

Development of Agent-Based Knowledge Discovery Framework to Access Data Resource Grid

¹Sanskruti Patel, ²Tejas Thakkar, ³Priya Swaminarayan and ⁴Priti Sajja

^{1,2,3}Lecturer, Computer Science Department, ISTAR, V V Nagar – 388120

²Lecturer, Computer Science Department, NVPAS, V V Nagar – 388120

³Reader, Computer Science Department, Sardar Patel Univerisity,
V V Nagar - 388120

E-mail: ¹sanskrutipatel@rediffmail.com, ²ththakkar@gmail.com,

³Priya_swaminarayan@yahoo.co.in and ⁴Priti_sajja@yahoo.com

Abstract

Knowledge discovery from heterogeneous data sources available on Data Grid environment is a challenging research and development issue. This paper covers all aspects of the knowledge discovery process and integrates this process with service-oriented grid application supported by agent framework for university domain. The GGF (Global Grid Forum), Open Grid Services Architecture (OGSA), and its associated specifications define consistent interfaces through web services to components of a grid infrastructure. The architecture presented in this paper, shows how different heterogeneous data resources in university domain are integrated via OGSA-DAI framework and how knowledge discovery process is performed in distributed data resource environment by using multiple task agents.

Keywords: Knowledge discovery, Database grid, OGSA-DAI, Data resources.

Introduction

The need for computer support is now a days necessary in education system where students and teachers interact with software components rather than with individual. Also, Grid technology is more and more useful in scientific as well as in academic environments. With the increasing focus on grid development, there is a need for proper abstractions for modeling grid applications. The presented data grid infrastructure facilitates a logical view of heterogeneous distributed resources that are shared between autonomous administrative domains. However, use of existing grid

computing environment is quite complex and requires a lot of human intervention. To avoid this intervention, enhancements are required in existing grid infrastructure. Viewed from a distributed AI perspective the most suitable abstraction is the concept of agents [1]. An agent is a software program that automatically performs tasks on behalf of the user [4]. Specific agent is defined to do particular task. The different task agents are performed knowledge discovery process on distributed environment via OGSA-DAI architecture and provide different services to the users of the university domain like searching, advisory and help desk system, course management and resource management, student and faculty communication and so on. Through knowledge-based multi-agent architecture, content is turned into something more than just a collection of data. i.e. understanding the context, format, and significance of the data called knowledge discovery.

Different agents providing additional facilities help applications and users to make better decisions about how to deal with the data [3].

Knowledge & Knowledge Management

Knowledge is as (i) expertise, and skills acquired by a person through experience or education; the theoretical or practical understanding of a subject, (ii) what is known in a particular field or in total; facts and information or (iii) awareness or familiarity gained by experience of a fact or situation. Philosophical debates in general start with Plato's formulation of knowledge as "justified true belief". Knowledge acquisition involves complex cognitive processes: perception, learning, communication, association and reasoning. The term knowledge is also used to mean the confident understanding of a subject with the ability to use it for a specific purpose if appropriate. Knowledge Management is the collection of processes that govern the creation, dissemination, and utilization of knowledge. These processes exist whether we acknowledge them or not and they have a profound effect on the decisions we make and the actions we take, both of which are enabled by knowledge of some type. Knowledge management is not a, "a technology thing" or a, "computer thing" If we accept the premise that knowledge management is concerned with the entire process of discovery and creation of knowledge, dissemination of knowledge, and the utilization of knowledge then we are strongly driven to accept that knowledge management is much more than a "technology thing" and that elements of it exist in each of our jobs.

Knowledge Discovery in Databases (KDD)

Knowledge discovery is defined as "the non-trivial extraction of implicit, unknown, and potentially useful information from data". Knowledge discovery describes the process of automatically searching large volumes of data for patterns that can be considered knowledge about the data. It is often described as deriving knowledge from the input data. This complex topic can be categorized according to 1) what kind of data is searched; and 2) in what form is the result of the search represented. Knowledge Discovery refers to the overall process of generating patterns from data.

The following steps are necessary for this. These steps are known as the "knowledge discovery cycle". They are: Acquisition, Representation, Storage, Analysis, Visualization, Interpretation and Deployment.

Traditional centralized knowledge discovery in databases (KDD) is reaching its limits due the fact of increasingly high volume of data, big calculation tasks and the geographical distribution of the involved/required (data) resources. The OGSA-DAI based KDD provides services used to extract knowledge from the data stored inside the heterogeneous database Grid. These services will be used both to build high-level knowledge discovery applications, as in the case of the Knowledge Grid, and to enhance existing basic Grid services.

