must be connected to the CE1 space of the EMIF. The SDRAM memory has a high access speed. It is used to store the system running code and temporary image data. The system uses four Samsung SDRAM K4S561632E, each of which is 16 bits, 32 MB, and CE0 connected to the DSP's EMIF interface. space. The CF card is connected to the CE2 space of the EMIF to store the original image data and the recognition result [7].

(4) Result output and display module

A display is a window in which a person interacts with a robot. In the design of the system, a touch panel, that is, a touch screen, is selected as a display device for sorting results. The system design uses the TPC1063H touch panel produced by Shenzhen Kunlun Tongzhou Technology Co., Ltd., and its composition is shown in Fig. 2.

This touch screen has high performance: it is equipped with Cortex-A8/1G Hz main frequency CPU, which has fast response speed and fast communication speed, which can bring more extreme and smooth operation experience; More serial ports: There is a 232 + 2 485 communication serial port, and the integration of multiple serial ports makes the product easier to use; Complete compatibility: adapt to the market's mainstream touch screen manufacturers opening size [8].

2.3 System Software Design Based on Differential Equation

Differential equation is a kind of mathematical equation which describes the relationship between function and its derivative. Its solution is usually function, while the solution of equation in elementary algebra is usually numerical. The differential equation must first reduce the dimensionality of environmental data that can undergo

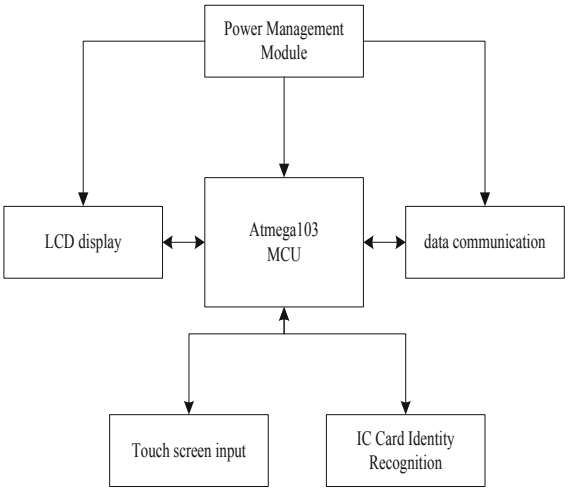

Fig. 2. TPC1063H touch panel

continuous changes, and minimize the possibility of data variation. After the data is reduced from high dimensionality to low dimensionality, each data is orthogonalized between the same dimensions, and the relationship between the data is judged by the cross result, and finally the same type of environmental protection data is grouped together and counted [9, 10].

In the case where the selected environmental data has different dimensions and the difference is relatively large, it is necessary to find the principal components between the different coefficients for analysis. Principal component analysis (PCA) is a statistical method. Through orthogonal transformation, a group of variables that may have correlation are transformed into a group of linear uncorrelated variables, and the transformed group of variables is called principal component. In practical projects, in order to analyze the problems comprehensively, many variables (or factors) related to this are often put forward, because each variable reflects some information of this project in varying degrees. Principal component analysis is first introduced by K. Pearson for non random variables, and then H. hotellin extended this method to the case of random vectors. The size of information is usually measured by the sum of squares or variance of deviations. The principal component analysis method makes the variables appear ellipsoidal distribution, and the distribution area is three-dimensional, and the linear relationship is extremely strong, and the analysis has significant significance. The formula for calculating the differential equation is as follows:

$$Y_i = k_i X = k_{i1} X_1 + k_{i2} X_2 + , \ldots, k_{in} X_n (i = 1, 2, \ldots, n) \tag{1}$$

In formula (1), Y represents the principal component data in i data, k_i represents the corresponding feature vector in the variable covariance matrix.

The data is put into the covariance matrix or the correlation coefficient matrix to find the eigenvector corresponding to the largest eigenvalue [11, 12]. The direction of

the eigenvector corresponds to the covariance matrix variability, and the matrix direction represents the direction of the main environmental data [13]. The direction corresponding to the second largest eigenvalue is the direction of data mutation, which is orthogonal to the first eigenvector, and the degree of orthogonality can reflect the relationship between environmental data. The eigenvectors are used to measure the proportion of data in different directions, and the largest eigenvector is used to form a reference frame for dimensionality reduction. It should be pointed out that the number of feature vectors selected is lower than the dimension of the original data.

