Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
March-April 2021 Page No. 226-236
Article History: Received: 05 January 2021, Revised: 02 February 2021, Accepted: 10 March 2021, Publication: 30 April 2021

EA Heuristic Optimization Algorithm for Resource Load Balancing Based on Random Forest Model in Web Cluster

Chunjuan Wang[1*]
[1]School of Information Engineering, Shaanxi Xueqian Normal University, Xi'an, Shaanxi, China
*Corresponding Author.

Abstract:

Web cluster system needs to use all kinds of resources of the system to meet the customer's request. How to make the system resource balanced distribution and make the optimal utilization of system resources is an urgent problem. Stochastic forest model improves the prediction accuracy of the model by summarizing a large number of classification trees. It is a new model to replace the traditional machine learning methods such as neural network. Aiming at the current application status and characteristics of Web cluster, a heuristic algorithm is proposed, which can balance the load of resources and make full use of system resources. It is an extension and optimization of set partitioning problem (SPP) and multiple choice multi dimension knapsack problem (MMKP). The heuristic algorithm can significantly reduce the computational complexity in the optimal allocation of resources, and make it meet the needs of real-time scheduling. The simulation results show that the method is effective.

Keywords: *Random forest model, Web cluster, resources, balanced allocation, resource utilization.*

I. INTRODUCTION

In the past few decades, network computing has become the mainstream technology in large-scale resource sharing and distributed integration system due to a new computing infrastructure and scientific research cooperation. In the server cluster, the user's request needs to be distributed to the background server by the load balancer for processing [1-2]. When the load balancer receives the resource request from the internal or external, according to the load situation of the server cluster, the balanced scheduling algorithm is used for reasonable allocation [3].

The purpose of load balancing mechanism is to provide a good solution for service cluster to

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
March-April 2021 Page No. 226-236
Article History: Received: 05 January 2021, Revised: 02 February 2021, Accepted: 10 March 2021, Publication: 30 April 2021

allocate tasks efficiently. Generally, load balancing algorithms can be divided into decentralized and centralized, static and dynamic algorithms. In the field of load balancing, service cluster has achieved a lot of research results, such as round robin scheduling, weighted minimum connection scheduling, source address hash scheduling and so on. Although the traditional scheduling algorithm is simple to operate, it is not suitable for the complex real environment, and the point-to-point algorithm often produces a large number of wrong peak points in the search process [4-5]. Thus, the judgment of the best result is affected. Zomaya and teh proposed to apply genetic algorithm (GA) to load balancing strategy. The essence of GA is a highly parallel global search algorithm. It can automatically acquire and accumulate knowledge about the search space in the search process, and automatically control the search process to obtain the optimal solution.

This paper is based on GA to distribute the load tasks reasonably and minimize the response time under certain resource utilization [6]. The mean variance model is suitable for the special relationship between the two in the fitness function. Therefore, it is of great practical significance to study the load balancing of Web Services Cluster Based on genetic algorithm, improve the traditional fitness function through the mean. Variance model, and shorten the waiting delay of user requests with a certain level of resource utilization, so as to obtain the optimal allocation combination and improve the user experience.

II. LOAD BALANCING TECHNOLOGY OF WEB SERVICE CLUSTER

Overview of Web Services Cluster

The definition of cluster is to deploy the same task on multiple servers, which can connect multiple servers and work like a machine. In order to improve the stability and reliability of the operation system, as well as optimize the data processing ability and service ability of the network center, the method of Web cluster system is generally used.

To put it simply, Web cluster is to bundle many servers together to achieve high-speed internal network by converting cable technology, forming high performance computing (HPC), so as to support multi-user parallel computing [7-9]. Basically, grid resources are distributed in computers or clusters. As a unified computing resource, grid resources are logically aggregated. Due to the arrival mode of uneven tasks and unequal computing power, computing nodes may be overloaded in one network site, while others may not be fully utilized in different network sites. Therefore, it is necessary to use this network system reasonably. Task scheduling and resource management represent the service level of network software infrastructure. Task allocation and load balancing is a common problem, which is the basic function of most network systems.

