Rough Set Clustering Approach to Replica Selection in Data Grids (RSCDG)

Rafah M. Almuttairi, Rajeev Wankar, Atul Negi, Raghavendra Rao Chillarige
Department of Computer Science,
University of Hyderabad
Hyderabad, India
rafahmohammed@gmail.com,{wankarcs, atulcs, crrcs}@uohyd.ernet.in

***Abstract* – In data grids, the fast and proper replica selection decision leads to better resource utilization due to reduction in latencies to access the best replicas and speed up the execution of the data grid jobs. In this paper, we propose a new strategy that improves replica selection in data grids with the help of the reduct concept of the *Rough Set Theory (RST)*. Using *Quickreduct* algorithm the unsupervised clustering is changed into supervised reducts. Then, *Rule algorithm* is used for obtaining optimum rules to derive usage patterns from the data grid information system. The experiments are carried out using *Rough Set Exploration System (RSES)* tool.**

Keywords- Data Grid; Replica Selection Strategies; Rough Set Theory (RST); K_means; Quickreduct; Rule Algorithm.

I. Introduction

An increasing number of scientific applications ranging from high-energy physics to computational genomics require access to large amounts of data with varied quality of service requirements. This diverse demand has contributed to the proliferation of storage system capabilities, thus making storage devices an integral part of the Grid environment and thereby constituting the *Data Grid* [13]. Data Grid architectures like *PRAGMA* [15] are efforts to standardize access to the multitude of storage systems spread across the grid environment. The architecture attempts to abstract these diverse elements of the data grid by providing a set of core services that can be used to construct a variety of higher-level services [13]. One of the most important benefits of using data grid infrastructure is *Data Replication*. The goal of replication concept is to minimize latency of data access. The data is replicated at several sites across the grid to avoid a single site to be flooded by requests [13]. 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* [14]. The best replica selection is a multi attribute decision making problem, because each *Replica Site* has its own capabilities, characteristics and values of replica attributes. At the same time every user has its own preference of attributes. Here a *Replica Site* $RS=\{ Si,\ i=1,2,\ldots.M\}$, where M is the number of replica. And also each replica is a vector of attributes $A=\{a_j , j = 1,..,N\}$, where N is the number of attributes. The grey value of attribute (a_j) in the site (S_i) is denoted by (V_{ij}). A *User Request* $UR=\{UR_h, h=1,2,\ldots,Q\}$ is also a vector of attributes, Q is the number of user attributes and $(Q\leq N)$. Therefore the replica selection problem can be transformed to a nearest match problem between UR_h and S_i. In this paper the nearest match problem is solved using *Grey based Rough Set Theory*. Knowledge extraction technique is used as a tool that can assist *Replica Manager (RM)* in analyzing vast amounts of replicas and turn the information contained in the data sets into successful decision making. In this paper, we first have used the *Reduct* concept of the rough set theory to construct the supervised clusters of replicas. Here, we consider a data grid information system without history of replicas, i.e. in the absence of *Decision* attributes of information system which might be found in the real life example of the replica selection process. So, a novel clustering technique is applied to cluster replicas by available attributes information to get different clusters labels which are used as labels of the *Class* (*Decision)* attribute. *QuickReduct* algorithm is used to get reducts. After that a *Rule* algorithm is used to get optimum rules which can be used as a guide to extract the best replica which has the best match to the user request in minimum search time. All that will lead minimize the searching space. This means, the time spent to get the best replica will be minimized using directed rules. And, the user can be served with best match to his request.

The rest of the paper is organized as following: *Section II* presents the preliminary concepts for the *Grey System Theory*, *Reduct algorithms, Rule Extraction*, *Grey-based Rough Set Theory* and *Grey based K-means* algorithm. *Section III* summarizes the related work. *Section IV* introduces the *Data Gird* mining process and proposed algorithm *RSCDG*. The application and analysis of the proposed approach are shown with an example of replicas selection in *Section V*. *Section VI* includes the simulation and its' result. The Conclusions are given in *Section VII*.

