Improved Particle Optimization Algorithm Solving Hadoop Task Scheduling Problem

Jun Xu
South China Normal University, College of computer, Guangdong Guangzhou 510631
Guangzhou Radio Group, GRG Banking, ATM Research Institute, Guangdong Guangzhou 510663
e-mail: xujun3447@163.com

Yong Tang
South China Normal University, College of computer, Guangdong Guangzhou 510631
e-mail: ytang@scnu.edu.cn

***Abstract*—Cloud computing to provide service for the user group is huge, so the number of cloud computer's tasks is enormous, the system handle large tasks all the time so that task scheduling is the key and difficult points in the cloud. This article make research on how to make full use of cloud resources for task efficiently scheduling. This paper proposes an Improved Particle Swarm-Estimation of Distribution optimization Algorithm (IPS-EDA) based on task allocation strategy. The task scheduling strategy is optimization strategy based on improved particle swarm algorithm, which introduce estimation of distribution algorithm (EDA) based probabilistic model and random sampling theory, the proposed algorithm does not fall into local optimum. The simulation results show that the performance of IPS-EDA has been greatly improved provides better load balancing and resource utilization.**

Keywords-Task scheduling; Estimation of Distribution; Particle Swarm Optimization; Cloud computing

I. INTRODUCTION

In recent years, cloud computing [1, 2] has become a hot topic of discussion. At present, IBM, Google, Amazon, Microsoft etc. in succession sortie cloud computing, providing a lot of cloud based services. Hadoop framework is designed in distributed computing environment of a MapReduce[3,4] computational model for large-scale data processing, Hadoop framework facilitates the development of distributed computing applications. Hadoop has three important parts, which are respectively HDFS, Map case (Mappers) and Reducer (Reducers) case. Although the overall architecture of Hadoop framework simplify process, three parts hide many complex low layer detail, including the hardware and software. The framework also provides a simple job scheduler FIFO (FIFO), in order to perform the job submission. Sequential scheduler can reduce the working management in a certain extent, and processing job queue is effective. However, some other factors also need to consider job scheduler. As everyone knows, many clusters are highly homogeneous environment running. For example, Hadoop uses an isomorphism cluster system in Yahoo, contains 4000 processor, 3TB of ram and a 1.5PB hard drive storage capacity. Published research results show the strong ability of Hadoop framework. In the distributed computing tasks, people focus on the extreme manifestation that the use of homogeneous environment may be desirable to avoid the load imbalance problem. Normally it is difficult to establish a number of nodes can reach thousands of homogeneous cluster system. The actual situation is the large numbers of heterogeneous nodes exist in most Hadoop clusters. Hadoop framework architecture has been designed to flexibly adapt to the heterogeneous resources. So we can see clearly resource heterogeneity affects the performance of the cluster.

This paper presents optimization strategy of task scheduling in heterogeneous MapReduce environment. The task scheduling strategy is the optimization strategy based on Improved Particle Swarm optimization (PSO) [5,6,7] algorithm, by introducing the estimation of distribution algorithm (EDA) 8 based probabilistic model and random sampling theory, the algorithm does not fall into local optimum. The simulation results show that the performance of simulations clusters has been greatly improved.

II. TASK SCHEDULING OPTIMIZATION PROBLEM OF HADOOP

In general, Hadoop's MapReduce mainly includes two user defined functions: Map functions and Reduce functions. Figure 1 illustrates the basic principle of Map-Reduce operation.

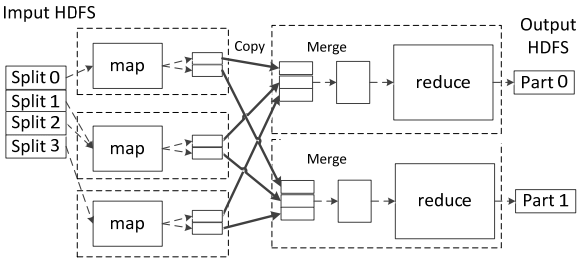

Figure 1. Map/Reduce Implementation Process

Map / Reduce is mainly divided into the following two stages:

(1) Map phase: each input split will make a map task to handle, a map generated data, which will be through a process of data shuffling ,and then partitions are allocated to different reduce tasks.

(2) Reduce phase: Reduce will receive data from different map tasks, each data from the map are ordered, and data will be collected and analyzed, the final results will be outputted.

Task scheduling problem for Map/Reduce is a complex combinatorial optimization problem. For convenience of description, firstly we will do mathematical modeling of Map/Reduce task scheduling. For a certain period of time, hypothesis that N tasks are firstly set up and wait for scheduling, there are M processing node for tasks , task i need computing power t_i , the computing capacity c_j per unit time of j processing node, maximum processing task parallel number $Maxp_j$. How to make n tasks are assigned to M processing nodes, so that total task completion time is the shortest? As the objective function:

$$\min \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{o_{i,j} t_i}{c_j} \tag{1}$$

In the formula, $o_{i,j}$: task i Occupation processing node j computing resources, the symbolic value is 1; otherwise, the value is 0.

