Research on the cloud platform resource management technology for surveillance video analysis

Yonglong Zhuang[1]*, Xiaolan Weng[2], Xianghe Wei[2]

[1]*Modern Educational Technology Center, Huaiyin Normal University, Jiangsu Huai'an 223300, China*

[2]*College of Computer Science and Technology, Huaiyin Normal University, Jiangsu Huai'an 223300, China*

Received 1 October 2014, www.cmnt.lv

Abstract

As the cloud computing provides the characteristics of the computing resources which can be randomly used or rented, it has become a topic issue to develop a method for dynamically adjusting the computing resources in terms of the service. The paper introduces a method for dynamically adjusting the computing resources on the basis of the surveillance video analysis result. The contents of the surveillance video in the method are cut into many segments, distributed into many computing nodes and to be analyzed. In the meantime, the estimated method for the workload can be sensed through the introduced contents and the system workload can be predicted. The numbers of the computing nodes can be dynamically adjusted without affecting the quality of the service, and the utilized computing resources can be reduced to the minimum as soon as possible. The experimental result shows that the method introduced by the paper can effectively predict the system workload, and can show its advantages in the computing cost and work completion under the situation of not affecting the service quality.

Keywords: cloud computing, video processing, resource management

1 Introduction

As to establish a surveillance video streaming platform, the cloud computing is a very potential technology. Users can install the IPCAM in the home or office, and then the surveillance video contents with the streaming method and network can be stored in the cloud storage system. The video analysis is often expressed as the value-added form in the cloud video. The surveillance video contents can be initiatively analyzed through the strong computing capacity of the cloud platform. When the suspicious events happen, the warming will be noticed to the users real-time. However, the suspicious events in the surveillance video contents appear occasionally, the computing time needed in the video analysis models can also be changed. Therefore, the computing resources needed in the video analysis models changes with the changing of the frequency and numbers of the suspicious events appearance. In addition, the cloud computing provides the characteristics of the computing resources which can be randomly used or rented, it has become an important issue to how to effectively adjust and use the cloud resources to save the operating cost in terms of the actual computing. The enterprises establish the private cloud, and the consumption of the power can be saved by concentrating the virtual machine on the seldom physical servers and closing the useless physical servers.

The contents analysis system of the real-time surveillance video should be established in the cloud platform. One of the methods is to fixedly assign the surveillance video streaming to the specific computing nodes. The disadvantages of the method are that the computing resources can not be dynamically adjusted n terms of the computing change needed in the surveillance video contents analysis. The contents of the surveillance video include more suspicious events or suspicious objects, the needed analyzed time is increasing. If many surveillance video streaming processed in a certain computing node appear large numbers of the suspicious events or suspicious objects, the computing numbers of the computing nodes can be increased so that the computing node is unable to be loaded.

The surveillance video streaming introduced in the [1] is cut into many segments for analyzing the changing of the computing number needed in the surveillance video contents. The complexity of these segments and the workload of each computing nodes decide which computing nodes can analyze these segments. As the work scheduling problem of many computing nodes is a NP-hard problem, these enlightened algorithms are introduced to solve the problem. When a work which has a high computing demand is needed to be handled, the system would cut it into many small work and distribute it into many computing node in the [2]. These computing nodes would be computed synchronously for completing the work rapidly. In the [3], aiming to the large numbers of the video format conversion system, the distributed computing structure based on the Map-Reduce is introduced. It assumes that different computing nodes have different computing capacity, and the complex work is distributed

* ***Corresponding author's*** e-mail: yonglongzhuang@yeah.net

to the computing nodes which has high computing capacity. Although the system has fixed numbers of the computing nodes in the above two researches, the characters of dynamically increasing or reducing the computing nodes in the cloud platform is not considered.

In order to effectively use the computing resources in the cloud platform, several researches have taken the important character into the consideration, that is, the system can dynamically decide whether the numbers of the computing nodes can be increased and reduced in terms of the computing request or workload. In the [4], when the system finds that the work in the queue cannot be finished in the specified time, the whole computing capacity in the system can be increased to ensure the work can be finished in the specified time by increasing the numbers of the computing nodes. In the [5], the method based on the policy is introduced to decide whether the computing nodes are increased or reduced. In the meantime, the method can apply to the different system workload. In the [6], users take the efficiencies request and budget in to consideration, dynamically increasing or reducing the computing nodes, and appointing the work to the Cost-Efficient computing nodes. The system can male the work be finished in the specified time under the situation of the optimal cost. If the handled time needed in each work is the known condition in the above researches, namely, when the users are required to submit a work to the system, the computing time of finishing the work should be provided. In many conditions, it is difficult for users to subjectively and accurately estimate the workload. Therefore, the assumption is not practical in the actual application.

