Online Test System Load Balancing Method Based on Learning Analytics and Prediction Model

SI Wen-xue, ZHUGE Bin, CAI Jia-qi, DONG Li-gang and JIANG Xian

School of Information and Electronic Engineering, Zhejiang Gongshang University, Hangzhou 310018, China
zhugebin@zjgsu.edu.cn, *19020090009@pop.zjsu.edu.cn

Abstract. Aiming at the problem of load imbalance such as database instability caused by high concurrency in the online test system, an online test system load balancing method based on learning analysis and prediction model is proposed. The method is divided into two steps. The first step is to build a curve of the number of questions answered according to the user test data. Simulate the test load change at different time intervals, reduce the number of concurrent connections by reasonably arranging the test; the second step is to analyze the user's usual learning data and extract multiple features and then construct a time prediction model through the convolutional neural network. When the user accesses the test system, we can test the time required for him to complete the exam based on the predictive model. Each type of user is evenly redirected to each server by length of time. The experimental results show that the number of concurrent connections in the database generated by this method is lower, and the load of each server can be effectively balanced. In addition, the method can predict the load situation in advance so that can provide a reference for purchasing server resources before the test.

1. Introduction
With the rapid rise of the mobile Internet, the in-depth development of network services and the high popularity of smart phones, online teaching has become an inevitable trend [1]. However, no matter if you use a computer or a mobile phone, when a large-scale online exam is implemented, because a large number of candidates enter the system at the same time, the server needs to process a large number of access requests in a short time, and the system is prone to crash under such a high amount of concurrency [2]. In order to avoid the impact of a large number of user requests on the system and ensure the response speed and stability of the system, a better solution is to use load balancing technology[3].

In the era of rapid development of large-scale cluster computing and cloud computing technology, load balancing technology has become particularly important. At present, a lot of research has been carried out on this at home and abroad, and various load balancing algorithms have been proposed[4]. Load balancing algorithms can usually be divided into static algorithms and dynamic algorithms, and the load balancer uses these specified algorithms to forward service requests to the optimal server.

2. Load balancing method of online examination system based on learning analysis and prediction model
The data set used in this article is the data generated during the use of students in the freshman education platform of our school. The platform is built in a cloud application (Sina App Engine, SAE)
under the Sina Cloud environment, which mainly provides students with functions such as free exercises and online exams. There are more than 4000 registered users, including 4,686,601 free practice records, 175,825 random practice records, 69826 repeated practice records, 45447 exercise collection records, 1061018 test answer records.

2.1. Load balancing experiment design
This article builds 7 identical new beginning education platforms (hereinafter referred to as "platforms") among 7 SAEs, and 7 SAEs are connected to the same exclusive database, so their configuration and processing capabilities are the same. This article takes the actual situation in the use of the platform as an example to design a load balancing scheme, and respectively study (1) Under inherent constraints, how to arrange exams so that the number of concurrent connections to the database is below the quota, and the total time to complete all exams is shorter; (2) How to make the load of each SAE more balanced under a certain test arrangement. That is, the first step: compare the number of concurrent connections to the database generated under different time intervals and the total time consumed at the end of all exams, and select the interval where the number of concurrent connections is less than the quota and the total duration is shorter to schedule the exam. Step 2: under a fixed examination arrangement, different types of students are evenly distributed to each SAE according to a certain method, so that the load of each SAE is more balanced. According to the above two steps, two load balancing experiments are designed respectively.

2.1.1. Load balancing method based on learning analysis.
In the online exam, every time students answer a question, they need to query the question information from the database and submit the answer information to the database, which will generate a database connection. Therefore, according to the test records, this article counts the number of minutes answered by all students in each college within 50 minutes of the actual test, and forms a curve of the number of answers with the test time, which are recorded as respectively. In the experiment, a total of 3,710 qualified students entered the examination system according to the college divided into 14 examinations, and each examination time was 50 minutes. Change the time interval between each exam, through the superposition of \( f \) to simulate the change curve of the amount of answers in all exams over time:

\[
F(t) = \sum_{t=1}^{T=50+(n-1)\Delta t} (f_{a_1}(t) + f_{a_2}(t-\Delta t) + \cdots + f_{a_n}[t-(n-1)\Delta t])
\]  

Among them, \( T = 50 + (n-1) \times \Delta t \) is the total time consumed by the test, \( n \) is the number of test sites, \( \Delta t \) is the test interval, and \( t \) is the test time. Compare the number of concurrent connections to the database generated under different time intervals and the total time \( T \) consumed at the end of all exams, and select the exam schedule with the concurrent connection below the quota and the shortest total time consumption. Among them, the number of concurrent connections to the database quota is 400. For the sake of safety, this article sets the maximum number of concurrent connections not to exceed 70% of the quota in the experiment.

