Procedia Computer Science

Volume 51, 2015, Pages 2779–2788

ICCS 2015 International Conference On Computational Science

# A critical survey of data grid replication strategies based on data mining techniques

Tarek Hamrouni, Sarra Slimani, and Faouzi Ben Charrada

Department of Computer Sciences, Faculty of Sciences of Tunis
Tunis El Manar University, Tunisia
tarek.hamrouni@fst.rnu.tn, sarra.slimani@gmail.com, f.bencharrada@gmail.com

**Abstract**

Replication is one common way to effectively address challenges for improving the data management in data grids. It has attracted a lot of work and many replication strategies have therefore been proposed. Most of these strategies consider a single file-based granularity and do not take into account file access patterns or possible file correlations. However, file correlations become an increasingly important consideration for performance enhancement in data grids. In this regard, the knowledge about file correlations can be extracted from historical and operational data using the techniques of the data mining field. Data mining techniques have proved to offer a powerful tool facilitating the extraction of meaningful knowledge from large data sets. As a consequence of the convergence of data mining and data grid, mining grid data is an interesting research field which aims at analyzing grid systems with data mining techniques in order to efficiently discover new meaningful knowledge to enhance data management in data grids. More precisely, in this paper, the extracted knowledge is used to enhance replica management. Gaps in the current literature and opportunities for further research are presented. In addition, we propose a new guideline to data mining application in the context of data grid replication strategies. To the best of our knowledge, this is the first survey mainly dedicated to data grid replication strategies based on data mining techniques.

*Keywords:* data grid, replication strategy, files correlation, data mining, mining grid data, guideline

## 1 Introduction and motivations

One of the biggest challenges that data grids users have to face today consists in the improvement of the data management. Indeed, the used techniques must scale up while addressing the autonomy, dynamicity and heterogeneity of the data sources. In this regard, it is not new that understanding the system features, user behavior, or frequent patterns can help on achieving both efficient management and a better performance.

We deem that in order to discover knowledge and to enhance grid systems performance, new techniques and tools for extracting these useful knowledge and information are required. In this

Selection and peer-review under responsibility of the Scientific Programme Committee of ICCS 2015



doi:10.1016/j.procs.2015.05.434

respect, data mining techniques have proved to be a powerful tool facilitating the extraction of meaningful knowledge, hidden patterns, associations or anomalies from large data sets. Hence, by mining historical grid data using the techniques of the data mining field, valuable knowledge can be extracted. The knowledge discovered would enable grid systems to an improved management in many areas. In this paper, specifically, we focus on how extracted useful knowledge enables enhancing data replication strategies. Replication is one way to effectively address the challenges for improving the data management in data grids. Many strategies have been proposed in the literature and several surveys are conducted [2, 10, 13, 23]. However, most of existing replication techniques are based on single file granularity. They are indeed confined to identify popular files based on file access patterns observed at application runtime, and do not take into account file correlations or file access patterns. Actually, in many applications, data files may be correlated in terms of accesses and have to be considered together. In this respect, the analysis of data usage in a large set of real traces from a high-energy physics collaboration [12] in a typical data grid, namely DZero[1], revealed the existence of correlations between files. This led the authors to propose a new granularity for data management, called "filecule". Filecule is defined as a group of files always accessed together by jobs. In the same context, Ko *et al.* [18] by analyzing a real data-intensive grid application, called Coadd, found that there is a strong correlation between requested files and that jobs tend to demand groups of correlated files.

In the general case, knowledge about file correlations, *i.e.*, groups of files always requested together, file access patterns as well as future needs of grid sites must be efficiently considered in the replication process. Hence, our main idea is that hidden patterns and associations that are discovered by data mining techniques would enable: (*i*) to predict the future usage patterns based on the historical data, (*ii*) to identify groups of correlated files that are always requested together by users, jobs or sites, (*iii*) to adapt dynamically with the user access pattern usage. The discovered knowledge can be used, for example, to pro-actively pre-fetch data. Numerous works on data replication and caching techniques, in different distributed contexts, have shown that pro-actively pre-fetching files can significantly improve overall distributed system performance [11]. On the other hand, identified groups of correlated files may be used as granularity for replication. Many works have shown the efficiency of replication strategies based on correlated files than traditional ones, *i.e.*, those based on file granularity. Indeed, co-location of correlated files can help not only improving response time but also minimizing resource consumption. For example, in [16], authors address the problem of minimizing average query span, *i.e.*, the average number of machines that are involved in the processing of a query through co-location of related data items. It results in improving the effectiveness and efficiency of the replication process.