Introduction to OGSA-DAI

Different data providers are providing Web Services to retrieve data from their databases but these web services are not dynamic or flexible enough to adopt the ever changing user needs. If an application requires huge amount of computing and storage resources which could not met by local cluster, the only choice is to use grid computing. In such a case, distributing and maintaining the heterogeneous data and retrieving it in secure mode becomes very important. OGSA-DAI could fill this role comfortably [5]. The Global Grid Forum (GGF) is a community initiated forum of individual researchers and practitioners working on distributed computing, or "grid" technologies (<http://www.ggf.org>). The Open Grid Services Architecture (OGSA) is a product of the Grid community at large, and it has a major focal point in GGF which is divided into several work groups dealing with different areas like applications and basic services [2]. Database Access and Integration Services Working Group (DAIS) is belonging to the field of data, and its main research content is how to apply the database into Grid. OGSA-DAI is a middleware product that allows data resources, such as relational or XML databases, to be accessed via web services [6]. The prime goals of OGSA-DAI are following [7]:

- To provide controlled exposure of physical data resources to the Grid.
- To support access to heterogeneous physical data resources through a common interface style while employing the underlying query mechanisms.
- To provide base services that allow higher-level data integration services to be constructed, e.g. distributed query processing and data federation services.
- To leverage emerging Grid infrastructure for security, management, accounting etc.
- To standardized data access interfaces through the GGF DAIS WG.
- To provide a reference implementation of the DAIS specification.

It is also provided following facilities: Data access - accessing structured data in distributed heterogeneous data resources; Data transformation - for example, exposing data in schema X to users as data in schema Y; Data integration - for example, exposing multiple databases to users as a single virtual database; Data delivery - delivering data to where it's needed by the most appropriate means: web service, e-mail, HTTP, FTP, GridFTP. OGSA-DAI supports relational databases like Oracle, DB2, SQL server, MySQL, and Postgres, and XML databases like xindice and eXist,

and files like CSV, EMBL, and SwissProt files. Databases are deployed as data service resources, which contain all the information about the databases like their physical location and ports, the JDBC drivers that are required to access the databases, and the user access rights [5]. With OGSA-DAI, Querying, updating, transforming, and delivering data is done via web services. Also, uniform request is made to OGSA-DAI web services irrespective of the exposed databases. It supports integration of data from various data resources.

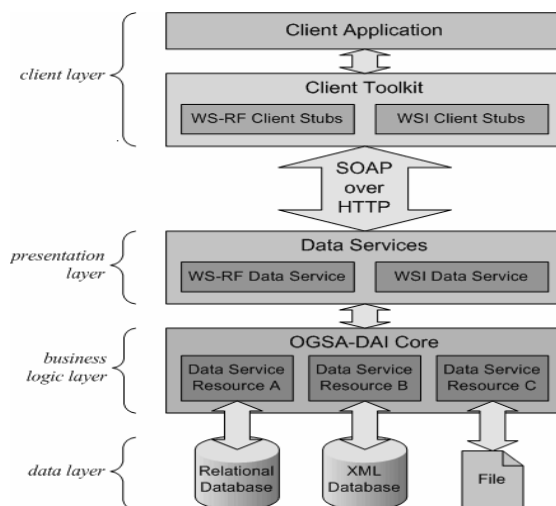


Figure 1: Architecture of OGSA-DAI.

Architecture of OGSA-DAI

The architecture of OGSA-DAI consists of a number of layers each serving a different purpose. A high-level schematic representation of this architecture is shown in Figure-1[8]. The data layer consists of the data resources that can be exposed via OGSA-DAI like Relational databases such as MySQL, SQL Server, DB2, Oracle, PostgreSQL., XML databases such as eXist, Xindice, Files and directories in formats such as OMIM, SWISSPROT and EMBL. This interface allows information to be communicated between the data layer and the business logic layer in both directions. It is realized by components known as data resource accessors. This layer encapsulates the core functionality of OGSA-DAI. This interface allows information to be communicated between the business logic layer and the presentation layer in both directions. It supports the invocation of OGSA-DAI functionality within the business logic layer in a way that is independent of a particular web environment. This layer encapsulates the functionality required to expose data service resources using web service interfaces. A client can interact with a data service resource via a corresponding data service. Depending on whether a WSRF or WSI data service has been deployed, the client application must be compliant with the WSRF or WSI standards. OGSA-DAI also include a Java Client Toolkit which provides a higher-level API for interacting with data services [8].

