iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 **Issue:** II **Month of publication:** February

DOI: http://doi.org/10.22214/ijraset.2019.2175

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue II, Feb 2019- Available at www.ijraset.com

A Survey of Dynamic Replication Strategies based on Data Mining Techniques

Ms Prema. R[1], Dr Antony Selvadoss Thanamani[2]
[1]*Head and Assistant Professor, Department of Computer Applications, Indo Asian Women's Degree College, Bangalore-560043.*
[2]*Head & Associate Professor, Department of Computer Science, Nallamuthu Gounder Mahalingam College, Pollachi – 642 001*

Abstract: Grid Computing is an emerging research field which is mainly concentrates on enterprise information systems. To improve the dynamic data management in grid computing the efficient replication strategy needs to be followed. In this paper, we focused on various data mining based replication strategies to enhance data replication which are important data management techniques commonly used in data grids.
Keyword: Grid Computing, Data Grid, Data Replication, Dynamic Threshold, Data Mining

I. INTRODUCTION

Data Grids provide geographically distributed resources for large-scale data-intensive applications which generate large data sets [4]. High speed of the internet determines the fast and efficient access of widely distributed data. Replication technique is one of the efficient factor for which affects the performance of data grids by replicating data in distributed data sources. Replica selection, replica placement, replica management are the three major key issues in all data replication algorithms [10]. Replica selecting is a process of selecting replicas among the huge number of copies spreads across the grid. The process of selecting the grid site to place the replica is known as replica placement.

The process of creating and deleting replicas in data grid is known as replica management. An efficient replication management services is required to address these problems which offers high data availability, low bandwidth consumption, increased fault tolerance, and improved scalability of the overall system [18].

II. RELATED WORKS

Data mining techniques, such as association rules, frequent sequence mining, are used mainly to identify file correlations 12]. Replication strategies with data mining approaches [7].

In order to gain conceptual clarity of the domain under study various articles, books, websites and some of other private reports were examined.

The review has been conducted by focussing on the key topic of dynamic data management in grid computing. Leyli Mohammad Khanli, Ayaz Isazadeh et al., proposed [8] a new dynamic replication method in a multi-tier data grid called predictive hierarchical fast spread (PHFS) which is an extended version of fast spread (a dynamic replication method in the data grid).

Gui Liu, HaiLaing et al., proposed [3] a strategy called replication strategy based on clustering analysis (RSCA), which confirms the correlation among the data files accessed according to the access history of users.

Sang-Min Park, Jai-Hoon Kim et al., proposed [16] a novel dynamic replication strategy, called BHR, which reduces data access time by avoiding network congestions in a data grid network. Sasi and Thanamani proposed [11], a Modified BHR algorithm to overcome the limitations of the standard BHR algorithm. The performance of the proposed algorithm is improved by minimizing the data access time and avoiding unnecessary replication.

Tian Tian, Junzhou Luo et al., addresses this problem by introducing a new replica value determination method through data mining, and based on this method they proposed a pre-fetching based replication algorithm in a virtual organization (VO) environment in order to carry out a better replication optimization [15].

Lakshmi and Thanamani proposed [7] a new dynamic data replication strategy, called DMDR, which consider a set of files as granularity. Their strategy gathers files according to a relationship of simultaneous accesses between files by jobs and stores correlated files at the same site. Jianhua Jiang, Huifang et al., proposed [6] an associated replica replacement algorithm based on Apriori approach in data grid.

III. DATA MINING

In this section we proposed a new concept for the application of data mining combined with data grid replication.
The proposed method describes all the following processes: (i) The conversion from data grid to data mining and vice-versa. (ii) Based on the choices the data is extracted from data grid. Replication strategy must select suitable data mining approach in order to obtain correlations between files. So that it must deal with following three basic questions:

1) What are the information used in the strategy to understand file correlations?
2) Which is the suitable data mining technique to obtain file correlations based on the user information? And what kind of patterns extracted through this data mining process?
3) Finally how to use this extracted knowledge to enhance replication strategy performance.

