Why Grid-based Data Mining Matters? Fighting Natural Disasters on the Grid: From SARS to Land Slides

A.K.T.P. Au, V. Curcin, M. M. Ghanem, N. Giannadakis, **Y. Guo**, M. A. Jafri, M. Osmond,

A. Oleynikov, A.S. Rowe, J. Syed, P. Wendel and Y. Zhang

Department of Computing, Imperial College London,

180 Queens Gate, London, SW7 2AZ

{aktp, vc100, mmg, ng300, yg, jafri, mo197, aio00, asr99, jas5, pjw4, yzhan}@doc.ic.ac.uk

Abstract

The Discovery Net UK e-Science project has built a framework and infrastructure for knowledge discovery services over data collected from high throughput sensors. In this paper we provide an overview of the Discovery Net approach and highlight some of the scientific applications constructed by end-user scientists using the Discovery Net system. These applications include genome annotation, the analysis of SARS evolution patterns, monitoring air pollution data and the analysis of earthquake and land slide satellite images.

1. The challenge of discovering new knowledge

In their simplest definition, e-Science platforms are Internet-enabled working environments allowing distributed scientists to form a virtual organization where they can share data and computing resources and collectively collaborate on the analysis of the data to derive new knowledge.

The vision of e-Science platforms, which are common in the UK and Europe, is closely related to the vision of computational grids in the US. However, current research into fundamental Grid technologies, such as Globus [1], has concentrated mainly on the provision of protocols, services and tools for creating co-ordinated, transparent and secure globally accessible computational systems. These technologies follow a service methodology for finding both computation and data services for performing computationally or data intensive tasks. The delivery of the low-level infrastructure is essential but does not provide end users with the easy-to-use tools that aid them in the creation of their scientific applications.

Compared to Grid computing platforms, e-Science platforms concentrate mainly on the provision of higher-level application-oriented platforms that are focused on enabling the end-user scientists in deriving new knowledge when devices, sensors, databases, analysis components and computational resources are all accessible over the Internet or the Grid. The Discovery Net system is an example of such e-Science platforms that are dedicated to empowering end users in conducting knowledge discovery activities, easily and seamlessly. The system is currently used by a number of application groups in different fields including life science, environmental monitoring and geo-hazard modeling.

In the remainder of this paper we describe our experience from building the Discovery Net platform for grid-based knowledge discovery and data mining, and using it in conducting data mining over scientific experimental data.

2. Knowledge Discovery in e-Science

e-Science concerns the development of new practices and methods to find knowledge. Non-trivial, actionable knowledge cannot be batch generated by a set of predefined methods, but

rather the creativity and expertise of the scientist is necessary to formulate new approaches. Whilst the dynamic nature of massively distributed service-oriented architectures provides much promise in providing scientists with powerful tools, it raises many issues of complexity.

New resources such as online data sources, algorithms and methods defined as processes are becoming available daily. A single process may need to integrate techniques from a range of disciplines such as data mining, text mining, image mining, bioinformatics, or chemoinformatics, and may be created by a multidisciplinary team of experts. A major challenge is to effectively coordinate these resources in a discovery environment in order to create knowledge.

As examples of e-Science data analysis processes, consider the following scenarios:

a. Scientists collaborating on the analysis of a newly discovered viral genome such as SARS and studying its evolution.

b. Scientists collaborating on the analysis of environmental air pollution data and correlating it with available medical records and traffic data.

c. Scientists collaborating on the analysis of satellite images for modelling the possible effects of earthquakes on populated regions.

In each of the above scenarios a scientific knowledge discovery process conducted in an open environment proceeds by making use of distributed data and resources. The main features of such processes can be summarised as:

1. The processes typically operate as data and application integration pipelines. At different stages of the knowledge discovery process, researchers need to access, integrate and analyse data from disparate sources, in order to use that data to find patterns and models, and feed these models to further stages in the process. At each stage, new analysis is conducted by dynamically combining new data with previously developed models.

2. There are typically many different data analysis software components that can be used to analyse the data. Such software components may be on the user's local machine, while others may be tied for execution on remote servers, e.g. via a web-service interface or even simply via a web page interface. New software components, services and tools are continually being made available, either as downloadable code or as remote services over the Internet for access by various groups. An individual researcher needs to be able to locate such software components and integrate them within their analysis procedures.

3. The discovery process itself is almost always conducted by teams of collaborating researchers who need to share the data sets, the results derived from these data sets, and, more importantly, details about how these results were derived. In this case, recording an audit trail of how a particular analysis result (or new knowledge) was acquired and used is essential since it allows researchers to document and manage their discovery procedures.