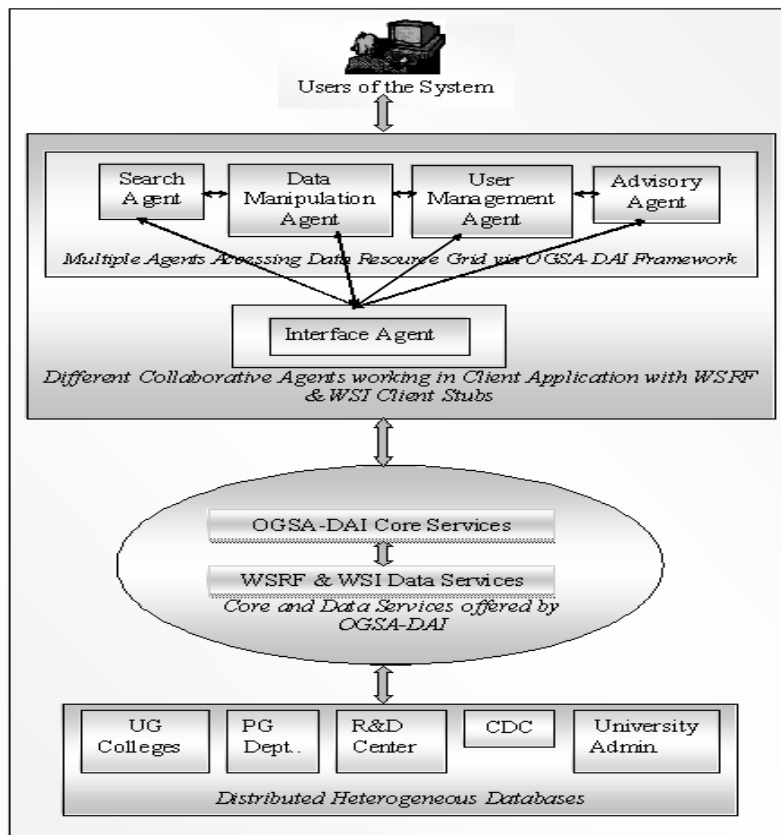


Figure 2: Architecture of Knowledge Discovery Framework with OGSA-DAI.

A Detailed Methodology of Prototype System

At present, most academic systems use single computers or servers. This research is aimed at applying Database grid approach to integrate idle computer resources in academic institutions. Also, it discovers a seamless information processing system across distributed, heterogeneous, dynamic, open organization that the user can access from any location. Besides giving advantages like effective retrieval and presentation of information, identifying user need, evaluation of user responses, and providing advisory, help and searching facilities, these results enable academic institutions like universities having limited infrastructure can implement this system to access vast distributed resources by using this solution. According to the various functionalities of university domain, we constructed the structure of our prototype model which has several metamodels including UG colleges, PG departments, R&D centers, Career Development Centers, etc. Take the UG College meta-model for example, which has a set of common attributes: {Course Type, Course Name, Fees, Duration, Scholarship, Placement and so on}. The figure 2 shows the detailed methodology of the prototype system. In one university domain, there can be various components are presented like UG colleges, PG departments, Research & Development center, Career Development center and university administration. The database of these all components is not

stored centrally. Rather, each component has its own location to store the database. So, in such scenario, the data resources are stored on different locations and they can be heterogeneous or homogenous. We are proposing OGSA-DAI framework for deployment a database grid environment to integrate all the distributed heterogeneous or homogeneous databases for university domain. The data layer integrates all the different data resources. The Business logic layer and presentation layer is provided by OGSA-DAI core and data services. On the client layer, there are two types of agents: interface agent, which interacts with individual user; task agents, which are responsible to carried out different problem-solving tasks and produces different functionalities. Users first have to contact interface facilitator in order to contact task agents. Different task agents facilitate the user by providing various facilities like search, advisory, user management, data manipulation, etc.

Knowledge Discovery Process in Distributed Data Resource Grid Environment for University Domain

The following figure 4 shows the example of knowledge discovery process. In one university, there can be several UG Colleges, PG department, R & D Center and Career Development Center. In the prototype system, four distributed databases are included with their sample entities and attributes. These databases are stored in distributed environments. For example, a user who is a student needs to have an advice and help to find an optimum choice for the course for getting an admission. There are several courses conducted by university but the information about them is stored in colleges' own databases which are located in distributed environment. So, generally a student has to access several databases of the different colleges to get the proper idea about the course which is suitable to him. But, by using the university portal (our prototype system), there is no need to access several databases individually because these all databases are now connected and formed data resource grid. He has to just pass out his choice and criteria's to an interface agent. These interface agent than choose the particular task agent (here it is an advisory agent) to do specific task. This specific task agent then accesses the distributed databases via OGSA-DAI environment and submits the result to the interface agent. Finally, the interface agent presents this result to the students for helping in his decision making process. The figure 3 shows only the sample databases with limited entitles and attributes and explains the workflow of the knowledge discovery process in distributed grid environment. In order to communicate with one another, a specific agent remains idle while no messages arrive. When a specific message is arrived from any other agent, it interprets the message and act accordingly by using a message queuing mechanism. All agents in the system work on this principle and remain idle while no messages arrived. Also, each agent can send a message to any other agent meanwhile. Web services are used implement the different task agents. Separate web services can be created to justify task of particular agent. Also, these web services can be crated using any language platform which is supported the creation of web services.

