Using Semantics in Peer Data Management Systems

Carlos Eduardo Pires[1], Damires Souza[1], Zoubida Kedad[2], Mokrane Bouzeghoub[2], Ana Carolina Salgado[1]

[1]Universidade Federal de Pernambuco (UFPE), Centro de Informática, Av. Prof. Luiz Freire, S/N 50.740-540 Recife, PE, Brazil

[2]Université de Versailles et Saint-Quentin-en-Yvelines (UVSQ), 45 Avenue des Etats-Unis, 78035 Versailles, France

{cesp,dysf,acs}@cin.ufpe.br, {zoubida.kedad, Mokrane.Bouzeghoub}@prism.uvsq.fr

Abstract. *Data management in Peer Data Management Systems (PDMS) is a challenging problem considering the excessive number of peers, their autonomous nature, and the heterogeneity of their schemas. To help matters, semantic knowledge in the form of ontologies has proven to be a helpful support for the techniques used for managing data in such systems. Ontologies can be used, for instance, to represent the semantic content of data sources as well as to unify the semantic relationships between their schemas. In this sense, the goal of this paper is to highlight the use of semantics in order to enhance data management issues in a PDMS. We present the current status of scientific cooperation between the database groups of Centro de Informática from Universidade Federal de Pernambuco (CIn/UFPE) and PRiSM laboratory from Université de Versailles Saint-Quentin-en-Yvelines (PRiSM/UVSQ). In addition, we point out further work to be done.*

1. Introduction

Peer Data Management Systems (PDMS) [Halevy *et al*., 2003; Valduriez and Pacitti, 2004; Mandreoli *et al*., 2007; Lodi *et al*., 2008; Kantere *et al*., 2009] came into the focus of research as a natural extension to distributed databases in the Peer-to-Peer (P2P) setting [Herschel and Heese, 2005]. They consist of a set of peers, each one with an associated schema (called exported schema) that represents the data to be shared with other peers. In such systems, schema matching techniques are used to establish schema mappings (i.e., correspondences between schema elements) which form the basis for query answering and peer clustering. Schema mappings are defined between pairs of semantic neighbor peers, i.e., peers that are semantically related as previously identified by a clustering process. Queries submitted at a peer are answered with data residing at that peer and with data that is reached through mappings over the semantic neighbors.

Data management in PDMS is a challenging problem considering the excessive number of peers, their autonomous nature, and the heterogeneity of their schemas. To help matters, semantic knowledge in the form of ontologies has proven to be a helpful support for the techniques used for managing data in such systems. For instance, ontologies can be used to represent the semantic content of data sources as well as to unify the semantic relationships between their schemas. Thus, the goal of this research project is to exploit the benefits provided by semantics through ontologies to enhance data management issues in PDMS. To this end, we present semantic-based approaches to support peer clustering, schema summarization, schema matching, and query reformulation. In the following, we present a description of such approaches as well as some obtained experimental results.

2. Work in Progress

The main categories of problems which have been particularly addressed by this research are described in the following.

Ontology-based PDMS

The establishment of schema mappings and consequently query answering in PDMS can be improved if semantically similar peers are put together in the overlay network. In this sense, we have proposed a semantic-based PDMS [Pires *et al.*, 2007] whose mixed network is mainly designed to assist the organization of peers according to their exported schema (represented by an ontology). Peers are grouped according to their knowledge domain (e.g., *Education* and *Health*), forming semantic communities. Inside a community, peers are organized in a finer grouping level, named semantic clusters, where peers share similar ontologies (schemas). A semantic cluster has a cluster ontology which represents the ontologies (schemas) of the peers within the cluster. Each cluster maintains a link to its semantic neighbors in the overlay network, i.e., to other semantically similar clusters. Regarding implementation, a PDMS simulator has been developed through which we were able to reproduce the main conditions characterizing the proposed system's environment.

Ontology Matching

We have proposed a semantic-based ontology matching process, named *SemMatch* [Pires *et al.*, 2009a], that considers, besides the traditional terminological and structural matching techniques, a semantic-based one. The process produces a set of semantic correspondences and a global similarity measure between two peer ontologies. The former is used to enhance query reformulation while the latter is used, for instance, to determine semantic neighbor peers in the overlay network. A tool implementing the semantic-based ontology matching process has been developed.

Ontology Summarization

We have proposed an automatic process to build summaries of cluster ontologies [Pires *et al.*, 2009b]. Such summaries are used as a semantic index to assist the identification of similar peers when a new peer joins the system. The summarization process is divided into several steps and is based on the notions of centrality and frequency. Centrality is used to capture the importance of a given concept within an ontology. The use of frequency is motivated by the fact that a cluster ontology is obtained by merging several different local ontologies. The summaries are used as a semantic index to indicate an initial cluster for new peers during their connection to the system. We have developed a preliminary implementation of an ontology summarization tool.

Ontology-based Peer Clustering

Peer connection in the proposed PDMS is mainly an incremental clustering process. When a new peer arrives, it searches for a corresponding semantic community in a DHT network. Then, within a semantic community, the new peer searches for a semantically similar cluster in an unstructured network. The search for a cluster starts when the new peer sends its exported schema (i.e., an ontology) to a promising initial cluster (provided by the semantic index) and proceeds by following the semantic neighbors of the initial cluster until a certain limit (TTL) is reached. At each visited cluster, *SemMatch* is executed taking as arguments the current cluster ontology and the exported schema of the new peer. Each cluster returns its global similarity measure to the new peer. The set of global measures are used by the new peer to determine if it will join an existing cluster or create a new one. The proposed process has been implemented in the mentioned simulator and submitted to experimental evaluation. Validation has been performed using clustering indices and by executing query answering simulations.