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
March-April 2021 Page No. 226-236
Article History: Received: 05 January 2021, Revised: 02 February 2021, Accepted: 10 March 2021, Publication: 30 April 2021

2.2 Principle of load balancing

Load balancing is the ability to effectively distribute the entire load to multiple servers or members of a computer cluster. As shown in Figure 1, in the load balancing environment of the Web cluster, the load balancer receives requests from various internal and external resources. According to the type of request, the task is processed and then assigned to the corresponding service cluster. The server cluster based on the dispatcher can accept and process all requests, and it can also achieve fine-grained load balancing [10]. In this system, the function of the front-end request allocator is to proxy the requests that arrive, that is to say, to receive the HTTP requests that arrive, and to distribute the client requests to the back-end server cluster in a balanced and transparent way. The whole cluster system has a single virtual IP address, that is, the cluster address, so the server in the cluster is transparent to the client. The reason why the server in the system is transparent to the client is that the request allocator provides the only virtual interface for the request instruction, and the cluster address is the only virtual IP address.

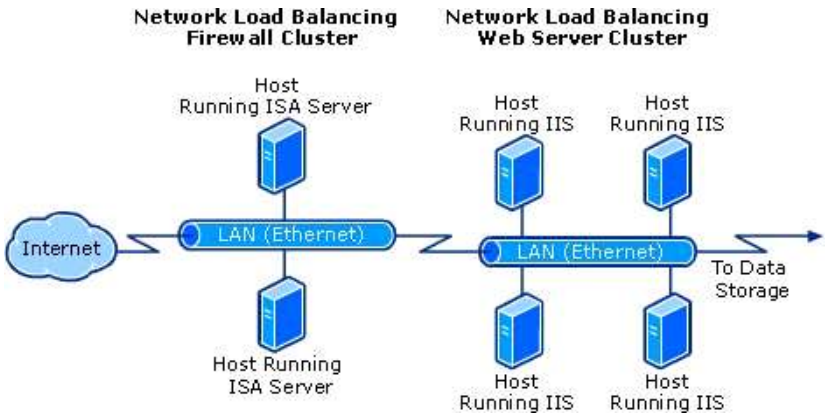

Fig 1: Load balancing environment of Web Cluster

In the service cluster environment, the purpose of load balancing mechanism is to balance the load allocated by each computing node, maximize the utilization of service nodes and reduce the total task execution time. In order to achieve the above goals, researchers have proposed many innovative load balancing technologies.

2.3 Key technologies of load balancing

(1) Direct routing for load balancing

Direct routing (DR) is a load balancing method in data link layer. In order to avoid the bottleneck of network card bandwidth in load balancing server. The main function of data link layer in TCP / IP protocol is MAC addressing and data exchange in the form of frame. The

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
March-April 2021 Page No. 226-236
Article History: Received: 05 January 2021, Revised: 02 February 2021, Accepted: 10 March 2021, Publication: 30 April 2021

server cluster can carry out load balancing by using Ethernet data to reach the computer network card through MAC addressing. In the data link layer, modify the MAC address in the packet to achieve the purpose of load balancing. The specific settings are shown in the figure.

