MapReduce based preprocessing on vibration data of wind turbine

Yang Jiming
School of Control and Computer Engineering, North China Electric Power University
Baoding, China
443089084@qq.com

Zhen Zhiguang
Shijiazhuang Liangcun thermoelectric Co. Ltd
Shijiazhuang, China
zhencg@163.com

Abstract—Analysis of the running state of the equipment is obtained from the vibration data acquisition of wind power plant ,which is very important to discover the critical wind turbine fault, and the driving end is one of the vibration data. The vibration data comes through multiple sensors by frequency 1.2KHz .Testing a 1~2 day will be more than GB of data. However, pretreatment with the traditional filtering method were used to be the single way of dealing, the efficiency is low, can not meet the actual demand. Cloud computing is a key technology to solve the above problem, which can be used for MapReduce model in large-scale data parallel operation. Based on the existing vibration data preprocessing method and MapReduce mechanism, realizes the parallel processing algorithm,consistency and also designed experiments to verify the validity of the method and the parallelresults on the platform of Hadoop.

Keywords-pretreatment; cloud computing; parallel filtering; MapReduce;

I. INTRODUCTION

Wind turbine construction in China is in a stage of rapid development, with the continuous enlargement in capacity of a wind-power generator, the stress status of wind turbines equipment such as gearbox becomes more complicated[2].In addition, wind turbine in which the work environment is usually particularly bad, when the wind turbine operating conditions change with wind speed, rotor speed and the components under load are also change at any time, therefore easily causes the damage to the components of transmission which affect the safety and reliability of the unit operation. First analysis of data from the sensor to get operational status of equipment, then conduct fault monitoring to ensure the normal operation of wind turbines[3]. However, the collected data generally contains noise, there are missing and inconsistent, meanwhile the low quality data analysis process will cause a low-quality analysis and processing results.Therefore, data preprocessing is crucial. The study of traditional pretreatment method combined with MapReduce programming model can solve the problem of efficient pretreatment for massive vibration data. At present in terms of preprocessing, the MapReduce programming model will be applied on vibration data preprocessing methods for parallel programming have not been reported.

II. PRETREAMENT METHODS

A. data smoothing

Vibration signal which get from data acquisition system was offen aliasing noise signal. In addition to periodic noise signal of the interference signal, there are also irregular random disturbance signal. Due to the wide-band random of noise signal, and sometimes share a large proportion of high-frequency components, so that the vibration signal waveforms drawn by the collected discrete data has many glitches, is not very smooth[12]. Furthermore, in the process of vibration test, the test instrument due to some unexpected interference, resulting in the sample signal of individual measuring point exits greater deviation from the baseline, whose shape is highly irregular[4]. Data smoothing mainly to weaken or eliminate the interference signal, improve the noise ratio and the smoothness of the curve. In this paper, five-spot triple smoothing algorithm[5] is taken to process the data.

Five-spot triple smoothing algorithm is a method for data processing based on equally spaced numerical, Taking y as a function of x, y will be expanded by Taylor's formula for the power series form, according to the precision requirements generally take the top four, namely:

$$y_i(x_i) = a_0 + a_1 x_i + a_2 {x_2}^2 + a_3 {x_3}^3 \tag{1}$$

The sum of variance as follows:

$$F(a_0, a_1, a_2, a_3) = \sum_{i=1}^{5} \left[(a_0 + a_1 x_i + a_2 {x_2}^2 + a_3 {x_3}^3) - y_i \right]^2 \tag{2}$$

Requirements (2) take minimum value, according to the principle of least square method, namely:

$$\frac{\partial F}{\partial a_k} = 0 \quad \text{k=0,1,2,3.} \tag{3}$$

Expand the above equation yields:

$$\begin{cases} 5a_0 + a_1\sum x_j + a_2\sum x_j^{\,2} + a_3\sum x_j^{\,3} = \sum y_j \\ a_0\sum x_j + a_1\sum x_j^{\,2} + a_2\sum x_j^{\,3} + a_3\sum x_j^{\,4} = \sum x_j y_j \\ a_0\sum x_j^{\,2} + a_1\sum x_j^{\,3} + a_2\sum x_j^{\,4} + a_3\sum x_j^{\,5} = \sum x_j^{\,2} y_j \\ a_0\sum x_j^{\,3} + a_1\sum x_j^{\,4} + a_2\sum x_j^{\,5} + a_3\sum x_j^{\,6} = \sum x_j^{\,3} y_j \end{cases} \quad (4)$$

