Research on the Design of Digital Signal Processing Algorithm Based on Spark Parallel Calculation

Yifan Ming1*, Rui Li 2
1School of Physics and Electronics, Central South University, Hunan, China
2School of Mathematics and Computational Science, Hunan University of Science and Technology, Hunan, China
*Corresponding author’s e-mail: Mingyifan2021@csu.edu.cn

Abstract. Digital signal processing as a key and difficult point of network technology research and development, currently commonly used content such as LabVIEW. But from a practical point of view, while these techniques can be used to process real-time signals, they can't handle historical offline data. The Spark parallel computing studied in this paper can be used to process offline signals. Therefore, on the basis of understanding the development trend of Spark parallel computing framework, the distributed Mallat algorithm is analyzed based on Spark parallel computing engine, and the application performance of the corresponding algorithm is verified.

1. Introduction
In the continuous innovation and development of modern information technology industry, how to integrate information technology and traditional manufacturing industry together and achieve excellent results in practical development has become the focus of the whole society. The digital signal processing technology with single chip microcomputer technology as the core is mainly used to solve the signal processing requirements under the conventional mode, but it cannot store large-scale data information, let alone process the historical data. Therefore, whether from an academic perspective or from an industrial economic development perspective, the future of digital signal processing will be studied using small batches of data as computing units. In the continuous development of signal technology, industrial technology equipment also began to modernize and refine the direction of innovation, at this time, the signal in many aspects have great changes, such as high value, many types, large number, etc. represent the era of social and economic development has produced new changes. In the context of big data era, how to use effective countermeasures to process signals and data has become the main topic discussed by current researchers. Hadoop big data analysis platform, as one of the most widely used platforms in technology research and development, is mainly used for parallel research of unstructured data information above PB level. Since digital signal processing algorithms are iterative computing processes, large intermediate results will appear in each iteration. Therefore, if the MapReduce
computing engine is used for research, the intermediate results will be recorded on disk, which will affect the overall system performance. The Spark technology is a new parallel computing engine specially proposed for large-scale data processing. It is universal and fast in practical applications. The data generated by the calculation is stored in the system preferentially to reduce the disk I/O or disk space usage. According to the analysis of the characteristics of the corresponding iterative algorithm, this technology can be used in the industrial big data platform. Although there is still a lack of suitable convolution algorithm at present, it is believed that with the continuous improvement of the technical level of practical research, appropriate solutions will certainly be proposed in the future research and development. In this paper, the application effect of distributed Mallat algorithm is tested and analyzed[1-3].

2. Methods
Spark is a distributed computing framework based on memory that has been widely promoted in recent years. The most common feature of Spark and the traditional distributed Hadoop framework is that they both introduce the basic concept of MapReduce. Therefore, Spark has all the advantages of Hadoop. The specific ecological structure is shown in Figure 1 below:

![Figure 1. Hadoop-based ecological structure](image)

2.1. Wavelet transform
Continuous wavelet transform algorithm and Fourier transform algorithm have consistency, both belong to the most fundamental application content of the same kind of algorithm, mainly used to solve the signal data of continuous periodic transformation. For the commonly used discrete wavelet, the scale parameter A should be extracted from the base 2 on the frequency axis, and the translation parameter should also be obtained according to the base 2 on the time axis. Thus, it can be obtained:

$$a = \frac{1}{2^i}, b = \frac{j}{2}, (i, j \in Z)$$

(1)

The corresponding discrete wavelet change data formula is:

$$W_q f(a, b) = \int_{-\infty}^{\infty} f(t) \left[ 2^i \varphi(2^i t - j) \right] dt = \langle f(t), \psi_{i,j}(t) \rangle$$

$$\psi_{i,j}(t) = 2^i \varphi(2^i t - j), (i, j \in Z)$$

(2)

and \(\psi_{i,j}(t)\) Represents discrete wavelets.

Mallat algorithm integrates the concept of multi-resolution analysis and is a fast implementation method of wavelet transform algorithm. The corresponding decomposition formula is as follows:
\[ c_{j,n} = \sum_{k \in \mathbb{Z}} c_{j-1,k} h(k - 2n)(n \in \mathbb{Z}) \]
\[ d_{j,n} = \sum_{k \in \mathbb{Z}} c_{j-1,k} g(k - 2n)(n \in \mathbb{Z}) \]  
(3)

In the above formula, \( C_{j-1k} \) represents the original data signal of the upper level waiting for decomposition; \( C_{j} \), \( n \cdot DJ \), \( n \) represents the result parameters of the lower level obtained by calculation; \( H \) is for low frequency filter; \( G \) stands for high frequency filter. These parameters are passed progressively in the calculation of the formula[4-6].

