International Journal of Grid Distribution Computing
Vol. 8, No.4, (2015), pp. 135-144
http://dx.doi.org/10.14257/ijgdc.2015.8.4.13

Cloud Computing Environments Parallel Data Mining Policy Research

Wenwu Lian, Xiaoshu Zhu, Jie Zhang and Shangfang Li

Yulin Normal University, No. 299 Jiaoyu Road (minddle), Yulin, Guangxi, China
jgxylww@126.com

Abstract

With the rapid development of computer science and technology, more and more data is stored in the computer storage media , data mining (DM) emerged as an interdisciplinary subject , it is based on a method previously used researcher science and algorithms . Cloud computing is an emerging shared infrastructure approach that open standards and service-based, Internet-centric , providing safe , fast and convenient data storage and network computing services. The cloud computing applications to data mining, you can tap the growing number of mass data solutions. This paper presents a cloud computing environment suitable for partitioning the data set allocation method and data sets; introduces improved Apriori algorithm based on its calculation of two parallel processes running on the platform in the cloud, the results of the simulation.

Keywords: *cloud computing, data mining, massive data, Apriori algorithm*

1. Introduction

With the rapid development of computer science and technology, more and more data is stored in computer storage medium, these data are often a large number of complexes, heterogeneous, there is noise, it is difficult to understand directly, which contains knowledge and patterns are more difficult to find. In other research and commercial areas, often need to analyze these data, and found valuable information and patterns or make predictions about the future. To meet these challenges, researchers from different disciplines to bring together, you can proceed with the development process more efficient, scalable technology and tools of different data types, data mining [1] (Data Mining, DM) emerged as an interdisciplinary it is based on the methodology and algorithm researchers previously used. Data mining using some thoughts from the following areas: (1) statistical sampling, estimation and hypothesis testing, (2) artificial intelligence, pattern recognition and machine learning search algorithms, modeling techniques and learning theory. Meanwhile, data mining also quickly accepted the idea from other areas; these areas include optimization, evolutionary computation, and information theory, signal processing, visualization and information retrieval. Data mining , also known as knowledge discovery in databases (Knowledge Discovery in Database, KDD), is a large , incomplete , there is noise , fuzzy and random practical application to extract data from hidden in them , people without prior know, but is potentially useful information and knowledge. Before its appearance makes a lot of history cannot be understood using data obtained on a number of areas such as scientific research and business decisions played a guiding role, with far-reaching social and economic significance [2].

Due to the huge expansion of information, data can be utilized too large, making the growing of large-scale data mining applications. Progress in the latest technology , so many fields of science using large data sets are also possible, but the speed of such data sets generated by them far more than the speed of manual analysis . Massive data mining

ISSN: 2005-4262 IJGDC

technology has become a hot research topic is a difficulty to be solved. However , mining large data sets requires tremendous computing resources because the data run on large data sets mining algorithm , if the traditional computer to solve the case, would be to spend time.

Thus, high-performance parallel computers and parallel data mining algorithms largely become an effective means to solve such problems [3]. Overall, for the following two reasons, the use of parallel computation requires data mining methods: First, because the analysis of data is complex and variable algorithm requires a high computational power, this calculation is only possible to provide a parallel computer; Second, handle large data sets and increasing rapidly, and parallel computers can be effectively organized to handle large amounts of data. Knowledge mining parallel computing clusters with large data sets can raise the quality of the analysis can also improve performance. Therefore, high-performance parallel data mining technology research to improve the efficiency and accuracy of mining massive data mining to solve the problem, has important practical significance.

Cloud computing is an emerging shared infrastructure approach that open standards and service-based, Internet-centric , providing safe , fast and convenient data storage and network computing services. The cloud computing applications to data mining, you can tap the growing number of mass data solutions. Cloud computing is facing massive TB and even PB level data, how to obtain valid information from the data, which will be one of the key decisions the success of cloud computing applications.

In addition to using parallel computing technology to accelerate the speed of data processing , but also need new ideas , methods and algorithms to perform more accurate, fast, powerful data mining[4] .

Cloud computing (Cloud Computing) is a new model of business computing. It distributed computing tasks in a large pool of computer resources constitute enable various application systems to obtain computing power needed storage space and a variety of software services. Definition of cloud computing has narrow and broad points. Cloud computing refers to the delivery of narrow and use of IT infrastructure mode refers to the demand, and scalable way to get resources (hardware, platform, software) required by the network. Provide resources network is called "cloud." "Cloud" of resources in the user appears to be infinitely scalable, and can be readily available, on-demand, any time extension, pay per use. This feature is often referred to as the use of water and electricity as the use of IT infrastructure. Cloud computing broadly refers to the delivery of services and usage patterns, refers to the demand, and scalable way to obtain the necessary services through a network. This service can be IT and software, Internet-related; it can be any other service.

