3

GridMiner: An advanced support for e-science analytics

Peter Brezany, Ivan Janciak and **A. Min Tjoa**

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. This chapter introduces the GridMiner framework, developed within a research project at the University of Vienna. The project's goal is to deal with all tasks of the knowledge discovery process on the grid and integrate them in an advanced service-oriented grid application. The GridMiner framework consists of two main components: technologies and tools, and use cases that show how the technologies and tools work together and how they can be used in realistic situations. The innovative architecture of the GridMiner system is based on the Cross-Industry Standard Process for Data Mining. GridMiner provides a robust and reliable high-performance data mining and OLAP environment, and the system highlights the importance of grid-enabled applications in terms of e-science 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 productive e-science analytics.

3.1 Introduction

The term e-science refers to the future large-scale science that will increasingly be carried out through distributed global collaborations enabled by the Internet. This phenomenon is a major force in the current e-science research and development programmes (as pioneered in the UK) and NSF cyberstructure initiatives in the US. Typically, the individual user scientists require their collaborative scientific enterprises to have features such as access to very large data collections and very large scale computing resources. A key component of this development is *e-science analytics*, which is a dynamic research field that includes rigorous and sophisticated scientific methods of data pre-processing, integration, analysis, data mining and visualization

Data Mining Techniques in Grid Computing Environments Edited by Werner Dubitzky

associated with information extraction and knowledge discovery from scientific data sets. Unlike traditional business analytics, e-science analytics has to deal with huge, complex, heterogeneous, and very often geographically distributed data sets which contain volumes measured in terabytes and will soon total petabytes. Because of the huge volume and high dimensionality of data, the associated analytic tasks are often both input/output and compute intensive and consequently dependent on the availability of high-performance storage, computing hardware, software resources and software solutions. This is one of the main reasons why high-performance computing analytics has become a regular topic of the programmes at the Supercomputing Conference Series since 2005. Further, the user community is often at different geographically distributed locations, and finally, a high level of security has to be guaranteed for many analytical applications, e.g. in finance and medical sectors. The grid is an infrastructure proposed to bring all these issues together and make a reality of the vision for e-science analytics outlined above. It enables flexible, secure, coordinated resource (computers, storage systems, equipments, database, and so forth) sharing among dynamic collections of individuals and institutions. Over the past four years, several pioneering grid-based analytic system prototypes have been developed. In most of these systems, analytic tasks are implemented as grid services, which are combined into interactive workflows orchestrated and executed by special services called workflow engines. In the following, we characterize some relevant developments.

The myGrid project is developing high-level middleware for data and legacy application resource integration to support *in silico* experiments in biology. Their developed workflow system Taverna (Wolstencroft, *et al.*, 2005) provides semantic support for the composition and enactment of bioinformatics workflows. Cannataro and Talia (2003) proposed a design of the Knowledge Grid architecture based on the Globus Toolkit (Foster and Kesselman, 1998). Discovery Net (Curcin, *et al.*, 2002) provides a service-oriented computing model for knowledge discovery allowing users to connect to and use data analysis software and data sources that are available on-line. The *Science Environment for Ecological Knowledge* (*SEEK*) project (Jones, *et al.*, 2006) aims to create a distributed data integration and analysis network for environmental, ecological and taxonomy data. The SEEK project uses the Kepler (Berkley, *et al.*, 2005) workflow system for service orchestration. The recently finished EU-IST project DataMiningGrid (Stankovski, *et al.*, 2008) developed advanced tools and services for data mining applications on a grid. It uses Web Services Resource Framework-compliant technology with Triana (Churches, *et al.*, 2006), an open source problem solving environment developed at Cardiff University, that combines an intuitive visual interface with powerful data analysis tools in its user interface.

In all the above-mentioned projects, the main focus has been put on the functionality of the developed prototypes and not on the optimization of their runtime performance. Moreover, the end user productivity aspects associated with their use have not been sufficiently investigated. Their improvement can have significant impact on many real-life spheres, e.g. it can be a crucial factor in achievement of scientific discoveries, optimal treatment of patients, productive decision making, cutting costs and so forth. These issues, which pose serious research challenges for continued advances in e-science analytics, motivated our work on a grid-technology-based analytics framework called GridMiner[1], which is introduced in this chapter. Our aim was to develop a core infrastructure supporting all the facets of e-science analytics for a wide spectrum of applications. The framework consists of two main components.

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

- *Technologies and tools.* They are intended to effectively assist application developers to develop grid-enabled high-performance analytics applications. GridMiner includes services for sequential, parallel and distributed data, text mining, on-line analytical processing, data integration, data quality monitoring and improvement based on data statistics and visualization of results. These services are integrated into interactive workflows, which can be steered from desk-top or mobile devices. A tool called workflow composition assistant allows semi-automatic workflow construction based on the SemanticWeb technology to increase the productivity of analytic tasks.
- *Use cases.* They show how the above technologies and tools work together, and how they can be used in realistic situations.

The presented research and development was conducted in cooperation with some of the worldwide leading grid research and application groups. The set of pilot applications directly profiting from the project results includes the medical domain (cancer research, traumatic brain injuries, neurological diseases and brain informatics) and the ecological domain (environment monitoring and event prediction).

3.2 Rationale behind the design and development of GridMiner

Since the 1990s, grid technology has traversed different phases or generations. In 2002 when the research on GridMiner started, all major grid research projects were built on the Globus Toolkit (Foster and Kesselman, 1998) and UNICORE (Romberg, 2002). Therefore, our initial design of the GridMiner architecture was based on the state-of-the-art research results achieved by Globus and its cooperating research projects, which provided grid software development kits exposing 'library-style APIs'. Very soon after the start of the project, the grid research programme and industry activities focused on architecture and middleware development aligned with the Web services standards. Further, we realized that our effort had to consider the achievements in the Semantic Web, workflow management and grid database technologies.

As a first step, we designed and prototyped a runtime environment and framework called GridMiner-Core based on the *Open Grid Service Architecture* (*OGSA*) concepts built on top of the Globus Toolkit Version 3 (Foster, *et al.*, 2002) and grid database access services provided by the *OGSA Data Access and Integration* (*OGSA-DAI*) (Antonioletti, *et al.*, 2005) middleware. This solution allowed the integration and execution of data pre-processing, data mining tools, applications and algorithms in a grid-transparent manner, i.e. the algorithm contributors could focus on knowledge discovery problems without the necessity to handle grid specifics. Thus, the framework abstracted from grid details such as platform, security, failure, messaging and program execution. The design decisions, scalability behaviour with respect to the data set size and number of users working concurrently with the infrastructure and performance of the prototype were evaluated. Based on the results and experience from the GridMiner-Core research tasks, a full service-oriented GridMiner architecture has been investigated (Hofer and Brezany, 2004).

