Estimating Median in The Multi-sourced Heterogeneous Data Set: A distributed implementation

Tian Gao\textsuperscript{a}, Yanyu Zhao\textsuperscript{b}, Xinyi Deng\textsuperscript{c}, Wenze Li\textsuperscript{d}, Hongjun Li\textsuperscript{e}, Xin Zhao\textsuperscript{f, *}

Beijing Normal University, Zhuhai, No 18 Jinfeng Road, Tangjiawan, Zhuhai, China
\textsuperscript{a}2796840832@qq.com, \textsuperscript{b}17688179779, \textsuperscript{c}2541529978@qq.com, \textsuperscript{d}13631276425, \textsuperscript{e}1357062882@qq.com, \textsuperscript{f}13631232848, \textsuperscript{g}651488039@qq.com, \textsuperscript{h}13631234771, \textsuperscript{i}867255105@qq.com, \textsuperscript{j}13724424422

\textsuperscript{f, *}Corresponding Author: zhaoxin@bnuz.edu.cn, 15992695676

ABSTRACT. The continuous running of enterprise applications produces a huge volume of business data that reside in different storage and system environments and owned by different companies or organizations, which forms a typical distributed, multi-sourced heterogeneous dataset. The multi-sourced heterogeneous data set provides big potential values for official statistics. While the median is a commonly used indicator in official statistics, it is not a trivial task to estimate the median in the distributed computing environment of multi-sourced heterogeneous data set due to its mathematical nature. In this paper, we proposed a distributed method to estimate the median value for the multi-sourced heterogeneous data set. Mostly considering the different size of multi-sourced data set and unevenly distribution of their data values, we first improve the traditional interpolation based median estimation method for the multi-sourced heterogeneous data set. Then, we propose distributed implementation for the proposed median estimation method based on web service technology. Finally, we evaluate the accuracy and performance of proposed method through experimental study.

CSC Concepts
Information systems $\rightarrow$ Data management systems $\rightarrow$ Information integration $\rightarrow$ Mediators and data integration

1. INTRODUCTION
The median is one of the most commonly used statistical indicators. Because of the special mathematical nature, in the scenario of distributed multi-source heterogeneous data sets, the calculation of the median is very difficult. At present, the related research on median calculation mainly focuses on the realization of median estimation of grouped data, such as interpolation method and linear equation interpolation method, and these methods focus on the process of mathematical calculation. It is required that the value range of each grouped data is in a regularly increasing state and uniformly distributed. Currently, few studies have proposed a median estimation method for multi-source heterogeneous data sets in real distributed computing environments.

In the multi-source heterogeneous data set, the data that can be used for median calculations are distributed across different system platforms, residing in different storage and computing environments, as compared to the traditional scenario of median calculation of packet data. More importantly, the data is not evenly grouped as in the traditional method, and the data size of different data sources is also different, which causes the value range of the overall data to be irregular and uneven. Therefore, it is not possible to simply apply the calculation formula for the median in the
existing grouping. In addition, because in the distributed computing scenario, the median of the multi-source heterogeneous data set may be distributed on different data sources, that the same median may be multiple values. Therefore, the group in which the median is located is uncertain. In this case, the result obtained by the traditional calculation method can only be an estimate.\cite{I}

Regarding the issue above, this paper firstly improves and optimizes on the basis of traditional interpolation method, and then proposes a median estimation method that is more suitable for adapting distributed heterogeneous data scenarios. On this basis, this paper based on webservice technology gives a distributed implementation framework of the method. Finally, we conducted experiments to verify the accuracy and performance of the median estimation of the proposed method. The following content of this paper is organized as follows: Section 1 introduces the problems and ideas studied in this paper. In section 2 we discuss related work and its differences with the method of this article.\cite{1}

2. RELATED WORK

Many researchers have proposed different methods for grouping data to solve the problem of median estimation. Lunzhi Gan tried to break through the assumption that median array data is evenly distributed of traditional formula for calculating the median of group groupings. He introduces the frequency information of the adjacent group of the median array in the calculation formula to improve the calculation accuracy of the formula\cite{2}. Ding Yuewei et al. studied and proved the in-depth solution method of the traditional linear equation interpolation method under the assumption of uniform distribution\cite{3}. Jie Cao et al. studied and discussed the current method of median calculation method in several special cases, including intermittent group spacing grouping, where the median group cannot be determined\cite{4}.

