Data Integration in Dynamic Environments

Ana Carolina B. Salgado[1], Bernadette F. Lóscio[1], Maria da Conceiçao M. Batista[2]
Rosalie B. Belian[3], Carlos Eduardo Santos Pires[4], Damires Y. Souza[5]

[1] Federal University of Pernambuco (UFPE), Center for Informatics
{acs, bfl}@cin.ufpe.br
[2] Universidade Federal Rural de Pernambuco (UFRPE)
ceca@deinfo.ufrpe.br
[3] Federal University of Pernambuco (UFPE), Center of Health Sciences
rosalie.belian@ufpe.br
[4] Federal University of Campina Grande (UFCG), Computer Science Department
cesp@dsc.ufcg.edu.br
[5] Federal Institute of Education, Science and Technology of Paraiba (IFPB)
damires@ifpb.edu.br

Abstract.

One of the major database research areas is Data Integration, which refers to providing users with a uniform view over a set of heterogeneous, distributed and autonomous data sources. Data Integration settings concern, for instance, mediator-based integration systems and P2P ones (named Peer Data Management Systems). In these settings, users pose their queries without having to spend time in searching the set of distributed data sources individually. To help matters, research on Data Integration environments has considered the use of semantic knowledge in the form of ontologies and contextual information. For instance, ontologies can be used to solve the heterogeneities between the data sources, while contextual information allows the system to deal with information that is dynamically acquired during the execution of a given query. In this sense, the goal of this paper is to present the Data Integration on Dynamic Environments Group, focusing our discussions on its two main research areas: Mediation-based data integration and P2P data management.

Categories and Subject Descriptors: H. Information Systems [**H.m. Miscellaneous**]: Databases

Keywords: Context, Data integration, Information Quality, Mediation, Ontologies, PDMS, Semantics, Schema mapping

1. INTRODUCTION

The era of the Internet in the 1990's changed the way information systems were implemented. One of the first challenges was the need to represent and to search the huge amount of non-structured data spread in the Web. This challenge was solved by the adoption of HTML and the use of search engines. The second challenging phase was to include more structure and semantics in the representation of Web data, and afterwards to integrate semi-structured and structured data stored on distributed, heterogeneous and autonomous data sources. The adoption of XML, as a standard language to represent Web data, came into the focus to help to solve such problems. Later efforts to overcome the obstacles created by the permanent growth in the amount of data available on the Web, associated with the desire of inserting some level of intelligence to the retrieval of documents, have motivated the development of the new generation of the Web: the Semantic Web [Berners-Lee et al. 2001].

The emergence of the Web and its permanent growth has caused a big impact on the database research community. New database research areas emerged and existing ones evolved in order to consider the new problems arising from the need of managing the huge volume of data available on the Web. One of such areas is Data Integration, which evolved from providing solutions for distributed and heterogeneous data sources to solutions for integration of huge volumes of semi-structured data [Halevy et al. 2006; Ziegler and Dittrich 2004]. In general, this problem consists in providing a

uniform view of a set of data sources in which users can pose their queries. This is done by resolving the heterogeneities and giving to the disparate sources a uniform view. Then users submit queries over the integrated view without having to spend a lot of time in searching the set of distributed data sources.

Despite a lot of researches done on data integration, several problems still remain open. Our research group, for example, has worked on the data integration subject for more than ten years, proposing solutions for different kinds of data integration problems. Our research has been influenced mainly by the advent of XML as the standard language for representing Web data [Abiteboul et al. 1999], the use of Peer-to-Peer (P2P) architectures as a platform for data integration and the use of semantic technologies to help solving semantic heterogeneity problems.

Initially, we have proposed and implemented Integra [Lóscio 2003], a prototype of a mediator-based data integration system, which adopts the Global-as-View (GAV) approach [Halevy 2000] to define mappings (mediation queries) between the integrated schema (mediation schema) and the data source schemas. Our main contributions were associated to the specification and implementation of some important processes, including: (i) the definition and evolution of mediation queries [Lóscio 2003]; (ii) the evaluation of schema quality criteria for the generated mediation schema [Batista 2008]; and (iii) a semantic name resolution approach to the mediation schema generation process [Belian 2008], [Belian and Salgado 2010]. Later, with the emergence of P2P architectures for structured data sharing and the evolution of the Semantic Web, we proposed SPEED (Semantic PEEr-to-Peer Data Management System), an ontology-based PDMS (Peer Data Management System) in which the content shared by peers (exported schema) is represented through ontologies. In this setting, we focused our research on the development of solutions regarding the following issues: the formation and maintenance of semantic clusters of peers [Pires 2009] and the query reformulation among peers [Souza 2009].