Query Reformulation

In our PDMS, a query posed at a peer is routed to other peers in order to find answers to the query. An important step in this process is reformulating a query issued at a peer into a new query expressed in terms of a target peer, considering the correspondences between them. In this light, we have worked on a semantic-based query reformulation approach, named *SemRef* [Souza *et al.*, 2009], which brings together both query enrichment and query reformulation techniques in order to provide users with a set of expanded answers. Exact and enriched query reformulations are produced as a means to obtain this set of answers. To this end, we make use of semantics which is mainly acquired from a set of semantic correspondences that extend the ones commonly found. Also, we take into account the context of the user, of the query and of the environment as a way to enhance the overall process and to deal with information that can only be acquired on the fly.

3. Further Work

There are a number of ongoing research issues concerned with the use of semantics in PDMS. Among them, we will focus in two relevant issues: (i) the maintenance of semantic communities; and (ii) query routing. Concerning the former, an issue to be studied in deep detail regards the evolution of cluster ontologies. In order to reflect the content available in a semantic cluster, cluster ontologies should be created and maintained dynamically, in an automatic way, according to peers' intermittence. A cluster ontology should be able to evolve not only when a requesting peer joins the cluster but also when a participating peer leaves it.

Regarding the latter, query reformulation strategies and query routing mechanisms [Montanelli and Castano, 2008] have a great influence on each other. In our approach, we consider that every peer P_i maintains a neighborhood $N(P_i)$ selected from the set of existing peers in the setting. In this sense, a submitted query must be reformulated in such a way that it is possible to ensure effective query routing, preserving the query semantics at the best possible level of approximation. Furthermore, we intend to use semantics to enhance the selection of relevant semantic neighbors and their ranking.

4. Cooperation Status

The scientific cooperation between the database groups of Centro de Informática from Universidade Federal de Pernambuco (CIn/UFPE) and PRiSM laboratory from

Université de Versailles Saint-Quentin en Yvelines (PRiSM/UVSQ) began in the nineties when two students from CIn/UFPE were accepted as PhD students in PRiSM/UVSQ supervised by Prof. Mokrane Bouzeghoub. This cooperation was intensified in 2002 when a PhD student from CIn/UFPE did a 'sandwich' stage in PRiSM/UVSQ. Since then it has been established a regular cooperation which has included research visits, cooperative projects, a sabbatical year of the CIn/UFPE database group leader, and another PhD 'sandwich' stage in PRiSM/UVSQ. Currently, we have a STIC/AMSUD project (2008-2009) which motivated the organization of a workshop in Recife last July with participation of researchers from the groups involved on the project. One of the main research areas of this cooperation project is the use of semantics to enhance data management in dynamic distributed environments.

References

Halevy, A. Y., Ives, Z. G., Mork, P., and Tatarinov, I. (2003) "Piazza: Data Management Infrastructure for Semantic Web Applications", In: World Wide Web Conference, pp. 556-567, Budapest, Hungary.

Valduriez, P., and Pacitti, E. (2004) "Data Management in Large-Scale P2P Systems", In: International Conference on High Performance Computing for Computational Science, pp. 104-118, Valencia, Spain.

Mandreoli, F., Martoglia, R., Penzo, W., Sassatelli, S., and Villani, G. (2007) "SUNRISE: Exploring PDMS Networks with Semantic Routing Indexes", In: 4th European Semantic Web Conference, Innsbruck, Austria.

Lodi, S., Penzo, W., Mandreoli, F., Martoglia, R., and Sassatelli, S. (2008) "Semantic Peer, Here are the Neighbors You Want!", In: 11th Extending Database Technology, pp. 26-37, Nantes, France.

Kantere, V., Tsoumakos, D., Sellis, T., and Roussopoulos, N. (2009) "GrouPeer: Dynamic clustering of P2P databases", In: Information Systems Journal, Volume 34, Issue 1, pp. 62-86.

Herschel, S., Heese, R. (2005) "Humboldt Discoverer: a Semantic P2P Index for PDMS", In: Int. Workshop Data Integration and the Semantic Web, Porto, Portugal.

Pires, C. E. S., Lóscio, B. F., and Salgado, A. C. (2007) "Semantic-based Connectivity in a Peer Data Management System", In: 6th Workshop of Thesis and Dissertation on Data Base, in conjunction with the 22th SBBD, pp. 65-72, João Pessoa, Brazil.

Pires, C. E. S., Souza, D., Pachêco, T., and Salgado, A. C. (2009a) "A Semantic-based Ontology Matching Process for PDMS", to appear in the 2nd International Conference on Data Management in Grid and P2P Systems (Globe'09), Linz, Austria.

Pires, C. E. S., Alencar, V. B., Kedad, Z., and Salgado, A. C. (2009b) "Building Ontology Summaries for PDMS", Submitted to the 28th Conference on Conceptual Modeling, Gramado, Brazil.

Souza D., Arruda T., Salgado A. C., Tedesco P., and Kedad, Z. (2009) "Using Semantics to Enhance Query Reformulation in Dynamic Environments", to appear in the 13th East European Conference on Advances in Databases and Information Systems (ADBIS'09), Riga, Latvia.

Montanelli S., Castano S. (2008) "Semantically Routing Queries in Peer-based Systems: the H-Link Approach", In: Knowledge Eng. Review 23(1): 51-72.