Research on Collaborative Acquisition of Multidimensional Massive Web Based on Fusion Credibility

ZHANG Ya-juan[1,a], HE Wen-qing[2]

[1]*Railway Telecommunication Department, Hunan Technical College of railway high-speed, China*
[2]*CNOOC Zhuhai Natural Gas Co., Ltd., China*

Abstract. In the era of rapid development of Internet technology and people's growing social needs, web information collection has been successfully applied to the major search engines and search areas. In this paper, the mass information collection is regarded as the dynamic task allocation problem based on the co-operation of the package. A multi-dimensional computer resource model is proposed, which uses the heuristic algorithm of the mutation to match the heuristic algorithm. Conditional cost objective function is optimized, so that the whole system in the process of dynamic changes, the time and cost are as small as possible. Finally, the experimental results show that the algorithm can meet the different user requirements on the basis of maximizing the total cost of the system.

1 Introduction

In the massive web page information collection task allocation problem, the concern is often how to divide the task, such as: Web division, regional division, domain name division, and then according to the size of the distribution of the corresponding physical nodes, few based on crawling nodes Self-attribute allocation method. Therefore, this paper proposes a multi-dimensional computer resource model which integrates the credibility, and then uses the mutation-matching heuristic algorithm to dynamically allocate the task. By optimizing the cost objective function with multiple constraints solving, making the whole system in the process of dynamic changes, the time and cost are as small as possible.

2 Basic concept

Dynamic task allocation problem is the core research problem of various kinds of complex collaborative systems in engineering project and practical application. During the operation of this kind of system, because of the external environment and some of the internal resources of the internal constraints, in the face of each task project And the unexpected situation, the system requirements will change accordingly, and the time cost and cost overhead will also change, seriously affect the stability of the entire system, the implementation of efficiency and overall cost, in order to solve these problems, the system needs to constantly Task sequence allocation and redistribution operation, this process is the dynamic task allocation, in a sense, this dynamic task allocation problem is an uncertain environment interdependent task allocation of the decision-making problem, also known as multi-objective decision Problem [1,2,3].

Multi-objective decision-making problem: Let the system have m objective functions: $f_1(x), f2(x),\ldots, f_m(x)$ and n decision variables composed of vector: $x=(x_1,x_2,\ldots,x_n)^T$. If these goals require the largest (or minimum), and the solution to meet the constraints of k constraints, the mathematical model Can be expressed as follows:

$$Z=F(X)=\begin{bmatrix}\max(\min) f_1(X)\\ \max(\min) f_2(X)\\ \vdots\\ \max(\min) f_m(X)\end{bmatrix} \quad (1)$$

$$s.t.\quad \phi(X)=\begin{bmatrix}\varphi_1(X)\\ \varphi_2(X)\\ \vdots\\ \varphi_k(X)\end{bmatrix}\le G=\begin{bmatrix}g_1\\ g_2\\ \vdots\\ g_k\end{bmatrix} \quad (2)$$

Equation (1) shows that there are m objective functions that need to be optimized. Equation (2) shows that there are k constraints that need to be satisfied.

2 The proposed technique

2.1 Multi-Dimensional computer resource model with fusion credibility

In the actual project implementation process, the demand is always changing, the number of tasks is dynamic, the resources required for each task is also uncertain, so we

[a] ZHANG Ya-juan: zhangyajuan2017@sina.com

assume that the project demand for the machine m_j is d_i . Then the vector $D=\{\mathrm{d}_1,\mathrm{d}_2,\ ...\ ,\mathrm{d}_\mathrm{n}\}$ indicates the project's demand for all types of machines. Under the influence of the dynamic execution of the project, the value of each element in D is uncertain. In this paper, the Gaussian distribution model is used to simulate the stochastic problem of demand change in the process of actual project implementation [5, 6].

The value of each element in the vector $D=\{\mathrm{d}_1,\mathrm{d}_2,\ ...\ ,\mathrm{d}_\mathrm{n}\}$ is different, assuming that $\mathrm{d}'_1,\mathrm{d}'_2,\ ...\ ,\mathrm{d}'_\mathrm{n}$ represents the value of each element in D, then the set $D'=\mathrm{d}'_1*\mathrm{d}'_2*\ ...\ *\mathrm{d}'_\mathrm{n}$ represents the Cartesian product of the set, then the relation between the set D and the set D' Can be formalized as:

$$\mathrm{D}\in\mathrm{D}' \text{ and } \sum_{D\in D'} p(D)=1 \tag{3}$$

At this point, when the number of reserved machines is $N^b_{j,i}$, the number of machines actually used during the operation of the project is $N^u_{j,i}(D)$, and the number of machines that may be attached is $N^a_{j,i}(D)$ for the specific requirements D mentioned above. Then the relationship between the three can be formalized as:

$$D=N^u_{j,i}+N^a_{j,i}(D),\ N^u_{j,i}(D)\le N^e_{j,i} \tag{4}$$

According to the universal collection fee standard, the total cost of the user needs to be paid in the case where the current demand is D:

$$\begin{aligned} C(D)=&\sum_{t_i\in T}\sum_{m_i\in M} {c_{j,i}}^{[b]}N^{[b]}_{j,i} \\ &+\sum_{D\in D'}\sum_{t_i\in T}\sum_{m_i\in M} P(D)({c_{j,i}}^{[u]}N^{[u]}_{j,i}(D)+{c_{j,i}}^{[a]}N^{[a]}_{j,i}(D)) \end{aligned} \tag{5}$$

