Smart Replica Selection for Data Grids using Rough Set Approximations (RSDG)

Rafah M. Almuttairi, Rajeev Wankar, Atul Negi, C. R. Rao.
Department of Computer and Information Sciences, University of Hyderabad,GridLab
rafahmohammed@gmail.com, {wankarcs, atulcs, crrsm}@uohyd.ernet.in

Abstract: *The best replica selection problem is one of the important aspects of data management strategy of data grid infrastructure. Recently, rough set theory has emerged as a powerful tool for problems that require making optimal choice amongst a large enumerated set of options. In this paper, we propose a new replica selection strategy using a grey-based rough set approach. Here first the rough set theory is used to nominate a number of replicas, (alternatives of ideal replicas) by lower approximation of rough set theory. Next, linguistic variables are used to represent the attributes values of the resources (files) in rough set decision table to get a precise selection cause, some attribute values like security and availability need to be decided by linguistic variables (grey numbers) since the replica mangers' judgments on attribute often cannot be estimated by the exact numerical values (integer values). The best replica site is decided by grey relational analysis based on a grey number. Our results show an improved performance, compared to the previous work in this area.*

Keywords: *Data Grid, Replica Selection Strategies, Rough Set theory, Lower and Upper approximation.*

1 Introduction:

Many scientific disciplines such as global climate change, high energy physics and computational genomics, generate large volumes of data on a terabyte scale annually. These huge amounts of data are shared among the researchers around the world. Data is replicated at several replica sites across the grid to avoid a single site to be flooded by requests. Therefore the goal is to minimize latency of data requests from a single organization. Data replication is the solution to this problem, where identical replicas of the same data are produced and stored at the different distributed nodes [1].

Selecting one specific replica site from many sites is an important and critical decision because it affects the total execution job time. It is generally called as a *Replica Selection Decision* [13]. The best replica selection is a multi attribute decision making problem, because each replica has its own capabilities, characteristics and values of attributes. At the same time every user has their own preference on attributes. Replica Sites $RS = \{S_1,S_2,...,S_M\}$, where S_{ij}, S_i *is a vector of attributes j where* $i=1,...,M$. M represents number of replicas, where $j =1,...,N$, N represents number of attributes. On the other hand a user request R_h is a vector of Q attributes. R_h, $h=1,2,...,Q$. Therefore the replica selection problem can be transformed to a nearest match problem between R_h and S_{ij}. It is well known that the nearest match problem is solved using approaches such as k-means, artificial neural networks [5] etc. In this work we differ from the previous literature in adapting powerful matching technique made possible due to rough set theory.

Let us consider an attribute like for instance *Security* with different linguistic values like, *Z={None, Low, Medium, Adequate, High}*, and let us say that *Replica Site-1* announced a *Medium Security* as its security level setting, on the other hand the user asks for a very high level of security by *Request-1*. The request is being processed by the *Replica Managers (RMs),* in a way that *RMs* always express their preferences on attributes of replicas according to the requirement of data grid user/application. In another words *RMs* rank the available replicas (files) to choose the best one to satisfy the users' requirements. Therefore this feature can be effectively used for ranking or selecting the most desirable replicas, but unfortunately the *RM*s judgment is often uncertain and cannot be estimated by the exact numerical value. Thus the replica selection problem has many uncertainties and becomes more difficult. Dynamic nature of the problem due to varying system and network load conditions added difficulty aspect to the decision making.

In our work the rough set theory [6, 7, 8] is adopted to deal with the selection problem under uncertainty [12]. To do that we use grey system theory with rough set theory to make the attribute values known precisely. So in the decision table of rough set theory, the attribute values must be represented precisely using upper and lower limits values [10, 11]. In another words the attribute values should not be integer values. This is required since the rank of the alternatives of ideal replicas will be decided by the lower approximation.

The replica selection problem has been investigated by many researchers but the closest work to our study is published by *Jaradat et al.* [5].This work has its own drawback and in related work in *Section 5* we clearly explain our way to overcome this drawback.

Here in our work, the replica selection problem is addressed as an important decision to guarantee efficiency and to ensure the satisfaction of the grid users, providing them with the best matching between available replicas and their requirements. To reach this aim, important attributes namely, *availability, security, distance, bandwidth* and *cost* of replica for each site should be utilized in the selection process.