4. Since the whole discovery process is executable, the end user may want to wrap it as an executable program (or software component) for access and use by other researchers. In this case it is essential to provide methods that allow such processes to be automatically converted into executable code, and that allow information about them to be published to allow users to locate and access such code.

5. Finally, with a large number of discovery processes being generated by different research groups, it is essential to be able to store such processes within a process warehouse, from which scientists can search, retrieve and re-use procedures developed from one scenario in similar scenarios. Furthermore, the availability of such a warehouse will help them in managing intellectual property activities such as patent applications, peer reviews and publications.

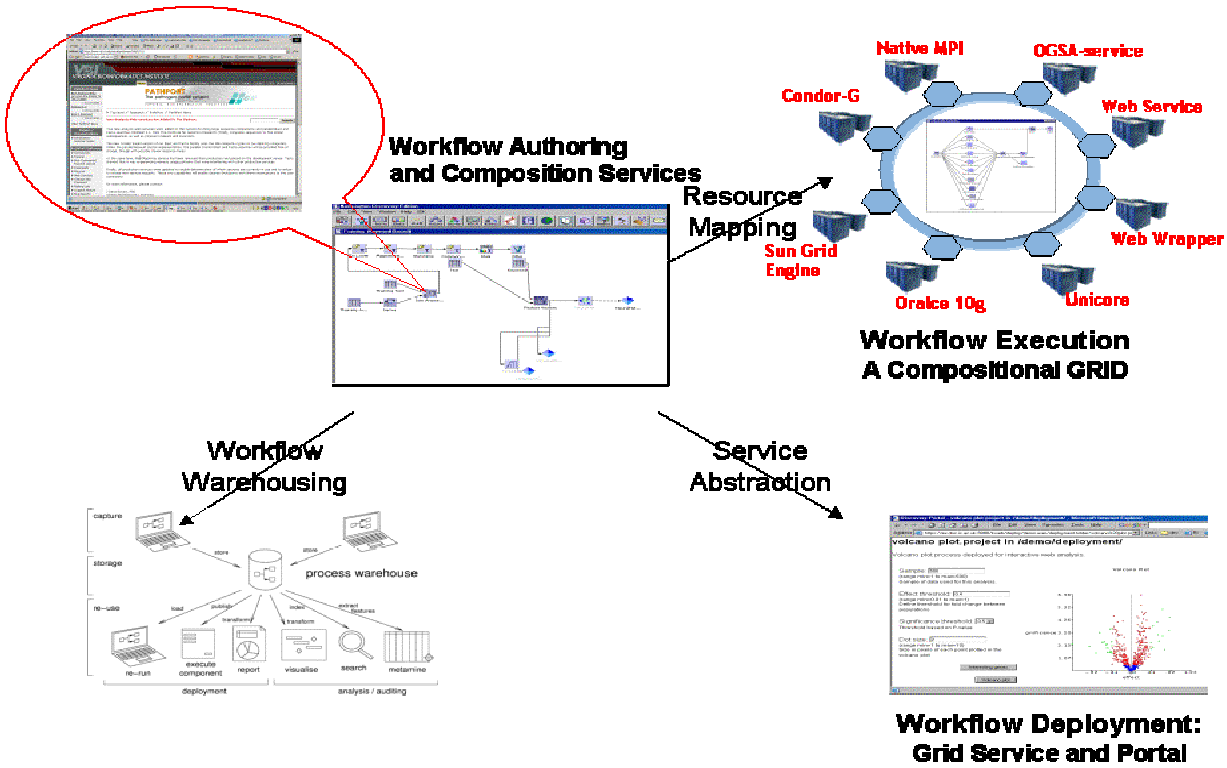

Figure1: Overview of Workflow Services within Discovery Net

3. Knowledge Discovery in Discovery Net

The requirements of e-Science data mining environments make it impractical to use traditional `closed' data mining systems that assume a centralised database or a data warehouse where all the data required for an analysis task can be materialised locally at any time, before feeding them to data mining algorithms and tools that themselves have been predefined at the configuration stage of the tool.

