

Research on the deployment tactics of workloads confliction based on the neural network in cloud computing

Wu Qinlan, Huang Yanmei

School of Internet of Things Engineering, Jiangxi College Of Engineering, JiangXi, 338029, China

Corresponding author's e-mail: wuqinlanjx@163.com

Received 10 September 2013, www.cmnt.lv

Abstract

Aiming at the degrading system performance that busy workloads bring in cloud computing, a resource deployment model based on error back-propagation neural network was proposed to resolve the problems referred to above. A network module is started automatically when the beginning of busy workloads is judged. The prediction of parameter adjustment value is carried out by using pertained network to achieve the purpose of tracking dynamically the changing of underlying resource and outside world task in cloud computing system. The results of simulation in CloudSim prove that the response speed of resource deployment can be improved efficiently by bringing neural network module.

Keywords: deployment tactics; required busy workloads; neural network; resource optimization; cloud computing.

1 Introduction

Cloud computing, or in simpler shorthand just "the cloud", also focuses on maximizing the effectiveness of the shared resources. Cloud resources are usually not only shared by multiple users but are also dynamically reallocated per demand. This can work for allocating resources to users. This approach should maximize the use of computing power thus reducing environmental damage as well since less power, air conditioning, rack space, etc. are required for a variety of functions. With cloud computing, multiple users can access a single server to retrieve and update their data without purchasing licenses for different applications.

The term "moving to cloud" also refers to an organization moving away from a traditional CAPEX model (buy the dedicated hardware and depreciate it over a period of time) to the OPEX model (use a shared cloud infrastructure and pay as one uses it).

Proponents claim that cloud computing allows companies to avoid upfront infrastructure costs, and focus on projects that differentiate their businesses instead of on infrastructure. Proponents also claim that cloud computing allows enterprises to get their applications up and running faster, with improved manageability and less maintenance, and enables IT to more rapidly adjust resources to meet fluctuating and unpredictable business demand. Cloud providers typically use a "pay as you go" model. This can lead to unexpectedly high charges if administrators do not adapt to the cloud pricing model.

A neural network for handwriting recognition is defined by a set of input neurons which may be activated by the pixels of an input image. After being weighted and transformed by a function (determined by the network's designer), the activations of these neurons are then passed on to other neurons. This process is repeated until finally,

an output neuron is activated. This determines which character was read.

Like other machine learning methods – systems that learn from data – neural networks have been used to solve a wide variety of tasks that are hard to solve using ordinary rule-based programming, including computer vision and speech recognition. Computational devices have been created in CMOS, for both biophysical simulation and neuromorphic computing. More recent efforts show promise for creating Nano-devices [1] for very large scale principal components analyses and convolution. If successful, these efforts could usher in a new era of neural computing [2] that is a step beyond digital computing, because it depends on learning rather than programming and because it is fundamentally analog rather than digital even though the first instantiations may in fact be with CMOS digital devices.

Between 2009 and 2012, the recurrent neural networks and deep feed forward neural networks developed in the research group of Jürgen Schmidhuber at the Swiss AI Lab IDSIA have won eight international competitions in pattern recognition and machine learning.[3] For example, multi-dimensional long short term memory (LSTM)[4][5] won three competitions in connected handwriting recognition at the 2009 International Conference on Document Analysis and Recognition (ICDAR), without any prior knowledge about the three different languages to be learned.

Deep learning feed forward networks, such as convolution neural networks, alternate convolution layers and max-pooling layers, topped by several pure classification layers. Fast GPU-based implementations of this approach have won several pattern recognition contests, including the IJCNN 2011 Traffic Sign Recognition Competition [6] and the ISBI 2012 Segmentation of Neuronal Structures in Electron Microscopy Stacks challenge [7]. Such neural networks also were the first artificial pattern recognizers to

achieve human-competitive or even superhuman performance[8] on benchmarks such as traffic sign recognition (IJCNN 2012), or the MNIST handwritten digits problem of Yann LeCun and colleagues at NYU.