II. Preliminary Concepts

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

A) Rough Set Theory

In rough sets theory [6,7,8], the data is organized in a table called *Decision Table (DT)*. *Rows* of the decision table correspond to objects, and columns correspond to

978-1-4244-8136-1/10/$26.00 ©2010 IEEE

attributes. In the data set, a class label is used to indicate the class to which each row belongs. The class label is called as *Decision attribute (D)*, the rest of the attributes are the *Condition attributes (C)*, where $C \cap D = \phi$. t_j denotes the j^{th} tuple of the data table. *Rough Set Theory* defines three regions based on the equivalent classes induced by the attribute values: *Lower approximation, Upper approximation, and Boundary*. *Lower approximation* contains all the objects, which are classified surely based on the data collected, and *Upper approximati*on contains all the objects, which can be probably classified, while the *Boundary* is the difference between the upper approximation and the lower approximation [9].

Definition 1. Let *U* be a non-empty finite set of objects called universe, *A* is a non-empty finite set of attributes and *R* is an equivalence relation on *U*. Then *T=(U,A)* is called an *approximation space*. The *indiscernibility* relation of *R* is:

$$IND(R) = \{(x_1, x_2) \in U^2 \mid \forall a \in R, a(x_1) = a(x_2)\} \qquad (1)$$

B) Reduct algorithms

Reducts algorithms are used to minimize the number of attributes in the *Decision (DT)* and keep only attributes which used to discriminate the replicas.

The reduction of attributes is achieved by comparing equivalence relation generated by sets of attributes. A *reduct* is defined as a subset of minimal cardinality R_{min} of conditional attribute set *C* such that

$$\gamma_R(D) = \gamma_C(D)$$

$$R = \{X : X \subseteq C; \gamma_X(D) = \gamma_C(D)\}$$

$$R_{\min} = \{X : X \in R; \forall Y \in R; |X| \leq |Y|\} \qquad (2)$$

The intersection of all the sets in R_{min} is called the core, the elements of which are those attributes that cannot be eliminated without introducing more contradictions to the dataset. In this method subset with minimum cardinality is searched for. *Quickreduct* algorithm used to compute a minimal reduct without exhaustively generating all possible subsets [19].

C) RULE EXTRACTION

Rule extraction algorithm is used to formulate the efficient rules [12].

D) Grey-Based Rough Set

Grey system theory [10], originally developed by *Deng in 1982*, has become a very efficient method to solve uncertainty problems under discrete data and incomplete information.

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 to be known precisely, for instance *Security* may be represented by different linguistic values like *{None, Low, Medium, Adequate, High}* and each linguistic value has upper and lower limits. So in the decision table of rough set theory, the attribute values would be represented precisely [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.

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

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) [3].

Definition 4. A grey number is the one the exact value of which is unknown, while the upper and/or the lower limits can be estimated. Generally if the lower and upper limits of *x* can be estimated then *x* is defined as interval grey number [10]. As it can be written as:

$$\otimes x = \left[\underline{x}, \overline{x}\right] \qquad (3)$$

Definition 5. Relationship between two grey sequences x_0 *and* x_k [10]

$$\Gamma_{ok} = \frac{1}{M} \sum_{i=1}^{M} \left(\frac{\Delta_{\max} - \Delta_{0i}(i)}{\Delta_{\max} - \Delta_{\min}}\right) \qquad (4)$$

where,

$$\Delta_{\max} = \max_{\forall i, \forall k} [\max D(x_0(k), x_i(k))] \qquad (5)$$

$$\Delta_{\min} = \min_{\forall i, \forall k} [\min D(x_0(k), x_i(k))] \qquad (6)$$

$$\Delta_{0i}(k) = D(x_0(k), x_i(k)) \qquad (7)$$

TABLE 1. THE CATEGORIES OF GREY ATTRIBUTES FROM 1-10

Category	Grey Values	Normalized Values
Very Low(VL)	[0,1]	[0,0.1]
Low (L)	(1,3]	(0.1,0.3]
Medium (ML)	(3,4]	(0.3,0.4]
Fair (F)	(4,5]	(0.4,0.5]
Medium Good (MG)	(5,6]	(0.5,0.6]
Good (G)	(6,9]	(0.6,0.9]
Very Good (VG)	(9,10]	(0.9-1]

E) Grey based K-means clustering algorithm

Here we present our version of a clustering algorithm that partitions the data sets into *K* clusters, where each cluster comprises data-vectors with similar inherent characteristics. The overall outcome of this stage is the availability of *K*-number of data clusters, which forms the basis for subsequent discovery of symbolic rules that define the structure of the discovered clusters.