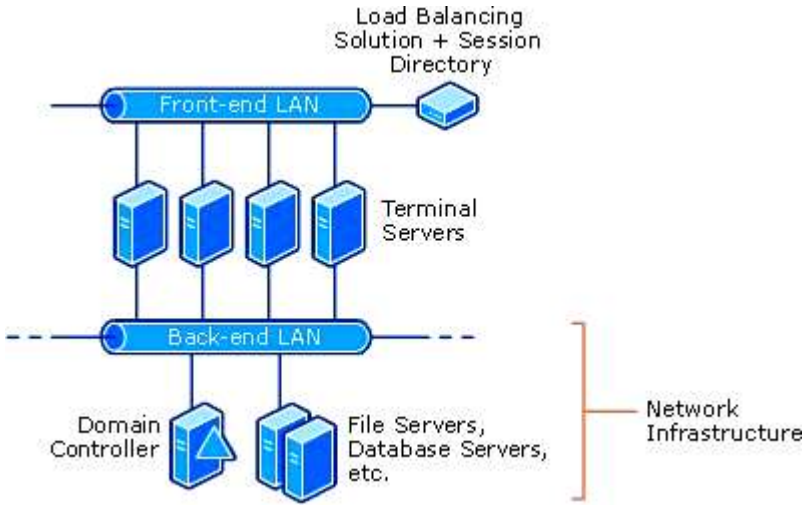

Fig 2: Load balancing of direct routing

(2) IP tunnel with load balancing

The principle of IP tunnel request forwarding mechanism is that the server sends the response data to the user. Before that, the scheduler should be used to repackage the received packet and send it to the server. At present, Linux mostly supports it and can be implemented by LVS, which is called LVS.tun (virtual server via IP tunneling). Different from LVS . DR, the actual server must have a legal port because it is not in the same WANt network segment as the scheduler. LVS . TUN configuration enhances LVS. DR function, and cluster nodes can be forwarded in different physical segments from Director. For example, in order to facilitate the transmission across the Internet, the data packets inside become valid data. If the server knows how to analyze these data packets, it can become a cluster node.

(3) Reverse proxy for load balancing

The load balancing technology of reverse proxy (RP) is applied to the application layer. The proxy server receives the client's request on Intemet, and calculates the appropriate service node in the cluster to process the request through the load balancer (RP) scheduling algorithm, and finally feeds back to the client. In the reverse proxy load balancing technology, when the reverse proxy server schedules the load balancing algorithm according to the user's request and makes the appropriate server selection, it forwards the request to the corresponding service

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
March-April 2021 Page No. 226-236
Article History: Received: 05 January 2021, Revised: 02 February 2021, Accepted: 10 March 2021, Publication: 30 April 2021

node, such as 192.168.1.1. When the service node 192.168.1.1 completes the user's request task, it will feedback the progress of the reverse agent and act as the reverse agent server of the transfer station to respond to the user.

III. LOAD BALANCING METHOD OF SERVICE CLUSTER BASED ON GENETIC ALGORITHM

3.1 Genetic algorithm

Most optimization methods determine the next point by a certain scheduling rule from a single point in decision space. This point-to-point approach is risky because it can find false peaks in multiple or multi peak search spaces. In contrast, GA climbs multiple peaks parallel from the point of a database (string set), so the relative probability of finding a false peak is reduced. GA mechanism is relatively simple, and does not involve more complex steps than copying strings and exchanging partial strings. Simple operation and fitness solution are two main factors in GA. The effectiveness of GA also depends on two factors: the standard of combination fitness function and the process of three-step genetic exploration.

Genetic algorithm can be defined as an array genetic algorithm (AGA):

$$AGA = (C, F, P_t, N, \Phi, \Gamma, \Psi, T)$$

Among them, C is the code of individuals in the candidate domain, F is the fitness function for the selection of excellent individuals, P_t is the initial population in the candidate domain, N is the size of the population, Φ, Γ and Ψ are the selection operator, crossover operator and mutation operator in the genetic operation respectively, and T is the end condition of the iteration of the genetic algorithm.

A load balancing method of server cluster based on genetic algorithm includes the following steps:

(1) The candidate solutions of the space are encoded, and an appropriate amount of initial string structure data is randomly generated as the initial population

(2) Resource utilization and execution time adaptability are evaluated and detected;

(3) According to the roulette method, the string with strong adaptability is selected, and the selected string is crossed and mutated, and a new string is generated for the next iteration;

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
March-April 2021 Page No. 226-236
Article History: Received: 05 January 2021, Revised: 02 February 2021, Accepted: 10 March 2021, Publication: 30 April 2021

(4) Genetic algorithm iterates with the initial population. When the difference between the fitness of the optimal string and the minimum fitness value is less than e (set a certain difference) or the iteration reaches the preset algebra, the algorithm will terminate; Go back to step 2.

