Using Ontologies to Enhance Data Management in Distributed Environments

Carlos Eduardo Pires[1], Damires Souza[2], Bernadette Lóscio[3], Rosalie Belian[4], Patricia Tedesco[3] and Ana Carolina Salgado[3]

[1] Federal University of Campina Grande (UFCG), Computer Science Department
Av. Aprígio Veloso, 882, Bodocongó - 58109-970 - Campina Grande, PB, Brazil
cesp@dsc.ufcg.edu.br

[2] Federal Institute of Education, Science and Technology of Paraiba (IFPB), Brazil
Av. Primeiro de Maio, 720, Jaguaribe - 58015-430 - João Pessoa - Paraíba
damires@ifpb.edu.br

[3] Federal University of Pernambuco (UFPE), Center for Informatics
Av. Jornalista Anibal Fernandes, s/n, 50.740-560, Recife, PE, Brazil
{bfl, pcart, acs}@cin.ufpe.br

[4] Federal University of Pernambuco (UFPE), Center of Health Sciences
Av. Prof. Moraes Rego, S/N, Cidade Universitária – 50670-901 – Recife, PE, Brazil
rosalie.belian@ufpe.br

Abstract. Data management solutions in distributed environments have been continuously evolving during the last years to answer users' needs and face new technology challenges. To help matters, ontologies have been used as a support for the techniques of managing data. For instance, ontologies may be used to describe the semantics of data at different sources, helping to overcome problems of semantic interoperability and data heterogeneity, and thus assisting schema integration and query answering over the distributed data sources. The goal of this paper is to highlight the use of ontologies in order to enhance data management issues in distributed environments. To this end, we describe a set of ongoing works which have been developed in our research.

Keywords: Ontology, Semantics, Data Management, Distributed Environment.

1 Introduction and Research Statement

The increasing use of computers and the development of communication infrastructures have led to a demand for high-level integration of autonomous and heterogeneous data sources. This fact caused the development of diverse distributed environments, including Data Integration Systems [Halevy *et al.* 2006], Peer Data Management Systems (PDMSs) [Sung *et al.* 2005], and Dataspaces [Hedeler *et al.* 2009]. While these types of data integration systems differ with respect to their level of coupling, all of them have in common the need of dealing with heterogeneity, mappings, and query answering. Particularly, these dynamic distributed environments are characterized by an architecture constituted by various autonomous data sources (e.g., websites, files, databases), here referred to as *peers*. These peers are linked to

each other by means of mappings (i.e. associations between schema elements), called hereafter as *correspondences*.

Data management in large distributed environments is a challenging problem given the heterogeneity of their schemas. Due to the fact that ontologies provide good support for understanding the meaning of data, there has been a growing interest in using *ontologies* for enhancing data management in such environments. In these settings, they have been used for some purposes, including: (i) metadata representation: in each data source are represented by a local ontology; (ii) global conceptualization: providing a conceptual view over the schematically heterogeneous source schemas; and (iii) support for high-level queries: given a global ontology, users can formulate queries without specific knowledge of the different data sources.

In addition, due to semantic heterogeneity, research on distributed environments has also considered the use of ontologies as a way of providing a domain reference. Considering a given knowledge domain, an agreement on its terminology can occur through the definition of a domain ontology which can be used as a semantic reference or background knowledge to enhance processes such as ontology matching.

In this light, in our research, we deal with ontology-based distributed environments, where various ontologies are developed (representing peer schemas) with meaningful content overlapping among them. We have mainly instantiated our research in a PDMS, named SPEED - **S**emantic **PEE**r **D**ata Management System, which adopts an ontology-based approach to assist relevant issues in peer data management, e.g., query answering and peer clustering.

Another kind of semantic knowledge we use is *context*. The term is concerned with some specific situation, usually perceived as a set of variables that may be of interest for an agent [Bolchini *et al.* 2007]. In order to store and use context, an important issue is how to represent its elements. Context ontologies have been considered an interesting approach because they enable sharing and reusability and may be used by different reasoning engines [Souza *et al.* 2008]. In this work, we have designed an ontology, named CODI - **C**ontextual **O**ntology for **D**ata **I**ntegration, to represent and store contextual information.

In summary, the goal of this paper is to exploit the benefits provided by semantics through ontologies to enhance data management issues in distributed environments. To this end, we present ontology-based approaches to support schema matching, peer clustering, query reformulation, schema summarization, schema merging, and data access. Furthermore, we present an approach that uses ontology as a means to represent and store contextual information. In the following, we present an overview of our approaches. Also, we describe the history of our research group and members.

2 Main Areas of Research

The main areas of research we have been working are: (i) ontologies in a PDMS; (ii) ontology as a means to represent contextual information; and (iii) ontology to provide data access. We provide an overview of them in the following.

