A Semantic Layer for Publishing and Localizing XML Data for a P2P XQuery Mediator

Florin Dragan
Université de Versailles
Saint-Quentin-en-Yvelines
45 avenue des États-Unis
78035 Versailles cedex,
France
fdragan@gmail.fr

Georges Gardarin
Université de Versailles
Saint-Quentin-en-Yvelines
45 avenue des États-Unis
78035 Versailles cedex,
France
ggardarin@gmail.fr

Laurent Yeh
Université de Versailles
Saint-Quentin-en-Yvelines
45 avenue des États-Unis
78035 Versailles cedex,
France
yeh@prism.uvsq.fr

ABSTRACT

In this poster, we present the P2P XML data localization layer of XLive, an XQuery mediation system developed at University of Versailles [2]. A major challenge in the evaluation of XQuery over a P2P network in the context of multiple XML sources is to abstract from structural heterogeneity. Most existing approaches have mainly exploited varied types of 1:1 semantic mappings between peer schemas and/or ontologies. Complex query algorithms are required to exploit these semantic links between peers. In contrast, our approach focuses on a simple semantic layer. It is built on a Chord [4] DHT [1] for indexing semantic descriptions of mediated data sources. A mapping process transforms an XML view of a mediator (each peer is equipped with a mediator) to a semantic form that is published into the P2P network and stored in the DHT. For a given user query, a search in the DHT retrieves all relevant data sources able to contribute to the result elaboration. Then, every peer that contains relevant data is directly queried to collect data for elaborating the final answer.

Categories and Subject Descriptors:
H.2.4. [Systems]: Query processing H.2.5. [Heterogeneous Databases]: Data translation

Keywords: XQuery, P2P, Semantic mediation.

1. INTRODUCTION

Data mediation in P2P system requires publishing and querying multiple data sources under different organization formats (e.g., XML, relational, LDAP). The structural heterogeneity of XML data sources is a bottleneck for localizing relevant data sources for a query. For instance, a fragment `authors` of books and reviews XML collections may be represented using the tag names `author` in books and `main_author` in reviews. We may find that both tags have a similar meaning by analyzing the descendants of the authors tag. Moreover, a tag with identical meaning in two XML collections can be addressed by different Xpaths (e.g., `book/author` and `review/paper/author`).

[1](Distributed Hash Table) is a method for storing hash tables in decentralized distributed systems. Any peer can efficiently retrieve the value associated with a given key.

WWW 2008, April 21–25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

Current approaches for querying several sources with heterogeneous data structure have mainly exploited varied types of 1:1 mappings between peers. Piazza [3] and SomeWhere [1] are two representative P2P infrastructures for sharing and mediating XML data sources. When a peer joins the network, it has to provide mappings between his local schema or ontology and some foreign schema or ontology. Piazza covers a large variety of mappings expressed in a language derived from XQuery with a complex query processing algorithm. SomeWhere uses description logic to define mappings between ontologies. Queries are routed according to the relevant mappings. When a user poses a query on a peer, it uses its local schema. The query is first mapped to the local data stored at the node. Next, the routing algorithm determines all the neighbors of the peer (i.e., nodes related by semantic mappings), reformulates the query for them, and passes the modified queries to them. The recursive processing of a query until no remaining useful links are discovered guarantees the exploration of all sources that are relevant.

For localizing and querying heterogeneous data sources, we have designed and prototyped a semantic layer for enhancing our existing XQuery mediator [2]. The main purpose of this semantic layer is to add an abstract data description model that is somehow independent of the structural heterogeneity of different XML data sources. For doing that, we map an XML view of a data source to a semantic form that is published into the P2P network. We introduce a simple semantic data model that is a simplification of the W3C RDF-S and OWL semantic languages. When extracting data an XML query is transformed to a semantic format that is then used for retrieving relevant data. In contrast with other approaches, we use the DHT `lookup` possibilities and a coloring algorithm for mediating all relevant data sources. Using this layer, the queries can be formulated with more precision and all the peers that contain relevant data can contribute to the final results.

2. SEMANTIC MODEL

We use a simple semantic model. Its main purposes are to force the peers to publish data in the same vocabulary and to facilitate the query execution process by eliminating the missing answers due to the structural heterogeneity of data sources. For simplicity and efficiency, the semantic model is close to the binary E-R model. Thus, it does not contain all the structures included in advanced semantic languages like RDFS or OWL, e.g. `is_a` and `part_of` relationships.

Our semantic model follows the organization of a bipartite graph. A bipartite graph is a graph whose vertexes can be divided in two disjoint sets (C,R). We consider that each component of the C set is an English word that defines a *concept*. Similarly each member of R is an English word that defines a *relation* between concepts. Every edge in the graph connects a vertex from the first set C with a vertex in the second set R. There is no edge between two vertexes in the same set.