3.2 Quantify load

When clients have a large number of access at the same time, the different users' requests consume different computing and network resources due to different task types. The client request type, the current network bandwidth and the utilization of the resources of the back-end server cluster are all factors that affect the quantitative load. For example, some requests only need to read an HTML page, and then calculate it simply to reduce the load to light load; Other database queries and calculations that require intensive data flow need to be quantified as heavy load. So we can quantify the different requests reasonably. Arilit et al. In 1996, the server log files of NASA Kennedy Space Center were analyzed effectively, and then different types of request documents involved were classified according to certain standards.

3.3 Adaptive threshold algorithm

Since the threshold is the task mapping that determines the adaptability of genetic iteration, they must adjust the threshold before task allocation in the sliding window. In order to determine the appropriate threshold, the average load value of nodes in load balancing system must be determined first. This is the sum of the load of the current system and the load of the tasks to be assigned in the sliding window, and then the average load value is obtained by the ratio of the total number of nodes in the system. The definition of average load in Appendix strategy is given as follows:

$$L_{ave} = \frac{CSL + NTL}{N} \quad (1)$$

Here, L_{ave} = average load, which is the average load;

CSL = the current system load, which is the node load in the current service cluster system;

NTL = the new tasks load, which is the total load to be allocated to the system in the sliding window;

N represents the total number of service nodes;

If a new set of tasks is assigned and the load of all processors is calculated to be the same as the average load, the system will be well balanced. However, using this value as the system threshold will be very strict, which is difficult to achieve as a system wide balance state, especially in a large load balancing system. Sometimes it is more realistic for load balancing to

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
March-April 2021 Page No. 226-236
Article History: Received: 05 January 2021, Revised: 02 February 2021, Accepted: 10 March 2021, Publication: 30 April 2021

relax the limit. Therefore, in the adaptive allocation process, the use of heavy threshold and light threshold adds more flexibility to the server load balancing system.

The "threshold" in the adaptive threshold strategy indicates that the processor is heavily loaded or lightly loaded. Each processor will directly report to the central scheduler when it reaches or completes the task, and then get the average load of a single server according to the current system load and the load of the new scheduling task. However, the adaptive object here is not the load state of a single server, but the index serving the whole load balancing system.

In the adaptive threshold method, the key step is to determine the appropriate threshold of the current service system, so as to achieve a good load balancing algorithm. If the threshold is set too low, excessive load balancers will increase and congestion will occur. However, if the threshold is set too high, the load balancing mechanism will fail.

Two kinds of threshold strategies can be regarded as load balancing algorithms, one is fixed threshold strategy, the other is adaptive threshold strategy in this paper. As the name implies, the fixed threshold policy has a predetermined threshold, which will not change when the system load changes. On the other hand, the threshold of adaptive threshold strategy is adjusted according to the system load. Then, the new threshold will be transmitted to the central dispatching management unit.

IV. DESIGN OF SERVICE CLUSTER LOAD BALANCING BASED ON MEAN-VARIANCE MODEL

4.1 Mean-Variance model

Mean-Variance model solves the problem of portfolio optimization in the field of economics, and maximizes the return rate and minimizes the risk by rationally allocating the path of assets and the weight of the path. Based on the theoretical basis of Mean-Variance model, the expected return and risk of investment portfolio are established, and the objective function of Mean-Variance model is set as follows:

$$min\sigma\left(r_p\right)=\sum\sum x_i x_j Cov\left(r_i r_j\right) \quad (2)$$

$$r_p=\sum x_i r_i \quad (3)$$

The limitations of the objective function:

$$1=\sum x_i \quad (4)$$

Mean-Variance mathematical model reveals the conclusion that "the expected return of assets is determined by its own risk", that is, the return of assets is determined by the path risk

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
March-April 2021 Page No. 226-236
Article History: Received: 05 January 2021, Revised: 02 February 2021, Accepted: 10 March 2021, Publication: 30 April 2021

of asset allocation, that is, by variance, and the return is determined by covariance. The idea of the Mean-Variance model is "mean-variance" or "mean-standard deviation". The following is the effective boundary of "mean-standard deviation", as shown in Figure 3:

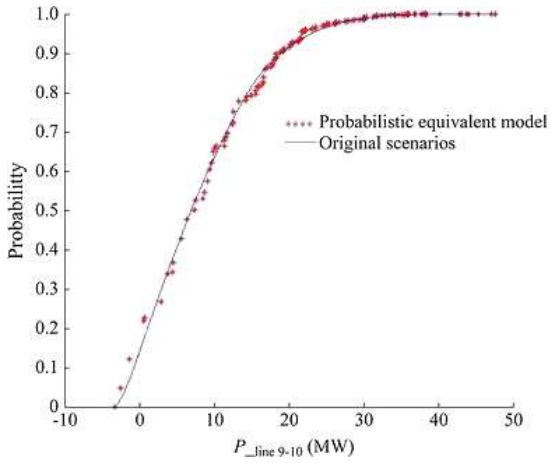

Fig 3: Efficient boundary of Mean-Variance model

4.2 Design of coding mechanism for load balancing

In GA, parameter coding is the first step in the main steps of load balancing. Coding is to transform the individuals in the solution space into an array represented by attributes into a search space that can be processed by genetic algorithm. When the optimal combination of deflection is assigned, the reverse operation is needed, which is called decoding.

There are several typical coding methods in GA:

(1) Binary code:

a. Definition: binary encoding is a string composed of binary symbols 0 and 1.

b. For example: $0 \leq x \leq 1023$, the precision is 1, M is the length of binary code. Make $2^{m-1} \leq 1000 \leq 2^{m-1}$. The binary code is related to the precision.

c. The advantages are that it conforms to the principle of minimum character set and is easy to use; The disadvantage is also obvious, which is easy to cause error in discretization.

(2) Gray code coding

a. definition: only one code bit in the coding corresponding to any two consecutive integers is different, which is called gray coding.

b. example: for interval [0. The two adjacent integers X_1=175 and X_2=176 in [1023] can be expressed as X_{11}:0010101111 and X_{12} : 0010110000 if they are coded by 10-bit binary code,

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
March-April 2021 Page No. 226-236
Article History: Received: 05 January 2021, Revised: 02 February 2021, Accepted: 10 March 2021, Publication: 30 April 2021

and they can be expressed as X_{21}: 00101111 and X_{22} : 0010101000, respectively, using Gray code of the same length

c. Advantages: compared with binary, gray coding has some advantages in local search ability.

In the coding mechanism, there are many coding methods of parameters, including binary coding, gray code, floating-point code, permutation coding, etc. among them, binary coding is the most common coding method because of its simplicity and processing mode number. However, the long binary encoding string makes the space search larger and will always occupy memory, which leads to the inefficient use of computer resources. Therefore, this paper proposes to use three-dimensional decimal system to encode the candidate solutions of space. Each string in the array has a fixed size, and each node in the search space is represented by a string, which is unique. Thus, 3D decimal coding will also improve the accuracy of genetic operation.

4.3 Implementation of optimized load balancing based on Mean-Variance model

This paper designs a simple example, which is the task allocation of service cluster load balancing based on mean variance. ① Current system status. For the current information of the processor at the current time, if the processor 3 task is empty, the new task needs to be reallocated. ② New tasks are assigned. According to the new round of tasks in the sliding window, the total time unit is calculated. ③ Genetic algorithm generates appropriate task mapping, namely task allocation. It mainly includes two problems: how to assign tasks in detail and whether to assign tasks according to objective function or adaptive function. ④ Determine the new threshold. According to the sum of the current system load and the tasks to be allocated, and then compare with the number of processors to get the average value, and then calculate the maximum threshold and minimum threshold. ⑤ The acceptable load of each processor is calculated. By subtracting the total load of load queue from the maximum threshold, the acceptable load is obtained. ⑥ Decision on task allocation. After getting the acceptable load of each processor, you can go back to the third step to see whether the load of each processor has exceeded the acceptable value. If it exceeds the acceptable value, it indicates that more serious load imbalance has been caused. ⑦ New system state. T + 1 load state. If processor L is detected to be idle and another batch of tasks are to be allocated, the sliding window will be filled with new tasks. ⑧ Update the sliding window. The task in the sliding

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
March-April 2021 Page No. 226-236
Article History: Received: 05 January 2021, Revised: 02 February 2021, Accepted: 10 March 2021, Publication: 30 April 2021

window will be replaced by a new task queue, and the new sliding window task will cycle.

