DartGrid: Semantic-Based Database Grid

Zhaohui Wu , Huajun Chen, Changhuang, Guozhou Zheng, and Jiefeng Xu

College of Computer Science, Zhejiang University, Hangzhou, 310027, China
{wzh,huajunsir,changhuang,zzzgz,xujf}@zju.edu.cn

Abstract. In presence of web, one critical challenge is how to globally publish, seamlessly integrate and transparently locate geographically distributed database resources with such "open" settings. This paper proposes a semantic-based approach supporting the global sharing of database resources using grid as the platform. We have built a semantic query system, called DartGrid, with the following features: a) database providers are organized as an ontology-based virtual organization; by uniformly defined domain semantics, database could be semantically registered and seamlessly integrated together to provide database service, and b)we raise the level of interaction with the data base system to a domain-cognizant model in which query request are specified in the terminology and knowledge of the domain(s), which enable the users to publish, discovery ,query databases only at a semantic or knowledge level. We explore the essential and fundamental roles played by data semantics, and implement some innovative semantic functionalities such as semantic browse, semantic query and semantic registration. We also reports on application results from Traditional Chinese Medicine (TCM) that requires data-intensive collaboration.

1 Introduction

In the next evolution step of web, termed semantic web[1], vast amounts of information resources (databases, multimedia, programs) will be enriched with uniformed semantics for automatic discovery, seamless communication and dynamic integration. In presence of such semantics defined for database integration or sharing, one critical challenge is how to transparently translate a semantically enriched query into a distributed query plan, and then properly locate and access geographically distributed database resources with such "open" settings.

This paper proposes a semantic query system, called DartGrid, using grid as the platform. DartGrid is designed to support the building of large-scale ontology-based database Virtual Organization (DB-VO), in which databases are organized by uniformly defined semantics, namely, domain ontologies. In a DB-VO, databases are semantically registered to a web service called Semantic Registry Service(SeRS) and the user query the system only at a semantic and knowledge level. Usually, the user dynamically generates a visual conceptual query when browsing the ontolgies stemming from Ontology Service, meanwhile, a semantic query is generated and submitted to Semantic Query Service (SeQS). After inquiring of SeRS about the mapping

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3036, pp. 59–66, 2004.

from shared ontologies to local database schemas, the semantic query is converted into a local database query languages (e.g., SQL , XQuery,etc.) . And then SeQS builds a distributed query plan to dispatch the query into proper database service .The result returned will be semantically wrapped again before they are presented for semantic browsing by the user.

Our work is essentially motivated and informed by requirements of communities of TCM researchers and professionals and our experience in building several TCM information systems [2]. Currently, in our deployed testbed, an ontology service, with about 10,000 records of TCM ontology instances contained, has been set up, and ten nodes with thirty TCM-related databases have been deployed. Reports from our partners, China Academy of TCM and its associated enterprises and institutes, show that our system significantly promotes the sharing and integration of their database resources and greatly facilitates their cooperation in their preferable web mode.

2 Ontology-Based Virtual Organization

As well-known, Virtual Organization enables disparate groups of organizations and/or individuals to share resource in a controlled fashion, so that members may collaborate to achieve a shared goal. We argue that ontology will play a significant role in constructing such Vos . Ontology defines the formal conceptuation model and standard terminology of the domain, which could significantly improve the sharing level within such VOs. In the following, we give a formal definition of the ontology-based VO as follows.