First, cloud computing provides the most reliable and secure data storage center; users do not have to worry about data loss, virus attacks and other problems. Secondly, cloud computing client devices require a minimum, it is also the most convenient to use. In addition, cloud computing data and applications can be easily shared between different devices. Finally, the use of cloud computing network provides us with an almost infinite number of possibilities [5].

Cloud computing out of the lab just around the corner towards commercialization , cloud computing and data storage features make commercialization, mass data storage and mining cloud computing environment is a theoretical and practical value of the research. Work in this paper will have a higher reference value theory and application.

2. Related Works

Enterprise is the main driving force to promote the rapid development of cloud computing. Currently , in addition to Amazon, Google, VMware, IBM, Microsoft, Cisco and Yahoo and other foreign IT giants have developed cloud computing platform , to

carry out cloud computing arrangement , the traditional Major OEM manufacturers, such as: Quanta, Foxconn , Asus also invested heavily in the development of cloud computing terminals and servers. The cloud computing industry applications such as: extraordinary development SALESFORCE, FACEBOOK and other companies market position has threatened to Microsoft, Google and other IT giants.

In China, the cloud computing industry is also booming. Such as: China Mobile, China Unicom, Century Internet , Dr. Peng Companies are also actively involved in the development of cloud computing technologies and products , the introduction of some products and test platforms.

From 2007, China Mobile began to conduct research and development of cloud computing, cloud computing is one of the earliest intervention research and practice of business. It is based on Hadoop open source software on independently developed the " big cloud" cloud computing systems , enabling key features of distributed file systems , massive databases, distributed computing framework , cluster management , virtual machine management , and has applied for more than 10 patents. By the end of 2008 , China Mobile and further construction of the 256 servers , the "big cloud" test platform 1000 CPU, 256TB storage composition , combined with existing network data mining, user behavior analysis and other needs in Shanghai , Jiangsu and other places for the application of the pilot, in improve efficiency, reduce costs, energy conservation and achieved very significant results. Currently, China Mobile is planning further expansion of the test platform, reaching 1,000 servers, 4,000 CPU, storage size 4000TB of [6] .

March 17, 2008, Google CEO Eric • Global Smit (Eric Schmidt) during his visit in Beijing, announced the launch of cloud computing (Cloud Computing) program in mainland China. In China’s "cloud computing” program, Tsinghua University is the first university participation and cooperation. Google provides course materials for finishing professor at Tsinghua University , provides laboratory equipment , and assist schools to build "cloud computing" experimental environment in the existing computing resources . 年初 2008, IBM has established cooperation with the Wuxi Municipal Cloud Computing Center in Wuxi Software Park , began in China 's cloud computing business applications. Rising in July 2008 launched the “cloud security " program. By the end of 2008 Institute of Computing has developed cloud -based parallel data mining platform PDMiner. This shows that a growing number of IT suppliers will China as a cloud computing business development hotspot, cloud computing business has great potential for development in the Chinese market.

Google’s cloud computing technology is actually for a particular web application Google custom. For large -scale internal network data features , Google proposed a set of distributed parallel infrastructure cluster approach , the use of software capabilities to deal with the frequent occurrence of the cluster node failure problems. Google cloud computing infrastructure model used consists of four independent and tightly coupled systems, including Google built on a cluster file system Google File System, according to the characteristics of the proposed application Google Map / Reduce (mapping / simplification , also known as MapReduce) simplified programming model, a distributed locking mechanism Chubby and Google developed the model of large-scale distributed database BigTable

IBM is another important enabler of cloud computing and cloud computing internally named “Blue Cloud “program. IBM has all the favorable factors for the development of cloud computing services, such as application servers, storage, management software, middleware, etc. Therefore, IBM also has a natural technological advantage. IBM recently introduced the idea of "New Enterprise Data Center”, which combines the advantages, envisaged Web-centric cloud computing model and the current corporate data center [7].