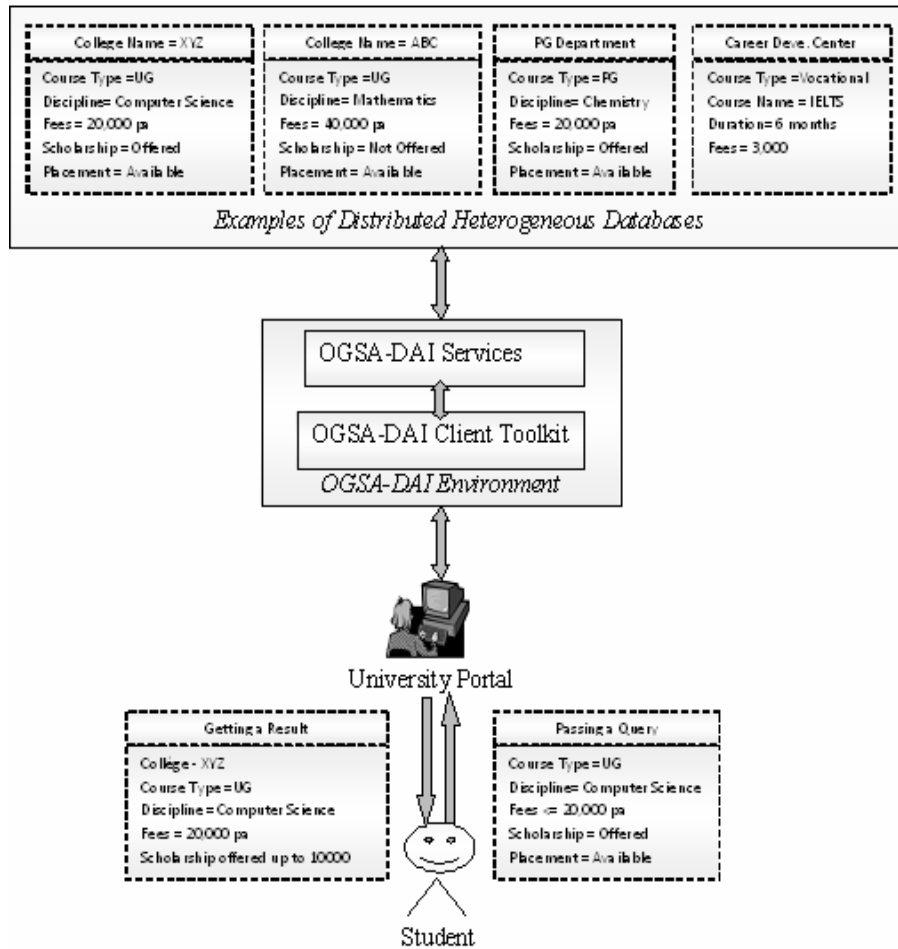


Figure 3: Knowledge Discovery in Distributed Data Resource Grid Environment

To run OGSA-DAI, we need to install java 1.4 or 1.5, Jakarta tomcat 5.5 or 5.0, and apache 1.6 or above. The OGSA-DAI client toolkit consists a JAVA API that provides the basic building blocks for OGSA-DAI client development. Using these APIs, a developer can construct anything from a basic query client to a complex distributed data integration client with relative ease. The Client Toolkit minimizes the specialist knowledge required to interact with OGSA-DAI services and provides some protection from future changes to the data service interfaces. It also operates transparently with OGSA-DAI data services compliant with both WS-I and WSRF. The main concept of the Client Toolkit is that of an activity. An activity dictates an action that is to be performed by a data service resource. OGSA-DAI provide many different types of activity to perform operations such as SQL queries, XSL transformations and FTP data delivery. The Client Toolkit provides a simple mechanism for assembling requests describing multiple activities, then submitting the request to data service resources for processing. To deploy a new activity onto an

OGSA-DAI server do the following. If the activity class is not already in a JAR on the server then we should create a temporary directory containing this JAR [9]. For example:

```
$ mkdir tmpActivityDir
$ cp my-activity.jar tmpActivityDir
```

Now, to deploy the new activity run the `deployActivity` command this takes the following arguments: [1] `dai.activity.jar.dir` - location of directory containing the activity implementation class. If the activity implementation class is already on the server then this value can be omitted. [2] `dai.activity.id` - activity ID. Unique ID by which the activity is known on the server. [3] `dai.activity.class` - name of activity implementation class. [4] `dai.activity.description` - human-readable description of the activity. This value is optional.

