DAV International Journal of Science

DATA REPLICATION IN GRID BY USING DATA MINING ON THE BASIS OF LEARNING

Mortaza Abbaszadeh
Department of Computer Engineering, Ilkhchi Branch, Islamic Azad University, Ilkhchi, Iran
Email: abbaszadeh@iauil.ac.ir

ABSTRACT
Data grid is considered as a method to manage distributed data with high volume and shared data resources. Distributed storage resources are used to request large scientific computations and to analyze data. In order to manage data, there is a key method called data replication. This strategy has been taken into account to increase access performance and reliability in data grid systems. In order to solve this problem, there are some methods such as mathematic method used to find the number of replication, and it requires some fixed parameters. These parameters are fixed, but sometimes they are not available, and this is one of the disadvantages of mathematic method. One of the problems of cache memory method is that it lacks compatibility. Therefore, in this paper, the method of data replication and transferring Master Engine with Deploy Methods having minimum cost have been considered. In addition, in this research, a decision making tree has been considered in data mining so that most widely used data can be determined in replication. This study is on the basis of inductive learning. The obtained results decrease time of data accessing to 28%.

KEYWORDS: entropy, data grid, data replication, decision making tree.

INTRODUCTION

One of the important grid technologies used to accelerate data access is data replication considered as an effective method to obtain higher performance and more accessibility by storing various copies in different locations. Data replication is used to reduce bandwidth consumption and access cost in grid environment (Tang, 2009); (Li, 2009) (Lamehamedi, 2002).

The advantages of data replication in grid by using replication method are as follows:

- Increasing reliability
- Increasing performance
- Decreasing time of data access
- Minimizing bandwidth consumption
- Replicate Multiple Master Engine With Deploy Method
- Remote Differential Compression
- Deploy Method
- Schedule Policy
- Minimum Cost
- Time saving
- Enforce Security
- Enforce Fail Tolerance
- Increasing the quality of service

Scheduling Models

Various scheduling models in grid are as follows:

- Centralized scheduling model
- Decentralized scheduling model
- Hierarchical scheduling model
- In centralized scheduling model, there is only one central coordinator coordinating the requests as a manager. In this model, all machines are same (Lin, 2008). In decentralized scheduling model, there is no central scheduler. Instead, several local schedulers are used, and they interact with each other to assign the tasks to machines. In hierarchical scheduling model, unlike centralized scheduling model, they manage the machines

themselves. If there are many machines, then it will be managed with difficulty, so it cannot be developed. The advantage of hierarchical architecture is that the central and local coordinator can separately apply policies. The model proposed in this paper is a hierarchical model.

The procedures of data mining

Data mining is an analytical process used to search data. The findings are validated by using the patterns. The main purpose of data mining is prediction. Data mining process involves three stages:

1) Initial search and exploration
2) Making a model or detecting a pattern by considering validation
3) Utilization

In the first stage, data are usually prepared, and it may involve data clearing, data conversion, and selecting subsets of records. This model requires simple prediction models or geographical and statistical models to detect considered variables and to determine models complexity, so that they can be used in next stage. In the second stage, various models are investigated, and the best model is selected. Different techniques have been developed to reach the purpose. For this purpose, different models are used in the same data set to compare their performance. Then, the model with the best performance is selected. Stage three is the last stage. In this stage, the model selected in the previous stage is used in new data so that expected output can be presented. Data mining has been generalized the information management tool to make decision (Domenici, 2004; Chakrabarti, 2004; Chakrabarti, 2004).

Various learning methods

It is clear that machine learning, as a combination of statistics and artificial intelligence, is an effective research field, and in this regard, various algorithms and problems have been presented to solve them. Different learning methods are as follows:

1. Supervised learning
2. Unsupervised learning

Supervised learning is used to satisfy unknown dependencies in input/output of known samples. In unsupervised learning, the instructor is detected, and it requests the learner to form the model. Decision making trees are an example of supervised learning.

Table 1. Diagram of comparing the modes presented in this paper with conventional studies and researches.