Input:

- *GIT : Grey Information Table*
- *K : Number of clusters*
- *Q: Set of user's attributes.*

Output:

- *Cluster's labels* $C_y=\{C_1,C_2,\ldots,C_K\}$

Step 1: Randomly select *K* centers identification $(c_1, c_2, \ldots, c_K)$ from replicas S_i.

Step 2: Assign point S_i, to C_y, *iff* $A_y < A_p$, $p,y=1,2,\ldots K$, and $j \neq p$ where,

$$A_y = Dist(S_i, c_y) = Dist(\{a_1, a_2, \ldots, a_Q\}, \{v_1, v_2, \ldots, v_Q\})$$

$$= \sum_{l=1}^{Q} \sqrt{(\underline{a_l} - \underline{v_l})^2 + (\overline{a_l} - \overline{v_l})^2}$$

where, l=1,2,…Q, Q is the number of user's request attributes. i=1,2,…M, M is number of Replicas. And $v_l = v_1, v_2, \ldots, v_Q$ *are the attribute values of the center* c_y *of cluster* C_y

Step 3: Compute new cluster centers: $c_1^*, c_2^*, \ldots, c_K^*$, as following, using *Equation 4*

$$c_y^* = \sum_{j=1}^{Q} \frac{1}{n_y} \sum_{\forall S \in C_y} (\underline{a_j}, \overline{a_j})$$

where n_y is the number of replicas in cluster C_y.

Step 4: *If* $c_y^* = c_y$ then terminate. Otherwise go to *step 2*.

Note: *Here the replicas will be clustered with respect to the users' set of attributes (Q) and not to all replica attributes (N).*

III. Related work

A new selection system called *K-means-D-System* that uses a *K-means* algorithm was proposed by *Jaradat at el* in [5]. The best replica site was the one which has high levels of *Security*, *Availability* and lowest level of response time in transferring requested files, and that is closest (using *Euclidian distance equation*) to the user request. *K-means-D-system* has some drawbacks. *First*, is that to get the best replica, *K-means-D-system* has to find *all* distances between *user request* R_h and replicas S_{ij} because it is assumed that *K* is equal to number of replicas. So, the best replica site is the one that has the shortest distance. The next, is that it cannot deal with sites having same attributes values because the distance would be the same value. Therefore the selection process of the best replica will be random because using the *Euclidian distance equation* for computations the distances are being equal. The same as, for instance, these attributes values of *(S_1, S_2)* in *Case1*. We take care of this problem using Grey values for representing the attributes.

	Attributes/Si	A	S	T	C
Case 1	S_1	60	75	50	40
	S_2	60	75	50	40

Another difficulty with that approach happens when sites having same distances with different attribute values get the same rank in the selection process. For example these two replicas *(S_1, S_2)*, with these values of attributes: The distance using the Euclidian equation distance will be equal for *(S_1, S_2)*. As we can see S_2 is far better than S_1, but *K-means-D-SYSTEM* from the previous work [5] might select S_1, with low availability and security levels, and with high cost and latency levels. This problem is taken care of using the normalization values concept in our proposed strategy (*RSCDG*).

	Attributes/Si	A	S	T	C
Case 2	S_1	50	50	99	99
	S_2	99	99	50	50

There is another possibility of *K-means-D-System*'s failure. Let's take below attributes values of two replicas (S_1, S_2) as an example,

	Attributes/Si	A	S	T	C
Case 3	S_1	50	60	99	99
	S_2	99	99	50	50

Using the *K-means-D-System* [5] the best replica is the one having the least distance. So, in this example, S_1 will be selected as a best replica site. But, this selection is incorrect, because the attributes values of S_2 are much better than 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 the concepts of *Rough Set Theory*.

IV. RSCDG Approach

In this study a new centralized and decentralized replica selection strategy using rough set theory (*RSCDG*) is being proposed. The new strategy considers the *QoS* of the data grid sites, whereas *QoS* itself is considered as a combination of multiple attributes such as *Security*, *Availability*, *Cost* and so on.

The *RSCDG* strategy can utilize many existing successful data grid core services, such us *Replica Location Service RLS* and *Network Weather Service (NWS)/Ipref* [13] as shown in *Figure 1*. *RLS* provides the *Physical File* locations (*PF*) for all available *Logical Files* names (*LF*) and *NWS* provides information about the network [1]. The *RSCDG* 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.