Our work procedure is shown in brief by the following four stages: First, the attribute values of decision table for all alternatives are decided by linguistic variables that can be expressed in grey number. Second, ideal replicas are decided by the lower approximation of rough set theory. Third, the most ideal supplier is decided by the grey relational analysis based on the grey number. Finally, an example of suppliers'

978-0-7695-4254-6/10 $26.00 © 2010 IEEE
DOI 10.1109/CICN.2010.94

selection problem is used to illustrate the proposed approach and the experimental result shows its effectiveness.

The rest of the paper is organized as follows: Section 2 presents the preliminary concepts of both the grey system theory and grey-based rough set. Section 3 introduces our proposed algorithm of replica selection strategy using grey-based rough set. In Section 4, the application and analysis of the proposed approach are shown by an example of replicas selection. Section 5 summarizes the related work. Section 6 includes the simulation. We conclude in Section 7.

2 Preliminary Concepts

In this section we presented the background concepts which are used in our strategy

2.1 Rough Set Theory

Rough Set theory is proposed by *Pawlak* [6,7] as an extension of conventional set theory that supports approximations in decision making.

Definitions and Notation:

Definition 1. Let *U* is a non-empty finite set of objects called universe. And let *A* is a non-empty finite set of attribute. Let *R* be an equivalence relation on *U*.

$$IND_{IS}(R) = \{(x_1, x_2) \in U^2 \mid \forall a \in R, a(x_1) = a(x_2)\} \quad (1)$$

where $IND_{IS}(R)$ is called the *R-indiscernibility relation.*

$$(x_1, x_2) \in IND_{IS}(R),$$

for any subset X□U , the pair T=(U,A) is called approximation space. The two subsets:

$$\underline{\underline{R}}X = \{x \in U \mid [x]_R \subseteq X\}$$
$$\overline{R}X = \{x \in U \mid [x]_R \cap X \neq \phi\} \quad (2)$$

are called *R-lower* and *R-upper* approximations of *X*, *respectively* $R(S) = \langle \underline{R}S, \overline{R}S \rangle$is called the rough set of *X* in *T*. The rough set *R(X)* denotes the description of *X* under the present knowledge, i.e., the classification of *U*.

We *use POSR(X)* =$\underline{R}X$ *to* denote *R-positive regi*on of *X*, NEGR(X) =U-$\overline{R}X$ to denote *R-negative region* of *X*, and $RN_R(X) = \overline{R}X - \underline{R}X$ to denote the *R-borderline region* of *X*. The positive region is the collection of those objects which can be classified with full certainty as members of the set *X*, using knowledge *R*. The negative region is the collection of objects which can be determined without any ambiguity, employing knowledge *R*, that they do not belong to set *X*. [9]

2.2 Grey-Based Rough Set

Grey system theory [11], originally developed by *Deng* in 1982, has become a very efficient method to solve uncertainty problems under discrete data and incomplete information.

Definition 2. *A grey system is defined as a system containing uncertain information presented by grey number and grey variables.*

Definition 3. *Let X be the universal set. Then a grey set G of X is defined by two mappings which are the upper and lower membership functions in G respectively. The scale of Grey attribute example declared in Table 1, where, x* $\in X$ *, X=R, (R:Real number set).*

$$\begin{Bmatrix} \overline{\mu_G}(x): x \rightarrow [0,1] \\ \underline{\mu_G}(x): x \rightarrow [0,1] \end{Bmatrix} \quad (3)$$

Definition 4. *A grey number is one of which the exact value is unknown, while the upper and/or the lower limits can be estimated. Generally grey number is written as* $\otimes x = (\otimes x = x \mid_{\underline{\mu}}^{\overline{\mu}})$

Definition 5. *If the lower and upper limits of x can be estimated then v is defined as interval grey number.*

$$\otimes x = [\underline{x}, \overline{x}] \quad (4)$$

Definition 6.*The Euclidean grey space distance for grey numbers v_1 and v_2 is defined as [2]*

$$D(x_1, x_2) = \sqrt{(\underline{x_1} - \underline{x_2})^2 + (\overline{x_1} - \overline{x_2})^2} \quad (5)$$

To measure the relationship between two sequences by calculating their correlative degrees which is called grey relational grade (GRG), a grey relational analysis (GRA) tool of quantitative analysis has been used [10].