3. SOURCE ANNOTATION

The goal of the annotation process is to generate a semantic image starting from the unified XML view provided by a mediator on a peer. A semantic image of a peer is a sub-graph of the global semantic graph (i.e., the ontology). In order to extract a semantic sub-graph from an XML view we consider that an XML document is composed of XML entities (i.e. elements, attributes) and relations between entities. The relation between entities corresponds to the child XML axis.

To extract a semantic sub-graph from an XML view, we developed a specific User Interface that proposes when possible a default mapping. The default mapping is based on the following rules: (i) An XML element or attribute is mapped to a concept in the common ontology in the C set. (ii) The XML child axis between two XML elements is mapped to a *relation* in the ontology in the R set. The relation depends on the XML element that is the origin of the axis and on the XML element that is the destination of the axis (i.e., the context XML elements). The same XML axis can be mapped to two different relations depending on the context XML elements. The result of this process is a set of semantic annotations where the first element contains two concepts and the second element contains a single relation (e.g `<(book,author), written_by>`).

4. P2P PUBLISHING AND QUERYING

For making the information available to all the peers in the mediation architecture, a peer must publish its semantic view into the P2P network. We consider that each piece of information is identified by one of the elements in the concept sets (this represents the key that will be further used for retrieving the information). To each key, we associate a value that is a semantic annotation. We use the value associated to the key for storing in the network all pieces of information required to retrieve the initial XML data.

The semantic information that is published for a concept is represented by all the relations in the R set that are linked to the concept and their correspondences in the C set. We have chosen to publish this kind of information for a given concept in order to facilitate the identification of a concept based on his links and neighbor concepts during the query evaluation phase.

A user XQuery submitted to the XLive mediator is transformed to a semantic query based on the ontology. The query is then decomposed in a set of semantic sub-queries. Each sub-query is evaluated using the P2P network. Each time a new answer is returned from a peer, a new XQuery can be generated and sent for evaluation to the peer that answered.

The process of query transformation based on a semantic graph is similar to the process of source annotation. We represent a semantic query as a bipartite graph. Then, we generate a DHT get operation. The key of the P2P get is represented by the concept. By executing the P2P search we can retrieve multiple messages. Each message contains a semantic part and a structural part (the source information). The two parts are then used in the result composition algorithm.

The result composition algorithm is based on the existence of the semantic query and on the XML query tree. For composing the final result, we propose an algorithm that is based on the coloring of the semantic query graph. The algorithm proceeds as follows. Based on the P2P information retrieved in the previous step, we give a color to all concepts and relations that come from the same document on the same peer. For each message retrieved from the P2P network we analyze the semantic information. The semantic information contains for each key concept the relations and concepts that it is linked to. All the relations and concepts that are found in the semantic query graph are colored with the same color. Next, for each partition of the graph identified by a given color we generate an XQuery than recomposes the local pieces of data according to the final form of the query. The XQuery is sent for evaluation at the peer where the document identified by the color resides. The XML results of sub-queries are recomposed by the mediator that initiated the query.

5. CONCLUSION

We propose a semantic layer for a P2P mediation network to improve the discovery of query relevant data sources. The main originality of the layer is the distribution of semantic views and ontologies in a Chord DHT, which allows fast retrieval of semantic concepts, relations, and mappings. In addition, the semantic layer helps solving the problems linked to the integration of data from structural heterogeneous XML sources. It is based on a data representation model that is similar to the core model of RDF and OWL, the semantic languages proposed by W3C to define ontologies. Extensions to fully support RDFS are on the way. A first version of the semantic layer is currently integrated in the XLive P2P mediation architecture (XLive has been used in several Europeen and French projects). The integration of such a semantic layer into an XML mediation architecture represents a new step towards semantic mediation of heterogeneous data sources.

6. REFERENCES

[1] P. Adjiman, P. Chatalic, F. Goasdou, M.-C. Rousset, and L. Simon. SomeWhere in the Semantic Web . In *International Workshop on Principles and Practice of Semantic Web Reasoning*, 2005.

[2] G. Gardarin and al. XLive : An XML Light Integration Virtual Engine . In *http://www.prism.uvsq.fr/ñtravers/xlive/*, 2007.

[3] Z. G. Ives, A. Y. Halevy, P. Mork, and I. Tatarinov. Piazza: mediation and integration infrastructure for semantic web data. *J. Web Sem.*, 1(2):155–175, 2004.

[4] I. Stoica and al. Chord: A scalable peer-to-peer lookup service for internet applications. In *Proceedings of the 2001 conference on applications, technologies, architectures, and protocols for computer communications*, pages 149–160. ACM Press, 2001.