Items	General Method	Our Method
Time Saving	NO	YES
Cost Saving	NO	YES
Security	NO	YES
Fail Tolerance	YES	YES

In table1, diagram of comparing the modes presented in this paper with conventional studies and researches has been presented, and the modes considered in this study have been demonstrated.

Fuzzy decision tree

A model combining of case base clustered process and Fuzzy decision tree has been developed to classify medical data. At first, the process of stepwise regression has been used to select the important factors in input set. In the next stage, the process of weighted clustering has been accommodated to divide and classify the cases into smaller cases. All homogeneous cases are classified in a class. Therefore, these data clearly show a reaction to diagnose diseases. Finally, genetic algorithm is implemented for developing Fuzzy sections of each factor to obtain the best Fuzzy decision tree for each case. The performance of these methods increases to 99.5% in terms of disease diagnosis and drug prescription. Also it increases to 85% in liver dysfunctions. At last, a set of Fuzzy tree rules are created for each cluster.

MATERIALS AND METHODS

In this paper, architecture has been hierarchically proposed. The proposed architecture has been shown in figure 1. In hierarchical model, there are some local coordinators and a central coordinator. Since this architecture has been

proposed in clustering form, cluster heads have been considered as the local coordinator. Central and local coordinators interact with each other to dedicate data.

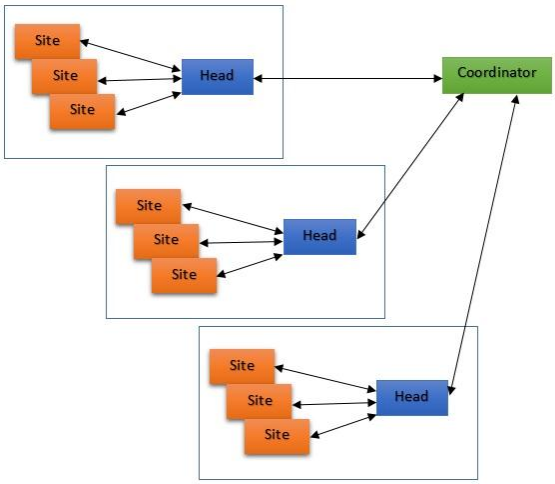

Figure 1. The proposed hierarchical architecture

Each cluster involving many computers and workstations is relatively strong, and they have connected to each other through a high speed dedicated network. A distributed operating system is usually placed on each cluster to manage the available computers and all resources. The computers of a cluster are managed under the management of a central node called cluster head. This node manages all data in a cluster by knowing the status of all nodes. Also, cluster head monitors the performance of all cluster machines so that it can be assured that the machines with more replications work properly. In order to obtain a comprehensive vision of whole system, it must be possible to remotely monitor all parameters required in all processor machines. When there are several sites in each cluster, each site involves thousands of nodes (Figure 2). There is a table in each cluster that involves the following fields:

- Data attribute shows data characteristics and attributes.
- The type of data determines the type of data, and it may be remote or local.
- The access number determines time of data dedication

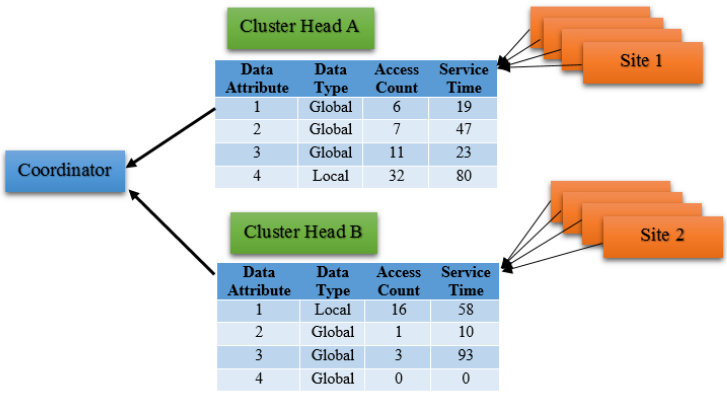

Figure 2. Available tables in clusters

RESULTS AND DISCUSSION

There is a coordinator in the center in the proposed hierarchical model. The coordinator is the manager of data replication, and defines a system. It can create, detect, manage and update the replications for virtual organization.
The tasks of a coordinator are as follows:

- Creating a connection for sending and receiving the message
- Computing the number of requests
- Computing duration of file transferring

- Detecting data type and transferring compressed files
- Transferring data to a cluster having more requests
- Dedicating TTL Time for sent packet to prevent high volume data transferring in the network to manage bandwidth
- Preventing replicated requests of DDOs
- The Best Solution For Secure Data Transfer by Backbone Structure
- Virtual Private Network Solution Using Subletting And Or Superheating
- Broadcast Blocking Service Management
- Prevent Broadcast Packet Data To Wireless Or Local Network With Isa Server Consol
- Optimize the network with Vlan Category To Best Quality And Hi Performance
- Living out a system that does not respond to live server packet And Inform To All Network
- Prevent foreign request by the management Number network card
- Observing and following requests
- Determining the type and number of replicated data by data mining

The central coordinator has some tables in this part:

Table 3. The available table in coordinator

Data attribute	Bandwidth	Replication cost	Data size	priority	Access number	Service time	The number of computed replication
1	medium	high	large	high	52	36	6
2	high	medium	medium	low	42	29	2
3	low	low	Large	low	30	18	5
4	medium	low	small	high	57	17	3

Finally, the coordinator stores transmission status and cost. The longer routes have more costs. Communication cost is computed by using hope count.

1) The first stage of data mining, decision making tree

There are various algorithms to make decision tree, and greedy algorithm has been taken into account in this paper. In this method, a trait is followed, and the samples are ideally classified. The trait classifying the samples as well as possible is called characteristic. Entropy 13 can be used in this part, and it shows the impurity degree. Entropy selects the best separator trait by using gain 14. Equation (1) (Mohammad Khanli, 2010) demonstrates the entropy computation formula:

$$Entropy(X) = -\sum_{x} P(x) \log P(x) \quad (1)$$

P(x) is X variable possibility. Log (px) is the possibility algorithm of x, and this algorithm is considered in base 2. Some points must be considered in this part:

- The objective function should involve limited discrete range
- The objective function must be stated by using a logical formula involving $\vee$ or $\wedge$.
- Learning data may have some errors.

Equation (2) (Mohammad Khanli, 2010) shows that how gain can be computed. This equation determines that which trait must be placed in the root of decision tree.

$$Gain\,(X, A) = Entropy(x) - \sum \frac{|S_V|}{|S|} Entropy\,(x_V) \quad (2)$$

The $|S|$ demonstrates the whole number of records, and $|S_V|$ shows the number of records belonging to class A.

Comparing entropy and gain

Entropy of the traits in table 3 is computed.

$$Entropy(S) = -\frac{8}{14} Log \frac{8}{14} - \frac{6}{14} Log \frac{6}{14} = 0.99 \quad (3)$$

The results show that entropy value is computed as 0.99. The gain of all traits in table 3 is computed so that a proper trait is selected for root.

$$Gain\,(S, Access\_number) = 0.99 - \frac{5}{14} Entropy\,(A_{Low}) - \frac{4}{14} Entropy\,(A_{Medium}) - \frac{5}{14} Entropy\,(A_{High}) = 0.24 \quad (4)$$

$$Gain\,(S, \Pr iority) = 0.99 - \frac{5}{14} Entropy\,(P_{Low}) - \frac{5}{14} Entropy\,(P_{High}) = 0.21$$

$$Gain\,(S, Service\_Time) = 0.99 - \frac{5}{14} Entropy\,(S_{Low}) - \frac{4}{14} Entropy\,(S_{Medium}) - \frac{5}{14} Entropy\,(S_{High}) = 0.24$$

$$Gain\,(S, Size\ \ of\ \ data) = 0.99 - \frac{5}{14} Entropy\,(S_{Low}) - \frac{5}{14} Entropy\,(S_{High}) = 0.13$$

$$Gain\,(S, Local) = 0.99 - \frac{5}{14} Entropy\,(L_{Low}) - \frac{5}{14} Entropy\,(L_{High}) = 0.11$$

The obtained results show that the trait of access number has the highest gain, so it must be located in decision tree. Gain is computed in each level of decision tree to determine that which trait is located in each level in other stages.

