

DBN Based Cloud Service Response Time Prediction Method

Yan Gao*, Bin Zhang**, Si Wang*, Anxiang Ma*

* School of Computer Science and Engineering, Northeastern University, China

** Software College Northeastern University, China

gaoyan@mail.neu.edu.cn, zhangbin@mail.neu.edu.cn, 378894183@qq.com, maanxiang@mail.neu.edu.cn

Abstract—Cloud resources are taken the pay-as-you-go model, so cloud service providers expect to provide better cloud service performance at fewer virtual machine resources. Therefore, how to predict cloud service performance under given virtual machine resources and concurrent requests has become a hot research issue. Aiming at the problem of cloud service performance prediction, this paper takes response time as an important indicator to measure cloud service performance, and proposes a DBN-based cloud service response time prediction method. This method establishes a DBN based mapping model between cloud service response time with concurrent requests and virtual machine resources. The cloud service response time is predicted based on the mapping model. The experiments show the effectiveness of the proposed method.

Keywords—Cloud Service, Response Time, DBN, Virtual machine resources, Cloud Computing

I. INTRODUCTION

With the rapid development of cloud computing, more and more cloud services are deployed on cloud platforms to provide users with a variety of information or computing and other resources. Cloud services are deployed on virtual machines, and cloud resources such as virtual machines resources are take the pay-as-you-go model. Cloud service providers expect to provide cost-effective services that meet user requirements at the lowest cost of cloud resource usage^[1]. So it is required to estimate the cloud service performance at a given virtual machines resources^[2]. Research shows that cloud services performance depends on virtual machine resources and concurrent requests. Therefore, how to predict cloud service performance under given virtual machine resources and concurrent requests has become a hot research issue.

The research shows that the application performance and allocation of resources are often not linear relationship^[3]. Many researchers use artificial neural network to solve this problem. In literature [4], the author uses the method of artificial neural network to model the key factors that affect the performance of the application under different virtualization architecture. In literature [5], the author presents a method for system performance modelling of system resource using statistical learning method, and uses a single hidden layer artificial neural network method to train the data sets which contain physical machine resources and virtual machine resources. In literature [6], the author discusses some

factors that affect the performance of application and uses artificial neural network to model, quantify and predict the performance of the application.

DBN(Deep Belief Network) was proposed by Hinton et al^[7], and has been widely applied in prediction field^[8-11]. This paper takes response time as an important indicator to measure cloud service performance, and proposes a DBN based cloud service response time prediction method. The method establishes a DBN based mapping model between cloud service response time with concurrent requests and virtual machine resources, predicts cloud service response time based on the DBN based mapping model. The method uses CPU, RAM, bandwidth and storage to measure the virtual machine resources, and takes the virtual machine resources and concurrent requests as the input of DBN. The top layer of DBN uses neural network to establish nonlinear regression, and the bottom layer uses RBM stack. The key parameters of DBN, such as the number of hidden layers, the number of neurons in each hidden layer, and the related parameters of RBM, are determined by experiments. The structure of DBN model is determined by experiments and the influence of training size on the experimental results is analysed. And the effectiveness of the proposed method is verified by experiments.

II. DBN BASED CLOUD SERVICE RESPONSE TIME PREDICTION PROCESS

The response time of cloud service is related to the resource allocation of the virtual machine. The response time of cloud services is different under different resource allocations. At the same time, the response time of cloud service is different for different concurrent requests. The relationship between response time of cloud services with virtual machine resources and concurrent requests is non-linear, so it is necessary to establish a performance model of cloud service first, and then use the performance model to predict the response time of cloud service in different resource allocations. The process of DBN based cloud service response time prediction is shown in Figure 1.

The DBN based cloud service response time prediction process can be divided into two stages.

(1) Modelling stage. Firstly, the DBN based cloud service response time mapping model is established. Then, according to cloud service historical benchmark training data, the model

parameters are trained, and the optimal DBN network structure is determined to form DBN based cloud service response time mapping model.

(2) Prediction stage. The prediction stage is to predict the response time of cloud service based on the DBN based cloud service response time mapping model for a given virtual machine resources allocation and concurrent requests.

