Simulation of big data balanced scheduling model in cloud computing environment

Zhou Jinyi
Sichuan Information Technology College, Sichuan guangyuan,628040

Keywords: cloud computing; big data; balanced scheduling

Abstract. in the process of the research on the modeling method of big data balanced scheduling, using the current algorithm to establish a big data balanced scheduling model, the data scheduling is easy to fall into local optimal solution, and there is a problem of big modeling error. To this end, a big data balanced scheduling modeling method based on cloud computing environment is proposed. In this way, the problem of big data balanced task scheduling in cloud computing environment is made formalized description. Through the formal derivation of the dynamic programming method, the heuristic priority allocation strategy of the earliest finish time is obtained. Based on this, using the improved genetic algorithm, the convergence rate of the optimal solution for large data equilibrium scheduling is accelerated. At the same time, with the dynamic heterogeneity of cloud computing environment, the fitness function is made optimization, and the search space of big data balanced scheduling is extended. Based on the optimal solution of the big data balanced scheduling, a big data balanced scheduling model is established. The simulation results show that the big data balanced modeling method based on cloud computing environment can effectively improve the efficiency of big data balanced scheduling and with strong robustness.

1 Introduction

With the rapid development of information technology, data management has become the core of many areas, playing an important role [1-3]. Using the reasonable scheduling method, the target data can be extracted from the massive data with the very low cost, and to improve the computing power and storage capacity [4-6]. Therefore, how to build the accurate model of big data balanced scheduling is the main problem to be solved in this field [7].

At present, the main stream of modeling method for big data balanced scheduling are based on neural network algorithm, greedy algorithm and particle algorithm [8-10]. Among them, the commonly used is the greedy algorithm. However, the algorithm is easy to fall into local optimal solution, and the modeling error is large. A modeling method for big data balanced scheduling in cloud computing environment can dynamically adjust the load of each data node, and avoid the conflict of data.

In view of the above problems, a big data balanced scheduling modeling method based on cloud computing environment is proposed. In this way, the problem of big data balanced task scheduling in cloud computing environment is made formalized description. Through the formal derivation of the dynamic programming method, the heuristic priority allocation strategy of the earliest finish time is obtained. Based on this, using the improved genetic algorithm, the convergence rate of the optimal solution for large data equilibrium scheduling is accelerated. At the same time, with the dynamic heterogeneity of cloud computing environment, the fitness function is made optimization, and the search space of big data balanced scheduling is extended. Based on the optimal solution of the big data balanced scheduling, a big data balanced scheduling model is established. The simulation results show that the big data balanced modeling method based on cloud computing environment can effectively improve the efficiency of big data balanced scheduling and with strong robustness.

2 Modeling principle of big data balanced scheduling

In the process of establishing the model of big data balanced scheduling, the minimum completion time of each task in the corresponding data resource is first calculated. Then, the minimum and maximum values from the minimum completion time are selected to consist to the task pairs. At this time, if the task pairs - resource combination is the optimal which is relative to other combinations, and then to complete the allocation; otherwise, the task will be assigned to other data resources. Furthermore, if there are many methods of distribution can make the current results optimal, then the task pairs are allocated to the resources with the minimum number of tasks. According to the above scheduling strategy, a big data balanced scheduling model is established. Specific steps are as follows:

In the process of building big data balanced scheduling model, using the following formula calculate the efficiency of big data scheduling tasks:

$$\varphi = \left|\theta - j\right| \cdot \sqrt{\left(\theta_1 - \sigma_1 j_1\right)^2 + \left(\theta_2 - \sigma_2 j_2\right)^2 + \mathrm{K}\ \left(\theta_q\right)} \qquad (1)$$

In the above formula, φ is the efficiency of the scheduling task, the smaller the relevance of θ and j is, the higher the efficiency of the scheduling task is.

In the process of constructing a big data balanced scheduling model, the following formula can be used to build massive data scheduling model in cloud computing environment:

$$\min \varphi = \sqrt{\sum_{i=1}^{q} \left(\theta_i - \sigma_i j_i\right)^2} \qquad (2)$$

3 modeling optimization principle of big data balanced scheduling

3.1 the heuristic priority allocation strategy of the earliest completion time

Assuming that the user submits the job to a set of n big data tasks, defining the task set $T = \left\{t_1, \mathrm{K}\ , t_i\ \mathrm{K}\ t_n\right\}$. For a task allocation scheme X, the scheduling load is VT_j, the expected completion time of all big data tasks assigned to the j-th virtual machine vm_j can be represented by the following formula.

