A Formal Data Model and Algebra for Resource Sharing in Grid*

Qiujian Sheng and Zhongzhi Shi

Key Laboratory of Intelligent Information Processing,
Institute of Computing Technology,
Chinese Academy of Sciences, P.O. Box 2704-28, Beijing 100080, China
&Graduate School of the Chinese Academy of Sciences, Beijing, China
shengqj@ics.ict.ac.cn

Abstract. Management of various complex resources is an important issue to be well done, in order to fulfill the potential of Grid. In this paper, we propose a formal data model and algebra for resource sharing in the Grid. In the model, domain ontology knowledge is described within knowledge schema and domain background knowledge is captured in integrity constrain rules. A query algebra is defined to manipulate the modeled resource base.

1 Introduction

It is envisioned that the success of using Grid for distributed system integration will rely on how to have the heterogeneous resources of the Grid, with its increasing scale and complexity, well managed[1]. Knowledge-based data management techniques will play a fundamental role in Grid, for examples, gird services that manipulate large data sets, especially those for data-intensive science, will need rich optimization techniques; as auto-service discovery became reality we will see "query" being made against huge volumes of intricate Grid-services descriptions, creating new challenges in indexing and query optimization. In the long run, a solid foundation, quasi relational theory, is desired. However, in order to effectively share the complex Grid resources, domain resource/data should be semantically structured and interrelated. In this respect, neither relational model[2] nor conventional object-oriented models are effective[3] for their simplicities, in some way. Recently a few models and query languages have been reported [4]. However, despite the works done so far, the research is far from being mature. An abstract and general foundation is still missing.

In this paper, we aim to develop a solid foundation for a scaleable and manageable facilitates of complex resource sharing in the Grid.

* Supported by the National High-Tech Program (Grant No.2001AA113121) of China, National Natural Science Foundation of China (90104021), and Beijing Natural Science Foundation (4011003)

M. Li et al. (Eds.): GCC 2003, LNCS 3033, pp. 231–235, 2004.

2 A Knowledge-Based Data Model

The model consists of knowledge schema, resource instance base and integrity constrain. The knowledge schema is used to capture the agreed domain knowledge, i.e., concepts and properties. It plays a similar role of database schema in guiding access to the data in the Grid.

2.1 Model Definition

Definition 1. Knowledge schema $KS=(C,D,P,T\prec_c,\prec_p)$, where

- C is a set of abstract concepts, e,g,. Person, Publication etc,
- D is a set of concrete domain concept, e.g., Integer, String, and alike predefined type, $C\cap D=\phi$.
- P is a set of properties of concepts in domain, $P=P_a\cup P_r$, $P_a\subseteq C\times D, P_r\subseteq C\times C$
- $T\subseteq P_r$ is a set of transitive property.

-Concept inclusion relationship $\prec_c \subseteq C\times C$ is a strict partial order. If $c_1\prec_c c_2$ then c_1 includes c_2, c_2 is a sub-concept of c_1. In KS, $\exists c_0\in C,\forall c\in C, c\neq c_0, c_0\prec_c c$

-Property inclusion relationship $\prec_p \subseteq P_r\times P_r$ is a strict partial order. If $p_1\prec_p p_2$ then p_1 includes p_2,or p_2 is a sub-property of p_1.In a knowledge schema KS, $\exists p_0\in P_r,\forall p\in P_r, p\neq p_0, p_0\prec_p p$.

The integrity constraint of knowledge schema captures the domain background knowledge that is not captured within concepts and properties of knowledge schema. It can be used in doing *semantic* query optimization during query process.

Definition 2. The integrity constrain of knowledge schema $R=R_p\cup R_i$, includes constrains on property and other domain background knowledge. There are two kinds of property constrains. One is constrains on concept type of domain and range concept of a relational property. The other is cardinality constrain. The property constrain $R_p=(\psi,\pi,Min,Max)$, where ψ,π are concept type constrains on domain concept and range concept of relational property *P* respectively. *Min* and *Max* are cardinality constrains.

- $\psi: P\rightarrow 2^C$, $\forall p\in P,\exists C_d\subseteq C, C_d=p.domain$
- $\pi: P\rightarrow 2^{C\cup D}$, $\forall p\in P,\exists C_r\subseteq C, C_r=p.range$
- $Min: C\times P\rightarrow N_0$
- $Max: C\times P\rightarrow\{N_0\}\cup\{\infty\}$

Another kind of domain constrain, like "Only those researchers whose title-level is higher than 8 can take position of Director in department.", is captured in our model by integrity rule with a generic form, $S^a\rightarrow S^c$, S^a, S^c are expressions constructed by concept and property. The example rule can be described as: $Director\rightarrow Researcher\wedge title-level:[8,10]$.