Consider we have two reference sequences of grey numbers *X_0={x_0(1), x_0(2), x_0(3)… x_0(N)}* and *X_i={x_i(1), x_i(2), x_i(3)… x_i(N)}, i=1,2,…,M: number of replicas sites*. Where *x_i(k)* represents the k^{th} attribute in x_i , *k=1,2,…,N: Number of attributes of each site*. The GRG *(Γ_{0k})* between each comparative sequence X_i and the reference sequence X_0 at the k^{th} attribute is calculated as [11,15]

$$\Gamma_{0k} = \frac{1}{M} \sum_{i=1}^{M} \left(\frac{\Delta_{\max} - \Delta_{0i}(i)}{\Delta_{\max} - \Delta_{\min}} \right) \quad (6)$$

Where, $$\Delta_{\max} = \max [\max_{\forall_i, \forall_k} D(x_0(k), x_i(k))] \quad (7)$$

$$\Delta_{\min} = \min [\min_{\forall_i, \forall_k} D(x_0(k), x_i(k))] \quad (8)$$

$$\Delta_{0i}(k) = D(x_0(k), x_i(k)) \quad (9)$$

Where, Γ_{0k} represents the degree of relation between each comparative sequence and the reference sequence. The higher degree of relation means that the comparative sequence is more similar to the reference sequence than comparative sequences.

Table 1. The scale of grey attributes from 1-10

Scale	Grey *Values*
Very Low(VL)	[0,1]
Low (L)	[1,3]
Medium (ML)	[3,4]
Fair (F)	[4,5]
Medium Good (MG)	[5,6]
Good (G)	[6,9]
Very Good (VG)	[9,10]

3 Rough Set Replica Selection Strategy for Data Grid (RSDG)

In this section we explain a brief definition for data grid components to give a clear picture where our strategy adds. In our approach a new centralized and decentralized replica selection strategy using rough set theory (*RSDG*) is being proposed. Our new approach considers the *QoS* of the data grid sites, whereas *QoS* itself is considered as a combination of multiple attributes like *Security*, *Availability*, *Cost* and *Time*. The *RSDG* strategy can utilize many existing successful data grid core services, such us *Replica location Service (RLS)* and *NWS/Ipref* [14] as shown in Figure1. *RLS* provides the *Physical File locations (PF)* for all available *Logical Files names (LF)* and *NWS* provides information about the network. The *RSDG* selects the best site location which houses the required replica. In this context, the best site is the site that provides the highest combined security and availability as well as the lowest cost and possible response time between local site (*CS*) and the remote site (*RS*) that houses the required replica. Henceforth, we use the term "best replica" to express the highest level of QoS- for both the replica and the site which houses this replica.

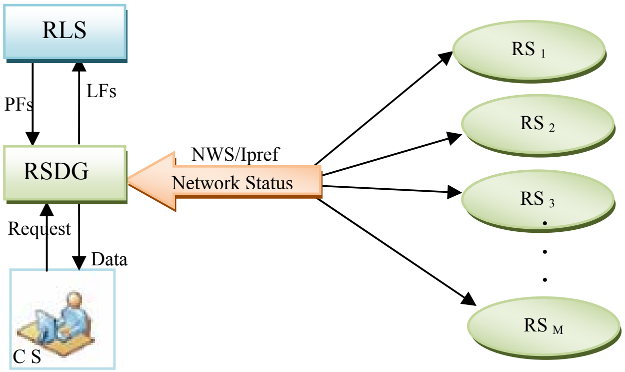

Figure 1. Rough Set Data Grid Strategy & related entities

The criteria set attributes are heterogeneous and conflicting with each other, making the problem quite complex to solve. Therefore the Rough Set Selection Theory is used to select the best replica. Once the best replica(s) is obtained, to start transfer data files the Data Transfer service is used, i.e. GridFTP. Here is a summary of the strategical steps of the proposed approach.

Algorithm:

Step 1: Receive the requests from the users or typically from *Resource Broker (RB)*. A user request R_h is a vector of Q attributes. R_h, $h=1,2,\dots,Q$.

Step 2: Contact the *RLS* to gather the replica location information. Let Replica Sites $RS = \{S_1, S_2, \dots, S_M\}$, where S_{ij}, S_i *is a vector of attributes j where* $i=1,\dots,M$. M represents number of replicas, where $j=1,\dots,N$, N represents number of attributes.

