Parallel computation of matrix norm based on MapReduce

Yuqiang Sun, Dongyu Zhang, Yan Chen, Bixia Chao, Yuwan Gu*

School of Mathematics and Physics, ChangZhou University, Jiangsu, Changzhou213164, China

**Corresponding author's e-mail: shisungu@126.com*

Received 24 November 2014, www.cmnt.lv

Abstract

A kind of parallel programming method based on MapReduce model is proposed, in allusion to data characteristic of having specific data partitioning requirement, parallel computation of matrix norm is implemented on the platform of high-performance MapReduce. Comparing with the traditional parallel programming model, MapReduce model parallel program can satisfy to requirement of high performance numerical calculations well, its programming for simplicity and readability can improve parallel programming efficiency in effect.

Keywords: MapReduce, Numerical computation, Matrix norm, Parallel computation, Data partitioning, High-performance

1 The definition and property of matrix norm

Definition 1: given $A \in C^{m\times n}$, prescribed a real-valued function of A on $C^{m\times n}$ according to a certain rule, marked $||A||$, it satisfies to the following 4 conditions:

(1) Non negative: if $A \neq 0$, then $\|A\| \succ 0$; if $A = 0$, then $\|A\| = 0$.

(2) Homogeneity: for arbitrary $k \in C$, $\|kA\| = |k|\|A\|$.

(3) Triangle inequality: for arbitrary $A, B \in C^{m\times n}$, $\|A+B\| \le \|A\| + \|B\|$.

(4) Compatibility: when the matrix product AB has meaning, if $\|AB\| \le \|A\|\|B\|$, then $\|A\|$ is called matrix norm.

Given $A = (a_{ij}) \in C^{n\times n}$, the real-valued function of the following provisions

$$\|A\|_{m_1} = \sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}| \quad , \quad \|A\|_{m_\infty} = n \bullet \max_{i,j}|a_{ij}| \quad ,$$

$\|A\|_{m_2} = (\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}|^2)^{\frac{1}{2}}$, they are all norm of matrix A.

Theorem 1: given $A = (a_{ij}) \in C^{m\times n}$, $x = (x_1, x_2, \cdots, x_n)^T \in C^n$, operator norm of three kinds of norms $\|x\|_1$, $\|x\|_2$, $\|x\|_\infty$ that belongs to the vector x is $\|A\|_1 = \max_j \sum_{i=1}^{m}|a_{ij}|$ (known as a column norm); $\|A\|_2 = \sqrt{\lambda_{\max}(A^H A)}$ (known as the spectral norm), where $\lambda_{\max}(A^H A)$ is the maximum of the absolute value of matrix $A^H A$ characteristic value; $\|A\|_\infty = \max_i \sum_{j=1}^{n}|a_{ij}|$ (called line norm) in turn.

2 Outline of MapReduce

MapReduce can be implemented in many ways, and indeed it has various implementations [1-4]. Here, we will outline MapReduce as described in [1]. In a nutshell, MapReduce computations consist in processing input data sets by creating a set of intermediate (key, value) pairs, and then reducing them to yet another list of (key, value) pairs. The computations are performed in parallel.

More precisely, MapReduce applications are divided into two steps. In the first step a Map function processes the input dataset (e.g. a text/HTML file), and a set of intermediate (key1, value1) pairs is generated. In the second step the intermediate values are sorted by key1, and a Reduce function merges the intermediate pairs with equal values of key1, to produce a list of pairs (key1, value2). Thus, the input dataset is transformed into a list of key/value pairs. Let us consider examples given in [1]. Counting occurrences of words in a big set of documents can be organized in the following way. Map function emits intermediate pair (word,1) for each word in the input file(s). The intermediate pairs are reduced by sorting them by word, summing 1s, and producing pairs (word, count). In the inverted index computation all documents comprising certain words must be identified. The Map function emits pairs (word, docID), where docID is a document identifier (e.g. a URL of a web page). In the Reduce function all (word, docID) pairs are sorted, and pairs (word, list_docIDs) are emitted, where list_docIDs is a sorted list of docIDs. There are many types of practical applications which can be expressed in the MapReduce model. More detailed and advanced examples are given in [1, 2, 3, 5].