2.1 Ontology-based PDMS

SPEED (**S**emantic **PEE**r **D**ata Management System) [Pires 2009] is a PDMS that adopts an ontology-based approach to assist relevant issues in peer data management. Its main goal is to cluster semantically similar peers in order to facilitate the establishment of semantic correspondences between peers and, consequently, improve query answering. Peers are grouped according to their knowledge domain (e.g., *Education*), 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. A simulator has been developed through which we were able to reproduce the main conditions characterizing the proposed system's environment. The main issues which have been particularly addressed in SPEED are the following:

Using Ontologies to Represent Peer Schemas

In SPEED, we use ontologies as uniform conceptual representations of peer schemas. The use of ontologies as a middle layer between the system's processes and the data sources adds a conceptual level over the data. In addition, it allows the system to uniformly deal with data without worrying about their specific restrictions (syntactic or semantic). We have been implementing a tool that automatically extract semantics from data sources and builds a peer ontology. Meanwhile, we are working with geographic data sources to instantiate SPEED. Due to the complex semantics of spatial data, we are implementing some new extraction rules for the spatial relations.

Ontology as Background Knowledge

We use domain ontologies (DO) as background knowledge in order to identify semantic correspondences between matching ontologies [Souza 2009]. The use of background knowledge through ontologies enhances the identification of other types of correspondences by extending the ones commonly found (e.g., equivalence and subsumption). For instance, we are able to find out other kinds of correspondences such as closeness and disjointness. Finding such degree of semantic overlap between ontologies becomes more useful for tasks such as query answering.

Ontology-based Schema Matching

We have developed a semantic-based ontology matching process, named *SemMatcher* [Pires *et al*., 2009], 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 of SPEED. A tool implementing the semantic-based ontology matching process has been implemented.

Ontology Merging

We have also implemented a merge tool, denoted *OntMerger* [Pires 2009], that takes as arguments two ontologies (i.e., a cluster ontology and a peer ontology) and the set of correspondences between them (generated by *SemMatcher*). As a result, the tool produces a new version of the cluster ontology containing the elements of both input ontologies as well as semantic correspondences between the new cluster ontology and the peer ontology.

Using Ontologies to Enhance Query Reformulation in PDMS

In SPEED, a query posed at a peer is routed to other peers in order to find answers to the query. An important step of this task 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 query reformulation approach, named *SemRef*, which brings together both query enrichment and query reformulation techniques in order to provide users with a set of expanded answers [Souza *et al.* 2009]. Exact and enriched query reformulations are produced as a means to obtain this set of answers. To this end, we make use of semantics acquired from a set of semantic correspondences between peer ontologies (e.g., closeness). Also, we take into account the context of the user, of the query and of the environment as a way to enhance the process and to deal with information that can only be acquired on the fly.

Ontology Summarization

We have developed an automatic process to build summaries of cluster ontologies [Pires *et al.* 2010]. 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 SPEED. We have developed *OWLSum*, a tool implementing the ontology summarization process.

Ontology-based Peer Clustering

Peer connection in SPEED is mainly an incremental clustering process [Pires 2009]. When a new peer arrives, it searches for a corresponding semantic community in a structured 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, *SemMatcher* 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 simulator and submitted to experimental evaluation. Validation has been performed using clustering indices.

2.2 Ontology to Represent and Store Contextual Information

CODI (**C**ontextual **O**ntology for **D**ata **I**ntegration) is an ontology for representing context according to some Data Integration (DI) and PDMS issues [Souza *et al.* 2008]. In our work, we consider that Contextual Elements (CEs) are used to characterize a given entity. Therefore, we determined six main domain entities around which we consider the CEs: *user*, *environment*, *data*, *procedure*, *association*, and *application*. We have already used CODI in query reformulation as a way to store the user and query contexts. CODI was also used for schema reconciling, to identify in which context the elements occur and thus, to ease spell-check and schema-level sense disambiguation tasks [Belian *et al*. 2010]. Element names can have different meanings depending on the semantic context to which they are related. Hence, CEs may provide a more accurate semantic interpretation, allowing restrictions or characterizations of an element name according to a specific semantic context.

Currently, we are using CODI to represent and store the user model. We are developing a CODI Data Service which will be responsible for storage and retrieval of the contextual elements. This service will be coupled to the SPEED query system.

2.3 Query Rewriting between Ontologies

The use of ontologies, as a conceptual representation for data sources, gives origin to relevant problems such as the query rewriting between ontologies [Calvanese *et al*. 2009]. Given the relevance of such problem, we have investigated this area and we have proposed a solution for query rewriting between heterogeneous ontologies. More specifically, we have proposed a solution for the following problem. Considering a target ontology O_T, a source ontology O_S and a set of correspondences between them, how to rewrite a SPARQL query Q, submitted to O_T, into a query Q', to be submitted to O_S, in such a way that query results are presented according to O_T and that O_T and O_S are heterogeneous?