IV. REPLICATION

Data replication is an efficient data management technique which is used for decades in many systems [1]. Data replication provides lot of benefits such as, we can get increased performance by strategic placement of replicas, by having multiple copies of data sets we can get improved availability and better fault-tolerance against possible failures of servers [2]. When tenant queries are submitted to the data management system, depending on the execution plan, e.g. the number of joins, they may require a number of relations in order to carry on with the execution. Naturally, in a large-scale environment where relations are fragmented and distributed geographically in multiple servers, not all required data may be present on the executing node itself. Considering that a query is processed on multiple servers according to inter-operator and intra-operator parallelism, the likelihood of some remote data to be shipped from faraway servers is a practical possibility [5]. In cases when the network bandwidth capability to the remote servers are not abundant, e.g. due to remote data being at a geographically separate location, a bottleneck that may ultimately lead to a response time dissatisfaction may occur during query execution process.
In order to ensure the satisfaction of query response time objective, the bottleneck data should be identified heuristically to be selected for possible replication before the query is even started executing. Also, when to trigger the actual replication event to start is another important decision that must be made for the same goal. Deciding how many replicas to create and how to retire the unused replicas must also be dealt with further down the road in the data replication decision process.
Strategic placement of the newly created replicas plays a key role in reducing data access latency and improving response time satisfaction. Undoubtedly, all of these replication decisions should be made from a cost-effective point of view to ensure the economic benefit of the provider, which is especially important in the economy-based large-scale systems such as cloud computing.
Dealing with the mentioned issues of data replication, a good data replication strategy must be able to decide in a meaningful way; (i) what to replicate to correctly determine which fragments of relations are in need of replication, (ii) when to replicate to be able to respond the change in demand of data in a timely manner to quickly resolve performance problems, (iii) how many replicas to create to avoid wasting precious resources such as storage to keep the costs down and retire unnecessary replicas accordingly, and finally (iv) where to replicate to strategically place newly created replicas to ensure tenant performance expectations are met and any possible penalties are avoided. Moreover, all of these decisions should be based on some criteria that are consistent with the aims of both the tenant and the provider.

V. DATA MINING BASED REPLICATION STRATEGIES

Data mining techniques based replication strategies are as follows. The strategies are grouped based on the data mining techniques they use.

A. Arra (Associated Replica Replacement Algorithm Based On Apriori Approach)

ARRA is indeed introduced in two parts. In first part, access behaviours of data intensive jobs are analyzed based on the APRIORI algorithm [13]. In another part, replica replacement rules are generated and applied. Data grid characteristics are transformed into data which is to be mined by APRIORI algorithm, accessed data files are considered as items, while the set of required data files by each data intensive job is regarded as a given transaction. The replica replacement rules can be summarized as follows:

1) If the frequency of an itemset is less than the minimum support, data files composing it should be replaced. Such itemsets having the least frequency will be replaced with high priority ones.
2) If some itemsets have the equivalent or almost the same frequency, itemsets with the smallest size will be replaced firstly.
3) The data file with the lowest confidence value will be replaced in its itemset.
4) Only one data file will be replaced at each time until there is enough storage size to satisfy the requirement of replica creation.

B. Bsca (Based On Support And Confidence Dynamic Replication Algorithm)

The main idea of this algorithm is to pre-fetch frequently accessed files and their associated files to the location near the access site. It is based on both frequently used association rule mining metrics, namely support and confidence, as well as on file access numbers. BSCA has two sub algorithms: replication algorithm and data mining algorithm [17].

1) The data mining algorithm applied to identify frequent sets of files is FP-GROWTH. Moreover, support and confidence of association rules between these frequent file sets are computed.

2) The replication algorithm is applied in each node as a decentralized manner to sorts the file serial containing the access history. It then finds out all files whose access numbers are greater than a predefined threshold. After that, the algorithm constructs sets of related files and replicates files that do not exist in the node. If free space is lacking, the algorithm deletes weakly correlated files. If the storage space is still insufficient, files whose access number is less than the predefined threshold will be deleted.

C. Pddra (A Pre-Fetching Based Dynamic Data Replication Algorithm)

A pre-fetching based dynamic data replication algorithm strategy consists of 3 steps [13]:

1) Storing file accessing patterns: In this step, file accessing sequences and data accessing patterns are stored in a database. A pre-fetching based dynamic data replication algorithm uses a structure for storing access sequences.

