Knowledge Grid Support for Treatment of Traumatic Brain Injury Victims*

Peter Brezany[1], A. Min Tjoa[2], Martin Rusnak[3]
Ivan Janciak[1]

[1] Institute for Software Science
University of Vienna, Liechtensteinstrasse 22, A-1090 Vienna, Austria
brezany@par.univie.ac.at
[2] Institute for Software Technology and Mutimedia Systems
Vienna University of Technology, Vienna, Austria
tjoa@ifs.tuwien.ac.at
[3] International Brain Trauma Foundation, Vienna, Austria
rusnakm@igeh.org

Abstract. Traumatic brain injuries (TBIs) typically result from accidents in which the head strikes an object. Among all traumatic causes of death, the TBI is the most serious one. Moreover, TBI can also significantly affect many cognitive, physical, and psychological skills of the cases that survive. It is clear from the data presented in many international studies, that improvement of outcomes is useful not only for the patients, but also for entire national economy.

This paper introduces a new software infrastructure, called TBI Knowledge Grid that will provide health professionals involved in the treatment and management of TBI patients with Information Technology tools, which will enable them to discover and access relevant knowledge and information from different distributed and heterogeneous data sources.

1 Motivation

Traumatic brain injuries (TBIs) typically result from accidents in which the head strikes an object. Among all traumatic causes of death, the TBI is the most serious one, besides injuries of the heart and great vessels. Moreover, TBI can also significantly affect many cognitive, physical, and psychological skills. Based on results from one Austrian study, in 1997 about 2.000 Austrian citizens experienced severe TBI. One third of those died at the scene, and about 2/3 were admitted alive into a hospital. Some 20 to 30 percent of all patients admitted to a hospital died, another 10 to 20 percent survived with severe disabilities, and approximately 50 to 60 percent of patients completely or partially recovered. Each year, abut two million people in the United States sustain a head injury; a lot of them are TBI cases. Children represent one special risk group.

The treatment of TBI patients is very resource intensive and frequently involves the patient staying in a Neurological Intensive Care Unit for an extended period. Besides

* The work described in this paper is being carried out as part of the research projects "Aurora" supported by the Austrian Research Foundation.

dealing with the initial accident (event), the medical team has to cope with secondary events such as swelling of the brain. As the brain is encased in the skull, treatment procedures are more limited than for the rest of the body. The primary clinical strategy appears to be alleviating extreme symptoms such as high intracranial pressure and relying on the self-healing mechanisms of the organ.

Predicting the outcome of seriously ill patients is a challenging problem for clinicians. There are still many situations where it is very hard to predict whether a patient will survive given his or her state on admission to hospital. This makes it difficult to choose the best course of treatment. Head injury has the added complication that it is very difficult to devise effective clinical trials, as the brain is largely inaccessible and experimentation can have serious consequences. One alternative to clinical trials is to analyze existing patient data in an attempt to predict the several outcomes, and to suggest therapies.

It is clear from what has been presented above that improvement of outcomes is useful not only for the patients, but also for entire national economy. TBI cases require highly technical evaluation and extensive attention to detail. This goals can only be achieved by a broad collaboration of people involved in the treatment and management of TBI patients. This collaboration has to be supported by appropriate information systems.

The general objective of the project, which is described in this paper is to provide health professionals involved in the management of TBI patients with information technology tools, which enable them to discover and access relevant knowledge and information from different distributed and heterogeneous data sources. Because of strong collaborative aspects of this application and geographical distribution of people and resources, the tools developed are integrated within an infrastructure, called **TBI Knowledge Grid** (TBI-KG), which is being developed on top of the existing Grid technology [11, 23]. TBI-KG allows accesses from different platforms (desktop, local wireless, and hand-held devices) used at the accident scene, ambulance, hospital bed, strategic specialist consultations, etc. The accesses are used for data generation (production), querying, and analysis.