In this paper, we present the Data Integration on Dynamic Environments Group, focusing our discussions on its two main research areas: Mediation-based data integration and P2P data management. Initially, in Section 2, we present the history of the group and members. Next, in Sections 3 and 4, we describe our main contributions for mediation-based data integration and P2P data management issues, respectively. In Section 5, we discuss the impact of our research on the society, presenting details about our international partnership. In Section 6, we describe our current research challenges, as well as the future ones. Finally, in Section 7, we present some conclusions.

2. HISTORY OF THE GROUP AND MEMBERS

Our Research Group, led by Ana Carolina Salgado, has started its research activities in Data Integration in late 1990's in the Center for Informatics (CIn) of Federal University of Pernambuco (UFPE). The main motivation to investigate this subject was the emergence of the Web, which has caused a huge interest on how to integrate heterogeneous and distributed data sources (databases and other structured or semi-structured documents). We have worked on data integration issues since then. The proposal of a mediator-based system architecture, an approach to mediation query evolution and the analysis of quality issues has led to the implementation of a data integration system prototype, called Integra. This prototype was developed as part of a research project, sponsored by the National Council for Scientific and Technological Development (CNPq)[1], whose team was composed by graduate and undergraduate students. Three of the PhD students who developed their PhD thesis in the context of this project are still part of our group: Bernadette Lóscio, Rosalie Belian, and Maria da Conceição Batista.

Currently, Bernadette Lóscio is an Associate Professor at the CIn/UFPE. After finishing her PhD, she continued to develop her researches on the data integration area, focusing mainly on the use of

[1]http://www.cnpq.br

ontologies and other Semantic Web technologies. Maria da Conceição Batista is an Associate Professor at the Federal Rural University of Pernambuco (UFRPE), where she leads a research group about Information Quality and Data Integration. Rosalie Belian is also an Associate Professor at UFPE, acting in the Health Informatics area. She continues doing research on semantic issues, which are relevant for the generation and evolution of inter-schema correspondences.

As data integration systems, Peer Data Management Systems (PDMSs) accomplish their services over data from existing heterogeneous sources, however making use of a peer-to-peer (P2P) architecture. Given the big volume of data sources in such environments, more semantic information about them becomes necessary in order to reconcile heterogeneity. Thus, we are currently working on the development of SPEED, a PDMS that clusters semantically similar peers using ontologies in order to facilitate the identification of semantic correspondences between peers'schemas (ontologies) and, consequently, improve query answering on a large number of data sources. Carlos Pires and Damires Souza were PhD candidate students concerned with the main architectural and structural definitions of SPEED. The SPEED group includes not only PhD students but also master and undergraduate students working in a complementary way to build a PDMS prototype that consolidates the main obtained results.

Currently, Carlos Pires is an Associate Professor at the Federal University of Campina Grande (UFCG), where he continues to investigate approaches to identify semantic relations between concepts. Damires Souza is an Associate Professor at the Federal Institute of Education, Science and Technology of Paraíba (IFPB), where she leads a research group and works on the development of solutions based on semantic technologies.

3. MEDIATION-BASED DATA INTEGRATION

Integrated access to multiple data sources ranging from traditional databases to semi-structured data repositories has been required in many applications. In this sense, data integration systems were proposed as tools which aim to offer a uniform access to distributed and heterogeneous data sources. Classical data integration systems are based on two approaches to data integration, each one with a specific implementation architecture:

—Virtual approach, in which the data remains in the sources and queries submitted to the data integration system are decomposed into queries addressed directly to the sources. In virtual data integration systems [Chawathe et al. 1994], a software module, called mediator [Wiederhold 1992] receives a query, decomposes it into sub-queries over the data sources and integrates its results;

—Materialized approach, in which data are previously accessed, cleaned, integrated and stored in a data warehouse [Widom 1995] and the queries submitted to the integration system are executed in this repository without direct access to the data sources.