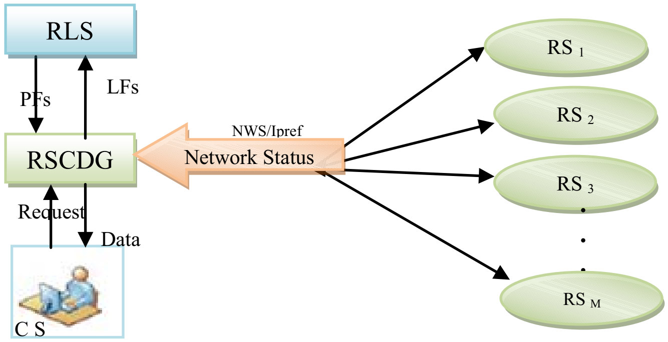

Figure 1. Rough Set Clustering Strategy in Data Grid & related entities

A) DataGrid mining Process

As it is shown in the diagram of Figure 2, there are four stages of Data Grid mining process:

Stage 1: Collect replicas with their attributes.
Stage 2: Use Clustering algorithm like *K-means*.
Stage 3: Use Reduction algorithm like *Quickreduct.*
Stage 4: Construct rule discovery.

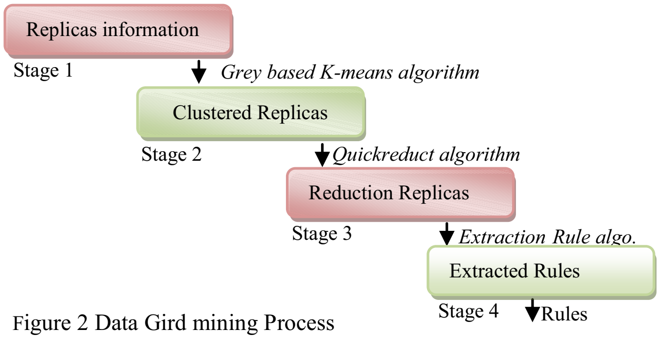

Figure 2 Data Gird mining Process

Stage 1: *Replicas' information Preparation:*

In the first stage, the replicas information is collected and tabulated by data grid services, such as *Replica Location Service (RLS)* and *Network Weather Service (NWS)*. As an example, we assumed that there are eight replica sites having copy of required file(s). Each replica has four different attributes as it is show in *Table 2 and* contains the required information about all replicas. For example *Security (the level of file security)*, *Availability (number of hours the file will be available)*, *Cost (cost of the file)*, *Response Time (time take to get response from replica site)* and so on.

TABLE 2 GREY INFORMATION TABLE (GIT)

S_i	a_1	a_2	a_3	a_4
S_1	[0.9,1]	[1.5, 2]	[0.9,1]	[3.6,4]
S_2	[0.9,1]	[3.8,4]	[1.5, 2]	[5.5,6]
S_3	[1.5, 2]	[3.8,4]	[0.9,1]	[3.9,4]
S_4	[2.5, 3]	[1.5, 2]	[0.9,1]	[5.8,6]
S_5	[2.5, 3]	[3.8,4]	[0.9,1]	[3.3,3.9]
S_6	[0.9,1]	[3.8,4]	[0.9,1]	[3,3.5]
S_7	[2.5, 3]	[3.8,4]	[1.5, 2]	[5,5.5]
S_8	[0.9,1]	[1.5, 2]	[1.5, 2]	[5.7,6]

Stage 2: *Clustering Replicas using Grey based K-means Clustering Algorithm*

Since the *Decision* attributes are absent, *Grey based K-means clustering algorithm* is used to partitioning replicas into *K* clusters to form the *Decision Table (DT)*. The names of clusters (labels) are used as a decision labels. The steps of *Grey based K-means clustering algorithm* has mentioned in *Section II-E*.

Stage 3: *Reduction Replicas*

To minimize the number of attributes in *(DT), Quickreduct* algorithm is used. *Reduct* keeps only those attributes that preserve the *indiscernibility* relation and, consequently, set approximation. There are usually several such subsets of attributes and those which are minimal are called *reducts.*

Stage 4*: Rule Extraction*

In this stage, reduced data obtained from *Stage 3* is applied to the *Rule Extraction algorithm* [12] to formulate the efficient rules. The rule extraction algorithm uses the Heuristic Approach which has mentioned in *Section II-C.*

Based upon the general process explained earlier we present the *RSCDG* algorithm in the following.