Even though data mining has been applied in numerous areas and sectors, the application of data mining to replication in data grids contexts is still limited. In this respect, only few works, when comparing their number with that of all replication strategies, have used data mining techniques to explore file correlations although the strength of data mining approaches in real-life applications. Our survey diverges totally from other surveys by mainly focusing on several replication strategies based on data mining techniques.

In this paper, we focus on how data mining techniques can improve performance of data grid replication strategies. To do this, in the next section, we survey the main replication strategies based on data mining techniques. In the third one, we analyze these strategies by focusing on their strengths and their drawbacks and how data mining techniques are applied. In the fourth section, we propose a new guideline to data mining application in the context of data

[1]The DZero experiments: http://www-d0.fnal.gov

grid replication strategies. This guideline describes finely directive lines to be followed when designing a new replication strategy based on the results of data mining techniques. Also, it describes in detail the main steps, the decisions to be made, the specificities and the constraints to be taken into account stemming from both data grid and data mining contexts. Finally, in the fifth section, we summarize our contributions and depict future work to be done. To the best of our knowledge, this is the first survey focusing on replication strategies based on data mining techniques.

# 2 Replication strategies based on data mining techniques

In this section, replication strategies based on data mining techniques are succinctly presented in ascending order by year of publication. We then mainly concentrate on analyzing in depth these strategies and highlighting the new proposed guideline in the following sections. We refer readers to [14, 31] for a detailed study on data mining techniques.

## 2.1 A pre-fetching based dynamic replication in data grid

The main idea of the PRA strategy [30] 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: when a site $S_i$ does not have a file locally, it requests a remote site $S_j$. This latter receives the request and transfer the file to the former 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. At last, a message containing the list of adjacent files will be sent to the site $S_i$ who will choose adjacent files to replication. PRA was compared with No Replication and Best Client strategies [25] and it is proved that it improves the average response time and the average bandwidth consumption.

## 2.2 Replication strategy based on clustering analysis

The RSCA strategy [21] is based on the existence of correlations among the data files accessed according to the access history of grid users. At the first stage, a clustering analysis is conducted on the file access history of all client nodes in the grid over a period of time. The outputs of this operation are correlated file sets related to the access habits of users. At the second stage, a replication is done on the basis of those sets, which achieves the aim of pre-fetching and buffering data. The clustering method adopted is used to group into equivalence classes all the files that are similar according to a given equivalence relation. The set of files in the same equivalence class are called correlative file sets. The experimental results using the OptorSim simulator show that RSCA is effective in term of average response time and bandwidth consumption compared to No Replication [25] and economy-based file replication strategies [5].

## 2.3 Predictive file replication on the data grids

In [20], Liao proposed a model to predict the popularity of a given file, *i.e.*, assessing its frequency (high or low), from some known file attributes. The model is built using the incremental decision tree technique. First, a trace analysis is performed with the aim of obtaining historical information of user behavior in the data grid in a given period. Then, from this analysis, a statistical measure of the association between each file attribute and the file access frequency is carried on using the chi-squared test in order to identify the most significant attributes, *i.e.*,

those having a strong correlation with the file access frequency. The replication decision rules are then derived from the built tree. When a file is requested and there is no sufficient space for storage, the proposed strategy checks the decision rule table converted from the decision tree model. If its predicted rank of future popularity is "higher", the file will be replicated. Otherwise, the file will be accessed remotely. This proposed predictive replication strategy outperforms the LRU strategy [4] and the economy-based file replication strategies [5] under sequential and Zipf access patterns.