2 The basic model of resource deployment based on BW

The core aim of cloud computing is to provide high quality service to the user. The required busy workloads bring unsteady shocks to the cloud computing system which may reduce the users' experience. Existing response model of deployment tactics of the required busy workloads to deal with the problem of the main ideas can be summarized as the following two aspects:

A. Beginning and end judgment for the required busy workloads

The priority task for the model is to judge the beginning and end for the required busy workloads, that is, by introducing load monitoring index I for quantitative analysis of task request quantity and rate of change in each time unit. The Eq. 1 is the basic equation for the beginning and end judgment for the required busy workloads:

$$I = SCV \left(1 + 2 \sum_{k=1}^{\infty} \rho_k \right) \tag{1}$$

When SCV is the squared coefficient of variation which represents the square coefficient for a fixed length requested task; ρ_k is the autocorrelation coefficient, it is used to look for a random variable relationship between the system itself and statistical data. Suppose $\{X_n\}$ is a set of random variable value of sequential system and can be described as Eq. 2, where $n=(0,1,2\dots\infty)$:

$$\rho_k = \frac{E \left[(X_t - \mu^{-1})(X_{t+k} - \mu^{-1}) \right]}{\sigma^2} \tag{2}$$

Existing algorithms is to judge both of the absolute value of the amount requested from the task and rate of change to compare the volume of arriving requests task and the defined required busy workloads at the length of time per unit to see if the definition of an outbreak task requests to meet their specific content, see reference [6]. Such algorithm can determine the whole story outbreak style task requests, at the same time, when the task requests can effectively avoid the amount of fluctuation in the appropriate range; it is mistaken for an outbreak of the type of required busy workloads.

The BP neural network algorithm is shown like this:

- (1) Initial the value.
- (2) Select a vector in random for calculation.
- (3) Calculating the input and output in the first class.

$$S^k = \sum_{i=1}^3 \omega_{ij} X^k - \theta_j \tag{3}$$

$$B^k = f(S^k), \quad j = 1, 2, 3. \tag{4}$$

- (4) Calculating the actual input and output of all neural nodes in output class.

$$L^k = \sum_{j=1}^3 v_j B^k - \gamma_t, \quad t = 1, 2. \tag{5}$$

$$C^k = f(L^k), \quad t = 1, 2. \tag{6}$$

Calculating the error according to the given expecting output.

$$d_t^k = -\frac{\partial E^k}{\partial t_t^k} = \frac{\partial E^k \partial c_t^k}{\partial E^k \partial t_t^k} \tag{7}$$

$$= (y_t^k - c_t^k) f'(t_t^k)$$

B. Design of the strategies of resource deployment

The current mainstream cloud resources deployment strategies are as follows: an improved simulated annealing algorithm proposed by HP Laboratory; First-Come-First-Service strategy used by OpcnNcbula and other open source software; the deployment model based on the multi-cast technology. The common features of these deployment strategies (models) are the performance are unilateral, aimed goal, the pursuit of optimal value. Such deployment strategy task requests in outbreak style environment will face the problem of unknown error in cloud computing systems under normal circumstances, that there was a substantial decline in the performance of an instantaneous moment of the system.

In response to the impact on system performance which is brought by required busy workloads, the current solution is mainly based on improved strategies for the proposed mainstream a (some) according to specific requirements of the users, which can be described in Table I.

TABLE 1 The weight set for testing

input
N, the number of available site.
K, the candidate sites ($1 < K < N$).
I, the state of system (burst or non burst).
The algorithm of Bursty Workloads Allocation
1. if (detect the start of burst)
2. { set K to Nub ; } // Nub is close to N; such as $Nub = 1/2N$.
3. set k to Nu_s ; // Nu_s to be a small value; such as 1.
4. end if
5. analysis all sites S_i ; // $1 < i < N$
6. select $S = \{S_1, S_2, \dots, S_k\}$ //select out the best K sites.
7. select $S' = \text{uniform}(1, K)$ //under the random measure.
8. submit the job to S' .