Then after, if using OGSA-DAI Axis on Tomcat run by following commands:

```
$ ant -Dtomcat.dir=/PATH/TO/TOMCAT \
-Ddai.activity.jar.dir=ACTIVITY-JAR-DIR -Ddai.activity.id=ACTIVITY-ID \
-Ddai.activity.class=ACTIVITY-CLASS \ -Ddai.activity.description=ACTIVITY-
DESCRIPTION deployActivity
```

This takes the following arguments:

`tomcat.dir` - location of OGSA-DAI Axis deployment on Tomcat

For example:

```
$ ant -Dtomcat.dir=/home/user/tomcatAxis \
-Ddai.activity.jar.dir=tmpActivityDir -Ddai.activity.id=MyActivityID \
-Ddai.activity.class=org.my.ActivityClass \
-Ddai.activity.description="This is my new activity" deployActivity
```

Also, the OGSA-DAI can be directly implemented by the Globus Toolkit, which is an open source software toolkit used for building Grid systems and applications. The Globus Alliance and many others all over the world are developing it. Globus allows people to share computing power, databases, and other tools securely online across corporate, institutions, and across geographic boundaries.

Conclusion

In this paper, we have proposed Agents Based Knowledge Discovery Framework Accessing Data Resource Grid for University Domain that can be used to provide generalized and flexible solutions applicable to all academic units. In education environment, institutions, colleges or schools may not have enough budgets to upgrade their resources like storage capacities. Grid computing offers them resource sharing in heterogeneous and graphically distributed environment. The proposed prototype system discovers a seamless information processing system across distributed, heterogeneous, dynamic, open organization that the user can access from any location. Besides giving advantages like effective retrieval and presentation of information, identifying user need, evaluation of user responses, and providing advisory, help and searching facilities, here we proposed a new academic and learning environment by integrating agents with OGSA-DAI and Data Grid environment. As future enhancement, we can use this system by adding University Ontology in the

system to generate the Knowledge Grid. The combined use of Knowledge Grid, Ontologies and Multi-agent technologies will enable the sharing of heterogeneous, autonomous knowledge sources in a capable, adaptable and extensible manner.

References

- [1] Amund Tveit (2002), jfipa - An Architecture for Agent-based Grid Computing. Retrieved 25 February 2008 from <http://citeseer.ist.psu.edu/tveit02jfipa.html>.
- [2] Bing Chen, Xuchu Weng, Changle Zhou, Xueqin Hu, Knowledge Grid Application in the Digital TCM Hospital System, Proceedings of 2008 IEEE International Symposium on IT in Medicine and Education
- [3] Martin C. Brown (2005), Freelance writer and consultant, MCslp. Retrieved 7 March 2008 from <http://www.ibm.com/developerworks/grid/library/gr-semgrid/main>
- [4] S. genesereth M. and Ketchpe (1994), Software agents, Comm. of ACM. Vol. 37, No. 7, pp. 48-53.
- [5] Samatha Kottha, Kumar Abhinav, Ralph Müller-Pfefferkorn, Hartmut Mix, Accessing Bio-Databases with OGSA-DAI -A Performance Analysis. Retrieved 5th September 2009 from http://www.dgrid.de/fileadmin/user_upload/documents/DGI-FG1.9/OGSA-DAI/GCCB06_OGSADAI_SCOPPI_Kottha.pdf
- [6] OGSA-DAI. Open grid services architecture data access and integration. Retrieved 10th September 2009 from <http://www.ogsadai.org.uk/>.
- [7] The Design and Implementation of Grid Database Services in OGSA-DAI. Retrieved 10th September from <http://www.nesc.ac.uk/events/ahm2003/AHMCD/pdf/156.pdf>.
- [8] <http://www.ogsadai.org.uk/documentation/ogsadai-wsi-2.2/doc/background/architecture.html>
- [9] <http://www.ogsadai.org.uk/documentation/ogsadai3.0/ogsadai3.0-axis/ServerConfigCH.html#DeployActivity>
- [10] www.wikipedia.com