2.1.2. Redirection load balancing method based on prediction model
This paper uses CNN to construct a prediction model of student answering time, and proposes a redirected load balancing method based on prediction model (PM-LB).
First, conduct a statistical analysis of the students’ historical learning records in the online examination system, and extract learning features including the total number of answers, the number of favorites, and the average answer time of daily practice as the input of the CNN model. The average answer time during the exam is used as the label value for model training and test. Use the features extracted from the learner's historical learning records to train the CNN model. Use the partial derivative of the loss function, that is, the derivative of the loss function with respect to the weight, is used to adjust and update the parameters, so that the loss function in the model is the lowest. Use
Adam The optimizer uses the learning rate to reduce the cross entropy. After calculating the gradient, subtract it from the initial weight to get the new optimized weight:

$$\theta_{i+1} = \theta_i - \alpha \times \nabla J(\theta_i)$$

(2)

Among them, $\theta_{i+1}$ is the optimized weight, $\theta_i$ is the initial weight, $\alpha$ is the learning rate, and $\nabla J(\theta_i)$ is the gradient of the loss function.

After completing the construction of the answer time prediction model, when students visit the examination system, they obtain their learning characteristics and predict their answering speed and the time required to complete the exam through the prediction model. All candidates are classified according to the length of completion time, and the candidates are redirected according to the category they belong to, that is, candidates of each category are equally allocated to 7 SAEs. Among them, this paper randomly divides the data set into two parts, 70% is used as the training set to train the CNN model, and the remaining 30% is used to test the trained model.

2.2. Load indicators and measurement methods

The load index is a characteristic value that measures the load status of the server and is used to describe the busyness of the server. In this paper, the number of concurrent database connections (DCC), page views (PV), and unique visitors (UV) are selected as load indicators, and the experimental results of the two load balancing methods are compared and analyzed.

In this experiment, the number of questions answered by students per minute during the exam is converted into the number of concurrent connections to the database and the PV value, respectively, according to a certain proportional relationship $k_1$ and $k_2$. The UV is naturally determined by the number of online people in the exam. Because the IP of the student's mobile phone changes during the exam, there is also a ratio $k_3$ when converting the current online number of people into the UV value. The formula is as follows:

$$DCC = k_1 \times F(t)$$  \hspace{1cm} (3)

$$PV = \frac{k_2 \times F(t)}{N}$$  \hspace{1cm} (4)

$$UV = \frac{k_3 \times S}{N}$$  \hspace{1cm} (5)

Substitute formula (1) into formula (3) (4), where $F(t)$ is the number of questions answered in minutes, $N$ is the number of servers, and $S$ is the current number of online users. Through the statistical analysis of the learning data and HTTP logs, this paper found that the conversion coefficients in this experiment are $k_1=0.05$, $k_2=3.06$ and $k_3=1.56$, respectively. In other environments, the coefficients may have certain changes, and the actual situation shall prevail.

3. Experimental results and analysis

This section compares and analyzes the results of the two experiments designed in the previous section.

3.1. Load balancing method based on learning analysis

Since each test lasts 50 minutes, the test interval $\Delta t$ is set to 5, 10, 15, 20, and 25 minutes. Under the fixed 14 exams, change the time interval between each exam, compare the number of concurrent Database connections generated under different time intervals and the total time consumed at the end of all exams.

It can be seen from Fig 1 that the maximum number of concurrent connections decreases as the test interval increases, and reaches the maximum value of 270 when the interval is equal to 5 minutes. When the interval exceeds 15, the maximum number of concurrent connections has dropped below 100. According to the variation of the maximum number of concurrent connections and the total test
duration with the test interval, all tests can be completed at the lowest cost under different restrictions. If the quota of concurrent connections to the database is 200, then the test interval needs to be set to 15 minutes, but the total test time will increase to 245 minutes. On the contrary, this method can predict the server load in advance, and on this basis, the required server resources can be purchased according to the load status.