The method for adjusting the cloud computing resources based on the surveillance video analysis result is proposed in the paper. It is unnecessary for the users to notice the needed computing numbers during the period of submitting the work, instead of adding a workload estimation models in the system. According to the submitted video contents, the work should be simply analyzed and the information for estimating the work should be extracted. Taking the estimation results of the work computing numbers and the unfinished work numbers into the consideration, the numbers of the computing nodes can be dynamically adjusted, the needed computing resources should be reduced to the minimums and the operating cost of the system can be reduced under the condition of not influencing the service quality.

2 System structure

Aiming to the cloud video surveillance system, the structure of dynamically adjusting the computing resources based on the video analysis result is introduced. The system can be used to deal with large numbers of the surveillance video streaming real time. Each video streaming entering the system can be real time into many small segments. The video contents of these segments should be initially analyzed through the Pre-analysis Module, the Metadata and the original video analysis should be packaged into many video analysis tasks, and be put into the whole domain work queue. Many computing nodes in the video analysis are derived from the whole domain work queue, and the specified video analysis computing should be conducted in advance, such as the detection of the suspicious objects or the identification of the license number. If the suspicious objects or the suspicious license number are found, the video analysis computing nodes would give the warming information to the related workers through the SMS or email, or the information related to the event would be wrote in the database for subsequent being processed. After the video analysis computing nodes finish a work, the whole work queue would capture the next work and repeat the above motions.

An estimation models for the workload are developed. The total workload of the whole system is estimated in terms of the analysis result of the Pre-analysis Module. The basic concept is that the estimation is the total workload needed in all video analysis work in the whole domain work queue. The limited time of finishing each work and the present computing node numbers in the system should be considered to decide whether the users can finish all work in the specified time. If the estimated results find that the present computing nodes can not make all video analysis work be finished in the specified time, the system should immediately increase the computing nodes numbers for improving the whole computing capacity and speeding up the finished time. Otherwise, if the estimated results find that the system has many computing nodes, parts of the computing nodes can be released to reduce the unnecessary resource waste and reduce the operating cost.

Compared with the methods proposed in the [4-6], the difference of the method proposed in the paper is that the system does not require the users to offer the computing information needed in finishing the work in advance before submit the work. The numbers of the video steams and is very large and the contents change at any time, the users can not estimate the computing numbers needed in the video analysis in advance. Therefore, the proposed method can improve the actual application feasibility and can be applied in the large video streaming analysis system based on the cloud computing.

The video analysis algorithm includes many phases, and the information produced in some specified phases can be used as the parameters of the workload estimation in the video analysis. For example, the license plate detection proposed in the [7] includes two phases: the first phase is License Plate Detection, and the second phase is Character Recognition. The main work in the License Plate Detection is to find out the license plate objects. The demands of the computing numbers are not big and the computing number can be finished in the short time for the objects of the license plate are some square objects.

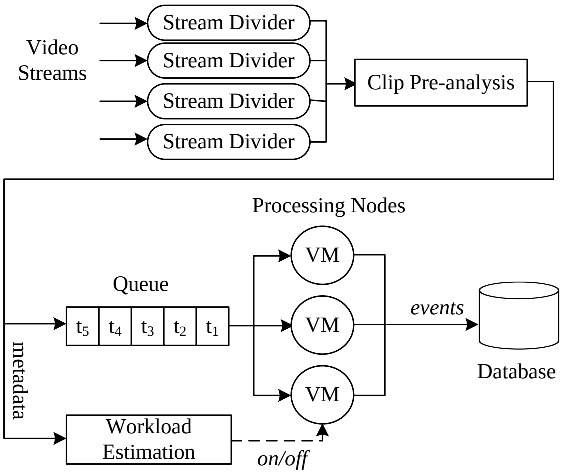


FIGURE 1 System structure

If the objects of the license plate is not found in the first phase, the phase can be directly ignored.

When the Character Recognition begins, the procedure of the whole license plate detection ends. If a license object or many license objects are found in the first phase, the subsequent Character Recognition needs more time to conduct the Character Recognition computing for the possible license objects. Some objects are actually the square objects in the picture, and are still thought as the license plate. Therefore, the objects would be transmitted to the second phase and be recognized. If more objects are recognized, the computing numbers and the computing time are also longer. A work may include the numbers of the license plate objects which can become the important indicator of the work computing complexity, and can be as the estimating principle of the Character Recognition phase.

