A Novel Heterogeneous Data Integration Approach for P2P Semantic Link Network

Hai Zhuge
China Knowledge Grid Research Group
Key Lab of Intelligent Information Processing
Institute of Computing Technology, CAS
+86-10-62562703, Beijing, 100080, China
zhuge@ict.ac.cn

Jie Liu
China Knowledge Grid Research Group
Key Lab of Intelligent Information Processing
Institute of Computing Technology, CAS
+86-10-62562703, Beijing, 100080, China
liujie@computer.org

ABSTRACT
This paper proposes a novel approach to integrate heterogeneous data in P2P networks. The approach includes a tool for building P2P semantic link networks, mechanisms for peer schema mapping, criteria for peer similarity degree measurement, and algorithms for heterogeneous data integration. The approach has three advantages: First, it uses semantic links to describe semantic relationships between peers' data schemas. Second, it deals with the semantic heterogeneity, the structural heterogeneity and the data value inconsistency. Finally, it considers the semantic similarity and structural similarity to forward queries to relevant peers.

Categories & Subject Descriptors: H.3.3
[**Information Storage and Retrieval**]: Information Search and Retrieval – *query formulation*, *search process*, *selection process*.

General Terms: Algorithms, Management, Measurement.

Keywords: Data Integration, P2P Computing, Semantic Link, Semantic Web.

1. INTRODUCTION
Data integration in Peer-to-Peer systems is a challenging issue due to its heterogeneous and autonomous characteristic [2]. An approach to answer queries in P2P networks was proposed in [1]. However, the approach in [1] only considers attribute semantics of peers. The proposed approach is based on our previous work on semantic links, soft-devices and semantic Web service integration [3, 4, 5].

2. GENERAL ARCHITECTURE
Figure 1 denotes an overview of P2P semantic link networks, where each peer is an active and intelligent soft-device [4]. Peers can dynamically establish connection and provide data and services to each other based on basic communication mechanisms. *Semantic Links* are used to specify semantic relationships locally between peers [3]. A P2P *Semantic Link Network* (SLN) is a directed network, where nodes are peers and edges are typed semantic links. A component-based tool for making semantic links between peers has been developed. When a peer enters into or leaves a P2P semantic link network, the relevant semantic links can be automatically established according to a set of semantic reasoning rules.

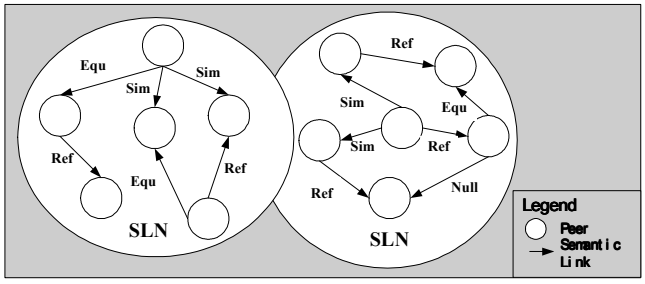

Figure 1. An overview of P2P semantic link networks.

As depicted in Figure 2, a peer in a P2P semantic link network mainly has two modules: a communication module and a data management module. Users can query a peer through GUI or SSeIQL — an SQL-like query language designed for data management. Queries between peers are through SOAP messages.

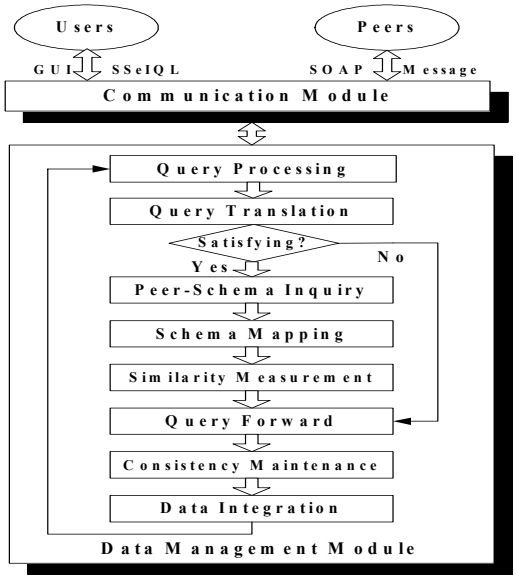

Figure 2. An overview of the proposed data integration approach.

Upon receiving a query, a peer will first check whether it can satisfy the requirement. If not, it will forward the query to its successors, who are likely to answer the query or forward the query further. Otherwise, the peer performs the following tasks:

WWW 2004, May 17-22, 2004, New York, New York, USA.
ACM 1-58113-912-8/04/0005.

Step 1. *Peer-Schema Inquiry*. To obtain peer schemas (i.e., the XML schema of the successors) by sending *Schema Inquiry* messages.