B) RSCDG Algorithm

Step 1: Receive *user request* $AR_{,}=\{a_1,a_2,\ldots,a_Q\}$, with $VR=\{vr_1, vr_2, \ldots, vr_Q\}$ and the priority of each attribute $P=\{p_1 p_2,\ldots p_Q\}$.

Step 2: Gather replicas information by contacting *RLS* to form *GIT.*

Step 3: Call *Grey based K-means clustering algorithm,* Input :(*GIT*, *K*) ; Output: (C_Y, C_Y, V_L)

Step 4: Form *Grey Decision Table (GDT*) using clusters labels (C_y).

Step 5: Using scale of grey attributes of *Table1*, Convert *(GDT)* to *Category Decision Table (CDT)*.

Step 6: Find *Reducts* by applying *Quickreduct algorithm* on *CDT* and form *Reduction Table (RT)* which having only *reduct* attributes.

Step 7: Apply *Rule Extraction* algorithm on *RT* using the following steps :

i. Merge identical rows having similar condition and decision attribute values.

ii. Compute the core of every row and form *Core Table (CoT).*

iii. Merge duplicate rows and compose a table with *reduct* value and form *Merged Rows Table (MT)*

Step 8: Formulate the efficient rules.

Step 9: Use $P=\{p_1 p_2,\ldots p_Q\}$ to find the *Closest Match Rule (CMR)* of the user's request.

Step 10: Get *Decision Label* (*Cluster Label*, C_g) of *CMR*.

Step 11: Normalize:

a. Grey attributes values of replica sites in *cluster* $C_g = \{S_1,S_2,\ldots S_{ng}\}$, where *ng* is total number of replicas in C_g.

b. User request to get VR^* .

- To *Benefit Attributes* where the highest values are preferable as *Security* and *Availability* attributes use the following equation[2]:

$$a^*_{ij}=\left[\frac{\underline{a}_{ij}}{a_j^{\max}},\frac{\overline{a}_{ij}}{a_j^{\max}}\right] \tag{8}$$

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

- *For cost attributes* where the lowest values are preferable like, *Cost (Price)* and response *Time* attributes use the following equation[2]:

$$a^*_{ij}=\left[\frac{a_j^{\min}}{\overline{a}_{ij}},\frac{a_j^{\min}}{\underline{a}_{ij}}\right] \tag{9}$$

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

Step 12: Compute the relationship *(Γ),* between the normalized user request VR^* *and* $C_g = \{S_1,S_2,\ldots S_{ng}\}$ using *Equation (6)*, The result is $\Gamma_{g=}\{\Gamma_1 ,\Gamma_2 ,\ldots,\Gamma_{ng}\}$.

Step 13: Find closest site to the user's request with maximum value of (Γ). $\Gamma_{b=}Max(\Gamma_g)$.

Step 14: Send the physical name of S_b to data transferring service like *GridFTP* to transfer requested file .

V. The application and analysis of proposed approach

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

Case study:

To select the best replica site by using our proposed algorithm use the following steps:

Step 1: Receive user request: $AR=a_h=_,\{F,ML,VL, VL\}$.

Step 2: Form *GIT* by contacting the *RLS*. Assume there are *10* replica sites having different attributes $a_j=\{a_1, a_2, a_{3,} a_4\}$, In our example assume that $\{a_1,a_2\}$ representing *benefit attributes* and $\{a_3,a_4\}$ are representing *cost attributes*.

Step 3: Call *Grey based K-means clustering algorithm* to drive *Decision attribute (D)*.
In our example input data are

- *GIT (Table 2).*
- $K=2$.
- $AR= \{a_1,a_2,a_3, a_4\}$.

And the outputs of algorithm are:

- $C_y= \{1,2\}$, *represents class labels of D,*

The result is : $C_1=\{S_1, S_3 ,S_5, S_6\}$ and $C_2=\{S_2, S_4,S_7, S_8\}$

Step 4: Form *GDT.*

Step 5: Convert *GDT* to *CDT* using *Table 1* categories as it is shown in *Table 3.*

Table 3 Category Decision Table (CDT)

S_i	a_1	a_2	a_3	a_4	C_y
S_1	VL	L	VL	ML	1
S_2	VL	ML	L	MG	2
S_3	L	ML	VL	ML	1
S_4	F	M	VL	MG	2
S_5	F	ML	VL	ML	1
S_6	VL	ML	VL	ML	1
S_7	L	ML	L	MG	2
S_8	VL	L	L	MG	2

Step 6: Apply *Quickreduct* algorithm on *CDT (Table 3*) to get *Reduction Table (RT)* as it is shown in *Table 4*.