2) Requesting a file and performing replication and pre-fetching: In this step, a grid site asks for a file and replication is accomplished for it, if it is beneficial. Adjacent files, which are determined by mining the obtained tree of past access sequences, are also pre-fetched.

3) Replacement: When the storage space is insufficient to perform replication, some existing files will be selected for replacement. In this regard, a pre-fetching based dynamic data replication algorithm computes a value for each replica. The value is based on three factors, namely the number of access to the replica, the replication cost, and the time interval between the current time and the last access time of the replica. Then a pre-fetching based dynamic data replication algorithm uses a fuzzy logic-based formula to determine this value.

D. PHFS (Predictive Hierarchical Fast Spread)

Data mining predictive techniques is used to predict the future usage of files and then pre-replicate them in hierarchical manner on a path from source to client. The strategy considers spatial locality that is related files to the currently accessed file are the likely future requests. In this way, the related files to the current accessed file can be replicated in advance, as they will probably be the subsequent requests. Dependencies between files are inferred from previous access patterns using association rules and clustering techniques.

Predictive Hierarchical Fast Spread algorithm operates in three phases [17]:

1) A software agent in the root node collects the file access information from all the clients in the system and puts them in a log files.

2) Data mining techniques are then applied on log files with the aim to find strongly related files that are accessed together.

3) Whenever a client requests a file, Predictive Hierarchical Fast Spread algorithm finds the Predictive Working Set (PWS) of that file. The PWS is composed by the most related files to the requested file or, in other words, the predicted subsequent requests. Then, Predictive Hierarchical Fast Spread algorithm replicates all members of PWS along with the requested file on the path from the source to the client.

E. Pra (Pre-Fetching Based Dynamic Data Replication Algorithm)

The main concept of this algorithm is to make use of the characteristics that members in a virtual organization have similar interests in files to carry out a better replication optimization. The algorithm is described as following [14]: when a site does not have a file locally, then it requests a remote site. The remote site receives the request and transfers the file to the local site. At the same time, it finds the adjacent files of the requested file by applying frequent pattern sequence mining technique on the file access sequence data base. A message containing the list of adjacent files will be sent to the local site too. At last, the local site will choose adjacent files to replication. Pre-fetching based dynamic data replication algorithm was compared with No Replication and best client strategies and it is proved that it improves the average response time and the average bandwidth consumption. However a major drawback of the algorithm that it does not distinguish the different requests arriving from the different grid sites and considers them as successive file access.

F. Rscp (Replication Strategy Based On Maximal Frequent Correlated Pattern Mining For Data Grids)

In order to discover file correlations, the RSCP strategy relies on a maximal frequent correlated pattern mining algorithm. The proposed strategy is composed by four main steps [17]:

1) Extracting file access history: at a given period, each site keeps track of the access history for all local and remote files by the jobs executed on it.

2) Converting the file access history into an extraction context: the extraction context is a table containing Boolean values where accessed files are considered as items, while the set of required files by each job is regarded as a transaction.

3) Maximal frequent correlated patterns are then mined: in this phase, a maximal frequent correlated pattern mining algorithm is introduced in order to discover the hidden correlations between files.

4) The set of patterns identified in the previous stage constitutes the input of the replication algorithm. For each group of correlated files to be replicated, if there is enough storage space to hold all files in the group, then the replication of these correlated files will always take place. Otherwise, a replacement process should be carried out. For this purpose, the candidate files for deletion are selected according to their weight.

G. RSCA (Replication Strategy Based On Clustering Analysis)

The strategy is done in two stages. Initially, a clustering analysis is conducted on the file access history of all client nodes in the grid over a period of time [3]. The clustering method is used to group all the files that are similar according to a given equivalence relation. The outputs of this operation are correlated file sets related to the access habits of users [13]. Then at second stage, a replication is done on the basis of those sets, which achieves the aim of pre-fetching and buffering data. The experimental results using the OptorSim simulator proves that RSCA is effective in terms of average response time and bandwidth consumption compared to No Replication and Economy-based File Replication Strategy. Data mining is effective in identifying clusters of files related to the access habits of users. Instead of replicating individual files, an advanced replication optimization may consider correlated files for replication.

