Grid-Based Data Mining and the KNOWLEDGE GRID Framework

DOMENICO TALIA
(joint work with M. Cannataro, A. Congiusta, P. Trunfio)

DEIS
University of Calabria
ITALY
talia@deis.unical.it

Minneapolis, September 18, 2003
Data mining is often a compute intensive task.

When

- large data sets are coupled with
- geographic distribution of data, users, and systems,

it is necessary to combine different technologies for implementing high-performance distributed knowledge discovery systems (PDKD).

Distributed data mining tools are available but most of them do not run on Grids.
“By providing scalable, secure, high-performance mechanisms for discovering and negotiating access to remote resources, the Grid promises to make it possible for scientific collaborations to share resources on an unprecedented scale, and for geographically distributed groups to work together in ways that were previously impossible”

Ian Foster
Grid middleware targets technical challenges in areas such as communication, scheduling, security, information and data access, and fault detection.

Efforts are needed for the development of knowledge discovery tools and services on the Grid.

Grid-aware PDKD systems
PARALLEL & DISTRIBUTED DM ON GRIDS

The basic principles that motivate the architecture design of the grid-aware PDKD systems

- Data heterogeneity and large data size
- Algorithm integration and independence
- Grid awareness
- Openness
- Scalability
- Security and data privacy.
WHAT THE GRID OFFERS

- Grid infrastructure tools, such as the Globus Toolkit and Legion, provide basic services that can be effectively used in the development of a data mining applications.

- Data Grid middleware (e.g. Globus Data Grid) implements data management architectures based on two main services: storage system and metadata management.

- Data Grids are useful, but are not sufficient for data mining.
KNOWLEDGE GRID - a PDKD architecture that integrates data mining techniques and computational Grid resources.

In the KNOWLEDGE GRID architecture data mining tools are integrated with lower-level Grid mechanisms and services and exploit Data Grid services.

This approach benefits from "standard" Grid services and offers an open PDKD architecture that can be configured on top of generic Grid middleware.
A **KNOWLEDGE GRID** application uses:

- A set of **KNOWLEDGE GRID**-enabled computers - **K-GRID nodes**
  
  declaring their availability to participate to some PDKD computation, that are connected by

- A **Grid infrastructure**
  
  offering basic grid-services (authentication, data location, service level negotiation) and implementing the **KNOWLEDGE GRID** services.
KNOWLEDGE GRID ENVIRONMENT

Basic Grid Infrastructure

K-GRAIN tools
Grid Middleware
Cluster containing data sets and/or DM algorithms
LAN
Cluster Element
Cluster Element
Cluster Element

Grid Middleware
Local Resources

K-GRAIN node

K-GRAIN tools
Grid Middleware
Local Resources

Generic Grid node
The KNOWLEDGE GRID services are organized in two hierarchic layers:

- **Core K-Grid layer** and
- **High-level K-Grid layer**.

The former refers to services directly implemented on the top of generic Grid services.

The latter is used to describe, develop, and execute PDKD computations over the KNOWLEDGE GRID.
KNOWLEDGE GRID ARCHITECTURE

High level K-Grid layer

- DAS: Data Access Service
- TAAS: Tools and Algorithms Access Service
- EPMS: Execution Plan Management Service
- RPS: Result Presentation Service
- KDS: Knowledge Directory Service

Core K-Grid layer

- KMR: Knowledge Metadata
- Resource Metadata
- Execution Plan Metadata
- Model Metadata

Generic Grid Services
KNOWLEDGE GRID SERVICES

Core K-Grid layer services:

• Knowledge directory service (KDS). Extends the basic Globus MDS and GIS services to maintain a description of all data and tools used in the KNOWLEDGE GRID.

• Resource allocation and execution management service (RAEMS). RAEMS services are used to find a mapping between an execution plan and available resources.

• The Core K-Grid layer manages metadata describing features of data sources, third party data mining tools, data management, and data visualization tools and algorithms.
High-level K-grid layer services:

- **Data Access**
  - Search, selection (*Data search services*), extraction, transformation and delivery (*Data extraction services*) of data to be mined.

- **Tools and algorithms access**
  - Search, selection, and downloading of data mining tools and algorithms.

- **Execution Plan Management**
  - Generation of a set of different execution plans that satisfy user, data, and algorithms requirements and constraints.

- **Results presentation**
  - Specifies how to generate, present and visualize the PDKD results (rules, associations, models, classification, etc.).
KNOWLEDGE GRID OBJECTS

- We use the Globus MDS model only for generic Grid resources, but extended it with an XML metadata model to manage specific KNOWLEDGE GRID resources.

- Metadata describing relevant K-Grid objects, such as data sources and data mining tools, are implemented using both LDAP and XML.