Definition 3. Resource Instance Base $RB=(I,L,\lambda,\rho)$,where

- I is a set of resource instances, $\forall\ i\in I$ is an instance of concept $c\in C$
- L is a set of literal values with type of Integer, String, etc., $I\cap L=\phi$
- λ:I$\rightarrow 2^{C}$ relates a concept to a set of instances
- ρ:$P\times$I$\rightarrow 2^{I\cup L}$ assigns to each property-instance pair a set of instances related through given property.

Ontology knowledge(i.e., concept and property) is a abstract of domain resources, while specific domain resources is an interpretation of conceptualized domain knowledge. In order to make a solid formal foundation, we have researched the denotation semantics of the model and its validation condition. Due to space limit, it is omitted here.

3 Algebra

Seven algebraic operators, including Project, Closure,Select,Join,Union,Intersect and Minus are defined to manipulate the modeled resource base. The algebra makes it possible to specify a high-level declarative query language with a clear semantics. Also it's the basis to do query optimization. Here, we mainly present the first four operators.

3.1 Projection

The projection operator gets a collection c and produces a collection with complex resources according to *P,* which specifies the projection behavior.

Definition $\pi_P(c)=\bigcup_{p_i\in P=\{p1,p2,..,pk\}}\{y|\exists x\quad (x,y)\in P^I(p_i)\wedge x\in C^I(c)\}$

Example 3.1. To query author, publish press and publish date of all publications.

$\pi_{authored-by,\,published-by,\,published-date}([Publication])$

Notably, if P is not specified then the resulting collection is just a copy of the input collection c.

3.2 Closure

The closure operator produces a collection through recursively traversing the specified property p over resource elements in c. Therefore the specified p must be a transitive property. We can decide the recursion depth by specified the iteration bound. The transitive closure iteration would end when there are no more paths to follow or a specified iteration limit is reached.

Definition $*_{(p[\max-iteration])}(c)=\{x|x\in p^*(c)\}$

Example 3.2 To query all prerequisite course for Grid research.

$*_{(based-on)}(\{Grid\})$

3.3 Selection

The selection operator evaluates condition expression F over input collection c and return those evaluated true.

Definition $\sigma_F(c)=\{x|x\in C^I(c)\wedge F(x)=True\}$,where F is a combination of logic operators, such as $\wedge,\vee,\neg,\forall,\exists$ and arithmetical operator, such as $>,\geq,<,\leq,=,\neq$ on property values. It has value of True or False.

Example 3.3 To query authors of publications whose at least one of authors is a Professor

$\sigma_{\exists(authored-by=[Professor])}([Publication])$

Example 3.4 Ask for author, publish house and publish date of all publications whose every author is Phd student.

$\pi_{authored-by,published-by,published-date}(\sigma_{\forall(authored-by=[Phd-student])}([Publication]))$

Obviously, it is a composition of two algebraic expressions.

3.4 Join

The Join operator returns a collection including the resource elements in c1 and c2 whose property values matching the specified p.

Definition Join $\triangleright\triangleleft(c1)_p(c2)=\{(x,y)|p(x,y)=True,x\in C^I(c1)\wedge y\in C^I(c2)\}$,p is a join predicate.

Example 3.5 To select Phd students and Master students who have common supervisor.

$\triangleright\triangleleft(p\quad[Phd-student])_{(p.supervised-by=m.supervised-by)}(m\quad[Master-student])$

4 Conclusion and Future Work

In this paper, we presented a knowledge-based data model and algebra to advance the state of the art of resource management in the Grid environment. To our knowledge, it is the first effort towards an algebra for complex resource sharing in the Grid domain. Currently our work is preliminary. There are a lot of works to be continued. In order to test the capacity of proposed model and its algebra, we are implementing the proposed model and query language with Semantic Web techniques. Moreover, we plan to extend the algebra to support inference and analysis on the schema and instance, based on Description Logic[5].At last, algebraic equivalence laws and heuristics would be studied to do query optimization.

References

1. Semantic Grid, http://www.semanticgrid.org/
2. Codd, E.F.,A relational model of data for large shared data banks, Comm. Of the ACM,13(6), June 1970, pp 377-387
3. Michael Halper *et al* Frameworks for incorporating semantic relationships into object-oriented database systems. Concurrency Computat.: Pract. Exper.2003, to appear.
4. Hai Zhuge, A Knowledge grid model and platform for global knowledge sharing, Expert Systems with Applications 22(2002) pp.313-320.
5. Franz Baader *et al*, The Description Logic Handbook --Theory, Implementation and Applications, ISBN: 0521781760, Cambridge University Press, January 2003