SDDG: a P2P-Semantic-Grid Enabled Information Sharing Infrastructure

Jingtao Zhou, Haicheng Yang, Mingwei Wang
The Key Laboratory of Contemporary Design and Integrated Manufacturing Technology, Ministry of Education, Northwestern Polytechnical University, Xi’an, China
Jtzhou.office@gmail.com
doi:10.4156/jnit.vol2. issue1.4

Abstract

SDDG (semantic desktop data grid) is a novel semantic peer-to-peer architecture built on top of previous work [1][2] to fully leverage the enterprise information resides in both enterprise and personal information sources. In SDDG, the combination of P2P data integration paradigm, semantic integration principle and grid service oriented framework provides both the required level of flexibility and scalability, and the high interoperability on information and semantic level. The semantic P2P model of SDDG is discussed in detail from the aspects of the definition of peer node, the master-slave P2P mapping mechanism, the classified P2P query. A semantic service oriented implementation framework based on OGSA-DAI is also discussed in this paper.

Keywords: Semantic Desktop, Semantic P2P, Service Oriented Framework

1. Introduction

Today’s enterprise information and knowledge pervade everywhere of the enterprise rather than only the enterprise-wide databases. This is especially true for that more and more knowledge achieved by individual may be stored into personal desktop besides common data sources in enterprise. To accomplish a complex task, knowledge workers need achieve any relative information not only from common data sources but also individual data sources by any possible way. However, the continuously increasing information and knowledge in personal computers managed by individual employees is neglected although enterprises have tackled information sharing inside their specific domains for more than a decade. Integrating and sharing all potentially useful information within entire enterprise, it must be of great benefit to both individual and group in collaborative work, research, even business decision and action. Unfortunately, current enterprise information infrastructure is poorly suited for dealing with the continuing, rapid explosion in data both in common and personal information sources.

We give a survey of related works of data integration on Grid, p2p data integration and semantic desktop in section 2. Section 3 discusses the architecture of SDDG. Section 4 and 5 discuss the basic principles for SDDG from P2P perspective and semantic grid perspective respectively. Section 6 gives the concluding remarks and future perspectives.

2. Related works

2.1. Data Integration on Grid

In the context of information integration and sharing, Grid technologies distinguish current information integration technologies in enterprise (eg. federated systems, data warehouse, etc.) by providing not a generic approach but also an open and standard-based infrastructure. Using of the proposed Grid technologies (Open Grid Services Architecture Data Access and Integration (OGSA-DAI) [1], OGSA-DQP [4] etc.) is becoming popular for standard-based access of heterogeneous resources. Some industry products, such as IBM WebSphere Information Integrator V8.2 has supported OGSA-DAI by implementing a Grid wrapper [5].

The current efforts of the Data Grid community mainly concentrate on providing a global, uniform access methodology for all database resources. However, the functional level
integration way of Grid-based Virtual Databases\[6\] has limited the exploitation of Data Grids in many real situations\[7\]. This motivates information grid projects to shift the emphasis on information integration and mediation. Moreover, the emerging of Semantic Grid is beginning to take this further, from information to semantic or knowledge. Some projects, to some extent, such as COG \[8\] and Dart-Grid \[9\] explore this trend in the context of information integration.

The COG project aims to integrate disparate data sources on semantic level by using a central Information Model (i.e. ontology). However, although COG means “Corporate Ontology Grid”, it does not seem to intend to use general Grid technologies. In essence, it is a solution following an ontology-based information integration approach \[10\]. Compared to COG, Dart-Grid is an OGSA-based Database Grid originally motivated by the application of web-based data sharing and database integration for Traditional Chinese Medicine. In particular, data sources integrated by Dart-Grid are mainly databases, other data sources such as documents, and data sources that stream data in real or pseudo-real time from applications are not supported by current Dart-Grid. Furthermore, details of some crucial issues concerned by enterprise, such as security, authorization, transaction, etc., are not addressed.

Above all, semantic interoperation based on single ontology such as in COG and Dart-Grid may fall short where one information source has a different view on a domain, or information sources are on changing \[10\].

In essence, both COG and Dart-Grid still follow the traditional centralized data integration architecture and rely on the notion of global schema. In this way, it is difficult to achieve semantic interoperability across personal and corporate boundaries, realize large-scale deployment and dynamic information sharing like personal information sharing. So, although semantic grid could provide an appropriate integration and operation infrastructure for enterprise information as discussed in our previous SGII approach \[11\], whether it is really adapted to dynamic enterprise integration environment that still requires substantial research. Fortunately, P2P technology has much to offer for this scenario.

2.2. P2P Data Integration

Differently from the traditional centralized data integration approach, peer-to-peer data integration architecture achieves information integration by establishing peer-to-peer mappings rather than using centralized schema. In order to solve the semantics problem pointed out in \[12\], many P2P data integration systems \[13\][14] provide semantic mappings between peers or super peers. Although P2P data integration doesn’t need to design and maintain a global schema, an ontology or vocabulary may be still necessary in some systems, such as the global RDF ontology in \[14\]. In essence, in spite of the using of single ontology or the same domain assumption in \[13\], it implies that all peers belong to the same domain and may encounter the same problem discussed in the end of section 3. In addition, since the hybrid architecture of P2P data integration is analogy with the hybrid architecture described in \[10\], it may have the analogous drawback that existing semantic representations of peers can not easily be reused.

Furthermore, P2P data integration by itself is insufficient to be a pragmatic architecture for EII (enterprise information integration), especially for information integration across several enterprises because P2P developers have worked mainly on vertically integrated applications \[15\], failed to define common protocols, standardized infrastructures for interoperability and enough control of data sources, which, on the contrary, are just the consideration of grids.

2.3. Semantic Desktop

Recently, some new research topics including semantic desktop based on semantic web technologies (e.g. Gnowsis \[16\], Haystack system at MIT \[17\], NEPOMUK \[18\]), and network semantic desktop based on both Semantic Web and P2P (e.g. \[19\]) are proposed, both topics of which are attempt to integrate desktop applications and the web to improve personal information management and collaboration. The semantic web centric approaches like \[16\][17][18] focus on personal data annotation and management using relative technologies.
such as RDF, whereas, the networked semantic desktop visions a new group collaboration infrastructure, which, in other works, concentrates on personal data sharing and succeeded personal collaboration apart from the management of personal data.

Differently with proposed approaches such as above, SDDG focuses on not only organizing and managing information on personal computers but also sharing and integrating information between personal and organizational data sources. SDDG aims to provide a pervasive information sharing infrastructure for any data residing in entire enterprise, in spite of organizational or personal data. In particular, since some data on personal computers inevitably involves some specific organizational domains of an enterprise, not all information on personal computers is symmetric. In other words, there exist, from logic perspective, some hierarchically arranged personal data sources in SDDG environment. Likewise, since some information on personal computers in an enterprise may also includes some sensitive individual or group data, corresponding structured robust security services are needed in SDDG besides intuitive trust mechanisms as used in P2P networks. In order to create a more standardized, generic, and pervasive infrastructure for entire enterprise information integration, we have proposed a Semantic Grid based architecture called ESD [1], which is built on top of OGSA and aims at providing a loosely-coupled semantic interconnection environment for both organizational and personal information. By fitting into existing Grid, Semantic Web and Semantic Web Services architectures (visions of Semantic Grid), ESD obtains more flexible and open characteristics than existed systems. However, whether this architecture is adapted to personal information sharing scenario requires substantial research since its inherited infrastructure from grid is not focused on a continuously changing information sharing environment in which sources may join and leave at any time. P2P systems, on the other hand, have proved their abilities of dealing with intermittent participation and highly variable behavior but failed to define common protocols and standardized infrastructures for interoperability [15]. In order to achieve a ubiquitous and pervasive information sharing platform for personal and corporate data sources, we believe such a complex application need the convergence of peer-to-peer and Grid (or semantic grid) computing as Foster has stated [20]. Undertaking the intersection fulfillment of these technologies in information sharing area will address scalability, selfadaptation, and failure recovery, while, at the same time, providing a persistent and standardized infrastructure for interoperability [8]. In this context, we extended our previous work ESD (enterprise semantic desktop)[1][2] by reflecting principles of P2P data integration in semantic grid., which takes the same semantic-grid-based framework of SGII [11][21] as foundation, to a p2p-semantic-grid enabled information sharing infrastructure (SDDG, Semantic Desktop Data Grid) in the work[22] by reflecting principles of P2P data integration in semantic grid environment. SDDG attempts to wrap both corporate and personal information sources as semantic grid services and create a semantic information interoperability environment following a p2p data coordination way.

3. Architecture of SDDG

From the point of view of information integration, there are two basic problems to be solved in a SDDG solution: dealing with disparate information content itself and dealing with disparate applications and interfaces to that data. So, The SDDG solution could be divided into two parts: 1) information content level integration, 2) system level integration.

The former integration concentrates on how to describe, explain, integrate, associate and infer some corresponding personal and group data, and how to realize the interoperability on information content level.

The latter aspect concern about how to locate and connect system, which we depend on to realize the interoperability among systems. System integration does not influence information integration directly but decides the access and connection capacity of information integration fundamental facilities and impacts on the scope, environment, performance, and scale of the entire information integration.
SDDG establishes P2P semantic integration at content level: with the adoption of high dynamic and distribution P2P technology while endow it with semantic interoperability, define multi-granularity semantic description. Through fusing semantic information integration idea into P2P data integration technology, define semantic mapping relationship among different kinds of peers, dispersing control and centralization control, and semantic interoperability in group and personal data, realize semantic-based dynamic loosely coupled integration.

SDDG establishes semantic grid service based integration at system level. The architecture introduces P2P data integration paradigms into semantic grid environment, is organized by constituents on different peers and grouped by functionality services. Since OGSA Grids is a suitable infrastructure for P2P computing [7], SDDG is built on top of OGSA to provide generic system architecture for semantic integration of information. SDDG sets P2P semantic interoperability on OGSA-DAI and transforms each peer into semantic grid service, and transform interoperability between peers into interoperability of semantic grid service, realize mapping from semantic P2P based logic integration to semantic grid service based physical integration, as well as service oriented dynamic loosely coupled integration system.

Combine the above two: realize loosely coupled information integration at information content level and system level, set up semantic-based dynamic semantic desktop infrastructure, and realize large scale dynamic integration and semantic interoperability.

4. Principles from P2P Perspective

From the P2P point of view, SDDG is a P2P information integration system \( I \) composed of a set of instances of two kinds of peers: data peer \( (dp) \), domain peer \( (dop) \), each of which has a schema describing the data held by the peer, and a set of P2P mappings that specify the semantic relationships with the data exported by other peers. Instances of each kind of peer construct corresponding sets called \( \text{DPSet} \) and \( \text{DoPSet} \) respectively. The formal framework of SDDG is \( I := (\text{DPSet}, \text{DoPSet}, F_{DP-DoP}, F_{DoP}) \), where \( F_{DP-DoP} : \text{DPSet} \rightarrow \text{DoPSet} \) represents P2P mappings between \( dp \in \text{DPSet} \), and \( dop \in \text{DoPSet} \); if \( \exists (dp, dop) \in F_{DP-DoP} \), it means that a P2P mapping between \( dp \) and \( dop \) exists; \( F_{DoP} \subseteq \text{DoPSet} \times \text{DoPSet} \) - \( I_{DoP} \). \( F_{DoP} \) represents P2P mappings between \( dop \), \( I_{DoP} = \{<x, x>| x \in \text{DoPSet}\} \).

To avoid the defects caused by P2P integration when it allows arbitrary network of relationships between peers [26], SDDG divides the entire information areas into more domains by domain peer. Each domain peer schema is regarded as a global ontology to associate the corresponding group data and personal data wrapped by data peer. Through the P2P mapping between domain peers, SDDG achieves further large-scale cross-domain integration of information.

4.1. Data Peer

Each data peer \( dp \) behaves as a control point of a group (enterprise) or personal data source, which is defined as a tuple \( dp = (S_{\text{dmax}}, D, M_{D_{Dop}}) \), where

- \( S_{\text{dmax}} \) is the set of shared data of the data source.
- \( D \) is the peer schema held by \( dp \) to represent the intensional description of the data controlled by \( dp \). \( D \) can be regarded as an export schema for \( S_{\text{dmax}} \) from data source.
- \( M_{D_{Dop}} := 2^D \rightarrow 2^{D_o} \), \( M_{D_{Dop}} \) is a P2P mapping between \( D \) and the peer schema \( D_o \) held by domain peer to connect \( dp \) and \( dop \) following master-slave P2P mapping paradigm as discussed in section 4.3.

4.2. Domain Peer

Each domain peer \( dop \) is defined as a tuple \( dop = (D_o, M^{D_o}) \), where
- Do is the peer schema or ontology held by dop to represent the intensional knowledge of a specific domain in enterprise. For most enterprises, there usually exist some well-defined set of meta-data standards for specific domains, which can be used as initiate part of Do.

- \( M^\text{Do} \) is a set of P2P mappings that express semantic relationships between two or more elements of ontologies or peer schemas of DoPs. Each \( M^\text{Do} \in M^\text{Do}, M^\text{Do} := 2^{dop^*} \rightarrow 2^{dop}, Dop \in dop, Dop^* \in dop^*, Dop \neq dop^* \).

We use Do as the common semantic model or ontology agreement for a (sub-) community. By Do and \( M^\text{Do} \), each dop works as a mediator for a group of dps. Therefore, dop and relative dps usually belong to the same (sub-) community in logic. The P2P network formed by connected DoPs creates an alignment between different DoPs' schemas, which is actually used to represent the intersection and semantic relationships among ontologies of different (sub-) communities. The existence of Do makes the establishment of the consensus of data sources in a relatively small area possible. It is not only to avoid the existence of a direct one to one mapping between dps, reduce the difficulty of cross-cutting integration, and improve adaptation of integration, but also to allow achieving a higher level of semantic integration.

4.3. Master-slave P2P mapping

Currently in the field of information integration, whether using a single global schema (including the global schema or global ontology) or using multiple schemas (such as federal schemas), or P2P data integration, the approaches of mapping between schemas can be categorized into three basic ways \[23\]: GAV (Global As View) approach, LAV (Local As View) approach and GLAV (Global-Local As View) approach.

However, in spite of GAV or LAV / GLAV mappings, there are critical gaps in the following when directly applying them to SDDG system.

1) Currently, mapping principle of GAV, LAV and GLAV is built on top of conjunction based view mapping principle\[23\], and not enough to describe the semantics of complex data integration\[24\].

2) The basic principle of these mapping patterns requires query posed on global schema must rely on the specific query rewriting calculation process and data translation steps \[23\] to translate this query into query posed on local schema, which can not support complex semantic rich query, limited reasoning ability of which is not enough to get all possible results;

3) Currently, GAV, LAV and GLAV mapping patterns all consider data integration and exchange issues in a common domain, not for multiple domains, which are lack of researches on how to establish semantic mapping across multiple domains with different semantic context. Some large-scale P2P integration systems using the GLAV mapping pattern assume that integration only occurs in one domain to avoid semantic mappings across multiple domains. So these systems can not be adapted to actual enterprise information integration environment.

To solve the above problems, we suggest a "master-slave" type of P2P semantic mapping.

Compared with traditional GAV / LAV / GLAV mapping patterns, master-slave P2P semantic mapping expands in the aspects of the mapping principle (one to one mapping), mapping content (semantic based mapping), and mapping results description.

As shown in Figure 1, master-slave mapping is the mapping of the two parties to participate, related content of "slave side" is described through the semantic view of "master side". Its basic principle consists of two parts: ① slave side initiates mapping; ② establish mapping based on semantic view.

Be noted that, "master-slave" is relative, not absolute. Peer node in a P2P mapping can serve as "master" side role, while in another P2P mapping as "slave " side role. It is consistent with the basic spirit of P2P.

1) Slave side initiates mapping

In order to provide the independence required by P2P mapping, mapping creation is initiated from slave node. That is, slave node considers how to describe the elements contained in slave
node using the elements of master node. Any change of the mapping initiator (slave side) will not affect the mapping reference side (master side). Changes of mapping arising from the changes of master will also be maintained by the mapping initiator. Therefore, slave side is responsible for the establishment and maintenance of the mapping. Mapping between master and slave is loosely coupled.

Figure 1. Basic principle of master-slave mapping

2) Mapping establishment based on the semantic view

As shown in Figure 1, the P2P mapping between peer nodes is built by establishing the mapping between the semantic views of the peer schemas. Semantics of slave peer node schema \( N \) is represented by the semantics of master peer node schema \( N^* \). Slave node schema’s parts involved in mapping form the semantic view \( V \), so does the semantic view \( V^* \) for master node. So, the mapping between peer nodes (e.g. Node \( N \) and node \( N^* \) shown in fig 5-1) is translated into semantic mapping between corresponding views: \( 2^V \rightarrow 2^{V^*} \). Intuitively, when the view \( V \) \((V^*)\) is the node model \( N \) \((N^*)\), the map is converted to \( 2^N \rightarrow 2^{N^*} \) \((2^V \rightarrow 2^{V^*})\). If all elements of the two peer models are likely to participate in mapping (but not required), the mapping can be expressed as \( 2^V \rightarrow 2^{V^*} \). Clearly, both \( 2^V \rightarrow 2^{V^*} \), \( 2^V \rightarrow 2^{V^*} \) and \( 2^V \rightarrow 2^{N^*} \) are special circumstances of \( 2^V \rightarrow 2^{V^*} \), so the mapping between peer nodes can be represented as a uniform expression \( 2^V \rightarrow 2^{V^*} \).

Compared with the existing mapping patterns, master-slave P2P mapping has the following characteristics:

First, the master-slave mapping only need to consider 1-1 mapping. The establishment and maintenance of mapping are all taken in charge by the mapping initiator. So, the master-slave mapping has more flexible modularity and loosely coupled properties, being consistent with the characteristics of P2P mapping.

Secondly, mapping principle based on semantic view mapping \((2^V \rightarrow 2^{V^*})\) supports to establish any semantic mapping between any semantic elements (concepts, attributes, relationships, and their logic operations). In principle, the master-slave mapping allows any logical language to describe the semantic mapping. It just chooses the appropriate modeling method and description language for semantic mapping to support rich semantic retrieval and reasoning.

For more detail about master-slave mapping, please refer to [27].

4.4. Classified Query

By reflecting the SGII query principle [21] in SDDG, SDDG establishes two kinds of queries.
- D-Query, query posed on the peer schema $D$ of $dp$. Locally, the query will be decomposed into sub-queries on relative data source controlled by this $dp$ according to the shared $S_{dmin}$ to get answer from local source. This local query process is called D-L-Query. Meanwhile, the source query is reformulated into a target query over the connected $dop$ through $dp$-$dop$ semantic mediation process based on the mapping $M_{D_{dp}}$. This query transfer process is called D-Do-Query. The D-Do-Query will trigger another kind of query, Do-Query.

- Do-Query, query posed on the peer schema $Do$ of $Dop$. Do-Query involves the relative $dps$ and $dops$ in SDDG. In this context, the P2P network composed of $dps$ and $dops$ can be looked as a super-peer network, where $dop$ is super peer and Do-Query is the query on super peer. First, the Do-Query is rewritten into sub-queries on $dps$ (called Do-D Query) through $M_{DD_{dp}}$ and sub-queries on neighboring $dops$ (called Do-Do Query) through $dop$ semantic mediation process based on $M_{D_{do}}$. Then, the Do-Query will trigger the D-L-Query process in each relative $dp$ to get answer from data sources while the Do-Do Query will start new Do-Query process (es) on the corresponding $dops$.

