A Framework for Semantic Gossiping[*]

Karl Aberer, Philippe Cudré-Mauroux, Manfred Hauswirth
École Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland
{karl.aberer, philippe.cudre-mauroux, manfred.hauswirth}@epfl.ch

Abstract

Today the problem of semantic interoperability in information search on the Internet is solved mostly by means of centralization, both at a system and at a logical level. This approach has been successful to a certain extent. Peer-to-peer systems as a new brand of system architectures indicate that the principle of decentralization might offer new solutions to many problems that scale well to very large numbers of users.

In this paper we outline how the peer-to-peer system architectures can be applied to tackle the problem of semantic interoperability in the large, driven in a bottom-up manner by the participating peers. Such a system can readily be used to study semantic interoperability as a global scale phenomenon taking place in a social network of information sharing peers.

1 Introduction

The recent success of peer-to-peer (P2P) systems once again has surfaced a key problem in information systems: the lack of semantic interoperability. Semantic interoperability is a crucial element for making distributed information systems usable by providing features such as distributed query processing. At the moment, P2P systems either impose a simple semantic structure a-priori (e.g., Napster) and leave the burden of semantic annotation to the user, or they do not address the issue of semantics at all (e.g., Gnutella, Freenet) but simply offer unstructured, i.e., textual, data representation and leave the burden of search to the skills of the user.

Classical approaches, on the other hand, to make information resources semantically interoperable, in particular in the domain of heterogeneous database integration, appear to have problems in scaling to very large and very dynamic integration scenarios, as they typically require some global focal point, be it in the form of global schemas or globally used ontologies.

Despite a large number of concepts and tools developed, such as the federated DB architecture, the mediator concept or ontology-based information integration approaches [Hull 97][Ouksel, Sheth 99], practically engineered solutions still are frequently hard-coded systems that require substantial support through human experts. For example, domain-specific portals such as CiteSeer (www.researchindex.com, publication data), SRS (srs.ebi.ac.uk, biology) or streetprices.com (e-commerce) integrate data sources on the Internet and store them in a central warehouse. They typically require substantial development effort for the automatic or semi-automatic generation of mappings from the data sources into a global schema. Approaches using global schemas (either constructed bottom-up using GAV approaches or top-down as in LAV approaches) are inherently difficult to apply in dynamically evolving application domains, such as scientific databases and in the large scale.

We argue that we can see the emerging P2P paradigm as an opportunity for semantic interoperability rather than as a threat. First, we observe that semantic interoperability is always based on some form of agreement. Approaches to establish such agreements at a global level seem to be doomed to fail since no global enforcement is possible in highly autonomous environments. Thus we impose more modest requirements by assuming only the existence of local agreements on mappings between different schemas to enable semantic interoperability, i.e., agreements established in a P2P manner. These agreements will have to be established in a manual or semi-automatic way. Once such agreements exist we establish on-demand relationships among schemas of different information systems that are sufficient to satisfy information processing needs such as distributed search.

A first natural application of our approach would be the introduction of meta-data support in P2P applications. Due to their decentraliza-

[*] The work presented in this paper was supported (in part) by the National Competence Center in Research on Mobile Information and Communication Systems (NCCR-MICS), a center supported by the Swiss National Science Foundation under grant number 5005-67332.

tion, imposing a global schema for describing data in P2P systems is nearly impossible. It will not work if not all users concisely follow the global schema. Here our approach would fit well: We let users introduce their own schemas which best meet their requirements and by exchanging translations between these schemas, they incrementally come up with a “consensus schema” which gradually improves the global search capabilities of the used P2P system. This approach is orthogonal to the existing P2P systems and could be introduced basically into all of them.

Conceptually, our approach is built on two pillars: semantic gossiping and analysis of networks of schema mappings.

For enabling search we build on the gossiping approach that has been successfully applied for creating useful global behavior in P2P systems. Search requests are broadcasted over a network of interconnected information systems, and in addition when different schemas are involved, local mappings among them are used to further distribute them. However, the quality of search results in such an approach depends on the quality of the local mappings. Our fundamental assumption is that these mappings can be incorrect.

Thus assessments need to be made whether mappings can be trusted or not. We do this by analyzing what amount of *agreement* exists in composed mappings that had been constructed while traversing the network. This allows us to extract as much (consistent) global information as possible from existing local mappings and thus to extract globally agreed semantics. We apply this knowledge for routing search requests more precisely.

We believe that this fundamentally new approach to semantic interoperability shifts the attention from problems that are inherently difficult to solve in an automated manner at the global level – namely how to interpret information in terms of real world concepts – to a problem that leaves vast opportunities for automated processing and has high potential for increasing the value of existing knowledge through processing of existing local knowledge on semantic relationships to raise local semantic interoperability to a global level.