The analytical part of the project focuses its effort on data mining [15] and On-Line Analytical Processing (OLAP)[4] [7], two complementary technologies, which, if applied in conjunction, can provide a highly efficient[5] and powerful data analysis and knowledge discovery solution on the Grid.

The following parts of the paper are organized as follows. Section 2 provides a brief description of the TBI application. Section 3 deals with the overall TBI-KG system design. Section 4 introduces the architecture and functionality of the Analysis Component of the TBI-KG. Data integration aspects are briefly discussed in Section 5. Section

[4] In OLAP, a specific data structure, called Data Cube, is used to provide a multidimensional view of a data warehouse where a critical value, e.g. intraventrical pressure in our application, is organized by several dimensions, for example, by sex, age, type and location of a catherer, pathophysical value, date of the measurement, etc.

[5] The OLAP data cube is a reduced representation of the mined dataset that is much smaller in volume, yet closely maintains the integrity of the original data; so the number of accesses and the volume of the I/O data is reduced.

6 outlines components of the TBI-KG architecture. Related work is discussed in Section 7. We briefly present our conclusions in Section 8.

2 Application Explication

The trajectory of the TBI patient management includes the following main points:

- *Trauma event* (e.g. injury at a car or sport accident). A first aid service is called; the initial data providing a brief description of the context of the accident and condition of the victim can flow into the database.
- *First aid.* The patient is treated by an Emergency Medical Service (EMS). If he is severe injured, his verbal and motoric communication ability is low or none. In some cases, his personal data can be identified. The data, representing the first examination results are recorded in a patient database, using remote access. Based on these data and mobile phone communication, the emergency control centre decides into which hospital the patient will be transported. Additional data about the patient (e.g. information about allergy to some drugs), can then be gathered, for instance by interviewing his relatives or his general practitioner, if necessary, and if his identity is known.
- *Transportation to hospital.* In the ambulance, the EMS team performes additional treatment and data recording.
- *Acute hospital care.* Immediately after the patient has been admitted, the first goal is to stabilize his state. During this phase, a large amount of the patient's data is collected by a set of examinations (e.g., basis laboratory tests, temperature, blood pressure and heart beat frequency measurements, brain imaging, and examination of other injured body parts). Selection of an optimal stabilization treatment and successive daily care procedures is extremely important because research has proven that all brain damage does not occur at the moment of accident, but rather evolves over the ensuing hours and days after the initial injury, due to brain swelling. Eliminating of risk factors for secondary damage is the objective for optimal care provided to a patient with severe TBI.
- *Home care.* After hospital leaving, a long-term outcome (e.g. psychological state, rehabilitation conditions, etc.) is monitored, and the data observed or measured are sent to the patient database.

All the above TBI patient management phases are associated with data collection into databases, which are currently autonomously managed by individual hospitals, and are, therefore, heterogeneous. Knowledge discovery in these databases can significantly improve the quality of the decisions taken. Moreover, in the future, this knowledge can assist TBI experts in preparation of scientific evidence-based treatment guidelines [4]. Further, various statistics and results of the analytical processing allow to evaluate and compare different treatment and management methods and performance of individual hospitals.

3 The System Design

Although rich and powerful OLAP and data mining functions form the core of the TBI-KG system, the design of its overall architecture is critically important.

The basic management of the project resources is performed by the Computational and Data Grid services provided by basic Grid infrastructures, e.g. the Globus toolkit [11]. To convert the large, often low-level, heterogeneous TBI data into powerful knowledge, additional application Grid layers have to be built on top of the existing Grid. We address this issue by extending and further developing the layered Grid architecture model proposed by K. G. Jeffery [16], who introduces Information and Knowledge Grid layers on top of Computational and Data Grid layers. The role of our *Information Grid* layer is to integrate heterogeneous information into a homogeneous presentation to the *Knowledge Grid* layer, and provide the appropriate OLAP functionality. The task of the Knowledge Grid layer is to extract knowledge from information provided by OLAP and/or from original data sources. The knowledge extracted is then being presented to the human user or to another application by an appropriate interface. Components within each layer share common characteristics, but can built on capabilities and behaviors provided by any lower layer.