Data integration systems can also be classified according to the approach used to define the mappings, called mediation queries, between the data sources and the mediated schema [Halevy 2000; Ullman 1997], which are used to compute each element in the mediated schema. The first approach is called Global-As-View (GAV) and requires that each object of the global schema be expressed as a view (i.e., a query) on the data sources. In the other approach, called Local-As-View (LAV), mappings are defined in an opposite way: each object in a given source is defined as a view on the mediated schema. The Global-Local-as-View (GLAV) [Friedman et al. 1999; Lenzerini 2002] and Both-as-View (BAV) [McBrien and Poulovassilis 2002] approaches combine features of both GAV and LAV approaches.

Our research on data integration has evolved over the years, owing to the technology evolution, particularly on the Web setting. Initially, our major focus was the proposal of a mediator-based data integration system, called Integra, and solutions for some of the most important data integration problems, including: mediated schema generation [Lóscio et al. 2003; Belian 2008], mediation queries

generation and maintenance [Lóscio 2003], and mediated schema quality evaluation [Batista 2008]. More recently, we had focused on the development of data integration solutions for more dynamic and flexible environments, as the dataspaces [Halevy et al. 2006] [Hedeler et al. 2009] and PDMS. The following sections provide additional information about our major contributions on mediation-based data integration, mainly related with the results obtained with the development of Integra. Our current research on dataspaces will be discussed in Section 6.

3.1 Integra Overview

One of the main contributions of our research was the proposal of Integra, a mediator-based data integration system which adopts the GAV approach and uses XML as a common model for data exchange and integration [Lóscio 2003]. Integra combines features of both data integration approaches supporting the execution of virtual and materialized queries. That way, some portions of data more intensively unavailable and static may be materialized in a data warehouse and more dynamic data may be accessed by virtual queries [Batista 2008]. Moreover, Integra uses a local cache, i.e., a repository to store prepared answers for the most frequently queries submitted to the integration system.

Figure 1 illustrates Integra's architecture, which is divided into five major spaces: (i) Common core: it offers information about local data source schemas while receives and answers local data source queries; (ii) Data integration space: its main component is the mediator which is responsible for restructuring and merging data from autonomous data sources and for providing an XML integrated view of the data. This space has other components, which are used to optimize the overall query response time of user queries, including the data warehouse and the cache; (iii) Mediator generation and maintenance space: this space is responsible for the definition and maintenance of mappings (mediation queries) between the mediated schema and the source schemas; (iv) User space: its components are used to specify and manage the evolution of user requirements, and (v) Information quality analysis space: this space is responsible for the IQ analysis and the maintenance of IQ metadata.

A major difficulty in integrating information from multiple data sources is their heterogeneous nature [Miller et al. 2001]. To overcome this limitation, Integra uses XML as a common data model for representing the sources'content and the integrated data, and it adopts XML Schema to represent sources'schemas as well as the mediated schema. Besides of XML, Integra also adopts a conceptual model for XML Schemas, called X-Entity [Lóscio et al. 2003]. The X-Entity model was proposed to provide a richer representation for XML schemas in such a way that an X-entity schema offers the necessary and relevant information required for Integra. We also proposed a formalism, based on the concept of correspondence assertions [Spaccapietra and Parent 1994] to specify correspondences between X-Entity elements. Such correspondences are required for the processes of mediation queries generation and maintenance, as it will be explained later. A detailed description about the X-Entity model and the formalism used to represent correspondences between X-Entity elements can be found in [Lóscio 2003].

3.2 Mediation Queries Definition and Maintenance

One of the main issues in a data integration system concerns the definition of mappings (mediation queries) between the mediated schema and the source schemas. A mediation query is a computing expression that describes how to compute an element of the mediated schema over the local data sources. Since we considered that an X-Entity schema represented a mediated schema and the fundamental component of an X-Entity schema is an entity, in [Lóscio and Salgado 2003] we proposed a process for mediation queries generation consisting in the discovery of a computing expression for each entity of the mediated schema. Based on such mediation queries, Integra is capable of computing at run-time the most appropriate rewriting for answering a user query by simply executing the necessary mediation queries and combining their results [Lóscio et al. 2006].

