Study of the Grid Resource Neural Network Reservation Technology Based on SLA

Xiaolan Xie[*1,2] and Aixue Wang[1]

[1]Institute of Information Science and Engineering, Guilin University of Technology, Guilin 541004, China
[2]Guangxi Key Laboratory of Spatial Information and Geomatics
xie_xiao_lan@foxmail.com, 694573108@qq.com

Abstract

Reasonable resource reservation is a base for providing a good service in the Grid computation. The technology using neural network resource reservation based on service level agreement is proposed in this paper. A framework for the resource reservation using neural network decision system is built and a resource reservation decision control strategy based on neural network is researched. According to the state of grid services, with resource reservation neural network prediction categories for decision-making basis, the algorithm is described and analyzed by the examples, and is verified by the experiments. The simulation results show that the proposed grid resources reservation decision technology is able to predict the reasonable grid resource reservation category value. Combining with the current state of the grid service decision-making out grid can actually offer service level, and the service request has the high response efficiency and the components of the gird service request and response is of certain practical reference value.

Keywords*: Grid, service level agreement (SLA), Neural Network, resource reservation*

1. Introduction

During the process of grid providing services, how to efficiently respond to the service requests of the service requester is the basis for ensuring good quality service. The reasonable resources reservation (resource reservation, RSV) is a precondition for providing good services. Currently, there are some scholars proposing a strategy based on service level agreement (SLA) which is used the static condition constraint condition [1] [2] [3] [4]. In fact, because the character of the services request is dynamic, the resources reservation of the Grid is also dynamic. Therefore, it is very important and significant to study the resources dynamic reservation technology of the Grid.

RSV technology provides a running environment for the tasks Collaboration configuration and tasks performance in the Grid. The demand for resources of the consumption and demand is a dynamically as the resources owner expect to achieve the highest economic return and the best choice is the full load operation of the resources. However, due to the randomness of resource requests, the new services revenue of resource requests may be higher than the committed service revenue, or the new high level SLA services cannot run because of the full load resulting in economic losses. The full load operation of the resources will significantly increase the consumption of resources or even the destruction of resources and increased maintenance costs as well. To ensure the service quality of high level SLA services, gaining higher economic returns and reduce resource consumption costs, the Grid resource reservation techniques must be studied[5][6]. Taking into account the dynamic nature of resource reservation, this paper proposes a SLA-based resource reservation neural network technology. It is on the basis of resource reservation neural network prediction type as decision-making and use the

rate of resource reservation and admission control request as the indicators of accepting and controlling, so as to provide efficient grid services for the users.