### 2.2. Algorithm analysis

Based on the above analysis, it can be seen that the essence of the selected algorithm is to use signal and filter to implement convolution calculation. In a single-machine convolution between vectors, the signal and filter can be placed in an array, creating an empty data set whose length is the size of the calculated result, and then storing the final result. At the same time, we use the pointer loop to compute all the resulting elements, and the number of calculations per loop is only directly related to the length of the filter array. Since parallel computing means that all the contents contained in the elastic distributed data (RDD) will be calculated in each iteration, there are differences between the parallel algorithm based on the wavelet transform algorithm and the single-machine serial algorithm.

First, matrix. In the Spark matrix, there is no clear solution based on vector convolution. Instead, according to the algorithm performance, the filter array is transformed into a matrix with the same length as the signal data to be processed, and then the calculation and analysis are carried out by multiplying. The concrete matrix is shown in Figure 2 below:

![Figure 2. Data transformation diagram in matrix multiplication](image)

Second, loop execution. Because the functions contained in the matrix cannot meet the needs of the wavelet transform algorithm, the cyclic memory method will not use the matrix algorithm, but use the cyclic method for calculation.

Third, transitional data. In the data model, the data are relatively independent, so it is difficult to directly select two data to implement the combined calculation, so extra memory must be used to store the intermediate data during the calculation. At the same time, in the process of calculation and analysis, the waiting signal and the filter data should be integrated, and the Cartesian product operation should be carried out. Then, all the calculation results should be provided with serial numbers, which represent a part of the corresponding position data of the final result. By taking
the data and combining it according to the sequence number, the result of the calculation can be output. Fourth, Mallat algorithm. The marginal continuation is regarded as the pre-processing link of the original data. It is assumed that the length of the original signal \( s(k) \) is \( L \), and \( k = 0,1..., L \) minus 1. Assuming that the length of the filter is \( M \), then each decomposition only needs to carry out boundary continuation on the left and right sides of the original signal \( \left\lfloor \frac{M}{2} \right\rfloor \). And ensure that this value is a positive function.

Under the condition that \( M \) is even, the signal formula after the continuation is:

\[
s\left(\frac{M}{2} - 1\right), s(0), s(1), ..., s(L - \left\lfloor \frac{M}{2} \right\rfloor)
\]

Under the condition that \( M \) is odd, the signal formula after the continuation is:

\[
s\left(\frac{M}{2}\right), s(1), s(0), s(1), ..., s(L - \left\lfloor \frac{M}{2} \right\rfloor - 1)
\]

As the application core of the wavelet transform algorithm, the convolution operation should carry out the following operations in practice: First, the RDD that completes the boundary continuation is used as the input data; Secondly, map operator is used to operate each piece of data and ensure that the data and elements in the filter meet the key value judgment before multiplying, so as to reduce the intermediate data on the basis of achieving the requirement of sample collection. Finally, transfer the operation results of the previous step to reduceByKey and then integrate the results. The corresponding data will be integrated locally before network transmission, and then implement global integration, and finally transfer the corresponding results to the file. Combined with the analysis of transition data, it can be seen that the reduceByKey operator can effectively control the network I/O and improve the application performance of the system in practice. The results generated in one step represent the final results of the wavelet transform algorithm, which can then be further explored using other tools[7-11].

3. Result analysis

In this paper, the wavelet transform algorithm and distributed Mallat algorithm are compared and analyzed. By using the same running data for verification analysis, the accuracy of different algorithms is verified according to the final results, and the variation rule of the running time of the algorithm is studied according to the data of different data volumes. The design experiment in this paper includes a cluster composed of five nodes, which have the same hardware configuration, 4G memory and one-core CPU. All of them are slave nodes, and one of them has implemented master layout.

First, accuracy. In the experimental test and analysis, the size of the data selected is controlled at 120 kB, the sampling frequency is 20kHz, and the sampling number is 20k. The data set is saved in the format of "*.mat", and the data are separated by "\n" representing newline characters. Research and analysis are carried out according to the Wavelet transform algorithm library in Python and
the algorithm proposed in this paper, and the obtained results are stored in the system, and then the relevant content is presented according to the visualization technology, and the results are as follows:

![Figure 3. Comparative analysis of the results of the two algorithms](image)

Combined with the analysis of the figure above, it can be seen that the error of the results obtained by the two algorithms is mainly controlled between 10 and 13 orders of magnitude. Compared with the original data information, the error can be regarded as zero, which proves that the wavelet transform algorithm described in this paper can accurately process digital signals.

