GridMiner: A Framework for Knowledge Discovery on the Grid - from a Vision to Design and Implementation

P. Brezany[1], I. Janciak[1], A. Wöhrer[1], and A M. Tjoa[2]

[1] Institute for Scientific Computing, University of Vienna
Nordbergstrasse 15/C/3, A-1090 Vienna, Austria
email:{brezany,janciak,woehrer}@par.univie.ac.at
phone: (+43 1) 4277 38825, *fax:* (+43 1) 4277 9388

[2] Institute of Software Technology and Interactive Systems, Vienna University of Technology, Favoritenstrasse 9-11/E188, A-1040 Vienna, Austria
email: tjoa@ifs.tuwien.ac.at
phone: (+43 1) 58801 18800, *fax:* (+43 1) 58801 18899

Abstract

Knowledge discovery in data sources available on Computational Grids is a challenging research and development issue. Several Grid research activities addressing some facets of this process have already been reported. The GridMiner project (www.gridminer.org) at the University of Vienna aims, as the first Grid research effort, to cover all aspects of the knowledge discovery process and integrate them as advanced service-oriented Grid application. The innovative architecture provides a robust and reliable high performance data mining and OLAP environment and strengths the importance of Grid enabled applications in terms of business intelligence and detailed analysis of very large scientific data sets. The interactive cooperation of different services - data integration, data selection, data transformation, data mining, pattern evaluation and knowledge presentation - within the GridMiner architecture is the key to high performance knowledge discovery on large datasets.

1 Introduction

It is not a simple matter to develop an integrative approach that exploits synergies between knowledge management and knowledge discovery in order to monitor and manage the full lifecycle of knowledge and provides services quickly, reliably and securely. Several Grid research activities addressing some facets of this process have already been reported, e.g. [4]. The GridMiner project[1] at the University of Vienna aims, as the first Grid research effort, to cover all aspects of the knowledge discovery process and integrate them as advanced service-oriented Grid application. The technology developed is being validated and tested on an advanced medical application addressing treatment of traumatic brain injury (TBI) victims [7]. Medicine is just one application area where an environment is needed for continuous knowledge discovery and management.

[1] http://www.gridminer.org

Fig. 1 pictures the knowledge life cycle – from discovery to processing, sharing and finally reusing of knowledge as input for a new discovery phase [8] – which represents the overall target of our research efforts. The GridMiner prototype is already covering and supporting a great portion of this cycle. In general, a knowledge discovery process consists of an iterative sequence of several steps: data cleaning/integration/selection/transformation can be summarized as data preprocessing, data mining, pattern evaluation and knowledge presentation and visualization. Afterwards, the patterns are getting processed and applied to appropriate data material. But to gain the most out of the discovered knowledge, that should not be the end of the usage-story. Other professionals will be interested in the already gained understanding, so it has to be shared/stored in a suitable way for later re-usage.

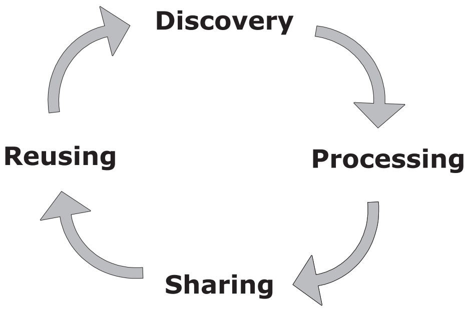

Fig. 1: Knowledge life cycle

The aim of the GridMiner application is to give to an expert (Dataminer) a tool which can ease the knowledge discovery process in the distributed Grid environment. So it is essential that the system provides a powerful, flexible and simple to use graphical user interface (GUI) which hides the complexity of the Grid but still offering possibilites to interfere during the execution phase, control the task execution and visualize results.

The remain part of the paper is organized as follows. Section 2 gives an overview of the 3-layer architecture of GridMiner and describes in more detail the Grid layer, the Web layer and the User Environment and reviews some of our current implementation approaches. We briefly discuss related work in Section 3 and finish the paper with our conclusions and a future work outline in Section 4.

2 Architecture

Fig. 2 shows a high-level abstraction view of the components in our architecture and how they are connected, as it has been implemented in the current GridMiner prototype. The GridMiner is a service oriented application and has been implemented as a research prototype completely built on top of the Globus Toolkit Version 3 and standard web technologies.