2. Resource Reservation Neural Network Decision System Framework

Resources reservation neural network decision system uses the Self-Organizing Feature Map network SOM and error back propagation network BP neural network technology compound. The advantage of this technical is that: when the neural network accepts outside input mode, through Unsupervised Learning method to aotodiscovery inner laws and essential attribute in the samples, self-organizing, self-adapting to changes to the network parameters and structure. The neural network is divided into different response areas, and imitates the real world of the human brain in the organization learning function. The system framework as shown in figure1 below:
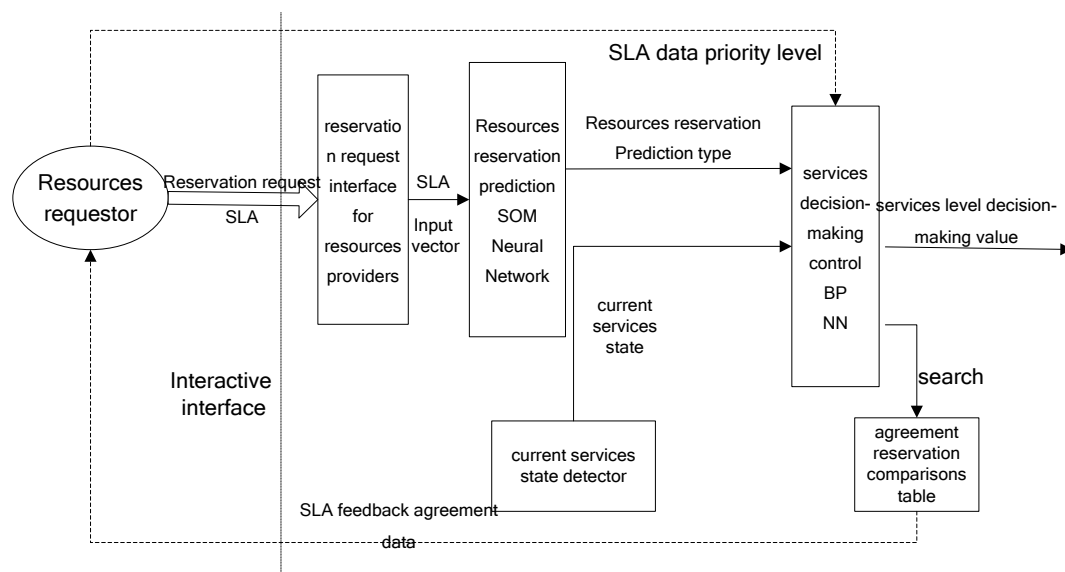
Figure 1. Resource Reservation Neural Network Decision System Framework

First of all, the resources requester submits the SLA data to the provider of resources by reservation request interface. Because the quantity of the input of the possible SLA data combining may be very huge, it is unpractical and unnecessary that a single service resource allocation plan is given to every combination. In order to improve the efficiency of resources reservation, we need to classify the huge quantity of the inputted SLA data combining, and the process of classification is realized by resources reservation prediction neural network. Resources reservation neural network uses the Self-Organizing Feature Map (SOM) neural network, and takes use of offering data in advance to offline training before using, the neural network can predict resources reservation categories after training is completing, then the decision will be sent to the service decision-making neural network to be decided.

The service decision-making neural network uses the type of the forward BP neural network to realize, and it deal with the current state and resources reservation categories, deciding the service level. If the service level meets the request of resources requesters, we can begin to service. Otherwise, we will get the SLA negotiation data similar to SLA request by checking the negotiation of reservation table and feedbacks to the resources requester. The SLA data submitted by the resources requester has the priority of agreement data, even the service level of decision-making cannot meet the request of the resources requester, we can go further negotiation according the priority of each

agreement data. If the resources requester accepts the agreement data then it will do the reservation service, or will not.

3. The Design and the Realizing of The Reserve Reservation Neural Network

The agreement data submitted by the resources requester will become the output of the resources reservation prediction neural network after normalization. During the process of the design, we find that the quantity of the input SLA data combining is very huge, for example, if the service level is 10, and the SLA parameters are 5 parameters which can range from 1 to 1000, so the input vector can be 10×10005=1016 combination, and it is unpractical that a single service resource allocation plan is given to every combination. At the same time the agreement data combination submitted by the resources requester is not the uniform distribution, therefore, it is necessary to make a classification for input vector. According to the specific situation of the grid resources, the specific classification number is given, and the corresponding clustering can be also obtained through the resources reservation prediction neural network.

The prediction Accuracy of the resources reservation prediction neural network is bound up with the number of setting of clustering of neural network. In a certain range, we can choose more clustering number to improve the prediction accuracy, but the number cannot be too big, we should adjust it according the Grid.

SOM neural network is good at assortment, and can classify the input vector, and each output of the neuron corresponds to a different clustering. After the way of Unsupervised Learning method to complete the network training, each neurons in every competitive layers of SOM to compete the opportunity of the response, and there is only one can be the winner from each clustering. In order to make the management of resources more convenient, meanwhile improving the speed of the response in the Grid Service, we use SOM neural network to predict the resources reservation, the output layer can be one dimensional linear array, as it is shown in Figure 2.