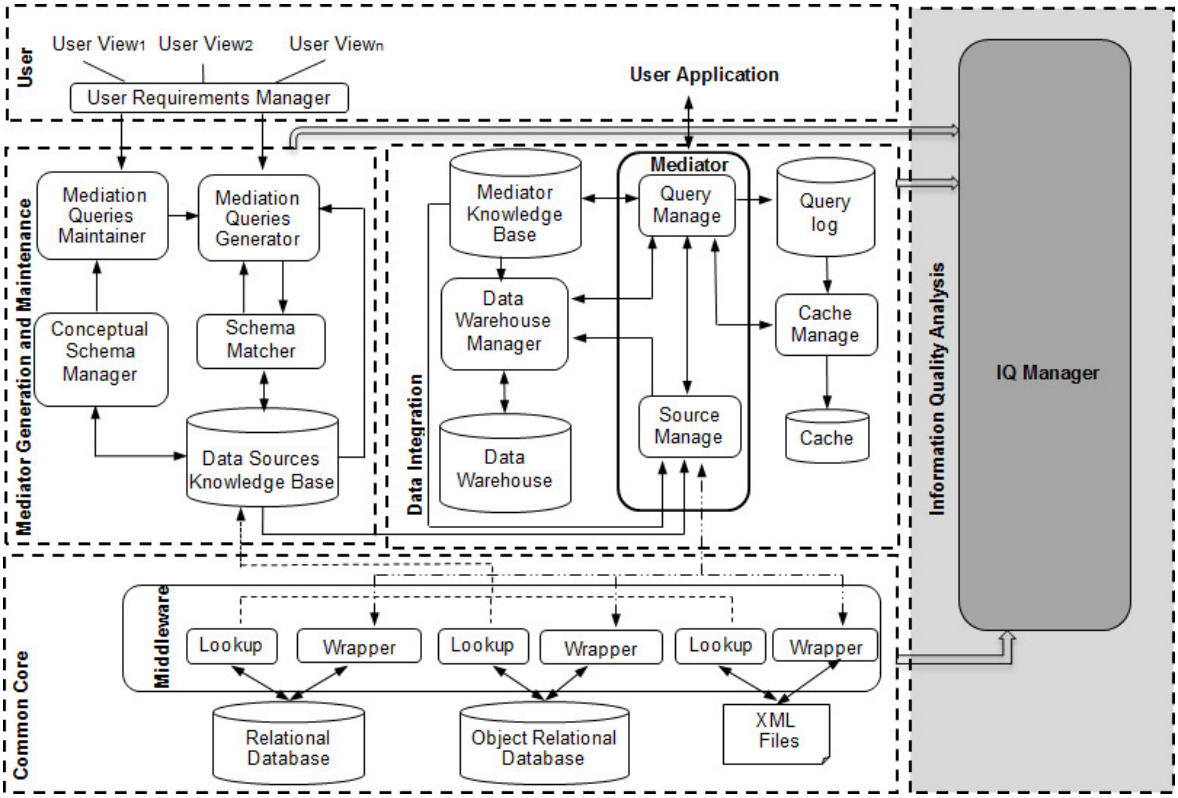

Fig. 1. The Integra Architecture

In a dynamic environment, mapping maintenance is an important task concerning data integration systems and it has been the subject of recent researches [Colazzo and Sartiani 2009; Meilicke et al. 2009; Kondylakis et al. 2009; An and Topaloglou 2008]. Mappings between the mediated schema and the source schemas must be flexible enough in order to accommodate new data. Each change at the source schema level may lead to the reconsideration and possibly the change of a set of existing mappings (mediation queries). In our research, we also have addressed this problem. In [Lóscio 2003] we have proposed an approach for propagating a change event occurring at the source level into the mediation level, in such a way that the mediation level may evolve incrementally and modifications can be handled easier increasing the system flexibility and scalability.

3.3 Semantic Issues in the Categorization of Schema Element Names

Other important data integration issue covered by our research concerns the generation of mediated schemas. In [Belian 2008] we proposed an approach for mediated schema generation based on the idea of clustering schema element names by identifying their semantic similarity [Zhou and Ram 2004]. In our proposal, the clustering process groups semantic similar names of schema elements according to available domain information and contextual knowledge [Dey 2001]. More specifically, we adopted a domain ontology [Guarino 1998] to help the identification of semantically similar elements. We also considered the use of contextual knowledge to provide a more accurate semantic interpretation, allowing restrictions or changes on the meaning of an element name according to a specific semantic context. In [Souza et al. 2008], we proposed an ontology-based solution to represent contextual knowledge relevant for data integration systems in general.

