AN OVERVIEW OF ONTOLOGY BASED APPROACH TO ORGANIZE THE METADATA FOR GRID DATABASE

Krisna Adiyarta and Naomie Salim
Fakulti Sains Komputer & Sistem Maklumat
Universiti Teknologi Malaysia

Abstract

Grid is a promising application infrastructure that promotes and facilitates the sharing and collaboration in the use of distributed heterogeneous resources through Virtual Organization (VO). The Semantic Grid refers to an approach to Grid computing in which information, computing resources and services are described in standard ways that can be processed by computer. This makes it easier for resources to be discovered and joined up automatically, which helps bring resources together to create virtual organizations. Semantic grid allow user to discover data based-on service. Data Publication is the process by which data sets and their associated attributes are stored and made accessible to a user community. A gap arise when data is published by its metadata that focused on the context of its data semantic, on the other hand, normally discovery process is focused to the service concept. This paper gives an overview that can be used as an overview approach to solve this gap. Our approach focuses on the usage of specifics of descriptive information about services, and relevant information objects as classification scheme. We also discuss the possibility of ontology usage for metadata organizing.

Keyword : Grid Database, Metadata, Ontology, Reliability

1. Introduction

Chetty *et. al*.(2002) shows Increased network bandwidth, more powerful computers, and the acceptance of the Internet have driven the on-going demand for new and better ways to compute. Commercial enterprises, academic institutions, and research organizations continue to take advantage of these advancements, and constantly seek new technologies and practices that enable them to seek new ways to conduct business. However, many challenges remain. Increasing pressure on development and research costs, faster time-to-market, greater throughput, and improved quality and innovation are always foremost in the minds of administrators - while computational needs are outpacing the ability of organizations to deploy sufficient resources to meet growing workload demands.

Grid is a system that coordinates resources that are nor subject to centralized control and delivers nontrivial qualities of service. On the other word we can say that Grid is used in the widest sense to describe the ability to pool and share Information Technology (IT) resources in a global environment in a manner which achieves seamless, secure, transparent, simple access to a vast collection of many different types of hardware and software resources, (including compute nodes, software codes, data repositories, storage devices, graphics and terminal devices and instrumentation and equipment), through non-dedicated wide area networks, to deliver customized resources to specific applications.

There are five major application classes for computational grids, which are Distributed supercomputing, high-throughput computing, On-demand applications, data-intensive applications, Collaborative applications. Our research is focused on data intensive application. Data intensive scientific applications promise dramatic progress in scientific discovery. These applications produce and analyze terabyte and petabyte data sets that may span millions of files or data objects.

Metadata services are required to support these data intensive applications. Metadata is data about data, and is important as it adds context to the data, aiding its identification, location, and interpretation. Key metadata includes the name and location of the data source, the structure of the data held within it, data item names and descriptions. Astronomy community provides a good example of how metadata services are used in collaborative environments. Initially when the data, in the form of images, is collected by an instrument, such as a telescope, it is pre-processed, calibrated and stored in an archive. Metadata about the images describing the location in the sky, the calibration parameters, etc., are stored as well. Additional processing may occur to extract interesting features of particular regions of the sky or to produce images focusing on particular celestial objects. The information about the processing is captured in metadata attributes and stored as well. Once the metadata and data are prepared in this fashion, they are released to the group of scientists within the collaboration. Researchers can then pose queries on the metadata to discover data relevant for them. Based on the results of the searches, scientists may want to organize the metadata in a way that is most appropriate for their research. During this phase of the data publication process, the data may be further annotated by the collaborators. After a certain

period of time, usually on the order of two years, the data and the metadata are released to the general public. At that time, users outside of the initial collaboration can search the metadata based on attributes that are important to them.

New information architectures enable new approaches to publishing and accessing valuable data and programs. So-called service-oriented architectures define standard interfaces and protocols that allow developers to encapsulate information tools as services that clients can access without knowledge of, or control over, their internal workings. A service is a network enabled entity that provides some capability through the exchange of messages. Grid services extend standard web services by providing support for associating state with services, managing the lifetime of service instances, and standard mechanisms for subscription and notification of state changes.

Grid services are used in many application domains today to deliver computing power and data management capabilities needed by large-scale science. Grid services extend standard web services by providing support for associating state with services, managing the lifetime of service instances, and standard mechanisms for subscription and notification of state changes. A Web service specifies the procedure's syntax in its Web Service Definition Language (WSDL) instance. Metadata adds method semantics, or meaning, in each service. Such enriched function calls (Grid service ports) are characteristics of the Semantic Grid. Roure *et. al.*(2005) defined The Semantic Grid is an extension of the current Grid in which information and services are given well-defined meaning, better enabling computers and people to work in cooperation.

The existence of metadata is basically used to aid user for data discovering process. Our research motivation is based on the lack of standardization in grid. We argue that standardization is a way to reach the reliability in discovery process. Musa *et. al.* (1987) defined reliability statistical terms as "The probability of failure free operation of computer program in a specified environment for a specified time". For an illustration, there is an application program which expected to discover a data about 'X' in a specified environment for a specified time. Let us assume, there are 100 relevant metadata items in catalog for a desire input ('X"). From the execution of the application, it finds all of those 100 item metadata. This result is fit with the actual circumstances regarding to the existence of the metadata which is relevant to the data 'X'. This circumstance shows that the application has a high reliability value.

Our research tries to find a mechanism for data publishing and data discovery that can produce a reliable metadata service for grid database. Based on the characteristic of semantic grid, we argue that we need an annotation of metadata that relates to the grid service. Data publication is focused only on data sets and their associated attributes which stored and made accessible to a user community. Practically, information about data in grid environment is based on service. This study would give a solution to answer the gap between data publication and data discovery in the context of grid computing. We begin with the assumption that user need information regarding to database that related to a particular service concept.

2. Ontology

The term ontology is used very differently in various areas of computer science. Gruber (1993) defined "An ontology is a formal explicit specification of a shared conceptualization". The keyword in this definition is conceptualization. A conceptualization is an abstract simplified view of a domain, it identifies the concepts relevant in representing the domain. This view should reflect consensual knowledge and thus be shared by a group. Ontology describes this conceptualization by making the concepts and relations explicit, i.e. by defining terms and axioms, in some formal language that is machine readable.

Our motivation to use ontology as the methodology for our research was based on the definition given by Partrige (2005) who defined that for the purposes of database integration, the traditional philosophical (metaphysical) notion of ontology is useful, where the ontology is "the set of things whose existence is acknowledged by a particular theory or system of thought." In this perspective, each database can be regarded as a 'theory' that acknowledges the existence of a set of objects. Some care needs to be taken to distinguish this traditional metaphysical use of the word ontology from one that has recently developed in Computer Science. Here ontology is regarded as a "specification of a conceptualization".

Several research used ontology in database environment. Brisaboa *et. al*(2002) show how ontologies used to integrate different database schemas. In their system, ontologies are conceptual models with an abstraction level higher than the schemas of any database integrated in the federation. Ontologies give a homogeneous description to different schemas of databases integrated in the system. It means that ontologies offer a way to make the necessary schema conciliation. Since we saw that metadata can be extracted from its schema Lee and Hwang (2001), then we argue ontology could be used in area of metadata field.

3. Metadata dan Catalog

Grid Database is a federated and distributed database system. Federated database system offers several databases in different format to work together for a particular application. Grid database is not only in a federated system but also geographically distributed. Each database is autonomously and managed by the

owner. Database metadata that it could be useful to have access to includes Content description and Capability description. Normally content description is a the metamodel of the data in the database describes what the database contains. The capabilities of a DBMS are many and varied, and a service architecture must be able to accommodate systems with diverse facilities. This is supported by individual services making their capabilities known.

Atkinson *et. al*, (2003) suggests that metadata should be composed into several types regarding to the role of these metadata in metadata service. To enhance the reliability of the discovering processes, metadata should be composed into three categories, those are technical metadata, contextual metadata, and mapping metadata. Technical metadata describes the information regarding to the physical form of the data. It includes location of data resource, physical structure, data organization, and the information of how to access the database. Each database will have a particular technical catalog. Contextual metadata is used as the logical references for several technical metadata which have a similarity. Metadata Catalog Service (MCS) (Singh *at. al*.,2003), EGEE gLite (EGEE, 2005) uses 'logical file names' (LFNs) and the collection of it which referred to several 'physical file name' (PFNs). Each PFNs describe the metadata of a database in technical or physical description. Because of the existence of technical metadata and contextual metadata, then a particular metadata is used as a mapping between them. Mapping metadata describes mapping between contextual metadata and technical metadata. Publication of database will create a technical metadata. Normally contextual metadata is organized by virtual organization or user.

Several approaches have been proposed in order to increase the reliability of the metadata service. MCS and EGEE gLite, Storage Resource Broker (SRB) (Baru *et. al*. 1998) give an opportunity to the user to create several collections regarding to their interest, e.g. based on the domain application. To improve the reliability of data discovery process, SRB use 'Descriptive Metadata' which used a standard schema that similar to the specification of Dublin Core (Dublin Core, 2004). The Natural Environment Research Council Data Grid (NGD) (O'Neill *et. al*. 2003) try to improve the reliability by making a standardization to the terminologies which used for the system. NDG try to limit the usage of terms to avoid a misunderstanding between system and the user.

Database semantic schema have to be analyzed in order to associate to a particular metadata. We argue that a mechanism for automatic cataloging could be exist in grid since we identified that grid middleware has controlled the database and consider the consistency for database schema. Database reverse engineering gives an opportunity to be use as an approach to the automatic metadata extraction from a particular database. Chiang *et. al*. (1994) define database reverse engineering (DBRE) as a process that obtains domain semantics about an existing database, then converts the schema from relational to conceptual, and finally represents the results as a conceptual schema. Lee and Hwang (2001) propose a reverse engineering agent for the Conflict Resolution Environment Autonomous Mediation (CREAM) system, called the Semantic Metadata Extracting & Visualizing Agent (SMEVA)

4. Classification Approach

Two important steps in data publication process which are content analyzing and classification process. Content analyzing what we mean in our research is to analyze the semantic of a database trough its metadata. Next, classification process could be done based on the semantic of metadata. Classification process would classify the database to be associated to a particular subject in metadata management. Subject refers to a concept for a particular database semantic. Many subjects would be organized trough several schema what we called a classification scheme. This section we would discus several classification techniques usually use in metadata management. In this section we assume that we have identified the metadata semantic from a database.

Mota and Lu, (2006) said that classification can be seen as the problem of finding the solution (class) which best explains a certain set of known facts (observables) about an unknown object, according to some criterion. In the following we will discuss each of these concepts in turn. We need for a set of intelligent tools that can process existing data by interpreting structure and content, extracting relevant metadata information, and populating definitions automatically. The idea of our research actually is to answer these needs. MCS has shown a simple classification in which a technical metadata that describes a database is mapped to the relevant application domain.

Classification process begins with analyzing the object content that focused on the terminology and ontology. The next step is to classify the object with a relevant metadata that logically referred to the object from a classification scheme. Several terms could be derived from particular database that construct the structure of data element. We argue that metadata could derive from the structure data element. If general classification scheme has already been exist in the system, then classification process can be performed. We need a classification technique to classify an object based on its metadata to the classification schema. In our initial study, we have identified two categories of classification techniques, terminology based and ontology based. Terminology based classification is the classification technique which classify object that focus on term used, such as controlled vocabulary, taxonomies, thesauri and faceted classification. Ontology based classification is close to classify the object by its semantic. Several classification techniques, such as :