2 Semantic Gossiping

We assume that there exists a communication facility among the participants that enables sending and receiving of information, i.e., queries, data, and schema information. The underlying system could typically be a P2P system, but also a federated database system or any system of information sources communicating via some communication protocol. We assume that the peers *P* are all using semantically heterogeneous schemas *S* to represent their information. The case where multiple peers share the same schemas leads to possible optimizations of the approach, but is not conceptually different; therefore we ignore it in the following. To be semantically interoperable, the peers maintain knowledge about the relationships among their schemas. This knowledge can be given in the form of views, for example. For peers $p_1, p_2 \in P$ with schemas $S_1, S_2 \in S$ the relationship is given by a query $q_{1,2}$ applicable to schema S_2 and producing results according to schema S_1. We assume that skilled experts supported by appropriate mapping tools are able to provide these mappings. The direction of the mapping and the node providing a mapping are not necessarily correlated. For instance, both node p_1 or p_2 might provide a mapping from schema S_1 to schema S_2, and they may exchange this mapping upon discretion. During the operation of the system, each peer has the opportunity to learn about existing mappings and add new ones. This means that a directed graph of mappings will be built between the peers along with the normal operation of the system (e.g., query processing and forwarding in a P2P system). Such a mapping graph has two interesting properties:

(1) based on the already existing mappings and the ability to learn about mappings, new mappings can be added automatically by means of transitivity, and

(2) the graph has cycles.

The first property essentially means that we can propagate queries towards nodes to which we have no direct translation link. This is what we denote as *semantic gossiping*. While doing so, the nodes may check whether the translated query is worth to be propagated at all. It may occur that the translated query will return no or no meaningful results because of schema mismatches. This can be checked at a syntactic level. We do this by introducing the concept of syntactic distance, which analyses to what extent a query is preserved after translation, and use the syntactic distance as a criterion to decide whether or not to propagate a query.

The second observation leads to a chance to assess the degree of semantic agreement among a set of peers along a cycle. An agreement corresponds to those parts of the schemas that are preserved along the cycle. We have the opportunity to automatically assess the degree of semantic agreement by analyzing the result of propagated queries. As soon as a peer detects that it has been reached by its own query again, it can investigate "how much of its original query is left," e.g., which is the result of the compound mappings on the original query. A second potential for analyses is provided by the fact that any node along the propagation path of a query may return results to the originator. In that case the received results can again be used to assess the quality of the mappings along a cycle.

Now the question remains how a peer can locally use such information to assess the quality of an outgoing mapping in future routing decisions? A peer will obtain both returned queries and data through multiple cycles. In case a disagreement is detected (e.g., a wrong attribute mapping at the schema level or the violation of a constraint at the data level), the peer has to suspect that at least one of the mappings involved in the cycle was incorrect, including the mapping it has used itself to propagate the query. Even if an agreement is detected it is not clear whether this is not the accidental result of compensating mapping errors along the cycle. Thus analyses are required that assess which are the most probable sources of errors along cycles, to what extent the own mapping can be trusted and therefore of how to use these mappings in future routing decisions.

Assuming every peer is doing the same, we may expect the peer community converging to a state where the correct mappings are increasingly re-enforced by adapting the routing decisions, which we then may consider as a state where the best possible global agreement on the semantics of the schemas has been reached.

3 An Illustrating Example

One immediate application of our approach in the context of today's P2P systems is the provisioning of structured metadata in order to annotate files (media files, documents, etc.). For example, using XML as syntax, we can annotate music files in database DB1 as illustrated in the following example.

```
<song>
  <name>On my way to nowhere</name>
  <artist>Vagabond</artist>
  <encoding type="mp3" />
  <copyright boolean="yes"
     year="2002">AD Inc.</copyright>
  <size unit="MB">3.8</size>
</song>
```

Using such metadata annotations would be a simple extension of the current practice of using textual strings for representing media content and could be adopted in existing implementations of P2P systems with minimal effort. Once such a representation is given, search requests can be not only be "flat" textual strings (which should still be supported to provide backward compatibility), but could then also be formulated in a structured query language such as XQuery. Likewise, queries are also used to translate between different annotation schemas.