The Discovery Net system is designed primarily to support analysis of scientific data based on a workflow or pipeline methodology. In this framework, services can be treated as black boxes with known input and output interfaces. Services are then connected together into a sequence or operations. The Discovery Net architecture provides a platform for open data mining allowing the integration of distributed data sources and distributed tools in knowledge discovery activities. The approach and its components are shown in Figure 1 and described briefly below

1. Users construct their data analysis workflows using the Discovery Net workflow authoring tools, which allow them to dynamically browse for, and integrate data services and analysis services and compose them dynamically as workflows. These services are accessed through a variety of access interfaces and protocols that are supported by the system, including native interfaces as well as web service interfaces. The workflows themselves are represented and stored using DPML [4] (Discovery Process Markup Language), an xml-based representation of the workflows. Each workflow is represented as a data flow graph of nodes, each representing a service. DPML is typically authored using a visual programming interface on a Discovery Net client, and each node descriptor contains information about three aspects: the service parameters, the service history within the context of the workflow (changes to parameter settings, user information, etc) and user added comments. A process created in DPML is reusable and can then be encapsulated and shared as a new service on the Grid for other scientists.
2. The execution of the composed workflows is delegated to a workflow execution engine, which handles resource allocation and data movement between different computing resources using a wide variety of methods [2]. This is achieved both through co-ordination between different services to achieve distributed execution across servers and also within each service

where distributed/parallel execution is handled by other engines (e.g. MPI over parallel machines, Condor over workstation cluster, etc).

3. In addition, the Discovery Net InfoGrid infrastructure [5] is to used provides the users with ability to dynamically access and integrate various heterogeneous data sets in their analysis workflows. This is achieved through various methods including providing interfaces to SQL databases, OGSA-DAI sources, and Oracle databases, as well as through the provision of specialised wrappers to a wide variety of web-based data sources.
4. The Discovery Net Deployment Tool allows users to abstract their workflows and encapsulate them as new services that can be published using either a portal interface for direct access by end users or through a web-service interface for programmatic access by other services. The tool allows users to specify which parameters of the services within a workflow can be changed by the end user, and also to specify the action points within a workflow.
5. Finally, Discovery Net provides a workflow warehouse [4] that acts as a repository for the user workflows. A wide range of query and analysis methods are available allowing users to query the warehouse to retrieve workflows relevant to their analysis, and also to perform meta-analysis over the workflows.

4. Discovery Net Applications

End user scientists are currently using the Discovery Net infrastructure in three different applications domains: Life Sciences, Environmental Monitoring and Geo-hazard Modelling. In this section, we provide a brief introduction to four end user applications in these areas, and describe briefly how they benefit from the Discovery Net infrastructure.

4.1 Distributed Genome Annotation

An early example of a complete Life Science applications conducted using the Discovery Net infrastructure is real-time genome annotation described in [6]. The genome annotation application is highly data and computer intensive and requires the integration of a large number of data sets and tools that are distributed across the Internet. Furthermore, it is a collaborative application where a large number of distributed scientists need to share data sets and interactively interpret and share the analysis results.

The first prototype of the application was constructed in 2002 where some of the analysis components were integrated as Web Services, some as local processes, some as Grid services. The creation of these services took less than three man weeks at the time. The Discovery Net client was then used to combine the distributed tools together, and to visualize and interpret the results. The application was presented at the IEEE SC2002 Supercomputing conference in Baltimore. The annotation pipelines were running on a variety of distributed computing resources including: high performance resources hosted at the London e-Science centre, servers at Baltimore, and databases distributed around Europe and the USA. The application was combined with a Real Time DNA sequencing platform provided by DeltaDot Ltd to produce a real time genome sequence and annotation pipeline. This application was subsequently awarded the “Most Innovative Data Intensive Application” at the conference’s High Performance Computing Challenge at the conference.

4.2 SARS Evolution Analysis

Another example of Life Science applications conducted within Discovery Net is the analysis of the evolution of the SARS virus [7] for establishing the relationship between observed genomic variations in strains taken from different patients, and the biology of SARS virus. The Discovery Net team, in collaboration with researchers from SCBIT [8] (Shanghai Centre for Bioinformation Technology), was actively involved in this research and the platform was used to capture the analysis based upon a service-based computing infrastructure [9] and to serve as framework for further investigations over the collected data. The application requires the integration of a large number of data sets and tools that are distributed across the Internet. It also requires the collaboration of distributed scientists and requires interactivity in the analysis of the data and in the interpretation of the generated results.

As opposed to the early genome annotation prototype described above, the process of integrating remote services within SARS analysis workflows had been mostly automated for this application and performed on-the-fly, taking on average 5 minutes per tool for adding

the components to the servers at runtime, thus increasing the productivity of the scientists at an impressive rate. Furthermore, by using the Discovery Net deployment tools, the SARS analysis workflows have been placed within a SARS portal environment that will be updated as new data sets and analytical techniques become available, providing the scientists with a flexible way to access and re-execute their analysis.

4.3 Urban Air Pollution Monitoring

The Discovery Net infrastructure is currently being used as knowledge discovery environment for the analysis of air pollution data within the context of the GUSTO project [10]. The aim of is to provide an infrastructure that can be used by scientists to study and understand the effects of pollutants such as Benzene, SO_2, NO_x or Ozone on human health. The more this relationship is understood, the better chance there is of controlling and ultimately minimising such effects. In order to achieve this, pollutant concentrations must be monitored accurately and ideally *in situ* so that sources may be identified quickly and the atmospheric dynamics of the process is understood.