## 2.4 Associated replica replacement algorithm based on Apriori approach

The ARRA strategy [15] is introduced in two parts. In the first part, access behaviors of data intensive jobs are analyzed based on the Apriori algorithm [1]. In the second part, replica replacement rules are generated and applied. Accessed data files are considered as items of the database mined through Apriori, while each transaction is composed by the required data files of each data intensive job. Simulation results show that ARRA has a relative advantage in mean job time of all jobs, number of remote file access and effective network usage compared with LFU [3]. Noteworthily, in [24], a more recent strategy based on Apriori for association rule mining has been proposed.

## 2.5 Predictive hierarchical fast spread

By considering spatial locality, PHFS [17] uses predictive techniques to predict the future usage of files and then pre-replicates them in hierarchical manner on a path from the source to the client in order to increase locality in access. File correlations are inferred from previous access patterns by association rules and clustering techniques of data mining. PHFS operates in three steps. In the first, file access information are collected in the root node. In the second step, data mining techniques are applied on log files. Finally, whenever a client requests a file, PHFS finds the predicted subsequent requests after this request. Let us notice that the authors just used instances to compare the access latency of their strategy with Fast Spread [25].

## 2.6 A pre-fetching based dynamic data replication algorithm

The main idea of PDDRA [26] is the same of the PRA strategy (*cf.* section 2.1). Based on file access history, PDDRA predicts future needs of grid sites and pre-fetches a sequence of files to the requester grid site. As a consequence, the next time that this site needs a file, it will be locally available. PDDRA consists of three phases: storing file access patterns, requesting a file, and finally performing replication, pre-fetching and replacement. The simulation results using the OptorSim simulator show that PDDRA has better performances in comparison with six other strategies namely No Replication [25], LRU [4], LFU [3], EcoModel, EcoModel Zipf-like distribution [5] and PRA [30], in terms of job execution time, effective network usage, total number of replications, hit ratio and percentage of storage filled.

## 2.7 Based on support and confidence dynamic replication algorithm in multi-tier data grid

The major idea of the BSCA strategy proposed in [8, 9] is to pre-fetch frequently accessed files and their associated files to the location near the access site. It finds out the correlation between the data files through data access number and data access serial. In addition of the

use of the access numbers, BSCA is also based on the support and the confidence measures used for mining association rules. This strategy has two sub algorithms: data mining algorithm and replication algorithm. Once the data mining algorithm is applied to identify frequent files, support and confidence of association rules between these frequent files are computed. If the support and the confidence values between files exceed respective minimum thresholds, frequent files and their associated ones are replicated. Using the OptorSim simulator, BSCA outperforms the SBU [29], ABU [29] and Fast Spread [25] replication strategies. This is done through giving the lowest average response time. The average response time is the only measure of evaluation used to compare BSCA with the other strategies.

### 2.8 Replication strategy based on maximal frequent correlated pattern mining for data grids

Slimani *et al.* proposed in [27] a new strategy, called RSBMFCP, which is composed by four steps: (*i*) extracting file access history. (*ii*) converting the file access history into an extraction context (*i.e.*, a logical file access history). (*iii*) mining maximal frequent correlated patterns in order to discover the hidden file correlations. Indeed, each maximal frequent correlated pattern represents a maximal set of files frequently appearing simultaneously and whose correlation degree exceeds a given minimal threshold of a dedicated correlation measure. (*iv*) performing replication process and replacement. RSBMFCP outperforms the DR2 [28], Periodic Optimiser [6] and DPRSKP [7] strategies in terms of job execution time and effective network usage.

## 3 Analysis and discussion

Table 1 summarizes the properties of the surveyed replication strategies based on data mining techniques. The criteria used for classifying them are as follows: (*i*) Periodicity: the strategy is periodic or not, *i.e.*, carried out after a given period of time or at each demand of a requested. (*ii*) Decision making process: the strategy is centralized or decentralized. (*iii*) Adopted data mining method. (*iv*) Data used in the data mining process. (*v*) Main parameters used in the replication strategy. (*vi*) Storage space capacity assumed as limited or unlimited.