The key technical Map / Reduce now cloud computing has been applied in a number of machine learning and data mining algorithms. 2006 , relying on a simple Map / Reduce programming model for multicore processors , CT Chu , who achieved a series of

machine learning methods, including locally weighted linear regression, K -means, logic regression, naïve Bayes, linear support vector, the independent variable analysis, Gaussian discriminant analysis, expectation-maximization and reverse spread;. D Gillick, who improved on the Hadoop MapReduce implementation mechanism, and in the medium-sized clusters tested a series of standard data mining tasks, indicating MapReduce mechanism suitable for large-scale data mining, In 2007, T. Liu et al describes a nearest neighbor search algorithm scalable version that in 2008, C. Moretti et al proposed a scalable model for mining abstract classification is based on MapReduce computation model and cloud computing environments. 2009, Wang and Li Ming E SPRINT describe the implementation process of parallel algorithms in Hadoop MapReduce programming model.

Objectively speaking, now parallel data mining has been widely and successfully applied parallel data mining applications are two potential reasons: ① high degree of complexity of the data mining algorithm itself had to use a high-performance parallel computing. This idea was born parallel data mining task parallelism. ② Due to the amount of data mining is usually analyzed data itself is very large, the large amount of data mining work to parallel processing has become a natural idea. This approach we call data parallelism. In parallel data mining, this is only the two most basic mining parallelisms in many parallel algorithms; these two mechanisms are always simultaneously present.

Parallel algorithms for processing massive data significantly better than the serial algorithm, for serial algorithm, the huge amounts of data to be processed are very difficult. But now, there are still many parallel algorithms imperfections, there are some issues to be addressed, including: 1) Data Distribution: One of the advantages of parallel algorithms is that each node can reduce the size of a subset of the database deal with. 2) To reduce I / O operations: parallel algorithms must also be taken to minimize the database I / O operations. It is also an important influence on the performance of the algorithm, in particular those shared memory systems, which are mostly the serial I / O, frequent I / O operations, will greatly reduce the efficiency of the algorithm. 3) Load Balancing: To maximize the efficiency of the algorithm in parallel, the workload of each node should be as equal. In the algorithms have been proposed, most of them are based on the initial data distribution static load balancing. The actual calculation of the workload of each node may be very different, but the overall performance of the system depends on the node to finalize work. Therefore, dynamic load balancing during operation is very important. 4) Reduce the communication and synchronization: good parallel algorithm should allow each node asynchronous operation, should not be subject to traffic delays and restrictions synchronization requirements and often stopped to wait [8].

3. Proposed Scheme

Google's MapReduce library input files were divided into N data segments, and then evenly allocated to M Nodes, due to cloud computing cluster environments are heterogeneous, so this method does not take full advantage of cluster computing resources, this paper parallel computing cloud computing environments demand conditions and parallel data mining and data set is divided allocation set to do the research. Also, this paper two improved Apriori algorithm, so that in the cloud computing environment can be performed in parallel mining frequent itemsets massive data.

Data sources often contain huge amounts of data, requires a lot of I / O overhead and enough memory space data mining algorithm, which has become a bottleneck algorithm scalability and response time. Although there are some parallel data mining algorithms , but the general data parallel computing on how the data are not to be considered a reasonable division , so dig quality is not high. In fact, the data set into the performance of parallel data mining algorithms have a tremendous impact, data partitioning strategy for

CPU and bandwidth can take full advantage of system resources , reduce communication overhead , balance the system load and reduce the amount of calculation , which can most good play crucial parallelism and system performance. Improve or design a new data set into ways to improve the efficiency of the data set is divided parallel mining first need to solve important problems.

3.1. Apriori Algorithm Description and Problem Analysis

1 An algorithm description

Apriori algorithm is the basic algorithm for mining frequent [9] item sets to generate the desired Boolean association rules, but also a very influential association rule mining algorithm. Apriori algorithm is based on a priori knowledge about the nature of frequent item sets named. The algorithm uses an iterative search method called layer by layer, using k item sets to generate (k + 1) item sets. Specific treatment consists of two main steps, namely the connection and delete. L_{k-1} To be used to produce L_k , for example, the mining process is as follows:

(1) Connection procedure: to discover L_k , connect L_{k-1} by itself produces a set of candidate k item sets C_k . Set I_1 and I_2 for the two items in the set L_{k-1} , $I_i[j]$ indicating that the jth entry in $I_i[1] < I_i[2] < \ldots < I_i[k-1]$. For convenience, Apriori assumes centralized transaction or item ordered by dictionary order items. For the (k-1) item set I_i , which means the items are sorted, so $I_i[1] < I_i[2] < \ldots < I_i[k-1]$. Referred to L_{k-1} as the connection operation $L_{k-1} \oplus L_{k-1}$, which means that if the I_1 and I_2 the (k-2) term is the same, the contents in I_1 and I_2 of the L_{k-1} can be connected together.