Step 7: Apply *Rule Extraction* algorithm, if any identical pair objects of (a_1, a_2, a_3) occur merge it, otherwise compute the core of *Table 4* and present it as in *Table 5*. In the next step, merge duplicate objects with same decision value and compose. The merged rows are $\{S_1,S_6\}$,$\{S_2 ,S_8\}$,$\{S_3,S_5\}$ and $\{S_4 ,S_7\}$ as presented in *Merged Table (MT),Table 6.*

Step 8: Formulate the efficient rules. Decision rules are often presented as implications and are often called "if…then…" rules.

We can express the rules as follows:

(1) If $a_3= VL$ THEN $C_g= 1$.
(2) If $a_3= L$ THEN $C_g = 2$.
(3) If $a_2= ML$ and $a_3= VL$ THEN $C_g = 1$.
(4) If $a_1= F$ THEN $C_g = 2$.

Table 4 Reduction Table (RT)

S_i	a_1	a_2	a_3	C_y
S_1	VL	L	VL	1
S_2	VL	ML	L	2
S_3	L	ML	VL	1
S_4	F	M	VL	2
S_5	F	ML	VL	1
S_6	VL	ML	VL	1
S_7	L	ML	L	2
S_8	VL	L	L	2

Table 5 Core Table (CoT)

S_i	a_1	a_2	a_3	C_y
S_1	VL	*	VL	1
S_2	VL	*	L	2
S_3	*	ML	VL	1
S_4	F	*	*	2
S_5	*	ML	VL	1
S_6	VL	*	VL	1
S_7	L	*	*	2
S_8	VL	*	L	2

Table 6 Mereged Rows Table (MT)

S_i	a_1	a_2	a_3	C_y
S_1	VL	*	VL	1
S_2	VL	*	L	2
S_3	*	ML	VL	1
S_4	F	*	*	2

Step 9: Use *P*, attributes priorities to find *CMR*.
$AR^*=\{ F,ML,VL, VL \}$, $P=\{50\%,20\%,90\%,10\%\}$
CMR= rule(1) because a_3 has a highest priority.

Step 10: *Class label of CMR is C_1*. It means the best sites with highest match attributes to the user request can be found in *Cluster 1 (C=1).*

Step 11: Normalize *C_1 to get C_1** as in *Table 7, and AR to get :* $AR^*=\{[0.4,0.49],[0.3,0.39],[0,0.1],[0,0.1]\}$.

Step 12: Use *Equation 4* and compute the relation between AR^* and C_1^*. $\Gamma_{g\,=}\{\Gamma_1 ,\Gamma_2 , \Gamma_{3,} \Gamma_4\}$

$$\Gamma_1 = \Gamma_{(AR^*,S1)} =0.25 \quad , \quad \Gamma_2= \Gamma_{(AR^*,S2)} =0.093$$
$$\Gamma_{3=} \Gamma_{(AR^*,S3)} =0.001 \quad , \quad \Gamma_{4=} \Gamma_{(AR^*,S4)} =0.097$$

Step 13: $\text{Max}(\Gamma_g) = \Gamma_1 = \Gamma_{(AR^*,S1)} =0.25$

Step 14: Send physical name of S_1 *to GFTP.*

Table 7 Normalised attribute Table (MT)

S_i	a^*_1	a^*_2	a^*_3	C_y
S_1	[0.3,0.333]	[0.375, 0.5]	[1,0.9]	1
S_3	[0.5, 0.6667]	[0.95,1]	[1,0.9]	1
S_5	[0.833, 1]	[0.95,1]	[1,0.9]	1
S_6	[0.3,0.333]	[0.95,1]	[1,0.9]	1