The service cluster is evaluated according to the composition fitness function, so as to realize the survival of the fittest in the distribution composition. Firstly, the adaptive strings are selected and copied according to roulette selection method, and the excellent strings are crossed and mutated. After the mutation operation, the string will generate new fitness values, so it needs to be re evaluated, and then new survival probability will be generated. These values will be used to define the slot value of the wheel in the next cycle. Finally, it is necessary to judge whether K cycles are reached. If so, the optimal string will be decoded and used for task allocation. Otherwise, the iteration will continue. For the new system state T + 1, check whether there are idle processors. If there are idle processors, load balancing will be started to allocate new tasks.

V. CONCLUSION

Nowadays, network service technology is trusted by more and more people. When there are a large number of requests at the same time, load balancing strategy is very important in effective task allocation. Many projects in the literature focus on this problem and propose a series of solutions. Through the research of the method of service cluster equilibrium, this paper finds that the weight of node utilization in service cluster is set by combining the mean variance model in portfolio selection theory. The optimal weight vector is obtained by minimizing the task completion time. That is, the weight is allocated to the resource utilization of each server, and more effective combination fitness function is obtained. As a method of simulating natural evolution to search the optimal solution, genetic algorithm has inherent implicit parallelism and better global optimization ability. It has been widely used in combinatorial optimization, machine learning and adaptive control. After deeply studying the model of genetic algorithm, aiming at the deficiency of the existing load balancing method of service cluster based on genetic algorithm, this paper introduces the portfolio mean variance model to the load scheduling based on genetic algorithm.

ACKNOWLEDGEMENTS

This research was supported by School Natural Science Foundation of China (Grant No. 2020ZDRS02).

REFERENCES

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
March-April 2021 Page No. 226-236
Article History: Received: 05 January 2021, Revised: 02 February 2021, Accepted: 10 March 2021, Publication: 30 April 2021

[1] Wei Yonglian, Yi Feng, Feng Dengguo, Yong W, Yifeng L. Network Security Situation Assessment Model Based on Information Fusion. Computer Research and Development, 2009, 46 (3): 353-362

[2] Xu Guoguang, Li Tao, Wang Yifeng. A Network Security Real-time Risk Detection Method Based on Artificial Immune. Computer Engineering, 2005,31 (12): 945-949

[3] Jiang Wei, Fang Binxing, Tian Zhihong. Network Security Evaluation and Optimal Active Defense Based on Attack Defense Game Model. Acta Computer Sinica, 2009, 32 (004): 817-827

[4] Miao Yongqing. Stochastic Model Method and Evaluation Technology of Network Security. China Science and Technology Investment, 2017, 4: 314

[5] Yi Hua Zhou, Wei Min Shi, Wei Ma. Research on Computer Network Security Teaching Mode for Postgraduates Under the Background of New Engineering. Innovation and Practice of Teaching Methods, 2020, 3 (14): 169

[6] Yang Yi, Bian Yuan, Zhang Tianqiao. Network Security Situation Awareness Based on Machine Learning. Computer Science and Application, 2020, 10 (12): 8

[7] Li Zhiyong. Hierarchical Network Security Threat Situation Quantitative Assessment Method. Communication World, 2016, 23: 70-70

[8] Hu Wenji, Xu Mingwei. Analysis of Secure Routing Protocols for Wireless Sensor Networks. Journal of Beijing University of Posts and Telecommunications, 2006, 29 (s1): 107-111

[9] Bao Xiuguo, Hu Mingzeng, Zhang Hongli. Two Quantitative Analysis Methods for Survivability of Network Security Management Systems. Acta Communication Sinica, 2004, 25 (9): 34-41

[10] Li Weiming, Lei Jie, Dong Jing. an Optimized Real-time Network Security Risk Quantification Method. Acta Computa Sinica, 2009 (04): 793-804