

Grid-based Knowledge Discovery Services for High Throughput Informatics

M. Ghanem, Y. Guo, A. Rowe, P. Wendel
Imperial College of Science, Technology and Medicine
180 Queen's Gate, London
{mmg,yg,asr99,pjw4}@doc.ic.ac.uk

Abstract

Discovery Net is an application layer for providing grid-based knowledge discovery services. These services allow scientists to create and manage complex knowledge discovery workflows that integrate data and analysis routines provided as remote services. They also allow scientists to store, share and execute these workflows as well as publish them as new services. Discovery Net provides a higher level of abstraction of the Grid for knowledge discovery activities, thus separating the end-users from resource management issues already handled by existing and emerging standards.

1. Knowledge Discovery Services

Discovery Net provides the middleware for knowledge discovery services for a wide range of high throughput informatics applications including drug discovery, remote sensing and geo-hazard prediction. The data sets, and analysis routines, used in such applications are increasingly becoming available as remote services on the Internet. Examples include gene and protein databases and DNA sequence similarity searches in the case of life sciences applications, and satellite images, map servers and spatial analysis routines in the case of remote sensing applications.

Knowledge discovery procedures in all these applications typically require the creation and management of complex, dynamic, multi-step workflows. At each step, data from various sources is integrated and fed into an analysis routine. Based on the output results, the analyst chooses which other data sets, and analysis components, can be integrated in the workflow.

Discovery Net supports such activities by providing mechanisms and higher level services for representing, creating and managing knowledge discovery procedures and for composing existing data services and data analysis services in a structured manner, allowing scientists to plan, store, document, verify, share and re-execute their workflows as well as their output results.

Representing Knowledge Discovery Workflows:

Knowledge Discovery Procedures (workflows) are defined using an XML-based language, *DPML* (Discovery Process Markup Language). *DPML* represents a workflow as a data flow graph of nodes, each representing either a data analysis service or a data service. For each service, *DPML* records the service type, properties, parameters, and past parameter settings. This XML representation allows the workflows for discovery procedures to be easily stored, validated, shared and re-executed.

Discovery Net Components: The Discovery Net architecture provides three types of servers: Knowledge Discovery Look-up and Registration Servers, Meta-Information Servers and Knowledge Servers. *Knowledge Discovery Look-up and Registration Servers* allow the publication and retrieval of data analysis services and provide a store of service descriptors including functionality and input/output types. *Meta-Information Servers* provide services for data type management including type checking and data composition. *Knowledge Servers* provide services for warehousing knowledge discovery workflows, generating reports from them and application generation services allowing users to deploy their own workflows as new services.

Discovery Net Prototype: The design and development of Discovery Net is funded under the UK government's core programme for e-science. The data analysis components and visual programming environment used in the prototype implementation are based on the tools provided by the Kensington distributed data mining system [1]. The lower-level grid-based resource management activities will be based on the emerging OGSA standard [2].

References

- [1] J. Chattratichat, J. Darlington, Y. Guo, and S. Hedvall. An Architecture for Distributed Enterprise Data Mining. *Lecture Notes in Computer Science*, 1593:573–582, 1999.
- [2] I. Foster, C. Kesselman, J. M. Nick, and S. Tuecke. The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration. Technical report, <http://www.globus.org/research/papers/ogsa.pdf>, 2002.