The general solution has been the main goal of differential equation in history. Once the expression of the general solution is found, it is easy to get the special solution needed by the problem. The expression of the general solution can also be used to understand the dependence on some parameters, so that the parameter value is appropriate, the corresponding solution has the required performance, and it is also helpful for other research on the solution. Later development shows that there are not many cases in which the general solution can be obtained. In practical application, it is necessary to find the special solution satisfying certain specified conditions. Of course, general solutions are helpful to study the properties of solutions, but people have shifted the focus of research to the problem of definite solutions. Differential equations can find and classify environmental data, and environmental data in the same category should have as high a homogeneity as possible, while categories should have as high a heterogeneity as possible [14]. The grouping is done by finding the distance and similarity between the environmental data. The specific steps are as follows:

(1) Create N observation points and seek K familiar data;
(2) Record the distance between two pairs of different observation points;
(3) The near observation points are unified into one category, and the distant observation points are counted into another category. Finally, the distance between groups is maximized and the distance within the group is minimized.

The data classification first uses the K-means classification method to obtain the ideal classification model, and classifies a large number of multidimensional data samples, and then uses hierarchical classification to find out the number of interaction classifications and analyze to provide accurate similarity information [15]. According to the similarity of the formation of big data, the hierarchical map is drawn to make the class division more intuitive and accurate. Different samples are self-contained. The differential equation is used to calculate the distance between the classes and the distance between samples.

$$M(x, y, z) = N(x, y, z) \cdot Z(x, y, z) \tag{2}$$

In formula (2), M represents the distance between classes, x, y, z represent different vector directions, $\mathrm{G(x,\ y,\ z)}$ represents the vector direction of the first type of data, $Z(x, y, z)$ represents the vector direction of the second data. Inter-sample distance calculation process:

$$T(x,y) = \frac{e}{2g} \tag{3}$$

In formula (3), $T(x,y)$ represents the distance between samples in the horizontal and vertical directions, e represents a regular vector constant, g indicates the number of types of data. After the above calculation is completed, the two types of data with the smallest distance are unified and the two types of data are the largest, until all the big data classification ends.

3 Experiment Analysis

3.1 Lab Environment

The data used in the experiment comes from the network information database. The system needs two computers. The system hardware configuration is: Intel Rean-core 3 GHz processor, 32 GB memory.

3.2 Parameter Settings

The data types used in the experiment are: meteorological data, geological data, economic data, transportation data, etc., and they are numbered as S 1, S 2, S 3, S4, etc.; the data size is 2000, respectively. 1500, 1700, 1800, etc. [12].

3.3 Result Analysis

The classification performance of distributed multidimensional big data classification system based on differential equation, big data classification system based on genetic algorithm and big data classification system based on orthogonal decomposition is compared. The experiment uses three systems to classify experimental big data.

(1) Classification accuracy

It can be seen from Table 1 that the average classification accuracy of big data in this system is 86.25%. Compared with the big data classification system based on genetic algorithm and the big data classification system based on orthogonal decomposition, the accuracy is improved by 8.75% and 6.75%. Because the system designed in this paper uses differential equation to design software content, and then improves the accuracy of data classification.

(2) System resource occupancy

As can be seen from Fig. 3, the CPU usage of the system is 15%, the memory usage is 20%, and the total occupancy is 35%. The total occupancy rate of the big data classification system based on genetic algorithm is 58%. The total occupancy rate of big data classification systems based on orthogonal decomposition is 70%. It can be seen that the resource occupancy rate of the system is significantly lower than the other two

Table 1. Big data classification accuracy results obtained by the three systems

Data type	Differential equation	Genetic algorithm	Orthogonal decomposition
S1 (%)	85	87	95
S2 (%)	88	90	95
S3 (%)	85	88	94
S4 (%)	87	88	96
Average value (%)	86.25	88.25	95

systems, which proves that the system has better performance. The reason for this result is that in the big data classification system designed in this paper, the content of hardware design specially designs data integration module, which reduces the occupancy rate of system resources.