Since the birth of the OGSA technology, based on the Web service concepts, the grid community has been focusing on the service-oriented architectures. This trend was confirmed by the latest OASIS specification, named *Web Service Resource Framework* (*WSRF*) (Czajkowski, *et al.*, 2004). Therefore, our goal was also to develop an infrastructure supporting all the phases of the knowledge discovery process in the service-oriented grid environments with respect to the scientific workflows and the latest available technologies. Furthermore, we have devoted significant effort to the investigation of appropriate modelling mechanisms. We adopted and extended the existing *CRoss-Industry Standard Process for Data Mining* (*CRISP-DM*) Reference Model (Chapman, *et al.*, 2000) and its phases as essential steps for the service-oriented scientific workflows. The model is discussed in Section 3.4.

3.3 Use case

Before describing the GridMiner framework, we present a practical use case taken from a medical application addressing management of patients with traumatic brain injuries (Brezany, *et al.*, 2003a). A traumatic brain injury typically results from an accident in which the head strikes an object. Among all traumatic causes of death, it is the most important one, besides injuries of the heart and great vessels. Moreover, survivors of a traumatic brain injury may be significantly affected by a loss of cognitive, psychological or physical function. Below, we discuss how data mining and its implementation in GridMiner can support treatment of traumatic brain injury patients.

At the first clinical examination of a traumatic brain injury patient, it is very common to assign the patient into a category, which allows one to plan the treatment of the patient and also helps to predict the final outcome of the treatment. There are five categories of the final outcome defined by the *Glasgow Outcome Scale* (*GOS*): dead, vegetative, severely disabled, moderately disabled and good recovery. It is obvious that the outcome is influenced by several factors, which are usually known and are often monitored and stored in a hospital data warehouse; frequently used factors include Injury Severity Score, Abbreviated Injury Scale and Glasgow Coma Score. It is evident that if we want to categorize a patient then there must be a prior knowledge based on cases and outcomes of other patients with the same type of injury. This knowledge can be mined from the historical data and represented as a classification model. The model can be then used to assign the patient to one of the outcome categories. In particular, using the model, one of the values from the GOS can be assigned to a concrete patient.

One of the basic assumptions in classification is that by considering a larger number of cases, the accuracy of the final model can be improved. Therefore, access to the data for similar traumatic brain injury cases stored in other hospitals would help to create a more accurate classification model. In the grid environment, a group of hospitals can share their data resources such as anonymized patients records or some statistical data related to the management of the hospitals. The group of hospitals can be then seen as a *virtual organization* (*VO*) (Foster, *et al.*, 2002). There can be also other partners in the VO offering their shareable resources, for example, analytical services or high-performance computing resources, as illustrated in Figure 3.1. In such a scenario, it is necessary to deal with other challenges such as secure access to the distributed data, cleaning and integration of the data and its transformation into a format suitable for data mining. There are also many other tasks related to this problem such as legal aspects of data privacy that have to be solved, especially for such sensitive data patients' records.

There exist several possibilities for how to deal with the above-mentioned challenges. Our approach is based on the principle that the movement of the data should be reduced as much

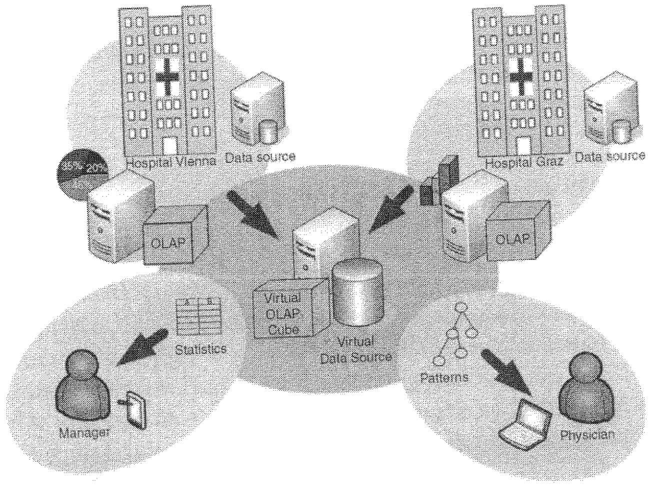


Figure 3.1 Use case scenario

as possible to avoid the leak of valuable information. This means that some tasks, such as data pre-processing, should be performed as close as possible to the data source. If the data movement is unavoidable then only patterns or aggregated data should be moved out of the hospital's data centre.

Another situation involves data integration prior to mining that requires all data records. To address this requirement problem we adopted an approach based on data virtualization and data mediation. This means that in order to reach a particular record a distributed query mechanism supported by the mediator service (Brezany, *et al.*, 2003b) is applied. The data integration problem has also several other aspects, which should not be ignored. The first one is semantic inconsistency of the integrating data sources, and the second one is the data format inconsistency.

Before a data mining technique can be applied, the data need to be preprocessed to correct or eliminate erroneous records. Problematic records may contain missing values or noise such as typos or outlier values that can negatively impact the quality of the final model.

From a health care control and governance point of view, hospital management data about the health care services and their quality are more interesting. The data of interest in this case include statistics, which can be represented as aggregated data. This can be supported by *online analytical processing* (*OLAP*) techniques.

3.4 Knowledge discovery process and its support by the GridMiner

Knowledge discovery in databases (*KDD*) can be defined as a non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data (Fayyad, Piatetsky-Shapiro and Smyth, 1996). KDD is a highly interactive process and to achieve useful results the user must permanently have the possibility to influence this process by applying different algorithms or adjusting their parameters. Initiatives such as CRISP-DM try to define the KDD project, its corresponding phases, their respective tasks and relationships between these