The license plate detection is used as an example to do the simulation in the paper. The proposed system structure can be easily applied to the video analysis system based on the object monitoring.

3 Estimation technology of the workload

The chapter estimates how to dynamically adjust the numbers of the computing nodes without influencing the service quality, where qi is the Waiting Time in the *i* work, *ei* is the Handling Time in the i work, *i=qi+ei* is the Turnaround Time in the i work. If the expected Turnaround Time is the Deadline, the object of the system is the computing node number k in the minimum operation. In the meantime, i should be ensured, and all work should be ensured to be handled completely in the expected Turnaround Time.

In order to reach the object, the work estimation models simulate all situations whose work is distributed and conducted in the whole domain work queue (as shown in the Figure 2), and then the Maximum Turnaround Time needed in all work is found (as shown in the Figure 3). If the found Maximum Turnaround Time is higher than the specified Threshold H, the system needs a higher computing capacity, that is, need more computing nodes. Therefore, the work estimation models would increase the numbers of the computing nodes, and then repeat the above motions. Conversely, if the found Maximum Turnaround Time is lower than the specified Threshold L, the numbers of the computing nodes would be reduced.

The appearance of the Threshold H and the Threshold L is to avoid the high frequency of increasing or reducing the computing nodes. There is a certain cost to establish the computing nodes or releasing the computing nodes in the cloud computing platform. If the virtual machine is used the computing nodes, the operating state and the off state of the virtual machine needs to take 1 to 2 minutes. Although the Live Virtual Machine Cloning in the [8,9] has reduced the operating time to Hundreds of milliseconds, increasing or reducing the virtual machine too frequently still needs expensive price.

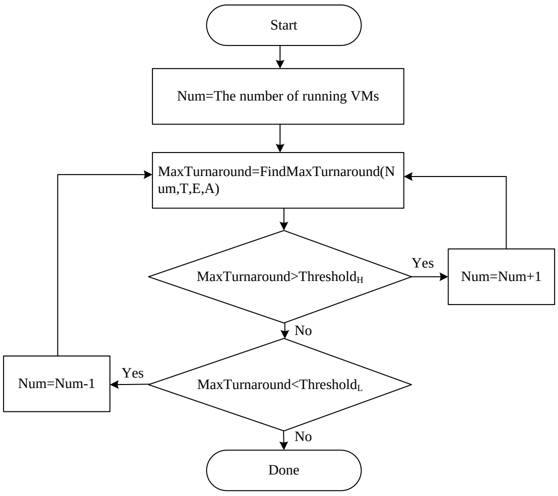

Start
Num=The number of running VMs
MaxTurnaround=FindMaxTurnaround(Num,T,E,A)
MaxTurnaround>Threshold_H
Yes
Num=Num+1
No
MaxTurnaround<Threshold_L
Yes
Num=Num-1
No
Done

FIGURE 2 Procedure figure of the work estimation

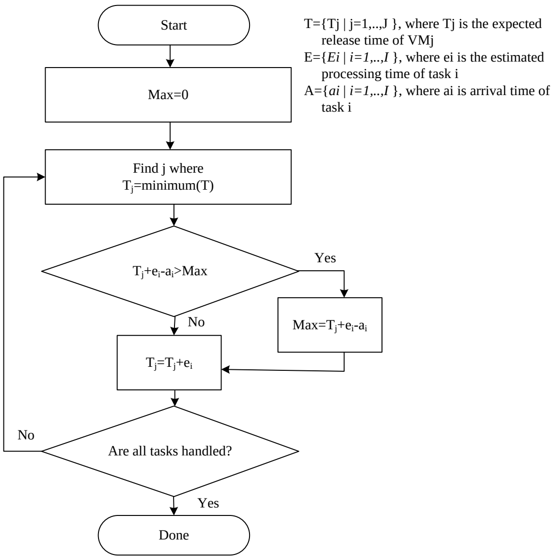

Start
Max=0
T={Tj | j=1,..,J }, where Tj is the expected release time of VMj
E={Ei | i=1,..,I }, where ei is the estimated processing time of task i
A={ai | i=1,..,I }, where ai is arrival time of task i
Find j where Tj=minimum(T)
Tj+ei-ai>Max
Yes
Max=Tj+ei-ai
No
Tj=Tj+ei
Are all tasks handled?
No
Yes
Done

FIGURE 3 The procedure figure of finding out the Maximum Turnaround Time

4 Efficiency estimation

The license plate recognition is used to do the efficiency estimation in the chapter. The license plate recognition algorithm in the [10,11] includes two phases: License Plate Detection and Character Recognition. The following is to show the implementation procedures of the two phases.