VI. ADVANTAGES OF REPLICATION STRATEGY

The advantages of replication strategies are [9]:

1) *Availability*: Suppose any failure happens in any site, immediately replicated data is stored as alternative. In this way replication strategy achieves the availability.

2) *Reliability*: The more the number of replicas increases the more probability so that user's request will be done completely.

3) *Scalability*: It is fully dependent on the type of architecture model which is selected for the Data Grid than replication strategy.

4) *Adaptability*: The user at any time can enter and quit a grid. The data replication strategy must be adaptive to the information of the grid environment in order to provide better results.

5) *Performance*: By storing replicas in multiple locations the user can easily access the data in data grid. So this is a way to increase the performance.

VII. PROPOSED GUIDELINE

At a glance, a strategy based on data mining technique should indeed be composed by three key steps:

First step: The grid data selection and pre-processing. In this respect, which data to consider in the grid data mining process is an important issue for which a right solution must be found. This indeed constitutes a key factor for the success of the whole process.

One must select the most relevant factors before starting the data mining process, to extract knowledge such as network performance prediction, file correlations, and associated replica. In the case of file correlations, for example, necessary information are mainly file access histories of grid sites.

Second step: The data mining step. Two main issues arise at this level: 1. What are the knowledge to extract by data mining and which type of learning to apply, i.e., supervised learning, unsupervised learning or semi-supervised data mining learning? This first issue is to decide which knowledge to extract from the data mining process. According to the availability of the historical data, we can select a proper type of learning. To get accurate result supervised learning algorithms required more number of labelled training data. Unsupervised learning methods are employed to discover structure in unlabeled data. Semi-supervised learning allows taking advantage of the strengths of both supervised and unsupervised learning by using simultaneously labelled and unlabeled data to build better learners, rather than using each one alone. 2. Which data mining technique to adopt in order to extract the desired knowledge, and what is the kind of patterns extracted through the data mining process? The second issue is related to the choice of the most suitable data mining technique to extract knowledge from historical data collected in the grid. However, these objectives

are often contradictory and closely depend on the available data to be mined. Moreover, according to the data mining technique adopted, the knowledge extracted can take many different forms: clusters, association rules, decision rules (resulting from decision tree), frequent (correlated) patterns, frequent sequences (resulting from sequential pattern mining), etc.

Third step: This is the last phase where replica selection or replication takes place based on the results of the data mining process. In this regard, how to use the extracted knowledge to enhance the data replication and replica selection strategies? The result of the data mining process constitutes the input of the replication or the replica selection algorithm. This step focuses on how the knowledge produced by the data mining technique will be re-injected through the replication or replica selection strategy in data grid, or in other words the transition from the data mining context to the data grid context. Obviously, the solution to this issue depends on the type of identified patterns. If, for example, the data mining technique used is decision tree, then the replication strategy must be able to translate the obtained decision rules to replication rules.

VIII. CONCLUSION

In this paper we have presented a survey of data mining based replica selection strategies and replica committed to data grids. The main aim of this work is to analysis how data mining techniques are applied to historical grid data to extract knowledge and use them to enhance data replication and replica selection strategies. A survey on replication strategies based on data mining techniques and a new guideline for data mining application behind data replication and replica selection strategies are the two main contributions are made in this work. Based on the survey we found that lot of data mining based replication strategies are available only thing is we need to choose an suitable technique so that we can enhance efficient data replication and replica selection strategies.

