A Framework to Support Distributed Data Mining on Grid

Kun Gao, Lifeng Xi, and Jifang Li

Computer Science and Information Technology College, Zhejiang Wanli University
Ningbo 315100, Zhejiang, R.P. China
gaoyibo@gmail.com

Abstract. In many applications fields, we can obtain benefits from analyzing large distributed data sets by using the high performance computational power. The Grid provides an unrivalled technology for large scale distributed computing as it enables collaboration over the global and the use of distributed computing resources, while also facilitating access to geographically distributed data sets. In this paper, we present a framework for high performance DDM applications in Computational Grid environments called DMGrid, which is based on Grid mechanisms and implemented on top of the Globus 4.0 toolkit.

Keywords: Distributed Computing, Grid Computing, Data Mining, Task Scheduling, Resource Management.

1 Introduction

Many applications are continuing to generate massive data all the time, but because of lack of computation power and collaboration mechanism, we can't obtain fully the potential knowledge in these data that have already spent the huge cost. Without feasible method to process and analyze it, the large cost will not bring the corresponding benefit for our decision. These dataset measured in terabytes or petabytes is very large and come from various applications, including commerce, medical science, scientific experiment, bioscience, etc. They have the inherent property of distribution and heterogeneity. The users who process the data with above characteristics are geographically distributed and large at the same time. The applications that process these data are also expected to have higher performance. At present, existing maturity computation mechanism and technology of data management can't meet requirement described above.

Grid computing [1] is a new platform for distributed process, constituted of heterogeneous computers, and accessed by a general interface. Grid computing, as an important computational model, aims to solve the problem of large scale resource sharing, nontrivial application, innovative applications and high performance application. It is the most advanced scientific research that the grid computing is applied at first. With the maturity of this technology, the grid computing is already applied to coordinated resource sharing and problem solving in dynamic, virtual organizations operating in the industry and business arena.

O. Castillo et al. (Eds.): Theor. Adv. and Appl. of Fuzzy Logic, ASC 42, pp. 489–494, 2007.
springerlink.com

Grid is a natural platform for deploying a high performance application for the knowledge discovery process. On one hand, what the data mining involved is enormous distributed data; On the other hand, its algorithms have higher complexity, need higher computing capability. The grid environment provides coordinated resource sharing, collaborative processing, and high performance computation. So, the combination of these two respects will bring enormous benefit, more rational composition and higher performance.

The outline of the paper is as follows. Section 2 introduces related works and point out the limitations of some previous works. Section 3 describes the rationale of design and development the grid enabled data mining system. Section 4 describes the distributed system framework and main components of DMGrid. Section 5 presents the architecture of tasks scheduling and resource allocation and Section 6 concludes the paper.

2 Related Work

In [2], it is from San Diego Supercomputing Centre which development a middleware to store and access datasets over networks. It is the category of data replication mechanism in fact, for example [3] [4], because it does not handle application implementing in real time. In [5] [6], they are grid computing problem solving environment constructed using MPI and CORBA but is limited to that domain.

In [7], it uses a central server to receive requests and dispatch tasks based on system real time parameters. The main shortcoming of this system is the lack of dynamics. In [8] [9], the systems specialize in parameter-sweep computation, especially supporting dynamic parameters, i.e. parameters whose values are determined at runtime. However, the systems aim at optimizing user parameters and budget for computational tasks only. It has no capability to access remote dataset and optimize the data transfer.

Like [9], [10] provide deployment of parameter-sweep applications on grid. The system emphasizes on data-reuse. The system can appraise the data file that all tasks need, duplicate the data from user node to computation node. When a lot of tasks are assigned to the same resources, it has a try to reuse the data duplicated to make data transmission reduce to minimally. However, the system doesn't support the multiple repositories of data; this method is not applicable to grid.

In this paper, we present a framework for high performance DDM applications in computational grid environments called DMGrid, which is based on grid mechanisms and implemented on top of the Globus 4.0 toolkit. It integrates Grid services by supporting distributed data mining, task scheduling and resource management services that will enlarge the application scenario and the community of Grid computing users.

3 DBGrid Requirements

The rationale of design and development the grid enabled data mining system is as follows:

- DMGrid adopts the standard, common and open grid service mode, follows OGSA norm, and offers unified support to the data mining applications.
- Based on Globus Toolkit and according to the existing networks system structure, DMGrid use the grid service to realize communication, operates each other and resource management.
- DMGrid is open, supports various data mining tools and algorithms, the extensibility is good.
- DMGrid is able to realize the improvement of performance by increasing network node, high performance computing node and cluster, the scalability is strong.
- DMGrid can deal with distributed huge volumes of high dimensional dataset, support heterogeneity data source.
- The main purpose to design and develop the DMGrid system is to improve the performance.
- Users carry out the data mining tasks in a transparent way; the concrete system structure, operation and characteristic in the grid environment is to be hidden.
- In the field of data mining, the security of the data and personal secrets are a sensitive topic. DMGrid supports the choice of place that the data mining execute.

4 DMGrid Architecture

Fig. 1. describes the distributed system framework that we designed and developed. It is mainly made up by following components:

- DMGrid Client Node: In consideration of ease of use, the system adopts Browser/Server mode. Grid client exchanges information with Grid portal through Internet Explorer browser. Users submit the requirement of data mining and receive the final result at Grid client.
- DMGrid Portal Node: It provides a single access way to distributed data mining application based grid. Users can make use of the whole grid resource transparently through the grid portal. This component is responsible for translating users' demand into the RSL language (Resource Specification Language) that can be recognized by grid, is used for grid resource discovery and grid resource allocation management. The final result is returned to grid portal first, and then returned to users by the portal.
- DMGrid Resource Broker Node and DMGrid Tasks Allocation Broker Node: user's data mining requirement has driven grid resource discovery. According to users' demand condition, DMGrid resource broker looks for the resources which meet the condition in a large number of grid resources, including algorithms, computing capability and data resource. It is an important job that finds appropriate resource [11] [12]. As to any application based on grid, it is first to find appropriate resource, then allocate tasks and management them. It can be predicted that there may be many nodes which meet a condition. Resource broker is used for finding available resource among MDS (Meta Directory Service); mapping between data resource and computing resource, i.e., the task allocation broker is responsible for dispatching a certain task on a certain node.

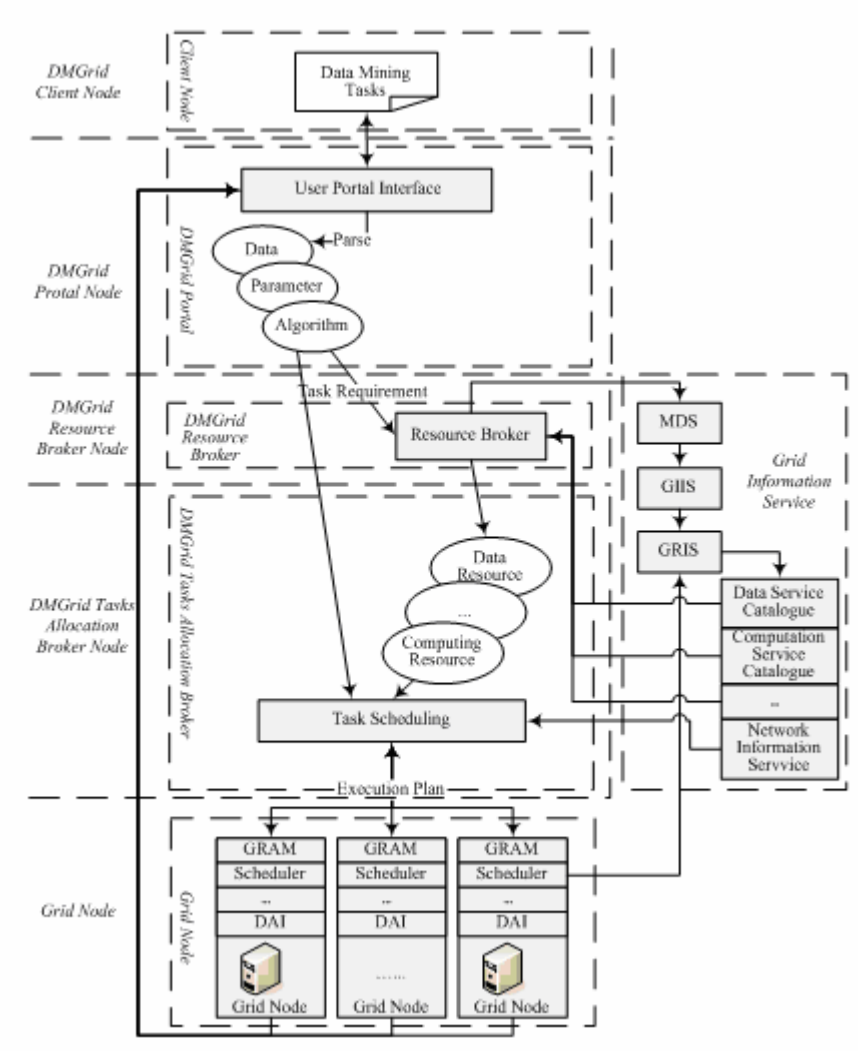

Fig. 1. The framework of distributed data mining on grid

- Grid Node: The Grid nodes are made up of personal computer, high performance computer and cluster. Each node is installed GLOBUS, as grid middleware. They are the data carrier and the computation implementation entity.

5 Tasks Scheduling and Resource Allocation

The broker is responsible for tasks scheduling and resource allocation. It is a core of the whole system. The task allocation procedure adopts the improved taboo search algorithms according to cost matrix, carry out resources mapping on the grid [13] [14].

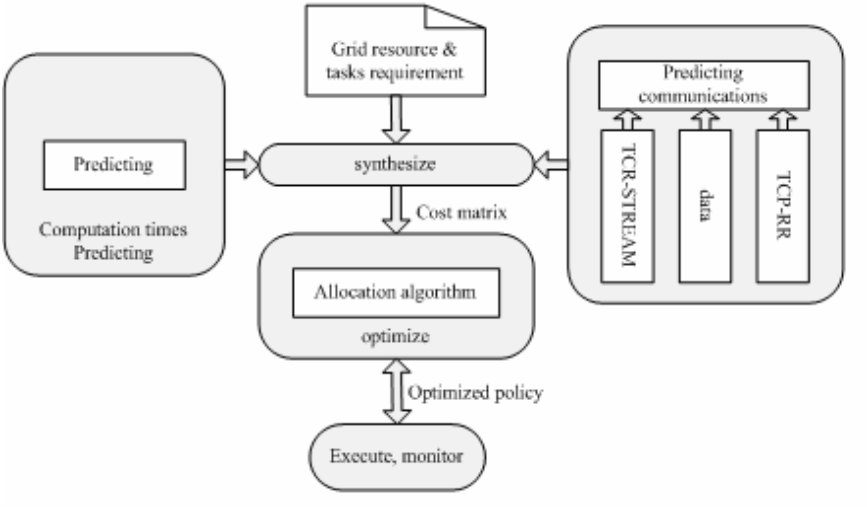

Fig. 2. Architecture of Tasks Allocation

Meanwhile, the establishment of a strategy is according to the real time state queried by the MDS. Broker exchanges information with each GRAM in the grid node to execute a task in coordination. If the grid node is the cluster or the high performance parallel machine, these grid node may have the autonomous local scheduling and allocation. Fig. 2 is the architecture of tasks allocation in the DMGrid.

6 Conclusions

Data mining is a nontrivial process of computation. An efficient data mining application should overcome many factors affecting its performance. With the emergence of globalization grid computing platform, it causes large-scale resource sharing and cooperation becomes possibly. This brings the new vitality for the development of distributed data mining. The grid can provide the high expensive computation resources which the data mining needs. At the same time, the grid environment is consistent with the inherent property of distribution and heterogeneity of data. It is a new trend to merge grid and data mining to meet demands of applications. In this process, optimizing the strategy of tasks allocation is extremely important feasible way to improve the performance of distributed data mining application.