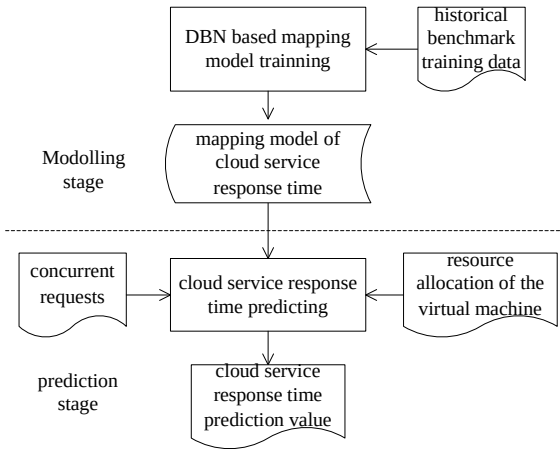


Figure 1. DBN based cloud service response time prediction process

III. CONSTRUCTION PROCESS OF DBN BASED CLOUD SERVICE RESPONSE TIME MAPPING MODEL

A. Cloud Service Response Time Mapping Model

There are many indicators for measuring the quality of cloud services, such as performance includes resource utilization ratio, throughput and response time, practicality and reliability. The response time is the most direct and most intuitive indicator to measure the performance of cloud services. Users are most concerned about is the response time of cloud services, it is a key factor affecting the quality of a website. Therefore, in this paper, the response time of cloud services is a measure of the performance of cloud services. If there are multiple cloud services in a virtual machine, the cloud services interact with each other, resulting in unstable response times for each cloud service. So this paper only considers the situation that a virtual machine deploying one cloud service. Each virtual machine has its own allocation resources. This paper uses resource amount of CPU, RAM, bandwidth and storage size to measure the resources of a virtual machine.

Definition 1. Virtual Machine Resource(VMR). VMR is represented by a four tuple:

$$VMR = \langle CPU, RAM, BW, HDD \rangle,$$

where, *CPU*, in general situation, adopt unified measurement (such as MIPS) to convert different architectures and computing power of CPU to CPU with the same computational capability; *RAM* represents memory size; *BW* represents bandwidth; *HDD* represents hard disk storage size. The resource is determined by the actual value of each attribute at a certain time.

Definition 2. Cloud service response time mapping model RT_CR_VMR . RT_CR_VMR is represented by a five tuple:

$$RT_CR_VMR = \langle CS, CR, VMR, RT, \tau \rangle,$$

where, *CS* represents cloud service; *CR* represents the average concurrent requests; *VMR* represents resource of virtual machine that *CS* deployed on; *RT* represents cloud service response time; τ represents mapping relationship $CR \times VMR \rightarrow RT$ on RT_CR_VMR of *CS*.

This paper use DBN to establish the mapping relationship τ . The DBN is usually stacked by multiple RBMs(Restricted Boltzmann Machines), and a classifier or a regression layer is established at the top level according to the problem domain and target^[8]. This paper uses the DBN based mapping model predict the response time of cloud service under a given virtual machine resources, so this paper constructs a non-linear regression layer at the top level of DBN using neural network, that is, the input layer is concurrent requests *CR* and virtual machine resource *VMR*, and the output is cloud service response time *RT*. The DBN based cloud service response time mapping model is shown in Figure 2.

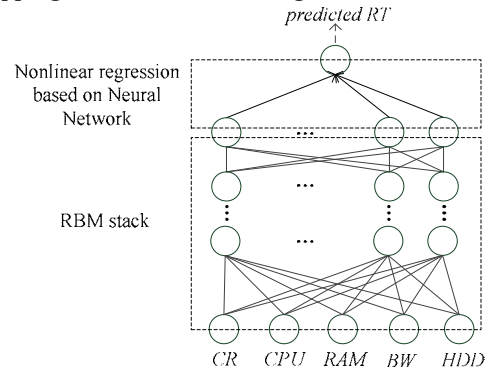


Figure 2. DBN based cloud service response time mapping model

B. Construction Process of Cloud Service Response Time Mapping Model

First, each RBM of DBN is trained and then fine-tuned by Error Back Propagation (BP) algorithm. The construction process of cloud service response time mapping model is shown in Figure 3.