Form Table 1, it can be seen that algorithm is deployed by adjusting the number of nodes to complete the optimization algorithms for the system which can avoid the decline to the system caused by the arrival of required busy workloads to make cloud computing systems generate local hot spots. But such algorithms are also evident in a lack of points. The Setting mode is too simple. The situation cannot be calculated in real time with the underlying system resources and cannot be combined with variation of cloud computing. The performance needs to be further improved.

3 The deployment model based on the BP neural networks

To solve the problem mentioned above, by introducing error back propagation neural network, the system can equip with adaptive, self-learning ability, to achieve the deployment parameters adjusted dynamically to improve the cloud computing system which is used to deal with the request in response to the performance of required busy workloads. The main adaptive workflow model proposed is shown in Fig. 1.

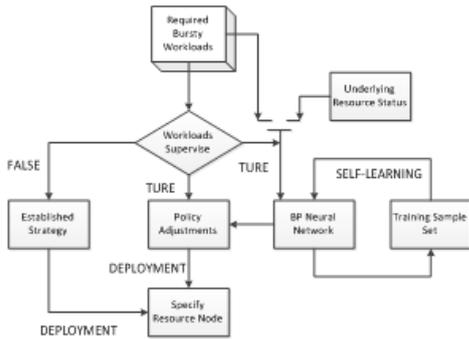
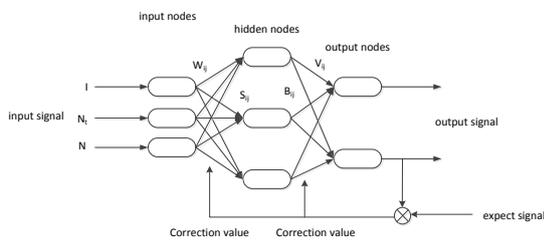


FIGURE 1 The main frame of adaptive workflow based on BP neural network.

STEP 1: Identified neurons. First, the output value of the item types and expectation input of cloud computing systems, types is analysed in accordance with the requirements of BP neural network, specifically with the type of data corresponding to neurons, the input and output neurons.



STEP 2: Training sample set. Through the sample set (selected by the rule set of samples described in the next section), it can be learnt how to adjust the connection weights based on the size and direction of the error. The output mode and network mode is the same as the expected output, so as to enable the prediction of the model of system parameters.

STEP 3: Trigger switch. Changes occur in real-time monitoring by the load monitoring module task requests, when the module detects when an outbreak of type task request arrives, the trigger switches and makes BP neural network retrieval task requests and cloud computing underlying resource-related information.

STEP 4: Quantify BW value. By quantifying the value of an outbreak of type task requests, it can make them more adapted to BP neural network input and can speed up the data disposal process.

STEP 5: Quantify the resource pool indicators. The underlying resource pool information which is collected from

cloud computing system should be standardized. Quantitative analysis of neural networks can change the underlying resources and to provide protection for the next parameter prediction.

STEP 6: Parameter prediction. Parameter prediction is core mission for the oriented model. By model training, input information collection, processing, the deployment model parameters with real-time forecasting resource deployment capabilities can be used to optimize cloud computing system capacity to respond to required busy workloads and to meet the requirement of dynamic and scalable cloud computing system.

4 The analysis and experiment

In cloud computing, three influencing factors can be selected as the input for neural networks, they were listening load index of task requests amount, I , the resources of nodes in cloud computing resource pool, N_t , the available nodes in resource pool at the current time which is noted as N . Where could be analysed as the quantitatively intensity per unit time in a period when tasks requested; N_t can determine whether the cloud computing system has capacity to respond to the required busy workloads; N can be used as the current system performance metrics.

The main purpose of this model is to make resource deployment strategy in response to the required busy workloads based on the BP neural network module for cloud computing system and to optimize the parameters to provide users with a better experience. BP neural network module can adjust the deployment of cloud computing solutions dynamically to meet the changes in the underlying real-time tracking resources. The structure is shown in Fig. 2 as following.