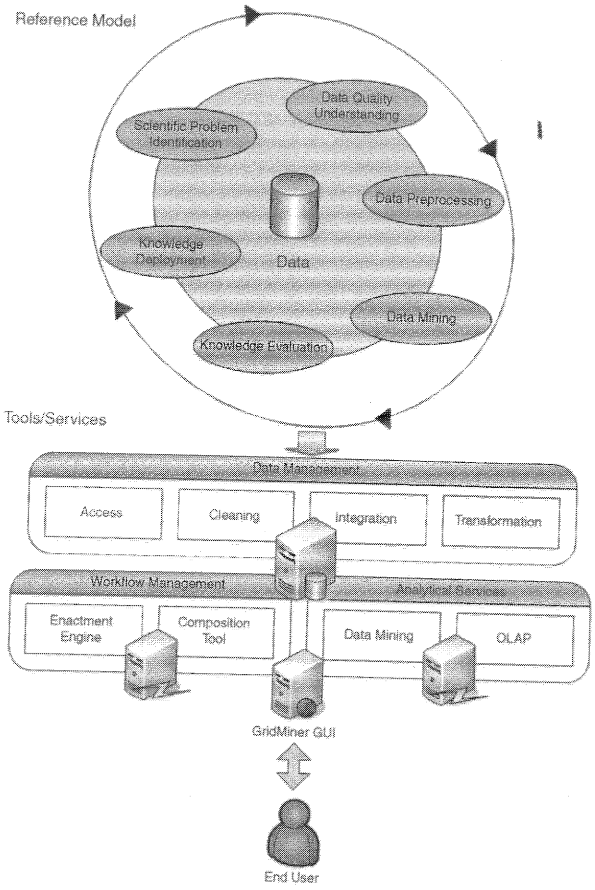


Figure 3.2 Data mining process covered by GridMiner's tools and services

tasks. Within the GridMiner project, we have adopted and extended the CRISP-DM model and its phases as essential steps to the service-oriented scientific workflows. The whole concept is depicted in Figure 3.2. The top part corresponds to our new reference model, and the bottom part shows an illustrative GridMiner framework and its supporting tools and services.

3.4.1 *Phases of knowledge discovery*

In the following paragraphs, we discuss the phases at the highest level of our new reference model for data mining projects as implemented in the GridMiner framework.

Scientific problem identification This initial phase focuses on clearly establishing goals and requirements of the scientific problem that is going to be solved, the role of data mining in the solution approach and selecting suitable data sets in the given data space and mining techniques. Within this phase, the main objectives of the data mining process are identified and their basic properties are specified in a preliminary workplan. The ways of solving the problem can include different methods and approaches, for example a scientific experiment, statistical confirmation of a hypothesis and so forth.

Data quality understanding The data and its quality understanding phase starts with an initial data collection and proceeds with activities allowing the user to become familiar with the data, to identify data quality problems, to obtain first insights into the data or to detect interesting subsets to form hypotheses.

Data pre-processing According to Pyle (1999), data pre-processing is usually the most challenging and time consuming step in the whole knowledge discovery process. The aim of this phase is to improve data quality, which has a significant impact on the quality of the final model and, therefore, on the success of the whole process. Data in operational databases are typically not clean. This means that they can contain erroneous records due to wrong inputs from users or application failures. Besides, these data may be incomplete, and essential information may not be available for some attributes. Hence, for data mining and other analysis tasks, it is necessary to clean data and to integrate the data from multiple sources such as databases, XML files and flat files into a coherent single data set. Because of the large size of typical grid data sources, building traditional data warehouses is impractical. Therefore, data integration is performed dynamically at the point of data access or processing requests.

Moreover, the size of the input data for modelling can enhance the accuracy of the predictive model as well as having significant impact on the time of model building. The data selection step allows the choice of an appropriate subset of the whole data set, which can be used as a training set to build a model as well as a test set for the model evaluation.

Data mining This phase deals with selecting, applying and tuning a modelling technique on the preprocessed data set. It involves the application and parameterization of a concrete data mining algorithm to search for structures and patterns within the data set. Typically, data mining has the two high-level goals of *prediction* and *description*, which can be achieved using a variety of data mining methods such as association rules, sequential patterns, classification, regression, clustering, change and deviation detection and so forth. Besides these most fundamental methods, the data mining process could, for example in bioinformatics (Wang, *et al.*, 2005), refer to finding motifs in sequences to predict folding patterns, to discover genetic mechanisms underlying a disease, to summarize clustering rules for multiple DNA or protein sequences and so on. An overview of other modern and future data mining application areas is given by Kargupta, *et al.* (2003).

The large size of the available data sets and their high dimensionality in many emerging applications can make knowledge discovery computationally very demanding to an extent that parallel computing can become an essential component of the solution. Therefore, we investigated new approaches enabling development of highly optimized services for management of data quality, data integration, data mining text mining and OLAP. These services are able to run on parallel and distributed computing resources and process large, voluminous data within a short period of time, which is crucial for decision making. There were several concurrency levels investigated as depicted in Figure 3.3.

(a) *Workflows.* Speculative parallelism can be used for parameter study of analytical tasks (for example, concurrent comparison of accuracy of neural network, decision tree and regression classification methods).

(b) *Inter-service parallelism.* Services can perform coordinated distributed processing.

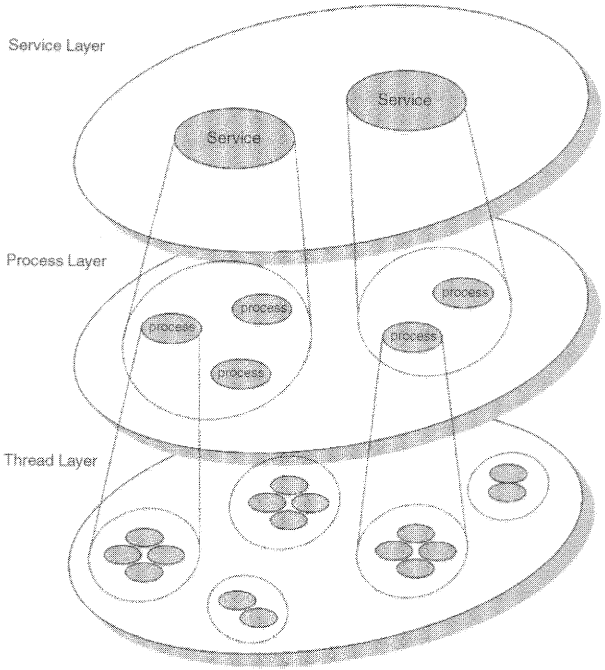


Figure 3.3 Inter and intra-service parallelism

(c) *Intra-service parallelism.* Service implementation is based on process and thread mechanisms, which are the concepts of parallel programming technology.

Within the data exploration processes we also investigated the impact of the data size on the models' accuracy to answer the question 'When is it necessary to mine complete data sets, as opposed to mining a sample of the data?' This opens a new dimension for data exploration methodologies. We also included OLAP in our research focus because we consider OLAP and data mining as two complementary technologies, which, if applied in conjunction, can provide efficient and powerful data analysis solutions on the grid.

Evaluation Data mining algorithms discover patterns in data, not all of which may be new or be of interest to the user. The evaluation phase is designed to address this issue and discriminate relevant from irrelevant patterns. The main goal of the presentation task is to give a user all discovered information in an appropriate form. In GridMiner, we addressed different visualization techniques for representing data mining models in, for example, tables, graphs and charts. These models are represented by the PMML so they can be imported to the other applications and further evaluated.