(2) Delete the steps of: C_k is a superset of the set L_k which are not necessarily frequent item sets, but all the frequent k item sets necessarily in L^p that there is $L_k \subseteq C_k$. Scan again database can decide in support count each candidate set, and thus get the individual elements of L_k (k frequent item sets). All support count is not less than the minimum support count of the candidate set is frequent item set belongs L_k . However, due to the candidate set in C_k a lot, so the amount of calculation involved in the operation is very large.

To improve the efficiency of frequent item sets produced layer by layer. Apriori algorithm is the nature of a frequent item sets for narrowing the search space: "All non-empty set of frequent item sets must also be frequent." That is, if a candidate k item sets any one subset (k-1) does not belong to item set L_{k-1} , then the candidate k item sets cannot be frequent, and thus it can be deleted therefrom C_k .

Advantage of Apriori algorithm is simple, easy to understand, no complicated derivation. In addition, the nature of the application Apriori algorithm in many cases greatly reduces the size of the candidate needs to be checked, so the algorithm efficiency is greatly improved [10].

2 Problem Analyses

(1) Multiple scans database

Apriori algorithm is performed at each iteration of the time needed to scan a database, usually when the length N excavated the largest frequent item sets, the need to scan N times the database. In practical applications, it is often necessary to tap length mode, repeatedly scans the database will have a tremendous overhead.

(2) May produce a large number of candidate item sets

Apriori algorithm to generate the iterative process in memory, processing and preservation of candidate item sets, this number is sometimes very great. For example, if there are 104 frequent a set, the Apriori algorithm to generate close to 5 × 107 candidates two sets. In addition, in order to find the length of frequent patterns 100, it must generate

up 2100≈1030 candidate, leading to poor adaptation algorithm in breadth and depth.

In short , Apriori by scanning the database multiple times to find all the frequent item sets , and multiple times to scan a database for storing massive amounts of data will spend a lot of time and memory space, which will become the bottleneck of Apriori algorithm. For this reason, research on parallel data mining algorithms in recent years continues to heat up.

3.2. Improved General Idea

Because of the characteristics of cloud computing environments with distributed, parallel execution of algorithms can support, thereby improving the efficiency of mining. In this paper, under the cloud computing environment to effectively implement a parallel association rule mining as the target, Apriori algorithm has been improved. Among them, many small number of nodes to perform Map task than the number m subset of data, execute Reduce the number of nodes are usually their task specified by the user, you can divide the data set method for specific data mining algorithms set during division, However, this article only data set evenly divided horizontally, except the last one data set for each of the remaining data set size 16MB.

3.2.1. The First Improvement

Specific improvement ideas are:

(1) Evenly into the size of data corresponding to a subset of the n level of the transaction database, a subset of the data sent to the node m.

(2) Each node scan its data subset to produce a collection of a local candidate k item sets, denoted by C_k^P , support count for each candidate k item set to 1.

(3) Use the partition function will generate m nodes r intermediate results C_k^P into different partitions, and then together with their support count is sent to r nodes.

(4) r nodes to count the same item set add up to produce a final practical support, with minimum support count min_sup compared to determine the set of local frequent k item sets L_k^P .

(5)The combined output r nodes that generate global frequent item sets k collection L_k

Advantage of the improved Apriori algorithm is to find and the process is completely independent, there is no link between them, it is a process cycle. k value increments from one until you find all the frequent item sets. You can also set the value of k, you only need to find L_k , or between all frequent item sets from L_1 and L_k

.

3.2.2. The Second Improvement

Despite these improvements algorithm efficiency is relatively high, but still need to repeatedly scan the transaction database , generating a set of frequent item sets to frequent k times need to scan the transaction database , spent a lot of time. So once again improved on Apriori algorithm, the improved algorithm only needs to scan the transaction database again, you can generate the entire frequent item sets, is still used in the cloud computing platform. Concrete improvement ideas are as follows:

(1) Evenly into the size of data corresponding to a subset of the n level of the transaction database, a subset of the data sent to the node m.

(2) Each node scan its data subset to generate a candidate set of options set (a set of candidate to candidate k item sets) , denoted by C^P , support count for each candidate set to 1.

(3) Use the partition function will generate m nodes r intermediate results C^P into different partitions, and then together with their support count is sent to r nodes.