In a distributed multi-source heterogeneous data set, data for median calculations is distributed across different data sources, similar to grouping data. But the distribution of data among the various data sources or data packets is more uneven. This paper proposes an improved algorithm based on the traditional interpolation method to solve the problem of low median calculation accuracy due to uneven data distribution under distributed heterogeneous data set environment. On this basis, based on Web-Service technology, we give a distributed implementation of the algorithm.

3. THE ESTIMATION METHOD OF MEDIAN FOR DISTRIBUTED MULTI-SOURCED HETEROGENEOUS DATA SET

3.1 THE ESTIMATION METHOD OF MEDIAN

Due to the nature of multi-source heterogeneous data sets cannot satisfy the requirement of the traditional interpolation method formula of calculation, we need to preprocess data from different data sources in specific ways, the data after processing meets the demand of interpolation method of calculation, using interpolation method so that we can achieve the purpose of multi-source heterogeneous data set digits statistical indicators. Firstly, we obtain the value range of the overall data of each data source, find the data with the smallest value and the data with the largest value, and then get the overall value range of all the data of the whole multi-source heterogeneous data set. Showing as table 1 we can easily conclude that the value range of the whole data set is 101-501 after knowing the value range of each data source.

| Grouped by data sources | Value range |
|-------------------------|-------------|
| data source1            | 112—435     |
| data source2            | 121—500     |
| data source3            | 101—417     |
Then the scope of the whole is divided into the number of copies data corresponding, grouping and form according to the different scope of "barrel", as shown in table 2, as you can see, this is also the key to solve problem, the uncertainty of data source value range of the multi-source heterogeneous data sets.

| Grouped by barrels | Value range |
|--------------------|-------------|
| Barrel 1           | 101—158     |
| Barrel 2           | 159—216     |
| Barrel 3           | 217—274     |
| Barrel 4           | 275—332     |
| Barrel 5           | 333—390     |
| Barrel 6           | 391—448     |
| Barrel 7           | 449—501     |

We use these barrels which have divided to specific data for each data source, and load the special data which is in each data source of the corresponding range into the corresponding barrels. Finally, it is concluded that different value scope of all of the data in the data source of specific frequency distribution by traversing the entire data according to the rule of all the data.

| Grouped by barrels | Value range | Frequency |
|--------------------|-------------|-----------|
| Barrel 1           | 101—158     | 524       |
| Barrel 2           | 159—216     | 464       |
| Barrel 3           | 217—274     | 623       |
| Barrel 4           | 275—332     | 742       |
| Barrel 5           | 333—390     | 543       |
| Barrel 6           | 391—448     | 379       |
| Barrel 7           | 449—501     | 673       |

At this point, the grouping form of the multi-source heterogeneous data set is processed into the grouping form of different groups with different value ranges, under which the calculation formula of interpolation method can be used to calculate the median statistical index.

Another problem of using the interpolation formula to solve the median index of multi-source heterogeneous data set is the improvement in accuracy. At present, the point that most affects the accuracy of calculation results is that the multi-source heterogeneous data group does not have absolute uniform distribution, and it can obtain the distribution group frequency of the median array data through the comparison between the median array and its neighbors, so as to obtain rough information. Therefore, we can improve the problem of uneven data distribution by adding frequency.
information of adjacent groups to the formula. The improved method refers to the improved calculation method of the median array distribution due to the improved calculation method of the median array of group spacing groups. The idea of "uneven" distribution of the data of the median array is as follows: coarse information can be obtained by comparing the frequency of the median array and the frequency of adjacent groups, so further increasing the frequency information bits of adjacent groups can improve the calculation accuracy to a certain extent. In the median array, the frequency of data falling into the left interval \([L, Me]\) is

\[
\frac{1}{2} \frac{S_{m-1}}{\sum f}
\]  

The frequency of data falling into the right interval \([Me, U]\) is

\[
\frac{S_{m-1} + f_m}{\sum f} - \frac{1}{2}
\]