$$VT_j = \sum_{j=1}^{n} x_{ij} \times c_{ij} \qquad (3)$$

In the modeling optimization process of the big data balanced scheduling, it is assumed that the task set of $T = \left\{t_1, \mathrm{K}\ , t_i\ \mathrm{K}\ t_n\right\}$ is assigned to the $VM = \left\{vm_1 \mathrm{K}\ vm_j \mathrm{K}\ vm_m\right\}$, the problem of representing m virtual machines is defined as to find the assignment scheme X, then the bellow formula is used to make that the task of the virtual machine in the allocation scheme is the earlist:

$$TS\left(n, m\right) = Min\left(\overset{m}{Max}\left(VT_J\right)\right) \times LB_X \qquad (4)$$

In the above formula, LB_X represents the shortest task and with the smallest load balancing degree.

In the process of modeling optimization of the big data balanced scheduling, for the k scheduling problems of tasks, assuming that the k tasks t_k are assigned to the z-th virtual machine vm_2, which means that the time span of the z-th virtual machine vm_2 is:

$$makespan_{kz} = vt_{(k-1)} + c_{ks} \qquad (5)$$

In the process of optimizing the big data balanced scheduling, in order to meet the recursive relation of big data equilibrium task scheduling in the cloud computing environment, the formula (4) is arranged as:

$$TS(k,m)=\min_{z=1}^{m}\left(TS(k-1,m),makespan_{ks}\right) \quad (6)$$

In the process of optimizing the big data balanced scheduling, we can know that the strategy of which k tasks t_k will be assigned to the virtual machine vm_z with the earliest completion time can be defined as the heuristic priority allocation strategy in the cloud computing environment,

3.2 Implementation of big data balanced scheduling modeling and optimization principle

In the process of the big data balanced scheduling optimization, the genetic algorithm is used to make pre-encoding for the chromosome. Assuming that it has m tasks, the task $ID=\{1,2,3\text{K}\}$, the resource $ID=\{1,2,3\text{K}\ .n\}$, then the total number of the length of the gene cluster of chromosome is m. The chromosome represents that the first task is assigned to the second resources, the second task is assigned to the third resources, the fourth, fifth, and sixth tasks are assigned to the sixth resources. For a chromosome task resource allocation scheme, the optimal span is the total task execution time for the resource nodes of the latest completion time in the assignment scheme. According to the dynamic heterogeneity of the cloud computing environment, the task processing time is mainly determined by the floating point operation ability of the resource nodes, the time of task mapping and the result gathering are mainly determined by the bandwidth between the host and resource nodes. Assuming that the time of a big data resource node k to complete the task i is expressed as follows:

$$T_{k.i}=\frac{t_i^f+t_i^0}{C_{k,i}^b}+\frac{t_i^l}{C_k^p} \quad (7)$$

Where t_i^f represents the input data file size of the task i, t_i^0 represents the output data file size of the corresponding result of the task i, t_i^l represents the length of the task i.

If the number of tasks assigned by the resource node k is AssignTasks, then the total time to complete the task is:

$$T_{k,total}=\sum_{i=1}^{\text{AssignTasks}}T_{k,i} \quad (8)$$

The selection probability of the individuals of the fitness function based on the standard deviation of the task allocation number is expressed by using the following formula:

$$p_2=\frac{f_2(j)}{\sum_{j=1}^{scale}f_2(j)} \quad (9)$$

The standard deviation of the optimal span fitness function value is used as the convergence condition, and the optimal solution can be found faster and better, the following formula is used to represent:

$$sd=\sqrt{\frac{\sum_{j=1}^{scale}\left(f(j)-f_{avg}\right)^2}{scale}}\le\xi \quad (10)$$

In the above formula, $f(j)$ is the fitness value of the j-th individual, and f_{avg} is the average fitness value of the contemporary population, *scale* is the population size, and ξ is the convergence threshold.

4 experiments and simulation

In order to prove the validity of the modeling method for big data balanced scheduling based on cloud computing environment, it needs to carry out an experiment. Simulation is performed in the cloud simulator Cloudsim. CloudSim is a function library developed in the discrete event simulation package SimJava, inheriting the programming model of Cridsim.

The traditional algorithm and the improved algorithm are used respectively for modeling experiment. Under different tasks, the proportionality of big data scheduling two algorithms, scheduling efficiency and stability are compared. The results are shown in Table 1 and table 2.