(4) r nodes to count the same item set add up to produce a final practical support, with minimum support count min_sup compared to determine the set of local frequent item

sets L^p .

(5) The output of the combined r -node set L which produce global frequent item sets .

Advantage of the improved Apriori algorithm is only need to scan it again transactional database to find all the frequent item sets.

4 The Experimental Results and Analysis

4.1. Hadoop Platform

Hadoop architecture point of view from the top is a typical Master / Slave configuration shown in Figure 1, there will be a Master primarily responsible Namenode work and Jobtracker work, Namenode stores metadata system. The main duties of Jobtracker are activated tracking and scheduling of tasks to perform each Slave. Because Jobtracker needs to read the information file blocks, so Jobtracter usually Namenode in the same node. How wills Taiwan Slave, each usually has Datanode Slave functionality and is responsible Tasktracker work. Datanode used for actual data storage. Hadoop subordinate task is called Tasktracker, subordinate tasks directly on the child node for data processing, complete the calculation to migrate Tasktracker store information on the status and completion report to Jobtracker.

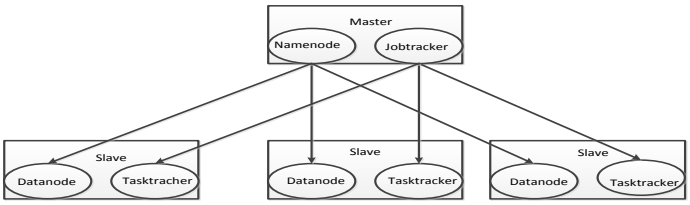

Figure 1. Hadoop Structure

Hadoop cluster supports three operating modes: single mode, pseudo-distributed mode and fully distributed mode.

4.2. Comparison of Two Improved Performance of the Algorithm

In the next experiment, we compare the first and second, an improved algorithm to improve performance, test two algorithms mining frequent item sets the time it takes data from one node increases to 10, the database is T20. I5.N100K.D100K. the minimum transaction support counts to 30. Figure 2 shows with the increase in the number of nodes, the time required to perform two improved algorithms.

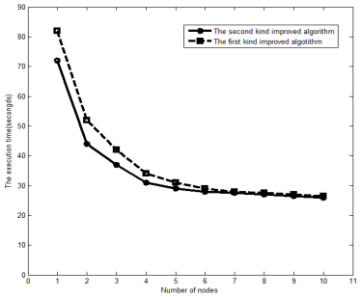

Figure 2. Two Improved Time Performance Comparison Algorithm

As can be seen from the figure, when the node is only one available, the second improved algorithm execution time than the first , an improved algorithm is much less , and the second improved algorithm is very obvious advantages , as the number of nodes

the increase in execution time of two improved algorithms are on the decline. Overall, the improved algorithm with respect to the second section, an improved algorithm has a clear advantage, but this advantage with the increase in the number of nodes is reduced.

4.3. Data Set Allocation Methods Is Applied To an Improved Algorithm First

In order to examine the data set allocation method proposed for association rule mining heterogeneous cluster environments affect performance of the algorithm, the paper also made experiments.

Figure 3 shows an improved algorithm in the first use and not to use the data set allocation method in both cases the execution time. As can be seen from the figure, when only one available node, the data set allocation method almost does not work , because there is only one node does not require data collection division and distribution of the data set ; but with the increase in the number of nodes in heterogeneous clusters environment, the data set allocation method proposed in this paper can significantly improve the efficiency of mining association rules .

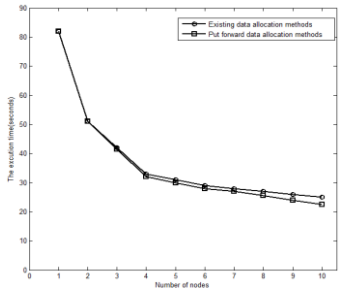

Figure 3. Data Set Allocation Method, an Improved Algorithm for the First

4.4. The Second Data Set Allocation Method Used In the Improved Algorithm

Figure 4 shows a second modification of the algorithm is not used in the use and distribution of the data set in both cases the method execution time. And the last section is similar to the experimental results for the second improved algorithm in heterogeneous cluster environments, the data set allocation method can also be good way to improve the efficiency of association rule mining algorithm, we can see that in a heterogeneous cluster environment, data set allocation algorithm can improve the design of two improved the efficiency of data mining algorithms.