License Plate Detection: the main object is to find out the possible license objects in the picture. The so-called possible license objects are some similar square objects. The demands of the computing numbers are not big in the license plate detection and it can be finished in the short time. The candidate area of the license plate has higher density edge information, especially the edge texture in the vertical direction. The paper adopts the Gradient in the vertical direction to do the pretreatment, and the pre-set Sliding Windows is used to scam the whole image. In the meantime, the edge density, the length-width ratio and the minimum length-width of the license sample is used to check whether the scanning area is the license candidate area in the image. The Gradient image also adopts the Integral Image method to accumulate each gradient numbers for speeding up the whole computing procedure.

Character Recognition: if the license detection finds the suspicious license objects, the character recognition phase can be entered. The character cut can adopt the X-Y cut [7] method. First of all, the license candidate area should be scanned from the X direction, and the highest and the lowest horizontal direction of the character area should be assigned. The 1/6 height of the license candidate area is the searching area and the Local Peak of the vertical gray-gradient projection is found out. The principle is to choose the length-with ratio in accordance with the character area in terms of the two peaks, then the area can be merged and the dash can be deleted. After the license character is obtained, the size of the image in each character is resized to 32x16, and the features of the four areas: Zone, Cross, Histogram, Profile is extracted. The character recognition engine adopts to the Support Vector Machine. The initial recognition results must conform to the filtering conditions of the license character stream method (2-3, 2-4 and 4-2 conditions). In addition, in order to improve the accuracy of recognizing character, the Voting method is adopted to give the grades in terms of different images and the recognized character streams whose positions are very close.

The multi fixed positions and the cameras which take photos in the close distance are adopted in the experiment. The video streaming contents are used as the origin of the video analysis. Each image in the streams is the close distance image which takes the vehicles photos (as shown in the Figure 4). In order to make the frequency of the vehicles have the increasing and reducing characters, the video contents include the on and off hours of the company, that is, from 8:00 am to 9:00 am and from 5:00 pm to 6:00 pm. The frequency of appearing the vehicles would be evidently increases in the period. All video contents would adopt the H.264/AVC code, and the resolution ratio is 720x480, the updating frequency of the image is 30 pieces/second. In the meantime, if a computing node does the video analysis from the start-up stage to the operating stage, the spending time is 400 milliseconds.(as shown in the [12,13])

5 Experimental results

The proposed method is compared with the following two algorithms. The first algorithm is the Fixed VM Quantity, FVQ. The numbers of the computing nodes are the pre-decided fixed values. The system cannot dynamically adjust the computing resources with the increasing or reducing of the whole workload. As to the general actual application, the numbers of the computing nodes can be decided by the budget. If the numbers are decided, the numbers of the computing nodes cannot be changed unless there are some special reasons. The advantage of the method is that the result is easily realized, and the disadvantage of the method is that the computing resources cannot be dynamically adjusted in terms of the changing of the system workload. Therefore, the advantages of the cloud computing cannot be really had a role play.

The second method is on the basis of Queue-Size-based Mechanism, QS. The system would increase the numbers of the computing nodes with the increasing the handled work numbers waited in the whole work queue. Otherwise, when the handled work numbers waited in the whole work queue are reduced, some computing nodes are closed. The advantage of the method is that it has the concept of dynamically adjusting computing resources in a certain degree. The method is just according to the work numbers waited in the whole work queue other than the workload (the work numbers are not equal to the workload). Therefore, the method may misestimate the system workload in some situations. For example, the work numbers are small and each video segment includes large suspicious objects, the system workload may be larger than the work numbers and each segment may not include or just include small system workload. When the system includes the same handled work numbers, the needed workload may be different and need to estimate the demanded workload in each work so that the actual system workload can be obtained.

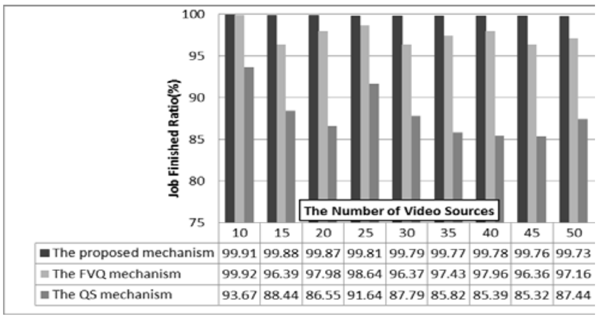
The following efficient estimation project is adopted to compare the proposed method with the above two algorithms.