The example given below illustrates this. It provides a translation from DB1 to another database DB2 (the type of the XML annotations translated to is clear from the query)

```
Q12 =
<mp3>
  <author> $s/artist</author>
  <title> $s/name</title>
  <size> $s/size</size>
</mp3>
FOR $s IN /song
```

This query now is used to translate a query against DB2 into a query against DB1. For example, the query

```
Q =
<alltitles>
  <title>$m/title</title>
  <album>$m/album</album>
</alltitles>
FOR $m IN /mp3
```

becomes the following query against DB1 after composition (and some simplification):

```
Q12(Q) =
<alltitles>
  <title>$s/name</title>
</alltitles>
FOR $s IN /song
```

Note that the element `<album>` is lost. This might be a reason to decide not to forward the query. Assuming that an inverse translation from DB2 to DB1 is given as

```
Q21 =
<song>
  <name>$m/author</name>
</song>
FOR $m IN /mp3
```

we would obtain the equivalence between

```
Q’ =
<title>$m/title</title>
FOR $m IN /mp3
```

and

```
Q12(Q21(Q’)) =
<title>$m/author</title>
FOR $m IN /mp3
```

which is obviously incorrect. This increases our uncertainty on translation Q12 when using it, but still leaves the possibility that the error is with Q21 (which is obviously the case in the example, but actually not immediately clear when looking at Q21 in isolation).

4 Implementation

All the tasks described in the previous section have been mapped onto an implementation architecture. The implementation uses a meta-data model expressed in XML and XQuery as the language to translate among schemas and it assumes the availability of a communication infrastructure. Without loss of generality, this could be JXTA [Gong 01], so that any JXTA compliant P2P infrastructure could be used. Based on these assumptions, Figure 1 shows the standard architecture used for semantic gossiping.

Incoming queries are registered at and handled by the *Incoming Query and Result Handler* whose task is to communicate with other peers, to forward the query for further processing and to gather partial results to assemble the final result of a query. The first step then is to detect whether a cycle has occurred. If so, semantic analysis of the cycle is triggered. Otherwise, the query is processed, first by querying the local database and then by handing it over to the *Query Router and Translator* to collect results from other peers.

Then the *Query Router and Translator* inquires for possible translations, evaluates the quality of the resulting queries, and if it is above a defined threshold, forwards the query to the respective peer in a different semantic domain. Queries are forwarded by the *Outgoing Query and Result Handler* which is also in charge of collecting the results and forwarding the results to the *Incoming Query and Result Handler* which returns them to the original requester. Additionally, it provides input data for semantic result analysis.

This was the main data processing flow of the architecture. In parallel, partly triggered by the ongoing data processing, there is also a semantic processing cycle as depicted on the right side of Figure 1. Its main tasks are semantic analyses of results based on the existing knowledge of schemas and their relationships and the semantic analyses of detected cycles. The results of these analyses are integrated again into the knowledge base and provide the basic decision criteria for query routing.

Additionally, the knowledge base is updated and improved by exploring the peer’s neighborhood and detecting new schemas and translations. The metadata repository will try to infer further translations and present new ones for human analysis as described in the previous section.

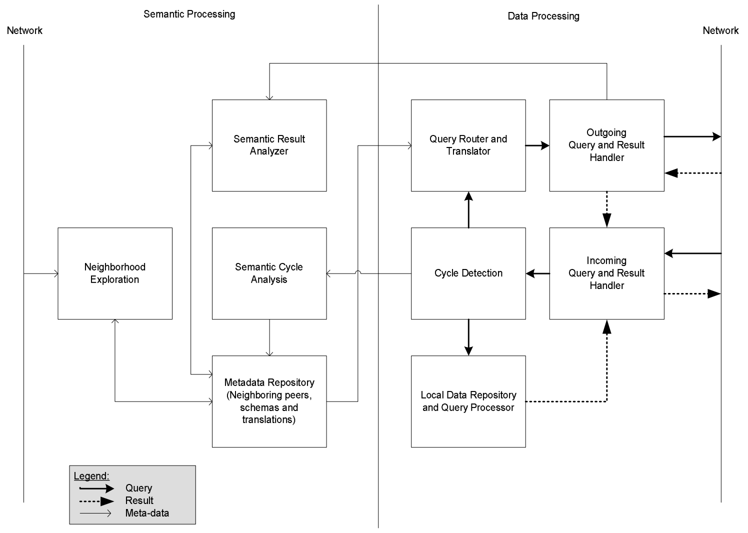
Semantic Processing
Data Processing
Network
Network
Semantic Result Analyzer
Query Router and Translator
Outgoing Query and Result Handler
Neighborhood Exploration
Semantic Cycle Analysis
Cycle Detection
Incoming Query and Result Handler
Metadata Repository (Neighboring peers, schemas and translations)
Local Data Repository and Query Processor
Legend:
Query
Result
Meta-data

Figure 1: Architecture for semantic gossiping

5 Current State of Work

We have developed algorithms for syntactic analysis of information loss and for determining the semantic correctness of mappings. Semantic correctness is checked by a maximum likelihood estimation of the correctness of compound queries along cycles and by dependency analysis of query results. The algorithms are given for a simple data model and query setting, consisting of relational tables with complex attribute types and single-table select-project-map queries. From the different features obtained through analysis, measures are derived that are used to route future queries. For performing practical experiments we have developed an interactive tool for automatic cycle analysis of networks of mappings between XML databases expressed as XQuery expression.