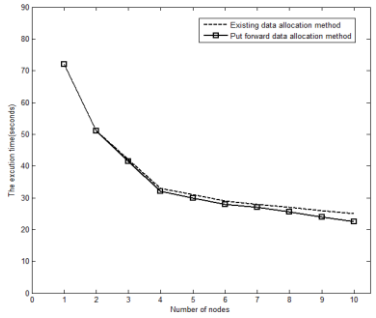

Figure 4. The Second Data Set Allocation Method for the Improved Algorithm

5. Conclusion

In this paper, an algorithm to improve the first, the second modification data set allocation algorithm and conducted experiments with the experimental results of the

performance analysis. First introduced the Hadoop platform, and on this platform , an improved algorithm for the first and second improved performance of the algorithm were compared to investigate the data set allocation method designed in this paper in heterogeneous cluster environment on association rule mining algorithm execution efficiency. Experimental results show that : for execution when two massive data mining algorithms have improved partitioning the data set allocation method and data sets better performance , this design can improve the heterogeneous cluster environments association rule mining algorithm in cloud computing environments efficiency.

Acknowledgements

The work in this paper has been supported by funding from Natural Science Foundation of Guangxi (No.2013GXNSFAA019337), from Key Project of Guangxi Education Department (No.2013ZD055), and from Special Project of Yulin Normal University (No. 2012YJZX04).

References

[1] J. Han and M. Kamber, "Data Mining, Southeast Asia Edition: Concepts and Techniques", Morgan kaufmann, **(2006)**.
[2] R. Jin, G. Yang and G. Agrawal, "Shared memory parallelization of data mining algorithms, Techniques, programming interface, and performance", Knowledge and Data Engineering, IEEE Transactions on, vol. 17, no. 1, **(2005)**, pp. 71-89.
[3] J. Ekanayake and G. Fox, "High performance parallel computing with clouds and cloud technologies", Cloud Computing, Springer Berlin Heidelberg, **(2010)**, pp. 20-38.
[4] J. Dean and S. Ghemawat, "MapReduce, simplified data processing on large clusters", Communications of the ACM, vol. 51, no. 1, **(2008)**, pp. 107-113.
[5] J. Dean and S. Ghemawat, "MapReduce, simplified data processing on large clusters", Communications of the ACM, vol. 51, no. 1, **(2008)**, pp. 107-113.
[6] C. Moretti, K. Steinhaeuser and D. Thain, *et al*., "Scaling up classifiers to cloud computers", Data Mining, 2008, ICDM'08, Eighth IEEE International Conference on, IEEE, **(2008)**, pp. 472-481.
[7] R. Agrawal and J. C. Shafer, "Parallel mining of association rules", IEEE Transactions on knowledge and Data Engineering, vol. 8, no. 6, **(1996)**, pp. 962-969.
[8] R. Perego, S. Orlando and P. Palmerini, "Enhancing the apriori algorithm for frequent set counting", Data Warehousing and Knowledge Discovery. Springer Berlin Heidelberg, **(2001)**, pp. 71-82.
[9] Y. Ye and C. C. Chiang, "A parallel apriori algorithm for frequent itemsets mining", Software Engineering Research, Management and Applications, 2006, Fourth International Conference on, IEEE, **(2006)**, pp. 87-94.
[10] W. Yu, X. Wang, F. Wang, *et al*., "The research of improved apriori algorithm for mining association rules", Communication Technology, 2008, ICCT 2008, 11th IEEE International Conference on. IEEE, **(2008)**, pp. 513-516.

Authors

Wenwu Lian, received his B.Sc. degree in Mathematics and Applied Mathematics from Yulin Normal University and M.Sc. degree in Operations Research and Control theory from Dalian University of Technology, China in 2004 and 2009 respectively. His current research interests on Cloud computing, and Data mining Mass data processing

Lingling Fu, received the B.Sc. degree in Mathematics and Applied Mathematics from Yulin Normal University and the M.Acc degree in Accounting Principles from Southwestern University of Finance and Economics, China in 2004 and 2009 respectively. His current research interests on Data mining and financial analysis.

Jie Zhang, received his B. Eng. and M. Eng. degree in System Engineering and management science and engineering from Xi'an University of Architecture and Technology, China in 1998 and 2004. He received the Ph.D. degree in computer science and technology from Xidian University, China in 2013. She is currently an associate professor at Yulin Normal University, Yulin, China. His current main research interests include data mining, evolutionary computation, et al.

Shangfang Li, received her M.Eng in Computer Application and Technique Research from Guangxi Normal University. Her current research interests on Optimization and Computation Technology.