Knowledge deployment It is obvious that the purpose of a data mining model, as a final product of the data mining process, is to increase knowledge about the analysed data. A model itself, without an appropriate description, can be of limited utility for the end user who is carrying out the deployment phase. Therefore, a kind of report describing used data and all the steps leading to the model is required. The *Predictive Markup Model Language* (*PMML*)

(Data Mining Group, 2004) facilitates not only the characterization of data mining models but also the description of data, metadata, statistics and data transformations. Hence, together with visualization, the PMML is a suitable format for knowledge representation. However, the representation of the models is usually not enough. Its interpretation by a domain expert is extremely helpful and can point to the weakness of the model to eliminate its misuse.

In the following sections we will introduce the services and tools supporting the phases of the knowledge discovery process. We will not go too deep into implementation details, which can be found in the cited papers, but rather introduce the design concepts we have used to develop the presented tools.

3.4.2 Workflow management

With respect to the discussion about the phases of the KDD, it is obvious that the GridMiner architecture highly depends on appropriate workflow management mechanisms. We came to the conclusion that GridMiner needs a dynamic workflow concept where a user can compose the workflow according to his individual needs. As a first step, an XML-based Dynamic Service Composition Language (Kickinger, *et al.*, 2004) was proposed; it serves as input to the *Dynamic Service Control Engine* (*DSCE*) (Kickinger, *et al.*, 2003), which was also developed within the project. The DSCE was implemented as an application-independent, stateful and transient OGSA (Foster, *et al.*, 2002) service consisting of a set of other OGSA services. Additionally, the engine allows the starting, stopping, resuming, modifying and cancelling of workflows and notifies a client about the status of the execution.

Workflow Enactment Engine The new scientific workflow concepts (Taylor, *et al.*, 2007), the acceptance of the WS-BPEL language (Sarang, Mathew and Juric, 2006) in the scientific community and standardization of WSRF have become the leading motivation for the development of a new workflow engine. A newly established subproject of the GridMiner project called *Workflow Enactment Engine Project* (*WEEP*) (Janciak, Kloner and Brezany, 2007) is aiming to implement a workflow engine for WSRF services using WS-BPEL 2.0 as a formalism to specify, process and execute workflows. The engine follows up ideas of its ancestor DSCE and is being developed as a central component for data mining workflow orchestration in GridMiner. The core of the engine provides a run-time environment in which process instantiation and activation occurs, utilizing a set of internal workflow management components responsible for interpreting and activating the process definition and interacting with the external resources necessary to process the defined activities. The main functionality of the engine can be summarized as follows.

- *Validation and interpretation.* The validation and interpretation of a BPEL process definition and its representation as a new resource with a unique identifier. The resource can be seen as a workflow template representing a stateless process.
- *Instantiation.* The instantiation of the workflow template to the stateful process represented by a service.
- *Execution.* The execution of the process instance together with controlling and monitoring of the whole workflow and its particular activities.

With the above-mentioned functionality, the engine is able to reuse deployed workflows, create new instances with different input parameters and represent the process as a single service.

GridMiner Assistant To support a semi-automatic composition of the data mining workflows we have developed a specialized tool called GridMiner Assistant (Brezany, Janciak and Tjoa, 2007), which assists the user in the workflow composition process. The GridMiner Assistant is implemented as a Web application able to navigate a user in the phases of the knowledge discovery process and construct a workflow consisting of a set of cooperating services aiming to realize concrete data mining objectives. The GridMiner Assistant provides support in choosing particular objectives of the knowledge discovery process and manages the entire process by which properties of data mining tasks are specified and results are presented. It can accurately select appropriate tasks and provide a detailed combination of services that can work together to create a complex workflow based on the selected outcome and its preferences. The GridMiner Assistant dynamically modifies the task composition depending on the entered values, defined process preconditions and effects, and existing description of services.

3.4.3 Data management

GridMiner needs to interoperate with data access and management technologies that are being developed by other communities. On top of these technologies, we developed services for data access, data integration, data preprocessing, computation of statistics and a novel concept of a grid-based data space. All these services and concepts are described in the following paragraphs.

Data access In order to avoid building a new proprietary solution and reimplementing solved aspects of *Grid Data Services* (*GDS*) (Antonioletti, *et al.*, 2003), we have decided to integrate our developed concepts into OGSA-DAI (Antonioletti, *et al.*, 2005). OGSA-DAI (see also Chapter 14 in this volume) is a middleware implementation of GDS for supporting access and integration of data from separate data sources within a grid. Moreover, the OGSA-DAI allows us to define extensions (called activities), which can internally transform data into the required form and deliver them as a file or as a data stream. The middleware uses XML as its native format for data representation. Therefore, XML was also adopted by data mining services implemented in GridMiner as standard input data format so no further conversion of the data is necessary. The latest release of the OGSA-DAI is implemented as a WSRF service and is also available as a core component for data access and integration in the Globus Toolkit. For all these reasons the OGSA-DAI was included in the GridMiner architecture as an essential service supporting data mining services with the access to the data sources on the grid.

Data integration Modern science, business, industry and society increasingly rely on global collaborations, which are very often based on large-scale linking of databases that were not expected to be used together when they were originally developed. We can make decisions and discovery with data collected within our business or research. But we improve decisions or increase the chance and scope of discoveries when we combine information from multiple sources. Then correlation and patterns in the combined data can support new hypotheses, which can be tested and turned into useful knowledge. Our approach to data integration is based on the data mediation concept, which allows for the simplification of work with multiple, federated and usually geographically distributed data sources. The realization of the 'wrapper/mediator' approach is supported by the *Grid Data Mediation Service* (*GDMS*) (Woehrer, Brezany and Tjoa, 2005), which creates a single, virtual data source. The GDMS allows integration of heterogeneous relational databases, XML databases and comma-separated value files into one logically homogeneous data source.

Data statistics, data understanding, data pre-processing The D^{3}G framework (Wöhrer, *et al.*, 2006), developed within the GridMiner project, describes an architecture for gathering data statistics on the fly and uses them in remote data pre-processing methods on query results. Additionally, it provides access to incrementally maintained data statistic of pre-defined database tables via the GDS. The implementation of the framework is based on an extension of OGSA-DAI with a set of new activities allowing for the computation of different kinds of data statistic. The framework provides a set of OGSA-DAI activities for extracting descriptive and advanced statistics, together with a specialized monitoring tool for continuously updating statistics.