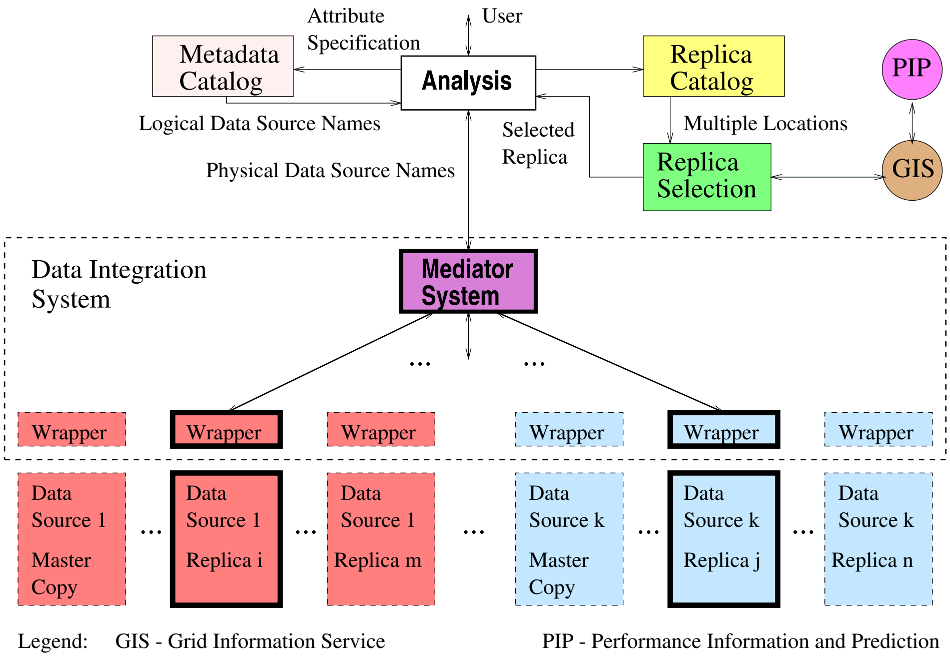

Fig. 1. The functional and data resource access model

The next step is the design of an appropriate functional and data access model. Fig. 1 shows how accesses to distributed data resources may be logically organized. The interactions of individual components are discussed in the context of a data mining query; processing an OLAP query can be explained in an analogous way. The user specifies appropriate knowledge discovery scenario by a graphical user interface or data mining query language. The query includes specification of the data mining task (e.g., mining association rules, classification of TBI cases, etc.) and specification of the data sources to be mined. The query is passed to the *Analysis System*, where it is analyzed, and then the following activities are initiated by the Analysis System:

(a) Determining physical dataset names – The data sources can be specified by logical names or by attributes. If an attribute specification is used, the corresponding logical name (or a collection of names) is found by means of a *Metadata Catalog*, and used for selection of the replica [25], which fulfills several requirements (e.g., minimal access time). The *Replica Selection* takes its decision on the basis of information provided by the *Grid Information Services* (GIS) component, which provides system configuration and status information such as computer server and network status, and performance information provided by the *Performance Information and Prediction (PIP)* component.

(b) Resource Allocation and Management – Based on results from query analysis results, computational and storage resource allocation requirements are addressed to the resource management component of the Grid.

(c) Data Mining – An essential process where intelligent methods are applied in order to extract data patterns from distributed data sources. The data mining component specifies accesses to data sources by queries, which are passed together with physical source data names to the *Mediator* system, which, with help of *wrappers* provides an integrated view of heterogeneous data to the data mining component. The Mediator centralizes the information provided by the wrappers in a unified view of all available data (stored in a metadata catalog), decomposes the user query in small queries (executable by the wrappers), gathers the partial results and computes the answer to the user query.

(c) Knowledge Presentation – Visualization and knowledge representation techniques are used to present the mined knowledge to the user.