- Controlled vocabulary, Controlled vocabulary is a rather broad term, but here we mean by it a closed list of named subjects, which can be used for classification. In library science this is sometimes known as an indexing language. The constituents of a controlled vocabulary are usually known as terms, where a term is a particular name for a particular concept. The purpose of controlling vocabulary is to avoid defining meaningless terms. French *et. al*. (2001) demonstrates a techniques for mapping user queries into the controlled indexing vocabulary that have the potential to radically improve document retrieval performance and also show how the use of controlled indexing vocabulary can be employed to achieve performance gains for collection selection.
- Taxonomies, Taxonomy is a form of classification that arranges the terms in the controlled vocabulary into a hierarchy. The benefit of this approach is that it allows related terms to be grouped together and categorized in ways that make it easier to find the correct term to use whether for searching or to describe an object. The Summary Schemas Model provides a user-friendly interface by allowing users to specify queries in their own terms and/or to use imprecise terms for data references. The summary schemas and the online taxonomy are used to map these terms to the semantically closest access terms that actually exist in the system. Bright, *et. al*.(1994)
- Thesauri, Thesauri basically take taxonomies as described above and extend them to make them better able to describe the world by not only allowing subjects to be arranged in a hierarchy, but also allowing other statements to be made about the subjects. Wattenbarger *et. al*. (1977) shown how thesauri is used to structure inter-relationships between terms, in order to clarify the scope and meaning of each term,
- Faceted classification, The Idea of Faceted classification is to classify documents by picking one term from each facet to describe the document along all the different axes. The facets can be thought of as different axes along which documents can be classified, and each facet contains a number of terms. This would then describe the document from many different perspectives. Prieto-Díaz. (1991)
- Ontologies, Main purposes of the classification is to link a particular catalog of an object to a relevant concept in the contextual metadata for easier discovery. Ontologies in computer science came out of artificial intelligence, and have generally been closely associated with logical inferencing and similar techniques. The term ontology has, need we say it, been applied in many different ways, but the core meaning within computer science is a model for describing the world that consists of a set of types, properties, and relationship types.

5. Approach

Every opportunity in meeting the requirements must be taken to ensure that, wherever possible, the metadata definition, publication and specification processes are automated and that the burden of manual metadata entry and editing is minimized. There is a need for a set of intelligent tools that can process existing data by interpreting structure and content, extracting relevant metadata information, and populating definitions automatically.

In metadata organization, discussing of metadata organizing lead us to focus on classification method. Based on some identified classifications discussed in previous section, we argue that ontology is the most appropriate technique to solve gap between data discovery and data publishing. Generally we identified two categories for classification techniques : terminology based and ontology based. Terms that used in publication and discovery has a distinction domain which, in the publication focused on the concept of data whereas discovery focused on the concept of service. This distinction causes a lack in classification process. To solve this problem, we proposed to use semantic based classification. We argue that it is better to classify the metadata in the term of concept model. Ontology has a better model to specify a concept.

5.1. Data Publishing

For our data publication approach consist of four primary functions. First, a function is used as a metadata extractor from a database. Foster *et. al*. (2001) shown that the lowest level of the architecture is fabric layer. The Grid Fabric layer provides the resources to which shared access is mediated by Grid protocols. Fabric components implement the local, resource-specific operations that occur on specific resources as a result of sharing operations at higher levels. Regarding to these architecture, Database and it's catalog can be a resource. Resource management mechanism is needed to provide some control of delivered quality of service. Based on the architecture, we argue that it is possible for grid middleware to control the databases. Control what we mean in here is including a process for technical metadata extraction automatically. The catalog extracted by the system is stored in local metadata repository.

Second, metadata harvesting. Similar to the concept of Open Archives Initiative Protocol for Metadata Harvesting which provides an application-independent interoperability framework based on metadata harvesting, OAI, (2004), our approach use a software agent to harvest metadata from local metadata repository. Local metadata repository that we mean is the database catalog repository that controlled by grid middleware.