Step 3: Gather the values of all attributes and establish the grey decision table like *Table 2*, with the current criteria values using linguistic values. V_{ij}. Represent the values of attributes which are linguistic variables as in *Table 1.* They are based on grey number and can be calculated as follows: $V_{ij} = \left[\underline{V}_{ij}, \overline{V}_{ij}\right]$

Step 4: Normalize the grey decision table as shown in *Table 4*. The normalization method is to preserve the property. The normalized grey number is belonging to [0, 1].

In our research *Security* and *Availability* attributes are called benefit attributes because the replicas with highest values of them are better than with low values. So, the normalization equation for benefit attributes is expressed as:

$$V_{ij}^{*} = \left[\frac{\underline{V}_{ij}}{V_j^{\max}}, \frac{\overline{V}_{ij}}{V_j^{\max}}\right] \qquad (10\ )$$

$$where\ , V_j^{\max} = \underset{i \geq 1}{\overset{M}{MAX}}\ (\overline{V}_{ij})$$

Whereas in *Cost* and *Time* attributes, the lowest values are better than highest values so the normalization equation is expressed as:

$$V_{ij}^{*} = \left[\frac{V_j^{\min}}{\overline{V}_{ij}}, \frac{V_j^{\min}}{\underline{V}_{ij}}\right] \qquad (11\ )$$

$$where\ , V_j^{\min} = \underset{i \geq 1}{\overset{M}{MIN}}\ (\underline{V}_{ij})$$

Step 5: Select the ideal replicas using the grey-based rough set lower approximation [3].

$\underline{R}S = \{S_i \mid [S_i]_R \subseteq S\}$ *, where i=1,2,..M*

Step 6: Select the *most ideal replica* from set of replicas *RS**, by calculating their correlative degrees using the following these two steps [2].

6.a) Form the most ideal referential replica S_0. It contains the maximum attributes values.

$$S_0 = \left\{ \overset{M}{\underset{i=1}{\forall}} \left[MAX\ \underline{V}_{i1}^{*}, MAX\ \overline{V}_{i1}^{*}\right], \overset{M}{\underset{i=1}{\forall}} \left[MAX\ \underline{V}_{i2}^{*}, MAX\ \overline{V}_{i2}^{*}\right], \right.$$

$$\left. \dots, \overset{M}{\underset{i=1}{\forall}} \left[MAX\ \underline{V}_{ij}^{*}, MAX\ \overline{V}_{iN}^{*}\right] \right\} \qquad (12)$$

6.b) Measure the relationship between S_0 and *RS*,* using the above equations (6,7,8,9). The best replica is the one with less difference from ideal replica S_0. i.e., the biggest value of (Γ_{0k}).

Step 7: Send Physical File locations of the best replica(s) to data transferring service like GridFTP to get the files from one or more replica sites to accelerate transferring time.

4 The application and analysis of proposed approach

In this section, we present a case study based on proposed approach to clarify the steps of our algorithm.

4.1 Example

To select the best replica site using a rough set approximation in our proposed algorithm the following steps should be followed:

Step 1: Get the attribute rating values for replica files by contacting the *RLS* to get all replicas with their attributes. Let us consider a data grid job J_1 asking to get dataset files from the best replica site of the ten replica sites which having the requested file(s) and each replica site has four different attributes: a_1, a_2, a_3 and a_4. Let us consider a_1, a_2 as attributes that have the characteristic such that the high value is better than low value. For example *Security* attribute: the high level value of security is better than the low level value. And let us consider a_3, a_4 to represent the attributes that have the characteristic such that the low value is better than the high value, for example the *Price* of the requested files, where the lowest prices are preferable. The attribute rating values for ten replicas sites are shown in *Table 2 below.*

Table 2.Linguistic attributes values

	a_1	a_2	a_3	a_4
S_1	VG	MG	G	G
S_2	MG	MG	F	G
S_3	F	ML	MG	G
S_4	MG	ML	G	G
S_5	F	L	G	MG
S_6	MG	L	G	L
S_7	G	G	G	G
S_8	G	L	G	L
S_9	MG	ML	G	G
S_{10}	G	L	G	L

Step 2: Establish the decision table as in *Table 3* below. Replica sites *(S_i)* in the rows whereas attributes of each site in the columns. The value of each attribute will be represented using grey real numbers to reflect the reality of linguistic variables. As we can see in *Table3* the decision values *(D_i, i=1,2,…,M)* are given by Replica Manger judgments. Replica manager judgment depends upon the history of the quality of services got from replica sites. Two values *{yes, no}* are used for this attribute. Numerically *{yes, no}* are represented as *{1, 0}*. The Universe is a finite set of objects $U=\{S_1,S_2,S_3,\ldots,S_{10}\}$ and the attributes $A=\{a_1, a_2, a_3, a_4\}$