We assume that the ratio of the frequencies of the medians left and right in the median array equals the ratio of the areas corresponding to these intervals in the histogram (see figure 1). Here, the "area area" method is different, will produce different results. Take the area ratio as \(\frac{\text{the area of rectangle ABMeL}}{\text{the area of rectangle BCUMe}}\), then

\[
\frac{1}{2} \frac{S_{m-1}}{\sum f} - \frac{1}{2} = \frac{x \cdot f_m}{(d-x) \cdot f_m}
\]

Solved the \(x\), we can get

\[
\frac{1}{2} \frac{S_{m-1}}{\sum f} - \frac{1}{2} = \frac{x \cdot f_m}{(d-x) \cdot f_m}
\]

Therefore, the calculation formula of the obtained median does not include the information of adjacent groups of the median array, and the result is the same. Now, add frequency information for adjacent groups of the median array. And the ratio of the area is \(\frac{\text{the area of trapezoid A2BMeL}}{\text{the area of trapezoid BC2UMe}}\). Then

\[
\frac{1}{2} \frac{S_{m-1}}{\sum f} - \frac{1}{2} = \frac{1}{2} \frac{x \cdot (f_m + f_{m-1})}{(d-x) \cdot (f_m + f_{m+1})}
\]

Solved the \(x\), we can get

\[
Me = L + \frac{\left(\frac{\sum f}{2} - S_{m-1}\right) \cdot (f_m + f_{m+1})}{f_m(f_m + f_{m-1}) + (f_{m+1} - f_{m-1}) \cdot \left(\sum f - S_{m-1}\right)} \times d
\]

Obviously, formula 11 is a special case of formula 12 when \(f_{m-1} = f_{m+1}\). Take the ratio of the area as \(\frac{\text{the area of trapezoid A1BMeL}}{\text{the area of trapezoid BC1UMe}}\). A1 and C1 are the midpoints of AA 2 and CC 2, respectively. And then we have
And then we can get

\[
\frac{1}{2} \frac{S_{m-1}}{\sum f} = \frac{1}{2} \left( f_m + \frac{f_m + f_{m-1}}{2} \right) \frac{d-x}{(f_m + f_{m+1})/2}
\]

(7)

Therefore, formula (5) also becomes a special case of formula 13 in which FM1 is equal to FM+1.

By adding frequency information of adjacent groups into the original interpolation formula, we can break through the restriction of uniform distribution of median array and improve the accuracy of median statistical index calculated in the case of uneven data distribution in the case of multi-source heterogeneous data set [2].

3.2 THE DISTRIBUTED IMPLEMENTATION OF PROPOSED MEDIAN ESTIMATION METHOD

In order to implement the median estimation method of proposed above, we firstly propose a distributed computing framework based on webservice technology, as shown in the figure 2. The framework is divided into two levels: central node and data source node.

Central nodes will interact with data source nodes, control the process of each algorithm calculation and collect the results of each data source with the global median estimate results.

Data source nodes will implement most of the data processing locally, including local computing module and algorithm interface module. Local computing module complete the process of basic calculation and local data process of estimation of median. To provide the basis for invoking the center nodes, algorithm interface module will open the service and function of encapsulating local computing based on webservice technology.

The contents that implement data processing process of data source node are traversing each data source, finding the maximum and minimum values of each data source, and returning them as parameters to the local; According to the obtained overall value range, divide different intervals on average, find out the specific data volume of all data sources in each interval, and return the divided data to the local in the form of a set. The local computing module mainly processes the parameters returned by the interface and performs the improved interpolation median formula for the final verification calculation. The main calculation and processing are as follows: process the parameters returned by the first algorithm interface, obtain the overall value range of the multi-source heterogeneous data set, divide the value range into the corresponding value range according to the number of data sources, and load it into the next algorithm interface; the estimate of the median is then calculated using the improved interpolation median formula based on the set returned by the second interface.
The overall algorithm flow is shown in figure 3. The calculation process of the algorithm is as follows: firstly, the algorithm of obtaining the maximum and minimum values is distributed to each data source to run, and the value range parameter of each data source is obtained, which is returned to local, and the overall value range of multi-source heterogeneous data set is obtained with the support of local computing. The calculation process of the algorithm is as follows: 1. The algorithm for obtaining the maximum and minimum values is distributed to each data source for operation, and the value range parameters of each data source are obtained, and the parameters are returned to the local, and the overall value range of multi-source heterogeneous data set is obtained with the support of local computing; 2. The corresponding local algorithm will be divided into the corresponding value range according to the number of data sources on average, and loaded into the next algorithm interface; 3. The interface of the interval partitioning algorithm is distributed to each data source for execution, and the specific distribution of each data source in different intervals is obtained. In other words, the multi-source heterogeneous data set is reorganized into the scene of traditional grouping median calculation, and returned to the local area in the form of set; finally, after locally obtaining the multi-source heterogeneous data set, which has been processed into the traditional grouping form, the interpolation method formula with improved calculation accuracy is used for calculation, and the median of experimental results is obtained[5].