Constraints that the processing task number of node j processing task cannot exceed it's maximum processing task parallel number, i.e.

$$s.t. \sum_{i=1}^{n} o_{i,j} \le Maxp_j \tag{2}$$

III. ALGORITHM DESCRIPTION IN THIS PAPER

In this paper, the performance of Hadoop MapReduce scheme as the position of the particle , based on particle swarm algorithm, the paper introduce the estimation of distribution algorithm to prevent falling into local optimal solution, guaranteeing the particles search the optimal solution in the global space, thus improving the performance of task scheduling in Hadoop.

Particle swarm optimization (PSO) algorithm is a stochastic optimization technique based on population, invented in 1995 by Dr. Eberhart [6, 7, 8]. PSO maintain populations of candidate solutions, and let these particles search space to meet the optimization objective function by moving. The particle motion is guided by the found best position to search the search space, the best position of the particle is updated after a better position found.

The system is initialized with a set of random solutions, search and update the optimal solution through the iterative. Each particle track their own position coordinates in the problem space, the particle find the best position called individual optimal value Pbest in each iteration. Another "best" value, the best solution is found in all the particles are tracked by far the optimal particle swarm optimization, this is called the global optimal value Gbest. The particle swarm optimization algorithm is towards the individual optimal values of Pbest and the global optimal values of Gbest to accelerate each particle search. The core formula of velocity and position updates of particle swarm algorithm is as follows:

$$v_i(k+1) = v_i(k) + c_1 \times r_1 \times (p_i(k) - x_i(k)) + c_2 \times r_2 \times (p_g(k) - x_i(k)) \tag{3}$$

$$x_i(k+1) = x_i(k) + v_i(k+1) \tag{4}$$

c_1 , c_2 is learning factor; r_1, r_2 is a random number between (0, 1); $p_i(k)$ is the individual optimal values of particles i; $p_g(k)$ is the global optimal value of particles i;

Estimation of distribution algorithm (EDA) [9] produces a new solution according to the probability distribution model of the evolutionary process quality solution information, which has probability analysis theory. The basic frame of algorithm is follow as:

First step: a number of high quality solutions are selected as the initial population from the random solutions;

Second step: using probability model to estimate the population, and producing new solutions by sampling;

Third step: these new solutions replace those old solutions in the new population;

Fourth step: determine whether the termination condition is satisfied. If meet the termination condition, the solution of new population is the final solution; otherwise, go to step second;

The core operator of EDA is to establish the probability model. In EDA, a probability vector $p(x) = (p(x_1), p(x_2), ..., p(x_n))$ express the solution probability model of spatial distribution, which $p(x_i)$ represents the probability of position i values taking 1. Probability calculation formula of the model is as follow:

$$p_{j+1}(x) = (1-a)\, p_j(x) + a \frac{1}{N} \sum_{j=1}^{N} x_i^j \tag{5}$$

$p_j(x)$ is probability vector of j population solution space, $x_i^1, x_i^2, ..., x_i^N$ are N good solutions, x_i^j is the value of i position, a is study factor.

The core operator of EDA was introduced to Improved particle swarm optimization algorithm (IPS-EDA), has overcome PSO's shortcoming that is easy to fall into local optimal solution. Because the core operator is to establish the probability model, the improved particle swarm algorithm has a theoretical basis of probability analysis. Improved particle swarm algorithm is proposed in this paper (IPS-EDA) , introducing the core operator of EDA to accelerate search for optimal solution , complete task scheduling in the smallest the time and efficient resource rates. The fitness function is a kind of load balancing based on resource capacity allocation of tasks to processors. Here is the entire IPS-EDA algorithm steps:

1. Initialized iteration counter (k = 0), the population size (Psize), the largest number of iterations (Maxgen).

2. Randomly generated the initial population of particles (Psize).

3. Calculate particle fitness function value, and find individual optimal value and global optimal value of the initialization value.

4. Update iteration counter k = k + 1.

5. According to the formula (3) and (4), updating speed and position of the particles.

6. According to the formula (5), establishing the probability model of population, and random sampling to generate new population.

7. According to formula (1), calculating fitness function value of particles, and sorted in ascending order.

8 Update the local optimum (p_i) and global optimum (p_g).

9. Continue steps 4 to 8 until the optimal solution converge.

10. Output the global best particle.

IV. SIMULATION EXPERIMENT AND RESULT ANALYSIS

In the Hadoop framework, using MapReduce programming to realize IPS-EDA algorithm. The experimental results are classified in the following aspects:

1) Performance analysis of IPS-EDA

Improved particle swarm optimization and estimation of distribution algorithm (IPS-EDA) compared with PSO and EDA, comparisons contain optimal resource allocation and time efficiency. Suppose there are 20 tasks and 5 processors. Figure 2 illustrates the faster convergence speed of HPSO-EDA algorithm, and can obtain better fitness value.