Issues related to the Grid catalogs, replica management, replica selection, etc. are addressed by several Grid projects. We focus our research effort on the Analysis System, Mediator System, and appropriate Grid extensions.

4 Analysis Component

Architecture. The architecture components and logical relations among them are depicted in Fig. 2. The system accepts on-line queries (or commands) via a graphical user interface API or command lines. If the query result is found in the user query result cache, it is answered immediately, otherwise, it is passed to the Grid nodes.

An OLAP query is processed by the *Grid OLAP Engines (GOEs)*. The *Query Processors (QPs)* of GOEs first try to answer this query using the contents of the Grid query cache. If this is not possible, the appropriate OLAP cubes must be searched. If the cubes are found in the Grid cube cache, the answer is computed from them, otherwise, either

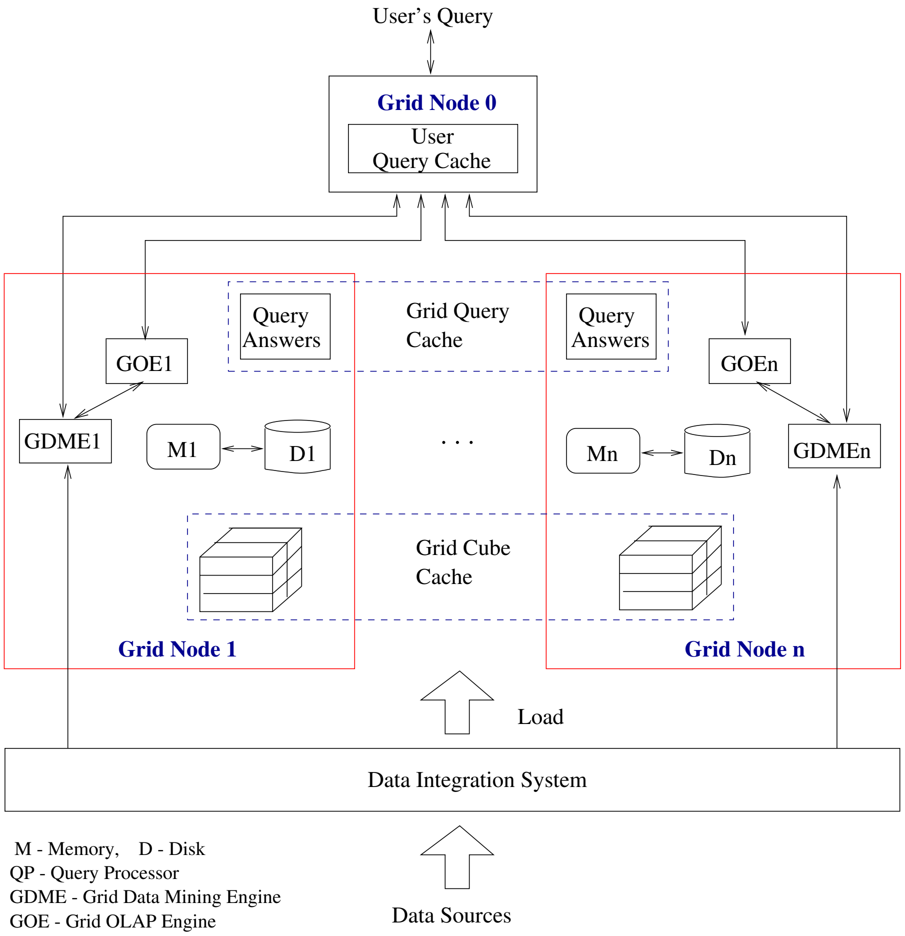

Fig. 2. Major components and structure of the Analysis System architecture

new cubes are computed from the cubes stored in the cache, or from the data sources, which are accessed by means of the *Data Integration System (DIS)*.

A data mining query is passed to the *Data Mining Engines (DMEs)*. The answer is computed either directly from the databases accessed by DIS or from the data cubes, which are accessed from DMEs using OLAP queries. In the last case, aggressive prefetching can be applied by GOEs, because cube access scenarios can be specified in advance.