Table 3. Decision table

	a_1	a_2	a_3	a_4	D_i
S_1	[5.75,6.00]	[5.50,5.50]	[6.50,7.50]	[7.50,8.00]	1
S_2	[5.00,6.00]	[5.50,6.00]	[4.75,5.00]	[6.75,8.25]	0
S_3	[4.75,5.00]	[3.25,4.00]	[5.25,6.00]	[7.50,9.00]	0
S_4	**[5.50,5.50]**	**[3.00,3.50]**	**[6.50,7.50]**	**[7.75,8.25]**	**1**
S_5	[4.50,5.00]	[2.50,3.00]	[6.50,8.50]	[5.25,6.00]	0
S_6	[5.25,6.00]	[2.30,3.00]	[6.5,8.50]	[1.25,3.0]	1
S_7	[8.70,9.00]	[6.50,7.50]	[7.50,7.50]	[7.50,8.00]	0
S_8	**[8.50,9.00]**	**[2.30,3.00]**	**[6.5,8.50]**	**[1.25,3.0]**	**0**
S_9	**[5.50,5.50]**	**[3.00,3.50]**	**[6.50,7.50]**	**[7.75,8.25]**	**0**
S_{10}	**[8.50,9.00]**	**[2.30,3.00]**	**[6.5,8.50]**	**[1.25,3.0]**	**1**

Step 3: Normalize *Table 3* using *(a_1)* and *(a_2)* using equation *(7)* and *(a_3)* and *(a_4)* using equation *(8)*. Refer *Table4.*

Step 4: In the decision table, some objects may have the same attribute values several times. The relation between these objects are called an indiscernibly relation for subset or all set of attributes. In our example we can see this relation between (S_4 and S_9) and also between (S_8,S_{10}) because they have same values of attributes $R=\{a_1,a_2,a_3,a_4\}$. This means, (S_4 and S_9) cannot be recognized by R attributes. So, to find elementary sets of U, $U=\{S_1, S_2, S_3, S_4, S_5, S_6, S_7, S_8, S_9, S_{10}\}$ in space $R=\{a_1, a_2, a_3, a_4\}$.

Table 4. Decision table with normalized grey attributes

Table 4. Decision table with normalized grey attributes

	a_1*	a_2*	a_3*	a_4*
S_1	[0.805556, 0.666667	[0.733333, 0.733333]	[0.653846154, 0.566666667	[0.166666667, 0.15625
S_2	[0.5555556, 0.666667	[0.733333, 0.8]	[0.894736842, 0.85]	[0.185185185, 0.151515152]
S_3	[0.5277778, 0.555556]	[0.433333, 0.533333]	[0.80952381, 0.708333333]	[0.166666667, 0.138888889]
S_4	[0.6111111, 0.611111]	[0.4, 0.466667]	[0.653846154, 0.566666667]	[0.161290323, 0.151515152]
S_5	[0.5 0.555556]	[0.333333, 0.4]	[0.653846154, 0.5]	[0.238095238, 0.208333333]
S_6	[0.583333, 0.666667	[0.4, 0.466667]	[1, 0.85]	[0.166666667, 0.138888889]
S_7	[0.9666667, 1]	[0.866667, 1]	[0.566666667, 0.566666667]	[0.166666667, 0.15625]
S_8	[0.9444444, 1]	[0.333333, 0.4]	[0.653846154, 0.5]	[1, 0.416666667]
S_9	[0.6111111, 0.611111]	[0.4, 0.466667]	[0.653846154, 0.566666667]	[0.161290323, 0.151515152]
S_{10}	[0.9444444, 1]	[0.333333, 0.4]	[0.653846154, 0.566666667]	[1, 0.166666667]

The indiscernibility classes defined by *U/R*, describes the universe *U* with respect to *R using (equation 1).*

$$U/R = \{\{S_1\}\{S_2\}, \{S_3\}, \{S_4, S_9\}, \{S_5\}, \{S_6\}, \{S_7\}, \{S_8, S_{10}\}\}$$

Step 5: Select the ideal replicas set of sites S* using lower approximation of rough set theory. Using lower approximation a set of replicas which has the closest range of attributes to requested attributes is determined. Using *equation 2* we get:

$\underline{R}S^* = \underline{R}X_1 = \{S_i \in U| [Si]_R \subseteq S^*\}$, *Where* $S^* = \{S_i|d_i = yes\}$

This will find all sites which their Decision attribute value is equal to "yes" i.e.

$$X_1 = \{S \mid D_i(S) = Yes\} \rightarrow X_1 = \{S_1, S_4, S_6, S_{10}\}$$

Lower and Upper approximations using *Equation (2)* are:

$$\underline{R}X_1 = \{S_1, S_6\}$$

$$\overline{R}X_1 = \{S_1, S_4, S_6, S_8, S_9, S_{10}\}$$

To check if the decision attribute value (d_i= *yes)* is a rough or not we have to check the boundary of X_1

$$RN_R(X) = \overline{R}X - \underline{R}X \xrightarrow{\quad} \{S_4, S_8, S_9, S_{10}\}$$

The decision class, *yes,* is rough since the boundary region is not empty, i.e.

$$RN_R(X) \neq \varphi$$

So, the ideal replicas sites will be contented in lower approximation set $\underline{R}S^* = \{S_1,S_6\}$. This means, S_1 and S_6 are ideal replicas sites which can be trusted to get the requested files from them.

Step 6: Select the most ideal replica using the following:

6.a) Form the S_0 from maximum attributes using *equation (12)*

In our example, S_0= *[0.940540541, 1], [0.86666667, 1], [1, 0.85], [1, 0.416667]*

6.b) Calculate Γ_{0k} between reference sequence (S_0) and comparative sequences (S_1 and S_6).

The value of Γ_{01} *(S_0,S_1)= 1.733484884* , whereas the value of Γ_{06} *(S_0S_6) = 1.617177469*

Thus result says: the S_1 is the most ideal replica. Because, Γ value of *(S_0,S_1)* bigger than other value of *(S_0S_6)*, in another words we can say that, *(S_1)* vector is closer to the ideal vector *(S_0)* than *(S_6)*.

Step 7: Send *(S_1)* to transport service, *GridFTP* to transfer the requested files.

5 Related works.

Selection best replica problem has been investigated by many researchers before. But as we mentioned earlier the most similar approach research to our work was published in 2009 by *A. Jaradat et al.*[5], they proposed an approach called ***K-means-D-System*** that also utilizes *availability, security* and *time* as selection criteria between different replicas by adopting k-means clustering algorithm concepts to create a balanced (best) solution. In their work, the best site does not mean the site with shortest time of file transfer, but the site which has three acceptable values: security level, availability and time of file transfer. To do selection process a Model Replica *(MR (100,100,100)) is* considered; replica with ideal attributes values. Then using Euclidian distance the closest replica to the ideal model (MR) is recognized. Their *K-means-D-system* has some drawbacks, mentioned below:

1- To get best replica *K-means-D-system* has to find all distances between *MR* and all replicas and then find the shortest distance which marks the best replica site.
2- Cannot deal with sites having same attributes values and sites having same distances with different attribute values.

We demonstrate that our approach better through a simulation in the next section.

6 Simulations and Result

The RSES 2.2 (Rough Set Exploration System 2.2) software tool and (Matlab 7.6.0) are used for the simulation. They provide means for analysis of tabular data sets with use of various methods, in particular those based on Rough Set Theory [4]. We simulate *99* replicas with different attributes and compare our work with the selection *K-means-D-System* proposed in [5]. The authors used K-means rule algorithm to select the best replica among multiple alternatives and the simulation. The results shown in *Figure 2,* implies that our approach is better in terms of speed of execution, as well as more accurate in choosing the best replica site. On the other hand our proposed strategy covers the drawbacks mentioned by *A. Jaradat* and others in [5], which risen up because of the non-consideration of the potential problems which are explained below.

Drawbacks:

In K-means-D-system proposed by *A. Jaradat* in [5] the clustering concept is weekly reflected by using number of cluster is equal to number of replicas *(K=M)*, whereas it should be *(K<M)*.