In Formula (4), $j = i-2, i-1, i, i+1, i+2$, Because it is equidistant values, it may take $x_i = 0$, then the corresponding five-point values of x are: $-2\Delta x, \Delta x, 0, \Delta x, 2\Delta x$. Accordingly, the following equation can be obtained:

$$\begin{cases} \sum x_j = \sum x_j^{\,3} = \sum x_j^{\,5} = 0 \\ \sum x_j^{\,2} = 10\Delta x^2 \\ \sum x_j^{\,4} = 34\Delta x^4 \\ \sum x_j^{\,6} = 130\Delta x^6 \\ \sum y_j = y_{i-2} + y_{i-1} + y_i + y_{i+1} + y_{i+2} \\ \sum x_j y_j = (-2y_{i-2} - y_{i-1} + y_{i+1} + 2y_{i+2})\Delta x \\ \sum x_j^{\,2} y_j = (4y_{i-2} + y_{i-1} + y_{i+1} + 4y_{i+2})\Delta x^2 \\ \sum x_j^{\,3} y_j = (-8y_{i-2} - y_{i-1} + y_{i+1} + 8y_{i+2})\Delta x^3 \end{cases} \quad (5)$$

Simultaneous equations (4) and (5) to work out the values of a_0, a_1, a_2, a_3 are:

$$\begin{cases} a_0 = \frac{1}{35}(-3y_{i-2} + 12y_{i-1} + 17y_i + 12y_{i+1} - 34y_{i+2}) \\ a_1 = \frac{1}{10\Delta x}(\frac{5}{6}y_{i-2} - \frac{20}{3}y_{i-1} + \frac{20}{3}y_{i+1} - \frac{5}{6}y_{i+2}) \\ a_2 = \frac{1}{14\Delta x}(2y_{i-2} - y_{i-1} - 2y_i - y_{i+1} + 2y_{i+2}) \\ a_3 = \frac{1}{12\Delta x}(-y_{i-2} + y_{i-1} - 2y_{i+1} + y_{i+2}) \end{cases} \quad (6)$$

The values -2 Δx , Δx ,0, Δx ,2 Δx of x are taken into (1) yields:

$$\begin{cases} y_{i-2} = a_0 - 2a_1\Delta x + 4a_2\Delta x^2 - 8a_3\Delta x^3 \\ y_{i-1} = a_0 - a_1\Delta x + a_2\Delta x^2 - a_3\Delta x^3 \\ y_i = a_0 \\ y_{i+1} = a_0 + a_1\Delta x + a_2\Delta x^2 + a_3\Delta x^3 \\ y_{i+2} = a_0 + 2a_1\Delta x + 4a_2\Delta x^2 + 8a_3\Delta x^3 \end{cases} \quad (7)$$

Then take (6) into (7), we obtain five-spot triple smoothing formula is:

$$\begin{cases} \bar{y}_{i-2} = \frac{1}{70}(69y_{i-2} + 4y_{i-1} - 6y_i + 4y_{i+1} - y_{i+2}) \\ \bar{y}_{i-1} = \frac{1}{35}(2y_{i-2} + 27y_{i-1} + 12y_i - 8y_{i+1} + 2y_{i+2}) \\ \bar{y}_i = \frac{1}{35}(-3y_{i-2} + 12y_{i-1} + 17y_i + 12y_{i+1} - 3y_{i+2}) \\ \bar{y}_{i+1} = \frac{1}{35}(2y_{i-2} - 8y_{i-1} + 12y_i + 2y_{i+1} + 2y_{i+2}) \\ \bar{y}_{i+2} = \frac{1}{35}(2y_{i-2} - 8y_{i-1} + 12y_i + 27y_{i+1} + 2y_2) \end{cases} \quad (8)$$

By the formula (8) also can be seen, the processed data $\bar{y}_i$ are only related to y_i and its five values before and after, but have nothing to do with x_i and interval Δx , therefore, so long as they are equally spaced can hardly be processed in this way.