- *Descriptive statistics.* Could be viewed as a metadata extraction activity providing basic information about the selected tables for nominal and numerical attributes.
- *Advanced statistics.* Allow for the enhancement of basic statistics into advanced data statistics based on additional input information about a data set, for example statistics for intervals.
- *Data source monitoring.* Provides an interface to the incrementally maintained statistics about whole tables returning more recent statistics.

All data statistics are presented in the PMML format to support service interoperability. The gathered statistical data are useful in deciding what data pre-processing technique to use in the next phase or whether it is necessary at all, which is especially interesting for very expensive pre-processing methods.

DataEx The aim of this research task is to develop a new model for scientific data management on the grid, which extends our earlier work on structural and semantic data integration (Woehrer, Brezany and Tjoa, 2005) to provide a new class of data management architecture that responds to the rapidly expanding demands of large-scale data exploration. DataEx follows and leverages the visionary ideas of *dataspaces*, a concept introduced by Franklin, Halevy and Maier (2005). A dataspace consists of a set of participants and a set of relationships. Participants are single data sources (elements that store or deliver data). Relationships between them should be able to model traditional correlations, for example one is a replica of another one, as well as novel future connections, such as two independently created participants contain information about the same physical object. Dataspaces are not a data integration approach – rather more a data co-existence approach.

3.4.4 Data mining services and OLAP

Sequential, parallel and distributed data mining algorithms for building different data mining and text mining models were proposed and implemented as grid services using the multi-level concurrency approach. Each service can be used either autonomously or as a building block to construct distributed and scalable services that operate on data repositories integrated into the grid and can be controlled by the workflow engine. Within the project, the following services providing data and text mining tasks were developed.

- *Decision trees.* Distributed version of grid service able to perform high-performance classification based on the SPRINT algorithm (Shafer, Agrawal and Mehta, 1996). The service is implemented as a stateful OGSA service able to process and distribute data provided by data sources represented by the OGSA-DAI interface. The algorithm and the service implementation details are described by Brezany, Kloner and Tjoa (2005).
- *Sequence patterns.* Implementation of the out-of-core sequence mining algorithm SPADE producing a sequence model and its evaluation. The work on the service is in progress.
- *Text classification.* Parallel implementation of the tree classification algorithm operating over text collections represented in the XML format. Implementation details and performance results can be found elsewhere (Janciak, *et al.*, 2006).
- *Clustering.* Service enabling discovery of a clustering model using the k-means algorithm and evaluation of the final model. The work on the service is in progress.
- *Neural networks.* Parallel and distributed version of the back-propagation algorithm implemented as a high-performance application in the Titanium language (a parallel Java dialect) and fully tested on HPC clusters. The used algorithm and implementation details together with performance results can be found elsewhere (Brezany, Janciak and Han, 2006).
- *Association rules.* A specialized service operating on top of OLAP using its data cube structure as a fundamental data source for association rule discovery. A detailed description of the design and implementation of the service can be found elsewhere (Elsayed and Brezany, 2005).

All the services were fully tested in our local test bed, and the performance results were compared with other implementations. Moreover, the visualization application was integrated into the graphical user interface and can visualize the models of the developed services, which are represented in the PMML format.

OLAP So far, in the grid community, no significant effort has been devoted to data warehousing and associated OLAP, which are kernel parts of modern decision support systems. Our research effort on this task resulted in several original solutions for scalable OLAP on the grid.

Because there was no appropriate open source OLAP system for proof-of-concept prototyping of developed concepts available, we decided to develop appropriate design and implementation patterns from scratch. First, we designed the architecture of a sequential grid OLAP Engine (Fiser, *et al.*, 2004) as a basic building block for distributed OLAP on the grid. This architecture includes the data cube structure, an index database and function blocks for cube construction, querying and connection handling. During investigation of existing cube storage and indexing schemes, we discovered that there was no suitable method available for management of sparse OLAP cubes on the grid. The existing methods, for example *Bit-Encoded Binary Structure* (*BESS*) (Goil and Choudhary, 1997), require that the number of positions within each cube dimension is known before the records are imported from data repositories into the cube. However, such exact information is often not available in grid applications. Therefore, we proposed a method called *Dynamic Bit Encoding* (*DBE*) as an extension of BESS. DBE is based on basic bit logic operations; coding and indexing are fully extendable during the cube construction.

With this concept it was possible to focus our research on a parallel and distributed OLAP solution. In our approach, the OLAP cube is assumed to be built on top of distributed computational and data storage resources – the goal is to merge independent systems into one large virtual federation to gain high computational power and storage capacity. However, the OLAP Management System has to virtualize all these resources. Therefore, for the end user and other applications, we consider this data cube as one virtual cube, which is managed by resources provided by cooperating partners. Cube segments (sets of chunks) are assigned to grid nodes, which are responsible for their management.

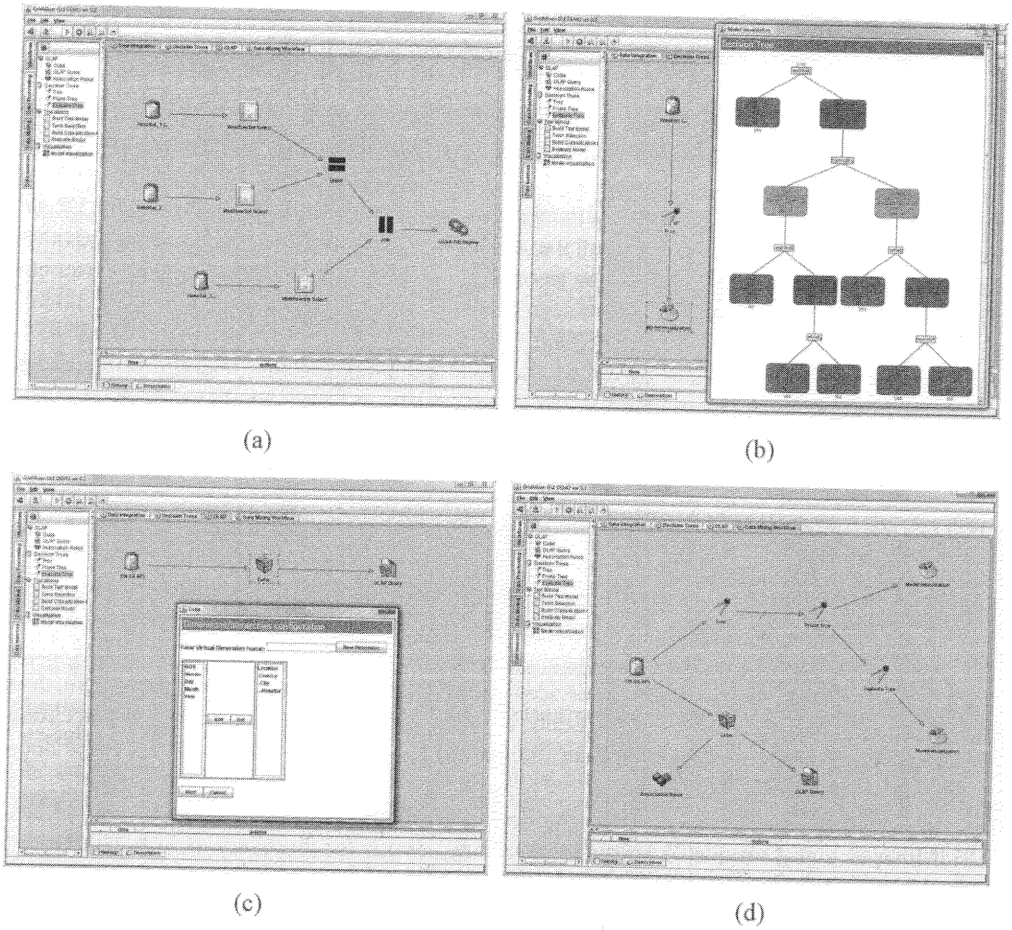
A most helpful programming model for the development of scalable scientific applications for distributed-memory architectures is the SPMD model (SPMD: single program–multiple data). This model can be applied to data parallel problems in which one and the same basic code executes against partitioned data. We took an analogous approach to the development of scalable OLAP applications on the grid. We proposed the *SSMC model (SSMC: single service–multiple cube data)*. In this model, there is only one OLAP service code, which is used by the grid OLAP Service Factory for generation of a set service instances. The instances are then initiated on appropriate grid nodes due to the configuration specification, which can be smartly designed in the administration domain of the GridMiner graphical user interface and can be arbitrarily extended. For each service instance, it is only necessary to get information about the locations of its immediate children's services due to the hierarchical system architecture, data distribution and indexing strategies. The cube data aggregation operations and query processing are performed in each node by a pool of threads, which can really run in parallel if the node includes multiple processors.