In [5] the distance between the Ideal replicas, the one having best attribute values and other replica sites is a criterion of preference. Therefore knowing all distances one can select the best replica site, which is the site with less distance value. But the system cannot accurately select the best replica in case of equal distances, which may occur in two the following two cases:

Attributes/Site	A	S	T	C
S_1	60	75	50	40
S_2	60	75	50	40

Using the Euclidian equation the distance will be the same and equal to *(91.24144)*. In this case it becomes impossible to select the best replica out of two equal.

Our strategy covers this problem using grey numbers, so the same attributes can rarely occur even if they have the same linguistic degree. For example, if both replicas have *Good Security (G)*, we can observe when we look at *Table1* that the *Good* level taking many different values between *[6, 9]*. In case of finding the same security value they still may not have the same decision value, therefore the strategy can distinguish between them.

Second: The case of different values with the two equal distances. For example, in case there are two replicas *(S_1, S_2)*, with these values of attributes:

Attributes/Site	A	S	T	C
S_1	50	50	99	99
S_2	99	99	50	50

The distance using the Euclidian equation will be the same for both and equal to (*70.72482*). Then it is hard again to select the best replica in this case. As we can see S_2 is far better than S_1 but the system cannot take a decision. If the system follows the authors in [5] selecting the replica arbitrary, it might select S_2, with low availability and security, at the same time with high cost and latency. This problem can be covered using normalization concept.

There is another possible case of the system failure. Let's say again there are two replicas *(S_1, S_2):*

Attributes/Site	A	S	T	C
S_1	50	60	99	99
S_2	99	99	50	50

D_1 is a distance between *MR* and S_1 equal to *(64.04686)*. D_2 is a Distance between *MR* and S_2 equal to *(70.72482)*.

In the earlier proposed selection system *(K-means-D-system)* the best replica will be the one closest to *RM,* i.e. the replica site that have the least distance. In this example S_1 will be selected as a best replica site. But this is incorrect selection since the attributes values of S_2 are much better than attributes values of S_1. The file(s) in S_1 is (are) more available, more secure, having less transfer time and low cost (price). This problem is resolved using lower approximation concept of rough set theory.

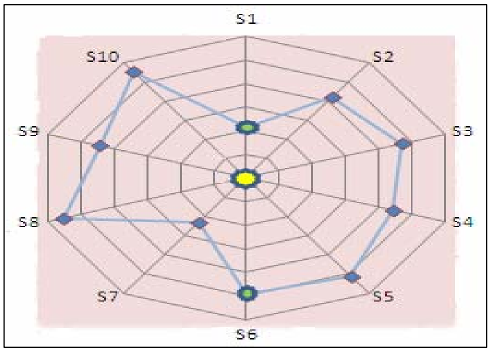

Figure 2. Distance between ideal replica (✪*S_0)* and available replicas(*S_1,S_2,..,S_{10})* Using RSDG algorithm

When no one of the previous issues appears in the data set, the previous approach, *K-means-D-system* will work properly and the selection might be the same as our strategy selection. Refer *Figures (2, 3)* for details of the simulation results.

As we can see below, both strategies *K-means-D-system* and our strategy *RSDG* have selected S_1 as a best replica site, but second best replica in *K-means-D-system* will be S_7 whereas in *RSDG* it will be S_6. We can see that even the distance of S_6 is less than S_7 but *RSDG* system has selected since it depends on decision values which make the selection more accurate and closest to the ideal.

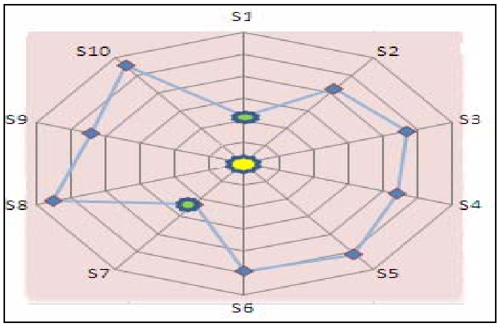

Figure 3. Distance between ideal replica (☼S_0) and available replicas (S_1,S_2,..S_{10}) Using K-means rule algorithm