The proposed mediated schema generation process creates a set of semantic clusters representing

schema elements and mappings from each semantic cluster to every related schema element in the corresponding data source. Hence, semantic clusters and mappings may be seen as drafts to the final version of the mediated schema. At this stage, the set of semantic clusters could be processed in order to make the proper adjustments to generate the resulting mediated schema, such as: (i) naming of entities and attributes of the mediated schema elements; (ii) definition of the final attribute element characteristics; (iii) definition of the resulting relationships between elements; and (iv) resolution of structural differences, and so on.

3.4 Mediated Schema Quality Evaluation

Information Quality (IQ) has become a critical aspect in organizations and, consequently, in Information Systems research. The notion of IQ has only emerged during the past ten years and shows a steadily increasing interest. IQ is a multidimensional aspect and it is based on a set of dimensions or criteria. The role of each one is to assess and measure a specific IQ aspect [Wang and Strong 1996; Tayi and Ballou 1998; Lima et al. 2009]. More specifically, in data integration systems, Naumann and Leser [Humboldt-University et al. 1999] define a framework addressing the IQ of query processing.

In [Batista 2008] we also have addressed IQ aspects for data integration systems. We proposed IQ criteria analysis in a data integration system, mainly related to the integrated schema. Data integration systems may suffer with lack of quality in query processing, in such a way that the results can be outdated, erroneous, incomplete, inconsistent, redundant, and so on. Moreover, schemas may be barely defined, with inconsistencies and/or redundancies. As a consequence, the query execution can become rather inefficient. To minimize the impact of these problems, we propose a quality approach that serves to analyze and tries to improve the integrated schema definition, and consequently the query execution. Our hypothesis is that an acceptable alternative to optimize query execution would be the construction of good schemas, with high quality scores. We focused on the formal specification of algorithms and definitions of three schema IQ criteria: schema completeness, minimality and type consistency [Batista and Salgado 2007c; 2007a; 2007b]. Our work was used for some specializations in [Duchateau and Bellahsene 2010] and [Wang 2010]. The work presented by Duchateau & Bellahsene [Duchateau and Bellahsene 2010], for example, uses our specifications of minimality and schema completeness to create a so called expert schema. In a similar way, Wang in [Wang 2010] also uses the minimality and schema completeness criteria specification of [Batista and Salgado 2007c] to define an ontology based quality framework that focuses on user requirements.

3.5 Main Results

Due to our research on data integration systems several results were obtained. In addition to an Integra prototype that allowed the experimentation and validation of our proposal, some PhD and MSc theses were concluded and the main results were published in international and Brazilian conferences. In this research area, several projects had financial support from Brazilian institutions (CNPq and FINEP[2]) from 2003 to 2006. Moreover, an international project in cooperation with France and Uruguay (STIC- AMSUD[3]) was also developed.

4. P2P DATA MANAGEMENT

Peer Data Management Systems (PDMSs) [Kantere et al. 2009; Halevy et al. 2003] came into the focus of research as a natural extension to distributed databases in the P2P setting. PDMSs are P2P applications where each peer represents an autonomous data source which exports either its entire data schema or only a portion of it. Such schemas, named exported schemas, represent the data to

[2]http://www.finep.gov.br
[3]http://www.sticamsud.org/

be shared with the other peers in the overlay network [Androutsellis-Theotokis and Spinellis 2004]. To enable data sharing, correspondences between elements belonging to these exported schemas are generated and maintained.

Data management in PDMSs is a challenging problem given the large number of peers, their autonomous nature, and the heterogeneity of their schemas. To help matters, semantics has shown to be a helpful support for the techniques used for managing data (e.g., query answering or schema matching). We consider that semantics concerns the task of assigning meaning to schema elements [Souza 2009; Kantere et al. 2009; Penzo et al. 2008] or expressions that need to be interpreted in a given situation. In our research, semantic knowledge is mainly obtained through ontologies and contextual information. More specifically, we use ontologies in a threefold manner: (i) as a standard model to represent peer's metadata; (ii) as a mechanism to represent and store contextual information; and (iii) as background knowledge to identify semantic correspondences between matching ontologies which represent peers'schemas.