III. THE PARALLED DATA PRETREATMENT ALGORITHM BASED ON MAPREDUCE

On the basis of the pretreatment method, parallel data preprocessing algorithm is proposed based on MapReduce .

A. The Hadoop distributed parallel computing framework

The parallel computing framework MapReduce of Hadoop, used for massive data (typically greater than 1TB) to parallel processing, which was originally proposed by Google Inc., running on Google's GFS, to provide background web search engine indexing process for the hundreds of millions of users worldwide, while it also provides services to thousands of Google internal applications and data processing[7-9].

MapReduce is a parallel programming model for processing massive data, which is to format the data into a number of key / value pairs, namely the parallel computation was carried out on the key/value. Through the programmer programming interface provided by MapReduce framework, we can easily write distributed applications running on several computers, which brings great convenience to the developer, simplifies the procedure of dealing with massive amounts of data. MapReduce uses small steps and simple things, it breaks a more complex problem into a number of smaller parallel computing problems, then merge them after parallel processing on these small problem. In a nutshell, through the MapReduce program our vast amounts of data can be converted into several small files operates separately, divide these tasks into several mapsand reduce tasks, which reduce task is to summarize these map tasks. However, the problems encountered during program execution are all handled by the MapReduce framework, without the intervention of programmer, greatly reduces the difficulty of writing distributed programs, and provides a better solution to handle massive data[10].