Both map and reduce operations are performed in parallel in a distributed computer system. Processing a MapReduce application starts with splitting the input files into load units (in [1] called splits). Many copies of the program start on a cluster of machines. One of the machines, called the master, assigns work to the other computers (workers). There are m map tasks and r reduce tasks to assign. In the further discussion the map tasks will be called mappers, and the reduce tasks reducers. A worker which received a mapper reads the corresponding input load unit and processes the data using the Map function. The output of this function is divided into r parts by the partitioning function and written to r files on the local disk. Each of the r files corresponds to one of the reducers. Usually the partitioning function is something like hash(key1) mod r. The information about local file locations is sent back to the master, which forwards it to the reduce workers.

When a reduce worker receives this information, it reads the buffered data from the local disks of the map workers.

After reading all intermediate data, the reduce worker sorts it by the intermediate keys in order to group together with all occurrences of the same intermediate key. Each key and the corresponding set of values are then processed by the Reduce function. Its output is appended to a final output file for a given reducer. Thus, the output of MapReduce is available in r output files. The execution of MapReduce is completed when all reducers finish their work [6, 7].

Basic framework of MapReduce is shown in figure 1:

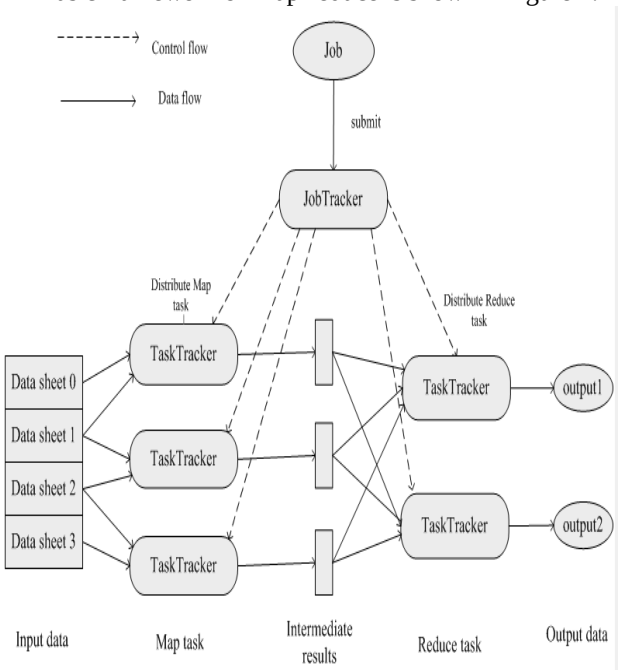

FIGURE 1 Basic framework of MapReduce

3 Parallel computation of matrix norm based on MapReduce

Given $A=(a_{ij})\in C^{n\times n}$, serial algorithm of real-valued function $\|A\|_{m_1}=\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}|$ is as follows:

Begin
  A=0
  For i=1 to n do
    For j=1 to n do
    A=A+ $|a_{ij}|$
    End for
  End for
  $\|A\|_{m_1}$ =A
End

Because of MapReduce Still belongs to the explicit data partition and parallel computing model, it shall be that the matrix will be divided into several blocks of data in accordance with the specific data partition strategy, the using partition method is according to line by continuous partition or according to column by continuous partition etc. Then according to serial computing semantic and partition method of matrix norm $\|A\|_{m_1}=\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}|$, so it can easily write treatment process of Map and Reduce, description is as follows:

(1) Map: Key=ij,Value=$|a_{ij}|$, if according to line by continuous partition, then the intermediate process is adding Value of the same i; if according to column by continuous partition, then the intermediate process is adding Value of the same j.

(2) Reduce: if according to line by continuous partition, key=i, value= the value corresponding to i obtained in the intermediate process; if according to column by continuous partition, key=j, value= the value corresponding to j obtained in the intermediate process. Finally, all value is added together, namely to obtain the value of matrix norm $\|A\|_{m_1}$.