7 Conclusions and Future Work

In this paper, we proposed a new replica selection strategy which uses a grey-based rough set approach to deal with replica selection problem under uncertainty environment of attributes. In data grid infrastructure, when the user/ application sends a request to execute a specific job, this job may consist of file or set of files that must be available in the computing site to make the job ready for execution. The user/application sends its request with a rating attributes that determine the user preferences for example: security level, availability, cost and the critical time which the execution must not exceed. Since the requested files are scattered in different distributed places of the world, the selection process of deciding from where the files should be brought becomes a sensitive issue because it has a large impact on the total time of the execution of the requested job. The sites that contain a copy of the requested files have different characteristics such as security, probability of data availability, cost (price) of the file and time taken to transfer the file from the replica site to the computing site. These attributes cannot be represented using the exact numerical numbers because they are linguistic variables which can be expressed better using grey numbers. Grey numbers are used for an accurate attribute expression and to prevent getting replicas with the same attributes values. This problem happened in the previous work well where the authors mentioned it as a drawback of their paper [5]. That is the important reason to select grey numbers to represent replicas attributes. On the other hand the Rough Set Theory is very useful for the selection of the set of replicas which have the closest attributes to the ones requested by the user/ application. Lower approximation concept being a classification of the domain objects (replicas) works to describe the replicas which are with certainty belong to the subset of interest. In other words, it works to distinguish a group of replicas that contains the closest values of attributes to the user request. Then the best replica site or a set of sites can be easily selected to share the transfer of the requested file(s) and to accelerate the execution of the job. An example of replica selection problem is used to illustrate the proposed approach. The experimental results shows an advanced performance, compared to the previous work in this area and this gives us a good opportunity being a node of PRAGMA Data Grid Infrastructure to develop our strategy as a service for the optimization component of our Data Grid Site[16,17].

8 Acknowledgements

Authors wish to register their sincere thanks to Prof. Arun Agarwal, from GridLabs Department of Computer and Information Sciences, University of Hyderabad, India for providing all the infrastructural and computational support required to carry out this work. His academic suggestions to improve the quality of the work are also highly appreciated/ acknowledged.

References:

[1] Foster, I. and C. Kesselman, 2001. The anatomy of the grid: Enabling scalable vimal organizations. Int. J. HighPerform. Comput. Appl., vol: 15: pp: 200-222.

[2] Li, G.D., Yamaguchi, D., and Nagai, M.: A Grey-Based Approach to Suppliers Selection Problem. In: Proc. Int. Conf. on PDPTA’06, Las Vegas, pp: 818–824, June.2006.

[3] Yamaguchi, D., Li, G.D. and Nagai, M: On the Combination of Rough Set Theory and Grey Theory Based on Grey Lattice Operations, Int. Conf. on RSCTC’06.

[4] logic.mimuw.edu.pl/~rses/RSES_doc_eng.pdf

[5] A. Jaradat, R. Salleh and A. Abid : Imitating K-Means to Enhance Data Selection, Journal of Applied Sciences 9 vol: 19, pp:3569-3574, 2009, ISSN pp:1812-5654.

[6] Pawlak, Z., 1982. Rough sets. International Journal of Computer and Information Sciences. Vol 11, pp: 341–356.

[7] Pawlak, Z.: Rough Classification. Int. J. Man-Machine Studies. Vol: 20, pp: 469–483, 1984.

[8] Polkowski, L. and Skowron, A. (Eds.): Rough Sets in Knowledge Discovery. Physica-Verlag 1(2) 1998.

[9] Wang, S.J. and Hu, H.A.. Application of Rough Set on Supplier’s Determination. The Third Annual Conference on Uncertainty. pp: 256–262, Aug. 2005

[10] Deng, J.L.: Control Problems of Grey System. System and Control Letters. Vol: 5, pp: 288–294, 1982.

[11] Nagai, M. and Yamaguchi,D.Elements on Grey System Theory and its Applications. Kyoritsu publisher, 2004.

[12] Wang, Y.X.: Application of Fuzzy Decision Optimum Model in Selecting Supplier. J. Science Technology and Engineering. Vol: 5 ,pp: 1100–1103, Aug. 2005.

[13] Ahmar Abbas, Grid Computing: A Practical Guide to Technology and Applications, 2006.

[14] Vazhkudai, S., S. Tuecke and I. Foster, 2001. Replica selection in the globus data grid. Proceeding of the First IEEE/ACM International Symposium, May 15- 18, Brisbane, Australia, pp: 106.1 13.

[15] Xia, J.: Grey System Theory to Hydrology. Huazhong Univ. of Science and Technology Press, 2000.

[16] http://goc.pragma-grid.net/pragmadoc.-

[17]http://rocks67.sdsc.edu/cgibin/scmsweb/probe_analysicgi ?cluster=venus.uohyd.ernet.in&grid=PRAGMA