A. Performance measurement and analysis

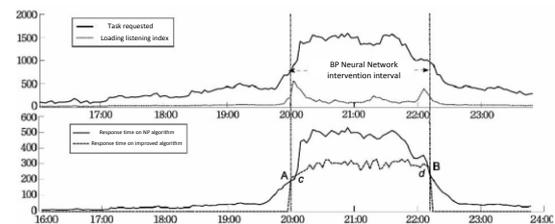


FIGURE 2 The structure diagram of error back propagation in neural network module.

In this paper, the cloud computing simulation software CloudSim is used for the simulation experiments. Compiled by CloudSim extension, the parallel increase BP neural network learning modules is added to achieve the proposed method.

First, the workload listening index is transformed according with the amount of requested tasks in different time and then determine the arrival time of the outbreak of the formula and the end of the task request (in the figure, A, B section). The response time which was found to increase the BP neural network module system response time was significantly shortened, but in [A-c] [d-B] segment, it

increased system performance slightly inferior to the existing network module system. The main reason for this phenomenon is due to the introduction of BP neural network module, which occupies system resources, resulting in a small decrease in system performance.

Task requests in the outbreak style strength is not very intense when brought into the network module to enhance the effectiveness of the system performance to be less than the performance of the module system resources, so it caused a slight increase in the response time of the system. But with the outbreak of the task to increase the strength of the request type, the proportion of their share of system resources continues to decrease system performance and brings its performance gradually. Therefore it can be concluded that the BP neural network module can effectively improve the performance of cloud computing system in response to the required busy workloads and the proposed model is feasible to solve the problem mentioned above.

References

- [1] C'anilf A, Lu Lei, Mi Ning-fang 2010 Fast track for Taming Business and Saving Power in Multi-Tiered Systems *International Telecom Congress(ITC: 22) September, Amsterdam, the Netherlands* 8(2) 520-31
- [2] Tai Jiang-zhe, Meleis W, Zhang Jue-min 2013 Adaptive Resource Allocation for Cloud Computing Environments under Busy Workloads *Northeastern University of Boston USA* 978-87
- [3] Tirado J M, Higuero D, Isaila F 2011 Predictive Data Grouping and Placement Ivor Cloud-based Elastic Server Infrastructures, *11th IEEE/ACM International Symposium on Cluster, Cloud and grid Computing IEEE* 281-94
- [4] U Labs S I 2013 Software system: private cloud computing with open nebula *Private cloud computing with open Nebula* 21-30
- [5] Uranovetter M 1973 The strength of weak ties *American Journal of Sociology*. 78(6) 1360-80
- [6] Huberman B A, Adamic I A 2004 Information Dynamics in the Networked World *Lect. Notes Phys* 650 371-98
- [7] Qin I, Yu J R, Chang I 2009 Keyword search in databases: the power of RDBMS, *SIGMOD Conference* 1 681-94
- [8] Illenberger J, Kowald M, Axhausen K W 2011 Spatially embedded social network from insights into a large-scale snowball sample *The European Physical Journal B-Condensed Matter and Complex Systems* 2 1-13

5 Conclusions

This paper analyses the current response deployment model of the required busy workloads, and summarizes the advantages and disadvantages of existing and subsequently methods and introduces BP neural network algorithm for the problem. From the application of the principles and processes deployment model, its network structure and learning algorithm is designed. A network module is started automatically when the beginning of busy workloads is judged. The prediction of parameter adjustment value is carried out by using pertained network to achieve the purpose of tracking dynamically the changing of underlying resource and outside world task in cloud computing system. The results of simulation in CloudSim prove that the response speed of resource deployment can be improved efficiently by bringing neural network module.

Authors	
	<p>Wu Qinlan, born in May 1982, Xinyu City, Jiangxi Prov, China</p> <p>Current position, grades: Lecturer University studies: Computer science Scientific interest: Cloud computing Experience: From 2006 to 2014, working in Jiangxi college of engineering.</p>
	<p>Huang Yanmei, born in January 1979, Nanchang City, Jiangxi Prov, China</p> <p>Current position, grades: Lecturer University studies: Computer science Scientific interest: Cloud computing Experience: From 2002 to 2014, working in the computer related</p>