Given $A=(a_{ij})\in C^{n\times n}$, serial algorithm of real-valued function $\|A\|_{m_\infty}=n\bullet\max_{i,j}|a_{ij}|$ is as follows:

Begin
  A=$|a_{11}|$
  For i=1 to n do
    For j=1 to n do
    If A<$|a_{ij}|$ then A=$|a_{ij}|$
    Else A=A
    End for
  End for
  $\|A\|_{m_\infty}=n*A$
End

Because of MapReduce Still belongs to the explicit data partition and parallel computing model, it shall be that the matrix will be divided into several blocks of data in accordance with the specific data partition strategy, the using partition method is according to line by continuous partition or according to column by continuous partition etc. Then according to serial computing semantic and partition method of matrix norm $\|A\|_{m_\infty}=n\bullet\max_{i,j}|a_{ij}|$, so it can easily write treatment process of Map and Reduce, description is as follows:

(1) Map: Key=ij,Value=$|a_{ij}|$, if according to line by continuous partition, then the intermediate process is getting the maximum Value of the same i; if according to column by continuous partition, then the intermediate process is getting the maximum Value of the same j.

(2) Reduce: if according to line by continuous partition, key=i, value= the value corresponding to i obtained in the intermediate process; if according to column by continuous partition, key=j, value= the value corresponding to j obtained in the intermediate process. Finally, obtaining the maximum Value, this value is multiplied by n times, then namely matrix norm $\|A\|_{m_\infty}$.

Given $A=(a_{ij})\in C^{n\times n}$, serial algorithm of real-valued function $\|A\|_{m_2}=(\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}|^2)^{\frac{1}{2}}$ is as follows:

Begin
  A=0
  For i=1 to n do
    For j=1 to n do
    A=A+ $|a_{ij}|^2$

End for
End for
$\|A\|_{m_2} = \sqrt{A}$
End

Because of MapReduce Still belongs to the explicit data partition and parallel computing model, it shall be that the matrix will be divided into several blocks of data in accordance with the specific data partition strategy, the using partition method is according to line by continuous partition or according to column by continuous partition etc. Then according to serial computing semantic and partition method of matrix norm $\|A\|_{m_2} = (\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}|^2)^{\frac{1}{2}}$, so it can easily write treatment process of Map and Reduce, description is as follows:

(1) Map: Key=ij,Value= $|a_{ij}|^2$,if according to line by continuous partition, then the intermediate process is adding Value of the same i; if according to column by continuous partition, then the intermediate process is adding Value of the same j.

(2) Reduce: if according to line by continuous partition, key=i, value= the value corresponding to i obtained in the intermediate process; if according to column by continuous partition, key=j, value= the value corresponding to j obtained in the intermediate process. Finally, all value is added together and the result is squared root, namely to obtain the value of matrix norm $\|A\|_{m_2}$.

Given $A=(a_{ij}) \in C^{m\times n}$, serial algorithm of Column norm $\|A\|_1 = \max_j \sum_{i=1}^{m}|a_{ij}|$ is as follows:

Begin
For j=1 to n do
A[j]=0
For i=1 to m do
A[j]=A[j]+ $|a_{ij}|$
End for
End for
A=A[1]
For j=1 to n do
If A<A[j] then A=A[j]
End for

Because of MapReduce Still belongs to the explicit data partition and parallel computing model, it shall be that the matrix will be divided into several blocks of data in accordance with the specific data partition strategy, the using partition method is according to column by continuous partition etc. Then according to serial computing semantic and partition method of matrix norm $\|A\|_1 = \max_j \sum_{i=1}^{m}|a_{ij}|$, so it can easily write treatment process of Map and Reduce, description is as follows:

(1) Map: Key=ij, Value=$|a_{ij}|$, the intermediate process is adding the Value of the same j.

(2) Reduce: key=j, value= the value corresponding to j obtained in the intermediate process. Finally, obtaining the maximum Value, namely to obtain the value of matrix norm $\|A\|_1$.

Given $A=(a_{ij}) \in C^{m\times n}$, serial algorithm of line norm $\|A\|_\infty = \max_i \sum_{j=1}^{n}|a_{ij}|$ is as follows:

Begin
For i=1 to m do
A[i]=0
For j=1 to n do
A[i]=A[i]+ $|a_{ij}|$
End for
End for
A=A[1]
For i=1 to m do
If A<A[i] then A=A[i]
End for

Because of MapReduce Still belongs to the explicit data partition and parallel computing model, it shall be that the matrix will be divided into several blocks of data in accordance with the specific data partition strategy, the using partition method is according to line by continuous partition etc. Then according to serial computing semantic and partition method of matrix norm $\|A\|_\infty = \max_i \sum_{j=1}^{n}|a_{ij}|$, so it can easily write treatment process of Map and Reduce, description is as follows:

(1) Map: Key=ij, Value=$|a_{ij}|$, the intermediate process is adding the Value of the same i.