REFERENCES

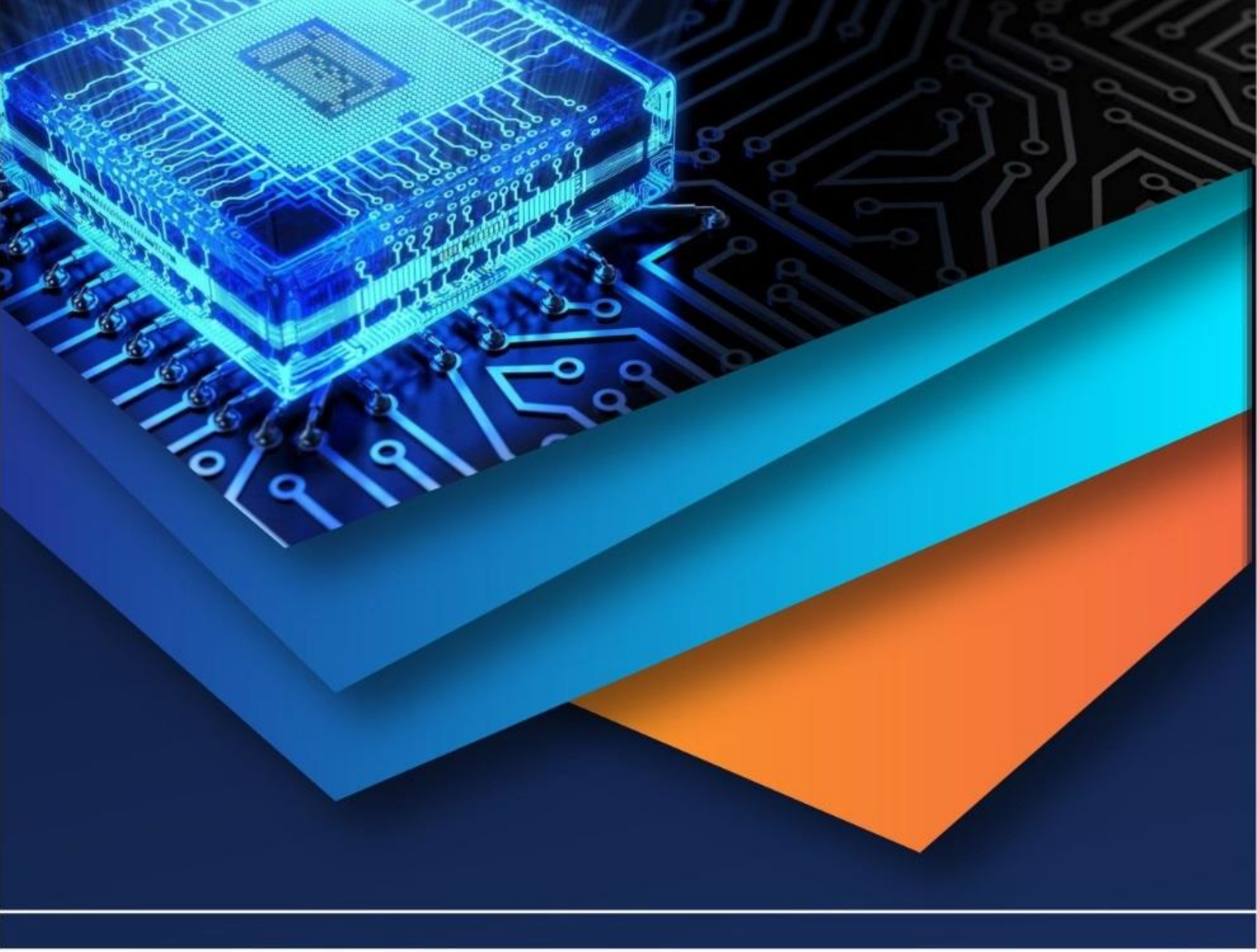
[1] Djebbara, Mohamed Redha et al., "Cloud data replication strategy using multiple-criteria decision analysis methods", Multiagent and Grid Systems, vol. 14, no. 2, pp. 203-218, 2018.
[2] Ghemawat, Sanjay & Gobioff, Howard & Leung, Shun-Tak. “The Google File System”, ACM SIGOPS Operating Systems Review. 37. 29-43, 2003.
[3] Gui Liu, HaiLiang Wei et al., “Research on Data Interoperability Based on Clustering Analysis in Data Grid”, International Conference on Interoperability for Enterprise Software and Applications China, IEEE, ISBN:978-0-7695-3652-1.
[4] Houda Lamehamedi, Ewa Deelman et al., "Data Replication Strategies in Grid Environments", Proc. 5th International Conference on Algorithms and Architecture for Parallel Processing, October 2002, IEEE Computer Society Press, pp. 378-383.
[5] Huebsch, Ryan & Garofalakis, Minos & Hellerstein, Joseph & Stoica, Ion. (2007). Sharing aggregate computation for distributed queries. 485-496. 10.1145/1247480.1247535.
[6] J. H. Jiang et al., "ARRA: An Associated Replica Replacement Algorithm Based on Apriori Approach for Data Intensive Jobs in Data Grid", Key Engineering Materials, Vols. 439-440, pp. 1409-1414, 2010.
[7] Lakshmi, Thanamani, “Performance Evolution of Dynamic Replication in a Data Grid using DMDR Algorithm”, International Journal of Engineering Research & Technology, ISSN: 2278-0181, Vol. 5, Issue. 10, 2016, pp. 389-394.
[8] Leyli Mohammad Khanli, Ayaz Isazadeh et al., “PHFS: A dynamic replication method, to decrease access latentcy in the multi-tier data grid”, Future Generation Computer Systems, Elsevier, 27(2011), pp.233-244.
[9] Najme Mansouri, Gholam Hosein Dastghaibyfard et al., "Combination of data replication and scheduling algorithm for improving data availability in Data Grids", Journal of Network and Computer Applications, ScienceDirect, Volume 36, Issue 2, March 2013, Pages 711-722.
[10] NajmeMansouri, Gholam Hosein, et al., "A dynamic replica management strategy in data grid", Journal of Network and Computer Applications, Volume 35, Issue 4, July 2012, pp.1297-1303.
[11] Sashi, Thanamani, “Dynamic replication in a data grid using a Modified BHR Region Based Algorithm”, Future Generation Computer Systems, Elsevier, 27(2011), pp.202-210.
[12] Seema Maitrey, C.K. Jha, "Association Rule Mining: A Technique for Revolution in Requirement Analysis, International Journal of Scientific and Research Publications, Volume 4, Issue 8, August 2014, ISSN 2250-3153.