1) Job Finished Ratio: the job finished ratio before the deadline.

2) VM Cost: the start-up computing node numbers in each time multiply the total simulated time. The finished deadline of the work is set as 3 seconds. When the work reaches to the system, it must be finished in the 3 seconds, including the waiting time in the work queue ad the operating time in the computing nodes.

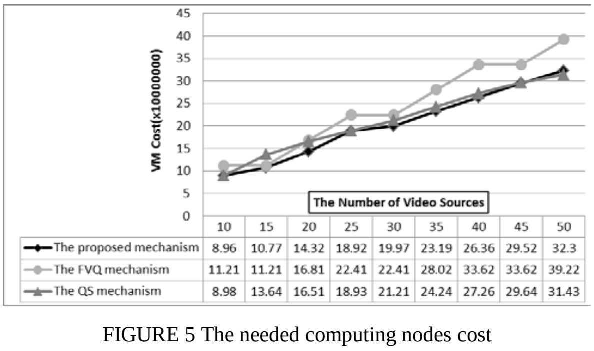
First of all, the job finished ratio in the different methods is estimated, as shown in the Figure 5. The proposed method has higher job finished ratio compared with other two methods. 99.81% job in the proposed method can be finished in the deadline, and 97.58% job in the FVQ method can be finished in the deadline, while the efficiency in the QS method is the worst and there is 85.32% to 93.67% job to be finished in the deadline. The experiment shows that the proposed method can successfully predict the actual workload in the system and the appropriate numbers of the computing nodes can be adjusted to reach the higher job finished ratio.

Then the computing nodes cost needed in the different methods are estimated, as shown in the Figure 6. No matter how much the numbers of the video sources are, the proposed method has the low computing nodes cost compared with the FVQ method. Take the 50 videos as the example, the proposed method can reduce 22% computing nodes cost compared with the FVQ method. On the other hand, when the numbers of the video sources are 15 and 20, the proposed method can reduce 21% and 13% computing nodes cost compared with the QS method. Although the proposed method and the QS method have the approaching cost demands in other experiments, as shown in the Figure 5, the QS method cannot ensure the certain service quality. Therefore, as to the surveillance analysis system which has the real time detected suspicious events, the low job finished ratio in the QS method cannot be accepted.



	10	15	20	25	30	35	40	45	50
■ The proposed mechanism	99.91	99.88	99.87	99.81	99.79	99.77	99.78	99.76	99.73
■ The FVQ mechanism	99.92	96.39	97.98	98.64	96.37	97.43	97.96	96.36	97.16
■ The QS mechanism	93.67	88.44	86.55	91.64	87.79	85.82	85.39	85.32	87.44

FIGURE 4 The comparison of the job finished ratio



	10	15	20	25	30	35	40	45	50
The proposed mechanism	8.96	10.77	14.32	18.92	19.97	23.19	26.36	29.52	32.3
The FVQ mechanism	11.21	11.21	16.81	22.41	22.41	28.02	33.62	33.62	39.22
The QS mechanism	8.98	13.64	16.51	18.93	21.21	24.24	27.26	29.64	31.43

FIGURE 5 The needed computing nodes cost

On the whole, the experimental results show that the cloud computing resources method can be dynamically adjusted based on the surveillance video image analysis results. The workload needed in the whole system can be accurately estimated and the computing resources in the whole system can be appropriately adjusted without sacrificing the system service quality. The utilized workload estimation models takes the suspicious objects numbers included in the video segments into consideration, the proposed method can be applied in the video analysis system based on the large numbers of the tracking objects, especially the video analysis engine which has some higher computing numbers, such as the dynamic video condensing system in the [14,15].

6 Conclusions

The paper introduces a method for dynamically adjusting the cloud computing resources based on the surveillance video image analysis results. The demanded computing numbers are estimated by analyzing video segments and the computing resources in the cloud platform are dynamically adjusted so that the operating cost can be saved. Take the license plate recognition as an example, the numbers of the possible licenses objects produced in the license plate detection phase can be used as an important index for estimating the workload in the character recognition phase so that the indicator can accurately estimate the total computing resources needed in the system. The experiment shows that the method proposed in the paper not only has a higher job finished ratio, but also has lower operating cost. Take the 50 video source as an example, the proposed method can reduce 22% computing nodes cost compared with the FVQ method. When the numbers of the video sources are 15 and 20, the proposed method can reduce 21% and 13% computing nodes cost compared with the QS method.

