XML Based Semantic Data Grid Service

Hui Tan and Xinmeng Chen

Computer School, Wuhan University, Wuhan 430072, China
journal@whu.edu.cn

Abstract. This paper introduces a novel wrapper-mediator based semantic data grid service mechanism to solve the problem of Semantic heterogeneity and few compatible data sources. It uses ontology based semantic information to wrap the heterogeneous data source, and employs mediator structure to supply accessing interface for the data sources, and it extends semantic query, mapping and fusion languages to support semantic grid communication mechanism. The extension of XML algebra with semantic query enhanced is discussed to enable semantic querying on data gird environment.

1 Introduction

Data grid technology is the standard means of realizing the needs of integrating and querying distributed and heterogeneity information, especially semi-structured and non-structured information. However, the studies in data grid technology still have the shortcomings as follows: 1)The flexibility of the grid technology is limited. Taking OGSA-DAI[1] for example, it only supports the limited related database and native XML database. However, most information on Internet comes from web-based semi-structured data environment, such as company web application and XML-based e-commerce platform; furthermore, OGSA-DAI does not have the effective mechanism for other data sources to be integrated into the grid environment. 2) The individual node in the grid environment may exist in varied semantic environment; different data resource is constructed in accordance with different semantic standard. The present data grid does not take into consideration the semantic heterogeneity among different nodes. Many projects are focusing on these two topics. GridMiner[2] and OGSA-WEB[3] are two novel projects focusing on the first one, and DartGrid II[4] and SemreX[5] are excellent projects focusing on the second topic.

This paper focusses on these two topics too. It employs a mediator-wrapper framework to support different information sources and enable semantic information operation on different grid nodes. And it uses XML query style language to retrieve information from different grid nodes, because XML is rapidly becoming a language of choice to express, store and query information on the web. The remainder of this paper is structured as follows. Section 2 gives the general discussion about framework of the mediator-wrapper based semantic data grid. Section 3 discusses the knowledge communication mechanism to support

Y. Shi et al. (Eds.): ICCS 2007, Part I, LNCS 4487, pp. 418–425, 2007.

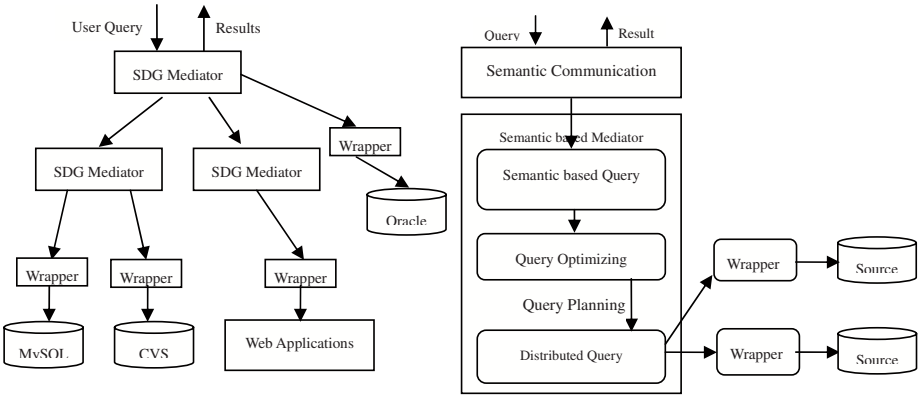
semantic querying and knowledge fusion. Section 4 discusses ontology enabled querying rewriting on XML based grid nodes. Section 5 summarizes the whole paper.

2 Mediator-Wrapper Based Semantic Data Grid Service

Semantic Data Grid (SDG) must satisfy the following requirements:

- The architecture must be opening and compatible with existing standard such as the framework of OGSA[6] or WSRF[7] considering compatible with OGSA-DAI;
- It must provide flexible method for integrating various data sources including relational databases, Native XML databases, or Web based application systems;
- It must support the global semantics to the users who access semantic data grid.

This paper uses a semantic grid adapter service to support semantic operation on the gird. This paper employs a mediator-wrapper method to construct the adapter service, which can be expressed by figure 1(a). The function of the wrapper of local grid nodes is to describe its semantics and its mapping relationship with other nodes, the information source of these nodes include both free and commercial databases, flat files services, web services or web based applications, HTML files and XML files, and the semantic information of every local gird node is described with the language based on its ontology. The mediator node constructs the global semantics of the local nodes, the semantic communication mechanism between the mediator and wrapper nodes is discussed in the following section.