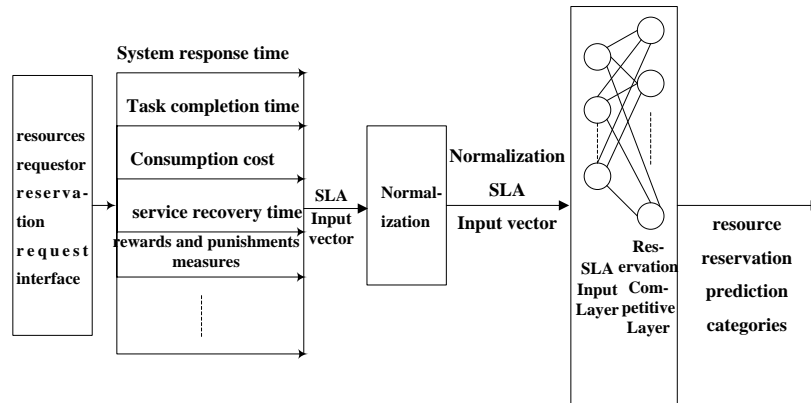The input of neural network is defined like the following:

Figure 2. The Structure of Self Organizing Map（SOM）Neural Network

Supposing the time at t+1, there is a new request, and the level of this request is Q, and Q belong to {1,2,..M}. The SLA data S={X1,X2,..Xn}, and Xi is the parameter values of the level index required by the SLA, there are n parameters altogether. For convenience of describing it, this paper only take one of the most important five parameters: system response time, task completion time, consumption cost, service recovery time and rewards and punishments measures, respectively for X1 ~ X5, that is, S = {X1,X2,X3,X4,X5}.

Set I={Q,S} as the input parameters of resource reservation neural network , that is: I={Q,X1,X2,X3,X4,X5}

In order to facilitate the competitive layer handling, pretreatment to SLA parameters are as follows:

$$Average\ unit\ task\ completion\ time = \frac{Task\ completion\ time}{Bull\ task\ number} \tag{1}$$

$$Average\ unit\ consumption\ \cos t = \frac{Consumption\ \cos t}{Bull\ task\ number} \tag{2}$$

$$Average\ unit\ rewards\ and\ punishment = \frac{Total\ rewards\ and\ punishment}{Bull\ task\ number} \tag{3}$$

Input the vector change interval and normalization, as shown in Table 1.

Table 1. Table Label

	change interval	Normalized formula	Normalized change interval
SLA level Q (grade)	(1,10)	(Q-1)/9	(0,1)
System response time RespT(One-tenth seconds)	(1,200)	(RespT-1)/199	(0,1)
Average unit task completion time PerT(ten minutes)	(1,1440)	(PerT-1)/1439	(0,1)
Average unit consumption cost PerC (Yuan)	(1,2000)	(PerC-1)/1999	(0,1)
service recovery time ResuT(seconds)	(1,600)	(ResuT -1)/599	(0,1)
Average unit rewards and punishment PerG (Yuan)	(1,1000)	(PerG-1)/999	(0,1)

Before use of neural network, network must be trained using the sample data. The more collection of sample data, the more categories accord with the actual situation, the better neural network training effects. Take the offline training first before put Grid into use, give the resource requester's SLA request vector as input with simulation method based on experience. After put into use of the Grid, use the collected real data for grid online training by collecting the actual resource requester's SLA request.

Connections method between input neurons and competition layer neurons is shown as Figure 3[7]. Output layer takes a one-dimensional linear array to improve the convergence speed of training.

The input mode for network is：$P_k = (P_1^k, P_2^k, \ldots, P_i^k, \ldots P_N^k)$ ， $k=1,2,\ldots,q$ ， $i=1,2,\ldots,N$

Competition layer neurons vector is： $A_j = (a_{j1}, a_{j2}, \ldots, a_{jm})$ ， $j=1,2,\ldots,m$

The connect power vectors between competition layer neurons j and input neurons is： $W_j = (w_{j1}, w_{j2}, \ldots, w_{ji}, \ldots w_{jN})$, $i=1,2,\ldots,N$, $j=1,2,\ldots,m$

The training process of neural network adopts the Kohonen learning rules and trains a certain neuron node just only responding to the special input vector P, consequently, it makes neural network be able to identify the vector P.