Step 2. *Schema Mapping*. To build node mapping, path mapping and tree mapping between the current schema and peer schemas.

Step 3. *Similarity Measurement*. To measure semantic similarity and structural similarity between peers so as to select appropriate successors to forward the query.

Step 4. *Consistency Maintenance*. To detect inconsistent data in returned data flows.

Step 5. *Data Integration*. To integrate relevant data satisfying query requirements to provide users with a single semantic image data usage mode.

3. P2P SEMANTIC LINK NETWORK MODEL

In a P2P semantic link network, a semantic link between two peers P_i (predecessor) and P_j (successor) can be denoted as $P_i \xrightarrow{\alpha} P_j$, where α is one of the following types:

(1) *Equal-to Link* — Semantics of P_i is equal to that of P_j;

(2) *Similar-to Link* — Semantics of P_i is similar to that of P_j;

(3) *Reference Link* — Semantics of P_i refers to that of P_j;

(4) *Empty Link* — No semantic relationships between P_i and P_j;

(5) *Null Link* — Uncertain semantic relationships between P_i and P_j.

We can chain relevant semantic links to obtain uncertain semantic relations between peers according to a set of reasoning rules [3].

4. DATA INTEGRATION

4.1 Peer Schema Mapping

Upon receiving peer schemas through SOAP messages, a peer will traverse the schemas recursively in the depth-first order to extract schema information. To solve the semantic heterogeneity, each node in a peer schema is associated with a semantic attribute set (i.e., a set of semantically related terms). Structural heterogeneity is solved through: (1) *Node Mapping* — To map nodes of the current schema into those of peer schemas; (2) *Path Mapping* — To map label paths in the current schema into paths in peer schemas; and (3) *Tree Mapping* — To transform the current schema as a tree into trees of peer schemas.

4.2 Peer Similarity Measurement

A peer determines where to forward a query according to semantic similarity and structural similarity between itself and the successors. The semantic similarity can be measured by cycle analysis, functional dependency analysis, etc. [1]. We define the structural similarity as:

$$Structural\text{-}Similarity\ (S_{P_i}, S_{P_j}) = \frac{\vec{W} \cdot \overrightarrow{FV_{str}}}{\left|\vec{W}\right| \left|\overrightarrow{FV_{str}}\right|}$$

Where S_{P_i}, S_{P_j} are schemas of P_i, P_j, $\vec{W} = (W_{P_i N_1}, ..., W_{P_i N_k})$ is a user-defined weight vector to denote importance of nodes in S_{P_i}, $\overrightarrow{FV_{str}} = (fv_{P_i N_1}, ..., fv_{P_i N_k})$ is a feature vector to express structural similarity between S_{P_i} and S_{P_j}. If there are no mapping nodes in S_{P_j} for node N_j, then $fv_{P_i N_j} = 0$,

Else $$fv_{P_i N_j} = \frac{1}{dist(N_j, mapping(N_j)) + 1}$$

where $dist(N_j, mapping(N_j))$ denotes the distance, i.e., the minimum total number of deletion, insertion and substitution operations required to transform N_j to its mapping node in S_{P_j}.

4.3 Heterogeneous Data Integration

Within a predefined timeout, the peer initiating a query will analyze the data flows returned and integrate relevant data to answer the query. To solve the problem of data inconsistency, we take into consideration QoP (i.e., Quality of Peers), which focuses on user-perceived qualities, such as response time, precision, recall, traffic overhead, etc.

5. CONCLUSIONS

The proposed approach provides a new way to integrate heterogeneous data in P2P semantic link networks, which incorporates the characteristic of P2P and semantic link network to provide users with a single semantic image data usage mode.

6. ACKNOWLEDGMENTS

The research work was supported by the National Science Foundation of China (NSFC). We thank all the team members in China Knowledge Grid Research Group (http://kg.ict.ac.cn, http://www.knowledgegrid.net).

7. REFERENCES

[1] K. Aberer, P. Cudre-Mauroux, M. Hauswirth. The chatty Web: emergent semantics through gossiping. WWW 2003, May, 2003, Budapest, Hungary.

[2] P. Bernstein et al. Data management for peer-to-peer computing: a vision. WebDB 2002, June, 2002, Madison, Wisconsin.

[3] H.Zhuge. Active e-document framework ADF: model and tool. Information and Management 41 (2003), 87-97.

[4] H.Zhuge. Clustering soft-devices in Semantic Grid. IEEE Computing in Science and Engineering 4 (2002), 60-63.

[5] H.Zhuge, J.Liu and L.Ding. Service integration based on componential process construction and service grid. WWW 2003, May, 2003, Budapest, Hungary. Available at http://www2003.org/cdrom/papers/poster/p137/p137-zhuge/p137-zhuge.htm.