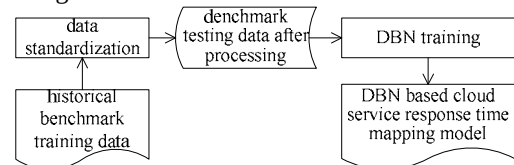


Figure 3. Cloud service response time mapping model construction process

The dimensions of each virtual machine resource attribute and concurrent requests in the input layer are different, for example, the dimension of bandwidth is in MB/S, and the dimension of memory is in GB, so the input layer data needs to be standardized. The smaller the resource attributes values, the lower the cost of the resource usage, so we can think the smaller the resource attribute values, the better. At the same time, the smaller the response time, the better. So in this paper, we take the "very small" normalized algorithm to standardize the concurrent requests, resources and response time. This paper use formula 1 to standardize the inputs of DBN to transform them to [0,1].

According to input parameters k , this paper uses the formula shown in formula 1 to standardize.

$$r_i^k = \frac{r_i^k - r_{\min}^k}{r_{\max}^k - r_{\min}^k} \quad (1)$$

where, r_i^k represents the standardization value of input parameters k ; r_i^k represents the value of input parameters k ; $r_{\max}^k$ represents the maximum value of input parameters k ; $r_{\min}^k$ represents the minimum value of input parameters k .

DBN has two important parameters, the number of DBN hidden layers and the number of neurons in each hidden layer. After data standardization, then train DBN in order to get optimum parameter values of the two parameters to generate cloud service response time mapping model.

IV. EXPERIMENT ANALYSIS

In this section, we get the optimum parameter values of DBN and verify the effectiveness of the proposed DBN-based cloud service response time prediction method through experiments.

A. Experiment Data and Experiment Schemes

Experiment data is an important part of the service response time prediction, the quality of the experiment data directly affects the objectivity and authenticity. In this paper, we assume that each virtual machine is running only one service. This paper presents sixteen kinds of virtual machine resources allocation schemes, and simulates nine kinds of distribution types of concurrent requests frequency to get the variety of concurrent requests. The nine kinds of distribution types are: Poisson Distribution, Normal Distribution, Catastrophe Distribution, Exponential Distribution, Gradual Distribution, Uniform Distribution, Chi-Square Distribution, Gradual Distribution, Random Distribution. We produced 100 data each assignment each distribution user access frequency, and totally obtained 14400 experimental data. We divide the dataset into two parts: training dataset and test dataset.

In order to avoid accidental error affecting the accuracy of the experimental results, we take the average value of 30 times experiment results for each experiment scheme. Firstly, DBN model is trained by training dataset, and then tested by test dataset. Cloud service response time predicted by DBN based cloud service response time mapping model is compared with actual response time. Root Mean Square Error ($RMSE$) is used as the evaluation criterion. The smaller the $RMSE$ is, the more accurate the prediction is. We use formula 2 to compute $RMSE$.

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (RT - RT^p)^2}{n}} \quad (2)$$

where, RT represents actual value of cloud service response time; RT^p represents prediction value of cloud service response time; n represents the data number of test dataset.

B. Determine the parameter values of DBN based cloud service response time mapping model

The neuron number of hidden layer will affect the modelling effect of DBN, so we use trial-and-error method to get the optimal number of hidden layer and neuron number of each hidden layer. We use 70% data of dataset as training dataset and 30% data of dataset as test dataset. The method has two steps.

Step 1. Consider the DBN model with only one hidden layer, and set the neuron number of hidden layer as 4, 6, 8, ..., and so on. Experiments were carried out on different neuron numbers to find the optimal number of neurons.

Step 2. Add a layer to DBN model on the basis of step 1, and the neuron number in this layer is still set as shown in step 1. Compare the results and select the optimal neuron number of the current layer. Then add another layer, and compare the obtained root mean square errors predicted by DBN models with different layers to determine the optimal hidden layer number.

Firstly, a hidden layer is set. The number of neurons in the hidden layer is 4, 6, 8, ..., 22, 24, respectively. Experiments are performed on the number of different neurons, and the prediction result of cloud service response time is obtained. The root mean square error between actual cloud service response times and the prediction one is calculated. The experiment results are shown in Figure 4, the horizontal axis represents the number of neurons and the vertical axis represents the root mean square error.