Figure 2: Algorithmic Architecture Diagram

Figure 3: Algorithm flowchart
4. EXPERIMENTAL STUDY

4.1 THE ACCURACY OF DISTRIBUTED ALGORITHM
In the accuracy evaluation of the experimental median, we mainly use accuracy as an evaluation index. During the experiment, we will first calculate the median of the overall data and get the control value before we divide the data into different data sources, which can be used to verify the accuracy of the comparison with the value obtained from the experimental results. Accuracy is calculated as 1 minus the percentage difference between the accuracy value and the calculated result. Accuracy rate will be the main reference index to verify the feasibility of the algorithm. Therefore, the main purpose of this accuracy experiment is to test the accuracy of the median index obtained by this algorithm in the environment of multi-source and heterogeneous data sets by conducting experiments of different data sizes in the same experimental environment, and then evaluate the feasibility of this algorithm.

The specific experimental methods and procedures are as follows: firstly, use webservice technology to load the relevant value range calculation algorithm, distribute it to the prepared data source for execution, and obtain the value range parameters of each data source; then, the returned parameter results are calculated locally to get the appropriate value range grouping, and the grouping results are loaded into the interface and then executed on each data source to get the grouping results of each data source. The returned grouping results are calculated locally using the median algorithm to get the median results of the experiment; Finally, the data sources used in the control group experiment were obtained locally, and the median result of the control group was calculated using traditional methods, and the accuracy was calculated by comparing with the experimental results.

After we use Java programming to implement the relevant algorithm of median calculation, we use webservice technology to load and form the algorithm interface, and distribute it to each data source to run. Table 4 is the result returned after running at 6 data sources (with a total data amount of 5000) using the interface that obtains the value range of data source. In this environment, the value range of the whole data source is between 15 and 75, which is within the reasonable range of the experimental data.

| Data source number | min | max |
|--------------------|-----|-----|
| 1                  | 20  | 75  |
| 2                  | 20  | 74  |
| 3                  | 20  | 68  |
| 4                  | 20  | 70  |
| 5                  | 15  | 75  |
| 6                  | 20  | 72  |

The values range is integrated into 6 groups, and the value range group is calculated. Then, using these already-divided groups finds the specific data of each data source separately, and loads the specific values in the corresponding value range of each data source into the corresponding groups. Table 5 shows the grouping results of the data source returned by the algorithm interface.

| Group | Data number |
|-------|-------------|
| 1     | 831         |
| 2     | 818         |
| 3     | 879         |
Finally, using traditional formulas and improved formulas respectively calculated the locally processed data, and comparing the estimated value of the calculated result with the experimental group obtained the accuracy rate. In order to further verify the reliability of the algorithm under different data quantities, we will change the amount of data of each data source to repeat the experiment and get experimental results under different scale and distribution data sets. The results of all experiments obtained after 5 experiments were shown in Table 6.

Table 6. Different data volume accuracy rate results

| Experiment | amount | Frequency | Improved algorithm accuracy |
|------------|--------|-----------|----------------------------|
| 1          | 500    | 98.5      | 98.7                       |
| 2          | 1000   | 97.9      | 98.1                       |
| 3          | 5000   | 90.2      | 97.3                       |
| 4          | 10000  | 82.4      | 96.5                       |
| 5          | 50000  | 77.5      | 95.9                       |

This experiment shows that the improved median calculation formula and webservice technology can achieve efficient and accurate median acquisition in the distributed computing environment. According to the experimental result data in Table 6, we can know that as the amount of data increases, the accuracy of the median estimate obtained by the algorithm is gradually decreasing. The accuracy rate of the improved algorithm is much smaller than the accuracy of the traditional algorithm. It shows that as the amount of data increases, the most important factor affecting the accuracy is the average degree of distribution of data. The improved algorithm reduces the impact of uneven data distribution on the calculation results.