Third, Ontology creation for each harvested metadata. Broekstra *et. al*.(2003) has shown the Ontology creation. The extracted structures are a starting point for the ontology creation process. They suppose that the extracted information and background knowledge like e.g. WordNet3 can be used to create these richer structures. This would mean adding formerly implicit semantics explicitly to the structure.

Fourth, classify the metadata to the classification schema using ontology-based classification technique. Prabowo *et. al*.(2002) describes an automatic classifier, which focuses on the use of ontologies for classifying Web pages with respect to the Dewey Decimal Classification (DDC) and Library of Congress Classification (LCC) schemes. They also explain how these ontologies can be built in a modular fashion, and mapped into classification scheme. Another techniques potentially could be used is by measuring the similarity of ontologies Marc *et. al*, (2004). Since we have metadata ontology and service ontology, we argue that similarity measurement could be used to map a metadata to a service concept. In this technique classification schema is based on the concept of service, which each service concept is represented by an ontology.

5.2. Data Discovery

Grid needs to support data discovery through interactive browsing tools, and from within an application when discovery criteria may be pre-defined. Browsing might be shown subject classification or index of physical metadata which relevant to particular metadata. Rose *et. al*, (1994) has shown how metadata is organize in term of hierarchical concept and can be browsed by user in order to find the relevant metadata is need by them.

In our research we, proposed two steps in data discovery process. First identified the concept and the second chose the relevant metadata. For the first steps, we are interested with IBM VideoAnnEx, (Wu *et. al*, 2004) In NIST TREC Video 2003 benchmark, 133 concept are defined in lexicon, organized hierarchical ontology based on the IBM MPEG-7 Video Annotation Tool. Since we can measure the similarity between two ontologies, it is possible for us to rank the metadata as the result for a particular concept chosen by user. Marc *et. al*, (2004) shown an approach to measure the similarity of two ontologies.

5. Conclusion

In our research, we try to solve the gap between discovery process and publication process. Terms gathers from metadata extraction are analyzed and classified to the relevant conceptual metadata. Our classification technique is used an ontology approach. Ontology gives a better approach to specify semantic of a metadata and service. Conceptual metadata is visualized in hierarchical way and at the end the technical metadata that refers to the concept is display in rank order.

REFERENCES

Atkinson *et. al*, 2003, "Grid Database Access and Integration: Requirements and Functionalities", Global Grid Forum (2003) http://www.gridforum.org/documents/GFD.13.pdf. , [Accessed, September 15th ,2005]

Baru. *et. al*. (1998) , "The SDSC Storage Resource Broker", Proc. Of CASCON'98 Conference, Toronto, Canada (1998)

Bright. *et. al*.(1994), "Automated resolution of semantic heterogeneity in multidatabases", June 1994 ACM Transactions on Database Systems (TODS), Volume 19 Issue 2 , ACM Press

Brisaboa. *et. al*.(2002), "Ontologies for Database Federation", The European Online Magazine for IT Profesional, Vol. III, No. 3, June 2002. http://www.upgrade-cepis.org

Broekstra. *et. al*.(2003), "A Metadata Model for Semantics-Based Peer-to-Peer Systems". In Proc. WWW'03 Workshop on Semantics in Peer-to-Peer and Grid Computing. 2003.

Chetty. *et. al*.(2002), "Weaving Computational Grids: How Analogous Are They with Electrical Grids ?", Computing in Science and Engineering (CiSE), ISSN 1521-9615, Volume 4, Issue 4, Pages: 61-71, IEEE Computer Society Press and American Institute of Physics, USA, July-August 2002

Chiang, *et. al*. (1994), "Reverse Engineering of Relational Databases: Extraction of an EER Model from a Relational Database", Data & Knowledge Engineering 12, 2 (March 1994), pp. 107-142.

Deelman. *et. al*,(2004), Grid-Based Metadata Service, In 16th International Conference on Scientific and Statistical Database Management (SSDBM04), page NA, 2004.