The sequential and parallel OLAP engines were implemented in Java and integrated into GridMiner as a grid service. A special data mining service for discovery of association rules from OLAP cubes created and accessed by the OLAP service was proposed and implemented. The results of data mining and OLAP are represented by the *OLAP Model Markup Language* (*OMML*), which was proposed by our project.

3.4.5 Security

Security of the data and service access is extremely important in GridMiner, which is designed for distributed data access and manipulation. In cooperation with the *Cancer Bioinformatics Grid* (*caBIG*) project, we integrated *Dorian* (Langella, *et al.*, 2006) as a federated identity management service into the GridMiner framework. Dorian provides a complete grid-enabled solution, based on public key certificates and the *Security Assertion Markup Language* (*SAML*), for managing and federating user identities in a grid environment. SAML has been developed as a standard for exchanging authentication and authorization statements between security domains.

To obtain grid credentials or a proxy certificate, a digitally signed SAML assertion, vouching that the user has been authenticated by his/her institution, has to be sent to Dorian in exchange for grid credentials. The user authorization is accomplished through the user's identity. Dorian can only issue grid credentials to users that supply an SAML assertion from a Trusted Identity Provider. Dorian also provides a complete graphical user interface.



(a) (b) (c) (d)

Figure 3.4 Graphical user interface. (a) Data integration, (b) decision tree visualization, (c) OLAP cube configuration, (d) workflow composition

Secure data mining services have their own credentials consisting of a private key and certificate assigned by a Grid Certificate Authority. The services are trusted based on the service identity; thus, services have full access authorization to other secure services.

Data access security The software architecture of the grid database access service OGSA-DAI provides the basic security mechanisms, which can be specified in a special security configuration file (RoleMap file). The application developers can extend the functionality of this approach. This idea is also followed in the GridMiner application of OGSA-DAI.

3.5 Graphical user interface

The aim of the *graphical user interface* (*GUI*) is to hide the complexity of the system architecture and offer an easy-to-use front end for the data mining experts as well as for the system administrators.

The GUI was designed as a stand alone Java application able to be remotely started by 'Java Web Start' practically in any operating system supporting Java. It provides a browser window allowing a user to interact with several Web applications used to configure and parameterize data mining tasks and interact with a specialized Web application for the workflow execution control. Another important contribution of the GUI is the possibility to interactively construct data mining workflows in a graphical representation on abstract and concrete levels (see Figure 3.4(d)). The workflow is created using drag-and-drop functionality and can be presented as a graph as well as being specified in the workflow language. Additionally, for the data integration tasks it is possible to construct mediation schemas used by GDMS (see Figure 3.4(a)). In summary, the main functionality of the GUI includes the following.

- Training and test data set selection based on query preparation and execution.
- Parameterization of tasks and configuration of the grid services.
- Data mining workflow composition and its execution.
- Controlling and monitoring of workflow execution.
- Visualizing data mining models, statistics and results of OLAP queries.

The layout of the GridMiner GUI is divided into three main panels as follows.

- *Resource panel.* The panel on the left hand side is used to manage grid services. They are clustered due to their relationship with the phases of the KDD process and are therefore categorized into data processing and data mining groups. The panel allows for the addition and selection of new data sources and management of created workflow instances.
- *Workflow composition panel.* This central panel of the GUI allows for the creation and modification of data mining workflow instances consisting of the resources selected from the resource panel using drag-and-drop mechanisms.
- *Log panel.* The bottom panel displays log and debug messages as well as notification messages during the workflow execution.

All the services of the workflow instance can be configured and parameterized by the user. Double-clicking on the resource icon in the *Workflow composition panel* opens a window displaying the Web application responsible for configuration of the particular service (see Figure 3.4(c)). Double-clicking on the visualization icon corresponding to the mined model opens a visualization application (see Figure 3.4(b)).

Visualization Several visualization methods for different data mining models have been applied in the GridMiner project, namely visualization for decision trees, association rules, neural networks, sequences and clustering. The visualization is supported by a Web application converting data mining models or statistics stored in PMML documents to their graphical or textual representations. The graphical representation is made by converting the models into the widely supported *Scalable Vector Graphics* (*SVG*) format, which can be displayed by standard Web browsers. A specialized interface for OLAP query visualization is also supported by a dynamic webpage displaying aggregated values and dimensions of the OLAP cube.

3.6 Future developments

From the previous paragraphs, we can see that we are already witnessing various grid-based data mining research and development activities. There are many potential extensions of this work towards comprehensive, productive and high-performance analysis of scientific data sets. Below we outline three promising future research avenues.