4.2 THE DISTRIBUTED IMPLEMENTATION OF PROPOSED MEDIAN ESTIMATION METHOD

In order to measure the feasibility of the algorithm in terms of performance and efficiency in practical application, we need not only to study the accuracy of the median of experimental results, but also to measure the performance. The specific experimental method is to record the time difference in different data sizes with traditional calculation methods. We firstly recorded the time spent on each experiment execution. The specific recording time periods were the weighted total running time of the two interfaces of each data source. With the same amount of data, the traditional median calculation method was used to obtain the data from all data sources and collect them at home. Then we used the general median calculation method to calculate the results and record the time. Finally, the total time obtained by the two methods is compared to get the results. The results are recorded as shown in table 7.

Table 7. Execution time cost for different data quantities

| Experiment numbers | amount | Time(ms) | T-time(ms) |
|--------------------|--------|----------|------------|
| 1                  | 500    | 17       | 80         |
| 2                  | 1000   | 43       | 753        |
| 3                  | 5000   | 546      | 10000      |
According to the data of table 7, the time consumption of both approaches increases with the amount of data. At the same time, the time consumption of traditional methods grows much faster than that of improved methods. However, the improved method takes less time than the traditional method at the same data size. Therefore, the calculation method improves the calculation efficiency and takes less time than the traditional method in general, and with the increase of data size, this method has more advantages in computing efficiency.

5. CONCLUSIONS
This research mainly improves the original traditional median calculation formula, and combines the data preprocessing of the multi-source heterogeneous data set with the support of webservice technology to efficiently and accurately obtain the multi-source heterogeneous data set statistical indicators. It also verifies the accuracy of the interpolation method in the calculation of the relevant median and Lunzhi Gan's improvement on the accuracy of the median calculation. Today, with the rapid development of information technology, massive data will still be generated and accumulated rapidly. Data analysis of multi-source heterogeneous data sets will become a very important research direction in the future. Therefore, when solving the problem of the acquisition of the median of a multi-source heterogeneous data set, this paper also supplements the shortcomings of the current research in this aspect, and reduces the overhead in the acquisition of statistical indicators of multi-source heterogeneous data sets. It is of great significance and value for the research direction of subsequent multi-source heterogeneous dataset data analysis. Because this study is based on the acquisition scenarios of statistical indicators of multi-source heterogeneous data sets, and selects the median indicators that are difficult to obtain and representative to conduct special research, there are no other relevant statistical indicators. The methods for obtaining indicators such as mean value and variance are studied, so relevant research needs to be supplemented and solved. In addition, because of the nature of the algorithm itself, the algorithm interface needs to determine the number of data sources before designing, and the interface for calculating the value range of each data source needs to manually input the corresponding data before the subsequent calculation. At present, the restrictions are large and the process is cumbersome, so there is still room for improvement in programming.

ACKNOWLEDGMENTS
This paper is supported by Guangdong province Science&Technology Plan (project Number:2016A040403029) and Guangdong province Philosophy&Social Science "12th five" plan Interdisciplinary project (project number:GD15XGL02).

REFERENCES
[1] Master's thesis of Beijing University of Posts and Telecommunication. The Design and Implementation of a Real-time Query System for Massive Data Based on Spark.2017
[2] Zhihun Gan. Improvement of calculation method for median of group spacing grouping sequence[J]. Statistics and Decision.2013(02):18-20.
[3] Yuewei Ding, Xiaofeng Zhang, Yueai Huang. The calculation of median and mode in grouping statistical dat[J]. Statistics and Decision. 2006(05):146-147
[4] Jie Cao, Lingling Ma, Rongrong Jiao. Some Thoughts on Median Computing under Group Spacing Grouping[J]. Industrial & Science Tribune,2017,16(05):67-68.
[5] Master's thesis of Inner Mongolia Agricultural University. The Research and Build of Massive Pictures Retrieval Platform Based on Hadoop.2007