However, the research just includes the video analysis system based on the tracking objects, whether other types of the video analysis system can be designed by using the same or similar concepts is needed a further research.

Acknowledgements

This study was supported in part by the Natural Science Research Project of Colleges and Universities in Jiangsu Province (No.11KJB520002), the Archives of science and technology project in Jiangsu province (No.2013-16), the Science and Technology Support (industrial) project in Huaian city (No.HAG2012046).

References

[1] Saini M, Wang X, Atrey P K, Kankanhalli M 2014 Adaptive Workload Equalization in Multi-Camera Surveillance Systems *IEEE Transactions on Multimedia* **14** (3) 555-62

[2] Lin S, Zhang X, Yu Q, Qi H, Ma S 2013 Parallelizing Video Transcoding With Load Balancing On Cloud Computing *Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS)* 2864-7

[3] Lao F, Zhang X, Guo Z 2012 Parallelizing Video Transcoding Using Map-Reduce-Based Cloud Computing *Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS)* 2905-8

[4] Marshall P, Keahey K, Freeman T 2009 Elastic Site Using Clouds to Elastically Extend Site Resources *Proceedings of IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing* 43-52

[5] Assuncao M D, Costanzo A D, Buyya R 2009 Evaluating the Cost-benefit of Using Cloud Computing to Extend the Capacity of Clusters *Proceedings of ACM international symposium on High Performance Distributed Computing* 141-50

[6] Mao M, Humphrey M 2011 Auto-Scaling to Minimize Cost and Meet Application Deadlines in Cloud Workflows *Proceedings of International Conference for High Performance Computing, Networking, Storage and Analysis (SC)* 1-12

[7] Kovantsov A, Krumbergs R 2012 Creation Of Graphs Of Functions With Use Of Theorems Of Elementary Geometry *Computer Modelling And New Technologies* **16**(2) 50-9

[8] Zyryn S, Tolkach I 2011 Dependence Of Retention Properties Of Memory Transistors On The Temperature Of Injected Electrons *Computer Modelling And New Technologies* **15**(3)

[9] Chang C C, Lin C J 2011 LIBSVM: a library for support vector machines *ACM Transactions on Intelligent Systems and Technology* **2**(3) 27:1-27:27

[10] Sugier J 2013 Computational Methods For Adaptation Of Markov Models To Requested Maintenance Policies *Computer Modelling And New Technologies* **17** (1)

[11] Jia H 2014 Research and application of cloud computing based on supply chain information system *Energy Education Science and Technology Part A: Energy Science and Research* **32**(6) 4681-90

[12] Jia H, Cao K 2014 Study on the calculation transition from parallel computing to cloud computing on operating system *Energy Education Science and Technology Part A: Energy Science and Research* **32**(5) 4353-60

[13] Shi H, Zhang S, Bai X 2014 Study of cloud storage mechanism based on improved dynamic programming method *Energy Education Science and Technology Part A: Energy Science and Research* **32**(5) 4367-72

[14] Jia H 2014 Study on the parallel web mining system based on cloud platform *Energy Education Science and Technology Part A: Energy Science and Research* **32**(4) 2165-72

[15] Wang X, Jia H 2014 Study on the simulation modeling of message-passing parallel applications based on cloud computing *Energy Education Science and Technology Part A: Energy Science and Research* **32**(4) 2225-32

Authors

	Yonglong Zhuang, May 23, 1978, in Huaian China. **Current position, grades**: Engineer in Modern Educational Technology Center, Huaiyin Normal University, China. **University studies**: Master's Degree in Computer Application Technology in Engineering Corps Command College, China, in 2000 **Scientific interest**: cloud computing, network security, data mining.
	Xiaolan Weng, November 11, 1977, in Taixing China. **Current position, grades**: Associate professor in College of Computer Science and Technology, Huaiyin Normal University, China. **University studies**: PhD candidate in Hohal University, China, in 2014. **Scientific interest**: distributed database, data mining.
	Xianghe Wei, April 18, 965, in Huaian China. **Current position, grades**: Associate professor in College of Computer Science and Technology, Huaiyin Normal University, China. **University studies**: PhD candidate in Nanjing University of Science and Technology, China, in 2010. **Scientific interest**: information security.