Our proposal for query rewriting between ontologies [Lopes 2010] combines the semantics and expressiveness of SPARQL with logic programming and considers the rule-based formalism for representing mappings between ontologies proposed in [Sacramento *et al*. 2010]. Our approach deals with some relevant questions, including: the structural heterogeneity between the target ontology and the source ontology and the prune of irrelevant parts of the rewritten query. A tool implementing the proposed query rewriting process, called *SQuOL*, has also been proposed.

3 History of the Group and Members

The SPEED project[1], directed by Ana Carolina Salgado, started in 2006 as an evolution of previous researches in traditional data integration systems. At this time, Carlos Pires and Damires Souza were PhD candidate students concerned with the main architectural and structural definitions of SPEED. Bernadette Lóscio and

[1] http://www.cin.ufpe.br/~speed

Rosalie Belian have developed their PhD thesis in related data integration problems and are research collaborators always interacting with the SPEED team. Patricia Tedesco is the Artificial Intelligence member of the group acting as co-advisor in some of the thesis. The SPEED group includes not only PhD students but also master and undergraduate students working in a complementary way to construct a PDMS prototype that consolidates the main obtained results.

References

Belian, R. B., Salgado, A. C. (2010): A Context-based Schema Integration Process Applied to Healthcare Data Sources. In Proc. of the International Conference *On the move to meaningful internet systems*, Springer-Verlag.

Bolchini C., Curino C.A., Quintarelli E., Tanca L., Schreiber F. (2007): A data-oriented survey of context models. SIGMOD Record, 2007.

Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., Poggi, A., Rodriguez-Muro, M., and Rosati, R. (2009): Ontologies and databases: The DL-lite approach. Reasoning Web. Semantic Technologies for Information Systems, pages 255–356.

Halevy A., Rajaraman A. and Ordille J. (2006): Data integration: the teenage years. In Proc. of the 32nd International Conference on Very Large Data Bases, Vol. 32, pages 9-16.

Hedeler, C., Belhajjame, K., Fernandes, A.A.A., Embury, S.M., Paton, N.W. (2009): Dimensions of Databases, In Proc. of 26th British National Conference on Databases, Birmingham, UK, pages 55-66.

Lopes, F. L. R. (2010): Acesso a Dados a partir de Ontologias Utilizando Mapeamentos Heterogêneos e Programação em Lógica. MSc. Thesis, UFC, Fortaleza, Brazil, Nov. 2010.

Pires, C. E.: Ontology-based Clustering in a Peer Data Management System. Ph.D. thesis, CIn/UFPE, Recife, Brazil, Apr. 2009.

Pires C. E, Souza D., Pacheco T., and Salgado A. C. (2009): A Semantic-Based Ontology Matching Process for PDMS. In 2nd International Conference on Data Management in Grid and P2P Systems, Linz, Austria, pages 124-135.

Pires, C. E., Sousa, P., Kedad, Z., and Salgado, A. C. (2010): Summarizing Ontology-based Schemas in PDMS. In International Workshop on Data Engineering meets the Semantic Web, 2010, Long Beach, CA, USA, pages 239-244.

Sacramento, E. R., Vidal, V. M., Macêdo, J. A., Lóscio, B. F., Lopes, F. L. R., and Casanova, M. A. (2010): Towards automatic generation of application ontologies. Journal of Information and Data Management (JIDM), 1(3):535–551.

Souza D. (2009): Using Semantics to Enhance Query Reformulation in Dynamic Distributed Environments. Ph.D. thesis, CIn/UFPE, Recife, Brazil, Apr. 2009.

Souza D., Arruda T., Salgado A. C., Tedesco P. and Kedad, Z. (2009): Using Semantics to Enhance Query Reformulation in Dynamic Environments. In Proc. of the 13th East European Conference on Advances in Databases and Information Systems (ADBIS'09), Riga, Latvia, pages 78-92.

Souza, D., Belian, R., Salgado, A. C., Tedesco, P. (2008): Towards a Context Ontology to Enhance Data Integration Processes. In Proc. of the 4th Workshop on Ontologies-based Techniques for DataBases in Information Systems and Knowledge Systems (ODBIS'08). Auckland, New Zealand, pages 49-56.

Sung, L. G. A., Ahmed, N., Blanco, R., Li, H, Soliman, M. A., and Hadaller, D. (2005): A Survey of Data Management in Peer-to-Peer Systems. School of Computer Science, University of Waterloo, 2005.