As can be seen from Fig 2, the maximum number of concurrent connections will decrease as the number of exam venues increases, and as the exam interval increases, the number of concurrent connections will be lower. Obviously, the total test time will increase as the test interval and the number of test venues increase. Through experiments, select a combination that meets the maximum number of concurrent connections below 70% of the quota and the total test duration is less than 120 minutes. The results are:

| Combination | Examination interval (minutes) | How many exams | Maximum concurrent connections (times) | Total test time (minutes) |
|-------------|-------------------------------|---------------|--------------------------------------|--------------------------|
| A           | 5                              | 14            | 270                                  | 115                      |
| B           | 5                              | 15            | 252                                  | 120                      |
| C           | 10                             | 7             | 276                                  | 110                      |
| D           | 10                             | 8             | 242                                  | 120                      |
| E           | 15                             | 5             | 268                                  | 110                      |
| F           | 20                             | 4             | 275                                  | 110                      |

From the above table, you can choose E combination as the test arrangement when the quota of concurrent database connections is 400. Compared with the arrangement of real exams, E combination makes full use of server resources and greatly reduces the total time consumed by exams.
3.2. **Redirection load balancing method based on prediction model**

After confirming the examination arrangement, the second step can be carried out: balance. In order to verify the effectiveness of this method, the experiment selects two applications a and b, using the method in this paper and the rotation training algorithm, respectively, to compare and analyze the change curve of their PV and UV values with the test time. Among them, the ordinate is the PV/UV value, and the abscissa is the actual test time.

![Fig 3. Comparison chart of the change of PV value with the test time under the two methods](image)

![Fig 4. Comparison chart of the change of PV value with the test time under the two methods](image)

![Fig 5. Comparison chart of the change of UV value with the test time under the two methods](image)
Fig 6. Comparison chart of the change of UV value with the test time under the two methods

As can be seen from Fig 3 to Fig 6, after using the predictive model-based redirection load balancing method proposed in this article, the PV/UV value of each application is more balanced, especially at 17:00: During 00-21:00:00, during the process of continuous participation of candidates, the fluctuation range of PV/UV over time is much smaller than that of the rotation training algorithm. It can be seen that the equalization effect of the algorithm in this paper is better than that of the rotation algorithm.

4. Concluding remarks
The curve of answer volume over time constructed by analyzing test data can provide more scientific and accurate guidance for reasonable test arrangements. The redirected load balancing method based on the predictive model makes the load of each server more balanced, and the fluctuations are smoother. The model and method in this article are suitable for large-scale online exam scenarios in universities and institutions, as well as other scenarios where high concurrent access causes uneven load.

Acknowledgements
This work was supported by the National Natural Science Foundation of China (61871468), the Key Research and Development Program of Zhejiang under Grant No.2017C03058, the Key Research and Development Program of Zhejiang under Grant No.2020C01079, Zhejiang Provincial Key Laboratory of New Network Standards and Technologies No.2013E10012, Education Science Planning Project of Zhejiang No. 2019SCG222, The Second Batch of Teaching Reform Research Projects in the 13th Five-Year Plan of Zhejiang Higher Education (jg20190181), Zhejiang Gongshang University Graduate Education Reform Project Funding (YJG2018216)

References
[1] Liu H, Li JJ, Zhang W, et al. Data visualization method and application oriented to online teaching platform[J]. Distance Education in China,2018(1):37-44
[2] Li Y P, Zou H. Load balancing technology and implementation scheme of application servers[J]. Telecommunications Science, 2013,29(10): 15-18.
[3] Dong B, Li X, Wu Q, et al. A dynamic and adaptive load balancing strategy for parallel file system with large-scale I/O servers[J]. Journal of Parallel & Distributed Computing, 2012, 72(10):1254-1268.
[4] Ramakrishna M, Kodati V K, Gratz P V, et al. GCA: Global Congestion Awareness for Load Balance in Networks-on-Chip[J]. IEEE Transactions on Parallel & Distributed Systems, 2016, 27(7):2022-2035.