[13] Tarek Hamrouni, Sarra Slimani et al., "A data mining correlated patterns-based periodic decentralized replication strategy for data grids", Journal of Systems and Software, ScienceDirect, Volume 110, December 2015, Pages 10-27.
[14] Tarek Hamrouni, SarraSlimani et al., "A Critical Survey of Data Grid Replication Strategies Based on Data Mining Techniques", Procedia Computer Science, ScienceDirect, Volume 51, 2015, Pages 2779-2788.
[15] Tian Tian, Junzhou Luo et al., “A Prefetching-based Replication Algorithm in Data Grid”, Third International Conference on Pervasive Computing and Applications, IEEE.
[16] Tinghuai Ma, QiaoqiaoYan et al., "Replica creation strategy based on quantum evolutionary algorithm in data gird", Knowledge-Based Systems, Volume 42, April 2013, Pages 85-96.
[17] T.Hamrouni, S.Slimani et al., "A survey of dynamic replication and replica selection strategies based on data mining techniques in data grids", Engineering Applications of Artificial Intelligence, ScienceDirect, Volume 48, February 2016, Pages 140-158.
[18] Yi Ren, Xun & Chuan Wang, Ru & Dong Ma, Xiao. (2011), “Improvement on Reading and Writing for Replica Consistency”, Advanced Materials Research. 187. 232-236. 10.4028/www.scientific.net /AMR.187.232.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue II, Feb 2019- Available at www.ijraset.com

About the Authors

Ms. Prema .R is presently working as Head, Dept of Computer Applications, Indo Asian Women's Degree College, India (affiliated to Bangalore University,Bangalore . She is pursuing her Ph.D in Bahrathiar University, Coimbatore. Her areas of interest include Algorithm Analysis and design, Data Mining, Data Sturctures.Operation Research. She has 10 years of teaching experience.

Dr. Antony Selvadoss Thanamani is presently working as Professor and Head, Dept of Computer Science, NGM College, Coimbatore, India (affiliated to Bharathiar University, Coimbatore). He has published more than 100 papers in international/ national journals and conferences. He has authored many books on recent trends in Information Technology. His areas of interest include ELearning, Knowledge Management, Data Mining, Networking, Parallel and Distributed Computing. He has to his credit 32 years of teaching and research experience. He is a senior member of International Association of Computer Science and Information Technology, Singapore and Active.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY