**American Journal of Engineering Research (AJER)**
e-ISSN: 2320-0847 p-ISSN : 2320-0936
Volume-4, Issue-7, pp-256-260
www.ajer.org
Research Paper Open Access

# DataMining with Grid Computing Concepts

Mohammad Ashfaq Hussain [1], Mohammad Naser [1], AhmedUnnisa Begum [1], Naseema Shaik [2], Mubeena Shaik [1]
*[1] Jazan University, Jazan, Kingdom of Saudi Arabia*
*[2] King Khalid University, Kingdom of Saudi Arabia*

***ABSTRACT:*** *Now days the organizations often use data from several resources. Data is characterized to be heterogeneous, unstructured and usually involves a huge amount of records. This implies that data must be transformed in a set of clusters, parts, rules or different kind of formulae, which helps to understand the exact information. The participation of several organizations in this process makes the assimilation of data more difficult. Data mining is a widely used approach for the transformation of data to useful patterns, aiding the comprehensive knowledge of the concrete domain information. Nevertheless, traditional data mining techniques find difficulties in their application on current scenarios, due to the complexity previously mentioned. Data Mining Grid tries to fix these problems, allowing data mining process to be deployed in a grid environment, in which data and services resources are geographically distributed belong to several virtual organizations and the security can be flexibly solved. We propose both a novel architecture for Data Mining Grid, named DMG.*

***KEYWORDS:*** *clusters, parts, rules, formulae, grid environment, Data Mining Grid.*

## I. INTRODUCTION

DataMining refers to the process of extracting useful, handy and survivable knowledge from data. The extracted knowledge useful in many areas such as business applications like financial business analysis, purchasing behavior scenarios and also in biology, molecular design, weather forecast, climate prediction, physics, fluid dynamics and so on. Now the challenge in these applications is to mine data located in distributed, heterogeneous databases while adhering to varying security and privacy constraints imposed on the local data sources.

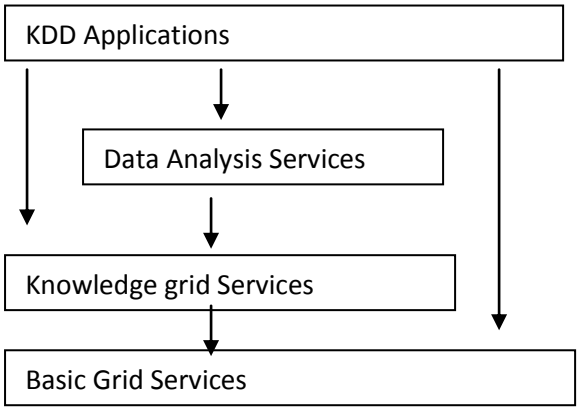
The term grid can be defined as a set of computational resources interconnected through a WAN, aimed at performing highly demanding computational tasks such as internet applications. A grid makes it possible to securely and reliably take advantage of widely dispersed computational resources across several organizations and administrative domains. The aim of grid computing is to provide an affordable approach to large-scale computing problems.

Grid technology provides high availability of resources and services, making it possible to deal with new and more complex problems. But it is also known that a grid is a very heterogeneous and decentralized environment [8]. It presents different kinds of security policy, system administration procedure, data and computing characteristic and so on. In this juncture we can't say that any grid is not just a data mining grid. It is a very important aspect in maintaining grid systems. Grid management is the key to providing high reliability and quality of service [9]. The complexities of grid computing environments make impossible to have a complete understanding of the entire grid. Therefore, a new approach is needed. Such an approach should pool, analyze and involve all relevant information that could be obtained from a grid. The insights provided should then be used to support resource management and system involvement. Data mining has proved to be a phenomenally robust tool, which smoothen the analysis and interpretation of large volumes of complex data. It results, given the complexities involved in operating and maintaining grid environments efficiently and the ability of data mining to analyze and interpret large volumes of data, it is obvious that 'mining grid data' could be a solution to improving the performance, operation and maintenance of grid computing environments.

## II. GRID ENVIRONMENT AND KDD

The Knowledge Grid framework makes use of basic grid services to build more specific services supporting distributed knowledge discovery in databases (KDD) on the grid Such services allow users to implement knowledge discovery applications that involve data, programming and mathematical resources available from distributed grid views. To this end, the Knowledge Grid defines mechanisms and higher-level services for publishing and searching information about resources, representing and managing KDD applications [7], and also for managing their results.

## III. THE SIMPLE WAY TO REPRESENT GRID WITH KDD




The above architectural approach is flexible in allowing KDD applications to be built upon data analysis services and upon Knowledge Grid services. These services can be composed together with basic services provided by a generic grid toolkit to develop KDD applications.

1. ***Basic grid services*** are core functionality provided by standard grid environments which include security services, data management services, execution management services and information services such as data access, file transfer, replica management, resource allocation, process creation, resource representation, discovery and monitoring ect.

2. ***The Knowledge Grid Services*** specifically designed to support the implementation of data mining services and applications they include resource management services, which provide mechanisms to explain, publish and retrieve information about data sources, data mining algorithms and computing resources, and execution management services, allow users to design and execute distributed KDD applications.

3. ***Data analysis service***s are ad hoc services that express the Knowledge Grid services to provide high-level data analysis functionalities. A data analysis service can expose either a single data pre-processing or data mining task which includes classification, clustering, association.

4. ***KDD applications*** are the knowledge discovery applications built upon the functionalities provided by the underlying grid environments, the Knowledge Grid framework or higher-level data analysis services. Different models, languages and tools can be used to design distributed KDD applications in this framework.

## IV. THE DATAMININGGRID (DMG) ARCHITECTURE

Mining Data is a complex process, which can be deployed by means of multiple approaches. The distributed nature of data and the extension of information sharing makes the Grid a suitable scenario in which data mining applications can be executed [1]. The Grid community has oriented its activity towards a service model [2]. The architecture OGSA (Open Grid Services Architecture) defines an abstract view of this trend in Grid environments. OGSA gives support to the creation, maintenance and lifecycle of services offered by different VOs (Virtual Organizations) [3]. Different domain problems can be solved by means of grid services. Our proposal is a vertical and generic architecture, named DMGA (Data Mining Grid Architecture) [4], which is based on the main data mining stages: pre-processing, data mining and post-processing. Within this framework, the main functionalities of every stage are deployed by means of grid services. Figure A shows the three stages:

(i) PP1 stage (preprocessing stage), (ii) DM stage (data mining stage) and (iii) PP2 stage (post-processing stage). All these data mining services use both basic data and generic grid services. Data Grid services are services oriented to data management in a grid. One of the best known data grid service is GridFTP [5]. Besides data grid services, data mining grid services also use generic and standard grid services. Generic services offer common functionalities in a grid environment [6].

Data Mining Grid services are intended to provide specialized and new data mining services. Whenever a data or generic grid service can be used for a purpose, it will be used. Only if the typical behavior of a service must be modified, a specialized service must be placed at the Data Mining Grid Services level. For instance, if we need to transfer files in the pre-processing stage, we can make use of the GridFTP service. However, we can also invoke a specific Data Access Service, which is an adaptation of the DAI (Data Access and Integration) Data Grid service to data mining applications on grid. This last service is the SDAS (Specific Data Access Service) Data Mining Grid Service.

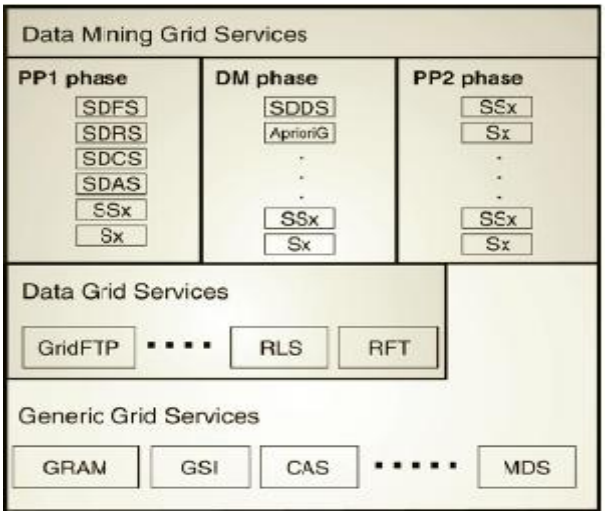



Fig: DataMiningGrid Architecture

New services oriented to data mining applications are included in the architecture. These kinds of services are usually linked to data mining techniques and algorithms. One example is the AprioriG service, which exhibits the Apriori algorithm functionality. Specialization services of the architecture are the following:

- SDFS, Specific Data Filtering Service: due to the huge amount of data involved in the process of data mining, filtering data is an important task, solved by means of this service.
- SDRS, Specific Data Replication Service: one important aspect related to distributed data is data replication. This service deals with this task.
- SDCS, Specific Data Consistency Service: its main purpose is maintaining the data consistency in the grid.
- SDAS, Specific Data Access Service: as has been previously mentioned, this service is an adaptation of the DAI (Data Access and Integration) Data Grid Service to data mining applications on grid.
- SDDS, Specific Data Discovery Service: this service improves the discovery phase in the grid for mining applications.
- SSx: additional specific services can be defined for the management of other features offered by the generic or data grid services, without changes in the rest of the framework.
- Sx: it represents new additional data mining services, which are not provided by basic grid Services. AprioriG is one example of such a kind of service.

## V. THE IMPLEMENTATION OF DMG (DATA MINING GRID)

Most grid classifiers have their foundations in ensemble way [10]. The ensemble approach has been applied in various domains to increase the classification accuracy of predictive models. It produces multiple models (base classifiers) – typically from "homogeneous" data subsets – and combines them to enhance accuracy. Generally, weighted or UN weighted schemes are employed to aggregate base classifiers.

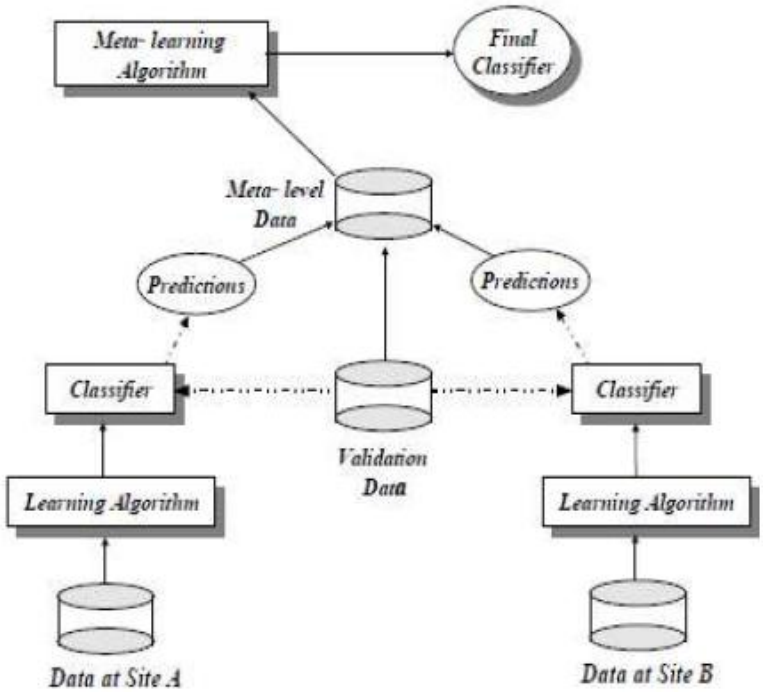



**Figure B Grid Meta Learning from Distributed Data Sites**

The ensemble approach is directly applicable to the distributed scenario. Different models can be generated at different sites and ultimately aggregated using ensemble combining strategies.

In this approach, supervised learning techniques are first used to learn classifiers at local data sites; then meta-level classifiers are learned from a data set generated using the locally learned concepts. The meta-level learning may be applied recursively, producing a hierarchy of meta-classifiers on grids.
Meta-learning follows three main steps

- Concrete base classifiers at each site using a classifier learning algorithms.
- Collect the base classifiers at a central site. Produce meta-level data from a separate validation set and predictions generated by the base classifier on it.
- Generate the final classifier (meta-classifier) from meta-level data grid.

Learning at the meta-level grid can work in many different ways. For example, we may generate a new data grid using the locally learned classifiers. We may also move some of the original training data from the local sites, blend it with the data artificially generated by the local classifiers, and then run any learning algorithm to learn the meta-level grid classifiers.
We may also decide the output of the meta-classifier by counting votes (weighted & non weighted) cast by different base classifiers.
The following discourse notes two common techniques for meta-learning grid from the output of the base classifiers are briefly described in the following.

- **The Grid Arbiter Scheme:** This scheme makes use of a special classifier, called arbiter, for deciding the final class prediction for a given feature vector. The arbiter is learned using a learning algorithm. Classification is performed based on the class predicted by the majority of the base classifiers and the arbiter. It there is a tie, the arbiter's prediction gets the grid preference.
- **The Grid Combiner Scheme:** The grid combiner scheme offers an alternate way to perform meta-learning. The combiner classifier is learned in either of the following ways.

One way is to learn the combiner from the correct classification and the base classifier outputs. Another possibility is to learn the combiner from the data comprised of the feature vector of the training examples, the correct classifications, and the data comprised of the feature vector of the training examples, the correct classifications, and the base classifier outputs.
Either of the above two techniques can be iteratively used resulting in a hierarchy of meta-classifiers.

The above figure (B) shows the overall architecture of the grid Meta learning framework.
The Grid Meta-learning illustrates two characteristics of DDM algorithms – parallelism and reduced communication.
All base classifiers of grid are generated in parallel and collected at the central location along with the validation set, where the communication overhead is negligible compared to the transfer of entire raw data.

## VI. CHALLENGES FACING BY DATAMININGGRID

The shift towards intrinsically distributed complex problem solving environments is prompting a range of new data mining research and development problems. These can be classified into the following broad challenges:

- **Distributed data:** The data to be mined is stored in distributed computing environments on heterogeneous platforms such as grids. Both for technical and for organizational reasons it is impossible to bring all the data to a centralized place. Consequently, development of algorithms, tools, and services is required that facilitate the mining of distributed data.
- **Distributed operations:** In future more and more data mining operations and algorithms will be available on the grid. To facilitate seamless integration of these resources into distributed data mining systems for complex problem solving, novel algorithms, tools, grid services and other IT infrastructure need to be developed.
- **Massive data:** Development of algorithms for mining large, massive and high-dimensional data sets (out-of-memory, parallel, and distributed algorithms) is needed.
- **Complex data types:** Increasingly complex data sources, structures, and types (like natural language text, images, time series, multi-relational and object data types etc.) are emerging. Grid-enabled mining of such data will require the development of new methodologies, algorithms, tools, and grid services.
- **Data privacy, security, and governance:** Automated data mining in distributed environments raises serious issues in terms of data privacy, security, and governance. Grid-based data mining technology will need to address these issues.
- **User-friendliness:** Ultimately a system must hide technological complexity from the User. To facilitate this, new software, tools, and infrastructure development is needed in the areas of grid-supported workflow management, resource identification, allocation, and scheduling, and user interfaces.

## VII. CONCLUSION

The DataMiningGrid needs frequently exchange of datamining models among the participating sites. Therefore, seamless and transparent realizations of DMG technology will require standardize schemes to represent and exchange models. Web search sites like Yahoo and Google are likely to start offering data mining services for analyzing the data they interconnecting data models from such sites will be treated as distributes data mining applications which are mostly implementing through grids. And so now we need to take the techniques that have been developed for things like business intelligence and data mining that goes on around that and think how we can apply those in these real time as well, how we can take every step of the process and have it be very visual and only require as much software understanding as is absolutely necessary.

## REFERENCES

[1] Mario Cannataro, Domenico Talia, Paolo Trunfio, Distributed data mining on the grid, Future Generation Computer Systems 18 (8) (2002) 1101–1112.
[2] Ian Foster, Carl Kesselman, Jeffrey M. Nick, Steve Tuecke, The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration, Tech. Report Globus Project, 2002.
[3] Ian Foster, The anatomy of the Grid: Enabling scalable virtual organizations, Lecture Notes in Computer Science 2150 (2001).
[4] Alberto S´anchez, Jos´e M. Pe˜na S´anchez, Mar´ıa S. P´erez, Victor Robles, Pilar Herrero, Improving distributed data mining techniques by means of a grid infrastructure, in: OTM Workshops, in: LNCS, vol. 3292, 2004, pp. 111–122.
[5] William Allcock, Joe Bester, John Bresnahan, Ann Chervenak, Lee Liming, Steve Tuecke, GridFTP: Protocol extensions to FTP for the Grid, Global Grid Forum Draft, 2001.
[6] Giovanni Aloisio, Massimo Cafaro, Italo Epicoco, Early experiences with the gridftp protocol using the grb-gsiftp library, Future Generation Computer Systems 18 (8) (2002) 1053–1059.
[7] Ian H. Witten, Eibe Frank, Data Mining: Practical machine learning tools with Java implementations, Morgan Kaufmann, San Francisco, 2000.
[8] Agrawal R. & Srikant, R. (1994, September). Fast Algorithms for Mining Association Rules. In Proceedings of the 20th International Conference on Very Large Databases (VLDB'94), Santiago, Chile, 487-499.
[9] Bauer, E., & Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. Machine Learning, 36 (1-2), 105-139.
[10] Dietterich, 2000; Opitz & Maclin, 1999; Bauer & Kohavi, 1999; Merz & Pazzani, 1999