All the above operations are supported by an appropriate *Metadata Repository* which, among others, includes information about cube objects.

Query Languages. Recently, Microsoft proposed languages for data mining [20] and OLAP [21]. Other similar research efforts are reported in [15]. In the example introduced in the next paragraph, we use the notation based on the DMQL language [15].

```
use database db1, db2
mine classifications
analyze outcome
using g_parsimony_class
display as decision_tree
```

Fig. 3. Query example.

Data Mining Engine. The data mining process is initiated by a data mining query that specifies the task-relevant data, the desired kinds of knowledge to be minded, the associated constraints, interestingness thresholds, and so on. For example, to mine patterns classifying victim outcomes, where the classes are determined by the attribute *outcome* from two Grid databases, the extended DMQL specification introduced in Fig. 3 can be used, where *db1* and *db2* are logical database names; each of them identifies one or a set of physical databases at different storage locations[6]. The (optional) clause **using** specifies that a classification algorithm called *g_parsimony_class* should be applied. It is our Grid extension of the OLAP cube based parallel algorithm developed by Goil [13]. A set of other classification algorithms can be considered, e.g. the distributed algorithm based on the *Collective Data Mining Method* [17] proposed by Kargupta et al. The query result is to be displayed by a decision tree (the **display as** clause).

A data mining query can be answered either by accessing the data sources specified in the query or by creating appropriate data cubes and then using the cube's aggregates in the data mining algorithm [6].

OLAP Engine. The Grid OLAP Engine (GOE) is based on a *multidimensional OLAP (MOLAP)* [27]. We use so called "chunked" organization [8] for storing OLAP cubes.

We investigate two cube construction and query processing strategies, as discussed below.

1. Persistent cubes: This approach is used by the traditional *data warehousing* technology, which follows the *eager* approach, where data integration occurs in a separate materialization step, *before* the actual user queries. The cube is constructed from the source data and updated periodically offline. Consequently, updates to the source databases are not immediately applied to the cube, resulting in a considerable *update window*, which cannot be acceptable in several use-cases of our application. Moreover, if many aggregation paths are considered, the size of the pre-materialized data cube could exceed the cumulated size of the original Grid database sources by orders of magnitude, even exceeding large available Grid storage space. Therefore, other approaches also have to be considered.

2. Virtual cubes: We have developed the concept virtual data cubes [24] distributed across a cluster of workstations or PCs. This approach corresponds to a *lazy* approach (also called *on demand* or *virtual*), i.e., OLAP queries are processed dynamically as they flow from the user to the data sources[7]. We are adapting this solution to the Grid.

By OLAP queries, various statistic information could be gained, which might be used by researchers or doctors as input for other queries, hypothesis testing, model development, etc.

[6] Naming conventions have already been proposed to describe some types of Grid datasets.

[7] This approach is similar to the *virtual data* idea applied in the GriPhyN project [14].

Caching. OLAP and data mining queries are typically repetitive and follow a predictable pattern. Therefore, efficient caching mechanisms have to be considered. Our caching methods extend the query caching mechanisms proposed by Deshpande et al. [9].