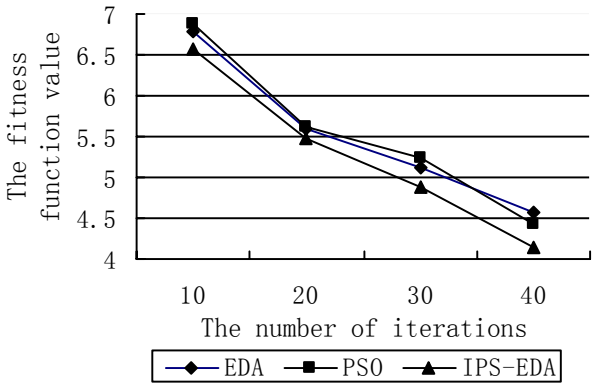

Figure 2. Comparison of algorithm performance

2) IPS-EDA load balancing

Figure 3 illustrates HPSO-EDA load balancing ability to provide better, because the algorithm takes full account of the processing capacity of resources. The figure shows in contrast to maximum processing capacity of the MAXMIPS processor, the utilization rate of resources in most HPSO-EDA algorithms on the processor are higher than that of PSO and EDA algorithm.

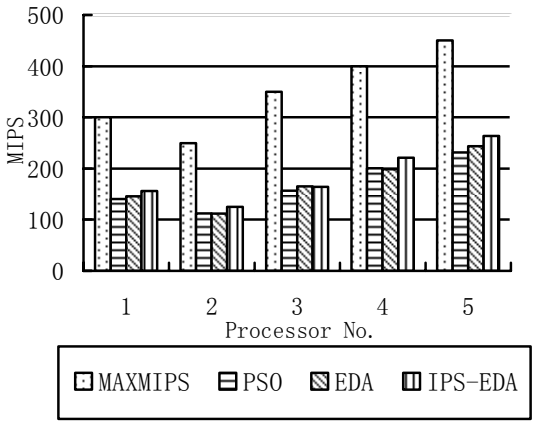

Figure 3. Comparison of load balancing

3) Fitness function value of IPS-EDA

Table 1 shows that IPS-EDA algorithm can achieve a smaller value of fitness function under different number of tasks when compared to IPS-EDA and EDA /PSO algorithm .

TABLE 1. OPTIMIZATION RESULTS OF DIFFERENT NUMBER TASKS

Serial number	Number of tasks	PSO	EDA	IPS-EDA
1	35	7.895	7.975	7.784
2	25	5.753	5.748	5.689
3	30	6.573	6.598	6.495
4	15	3.875	3.783	3.695

V. CONCLUSIONS

Particle swarm optimization algorithm uses each particle's memory and accumulation knowledge of the whole population to search the global optimal solution. In this paper, proposed IPS-EDA finds better solutions through avoid falling into local optimal value. This is because when the PSO particle stopped searching for better solutions, EDA random sampling will be dispersed particle position. The proposed IPS-EDA algorithm avoids falling into local optimal solution of the premature phenomena in the search process, and to expand the search space during the searching process. In addition, this paper use MapReduce programming paradigm to implement IPS-EDA algorithm. The experimental results show that compared with PSO and EDA, IPS-EDA provides better load balancing and search optimal capacity in grid environment. IPS-EDA algorithm shows high timeliness, scalability and reliability, bring good prospects for the use of algorithm in engineering.

ACKNOWLEDGMENT

The paper is supported by National High-tech R&D Program of China (863 Program) (No. 2013AA01A212), National Science and Technology Ministry of China (No. 2012BAH27F05), the Guangdong Nature Science Foundation (No. S2012030006242).

REFERENCES

[1] HayesB. CloudComputing.Communications of the ACM, 2008,51 7:9-11.

[2] LING, DASMALCHIG, ZHUJ. Cloud computing and IT as a service: opportunities and challenges. Proc of the IEEE 6th International Conference on Web Services(ICWS'08). Los Alamitos: IEEE Computer Society, 2008:1-5.

[3] Wang GZ, Salles MV, Sowell B, Wang X, Cao T, Demers A. Behavioral simulations in MapReduce[C]//PVLDB, 2010: 952-963.

[4] Jeffrey Dean, Sanjay Ghemawat. MapReduce: simplified data processing on large clusters.Communications of the ACM. 2008, 51(1):107-113.

[5] Kennedy J, Eberhart R C. Particle swarm 0ptimization .Proceedings of IEEE International Conference on Neural Networks, 1995:1942-1948.

[6] Clerc M, Kennedy J. The particle swarm-explosion, stability, and convergence in a multidimensional complex space. IEEE Trans. on Evolutionary Computation, 2002, 6(1):58-73.

[7] Eberhart R C, Shi Y. Comparing inertia weights and constriction factors in particle swarm optimilation .Proc. 2000 Congress Evolutionary. Computation. Piscataway, NJ: IEEE Press, 2000: 84-88.

[8] Shapiro J L. Drift and scaling in estimation of distribution algorithms. Evolutionary Computation, 2005, 13(1):99-123.

.