3.6.1 High-level data mining model

The most popular database technologies, like the relational one, have precise and clear foundations in the form of uniform models and algebraic and logical frameworks on which significant theoretical and system research has been conducted. The emergence of the relational query language SQL and its query processors is an example of this development. A similar state has to be achieved in data mining in order for it to be leveraged as a successful technology and science. The research in this field will include three main steps: (i) identifying the properties required of such a model in the context of large-scale data mining, (ii) examining the currently available models that might contribute (e.g. CRISP-DM) and (iii) developing a formal notation that exposes the model features.

3.6.2 Data mining query language

From a user's point of view, the execution of a data mining process and the discovery of a set of patterns can be considered as either the result of an execution of a data mining workflow or an answer to a sophisticated database query. The first is called the procedural approach, while the second is the descriptive approach. GridMiner and other grid-based data mining developments are based on this model. To support the descriptive approach several languages have been developed. The *Data Mining Query Language* (*DMQL*) (Han, 2005) and OLE DB for Data Mining (Netz, *et al.*, 2001) are good examples. A limitation of these languages is their poor support for data preparation, transformation and post-processing. We believe that the design of such a language has to be based on an appropriate data mining model; its features will be reflected by the syntactic and semantic structure of the language. Moreover, implementation of queries expressed in the language on the grid, which represents a highly volatile distributed environment, will require investigation of dynamic approaches able to sense the query execution status and grid status in specified time intervals and appropriately adapt the query execution plan.

3.6.3 Distributed mining of data streams

In the past five years with advances in data collection and generation technologies, a new class of application has emerged that requires managing data streams, i.e. data composed of continuous, real time sequence of items. A significant research effort has already been devoted to stream data management (Chaudhry, Shaw and Abdelguerfi, 2005) and data stream mining (Aggarwal, 2007). However, in advanced applications there is a need to mine multiple data streams. Many systems use a centralized model (Babcock, *et al.*, 2002). Here, the distributed data streams are directed to one central location before they are mined. Such a model is limited in many aspects. Recently, several researchers have proposed a distributed model considering distributed data sources and computational resources–an excellent survey was provided by

Parthasarathy, Ghoting and Otey (2007). We believe that investigation of grid-based distributed mining of data streams is also an important future research direction.

3.7 Conclusions

The characteristics of data exploration in modern scientific applications impose unique requirements for analytical tools and services as well as their organization into scientific workflows. In this chapter, we have introduced our approach to the e-science analytics that has been incorporated into the GridMiner framework. The framework aims to exploit modern software engineering and data engineering methods based on service-oriented Web and grid technologies. An additional goal of the framework is to give the data mining expert powerful support to achieve appealing results. There are two main issues driving the development of such an infrastructure. The first is the increasing volumes and complexity of involved data sets, and the second is the heterogeneity and geographic distribution of these data sets and the scientists who want to analyse them. The GridMiner framework attempts to address both challenges, and we believe successfully. Several data mining services have been already deployed and are ready to perform the knowledge discovery tasks and OLAP.

References

Aggarwal, C. (2007), *Data Streams: Models and Algorithms*, Advances in Database Systems, Springer.

Antonioletti, M., Atkinson, M., Baxter, R., Borley, A., Hong, N. P. C., Collins, B., Hardman, N., Hume, A. C., Knox, A., Jackson, M., Krause, A., Laws, S., Magowan, J., Paton, N. W., Pearson, D., Sugden, T., Watson, P. and Westhead, M. (2005), 'The design and implementation of grid database services in OGSA-DAI: Research articles', *Concurrency and Computation: Practice and Experience* **17** (2–4), 357–376.

Antonioletti, M., Hong, N. C., Atkinson, M., Krause, A., Malaika, S., McCance, G., Laws, S., Magowan, J., Paton, N. and Riccardi, G. (2003), 'Grid data service specification', http://www.gridpp.ac.uk/papers/DAISStatementSpec.pdf

Babcock, B., Babu, S., Datar, M., Motwani, R. and Widom, J. (2002), Models and issues in data stream systems, *in* 'Proceedings of the 21st ACM SIGMOD–SIGACT–SIGART Symposium on Principles of Database Systems (PODS'02)', ACM, New York, NY, pp. 1–16.

Berkley, C., Bowers, S., Jones, M., Ludaescher, B., Schildhauer, M. and Tao, J. (2005), Incorporating semantics in scientific workflow authoring, *in* 'Proceedings of the 17th International Conference on Scientific and Statistical Database Management (SSDBM'05)', Lawrence Berkeley Laboratory, Berkeley, CA, pp. 75–78.

Brezany, P., Janciak, I. and Han, Y. (2006), Parallel and distributed grid services for building classification models based on neural networks, *in* 'Second Austrian Grid Symposium', Innsbruck.

Brezany, P., Janciak, I. and Tjoa, A. M. (2007), Ontology-based construction of grid data mining workflows, *in* H. O. Nigro, S. E. G. Císaro and D. H. Xodo, eds, 'Data Mining with Ontologies: Implementations, Findings, and Frameworks', Hershey, New York, pp. 182–210.

Brezany, P., Kloner, C. and Tjoa, A. M. (2005), Development of a grid service for scalable decision tree construction from grid databases, *in* 'Sixth International Conference on Parallel Processing and Applied Mathematics', Poznan.

Brezany, P., Tjoa, A. M., Rusnak, M. and Janciak, I. (2003a), Knowledge grid support for treatment of traumatic brain injury victims, *in* '2003 International Conference on Computational Science and its Applications', Vol. 1, Montreal, pp. 446–455.

Brezany, P., Tjoa, A. M., Wanek, H. and Woehrer, A. (2003b), Mediators in the architecture of grid information systems, *in* 'Fifth International Conference on Parallel Processing and Applied Mathematics, PPAM 2003', Springer, Czestochowa, Poland, pp. 788–795.

Cannataro, M. and Talia, D. (2003), 'The knowledge grid', *Communications of the ACM* **46** (1), 89–93.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R. (2000), *CRISP-DM 1.0: Step-by-Step Data Mining Guide*, CRISP-DM Consortium: NCR Systems Engineering Copenhagen (USA and Denmark) DaimlerChrysler AG (Germany), SPSS Inc. (USA) and OHRA Verzekeringen en Bank Groep BV (The Netherlands).

Chaudhry, N., Shaw, K. and Abdelguerfi, M. (2005), *Stream Data Management (Advances in Database Systems)*, Springer.

Churches, D., Gombas, G., Harrison, A., Maassen, J., Robinson, C., Shields, M., Taylor, I. and Wang, I. (2006), 'Programming scientific and distributed workflow with Triana services: Research articles', *Concurrency and Computation: Practice and Experience* **18** (10), 1021–1037.