The second stage of data mining, creating decision making tree

Now decision making tree can be drawn. If all the traits in a branch of tree belong to a class, then dividing that branch is not necessary.

If all the traits do not belong to a class in leaves, voting policy can be used. For example, if the number of a class is more than other classes, then that class will be selected.

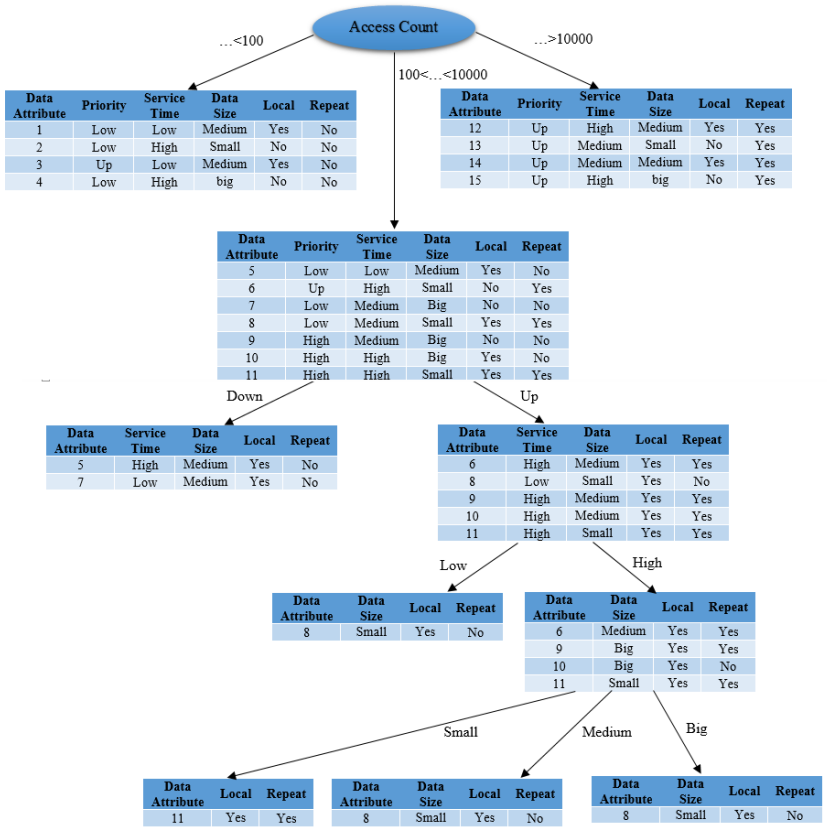

Figure 3: decision making tree

Figure 3 shows the hierarchical decision making tree in this paper. Decision trees are divided in each level on the basis of a trait. The first stage of decision making tree is selecting the root node showing all traits available in learning system. In next stage, if all traits belong to a class, dividing them is not required, and the tree will be complete. If traits do not belong to a trait, the process of dividing continues for each node until the traits belong to a class. In traits mentioned in this paper, root is divided into three groups, and two branches involve same classes, while the middle group does not involve the same class. The trait of next level is selected through obtaining the gain of other traits.

The third stage of data mining, presenting a rule

Bayesian possibilities were computed in the previous stage. Now maximum possibility affecting replication can be found, and the following rules can be presented:

Rule 1: If (access_number > 10000) then replicate=yes
Rule 2: If (access_number < 100) then replicate=no
Rule 3: If (100>access_number > 10000 and priority=high) then replicate=yes
Rule 4: If (100>access_number > 10000 and priority=low) then replicate=no
Rule 5: If (100>access_number > 10000 and size of data=high) then replicate=no
Rule 6: If (100>access_number > 10000 and priority=high and Service of Time =high and size of data=high) then replicate=no
Rule 7: If (100>access_number > 10000 and priority=high and Service of Time =high and size of data=low) then replicate=yes
Rule 8: If (100>access_number > 10000 and priority=high and Service of Time =high and size of data=medium) then replicate=yes
Rule 9: If (100>access_number > 10000 and priority=low and Service of Time=high) then replicate=yes
Rule 10: If (100>access_number > 10000 and priority=low and Service of Time=low) then replicate=no

For instance, the first rule shows that if access number is more than 10000, then replication must be performed. The third rule demonstrates that if access number is more than 10000, and data priority is low, then replication must not be performed.