Due to lack of available space, we only focus on three main observations that emerge from Table 1. A first observation is that the association rule mining is the most used technique for exploring file correlations with three works adopting this technique. In these strategies, two measures of association rule quality assessment are used, namely the support and the confidence measures. However, various studies have shown the limits of association rule mining based on these quality measures. Indeed, the resulting set of association rules has an excessively large size, with a majority of the mined rules either redundant or do not reflect the true correlation relationship among data [19]. The RSBMFCP [27] tries to overcome this limitation by proposing the use of the *all-confidence* correlation measure in the data mining process. A second observation concerns the data used to infer file correlations. We note that most strategies, consider the file access pattern of sites. In this situation, the database to be mined is composed by sites in lines and accessed files in columns. Actually, jobs executed in sites access the files. So considering a history of file access job is in our opinion a more reasonable choice to infer semantic relationships between files. A third observation concerns the parameters taken into account by strategies. We note that there are two types of parameters. On the one hand, we find file associated parameters like number of requests and file size. On the other hand, grid topology parameters like bandwidth and throughput are used. We can note that some strategies like RSCA neglect parameters of this latter type although they are of great importance.

| Strategy | Periodicity | Type of decision making | Data mining technique | Data used in the data mining process | Parameters | Storage space assumed |
|---|---|---|---|---|---|---|
| A pre-fetching-based replication algorithm: PRA (2008) | non periodic | decentralized | frequent sequence mining | sites/accessed files | file request number, minimum frequency threshold, minimum confidence threshold | limited |
| Replication strategy based on clustering analysis: RSCA (2009) | periodic | centralized | clustering | sites/accessed files | file request number, frequency threshold | unlimited |
| Decision tree based predictive model (2010) | non periodic | decentralized | decision tree | file access frequency, userID, file type | file access number, access delay, number of transmitted bits, transmission rate | limited |
| Associated replica replacement algorithm based on Apriori Approach: ARRA (2010) | non periodic | decentralized | association rules mining | jobs/accessed files | file request number, minimum frequency threshold, minimum confidence threshold | limited |
| Predictive hierarchical fast spread: PHFS (2011) | periodic + non periodic | decentralized | association rules mining, clustering | sites/accessed files | file access number, minimum confidence threshold | limited |
| A pre-fetching based dynamic data replication algorithm: PDDRA (2012) | non periodic | decentralized | frequent sequence mining | sites/accessed files | file size, bandwidth, file access number, minimum frequency threshold | limited |
| Based on support and confidence dynamic replication algorithm: BSCA (2013) | periodic | decentralized | association rules mining | sites/accessed files | file access number, minimum frequency threshold, minimum confidence threshold | limited |
| Replication strategy based on maximal frequent correlated pattern mining: RSBMFCP (2014) | periodic | decentralized | maximal frequent correlated pattern mining | jobs/accessed files | file request number, minimum frequency threshold, minimum *all-confidence* threshold, bandwidth, file size | limited |

Table 1: Comparative table of replication strategies based on data mining techniques

# 4 Guideline for data mining based replication strategies

In this section we propose a new guideline to present a roadmap for the application of data mining in data grid replication. The proposed guideline fully describes all steps of the process including: (*i*) the translation from the data grid context to the data mining one and vice-versa (*ii*) the choices to be taken into account and constraints stemming from the data grid and the data mining contexts. It models the answers to three basic questions that a replication strategy taking into account file correlations and based on a data mining approach must deal with:
1. What are the information used in the strategy to infer file correlations?
2. Which data mining technique is the most adopted to infer file correlations according to the used information? and what is the kind of patterns extracted through the data mining process?
3. How to use the extracted knowledge to enhance the replication strategy performance?

In fact, before starting a data mining process in order to extract useful knowledge, such as file correlations in grid, one must select the most appropriate information on which the process of data mining will be based. These information, are basically, file access histories in the grid. This key step represents the transition from the data grid context to the data mining context.

Hence, this information may require being preprocessed. For example, in case of association rules, an extraction context composed by rows and columns should be constructed starting from the collected information using for example the access history of files by jobs.

The second question is related to the choice of the most appropriate data mining technique in order to extract knowledge from historical data collected in the grid. We deem that this choice results in three main objectives to be satisfied: extracting knowledge with high quality, as rapidly as possible, while improving the performance of the replication process. However, these objectives are often contradictory. It is therefore of paramount importance to find the best trade-off between the quality of the extracted knowledge and the performance of the strategy. Moreover, according to the data mining method adopted, the knowledge extracted can take many different forms: association rules, decision rules (resulting from decision tree), frequent (correlated) patterns, sequences (resulting from a sequences mining algorithm), etc.