5. Principles from Semantic Grid Perspective

5.1. Grid service oriented framework

From the grid point of view, SDDG is a semantic grid service oriented architecture (as shown in figure 2), i.e., every peer in SDDG is independently realized as a grid service based on the Open Grid Service Infrastructure and will be a semantic grid service by semantic enrichment. In particular, each peer service, by now, is designed as OGSA-DAI grid data service (GDS) [3] compatible by extending the GDS port types of OGSA-DAI, and enriched by describing the corresponding P2P mapping relations using the metadata (represented using WSML) of GDS. Besides data access, OGSA-DAI enables some data integration functions by providing facilities for combining or transforming data from multiple data access components through the document-oriented interface of GDS. As a result, each P2P operation will be described as the activity of perform documents [3] of GDS and support complex P2P interactions by the composition of activities.

As a GDS, each peer service in SDDG implements the port types GDS and GDT (Grid Data Transport) from OGSA-DAI, and the Grid Service port type from OGSAG. Other port types from OGSA, such as NSnk (NotificationSink) and NSrc (NotificationSource) can also be optionally implemented according to the requirements. Each peer’s schema, relative P2P mappings and data service descriptions are described in the metadata (represented using WSML) of corresponding peer service.

Therefore, we alter our previous definition of peer services in [25] as follows:
- The Mapping (M) port type, which defines a facility for the P2P mapping between peers by interacting with the interfaces and tools provided by WSMT. Previous GAV/LAV mapping port types [25] are replaced by unified Mapping port type.
- The GetMapping (GM) port type, which is used to get the P2P mapping metadata from Peers.
- According to the different kinds of queries in SDDG, each kind of peer defines its own Query (Q) port type rather than GAV/LAV Query port types defined in [25] to perform queries posed on it. The Query port type accepts conjunctive queries as input.

In the context of data service description, the semantics of peer schema is a more powerful criterion than the similar port types, so we implement peer service as semantic grid service by enriching it using its peer schema itself. We can discover the peer service through the query posing on relative peer schemas, such as find the $dp$ service by the query on $dop$ schema, which may have direct or transferable connection with the $dp$ service expressed by P2P mappings.
5.2. Query implementing based on peer services

Since each peer service is a GDS, its instances are created based on the same mechanism of GDS instances by invoking the GDSF (Grid Data Service Factory) [3], whereas the DAISGR (Database Access and Integration Service Group Registry) [3] allows clients to search for GDSs and GDSFs. Here, the clients can be either semantic desktop application (e.g. semantic desktop in figure 2) or peers that need invoke sub-peers. In the context of query, it means that the originating peer is responsible for the instantiation of neighboring peers involved in the query process. Therefore, from the service point of view, a query process in SDDG is a process of peer services instantiation and interaction through the Query port types along with the query reformulation. In brief, the instantiation mechanism of peer service provides an easy way to implement query process based on services for SDDG.

6. Conclusions

Fusion-applying Semantic Grid and p2p data integration paradigms on desktop grid network is expected to effectively organize, share, cluster, fuse, and manage all the industrial information resources including both organizational and personal information sources, and enables new integrative group collaboration. By fitting into existing Grid, P2P and Semantic Grid Services architectures, SDDG obtains more flexible and open characteristics. To achieve pervasive semantic interoperability each resource is enriched by machine-understandable description and reconstructed as a semantic-grid enabled information repository. P2P service-oriented backbone and semantic information organization based on share ontologies construct a loosely-coupled operation environment for data and service. The application based on Semantic Grid and P2P frameworks for provision of effective information resource sharing in and across industrial enterprises, both on organizational and personal level, is promising. However it needs
more industrial case studies to be carried out in order to evaluate finally the utility of information and semantic interoperability based on SDDG.

7. Acknowledgment

The work is supported by China Postdoctoral Science Foundation through approval no. 20090450169, by National 863/CIMS of China through approval no. 2009AA04Z122, and by National Natural Science Foundation of China through approval no. 70931004.

8. References