Evaluation and simulation

Evaluation of this algorithm has been presented by comparing various modes. In diagram figure 4, comparison of three modes, involving 4 clusters, 8 clusters and 12 clusters, has been presented in two models with repetition and without repetition. When the proposed environment involves 4 clusters, service time reduces to 25%. The proposed environment with 8 and 12 clusters respectively show 28% and 37% reduction of service time.

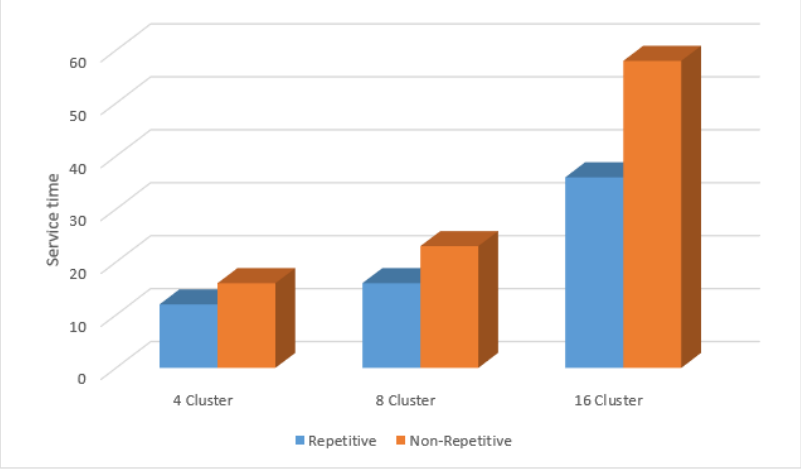

Figure 4. Comparing three modes diagram

CONCLUSION

Data grid manages geographically distributed resources for large scales of various applications creating large data set (Lin, 2008; Jianfeng, 2009; Mansouri, 2008). There is a key method called Advanced Replication in Grid Environment to manage data. In this method, data replication is a strategy to increase access performance and reliability in data grid

systems. In this paper, the proposed architecture model is a hierarchical model, and environment has been considered as clustering. This algorithm has been simulated without repetition and with repetition. The results of this evaluation show that minimum service time is required when LALR algorithm is taken into account in evaluation.

REFERENCES

Chakrabarti A R. S. (2004). Integration of Scheduling and Replication in Data Grids. *Springer- Verlag Berlin Heidelberg.*

Domenici A F. G. H. K. (2004). Replica Consistency in A Data Grid. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 534: 24–28.

Lamehamedi H. B. Z. (2002). Data Replication Strategies in Grid Environments. IEEE Computer Society Press, Bejing, China.

Li J. (2009). Fast Parallel File Replication in Data Grid. IITA International Conference On Services Science Science, IEE.

Mohammad Khanli L F. B. M. n. (2010). A hybrid model combining decision tree and Bayesian network for data replication In Grid environment. Journal of telecommunications, 3: 55-60.

Tang M, B. C. X. (2009). A Replication Strategy in Data Grid Environment. IEEE Computer Society Press, Bejing, China.

Mansouri.Y, G. M. ,. S. M. and M., S., (2008). Optimal Number Of Replicas In Data Grid Environment. *Distributed Framework and Applications, DFmA First International Conference.* pp. 96 - 101 .

Lin.Y, J. P., (2008). A List-Based Strategy for Optimal Replica Placement in Data Grid Systems. *37th Int. Conf. Parallel Processing IEEE.*

Jianfeng.Z, Q. Z. Z., (2009). A Duplicate-Aware Data Replication. In procceding of "Frontier of Computer Science and Technology, FCST '08", Nagasahi, pp. 112 - 117.