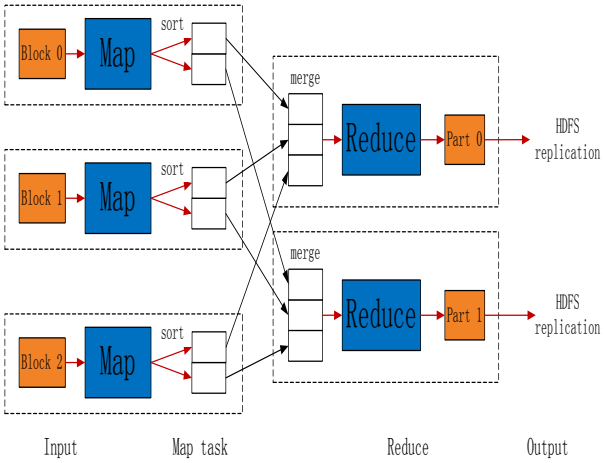

Figure 1. data flow of MapReduce

B. *Parallel Data smoothing algorithm*

Parallel data smoothing algorithm is shown in Figure 2

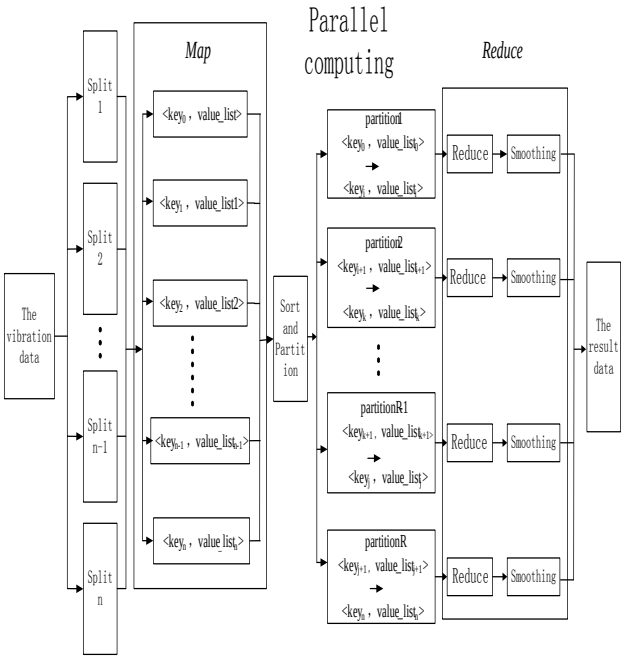

Figure 2. Parallel Data smoothing algorithm

In the process of parallel data smoothing, each corresponding output data set $value\_list\{x_1, x_2, ..., x_n\}$ of key_0 in the MapReduce phase, while been read in the Reduce stage, the sequence of data values in the data set $value\_list\{x_1, x_2, ..., x_n\}$ will change. For example, the key Map output whose corresponding data set is $value\_list\{x_1, x_2, x_3, x_4, x_5\}$,while read in the Reduce stage its corresponding data set will change into $value\_list\{x_1, x_3, x_4, x_5, x_2\}$.Different data in the data set of key has different change rule, So if you use a smoothing weighted formula in Reduce stage may not meet the conditions of the original formula because of the change of the order of the data values. so we can not use smooth weighted formula in the Reduce phase.

Here a practical example to illustrate the problem.Data set in key_3 is $\{x_1, x_2, x_3, x_4, x_5\}$, it will change in Reduce stage, which can not be processed by using a fixed weighting formula $\{-3x_1, 12x_2, 17x_3, 12x_4, 3x_5\}$ in Reduce stage. The solution is weighted different value before outputting in the Map stage, then in Reduce stage only need to work on average processing can complete data smoothing.

1) data smoothing in Map Steps

Read the data file, weight each x_i, deposit in the neighboring five corresponding data values, form of key-value pairs

$\{<key_{i-2}, -3x_i>, <key_{i-1}, 12x_i>,$
$<key_i, 17x_i>, <key_{i+1}, 12x_i>, <key_{i+2}, -3x_i>\}$

Algorithm for pseudo-code is shown in Figure 3

```
**************************************************
Pseudo-code1: Map Process
**************************************************
Input:
<key, V>  //key:the vibration data file name
          //V: vibration data set
Output: q
<k, v>    //k:data identification
          //v:The vibration data values of the k
begin
          set k=0;
          for each v belong to V   do
                    k++;
                    output.collect<k-2, 3v>;
                    output.collect<k-1, 12v>;
                    output.collect<k, 17v>;
                    output.collect<k+1, 12v>;
                    output.collect<k+2, 3v>;
          end
end
```

Figure 3. Smoothing process of map

2) data smoothing in Reduce Steps

Calculate sum_i using the data set corresponding to each key_i, $sum_i = v_{i-2} + v_{i-1} + v_i + v_{i+1} + v_{i+2}$,get the smoothed document $x_i = sum_i / 35$. Algorithm for pseudo-code is shown in Figure 4.

```
**********************************************
Pseudo-code2: Reduce Process
**********************************************
Input:
<k,V>   //key:the vibration data file name
        //V: vibration data set
Output:
<k,v>   //k is null,
        //v:the smoothing data
begin
        set value=0;
        for each v belong to V
                value += value;
        end
        value = value/35;
        set k = null;
        output.collect<k,v>;
        end
end
```

Figure 4. Smoothing process of reduce

C. *Experimental analysis of parallel preprocessing algorithm*

Experiment uses parallel Speedup algorithm performance evaluation index to evaluate data preprocessing parallel algorithm[11]. This experiment uses 6 host form a Hadoop cluster. Where a host is NameNode, another 5 hosts is DataNode. Hardware configuration: network bandwidth is 100Mbps Ethernet, 4-core CPU, clocked at 3.1GHz, memory 4G. Software environment: the operating system is CentOS 6.3, jdk version 1.7.0_11, Hadoop version is 1.0.4.

The data used in the experiment is a drive vibration data set of a wind power plant, the details are shown in Table 1.

Table 1. vibration data sets of wind power plants

vibration data set	Records munber	Files number	File Size （MB）
OR007@01	8225630	1	102
OR007@02	16451260	1	204
OR007@03	24676890	1	306
OR007@04	32902520	1	408
OR007@05	41128150	1	510

In Speedup experiments, we we experiment by adding nodes (DataNode) number and keep the size of the data to test the characteristics, calculated as follows:

$$Speedup(p) = \frac{T_1}{T_p}$$

Among them, p is the number of nodes (DataNode), T1 is the time for the system to perform a node participation in the experiment, TP is the time involved in the case where the execution node p experimental system. With the increasing amount of data, the time overhead grows linearly when read rhe data in Map stage, and rate of preparation time is less than the growth rate of the system reads the data, so with the increasing amount of data, the efficiency of parallel processing improved accordingly.