For more clarification of the meaning of the rules first, let us declare the meaning of attributes in our example. a_1 represent the security level, a_2 represents the availability of the file, a_3 represents the cost (price) of the file in each site, and a_4 represents the response time of replica site. When the user/application of data grid asks for getting a file with specific attributes values (Security, Availability, Cost and Time), the *Replica Manager* broker (*RM*) using our proposed strategy tries to serve the user with the best match of his requirement and in shortest time. Let us assume that the user requests a file with: *Fair level* of *Security, Medium level* of *Availability* and *Low Price Cost*. User gives the highest priority to the cost of the file attribute and less to others. In this case the *RM* checks the four extracted rules and selects the rules where the *Cost (Price)* attribute is *Low* as in *rule(1,3)* since it serves user's request. Both rules (1,3) point to cluster (C_1) so the best replica site is one of cluster (C_1) sites. In case of the extracted rules point to different clusters in this case the second priority attributes comes into the picture to decide one cluster and so on.

VI. Simulations and Result

The *RSES 2.2 (Rough Set Exploration System 2.2)* software tools [4] and (*Matlab 7.6.0*) are used for the simulation using random values of replicas attributes. They provide the 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 results have shown that our method 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 of *A. Jaradat at el.*[5], which we mentioned in previous section.

VII. Conclusions and Future Work

In this paper, we proposed a new replica selection strategy which uses:

1- A *Grey-based Rough Set* approach to deal with replica selection problem under uncertainty environment of attributes [18]. Grey numbers help in compilation number of sites which values are close together in one group grey category.

2- *Grey based K-means* clustering algorithm to cluster replicas into classes and use classes' labels as decision attributes in case they are unavailable by replica manager [17].

3- *Quickreduct algorithm,* since the decision table may have more than one reducts. Anyone of them can be used to replace the original table. Fortunately, in data grid selection applications it is usually not necessary to find all of them and it is enough to compute one such reduct is sufficient [16].

4- *Rule Algorithm,* we used rule algorithm to formulate the efficient rules which are used to minimize the searching space.

To select the best replica site, the result shows that our proposed approach *(RSCDG)* is faster than the previous method [5]. The reason is that the searching space in our method is minimized. And also our method is more accurate because the clusters *in RSCDG* are supervised clusters.

The experiments are carried out on randomly generated data sets. Our results show an improvement of performance, comparing 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 [15]. The proposed work can be improved by introducing the Neural Network in order to train the system and this is the direction for further research work.

Acknowledgements

Authors wish to register their sincere thanks to Prof. Arun Agarwal, from Grid Lab 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] I. Foster, and C. Kesselman, 2001. The anatomy of the grid: Enabling scalable vimal organizations. Int. J. HighPerform. Comput. Appl.,vol: 15: pp: 200-222.

[2] G.D. Li, D.Yamaguchi, , 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] D.Yamaguchi, G.D Li, 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] Z. Pawlak, 1982. Rough sets. International Journal of Computer and Information Sciences. Vol 11, pp: 341–356.

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

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

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

[10] D. Ju-Long.: Control Problems of Grey System. System and Control Letters. Vol: 5, pp: 288–294, 1982.

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

[12] Y. Yang and T. C. Chiam. Rule Discovery Based On Rough Set Theory. Proceedings of ISIF, TuC4, 11-16, 2000

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

[14] S. Vazhkudai, S. Tuecke, I. Foster, "Replica Selection in the Globus Data Grid", Proceedings of the IEEE International Conference on Cluster Computing and the Grid (CCGRID 2001), pp. 106-113, Brisbane, Australia, May 2001.

[15] http://goc.pragma-grid.net/

[16] X. Hu, T.Y. Lin and J. Jianchao. A New Rough Sets Model Based on Database Systems. Fundamenta Informaticae, 1-18, 2004.

[17] R. M. Almuttairi, R. Wankar, A. Negi and C.R. Rao, Replica Selection in Data Grids using Preconditioning of Decision Attributes by K-means Clustering (K-RSDG), VCON2010, Vaagdevi College of Engineering, Warangal, Andhra Pradesh, India (Recently Accepted). IEEE Computer Society Press, Los Alamitos, CA USA.

[18] R. M. Almuttairi, R. Wankar, A. Negi and C.R. Rao,Smart Replica Selection for Data Grids using Rough Set Approximations(RSDG), 2010 IEEE International Conference on Computational Intelligence and Communication Networks, 26-28 Nov 2010, Bhopal, India(Recently Accepted). IEEE Computer Society Press, Los Alamitos, CA USA.