Finally, the results of the data mining process form the input of the replication algorithm. This step focuses on how the knowledge produced by the data mining technique will be re-injected through the replication strategy in data grid, or in other words the transition from the data mining context to the data grid context. Obviously, the answer to this question depends on the type of identified patterns. If, for example, the data mining technique used is decision tree, then replication strategy must translate the obtained decision rules to replication rules.

It is worth to mention that the most of the existing strategies do not respond carefully and simultaneously to all questions that we had set in this guideline. A replication strategy based on data mining technique should indeed be composed by three key steps: a data selection and preprocessing step, a data mining step, and finally a replication step based on results of the data mining process. Now, let us describe the choices that should be carefully made in each step of the replication strategy. There is at least one choice per step that we describe in the following:

• **First step**: two choices: (*i*) choice of data to be considered for file correlation extraction. (*ii*) choice of data preprocessing method: how to transform grid data in order to make them suitable for a data mining process, especially that results of data mining closely depend on the quality of the resulting transformed data.

• **Second step**: in this step, choices are as follows:

1. Choice of the most adequate data mining algorithm.

2. Choice of the correlation measure: this choice is closely tied to the data mining algorithm performance. For this reason a multitude of correlation measures were proposed in data mining literature. In short, we deem that the correlation measure must afford to properly assess the correlation degree between the files and be as simple as possible to provide opportunities for optimized calculations of correlated files sets.

3. Choice of correlation measure thresholds: most of the works do not attach importance to these crucial parameters. However, the choice of thresholds associated with the measures used is very important and must be done carefully in order to avoid grouping files that are not correlated enough and therefore degrade the performance of the replication strategy.

4. Adopting a decentralized or centralized data mining (*i.e.*, executed in a central grid site): the data grid infrastructure is very adequate to the execution of decentralized data mining applications. On the contrary, in the data grid context, the choice of a centralized approach for executing the adopted data mining technique does not appear suitable. Indeed, several gaps arise like: (*i*) how to collect all information from all sites of the grid? (*ii*) how to maintain them in a central site? (*iii*) how to deal with the storage space of the central site that will store the history of all the sites of the grid? Indeed, this history requires a large storage space. The chosen site must thus have a large storage capacity. (*iv*) how to face the data mining algorithm

response time which will be influenced by the large size of the history?

5. Adopting a periodic or non periodic data mining: the data mining algorithm can be triggered at each file request or at each period. In the first case, the data mining algorithm will be executed frequently. Indeed, commonly applications access to a set of files which are distributed over the various nodes of the grid. Often these files are not available in the storage element where the application runs. Hence, many file requests are generated in the grid. In this case, invoking the data mining algorithm a number of times equal to the number of file requests may obviously negatively affect the performance of the replication strategy and, in particular, its response time. In this sense, PRA and PDDRA strategies have not paid attention to this point. Indeed, at each file request, the frequent sequence mining algorithm is invoked to predict the future requests. It is then preferred to trigger the data mining algorithm at each period of time whose the duration must be finely determined.

• **Third step**: choice of the way of using the extracted knowledge through the data mining algorithm in the replication strategy. Generally, the extracted knowledge will be used for pro-actively pre-fetch data or for data replication. In the first case, in each file request, the strategy will predict future file requests based on the results of the data mining algorithm. In the second case, file replication can be decided based on predicted value of an important aspect of file (such as popularity) or periodically correlated files will be used as granularity for replication.

# 5 Conclusion and future work

We have presented in this paper a critical survey of data mining-based replication strategies dedicated to data grids. The main objective of this work consists in the study of how the data mining techniques can be applied to access historical data of data grids and how do they infer file correlations knowledge and use them to enhance replication strategies performance. Two contributions are made in this article: on the one hand, a critical survey of the main replication strategies based on data mining techniques and, on the other hand, a new guideline to data mining application in the context of data grid replication strategies. We suggest that this guideline would facilitate further research works in this promising area and to give hints to other works to be done in the area of data mining and data grid replication. From this survey, it can be seen that the number of the proposed replication strategies based on data mining techniques is limited and so there is still a lot of work to be done in the field of data replication based on data mining. Some future research works are discussed below.