- The (Knowledge Metadata Repository) KMR is implemented by LDAP entries and XML documents. The LDAP portion is used as a first point of access to more specific information represented by XML documents.
APPLICATION COMPOSITION STEPS

1. Search and selection of resources
2. Design of the PDKD computation
3. Execution Plan

Metadata about K-grid resources
Metadata about the selected K-grid resources
KEPR
TMR
KMRs

DAS / TAAS
EPMS
APPLICATION EXECUTION STEPS

RAEMS
- Execution Plan
- Execution Plan optimization and translation
- RSL script
- Execution of the PDKD computation
- Computation results
- Results presentation

GRAM

RPS

KEPR

KBR
A prototype version of the KNOWLEDGE GRID architecture have been implemented using Java and the Globus Toolkit 2.x.

To allow a user to build a grid-based data mining application, we developed a toolset named **VEGA (a Visual Environment for Grid Applications)**.

**VEGA** offers users support for:

- *task composition* - definition of the entities involved in the computation and specification of relations among them;
- *checking of the consistency* of the planned task;
- *generation of the execution plan* for a data mining task.
- *execution of the execution plan* through the resource allocation manager of the underlying grid.
VEGA: OBJECTS and LINKS

**Objects:**
- Hosts
- Software
- Data

**Links:**
- File Transfer
- Execute
- Input
- Output

*Objects* represent resources

*Links* represent relations among resources
VEGA

Hosts pane

Resources pane
A KGrid application can be composed of several workspaces.
<Software>
  <name>AutoClass</name>
  <description>Unsupervised Bayesian Classifier</description>
  <release>
    <number major="3" minor="3" patch="3"/>
    <date>01 May 00</date>
  </release>
  <author>Nasa Ames Research Center</author>
  <hostname>icarus.isi.cs.cnr.it</hostname>
  <executablePath>/share/software/autoclass-c/autoclass</executablePath>
  <manualPath>/share/software/autoclass-c/read-me.text</manualPath>
  ...
</Software>
XML EXECUTION PLAN

<ExecutionPlan>
 ...
 <Task ep:label="ws1_dt2">
  <DataTransfer>
   <Source ep:href="g1../Unidb.xml" ep:title="Unidb on g1.isi.cs.cnr.it"/>
   <Destination ep:href="k2../Unidb.xml" ep:title="Unidb on k2.deis.unical.it"/>
  ...
  </DataTransfer>
  </Task>
 ...
 <Task ep:label="ws2_c2">
  <Computation>
   <Program ep:href="k2../IMiner.xml" ep:title="IMiner on k2.deis.unical.it"/>
   <Input ep:href="k2../Unidb.xml" ep:title="Unidb on k2.deis.unical.it"/>
   ...
   <Output ep:href="k2../IMiner.out.xml" ep:title="IMiner.out on k2.deis.unical.it"/>
  </Computation>
  </Task>
 ...
 <TaskLink ep:from="ws1_dt2" ep:to="ws2_c2"/>
 ...
</ExecutionPlan>
A GENERATED RSL SCRIPT

+...
(&resourceManagerContact=g1.isi.cs.cnr.it)
(subjobStartType=strict-barrier)
(label=ws1_dt2)
(executable=$(GLOBUS_LOCATION)/bin/globus-url-copy)
(arguments=-vb -notpt gsiftp://g1.isi.cs.cnr.it/.../Unidb
         gsiftp://k2.deis.unical.it/.../Unidb
    )

...)

(&resourceManagerContact=k2.deis.unical.it)
(subjobStartType=strict-barrier)
(label=ws2_c2)
(executable=.../IMiner)
...

)...
APPLICATION EXECUTION
Some things we have done recently

**VEGA**:  
- Support for more complex computation layouts,
- Execution plan optimization,
- Abstract resources definition and use.

**KNOWLEDGE GRID**:  
- A peer-to-peer system for presence management and resource discovery on the Grid,
- A tool for optimized file transfer on the Grid based on GridFTP,
- A data mining ontology and an associated tool.
ON GOING WORK

OGSA and KNOWLEDGE DISCOVERY SERVICES

- The KNOWLEDGE GRID is an abstract service-based Grid architecture that does not limit the user in developing and using service-based knowledge discovery applications.

- We are defining a set of Grid Services that export functionalities and operations of the KNOWLEDGE GRID.

- Each of the KNOWLEDGE GRID services is exposed as a persistent service, using the OGSA conventions and mechanisms.

- We intend to offer those OGSA-Compliant services for implementing distributed Data Mining applications and Knowledge Discovery processes on Grids.
CONCLUSION

- Parallel and distributed data mining suites and computational grid technology are two critical elements of future high-performance computing environments for
  - e-science (data-intensive experiments)
  - e-business (on-line services)
  - virtual organizations support (virtual teams, virtual enterprises)

- Knowledge Grids will enable entirely new classes of advanced applications for dealing with the data deluge.

- The Grid is not yet another distributed computing system: it is a medium to dynamically share heterogeneous resources, services, and knowledge.
CONCLUSION

- Grids are coupling computation-oriented services with data-oriented services and knowledge-based services.

- This trend enlarges the Grid application scenario and offer new opportunities for high-level applications.

- **We are much more able to store data than to extract knowledge from it.**

- The **KNOWLEDGE GRID** is a framework for the unification of knowledge discovery and grid technologies helping us to climb some mountain of data.


www.icar.cnr.it/kgrid