In the dynamic task allocation, because the actual operation of the project will inevitably change the task, the machine node without feedback, node crashes and other factors, in order to better maintain the stability of the system to enhance the system's efficiency and task execution rate. We also consider the cost of the time also need to consider the cost of time. Therefore, in this paper, when we optimize the dynamic task allocation model of fusion confidence, we first set a time threshold of the project itself, under which the cost is minimized, which ensures the actual project operation time Overhead, but also to achieve the purpose of minimizing the cost.

The optimization of the dynamic task allocation method for fusion confidence is carried out by using the maximal minimum method to convert the total cost of the target function $C(D)$ and the total time T_{total} to the single objective optimization problem which only optimize the total cost of $C(D)$. To the target function total time T_{total} set a threshold, the original constraints remain unchanged, respectively, the reserve phase, the use of stages and additional stages of the machine price, the number of machines used in each stage, processing time.

Minimize:

$$\begin{aligned} C(D)=&\sum_{t_i\in T}\sum_{m_i\in M} {c_{j,i}}^{[b]}N^{[b]}_{j,i} \\ &+\sum_{D\in D'}\sum_{t_i\in T}\sum_{m_i\in M} P(D)({c_{j,i}}^{[u]}N^{[u]}_{j,i}(D) \\ &+{c_{j,i}}^{[a]}N^{[a]}_{j,i}(D)) \end{aligned} \tag{6}$$

$$s.t.\ \ N^u_{j,i}(D)\le N^b_{j,i}, t_i\in T, \mathrm{m}_i\in M, D\in D' \tag{7}$$

$$\sum_{t_i\in T} n^{cpu}_{j,i}(N^u_{j,i}(D)+N^a_{j,i}(D))\le \max{}^{cpu}_j \tag{8}$$

$$\sum_{t_i\in T} n^{m}_{j,i}(N^u_{j,i}(D)+N^a_{j,i}(D))\le \max{}^{m}_j \tag{9}$$

$$\sum_{t_i\in T} n^{br}_{j,i}(N^u_{j,i}(D)+N^a_{j,i}(D))\le \max{}^{br}_j \tag{10}$$

$$\begin{aligned} &N^b_{j,i}, N^u_{j,i}(D), N^a_{j,i}(D)\in Z, \\ &\mathrm{t}_\mathrm{i}\in T, m_i\in M, D\in D' \end{aligned} \tag{11}$$

$$T_{total}=T(r)+T(a) \tag{12}$$

$$T_{total}\le T_{value} \tag{13}$$

Equation (7) indicates that the number of machines used when demand D cannot be greater than the total number of machines reserved. Equation (8) ~ (10) indicates that when the demand is D, the total amount of resources used in the use phase and the reservation phase cannot be greater than the maximum amount of resources of the machine. Equation (11) indicates that the number of machines used at each stage is a nonnegative integer. Equation (12) ~ (13) indicates that the total processing time cannot be greater than the threshold T_{value} we set.

2.2 A heuristic task allocation algorithm for mutation priority matching

Variation of the priority match Heuristic algorithm is described as follows:

(1) Selecting a node from all the used physical node sequences;

(2) Determine whether the confidence of the physical node Confidence value is greater than the set threshold;

a. Less than, to perform all the tasks of its revocation, back to the queue to be executed in the task sequence, continue to traverse the next has been used in the physical node sequence;

b. Not less than, to determine the physical nodes of multiple dimensions (CPU, memory, network bandwidth, etc.) can meet the requirements of the current task. Meet the requirements, the current task will be assigned to the physical node; do not meet the requirements, back to (1), continue to implement;

(3) If all of the used physical nodes cannot meet the requirements of the current task, then select the first from the list of physical nodes that are not used;

(4) To determine the physical nodes of multiple dimensions (CPU, memory, network bandwidth, etc.) can meet the requirements of the current task;

a. To meet the requirements, the current task will be assigned to the physical node;

b. Does not meet the requirements, then continue to traverse the current unused physical node sequence until you find the physical node to meet the requirements of the current task assigned to the physical node;

(5) Repeat the above steps until all tasks have been assigned.

3 Experiment results

Experiment 1 The number of physical nodes is fixed, and the relationship between the number of user requirements and the total cost is simulated.

Set the number of physical nodes to be 30, the number of physical nodes required for users is 1 to 60, the reservation phase, the use phase, the additional phase of the price were 3 Yuan, 8 Yuan, 15 Yuan. The values of each variable are shown in Table 1.

Table 1. Experimental variable value table.

Variable name	Variable value
Reserve physical nodes	30
User desired physical nodes	1-60
Reserve price	3
Use price	8
Additional price	15

Then, according to the above table the value of each variable and the formula (6) ~ (13) can be obtained, the reserve phase, the use of stages and additional stages of the cost of the situation shown in Table 2 below.