References

1. I. Foster and C. Kesselman, editors. The Grid: Blueprint for a Future Computing Infrastructure. 1999.
2. Asad Samar, Heinz Stockinger. Grid Data Management Pilot (GDMP): A Tool for Wide Area Replication,IASTED International Conference on Applied Informatics (AI2001), Innsbruck, Austria, February 2001.

3. C. Baru, R. Moore, A. Rajasekar, and M. Wan, The SDSC Storage Resource Broker, in CASCON'98 Conference, Toronto, Canada, 1998.
4. A. Chervenak, E. Deelman, I. Foster, L. Guy, W. Hoschek, A. Iamnitchi, C. Kesselman, P. Kunst, M. Ripeanu,B, Schwartzkopf, H, Stockinger, K. Stockinger, B. Tierney. Giggle: A Framework for Constructing Scalable Replica Location Services, Proceedings of Supercomputing 2002 (SC2002), November 2002.
5. G. Allen, W. Benger, T. Goodale, H. Hege, G. Lanfermann, A. Merzky, T. Radke, E. Seidel, J. Shalf, The Cactus Code: A Problem Solving Environment for the Grid, Proceedings of the Ninth International Symposium on High Performance Distributed Computing (HPDC), Pittsburgh, USA, IEEE Press.
6. K.Marzullo, M. Ogg, A. Ricciardi, A. Amoroso, F. Calkins, E. Rothfus, NILE: Wide-Area Computing for High Energy Physics, Proceedings of 7th ACM SIGOPS European Workshop, Connemara, Ireland, 2-4 Sept. 1996, ACM Press.
7. W. Hoschek, J. Jaen-Martinez, A. Samar, H. Stockinger, K. Stockinger, Data Management in an International Data Grid Project, Proceedings of the 1st International Workshop on Grid Computing (Grid 2000, Bangalore, India), Springer-Verlag, Berlin, Germany, 2000.
8. D. Abramson, J. Giddy, and L. Kotler, High Performance Parametric Modeling with Nimrod/G: Killer Application for the Global Grid?, Proceedings of the International Parallel and Distributed Processing Symposium (IPDPS 2000), May 1-5, 2000, Cancun, Mexico, IEEE CS Press, USA, 2000.
9. R. Buyya, D. Abramson, and J. Giddy, An Economy Driven Resource Management Architecture for Global Computational Power Grids, Proceedings of the 2000 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA 2000), June 26-29, 2000, Las Vegas, USA, CSREA Press, USA, 2000.
10. H. Casanova, G. Obertelli, F. Berman, and R. Wolski, The AppLeS Parameter Sweep Template: User-Level Middleware for the Grid, Proceedings of the IEEE SC 2000: International Conference Networking and Computing, Nov. 2000, Dallas, Texas, IEEE CS Press, USA.
11. Kun Gao, Semantics Based Grid Services Publishing and Discovery, Proceedings of The 1st International Symposium on GRID COMPUTING, Corfu, Greece, August 18, 2005, ISBN 960-8457-32-7, pages 89-93.
12. Kun Gao, Towards Semantic-Driven Grid Resource Discovery, WSEAS TRANSACTIONS on SYSTEMS, Issue 10, Volume 4, October 2005, ISSN 1109-2777, Pages 1668-1675.
13. Kun Gao, Youquan Ji, Meiqun Liu, Jiaxun Chen, Rough Set Based Computation Times Estimation on Knowledge Grid, Lecture Notes in Computer Science, Volume 3470, Jun 2005, Pages 557 - 566
14. Kun Gao, Kexiong Chen, Meiqun Liu, Jiaxun Chen, Rough Set Based Data Mining Tasks Scheduling on Knowledge Grid, Lecture Notes in Computer Science, Volume 3528, May 2005, Pages 150 – 155