Table 1 the overall effectiveness of using the traditional algorithm for big data balanced scheduling modeling

Experiment number (times)	Equilibrium of using traditional algorithm for big data balanced scheduling (%)	Efficiency of using traditional algorithm for big data balanced scheduling (%)	Stability of using traditional algorithm for big data balanced scheduling (%)
15	73	69	65
25	73	69	65
35	73	69	65
45	73	69	65
55	73	69	65
65	73	69	65
75	73	69	65
85	73	69	65
95	73	69	65

Table 2 overall effectiveness of using improved algorithm for big data balanced scheduling modeling

Experiment number (times)	Equilibrium of using improved algorithm for big data balanced scheduling (%)	Efficiency of using improved algorithm for big data balanced scheduling (%)	Stability of using improved algorithm for big data balanced scheduling (%)
15	95	96	97
25	95	96	97
35	95	96	97
45	95	96	97
55	95	96	97
65	95	96	97
75	95	96	97
85	95	96	97
95	95	96	97

Table 1 and table 2 show that the overall effectiveness of the established big data balanced scheduling model by using the improved algorithm is better than the traditional algorithm. This is mainly because by using the improved algorithm, the problem of big data balanced task scheduling in cloud computing environment is made formalized description firstly. Through the formal derivation of the dynamic programming method, the heuristic priority allocation strategy of the earliest finish time is obtained. Based on this, using the improved genetic algorithm, the convergence rate of the optimal solution for large data equilibrium scheduling is accelerated. At the same time, with the dynamic heterogeneity of cloud computing environment, the fitness function is made optimization, and the search space of big data balanced scheduling is extended. Based on the optimal solution of the big data balanced scheduling, a big data balanced scheduling model is established, to ensure the efficiency of using improved algorithm to establish big data balanced model.

The above experiment can prove that the big data balanced modeling method based on cloud computing environment can effectively improve the efficiency and the robustness is strong.

5 Conclusions

In view of the use of the current algorithm in the establishment of a big data balanced scheduling model, data scheduling is easy to fall into local optimal solution, there is a big problem of modeling error. A big data balanced scheduling modeling method based on cloud computing environment is proposed. In this way, the problem of big data balanced task scheduling in cloud computing environment is made formalized description. Through the formal derivation of the dynamic programming method, the heuristic priority allocation strategy of the earliest finish time is obtained. Based on this, using the improved genetic algorithm, the convergence rate of the optimal solution for large data equilibrium scheduling is accelerated. At the same time, with the dynamic heterogeneity of cloud computing environment, the fitness function is made optimization, and the search space of big data balanced scheduling is extended. Based on the optimal solution of the big data balanced scheduling, a big data balanced scheduling model is established. The simulation results show that the big data balanced modeling method based on cloud computing environment can effectively improve the efficiency of big data balanced scheduling and with strong robustness.

References

[1] Liu Yajie, Li Zhongmeng, Xie Jun. Method of Carrier-borne Aircrafts Exporting Scheduling Modeling Based on Petri Net [J]. Fire Control & Command Control, 2015 (09): 152-156.

[2] Chen Gonggui, Chen Jinfu. Environmental/Economic Dynamic Dispatch Modeling and Method for Power Systems Integrating Wind Farms [J]. Proceedings of the CSEE, 2013, 33 (10): 27-35.

[3] Jiang Xingwen, Zhou Jianzhong, Wang Hao, et al. Modeling and Solving for Dynamic Economic Emission Dispatch of Power System [J]. Power system technology, 2013, 37 (02): 385-391.

[4] Wang Chengshan, Hong Bowen, Guo Li, et al. Modeling and Solving for Dynamic Economic Emission Dispatch of Power System [J]. Proceedings of the CSEE, 2013, 33 (31): 26-33.

[5] Hu Xiangpei, Sun Lijun, Wang Zheng. Online Intelligent Scheduling Based on Internet of Things (IoT) [J]. Management science, 2015 (2): 137-144.

[6] Wang Xianghua, Chen Te fang. Modeling and implementation of intelligent bus dispatch system based on MAS [J]. Computer engineering and science, 2014, 36 (5): 986-990.

[7] Chen Rong, Liu Peipei. Modeling and analysis of vehicle scheduling system based on Multi-Agent [J]. China electronic commerce, 2013 (15): 37-37.

[8] Tian Feilong, Zhang Shangfeng, Zheng Houlei, et al. Optimization modeling and calculation of armed police troop transport scheduling [J]. Science and technology innovation and application to 2015 (23): 24-25.