6 Related Work

A number of approaches for making heterogeneous information sources interoperable are based on mappings between distributed schemas or ontologies without making the canonical assumption on the existence of a global schema.

For example, in OBSERVER [Mena et al 00] each information source maintains an ontology, expressed in description logics, to associate semantics with the information stored and to process distributed queries. In query processing they use local measures for the loss of information when propagating queries and receiving results. Similarly to OBSERVER, KRAFT [Preece et al 01] proposes an agent-based architecture to manage ontological relationships in a distributed information system. Relationships among ontologies are expressed in a constraint language. [Bernstein et al 02] propose a model and architecture for managing distributed relational databases in a P2P environment. They use local relational database schemas and represent the relation between those through domain relations and coordination formulas. These are used to propagate queries and updates. The relationships given between the local database schemas are always considered as being correct. In [Ouksel, Ahmed 99] a probabilistic framework for reasoning with assertions on schema relationships is introduced. Thus their approach deals with the problem of having possibly contradictory knowledge on schema relationships. [Mukeheriee 02] propose an architecture for the use of XML-based annotations in P2P systems to establish semantic interoperability.

Approaches for automatic schema matching – for an overview see [Rahm, Bernstein 01] – would ideally support the approach we pursue in order to generate mappings in a semi-automated manner. In fact, we may understand our proposal as extending approaches for matching two schemas to an approach of matching multiple schemas in a networked environment.

Finally we see our proposal also as an application of principles used in Web link analysis, such as [Kleinberg 99], in which local relationships of information sources are exploited to derive global assessments on their quality (and eventually their meaning).

7 Conclusions

By putting in place an infrastructure for semantic gossiping we can expect to establish a laboratory for studying how peers (which are of course instantiations of human users) interact, if they have the opportunity to interact in a semantically more meaningful manner. For example, it would be interesting to see whether specific schemas start to dominate the network (for example, the schema distribution would follow a power-law distribution, which emerges frequently in networked interactions), or multiple schemas connected by gateways could co-exist, or whether, for example, the network would partition into completely disconnected sub-networks. Essentially these processes will be driven by the individual decisions of peers. The peers will be taking into account the basic trade-off between the cost of adapting their own schema to some other (and so adhering to some established schema), or producing the necessary translation to some other schema in order to remain connected with the rest of the network. Scientific data management with it’s high dynamics, rich semantics and high demand of expert knowledge might be a primary application for setting up and evaluating an infrastructure implementing the principles of semantic gossiping.

References

[Bernstein et al 02] P.A. Bernstein, F. Giunchiglia, A. Kementsietsidis, J. Mylopoulos, L. Serafini, I. Zaihrayeu: Data management for peer-to-peer computing: A vision. In: Workshop on the Web and Databases, WebDB 2002.

[Gong 01] Li Gong. JXTA: A Network Programming Environment. IEEE Internet Computing, 5(3):88–95, May/June 2001.

[Hull 97] Richard Hull: Managing Semantic Heterogeneity in Databases: A Theoretical Perspective. PODS 1997: 51-61.

[Kleinberg 99] Jon M. Kleinberg: Hubs, authorities, and communities. ACM Computing Surveys 31(4es): 5 (1999)

[Mena et al 00] Eduardo Mena, Vipul Kashyap, Amit P. Sheth, Arantza Illarramendi: 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.

[Mukeheriee 02] A. Mukherjee, B. Esfandiari, N. Arthorne: U-P2P: A Peer-to-peer System for Description and Discovery of Resource-sharing Communities , RESH 2002.

[Ouksel, Ahmed 99] Aris M. Ouksel, Iqbal Ahmed: Ontologies are not the Panacea in Data Integration: A Flexible Coordinator to Mediate Context Construction. Distributed and Parallel Databases 7(1): 7-35 (1999)

[Ouksel, Sheth 99] Aris M. Ouksel, Amit P. Sheth: Semantic Interoperability in Global Information Systems: A Brief Introduction to the Research Area and the Special Section. SIGMOD Record 28(1): 5-12 (1999)

[Preece et al 01] Alun D. Preece, Kit-ying Hui, W. A. Gray, Philippe Marti, Trevor J. M. Bench-Capon, Zhan Cui, Dean Jones: Kraft: An Agent Architecture for Knowledge Fusion. IJCIS 10(1-2): 171-195 (2001)

[Rahm, Bernstein 01] Erhard Rahm, Philip A. Bernstein: A survey of approaches to automatic schema matching. VLDB Journal 10(4): 334-350 (2001)