Table 2. The stage of the experiment, the stage of the use, the stage of the additional stage.

	d_1: 1~30	d_2: 30~60
Reserve phase	30*3	30*3
Use phase	d_1*8	30*8
Additional phase	0(No additional)	（d_2-30）*15

The results of Matlab simulation experiment 1 are shown in Fig 1.

From the experimental results in Figure 2 shows that the number of fixed physical nodes in the fixed, with the number of users need to increase the number of physical nodes, the reserve stage of the cost remains unchanged; the use of stage costs first increase in the demand exceeds The cost of the additional phase is always zero when the number of reserved physical nodes is greater than or equal to the user's demand. When the user's demand exceeds the reserve number, the additional cost increases; the total cost is always on the rise.

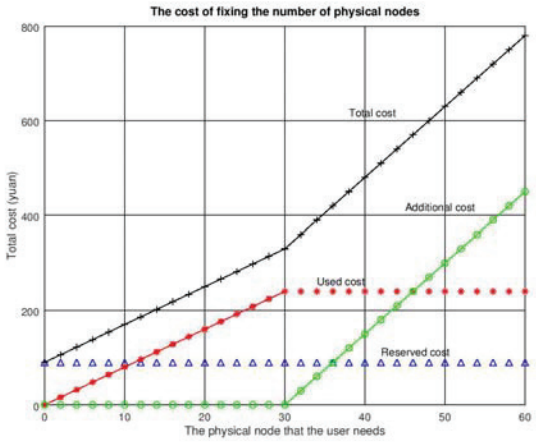

Figure 1. Fixed total cost changes when reserving physical nodes.

Experiment 2 The total number of physical nodes required by the user is the same, and the relationship between the number of physical nodes and the total cost is simulated.

Set the total demand for 50 physical nodes, the number of physical nodes reserved for the 1 to 150, the reserve phase, the use of stage, the additional phase of the price were 3 Yuan, 8 Yuan, and 15 Yuan. The values of each variable are shown in Table 3.

Table 3. Experimental variable value table.

Variable name	Variable value
Reserve physical nodes	1-150
User desired physical nodes	50
Reserve price	3
Use price	8
Additional price	15

Then, according to the value of each variable in Table 3 and the formulas (6) ~ (13) in Section 2.1, the cost of the reserve phase, the use stage and the additional stage are shown in Table 4 below.

Table 4. The stage of the experiment, the stage of the use, the stage of the additional stage.

	d_1: 1~50	d_2: 50~1500
Reserve phase	d_1*3	d_2*3
Use phase	d_1*8	50*8
Additional phase	50*15-(d_1*3+d_1*8)	0(No additional)

The results of Matlab simulation experiment 2 are shown in Fig 2.

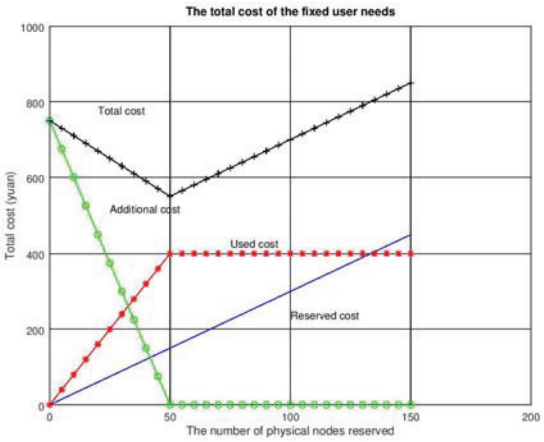

Figure 2. Fixed total cost changes when the user is required.

4 Conclusion

Based on the self-attribute of the computer, the multi-dimensional computer resource model is established. At the same time, in order to consider the timeliness of the physical node in practical application, the concept of credibility is put forward, and this constraint condition Integrated into the computer resource model, can better adapt to the dynamic changes in the system. The simulation results show that the proposed algorithm can meet the needs of users in a limited time range, and the total cost is the smallest.

References

1. Jadidi O, Zolfaghari S, Cavalieri S. International Journal of Production Economics. A new normalized goal programming model for multi-objective problems: A case of supplier selection and order allocation, 148(1):158-165(2014).
2. Choudhary D, Shankar R. Computers & Industrial Engineering. A goal programming model for joint decision making of inventory lot-size, supplier selection and carrier selection, 71(1):1-9(2014).
3. Jadidi O, Cavalieri S, Zolfaghari S. Applied Mathematical Modelling. An Improved Multi-Choice Goal Programming Approach for Supplier Selection Problems, 39(14):4213-4222(2014).
4. Luo J, Lan C E. Journal of Guidance Control & Dynamics. Determination of weighting matrices of a linear quadratic regulator, 18(6):1462-1463(2015).
5. Ejsmont W. Statistics & Probability Letters. A characterization of the normal distribution by the independence of a pair of random vectors, 114:1-5(2016).
6. Qun-ying. Acta Mathematicae Applicatae Sinica. Laws of the Iterated Logarithm for ρ -mixing Random Variables with Normal Distribution, 2:385-394(2016).