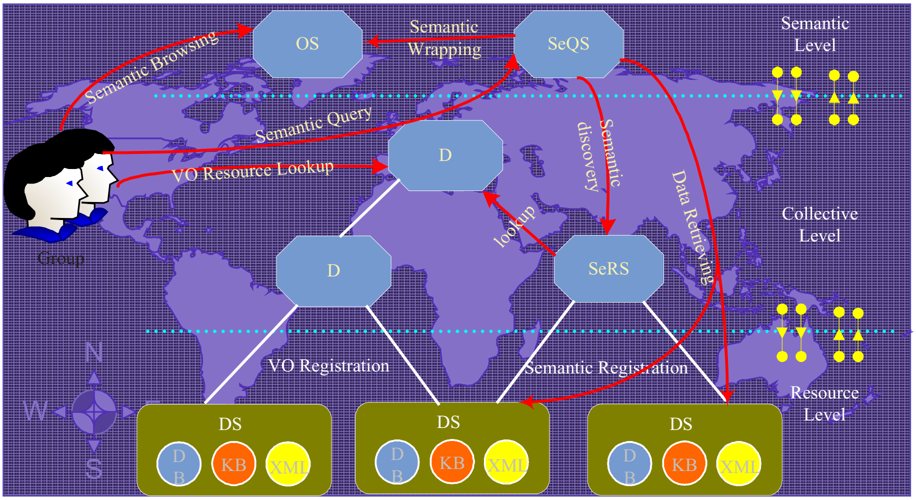


Fig. 1. A Formal Model of Ontology-based Virtual Organization

Definition 1: *an Ontology-based Virtual Organization OntoVO is a four-tuple OntoVO = (O+, SeRS, DS, FS) and*

- *O+ is a set of ontology services related to VO, "+"means every OntoVO must has at least one ontology;*
- *SeRS is Semantic Registration Service maintaining the mapping from the global semantics to local schemas, and is also a ontology-based index for classifying data objects*
- *DS is Data Service providing data objects*
- *FS is a set of optional functional services such as Semantic Query Service(SeQS) providing semantic query parsing and dispatching and address directory service (D in the figure) maintaining the physical address of all networked entities.*

Definition 2: *an Ontology Service OS is a two-tuple OS=(T_{os}, P_{os}), and*
- *T_{os}, is a set of terminologies which define the domain of this VO;*
- *P_{os} is service porttype which specifies a set of necessary knowledge-level operations on T_{os}.*

Definition 3: *an Data Service DS is a two-tuple DS= ($M_{ds\,,}$ $A_{ds,}$) where*
- *M_{ds} is meta data about the data service ;*
- *A_{ds} is the data objects provided by data service;*

3 Implementation

DartGrid is a referential implementation of the OntoVO model. The principal technical characteristics of DartGrid are highlighted as below:

1. We develop it on Globus 3.0, the de facto standard platform to construct VO in Grid Computing research area.
2. RDF, the standard data model for web semantics defined by W3C, is adopted as the universal data model for defining protocols such as protocol for semantic registration.
3. Ontologies used in DartGrid comply with the syntax and semantic of OWL, the standard ontology description language proposed by W3C.

Firstly, we introduce the core components developed in DartGrid as follows.

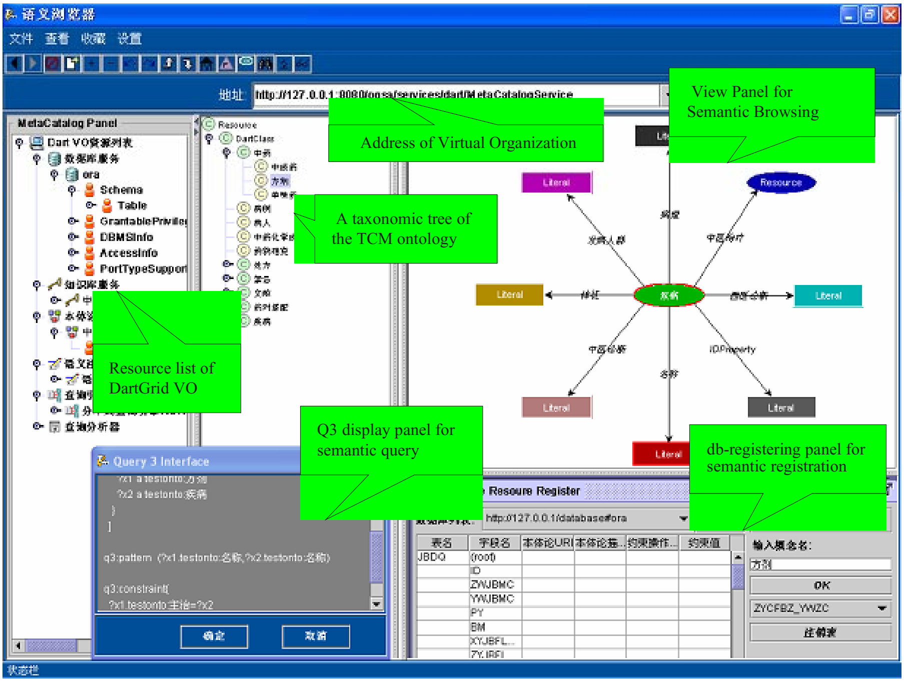
3.1 Building Blocks of DartGrid

3.1.1 Semantic Browser

Current web browsers are designed for human to browse web documents, and they only know how to interpret the HTML tags and present it as a plain text document. We proposed and developed a general-purpose browser, called the Semantic Browser [4], as the uniform user interface that enables the user to manipulating data semantics in DartGrid. The semantic browser we developed has the following characteristics. Figure 2 is a snapshot of the semantic browser.

(1) Improved navigation. User could use semantic browser to visit an ontology service and visualize the ontologies maintained in it. Using of ontologies provides the improved navigation. The user gets easy access to relevant information by browsing through the modeled concepts and their relations. An example is the navigation from a medicine to its relevant diseases.

(2) Visual Semantic Query Generation. User could visually generate a conceptual level query by interacting with the semantic browser when browsing domain ontologies.
(3) Visual Semantic Registration. Semantic browser provides the data vendor with a tool for visually doing mapping from local data semantics to shared domain ontologies.



View Panel for Semantic Browsing

Address of Virtual Organization

A taxonomic tree of the TCM ontology

Resource list of DartGrid VO

Q3 display panel for semantic query

db-registering panel for semantic registration

```
  ?x1 a testonto:方剂
  ?x2 a testonto:疾病
 }
]

q3:pattern (?x1.testonto:名称,?x2.testonto:名称)

q3:constraint(
 ?x1.testonto:主治=?x2
```

http://127.0.0.1/database#ora

表名	字段名	本体论URI	本体论属...	约束操作...	约束值
JBDQ	(root)				
	ID				
	ZWJBMC				
	YWJBMC				
	PY				
	BM				
	XYJBFL...				
	ZYJBFL				

Fig. 2. A Snapshot of the Semantic Browser

3.1.2 Semantic Services

DartGrid has implemented several semantic-level services, they are:

(1) Data Semantics Service (DSS):

Data resource vendors publish the information about local data semantics by this service. For database resource, the local semantic information is right the schema information about the tables contained in that database. Others could inquire of this service about the local data semantics to fulfill some tasks such as doing semantic mapping or integration.

(2) Ontology Service (OS):

Ontologies define the standard vocabularies/ terminologies/concepts and models of the domain of a VO. Thereby, ontologies could be viewed as public-agreed global data semantics. In the new VO model we defined for DartGrid, a VO should have at least one shared ontology. If a data vendor wants to join in the VO, he/she should map his local data semantics to the ontology service, which guarantee the data sharing and integration.

(3) Semantic Registration Service (SeRS):

In current DartGrid prototype, SeRS is designed specially for database resource sharing use and distinguishes themselves by the following characteristics from other index solutions such as UDDI or Grid Index Service:

- It maintains the information about semantic mapping from local data schemas to global ontologies. Any data resource provider should register their local schema to SeRS and finish the mapping process.
- It maintains an ontology-based taxonomy and is responsible for classifying all data objects by this taxonomy. In this case, SeRS determines which data service could answer a specific query and where they locate.

3.2 Semantic Query

Informally, we specify a semantic query herein as:

(1) a query expressed at a knowledge level, namely, the query is specified by formally defined concepts and their relationships, for example, the concept of *Disease* and *Medicine*, and their relationship *curedBy.* A typical knowledge query could be expressed by LISP syntax :*(curedBy HeartDisease ?Medicine)*,Which means we want to query all medicines which could help treat heart diseases.

(2) a query whose terms are specified by uniform shared global semantics, namely, all terms used in a semantic query are defined by public-available, widely-agreed shared ontologies.

(3) a query whose result returned should also be semantically enriched.

Normally, a semantic query will be converted into some local query languages such as SQL or XQuery. Afterward, a distributed query plan will be generated and the query is dispatched to proper data service to retrieve data of interest. In DartGrid, three key components are developed to implement the semantic query, they are:

(1) Q3 language:

We devise a formal semantic query language called Query3(Q3). Q3 adopts RDF data model and N3(Notation3) syntax, and could also be used to query description-logic knowledge base.

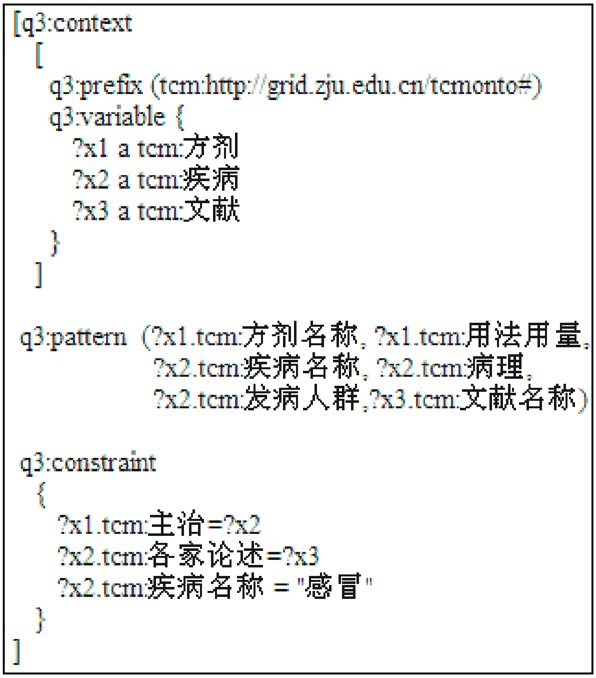
Definition 4*: an Semantic Query is a triple SeQ=(Cxt, Pat, Cst) and*

- *Cxt is the context of the query and Cxt=NS U VB where NS is the namespace of the term used and VB is the variable biding and scoping;*
- *Pat is the concept pattern of the result returned;*
- *Cst is the constraint of the query and one constraint is a statement (S,P,O) in which S is subject, P is predicate and O is object, all of them could be bound with a variable*

Figure 3 illustrate an example of a typical Q3 query where :

- *q3:prefix* specifies the namespace *"http: //grid.zju.edu.cn/tcmonto#"* used in this query, and its corresponding QName *"tcm"*;

- *q3:variable* specifies the variables used in portions of *q3:pattern* and *q3:constraint*. For example, *?x1* stands for the concept tcm:方剂 (*tcm:CompoundMedicine*);
- *q3:pattern* specifies the concept pattern for the result. For example, the occurrence of the term ?x1.tcm:方剂名称 implies that the result should contain the name of the *tcm:CompoundMedicine*.
- *q3:constraint* specifies the query constraints that the result should satisfy.



```
[q3:context
  [
    q3:prefix (tcm:http://grid.zju.edu.cn/tcmonto#)
    q3:variable {
      ?x1 a tcm:方剂
      ?x2 a tcm:疾病
      ?x3 a tcm:文献
    }
  ]

 q3:pattern  (?x1.tcm:方剂名称, ?x1.tcm:用法用量,
              ?x2.tcm:疾病名称, ?x2.tcm:病理,
              ?x2.tcm:发病人群,?x3.tcm:文献名称)

 q3:constraint
  {
    ?x1.tcm:主治=?x2
    ?x2.tcm:各家论述=?x3
    ?x2.tcm:疾病名称 = "感冒"
  }
]
```

Fig. 3. A Q3 Example

(2) Visual Semantic Query Generator:
Although we have provided programming interface for user to write and issue a Q3 query, it is a non-trivial task to manually write a Q3query. We have developed a visual query generator, a core component of our semantic browser, to facilitate the query construction. Normally, user browses the ontologies graphically, select the concepts of interest, and specify the constraints dynamically, afterward, submit the query to a semantic query service.

(3) Semantic Query Service(SeQS):
In our implementation, SeQS plays two significant functionalities, they are:

- Receiving a Q3 query generated by Semantic Browser and then converting it into a distributed query plan in SQL-syntax;
- Wrapping the results returned from DB resources with semantics, which enables the users to browse the results semantically.

3.3 Semantic Registration

The following TCM scenario illustrates how to register a db-resource to SeRS. In this case, a TCM data provider wants to add his compound-medicine (方剂) database resources into the TCM-VO for sharing.

(1) Publish his databases as a Data Service;

(2) Visit the data service from semantic browser. This will retrieve the data schema of the databases and display it in the db-registering panel.

(3) At the same time, he opens the TCM ontology service and locates the concept 方剂 in it. Firstly, he maps the concept to the 方剂 table, and then maps the properties of the concept of 方剂 to the corresponding column name of the 方剂 table. This will construct a registration entry in XML format.
(4)Before the final submitting, the registration entry will be sent to the ontology service for semantic verification. This will verify that the concepts included in the entry are valid.

4 Related Work

In a wider technical context, the proposal presented in this paper is part of a collection of results on knowledge-based query processing in distributed information system such as SIMS[5], OBSERVER[6] , TAMBIS[7] etc.. However, the essential difference between such proposals and ours is that DartGrid enables semantic processing in such a world-wide open setting. Although the knowledge-based "global as view or model" is adopted in all of those proposals, there was no consideration with related to semantics.

In a technical context of semantic web, DartGrid is also significantly different with those semantic web query systems such as SESAME (RQL) [8], HP's Jena (RDQL) [9], or DARPA's DQL [10]. The other characteristic of an open system such as web is "dynamic"; all of the proposals aforementioned take no consideration with that issue. In our approach, the SeRS enables the data providers to join in or drop out from the VO dynamically. Particularly, SeRS maintain the current status of the providers, and if some providers are become unavailable at some time, SeRS will mark them as inactive. User is not aware of it at all and there is no need for him to care which data service could answer it or where they locate.

In a technical context of Data Grid efforts such as EU's DataGrid[11], GridPhyN project [12], GGF's DAIS working group and so on, the significant difference is the semantic-based approach adopted in DartGrid. We have not seen such approach adopted in those efforts.

5 Conclusion

DartGrid catches the aims of enabling to build ontology-based database virtual organizations in such open settings. The significance of semantic in DartGrid is reflected by the following notions:

(1) Semantics guarantees the scalability of the system. This is very important for a web-based open query system.
(2) Semantics enables that the user only need to interact with the system at a semantic level.

DartGrid has been successfully applied in data sharing for Traditional Chinese Medicine in China. In the future, more types of resources such as pictures, audios, etc. will be added into our prototype.

Acknowledgements. This work is supported in part by the project Data Grid for Traditional Chinese Medicine, subprogram of the Fundamental Technology and Research Program, China Ministry of Science and Technology, and the China 863 Research Program on Core Workflow Technologies supporting Components-library-based Coordinated Software Development under Contract 2001AA113142, and Chinese 211 core project: Network-based Intelligence and Graphics.

References

1. Tim Berners-Lee, James Hendler, Ora Lassila. The Semantic Web. Scientific American May 2001.
2. Huajun Chen, Zhaohui Wu, Chang Huang, Jiefeng Xu: TCM-Grid: Weaving a Medical Grid for Traditional Chinese Medicine. Lecture Notes in Computer Science, Volume 2659, Jan. 2003.
3. Ian Foster, Carl Kesselman, and Steven Tuecke. The Anatomy of the Grid: Enabling Scalable Virtual Organizations. Lecture Notes in Computer Science, 2001, Vol. 2150: 1-26.
4. Mao Yuxin, Wu Zhaohui, Chen Huajun: SkyEyes: A Semantic Browser For the KB-Grid. International Workshop on Grid and Cooperative Computing, 2003, Shanghai.
5. Y. Arens, C.A. Knoblock, and W-M. Shen. Query Reformulation for Dynamic Information Integration. J. Intelligent Information Systems,6(2/3):99–130, 1996.
6. E. Mena, A. Illarramendi, V. Kashyap, and A.P Sheth. OBSERVER: An approach for query processing in global information systems based on interoperation across pre-existing ontologies. Distributed and Parallel Databases, 8(2):223–271, 2000.
7. Nardi, and R. Rosati. Information integration: Conceptual modeling and reasoning N. W. Paton, R. Stevens, P. Baker, C. A. Goble, S. Bechhofer, and A. Brass. Query Processing in the TAMBIS Bioinformatics Source Integration System. In Proc. SSDBM, pages 138–147. IEEE Press,1999.
8. Gregory Karvounarakis Sofia Alexaki Michel Scholl: RQL: A Declarative Query Language for RDF*. WWW2002, May 7–11, 2002, Honolulu, Hawaii, USA.
9. Libby Miller, Andy Seaborne, Alberto Reggiori: Three Implementations of SquishQL, a Simple RDF Query Language. HP Technical Report :
   http://www .hpl.hp.com/techreports/2002/HPL-2002-10.htmll
10. Richard Fikes, Pat Hayes, Ian Horrocks : DQL – A Query Language for the Semantic Web WWW 2003, May 20-24, 2003, Budapest, Hungary.
11. Kunszt, P. (CERN, IT Division): European DataGrid project: Status and plans. Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, v 502, n 2-3, Apr 21, 2003, p 376-381.
12. Ewa Deelman, Carl Kesselman et al. GriPhyN and LIGO, Building a Virtual Data Grid for Gravitational Wave Scientists. Proceedings of the 11 th IEEE International Symposium on High Performance Distributed Computing HPDC-11 2002 (HPDC'02).