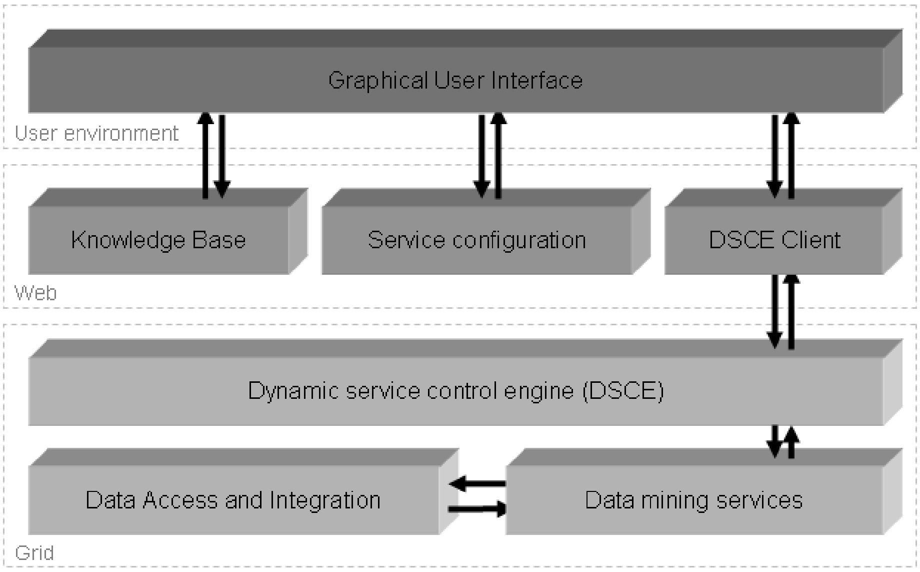

Fig. 2: 3-layered architecture of GridMiner

2.1 Grid Layer

Knowledge discovery is a highly interactive process and to achieve appealing results the user must permanently have the possibility to influence this process by applying different algorithms or adjusting their parameters. Therefore Gridminer supports highly **dynamic workflow concept**, where an user can compose a workflow according to its individual needs. A special research effort of our project deals with the integration of all needed services into a workflow, which is executed by the Dynamic Service Composition Engine [5]. In our approach, we designed a new specification for dynamic service composition called the Dynamic Service Composition Language (DSCL), which is based on XML notation. DSCL allows the description of a workflow consisting of various Grid services and the specification of their parameter values. DSCE is implemented as a Grid service and can be controlled interactively by a client, which has the possibilities to execute, stop, resume or even to change the workflow and its parameters.

The **data integration** in the Gridminer is based on the wrapper mediator approach supported by the Grid Data Mediation Service (GDMS) [10], which allows integrating heterogeneous relational databases, XML databases and comma separated value files into one logically single homogeneous virtual data source. The newly developed concepts for the mediation service have been implemented by reusing the standard reference implementation of Grid Data Services (GDS), namely OGSA-DAI [2], proposed by the DAIS Working Group.

Currently, the **data mining** process within the GridMiner is supported by several services able to perform data mining tasks and OLAP. The suite of data mining services (DMS) includes sequential, parallel and distributed implementations of data mining algorithms which are able to deal with data provided by Data Access and Integration Service. Each data mining service is implemented as a standalone grid service specified by Open Grid Service Architecture (OGSA), able to deal with huge amount of data and present its result in the standard format. The input data for DMS are in the XML WebRowSet format and are delivered to the service in a file or as a data stream.

The DMS present they results in Predictive Model Markup Language (PMML), a standard developed by the Data Mining Group[2], for representating data mining models. This allows to make the results compatible with third party visualization applications and also to use them as an input for another data mining task.

Following list presents DMS currently implemented in the GridMiner infrastructure and appropriate algorithm:

- Sequential Clustering Service (SimpleKMeans)
- Sequential Sequences Service (SPADE)
- Distributed Decision Rules Service (SPRINT)
- Parallel OLAP Service,
- Sequential Association Rule Mining Service on OLAP Cubes

In our research we focus on **On-Line Analytical Processing** (OLAP) - where so far, no data warehouse and scalable OLAP investigations on the Grid have been reported [3] . The usage of new data indexing, data materialization and querying techniques will allow us a distributed/parallel OLAP implementation.

2.2 Web layer

The **Knowledge Base** (KB) allows to store and share all the information needed by the other components in the process of the knowledge discovery and is incrementally extended by newly discovered knowledge for its future reuse. KB consists of (1) ontologies - describing data mining domain, data sources and activities used in the process of knowledge discovery, (2) metadata - holds information about data in data sources, (3) rules - discovered results of data mining tasks and (4) facts - explicit knowledge generated as a result of applying rules on the domain ontology. KB is also used as a central registry of the services and their locations and also stores information about users and their projects. All the information in the KB are stored XML and ontologies are described by OWL[3]

For **service configuration** we are utilizing a set of web applications able to interact with the user in the process of preparing data mining tasks. They allow to configure services (e.g. select algorithm), setup input parameters (select attributes etc.) and prepare workflow parameters (DSCL document) for the

[2] http://www.dmg.org

[3] http://www.w3c.org

DSCE Client. Service configurators are kind of wizards able to setup and confirm the task for the service by the user.