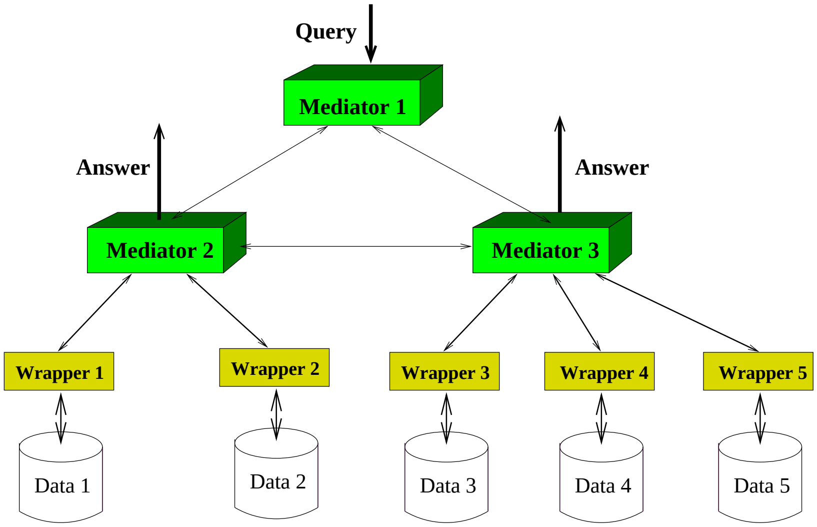

Fig. 4. Data Integration System - a distributed approach

5 Data Integration System

The wrapper-mediator framework [26] has become a standard architecture for information integration systems.

We are developing a mediation system distributed across Grid nodes as depicted in Fig. 4. In this system, data exchange and mediation relies on the XML standard. Our solution extends the approaches presented in [2], which address a traditional data integration architecture based on a centralized mediator. In Fig. 4, *Mediator 1* accepts and analyzes queries issued by OLAP and data mining engines and arranges that the query answers directly flow from *Mediator 2* and *Mediator 3* to the appropriate engines.

6 Grid Extensions

The TBI Knowledge Grid (TBI-KG) is organized in hierarchical layers, and is built on top of protocols and services provided by the Globus project (www.globus.org).

The new services are organized in two hierarchical layers: *Core Level TBI-KG Services* and *High Level TBI-KG Services*. The former has to support the definition, composition and execution of data mining and OLAP over the Grid. It performs the management of metadata describing characteristics of data sources, data mining, OLAP, and visualization tools and algorithms (*Knowledge Information Services*). Further, this layer includes services associated with data integration, cube computation and management, and cache management. Moreover, this layer coordinates data analysis executions, attempting to match the application requirements and the available Grid resources (*Resource Selection Service*).

The *High Level TBI-KG Services* allow to publish analysis tools and data sets, specify, develop, and execute data analysis over the TBI-KG, and present results of the analysis. The *Knowledge Source Discovery Services* are responsible for discovery, extraction and preprocessing of data sets to be analyzed, and *Tools and Algorithms Discovery Services* search and prepare tools and algorithms for data analysis execution; both of these service types are based on the Knowledge Information Services of the Core Layer.

The first TBI-KG prototype, we are implementing now is based on the Globus Toolkit 2.0 (www.globus.org) and will be reengineered to the Globus 3.0, which is based on the Open Grid Service Architecture [12], in the future, as we proposed in [3].

7 Related Work

McQuatt et al. [19] use decision tree techniques to predict the outcome of TBI patients. They only deal with sequential algorithms. There are already many publication on parallel and distributed data mining and OLAP (e.g., [1, 17, 18, 24]). A high-level description of an architecture for performing distributed data mining on the Grid was presented in [5], however, no data integration and OLAP aspects are considered there. R. Moore presents the concepts of knowledge-based Grids in [22]. Mahinthakumar et al. [10] report about the first clustering algorithm implementation on the Grid. A lot of valuable data integration concepts have been developed in the project “Federated Database for Neuroscience” [2]. So far, no research effort addressing the development of an OLAP-based data analysis infrastructure on the Grid has been reported.

8 Conclusions

In this paper we have described the project effort, which focuses on the application and extension of the Grid technology to a completely new and societally important category of applications. The architecture of the Knowledge Grid system which will support the management and treatment of traumatic brain injury victims was proposed and the detailed design of its components has been elaborated. The system is being implemented.

References

1. Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, A.: Fast discovery of association rules. In *U. Fayyad (ed.), Advances in Knowledge Discovery and Data Mining, AAAI Press*, pages 307–328, Menlo Park, CA, 1996.
2. Baru, C., et al.: XML-based information mediation with MIX. SIGMOD ’99, 1999.