(a) Mediator-wrapper based semantic data grid (b) Semantic Query mechanism with GAV style request

Fig. 1. Mediator-Wrapper based Semantic Data Grid

3 Communication Mechanism with Semantic Grid

It is very important to develop a knowledge communication and coordinating mechanism to support the ontology fusion and semantic query on different data grid nodes. This paper employs a Knowledge Communication and Manipulation Language for Semantic Grid, or KCML for short to support this mechanism, which is an extension of the KGOL[8] language. One function of KCML is to coordinate with each grid node to build the mediator-wrapper architecture dynamically. The other function is to build global knowledge on the mediator and enable semantic query. The communication language is build on SOAP, supporting SOAP over HTTP, HTTPS or other rock-bottom communication protocol. The language could describe as:

$$KCML ::= Ver|Operation|Sender|Receiver|Language|Content.$$

The field *Ver* is for keeping Expanding, showing which version language was used. The new version language has compatibility downwards, supporting the old communication mechanism; *Operation* gives basic communication atom which will be described next; *Content* describes what is communicated; *Sender* defines sender's information, including user, address (such as IP ,e-mail,URL, port); *Receiver* defines receiver's information (usually, receiver should be Web Service or Grid Service), ,including type (HOST, Web Service or Semantic Web Service), address(such as IP address, e-mail, URL, port, if receiver is Web Service, also including service address), identifier; *language* defines which language is used this communication, including RDF/RDFs, DAML+OIL, OWL etc.

3.1 Basic Communication Atom

To illustrate the algorithm, we first define the ontology based knowledge on the mediators and wrappers.

Definition 1. *A knowledge schema is a structure* $KB := (C_{KB}, R_{KB}, I, \iota_C, \iota_R)$ *consisting of (1) two sets* C_{KB} *and* R_{KB}*, (2) a set* I *whose elements are called instance identifiers or instances, (3) a function* $\iota_C : C_{KB} \to \Re(I)$ *called concept instantiation, (4) a function* $\iota_R : R_{KB} \to \Re(I^+)$ *called relation instantiation.*

To simplify the content of this paper, we only discuss the atom of KCML language which support ontology fusion and semantic querying. The atom includes query operation, join operation and union operation etc. as following[9]:

- **Selection.** $\sigma_F(c) = \{x|x \in \iota_C(c) \wedge F(x) = true\}$ where F is composed of logic expression, supporting logic operation $\wedge$, $\vee$, $\neg$, $\forall$, $\exists$, $<$, $>$, $\leq$, $\geq$, $\neq$, $=$ and $\in$. c is concept element of knowledge instance;
- **Join.** $\bowtie (c_1, p, c_2) = \{x, y|x \in \iota_C(c_1) \wedge y \in \iota_C(c_2) \wedge p(x, y) = true\}$, where p is join condition, c_1 and c_2 is concept element;
- **Union.** $c_1 \cup c_2 = \{x|x \in \iota_C(c_1) \wedge x \in \iota_C(c_2)\}$, c_1 and c_2 is the same as above;
- **Minus.** $c_1 - c_2 = \{x|x \in \iota_C(c_1 \wedge \neg c_2)\}$, c_1 and c_2 is the same as above;
- **Projection.** $\pi_P(c) = \bigcup_{p_i \in P} \{y|\exists x, (x, y) \in \iota_R(p_i) \wedge x \in \iota_C(c)\}$, where c is concept element, P is a set of relationship and $P = \{p_1, p_2, \ldots, p_k\}$;

3.2 Semantic Fusion Atom

The mediator node constructs the global semantics of the local nodes based on ontology via ontology fusion mechanism[10] based on the ontology mapping patterns in gird environment, the patterns of ontology mapping can be categorized into four expressions: direct mapping, subsumption mapping, composition mapping and decomposition mapping[11], a mapping can be defined as:

Definition 2. *A **Ontology mapping** is a structure $\mathcal{M} = (\mathcal{S}, \mathcal{D}, \mathcal{R}, v)$, where $\mathcal{S}$ denotes the concepts of source ontology, $\mathcal{D}$ denotes the concepts of target ontology, $\mathcal{R}$ denotes the relation of the mapping and v denotes the confidence value of the mapping, $0 \leq v \leq 1$.*

The KCML language must support the mapping patterns between different semantic nodes on gird, we use **Match** atom to support it, it can be defined as $M(c, d, r) = \{(x, y) | x \in \iota_C(c) \wedge y \in \iota_C(d) \wedge (x, y) \in \iota_R(r)\}$, where c is different concept from d, r is relationship of mapping. The knowledge stored at mediator can be described as the ontology fusion connections list, which can be described as definition 3.

Definition 3. ***Fusion Connection** is a structure $\mathcal{F}_c(O_1 : C_1, O_2 : C_2, \ldots, O_n : C_n, \mathcal{M})$, where C_1 denotes a concept or concept set of ontology O_1, C_2 denotes a concept or concept set of Ontology O_2, $\mathcal{M}$ denotes the mapping patterns between C_1 , C_2 , ... and C_n.*

4 Semantic XML Query Rewriting

The semantic query in a mediator-based SDG can be express as figure 1(b). The user's request is rewritten and modified accordingly based on the global semantics, and is due processed optimally. Corresponding operation plan is made and passed by the wrapper to each data source node for operation. From above description, we know that this paper employs the GAV(Global as View) method to process the user's query[12]. The query can be described as an XML query with semantic enhanced, which can be described as an extension of XML algebra, and it will be discussed in the following subsection. Because common XML query languages such as XQuery and XUpdate can be transferred into XML query algebra, so the extension is manageable.

This paper extended XML algebra TAX[13] to enable semantic querying on mediated gird nodes, TAX uses *Pattern Tree* to describe query language and *Witness Tree* to describe the result instances which satisfy the Pattern Tree. The extension of XML query algebra is discussed in paper [14]. The query planning is based on the semantic XML query rewriting technology. In order to simplify the discussion, this paper just pays attention to the query planning mechanism of the selection operation. Briefly, a selection operation can be expressed as $\sigma(X : S, Y)\ \{X \subseteq P_i \cup P_o, Y \subseteq PE\}$, where P_i is the input pattern tree, P_o is output pattern tree, PE is predication list, S denotes the site in which the query will be executed. We define two operators $\cup$ and $\bowtie$ to represent $Union$ and $Join$

operation separately, and define the operator $\Rightarrow$ to represent the query rewriting operation, and we use $\sigma(X : S_0, Y)$ or $\sigma(X, Y)$ to denote the user's query from the mediator site.

Firstly, we propose how to rewrite pattern tree (which is the X element of expression $\sigma(X, Y)$), there maybe several cases as follows:

1. X is one of the elements of input pattern tree or output pattern tree, and it is also a concept in the global ontology hierarchy. $X_i(1 \leq i \leq n)$ are the concepts for different local ontologies. X and X_i were combined into one concept in the integrated global ontology with strong direct mappings, which means that X and X_i can match each other, then we can rewrite X as $X \cup \bigcup_{1 \leq i \leq n} X_i$. The responding selection rewriting can be expressed as:

$$\sigma(X, Y) \Rightarrow \sigma(X, Y) \cup \sigma(X_1 : S_1, Y) \cup \sigma(X_2 : S_2, Y) \ldots \cup \sigma(X_n : S_n, Y) \quad (1)$$

2. The concept of X is generated by the subsumption mapping or composition mapping of $X_i(1 \leq i \leq n)$, then we can rewrite X as $\bigcup_{1 \leq i \leq n} X_i$. The responding selection rewriting can be expressed as:

$$\sigma(X, Y) \Rightarrow \sigma(X_1 : S_1, Y) \cup \sigma(X_2 : S_2, Y) \ldots \cup \sigma(X_n : S_n, Y) \quad (2)$$

And then, we propose how to rewrite the predication expressions (which is the Y element of the expression $\sigma(X, Y)$, there are also several cases, which can be described as follows:

1. If there are lots of concept $Y_i(1 \leq i \leq n)$ combined in the concept Y of global Ontology, we can rewrite Y as $Y \cup \bigcup_{1 \leq i \leq n} Y_i$. The corresponding selection rewriting can be described as:

$$\sigma(X, Y) \Rightarrow \sigma(X, Y) \cup \sigma(X : S_1, Y_1) \cup \sigma(X : S_2, Y_2) \ldots \cup \sigma(X : S_n, Y_n) \quad (3)$$

2. If the concept Y is generated by the subsumption mapping of $Y_i(1 \leq i \leq n)$, we can rewrite Y as $\bigcup_{1 \leq i \leq n} Y_i$. The corresponding selection rewriting can be described as:

$$\sigma(X, Y) \Rightarrow \sigma(X : S_1, Y_1) \cup \sigma(X : S_2, Y_2) \ldots \cup \sigma(X : S_n, Y_n) \quad (4)$$

3. If the concept Y is generated by the composition mapping of $Y_i(1 \leq i \leq n)$, suppose the composition condition is F, we can rewrite Y as $(Y_1 + Y_2 + \ldots Y_n) \cap F$. The corresponding selection rewriting can be described as:

$$\sigma(X, Y) \Rightarrow \sigma(X : S_1, Y_1 \wedge F) \bowtie \sigma(X : S_2, Y_2 \wedge F) \ldots \bowtie \sigma(X : S_n, Y_n \wedge F) \quad (5)$$

Algorithm 1. SEL_Rewrite_X(X)

Input: X is the pattern tree of selection query $\sigma(X, Y)$.
1 **foreach** $x \in X$ **do**
2 **switch** *Mappings of* X *node* **do**
3 **case** $funsion\_node$
4 $x \leftarrow x \cup \bigcup_{1 \le i \le n} x_i$;
5 $\sigma(X, Y) \Rightarrow \sigma(X, Y) \cup \sigma(X_1, Y) \cup \sigma(X_2, Y) \ldots \cup \sigma(X_n, Y)$;
6 **foreach** x_i **do**
7 SEL_Rewrite_X(x_i);
8 **end**
9 **case** *subsumption or composition*
10 $x \leftarrow \bigcup_{1 \le i \le n} x_i$;
11 $\sigma(X, Y) \Rightarrow \sigma(X_1, Y) \cup \sigma(X_2, Y) \ldots \cup \sigma(X_n, Y)$;
12 **foreach** x_i **do**
13 SEL_Rewrite_X(x_i);
14 **end**
15 **end**
16 **end**
17 **end**

It is worth to point out that rewriting process may require a recursion in the transitivity property of semantic mapping. The process of rewriting pattern tree and predication expressions can be described as algorithm 1 and 2.

The query planning is a sequence, each node of the sequence can be denoted as $P_n = (Q_n, S_n, C_n, F_n)$, where Q_n is the query which is needed to rewrite, S_n is a set of sub query executed on different sites, C_n denotes the connection operator, in most time, it is $\cup$ or $\bowtie$ operator, F_n is the predication which denotes the connection conditions. P_n represents the query rewriting procedure of query Q_n. The query planning procedure of user's query $\sigma(X, Y)$ can be expressed in algorithm[14].

5 Discussion and Conclusion

Semantic data grid service mechanism we present in this paper wrapped various information source through ontology semantic, and used Mediator-Wrapper to support the heterogeneous data source, employed mediator structure to realize virtual data gird service which supports semi-structured information retrieving language. The extension of XML algebra with semantic query enhanced and semantic grid communication mechanism are also discussed to enable semantic accessing on data grid environment. However, query optimizing in distributed web sites and the capability of different nodes and network were not considered in the query planning mechanism discussed in this paper, future research will be focused on this topic.

Algorithm 2. SEL_Rewrite_Y(Y)

Input: Y is the predication list of selection query $\sigma(X, Y)$.

foreach $y \in Y$ **do**
  switch *Mappings of* Y *concept* **do**
    case *funsion_node*
      $y \leftarrow y \cup \bigcup_{1 \le i \le n} y_i$;
      $\sigma(X, Y) \Rightarrow \sigma(X, Y) \cup \sigma(X, Y_1) \cup \sigma(X, Y_2) \ldots \cup \sigma(X, Y_n)$;
      foreach y_i **do**
        SEL_Rewrite_Y(y_i);
      end
    case *subsumption*
      $y \leftarrow \bigcup_{1 \le i \le n} y_i$;
      $\sigma(X, Y) \Rightarrow \sigma(X, Y_1) \cup \sigma(X, Y_2) \ldots \cup \sigma(X, Y_n)$;
      foreach y_i **do**
        SEL_Rewrite_Y(y_i);
      end
    case *decomposition*
      $y \leftarrow (y_1 + y_2 + \ldots y_n) \cap F$;
      $\sigma(X, Y) \Rightarrow \sigma(X, Y_1 \wedge F) \bowtie \sigma(X, Y_2 \wedge F) \ldots \bowtie \sigma(X, Y_n \wedge F)$;
      foreach y_i **do**
        SEL_Rewrite_Y(y_i);
      end
    end
  end
end

Acknowledgment

This work was partially supported by a grant from the NSF (Natural Science Fundation) of Hubei Prov. of China under grant number 2005ABA235, and it was partially supported by China Postdoctoral Science Foundation under grant number 20060400275 and Jiangsu Postdoctoral Science Foundation under grant number 0601009B.

References

1. Antonioletti, M., Atkinson, M., Baxter, R., et al.: The design and implementation of Grid database services in OGSA-DAI. Concurrency and Computation: Practice and Experience **17** (2005) 357–376
2. Wöhrera, A., Brezanya, P., Tjoab, A.M.: Novel mediator architectures for Grid information systems. Future Generation Computer Systems **21** (2005) 107–114
3. Pahlevi, S.M., Kojima, I.: OGSA-WebDB: An OGSA-Based System for Bringing Web Databases into the Grid. In: Proceedings of International Conference on Information Technology: Coding and Computing (ITCC'04), IEEE Computer Society Press (2004) 105–110

4. Chen, H., Wu, Z., Mao, Y.: Q3: A Semantic Query Language for Dart Database Grid. In: Proceedings of the Third International Conference on Grid and Cooperative Computing (GCC 2004), Wuhan, China, LNCS 3251, Springer Verlag (2004) 372–380
5. Jin, H., Yu, Y.: SemreX: a Semantic Peer-to-Peer Scientific References Sharing System. In: Proceedings of the International Conference on Internet and Web Applications and Services (ICIW'06), IEEE Computer Society Press (2006)
6. Foster, I., Kesselman, C., Nick, J.M., Tuecke, S.: Grid Services for Distributed System Integration. IEEE Computer **35** (2002) 37–46
7. Czajkowski, K., Ferguson, D.F., Foster, I., et al.: The WS-Resource Framework. http://www.globus.org/wsrf/specs/ws-wsrf.pdf (2004)
8. Zhuge, H., Liu, J.: A Knowledge Grid Operation Language. ACM SIGPLAN Notices **38** (2003) 57–66
9. Sheng, Q.J., Shi, Z.Z.: A Knowledge-based Data Model and Query Algebra for the Next-Gereration Web. In: Proceedings of APWeb 2004, LNCS 3007 (2004) 489–499
10. Gu, J., Zhou, Y.: Ontology fusion with complex mapping patterns. In: Proceedings of 10th International Conference on Knowledge-Based, Intelligent Information and Engineering Systems, Bournemouth, United Kingdom, LNCS, Springer Verlag (2006) 738–745
11. KWON, J., JEONG, D., LEE, L.S., BAIK, D.K.: Intelligent semantic concept mapping for semantic query rewriting/optimization in ontology-based information integration system. International Journal of Software Engineering and Knowledge Engineering **14** (2004) 519–542
12. Levy, A.Y., Rajaraman, A., Ordille, J.J.: Query heterogeneous information sources using source descriptions. In: Proceedings of the 22nd VLDB Conference, Mumbai, India, Morgan Kaufmann Publishers Inc (1996) 251–262
13. H.V.Jagadish, L.V.S.Lakshmanan, D.Srivastava, et al: TAX: A Tree Algebra for XML. Lecture Notes In Computer Science **2379** (2001) 149–164
14. Gu, J., Hu, B., Zhou, Y.: Semantic Query Planning Mechanism on XML based Web Information Systems. In: WISE 2006 Workshops, LNCS 4256 (2006) 194–205