We can note that all of the existing methods for mining file correlations either rely on file access sequence or file attributes. We suggest that more complex file correlations by judiciously combining access sequence mining with semantic file attribute mining can be inferred. It will be very interesting to future proposals to be oriented towards this direction. Another original future direction that we propose is to use data mining graph techniques in data grid to explore grid site clusters, *i.e.*, groups of sites (profiles of users) or clusters that have common interests on grid files. Indeed, such techniques have been proved to be very efficient for resolving problems that can be modeled using graphs like those related to social networks, etc. In this respect, the data grid sites can be considered as nodes of a graph while links between sites as edges.

From an experimental point of view, although file correlations are a common patterns observed in real data grid systems [22], like high-energy physics collaboration, and that there are a variety of tools that enable monitoring these grid systems, all strategies use simulation to evaluate and test the replication algorithms. Hence, it would be very interesting as a next step to test their strategies in a real grid environment. In addition, it has been observed that most of strategies compare the results with the very basic ones. Hence a lot of experiments

are still required for thoroughly assessing their performances. This issue is included in our future work. Indeed, we plan in this respect to perform a quantitative study of these strategies through re-implementing them. This will allow setting an open access platform for replication strategies implementations that can be extended through researchers aiming at making their implementations easily available for interested users. In this same context, the most of the strategies discussed in this paper do not study the impact of different considered parameters on performance evaluation results. Indeed, parameters such as period or in other words the history length (a small history windows or a long one) and thresholds for valid correlation inferring are very important and have a great impact on the obtained results. Furthermore, since the execution of replication strategy based on data mining technique depends also on the execution time of data mining algorithm, data mining overhead on strategy performances must be assessed.

# References

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very Large Data Bases, Santiago, Chile*, pages 478–499, 1994.

[2] T. Amjad, M. Sher, and A. Daud. A survey of dynamic replication strategies for improving data availability in data grids. *Future Generation Computer Systems*, 28(2):337 – 349, 2012.

[3] W. H. Bell, D. G. Cameron, L. Capozza, A. P. Millar, K. Stockinger, and F. Zini. Simulation of dynamic grid replication strategies in OptorSim. In *Journal of High Performance Computing Applications*, pages 46–57, 2002.

[4] W. H. Bell, D. G. Cameron, L. Capozza, A. P. Millar, K. Stockinger, and F. Zini. OptorSim - A grid simulator for studying dynamic data replication strategies. *International Journal of High Performance Computing Applications*, 17(4):403–416, 2003.

[5] W. H. Bell, D. G. Cameron, R. Carvajal-Schiaffino, A. P. Millar, K. Stockinger, and F. Zini. Evaluation of an economy-based file replication strategy in data grids. In *Proceedings of Third International Symposium on Cluster Computing and the Grid*, pages 661–668, 2003.

[6] F. Ben Charrada, H. Ounelli, and H. Chettaoui. An efficient replica placement strategy in highly dynamic data grids. *International Journal of Grid and Utility Computing*, 2(2):156–163, 2011.

[7] H. Chettaoui and F. Ben Charrada. A decentralized periodic replication strategy based on knapsack problem. In *Proceedings of the 13th International ACM/IEEE Conference on Grid Computing*, pages 3–13, 2012.

[8] Z. Cui, D. Zuo, and Z. Zhang. Based on support and confidence dynamic replication algorithm in multi-tier data grid. *International Journal of Computational Information Systems*, 9(10):3909–3918, 2013.

[9] Z. Cui, D. Zuo, and Z. Zhang. Based on the correlation of the file dynamic replication strategy in multi-tier data grid. *International Journal of Database Theory and Application*, 8(1):75–86, 2015.

[10] S. Dayyani and M. R. Khayyambashi. A comparative study of replication techniques in grid computing systems. *International Journal of Computer Science and Information Security*, 11(9), 2013.

[11] B. Dong, X. Zhong, Q. Zheng, L. Jian, J. Liu, J. Qiu, and Y. Li. Correlation based file prefetching approach for hadoop. In *Proceedings of the 2nd IEEE International Conference on Cloud Computing Technology and Science*, pages 41–48, 2010.