3. Brezany, P., Hofer, J., Tjoa, A Min, Wöhrer, A.: Towards an Open Service Architecture for Data Mining on the Grid. Submitted to Dexa 2003, September 2003, Prague, Czech Republic.
4. Bullock, R., et al.: Guidelines for the management of severe head injury. Brain Trauma Foundation, New York, 1996.
5. Cannataro, M., Talia, D., Trunfio, P.: Knowledge grid: high performance knowledge discovery services on the grid. In *2nd Int. Workshop, Denver, USA*, pages 38–50, Nov. 2001.
6. Chen, Q.: Mining exceptions and quantitative association rules in olap data cube. Technical Report, Simon Fraser University,Canada, July 1999.
7. Codd, E.F.: Providing OLAP (on-line pnalytical processing) to user-analysts: An IT-mandate. Technical report. E.F. Codd and Associates, 1993.
8. Colliat, G.: OLAP, relational, and multi-dimensional database systems. *SIGMOD Record*, 25(3), September 1996.
9. Deshpande, P., Ramaswamy, M. K., Shukla, A., Naughton, J.F.: Caching multidimensional queries using chunks. In *ACM SIGMOD International Conference on Management of Data*, June 1998.
10. Mahinthakumar, G., et al.: Multivariate geographic clustering in a metacomputing environment using globus. In *Supercomputing'99*, Orlando, USA, November 1999.
11. Foster, I., Kesselman, C., Tuecke, S.: The anatomy of the Grid: Enabling scalable virtual organizations. Intl. J. Supercomputer Applications, 15(3), 2001.
12. Foster, I., Kesselman, C., Nick, J., Tuecke, S.: Open Grid Service Infrastructure, Global Grid Forum, June 22, 2002.
13. Goil, S., Choudhary, A.: High performance multidimensional analysis and data mining. In *Proc. SC98: High Performance Networking and Computing Conference (SC'98)*, Orlando, November 1998.
14. GriPhyN. http://www.griphyn.org.
15. Han, J., Kamber, M.: *Data Mining. Concepts and Techniques*. Morgan Kaufmann, 2000.
16. Jeffery, K. G.: GRIDs in ERCIM. ERCIM News, April 2001.
17. Kargupta, H., Park, B., Hershberger, D., Johnson, E.: Collective data mining: a new perspective toward distributed data mining. In *H. Kargupta and P. Chan (eds.) Advances in Distrib. and Parallel Knowledge Discovery (AAAI Press)*, 1999.
18. Kimm, H., Ryu, T.-W.: A framework for distributed knowledge discovery system over heterogeneous networks using CORBA. In *Proceedings of the KDD2000 Workshop on Distributed and Parallel Knowledge Discovery*, 2000.
19. McQuatt, A., Andrews, P.J.D., Aleeman, D., Corruble, V., Jones, P.A.: The analysis of head injury data using decision decision tree techniques. In *Horn, W., et al. (eds), Artificial Intelligence in Medicine, LNCS 1620, Springer-Verlag*, June 1999.
20. Microsoft. OLE DB for data mining. http://www.microsoft.com/data/oledb/ dm.htm.
21. Microsoft. OLE DB for OLAP. http://www.microsoft.com/data/oledb/olap/olap.htm.
22. Moore, R.: Knowledge-Based Grids. Technical Report TR-2001-02, San Diego Supercomputer Center, January 2001.
23. European DataGrid Project. http://www.cern.ch/grid/.
24. Rauber, A., Tjoa, A Min, Tomsich, P.: An architecture for modular On-Line Analytical Processing sytems. In *In A. M. Tjoa er. at., eds., 10th International Workshop on Databases and Expert Systems Applications (DEXA 1999)*, September 1999.
25. Stockinger, H.: Database replication in world-wide distributed data grids. Phd Thesis, University of Vienna, November 2001.
26. Wiederhold, G.: Mediators in the architecture of future information systems. The IEEE Computer Magazine, March 1992.
27. Zhao, Y., Deshpande, P., Naughton, J.: An array-based algorithm for simultaneous multi-dimensional aggregates. In *ACM SIGMOD International Conference on Management of Data*, pages 159–170, 1997.