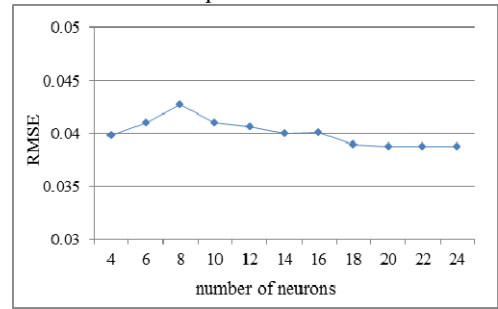


Figure 4. Experimental results of the first hidden layer

As shown in Figure 4, the overall trend of $RMSE$ gradually decreases with the increase of the number of neurons, but the $RMSE$ increases with the number of neurons between 4 and 8. Then, as the number of neurons increases, the $RMSE$ gradually decreases. The number of neurons starts from 18, and the $RMSE$ begins to stabilize and gradually converges. Therefore, the number of neurons in the first hidden layer is determined to be 18 in this paper.

After determining the number of neurons in the first hidden layer, the DBN model is added a new hidden layer, and the number of neurons is still set to 4, 6, 8, ..., 22, 24. The experiment results are shown in Figure 5.

As shown in Figure 5, when the number of neurons in the second hidden layer is varied from 4 to 14, the $RMSE$ gradually decreases with the increase of the number of neurons. When the number of neurons in the second hidden layer reaches 14, the $RMSE$ tends to be stable. Therefore, the number of neurons in the second hidden layer is determined to be 14.

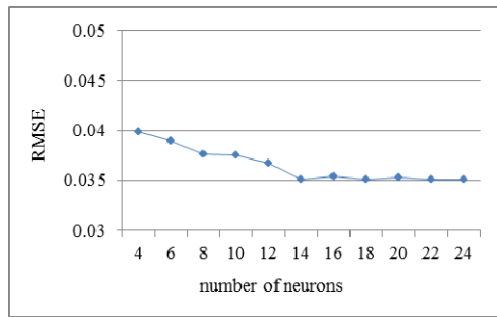


Figure 5. Experimental results of the second hidden layer

According to the third hidden layer, we get the experiment result by using the same method. The experiment results are shown in Figure 6.

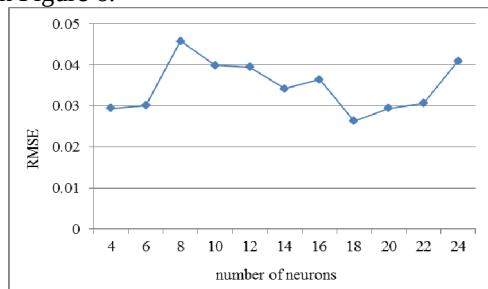


Figure 6. Experimental results of the third hidden layer

As shown in figure 6, when the hidden layers of DBN are increased to three layers, the trend of *RMSE* does not decrease with the increase of the number of neurons, but it fluctuates, indicating that the DBN model is not very suitable. We choose the best performance number of neurons as this hidden layer, that is 18. Since the *RMSE* of three hidden layers is not very suitable, so we do not need to increase the number of the hidden layers. Therefore, for the training dataset we use, the number of hidden layers of the DBN model is determined to be three layers, each layer containing 18, 14 and 18 neurons, respectively.

C. Comparison Experiments Analysis

The ratio of training dataset and test dataset will affect the modelling effect of DBN, so we do experiments under different ratios of dataset. We use the DBN model parameters that have already determined from above experiments. The ratio of training dataset is set to 50, 55, 60, 65, 70, 75, 80 and 85 respectively. The experiment results are shown in Figure 7.

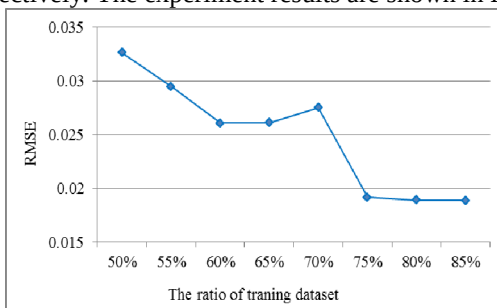


Figure 7. Effect of different ratio of training dataset

As shown in figure 7, with the increase ratio of training dataset, the *RMSE* gradually decreases. When the ratio of training dataset is in the range of 60% to 70% and in the range of 75% to 85%, the *RMSE* of the training results tend to be stable.