(2) Reduce: key=i, value= the value corresponding to i obtained in the intermediate process. Finally, obtaining the maximum Value, namely to obtain the value of matrix norm $\|A\|_\infty$.

Given $A=(a_{ij}) \in C^{m\times n}$, parallel computation process based on MapReduce of spectral norm $\|A\|_2 = \sqrt{\lambda_{\max}(A^H A)}$ is as follows:

There are two Mapreduce processes, $A^H A$ is implemented in the first, λ value of $A^H A$ is solved in the second.

The Map and Reduce treatment process of implementing $A^H A$ is as follows:

A^H is partitioned by line, A is partitioned by column.

(1) Map: $\text{Key}_1 = \text{i}$, $\text{Value}_1 = (\text{a}_{i1},\text{a}_{i2},\cdots,\text{a}_{in})$, $\text{Key}_2 = j$, $\text{Value}_2 = (\text{a}_{1j},\text{a}_{2j},\cdots,\text{a}_{mj})$, then the intermediate process is that the vector corresponding i and the vector corresponding j multiplied by two.

(2) Reduce: key=ij, value= a_{ij} , namely to obtain matrix $A^H A$.

The Map and Reduce process of treatment process of implementing λ value of matrix $A^H A$ according to a lower triangular in the parallel algorithm of LU decomposition in reference[8], then the diagonal is multiplied, namely to obtain all λ value, getting the maximal λ value, and the maximal λ value is squared root, namely to obtain the value of spectral norm $\|A\|_2$.

4 Conclusion

A parallel computation method of matrix norm based on the MapReduce model is proposed in the paper, in some areas related to computation of the matrix norm, parallel computation method in the paper brings convenient. As a new

type of parallel and distributed programming model, MapReduce model has a high parallel representation abstract [9, 10], can effectively reduce the difficulty of parallel programming, and upgrades the parallel programming productivity. The next step for the research work is that the model is introduced to high performance computing area of more numerical value / non numerical value.

Acknowledgment

Supported by The National Natural Science Fund (11271057, 51176016) and the project of general office of Broadcasting and Television (GD10101) and Natural Science Fund in JiangSu (BK2009535) and Natural Science Fund in ZheJiang (Y1100314) and Jiangsu Province ordinary university innovative research project (SCZ1412800004).

References

[1] Dean J, Ghemawat S 2004 MapReduce: simplified data processing on large clusters *in: Proc. OSDI'04: Sixth Symposium on Operating System Design and Implementation*, San Francisco, CA 137–50 http://labs.google. com/papers/mapreduce.html

[2] Lin J, Dyer C 2010 Data-Intensive Text Processing with MapReduce *Morgan &Claypool* 2010

[3] Pike R, Dorward S, Griesemer R, Quinlan S 2005 Interpreting the data: parallel analysis with Sawzall *Scientific Programming* **13**(2005) 277–98

[4] Wikipedia, MapReduce, *http://en.wikipedia.org/wiki/MapReduce* [Online; 17-February-2010]

[5] Ranger C, Raghuraman R, Penmetsa A, Bradski G, Kozyrakis C 2007 Evaluating MapReduce for multi-core and multiprocessor systems *in: Proceedings of International Symposium on High Performance Computer Architecture* HPCA 13–24

[6] Wang W, Li J H, Ding R F 2011 Maximum likelihood parameter estimation algorithm for controlled autoregressive autoregressive models *International journal of computer mathematics* **88**(16) 3458-67

[7] Bonettini S, Landi G, Piccolomini E L, Zanni L 2013 Scaling techniques for gradient projection-type methods in astronomical image deblurring *International journal of computer mathematics* **90**(1) 9-29

[8] Zheng Qi-long etc. 2010 Application of HPMR in Parallel Matrix Computation *Computer Engineering* **36**(8) 49-51

[9] Zhiyuan Shi, Volker Gruhn, Yuwan Gu, Yuqiang Sun 2013 The Study of Reuse Mashup Technology Based on Using Frequency *Information Technology Journal* **12**(14) 2669-72

[10] Yihong Cao, Yuwan Gu, Huanhuan Cai, Yuqiang Sun 2013 An Improved Decision Tree Algorithm Based on The Attribute Set Dependency *Information Technology Journal* **12**(22) 6641-5