Using the Discovery Net infrastructure, the scientists are able to construct a sensor grid that is characterized by the physical distribution of the sensors themselves, the vast amounts of data generated by each sensor (up to 8GB/day), the ability to access and use high performance computing resources, the ability to use a wide variety of distributed data analysis and mining components, and by the ability to correlate the analysis of the collected data with other distributed data sources (e.g. traffic data, weather data and health data).

4.4 Geo-hazard Modelling in Discovery Net

The Discovery Net infrastructure is also used by remote sensing scientists [11] to analyse co-seismic shifts of earthquakes using cross-event Landsat-7 ETM+ images. This application is mainly characterized by the high computational demands for the image mining algorithms used to analyse the satellite images (execution times for simple analysis to analyse a pair of images takes up to 12 hours on 24 fast Unix processors). In addition, the requirement to construct and experiment with various algorithms and parameter settings has meant that the provenance of the workflows and their parameter settings becomes an important aspect to the end user scientists.

Using the system, the remote sensing scientists analysed data from an Ms 8.1 earthquake that occurred in 14 Nov 2001 in a vast uninhabitable area along the eastern Kunlun Mountains in China. The scientific results of their study provided the first ever 2-D measurement of the regional movement of this earthquake and revealed stunning patterns that were never studied before on the co-seismic left-lateral displacement along the Kunlun fault in the range of 1.5-8.1 m.

5. Conclusions and Discussion

The accessibility of data and compute resources to scientists enables them to conduct more complex analyses over larger data sets in less time. It allows them to become more efficient in what they do best – discovering new knowledge. The availability of the end-user oriented frameworks that provide these scientists with full access to the benefits of grid commuting technologies while shielding them from the complexities of the underlying protocols, is essential.

Our experience indicates that achieving the balance is indeed possible by providing the scientists with tools at the appropriate level of abstraction that suits their problem solving methods, and that suits their modes of scientific discovery.

Our experience from our end user case studies is positive: e-Science can indeed be used effectively and easily to develop important knowledge discovery applications that save lives.

References

1. Foster I. and Kesselma C. Globus: a Metacomputing Infrastructure Toolkit. International. *Journal of Supercomputer Applications*, 11 (2), 115–128.

2. AlSairafi S., Emmanouil F. S, Ghanem M, Giannadakis N, Guo Y, Kalaitzopoulos G, Osmond M, Rowe A, Syed J and Wendel P. The Design of Discovery Net: Towards Open Grid Services for Knowledge Discovery. Special issue of The *International Journal on High Performance Computing Applications* on Grid Computing: Infrastructure and Applications, Vol. 17 Issue 3. 2003

3. Curcin V, Ghanem M, Guo Y, Kohler M, Rowe A, and Wendel, P. Discovery Net: Towards a Grid of Knowledge Discovery. In Proceedings of KDD-2002, the *Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. July 23-26, 2002 Edmonton, Canada.

4. Syed J, Guo Y and Ghanem M. Discovery Processes: Representation And Re-Use, *UK e-Science All Hands Meeting*, Sheffield UK, September, 2002.

5. Giannadakis N, Rowe A, Ghanem M and Guo Y. InfoGrid: Providing Information Integration for Knowledge Discovery. *Information Science*, 2003: 3:199-226.

6. Rowe A, Kalaitzopoulos D, Osmond M, Ghanem M and Guo Y. The Discovery Net System for High Throughput Bioinformatics. In *Proceedings of the Eleventh International Conference on Intelligent Systems for Molecular Biology*, 2003. Also appears in *ISMB (Supplement of Bioinformatics)* 2003: 225-231

7. The Chinese SARS Molecular Epidemiology Consortium. Molecular Evolution of the SARS Coronavirus During the Course of the SARS Epidemic in China. *Science*, Vol. 303, Issue 5664, 1666-1669, 12 March 2004.

8. SCBIT, http://www.scbit.org.

9. Curcin V, Ghanem M and Guo Y. SARS Analysis on the Grid. *UK e-Science All Hands Meeting*, Nottingham UK, September 2004.

10. Ghanem M, Guo Y, Hassard J, Osmond M and Richards R. Sensor Grids for Air Pollution Monitoring *UK e-Science All Hands Meeting*, Nottingham UK, September 2004.

11. Liu J. G and Ma J. Imageodesy on MPI & GRID for Co-seismic Shift Study Using Satellite Optical Imagery. *UK e-Science All Hands Meeting*, Nottingham UK, September 2004.