Ćurčin, V., Ghanem, M., Guo, Y., Köhler, M., Rowe, A., Syed, J. and Wendel, P. (2002), Discovery Net: towards a grid of knowledge discovery, *in* 'Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'02)', ACM Press, New York, NY, pp. 658–663.

Czajkowski, K., Ferguson, D., Foster, I., Frey, J., Graham, S., Maguire, T., Snelling, D. and Tuecke, S. (2004), 'From Open Grid Services Infrastructure to WS-Resource Framework: Refactoring and Evolution, version 1.1', Global Grid Forum.

Data Mining Group(2004), 'Predictive Model Markup Language', http://www.dmg.org/

Elsayed, I. and Brezany, P. (2005), Online analytical mining of association rules on the grid, Technical Report Deliverable of the TU/Uni-Vienna GridMiner Project, Institute for Software Science, University of Vienna.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996), 'From data mining to knowledge discovery in databases', *Ai Magazine* **17**, 37–54.

Fiser, B., Onan, U., Elsayed, I., Brezany, P. and Tjoa, A. M. (2004), Online analytical processing on large databases managed by computational grids, *in* 'DEXA 2004'.

Foster, I. and Kesselman, C. (1998), The Globus Project: a status report, *in* 'Proceedings of the 7th Heterogeneous Computing Workshop (HCW'98)', IEEE Computer Society, Washington, DC, p. 4.

Foster, I., Kesselman, C., Nick, J. M. and Tuecke, S. (2002), 'The physiology of the grid: an open grid services architecture for distributed systems integration', http://www.globus.org/research/ papers/ogsa.pdf

Franklin, M., Halevy, A. and Maier, D. (2005), 'From databases to dataspaces: a new abstraction for information management', *SIGMOD Record* **34** (4), 27–33.

Goil, S. and Choudhary, A. (1997), 'High performance OLAP and data mining on parallel computers', *Journal of Data Mining and Knowledge Discovery* **1** (4), 391–417.

Han, J. (2005), *Data Mining: Concepts and Techniques*, Morgan Kaufmann, San Francisco, CA.

Hofer, J. and Brezany, P. (2004), DIGIDT: distributed classifier construction in the grid data mining framework GridMiner-Core, *in* 'Workshop on Data Mining and the Grid (DM-Grid 2004), held in conjunction with the 4th IEEE International Conference on Data Mining (ICDM'04)', Brighton, UK.

Janciak, I., Kloner, C. and Brezany, P. (2007), 'Workflow Enactment Engine Project', http://weep.gridminer.org

Janciak, I., Sarnovsky, M., Tjoa, A. M. and Brezany, P. (2006), Distributed classification of textual documents on the grid, *in* 'The 2006 International Conference on High Performance Computing and Communications, LNCS 4208', Munich, pp. 710–718.

Jones, M. B., Ludaescher, B., Pennington, D., Pereira, R., Rajasekar, A., Michener, W., Beach, J. H. and Schildhauer, M. (2006), 'A knowledge environment for the biodiversity and ecological sciences', *Journal of Intelligent Information Systems* **29** (1), 111–126.

Kargupta, H., Joshi, A., Sivakumar, K. and Yesha, Y., eds(2003), *Data Mining: Next Generation Challenges and Future Directions*, AAAI/MIT Press, Menlo Park, CA.

Kickinger, G., Brezany, P., Tjoa, A. M. and Hofer, J. (2004), Grid knowledge discovery processes and an architecture for their composition, *in* 'IASTED Conference', Innsbruck, pp. 76–81.

Kickinger, G., Hofer, J., Tjoa, A. M. and Brezany, P. (2003), Workflow management in GridMiner, *in* '3rd Cracow Grid Workshop', Cracow.

Langella, S., Oster, S., Hastings, S., Siebenlist, F., Kurc, T. and Saltz, J. (2006), Dorian: grid service infrastructure for identity management and federation, *in* 'Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)', IEEE Computer Society, Washington, DC, pp. 756–761.

Netz, A., Chaudhuri, S., Fayyad, U. M. and Bernhardt, J. (2001), Integrating data mining with SQL databases: OLE DB for data mining, *in* 'Proceedings of the 17th International Conference on Data Engineering', pp. 379–387.

Parthasarathy, S., Ghoting, A. and Otey, M. E. (2007), A survey of distributed mining of data streams, *in* C. C. Aggarwal, ed., 'Data Streams: Models and Algorithms', Springer, pp. 289–307.

Pyle, D. (1999), *Data Preparation for Data Mining*, Morgan Kaufmann San Francisco, CA.

Romberg, M. (2002), 'The UNICORE grid infrastructure', *Scientific Programming* **10** (2), 149–157.

Sarang, P., Mathew, B. and Juric, M. B. (2006), *Business Process Execution Language for Web Services, 2nd Edition. An Architects and Developers Guide to BPEL and BPEL4WS*, Packt.

Shafer, J. C., Agrawal, R. and Mehta, M. (1996), SPRINT: A scalable parallel classifier for data mining, *in* T. M. Vijayaraman, A. P. Buchmann, C. Mohan and N. L. Sarda, eds, 'Proceedings of the 22nd International Conference on Very Large Databases (VLDB'96)', Morgan Kaufmann, pp. 544–555.

Stankovski, V., Swain, M., Kravtsov, V., Niessen, T., Wegener, D., Kindermann, J. and Dubitzky, W. (2008), 'Grid-enabling data mining applications with DataMiningGrid: an architectural perspective', *Future Generation Computer Systems* **24**, 259–279.

Taylor, I. J., Deelman, E., Gannon, D. B. and Shields, M. (2007), *Workflows for e-Science: Scientific Workflows for Grids*, Springer, Secaucus, NJ.

Wang, J. T.-L., Zaki, M. J., Toivonen, H. and Shasha, D., eds(2005), *Data Mining in Bioinformatics*, Springer.

Woehrer, A., Brezany, P. and Tjoa, A. M. (2005), 'Novel mediator architectures for grid information systems', *Future Generation Computer Systems* **21** (1), 107–114.

Wöhrer, A., Novakova, L., Brezany, P. and Tjoa, A. M. (2006), D^3G: novel approaches to data statistics, understanding and pre-processing on the grid, *in* 'IEEE 20th International Conference on Advanced Information Networking and Applications', Vol. 1, Vienna, pp. 313–320.

Wolstencroft, K., Oinn, T., Goble, C., Ferris, J., Wroe, C., Lord, P., Glover, K. and Stevens, R. (2005), Panoply of utilities in Taverna, *in* 'Proceedings of the First International Conference on e-Science and Grid Computing (E-SCIENCE'05)', IEEE Computer Society, Washington, DC, pp. 156–162.