V. CONCLUSIONS

How to predict cloud service performance under given virtual machine resources and concurrent requests in order to provide better cloud service performance by using fewer virtual machine resources has become a concern for service providers. Aiming at the problem of cloud service performance prediction, this paper proposes a DBN-based cloud service response time prediction method. The method takes response time of cloud service to measure cloud service performance, and establishes a DBN based mapping model between cloud service response time with concurrent requests and virtual machine resources. This paper gives construction process of DBN based cloud service response time mapping model. The cloud service response time is predicted based on the mapping model. Experiments show that the method has high prediction accuracy and stability.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation Program of China (61572116, 61572117, 61502089) and by the Fundamental Research Funds for the Central Universities (N171604009).

REFERENCES

- [1] Mega C, Waizenegger T, Lebutsch D, Schleipen S, Barney J. Dynamic cloud service topology adaption for minimizing resources while meeting performance goals[J]. IBM Journal of Research and Development, 2014, 58(2): 1-10.
- [2] ZHANG Bin, WANG Lin, ZHAO Xiu-tao, ZHANG Chang-sheng. A Novel Modeling Method for Relationships Between Resources and Service Performance [J], Journal of Northeastern University(Natural Science), 2015, 36(6): 773- 776.
- [3] Rao J, Wei Y D, Gong J Y, et al. QoS guarantees and service differentiation for dynamic cloud application[J]. IEEE Transactions on Network and Service Management, 2013, 10(1): 43-55.
- [4] Kundu S, Rangaswami R, Dutta K, et al. Application performance modeling in a virtualized environment[C], proceeding of IEEE 16th International Symposium on High Performance Computer Architecture. Bangalore, 2010: 1-10.
- [5] Lloyd W, Pallickara S, David O, Lyon J, Arabi M, Rojas K. Performance modeling to support multi-tier application deployment to infrastructure-as-a-service clouds[C]. Proceeding of the 2012 IEEE/ACM Fifth International Conference on Utility and Cloud Computing, 2012: 73-80.
- [6] Kousiouris G, Cucinotta T, Varvarigou T. The effects of scheduling, workload type and consolidation scenarios on virtual machine performance and their prediction through optimized artificial neural networks[J]. Journal of Systems and Software, 2011, 84(8): 1270-1291.
- [7] G. E. Hinton, S. Osindero, and Y.-W. The. A fast learning algorithm for deep belief nets[J]. Neural Comput, 2006, 18(7): 1527-1554.
- [8] Wenhao Huang, Guojie Song, Haikun Hong, and Kunqing Xie. Deep Architecture for Traffic Flow Prediction: Deep Belief Networks With Multitask Learning[J], IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, 2014, 15(5): 2191-2201.
- [9] Joe Lemieux, Yuan Ma. Vehicle Speed Prediction using Deep Learning[C], proceeding of 2015 IEEE Vehicle Power and Propulsion Conference, 2015: 1-5.

- [10] Feng Qiu, Bin Zhang, Jun Guo. A deep learning approach for VM workload prediction in the cloud[C], proceeding of 17th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2016: 319-321.
- [11] Yaying Zhang, Guan Huang. Traffic flow prediction model based on deep belief network and genetic algorithm[J], IET Intelligent Transport Systems, 2018, 12(6):533-541.

Yan Gao received the Ph. D degree in Computer Application Technology from Northeastern University, China, in 2006. Since 2008, he has been an Associate professor in Northeastern University, China. His research area includes Cloud Computing, Service Computing.

Bin Zhang received the Ph. D degree in Computer Science Theory from Northeastern University, China, in 1998. Since 2000, he has been a professor in Northeastern University, China. His research area includes Cloud Computing, Service Computing.

Si Wang received the master degree in Computer Technology from Northeastern University, China, in 2017. Her research area is Cloud Computing.

Anxiang Ma received the Ph. D degree in Computer Application Technology from Northeastern University, China, in 2009. Since 2008, she has been a lecture in Northeastern University, China. Her research area includes Cloud Computing, Service Computing.