The main goal of **Dynamic Service Control Engine Client** is to bridge the Web and the Grid enviroments. It allows to start Dynamic Service Control Service and control it execution and deliver notification messages from services to the client.

2.3 User Environment

The GUI allows to interactively construct workflow descriptions at a high abstraction level and visualize the results from data mining tasks. As depicted in Fig. 2, it lies in the client environment what can be any operating system supporting Java. The GUI is currently deployed as a Java standalone application able to be started by Java Web Start [4]. It allows an easy integration of existing and newly developed data preprocessing and data mining services into the Grid.

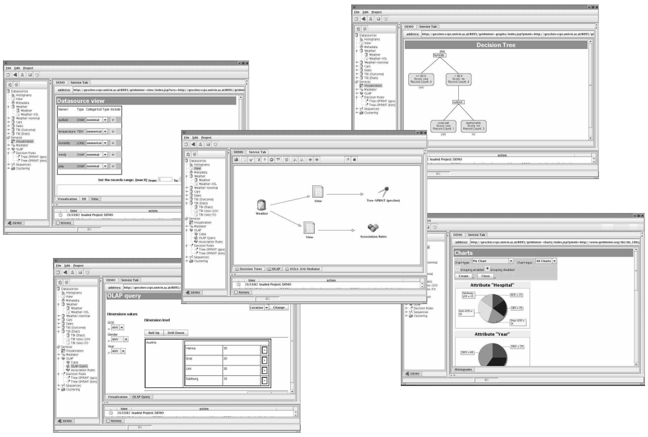

Fig. 3: Graphical User Interface.

3 Related Work

So far, only a little attention was devoted to knowledge discovery on the Grid. There are already many publications on parallel and distributed data mining [11]. An attempt to design an architecture for performing data mining on the Grid was presented in [1]. The authors present design of a Knowledge Grid architecture based on the non-OGSA-based version of the Globus Toolkit, and don't consider any concrete application domain. R. Moore presents the concepts of knowledge-based Grids in [9]. Mahinthakumar [6] report about the

[4] http://java.sun.com/products/javawebstart

first clustering algorithm implementation on the Grid. The WP47 of the OGSA-DAI project is working on the design of a distributed query processing service for the Grid[5].

4 Conclusions and Future Work

In this paper we have described our research effort, which focuses on the application and extension of the Grid technology to knowledge discovery in Grid databases. We described the service oriented architecture and its components implemented in the GridMiner application. Several data miningand OLAP services have been already deployed and are ready to perform the knowledge discovery tasks. The future work is to focus on the performance results and usability of the GridMiner application.

Acknowledgements. This research is being carried out as part of the research projects “Modern Data Analysis on Computational Grids” and “Aurora”.

References

1. M. Cannataro and D. Talia. Parallel and distributed knowledge discovery on the grid: A reference architecture. In *Fourth International Conference on Algorithms and Architectures for Parallel Processing ICA^{3}PP, Hong Kong, Dec. 11-13, 2000, World Scientific 2000*, pages 662–673, December 2000.
2. M. Antonioletti et al. OGSA-DAI: Two Years On. In *The Future of Grid Data Environments Workshop at GGF10*, March 2004.
3. B. Fiser, U. Onan, I. Elsayed, P. Brezany, and A Min Tjoa. On-line analytical processing on large databases managed by computational grids. Invited paper for the DEXA 2004, Zaragoza, Spain.
4. N. Giannadakis, A. Rowe, M. Ghanem, and Y. Guo. Infogrid: providing information integration for knowledge discovery, 2003.
5. G. Kickinger, J. Hofer, A Min Tjoa, and P. Brezany. Workflow management in GridMiner. In *3rd Cracow Grid Workshop*, 2003.
6. G. K. Mahinthakumar, F. M. Hoffman, W. W. Hargrove, and N. T. Karonis. Multivariate geographic clustering in a metacomputing environment using Globus. In *Supercomputing'99*, Orlando, USA, November 1999.
7. W. Mauritz, M. Rusnak, and I. Janciak. Implementing scientific evidence-based guidelines: Case study of severe traumatic brain injuries. *Clinical Research and Regulatory Affairs*, 20(1):81–88, January 2003.
8. Mark W. McElroy. The new knowledge management. *Journal of the KMCI*, 1(1):43–67, 2000.
9. R. Moore. Knowledge-Based Grids. Technical Report TR-2001-02, San Diego Supercomputer Center, January 2001.
10. A. Woehrer and P. Brezany. Mediators in the Architecture of Grid Information Systems. Technical report, Institute for Software Science, February 2004.
11. Mohammed J. Zaki. Parallel and distributed association mining: A survey. *IEEE Concurrency*, 7(4):14–25, /1999.

[5] http://www.ogsadai.org.uk/dqp