Dublin Core, (2004), "Dublin Core Metadata Element Set, Version 1.1: Reference Description", http://www.dublincore.org/documents/dces/, December 2004

EGEE, (2005), "EGEE gLite Metadata Catalog User's Guide **:** **G**Lite Metadata Catalog Interface Description", https://edms.cern.ch/document/573725

Foster, *et. al*. (2001), " The Anatomy of The Grid: Enabling Scalable Virtual Organizations". International Journal of High Performance

Computing Applications, 2001. **15**(3): p. 200-222.

Foster. (2002), "What is the Grid? A Three Point Checklist". *Grid Today*, **1**(6), July 22, 2002.

Foster, (2005), "Service-Oriented Science", 6 MAY 2005 VOL 308 SCIENCE

Foster and Kasselman (1998), "The Grid: Blueprint for a Future Computing Infrastructure", Morgan Kaufmann Publishers 1998

Fox (2003), "Data And Metadata On The Semantic Grid", Computing in Science & Engineering [see also IEEE Computational Science and Engineering] Volume 5, Issue 5, Sept.-Oct. 2003 Page(s):76 – 78

French. *et. al*. (2001), "Distributed Information Retrieval: Exploiting a controlled vocabulary to improve collection selection and retrieval effectiveness", October 2001 Proceedings of the tenth international conference on Information and knowledge management , ACM Press

Gruber. (1993), "Towards principles for the design of ontologies used for knowledge sharing". Technical report, Stanford University, Palo Alto, CA, 1993.

Lee and Hwang. (2001), "Extracting semantic metadata and its visualization", March 2001 Crossroads, Volume 7 Issue 3 , Publisher: ACM Press

Marc. *et. al*, (2004), "Similarity for ontologies - a comprehensive framework". In Workshop Enterprise Modelling and Ontology: Ingredients for Interoperability, at PAKM 2004, DEC 2004

Mota and Lu., (2006), "A Library of Components for Classification Problem Solving", http://kmi.open.ac.uk/projects/ibrow/Publications/Motta_pkaw00.pdf, accessed [23 Feb 2006]

Musa, *et. al*. (1987)," Engineering and Managing Software with Reliability Measures", McGraw-Hill, 1987

O'Neill, *et. al.* (2003), "The Metadata Model of the NERC DataGrid", Proceedings of the U.K. escience All Hands Meeting. S.J.Cox (Ed) ISBN 1-904425-11-9

Oracle, (2005), Oracle Grid Computing: An Oracle Business White Paper February 2005, http://www.oracle.com/technologies/grid/OracleGridBWP0105.pdf

Paton. *et. al*.(2002), "Database Access and Integration Services on the Grid", UK e-Science Programme Technical Report Series Number UKeS-2002-03, National e-Science Centre, UK.

Prabowo. *et. al*.(2002), "Ontology-based automatic classification for Web pages: design, implementation and evaluation", Web Information Systems Engineering, 2002. WISE 2002. Proceedings of the Third International Conference on 12-14 Dec. 2002 Page(s):182 - 191

Prieto-Diaz, (1991). "Implementing faceted classification for software reuse". *Commun. ACM 34*, 5 (May), 88–97.

Roure, *et. al*.(2005), "The semantic grid: past, present, and future", Proceedings of the IEEE Volume 93, Issue 3, Mar 2005 Page(s):669 – 681

Rose, *et. al*, (1994). "Hierarchical classification as an aid to database and hit-list browsing ", November 1994 Proceedings of the third international conference on Information and knowledge management , ACM Press

Singh, *at. al*. (2003), "A Metadata Catalog Service for Data Intensive Application", Procceding of the ACM/IEEE SC2003 Conference SC'03, November 15-21, 2003, Phoenix, Arizona, USA

Wattenbarger, *et. al*. (1977), "Interactive system for controlled vocabulary maintenance", January 1977 Proceedings of the 1977 annual conference, ACM Press

Watson (2001), "Database and The Grid", Technical Report CS-TR-755, University of Newcastle, 2001.