In this experiment, data smoothing preprocessing method is used. By changing the size and number of nodes in the data set to evaluate the vibration Speedup, the number of data nodes increase from one to five. Experimental noise data sets were OR007 @ 01 to OR007 @ 05. Figure 5 shows the Speedup indicators using data smoothing approach to vibration data sets. Experimental results show significant performance indicators, the number of nodes participating in the experiment, the better Speedup enhance.

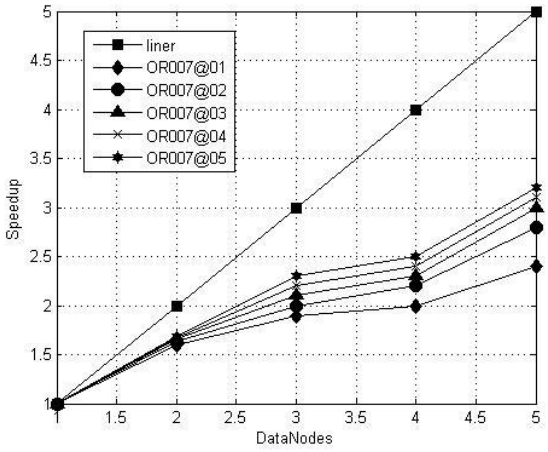

Figure 5. Speedup index under the Parallel smoothing algorithm

IV. CONCLUSION

This paper briefly describes Hadoop, MapReduce ect. basic knowledge of cloud computing,and introduces the traditional smoothing process method of vibration data preprocessing.Furthermore, through the traditional vibration data preprocessing method combined with a parallel programming model, wind turbine vibration data parallel preprocessing algorithm based on MapReduce was proposed, then develop the corresponding procedures. Application is deployed to the Hadoop platform for experimental verification and vibration data sets in different sizes is conducted for parallel index test. The results show that when the larger data sets, the higher efficiency the smoothing parallel algorithms of vibration data, which shows that the proposed method can effectively handle concurrent massive wind turbine vibration data.

REFERENCES

[1] Wang Peng. Key technologies of cloud computing and its application [M]. Beijing: People's Posts and Telecommunications Press, 2010

[2] Peng Huadong, Chen Xiaoqing, Ren Ming et al. Wind turbine intelligent fault diagnosis technology and system for [J]. power grid and clean energy, 2011,2 (27): 61-70.

[3] Xu Yan. Finite element analysis of key components of wind turbine [D]. Urumqi: Xinjiang University, 2005:4-8.

[4] Lv Yuegang, Guan Xiaohui, Liu Juncheng. Wind turbine condition monitoring system (J).Automation and instrumentation, 2012, 27 (1):6-10

[5] Sun Miaozhong. Vibration of electronic measurement technology [J]. smoothing methodbased on MATLAB signal processing, 2007,30 (6): 55-57.
[6] Du Haofan, Cong. Research on [J]. computer denoising for MATLAB wavelet simulation based on.2003, 7 (20): 119-122.
[7]http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-55.html
[8] Zaharia M, Borthakur D, Sarma J S, et al. Job scheduling for multi-user MapReduce clusters[J]. EECS Department, University of California, Berkeley, Tech. Rep. USB/EECS-2009-55, 2009
[9] Yan R, Fleury M O, Merler M, et al. Large-scale multimedia semantic concept modeling usingrobust subspace bagging and MapReduce[C]//Proceedings of the First ACM Workshop onLarge-scale multimedia retrieval and mining. ACM, 2009: 35-42.
[10] Wierman A, Nuyens M. Scheduling despite inexact job-size information[C].//ACMSIGMETRICS Performance Evaluation Review. ACM, 2008, 36 (1): 25-36.
[11] XU X W, Kriegel H P, JAGER J.A fast paralled clustering algorithm for large spatialdatabase[J].Data Mining and Knowlege Discovery, 1999,3 (3): 263-290
[12] Du Haofan, Cong. Research on [J]. computer denoising for MATLAB wavelet simulation based on.2003,7 (20): 119-122.