[12] S. Doraimani. *Filecules: A New Granularity for Resource Management in Grids*. Master thesis, University of South Florida, USA, 2007.

[13] R. Kingsy Grace and R. Manimegalai. Dynamic replica placement and selection strategies in data grids a comprehensive survey. *Journal of Parallel and Distributed Computing*, 74(2):2099 – 2108, 2014.

[14] J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, 2011.

[15] J. Jiang, H. Ji, G. Xu, and X. Wei. ARRA: an associated replica replacement algorithm based on Apriori approach for data intensive jobs in data grid. *Key Engineering Materials*, 439-440:1409–1414, 2010.

[16] A. K. Kayyoor, A. Deshpande, and S. Khuller. Data placement and replica selection for improving co-location in distributed environments. *Computing Research Repository (CoRR)*, 2013.

[17] L. M. Khanli, A. Isazadeh, and T. N. Shishavanc. PHFS: A dynamic replication method, to decrease access latency in the multi-tier data grid. *Future Generation Computer Systems*, 27(3):233–244, 2011.

[18] S. Y. Ko, R. Morales, and I. Gupta. New worker-centric scheduling strategies for data-intensive grid applications. In *Proceedings of the 8th ACM/IFIP/USENIX International Conference on Middleware, Newport Beach, CA, USA*, pages 121–142, 2007.

[19] Y. K. Lee, W. Y. Kim, Y. D. Cai, and J. Han. CoMine: efficient mining of correlated patterns. In *Proceedings of the 3rd IEEE International Conference on Data Mining, Melbourne, Florida, USA*, pages 581–584, 2003.

[20] C. Liao, N. Helian, S. Wu, and M. Rashid. Predictive file replication on the data grids. *International Journal of Grid and High Performance Computing*, 2(1):69–86, 2010.

[21] G. Liu, H. Wei, X. Wang, and W. Peng. Research on data interoperability based on clustering analysis in data grid. *Proceedings of the International Conference on Interoperability for Enterprise Software and Applications, Beijing, China*, pages 97–103, 2009.

[22] J. Ma, W. Liu, and T. Glatard. A classification of file placement and replication methods on grids. *Future Generation Computer Systems*, 29(6):1395 – 1406, 2013.

[23] R. Mokadem and A. Hameurlain. Data Replication Strategies with Performance Objective in Data Grid Systems: A Survey. *International Journal of Grid and Utility Computing*, 6(1):30–46, 2015.

[24] M. S. Q. Z. Nine, M. A. K. Azad, S. Abdullah, M. A. H. Monil, I. Zahan, A. Bin Kader, and R. M. Rahman. Application of association rule mining for replication in scientific data grid. In *Proceedings of the IEEE 8th International Conference on Electrical and Computer Engineering*, pages 345–348, 2014.

[25] K. Ranganathan and I. Foster. Identifying dynamic replication strategies for a high performance data grid. In *Proceeding of the 2nd International Workshop on Grid Computing*, pages 75–86, 2001.

[26] N. Saadat and A. M. Rahmani. PDDRA: A new pre-fetching based dynamic data replication algorithm in data grids. *Future Generation Computer Systems*, 28(4):666–681, 2012.

[27] S. Slimani, T. Hamrouni, and F. Ben Charrada. New replication strategy based on maximal frequent correlated pattern mining for data grids. In *Proceedings of the 15th international Conference on Parallel and Distributed Computing, Applications and Technologies*, pages 144–151, 2014.

[28] P. K. Suri and M. Singh. DR2: A two-stage dynamic replication strategy for data grid. *Journal of Recent Trends in Engineering*, 2(4):201–203, 2009.

[29] M. Tang, B-S. Lee, C. K. Yeo, and X. Tang. Dynamic replication algorithms for the multi-tier data grid. *Future Generation Computer Systems*, 21(4):775–790, 2005.

[30] T. Tian, J. Luo, Z. Wu, and A. Song. A pre-fetching-based replication algorithm in data grid. In *Proceedings of the 3rd International Conference on Pervasive Computing and Applications*, pages 526–531, 2008.

[31] M. J. Zaki and W. Meira Jr. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, 2014.