For example, a group of data that contains 100 training sample data are given as following Table 2(because data is too much, we would not list all data here),data has been sorted by the priority of Q、RespT、PerT、PerC、ResuT、PerG. The normalized data is as following Table 3.

Table 2. Training Sample Data

Q	1	1	1	…	6	6	…	10
RespT	196	196	192	…	63	62	…	3
PerT	1434	1432	1421	…	524	522	…	11
PerC	7	35	46	…	1266	1279	…	1998
ResuT	598	589	572	…	202	200	…	3
PerG	11	18	38	…	709	718	…	971

Table 3. The Normalized Processed Training Sample Data

| Q	0	0	0	…	0.5556	0.5556	…	1.0000
RespT	0.9794	0.9788	0.9577	…	0.3094	0.3045	…	0.0074
PerT	0.9955	0.9939	0.9861	…	0.3631	0.3618	…	0.0068
PerC	0.0025	0.0169	0.0222	…	0.6326	0.6389	…	0.9988
ResuT	0.9965	0.9806	0.9519	…	0.3349	0.3320	…	0.0031
PerG	0.0097	0.0167	0.0365	…	0.7081	0.7172	…	0.9701

In order to facilitate the drawing shows, for training samples data gives a very small number of clustering as a resource reservation prediction categories as output. Take the number of clusters is 20(The actual could reach 103～104), On average, the number of each cluster data point is 5. After trained, the relationship between the SOM neural networks each normalized clustering center and resource reservation prediction categories as shown in Table 4. Convert to the relationship of a reservation request SLA cluster center and categories as shown in Table 5. In the negotiating decision process, this table is used for negotiation reservation table of comparisons. Here, RRPC stands for resource reservation prediction categories.

**Table 4. Resources Reservation Prediction Normalized Category Relation Table**

Q	0.0013	0.0270	…	0.1890	…	0.4508	0.4972	…	0.8466	0.9219	0.9652
RespT	0.9265	0.8981	…	0.7411	…	0.3667	0.3248	…	0.1466	0.0827	0.0520
PerT	0.9629	0.9291	…	0.7698	…	0.3982	0.3666	…	0.1444	0.0915	0.0623
PerC	0.0589	0.0956	…	0.2497	…	0.5727	0.6121	…	0.8384	0.9123	0.9523
ResuT	0.9402	0.9132	…	0.8019	…	0.4330	0.3605	…	0.1437	0.0819	0.0454
PerG	0.0712	0.1109	…	0.3253	…	0.6584	0.6995	…	0.8217	0.8765	0.9153
RRPC	20	19	…	16	…	9	8	…	3	2	1

Table 5. Resources Reservation Prediction Category Relation Table (Negotiation Table Reserved)

Q	1	2	2	3	3	…	6	6	…	9	10	10
RespT	186	180	170	161	150	…	74	66	…	30	17	11
PerT	1386	1338	1265	1201	1116	…	575	529	…	205	131	86
PerC	121	194	308	399	492	…	1146	1225	…	1686	1829	1914
ResuT	564	548	527	511	484	…	261	217	…	86	50	27
PerG	74	113	175	238	320	…	659	700	…	824	879	920
RRPC	20	19	18	17	16	…	9	8	…	3	2	1
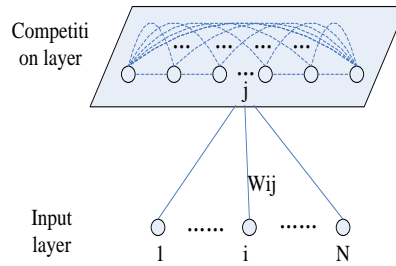
Figure 3. Connections between Input Neurons and Competition Layer Neurons
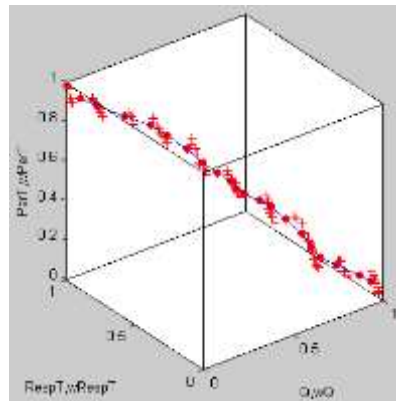
Figure 4. Input and Output Sample Clustering Figure

Because of variables of the input vector more than three, we cannot directly draw the multidimensional sample figure, but can take 3 variables projection to three-dimensional space for observation. Here take the Q, RespT, PerT three parameters projected into three-dimensional space, through Matlab simulation output, as shown in Figure 4.

In Figure 4, the input vector of Q, RespT, PerT three parameters projected onto the three-dimensional space and each point uses "+". The 20 solids ● is the 20 cluster centers of the trained neural network. The lowest right corner of the clustering point corresponds resources reservation prediction category 1, and the upper left of the adjacent clustering point corresponds resources reservation prediction category 2, and so on, the most in the upper left corner of the clustering point corresponds resources reservation prediction 20 categories, the weights in the vector distributes each cluster center of input vector. When the new SLA resource request input vector closes to the vector of a clustering point, input will produce close to one of the weighted input and, thus can get the corresponding clustering vector output. For example, the new SLA resources input vector request for:

I={Q,S}={Q, RespT, PerT, PerC, ResuT, PerG}={ 10,10,100,1910,30,910}

The normalized input vector for:

I={Q,S}={Q, RespT, PerT, PerC, ResuT, PerG}={1.0000, 0.0452,0.0688,0.9550,0.0484,0.9099}

The SLA resources request input vector is close to the clustering center of the resources reservation prediction categories 1 and resource reservation prediction categories value for 1.

4. The Implementation of the Service Level Decision Neural Network

The decision must be made only by integrating resources reservation prediction categories with the current service state of the Grid. The state detector can check the service state of the Grid. The definition of the current service state vectors is following: $S = [\lambda_{hit}, \lambda_{use}, \lambda_{find}, T_{rep}, T_{resp}, \lambda_r]$,and define each data like these:

（1）The rate of successful request : $\lambda_{hit} = \dfrac{Q_{hit}}{Q_{all}} \times 100\%$

Q_{hit} stands for the request number of gaining the result of resource service successful;

Q_{all} stands for the request number of the whole service.

（2）Resource Utilization: $\lambda_{use} = \dfrac{R_{use}}{R_{all}} \times 100\%$

R_{use} stands for the number of Grid resource entities by being used;

R_{all} stands for the whole number of Grid resource entities.

（3）The rate of request matching: $\lambda_{find} = \dfrac{R_{find}}{R_{allFind}} \times 100\%$

R_{find} stands for the number of resource based on request.

$R_{allFind}$ stands for the whole number of resource

（4）mean-time-to repair: $T_{rep} = \dfrac{1}{n} \times \sum\limits_{i=1}^{n} t_i$

n stands for the number of fault; t_j stands for the time of fault repairing.

（5）Average Response Time: $T_{resp} = \dfrac{1}{m} \times \sum\limits_{j=1}^{m} t_j$

m stands for the number of request by users; t_j stands for the response time of request by the j user.

（6）The rate of resources reservation: $\lambda_r = \dfrac{R_{leave}}{R} \times 100\%$

R_{leave} stands for the resources surplus quantity; R stands for resources capacity.

The arithmetical unit used by R_{leave} and R is the unit resources. λ_r stands for the parameter designed by Grid management personnel based on the current state of the Grid.

Service level decision-making neural network uses BP neural network to realize. Because part of the input of the network is from the output of the resources reservation prediction neural network, therefore, training data must be provided for training after the completion of resources prediction neural network training. The topology of the BP neural network is composed of input layer, hidden layer, output layer three parts, and the concrete parameters in each layer for:

Input layers: the input vector includes resources reservation prediction categories and Grid current service status. Neural network input node number is 7, namely input layer 7 neurons. The input vector for: BP$_{in}$={ Q$_{in}$, λ_{hit} , λ_{use} , λ_{find} , T_{rep} , T_{resp} , λ_r }

Output layer: output layer unit number for identification of the species number of the decision level is 10.

Hidden layer: the bigger the total number of the hidden layers, the stronger the ability of the network, but "over-fitting" may appear , and the error rate of the testing samples can be high to the point where it is unacceptable. Conversely, if the number of hidden layers unit is too small, it also can appear that the network fitting data is not good and the error rate is high. There is not definitive formula to determine the number of hidden layers unit. The number of hidden layer unit is determined by the empirical formula:

$$n = sqrt(d + k) + m \qquad (4)$$

Among them, the d is the input vector dimension 7, k is the output unit number 10, m is the constant between 0 and 10 , here m is 5, and calculate and get n = 9.

The BP neural network back propagation learning rule is based on the use of the traditional gradient descent algorithm. Considering the traditional BP neural network used very extensive, so the related formula, specific training method and formulas can refer to the neural network reference foundation books.

The output of the service level decision neural network is the service level according to the current SLA. If the service level meets the request of the resources requester, the service can begin; otherwise the negotiation data will feedback to the requester for further negotiation.

Negotiation data is generated by querying the table. According to the Table 5 resources reservation prediction category relation table, indexed in accordance with the priority of the resource requester submitting SLA data included in the negotiation data on Table 5, finding a few lines of the most proximal data to service levels back to the requester for further negotiation. If negotiation success, it can undertake service, otherwise not provide service. Negotiation process is shown in Figure 5.
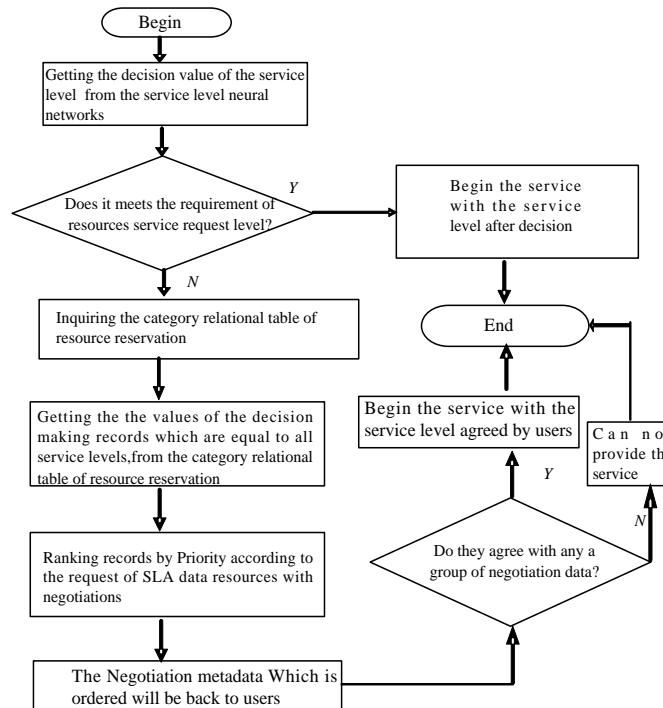
5. The Example of the Decision for Resources Reservation

The current state of a network service is given, such as the Table 6.

Table 6. The Current State of a Network Service

λ_{hit}	λ_{use}	λ_{find}	T_{rep}	T_{resp}	λ_r
0.89	0.92	0.87	70s	300ms	0.70

Given a resource reservation request SLA data such as shown in the first line, and the second line of Table 7. is normalized data.

Figure 5. SLA Processing Diagram

Table 7. Resources Reservation Request SLA Data

Q	RespT	PerT	PerC	ResuT	PerG
3	148	1104	505	480	329
0.2222	0.7387	0.7665	0.2521	0.7997	0.3283

Use the trained neural network decision-making system to make decision of the SLA data, resources reservation prediction output categories is 16. After service level decision-making neural network, the service level is level 3, and it shows that the system can provide service immediately.

A resource reservation request SLA data is given as the first line on the table 8 and the second line is Normalized data.

Table 8. Resources Reservation Request SLA Data

Q	RespT	PerT	PerC	ResuT	PerG
9	31	209	1677	88	822
0.8889	0.1508	0.1445	0.8384	0.1452	0.8218

Use trained neural network decision-making system to make the SLA data decision, resources reservation prediction categories output 3. After service level decision-making neural network, the service level is level 6, and shows that the system could not immediately provide 9 levels of service. Consult table 5 we can know, there were two groups of the service level Q for level 6, and the negotiation data is as shown in table 9. According to the priority data which is belong to the negotiation data attached by SLA data submitted by resource requester, two groups of data are indexed and feedback to users according to priority order from high to low arrangement. If users agree to any group of negotiation data, the service can be done immediately, otherwise can't provide services.

Table 9. Service Negotiation SLA Data

Q	6	6
RespT	74	66
PerT	575	529
PerC	1146	1225
ResuT	261	217
PerG	659	700

6. Summary

Wherever This paper puts forward the SLA based on grid resources reservation decision neural network technology, in the resources reservation neural network based on the framework decision system, analyzes the SOM neural network based on the resource reserve method of predicting and based on the BP neural network service level decision-making basic idea, according to the state grid services, with resources reservation neural network prediction categories for decision-making basis, the algorithm was described and arithmetic analysis by example, and was verified. Through the Matlab simulation results show that the proposed grid resources reservation decision technology has good application potential, to be able to predict the reasonable grid resources reservation category value, and combining with the current state of the grid service decision-making out grid can be actually offers service level, to the service request has the high response efficiency and

service request and response in a grid of the construction of the components is of certain practical reference value.

Acknowledgments

References

[1] X. –L. Xie, R.-K. Liu, G.-Q. Zhou and J.-S. Ni, "Research Of Job Scheduling With Cloud Based On Trust Mechanism And SFLA", International Journal of Grid and Distributed Computing, vol. 1, no. 8, **(2015)**, pp. 2005-4262.

[2] S. Bin, L. L. Lan and Y. Tao, "SLA-based resource management model in manufacturing grid.journal of Computer Applications", vol. 2, no. 26, **(2006)**, pp. 512-514.

[3] L. L. Lan, Y. Tao and S. H. Ya, "Service level agreement in quality of service management system in manufacturing grid", Computer Integrated Manufacturing Systems-CIMS, vol. 8, no. 12, **(2006)**, pp. 1322-132.

[4] L. Bo, "Research on Job Scheduling Algorithms with Resource Reservations for Grid Computing", Huazhong university of science and technology, Ph.D. Thesis, **(2006)**.

[5] Y. C. Xing and L. Z. Heng, "A Unified Resource Reservation Policy. Computer Engineering and Applications", vol. 24, no. 41, **(2005)**, pp. 144-146.

[6] S. Kumar, "Neural netwoks", Tsinghua University Press, **(2003)**.

[7] W. Xin, "MATLAB simulation and application of neural network", Science press, **(2003)**.

Authors

Xiaolan Xie, PhD, Professor, School of information science and engineering, Guilin University of Technology, research direction: Cloud computing, Grid computing, Intelligent decision system.

Aixue Wang, Assistant professor, School of information science and engineering, Guilin University of Technology, research direction: Artificial Intelligent.