Second, algorithm performance. In the process of studying the performance of the algorithm, it is necessary to make the algorithm perform calculation and analysis under different data volumes, and then study the corresponding running time. In the experimental operation, the input signal belongs to the acceleration vibration signal obtained by the signal collector, which needs to be stored in the format of "mat ", and the form of "\n" is used to represent separation. The corresponding parameters are shown in the following table:

| Serial number | File size / | Sampling points per group/W | Group number |
|---------------|-------------|-----------------------------|--------------|
| 1             | 12          | 2                           | 100          |
| 2             | 290         | 2                           | 2156         |
| 3             | 590         | 2                           | 4312         |
| 4             | 691         | 2                           | 5200         |

The experimental algorithm starts by reading the data and ends by storing the final result in an object file. According to the calculation and analysis of the parameters in the above table, the results can be obtained as shown in Figure 4 below:
Combined with the analysis above, it can be seen that the final application time does not increase with the change of data volume whether the computing analysis is carried out on a stand-alone or distributed system. However, when the data volume continues to increase, the computing performance will be limited by the hardware conditions, and the storage capacity and computing level of the cluster will gradually show its own advantages.

Nowadays, researchers have put forward few research projects and related achievements on digital signal processing technology in the research of big data framework, which will inevitably produce certain restrictions on the development of practical technology. This paper proposes three algorithms based on Spark, a distributed memory computing engine, and compares and analyzes the problems that may occur under different computing conditions, such as performance degradation and algorithm performance. The wavelet transform algorithm is further implemented by improving the excessive data. The final verification results show that the algorithm proposed in this paper not only meets the accuracy requirements, but also can make the performance of relevant algorithms meet the expected requirements, which has a positive effect on the development of modern industrial big data, and can further improve the computing level of data signals. Therefore, in the future exploration, researchers will inevitably focus on the wavelet algorithm, focus on Spark as the core of the big data computing framework and the application direction of related technologies, so as to provide better technical ideas and solutions for the increasingly innovative industrial big data environment[12].

4. Conclusion
To sum up, wavelet transform algorithm, as a new branch of science proposed in recent years, is a digital signal processing technology concept with Fourier theory as the core. Compared with traditional Fourier analysis, this kind of algorithm has an essential improvement in practical application. It can not only obtain more valuable features from signals, but also share the same processing framework with other signal processing methods. Although the application and research of this content in digital signal processing are not much in China at present, it is believed that with the continuous improvement of scientific research and technology level, better research results will certainly be obtained from it in the future. Therefore, researchers should put forward more representative research topics according to their own accumulated experience on the basis of learning from the results of relevant research projects at home and abroad.
Reference

[1] Qiao, X., Liu, F., Yu, B. H. (2018) Design of Distributed Digital Signal Processing Algorithm Library based on Spark. Application of Computer Systems, (8):214-218.

[2] Liu, Z. S., Pan, Z. S. (2016) Research on Parallel SVM Algorithm based on Spark. Computer Science, 43(005):238-242.

[3] Liu, P., Wang, X. K., Huang, Y. H., et al. (2017) Computer Science, (12):33-37. (in Chinese)

[4] Li, S., Wu, B., Du, X. M., et al. (2017) Design and Implementation of BIRCH Algorithm Parallelization Based on Spark. Computer Engineering and Science, (39):35-41.

[5] Zhu, Z. L., Li, L. J. (2018) Research on parallelization of Density Clustering Algorithm based on Spark. Computer Technology and Development, 2018, 028(006):80-84.

[6] Zheng, Y. F., Du, T., Zhu, L. J., et al. (2016) A new method for data mining based on Spark. Data Mining, 006(004):P.158-167.

[7] Lu, J. Y., Li, L. J. (2019) Journal of Computer Technology and Development, 029(001):85-89.

[8] Zhuang, R., Li, L. J. (2018) Parallelization of CVFDT classification Algorithm based on Spark. Computer Technology and Development, 2018, 028(006):35-38.

[9] Keary, M. (2013) Collaborative Research in the Digital Humanities: A Volume in Honour of Harold Short, on the Occasion of His 65th Birthday and His Retirement, September 201020132Edited by Marilyn Deegan and Willard McCarty. Collaborative Research in the Digital Humanities. online information review, 2013, 37(3):480-481.

[10] Osterman, A., Benedicic, L., Ritosa, P. (2014) An IO-efficient parallel implementation of an R2 viewshed algorithm for large terrain maps on a CUDA GPU. International Journal of Geographical Information Science, 28(11-12):2304-2327.

[11] Jiang, Y., Zhuo, J., Zhang, J., et al. (2018) The Optimization of Parallel Convolutional RBM Based on Spark. International Journal of Wavelets, Multiresolution and Information Processing, S0219691319400113-.

[12] Al-Kharusi, H., Al-Bahadly, I. (2014) Intelligent Parking Management System Based on Image Processing. World Journal of Engineering & Technology, 02(2):55-67.