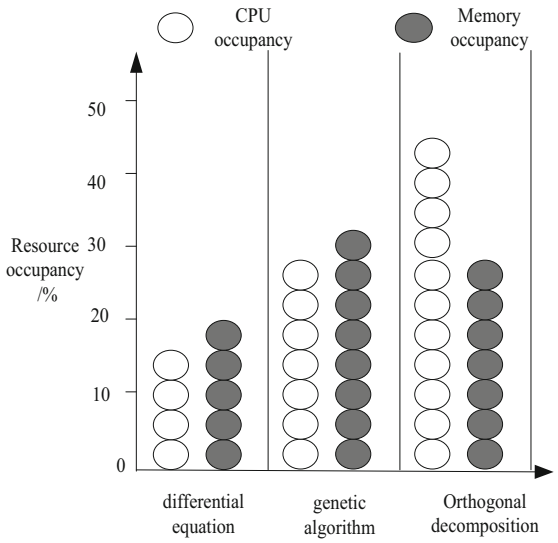

Fig. 3. System resource occupancy rate

4 Conclusion

In summary, for the big data classification system based on genetic algorithm and the big data classification system based on orthogonal decomposition, the classification accuracy is low and the system occupancy rate is high. The distributed multidimensional big data classification system based on differential equation is designed. The biggest feature of this system design is the application of differential equations. It has been verified that the accuracy of system classification is improved and the system resource occupancy rate is reduced. It can be seen that the performance of the system is improved. For the current distributed multi-dimensional data classification accuracy

can not be significantly improved, the possible reason is that the data structure is different, resulting in the data structure can not meet the requirements of the current classification system. In the future, we can use the integrated heterogeneous classifier to meet the data samples of different data structures, so that the classification accuracy can be better improved.

References

1. Yin, S.: Research on the classification technology of large environmental data based on differential equation. Environ. Sci. Manag. **43**(247(6)), 126–129 (2018)
2. Mao, G., Hu, D., Xie, S.: Large data classification model and algorithms based on distributed data flow. J. Comput. Sci. **1**, 161–175 (2017)
3. Huang, S., Lyu, Y., Peng, Y., et al.: Analysis of factors influencing rockfall runout distance and prediction model based on an improved KNN algorithm. IEEE Access **7**, 66739–66752 (2019)
4. Xiaofeng, Z., Yingtao, C.: Improved technology of association mining based on mathematical model of partial differential classification. Mod. Electron. Technol. **40**(8), 36–38 (2017)
5. Min, F., Jun, L.: Design and implementation of big data classification system based on web network. Electron. Des. Eng. **26**(8), 106–109 (2018)
6. Zhe, X.: Design and research of web big data classification system. Comput. Knowl. Technol. **13**(17), 216–217 (2017)
7. Kexing, Z.: Design and implementation of feature data classification system in network big data platform. Mod. Electron. Technol. **40**(8), 25–28 (2017)
8. Luo, X., Cha, Z., Xu, H., et al.: Design of large data automatic classification and processing system based on cloud computing. Comput. Meas. Control **25**(10), 278–280 (2017)
9. Liu, S., Liu, D., Srivastava, G., et al.: Overview and methods of correlation filter algorithms in object tracking. Complex Intell. Syst. (2020). https://doi.org/10.1007/s40747-020-00161-4
10. Liu, S., Bai, W., Liu, G., et al.: Parallel fractal compression method for big video data. Complexity **2018** (2018)
11. Lu, M., Liu, S.: Nucleosome positioning based on generalized relative entropy. Soft Comput. **23**(19), 9175–9188 (2018). https://doi.org/10.1007/s00500-018-3602-2
12. Liu, B., Liu, C.: Automatic classification of large data stored in cloud data management system using content text categorization method. Electron. Technol. Softw. Eng. (20), 179–180 (2017)
13. Tang, Z., Srivastava, G., Liu, S.: Swarm intelligence and ant colony optimization in accounting model choices. J. Intell. Fuzzy Syst. **38**, 2415–2423 (2020)
14. Weihs, C., Ickstadt, K.: Data science: the impact of statistics. Int. J. Data Sci. Anal. **6**(3), 189–194 (2018)
15. Thanigaivasan, V., Narayanan, S.J., Iyengar, S.N., et al.: Analysis of parallel SVM based classification technique on healthcare using big data management in cloud storage. Recent Pat. Comput. Sci. **11**(3), 169–178 (2018)