We also consider that contextual information is a major source of semantic knowledge. Context may be defined as a set of elements surrounding a domain entity of interest (e.g., user, query, and peer) which is considered relevant in a specific situation during some time interval [Bolchini et al. 2009; Souza et al. 2008]. In our research, context is used as a kind of semantic knowledge that enhances the overall query reformulation process by assisting it to deal with information that can only be acquired on the fly (e.g., according to the availability of data sources, which is perceived at run time, specific query routing strategies can be executed).

There are a number of relevant issues concerned with the use of semantics in PDMSs, which have been the focus of our research. In particular, in a PDMS scenario, a key challenge is query answering, where queries submitted at a peer are answered with data residing at that peer and with data acquired from neighbor peers through the use of schema correspondences. An important step in this task is the reformulation of a query posed at a peer into a new query expressed in terms of a target peer. Another important point is that query answering in PDMS can be improved if semantically similar peers are put together in the overlay network. In order to address such problems, in [Souza 2009] we proposed a semantic query reformulation process, while in [Pires 2009] we proposed an incremental peer clustering process considering the dynamic aspects of P2P. To offer support for these solutions, a set of ontology-based services to manipulate peer schemas was also developed. Additional information about such contributions is described in the following sections.

4.1 SPEED Overview

One of our major contributions on P2P Data Management was the proposal of SPEED (Semantic PEEr Data Management System) [Pires 2009], 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'schemas 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 (exported 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 PDMS simulator has been developed through which we were able to reproduce the main conditions characterizing SPEED's environment. The main issues, which have been particularly addressed in SPEED are described in what follows.

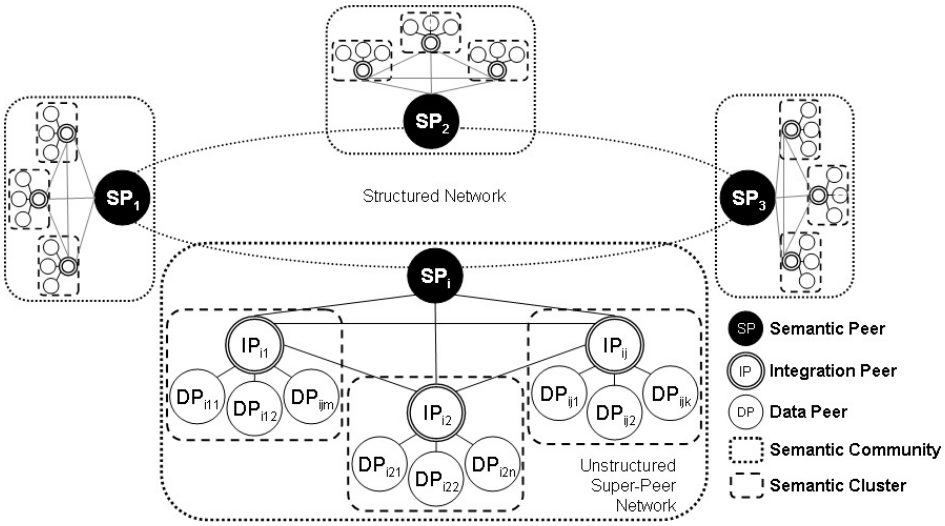

Fig. 2. The SPEED Architecture

4.2 Semantic-based Peer Clustering

Peer connection in SPEED is mainly an incremental clustering process. 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, an ontology matching service (described in Section 4.4) 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 is 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 SPEED's simulator and submitted to experimental evaluation. Validation has been performed using clustering indices.

4.3 Semantic-based Query Reformulation

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 (i.e., schema mappings) 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 2009]. Exact and enriched query reformulations are produced as a means to obtain this set of answers. To this end, besides equivalence, we use other correspondences which go beyond the ones commonly found (specialization, generalization, aggregation), proposing disjointness and closeness, identified by an ontology matching process (described in Section 4.4). Furthermore, 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. We have developed the SemRef approach within a query submission and execution module for SPEED [Souza 2009].

4.4 Ontology-based Services

As schemas are represented as ontologies in SPEED, the processes executed in the system make intensive use of ontology management services, such as matching and summarization. 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. We have also 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. Since we also use context as another kind of semantic knowledge, we have designed an ontology, named CODI - Contextual Ontology for Data Integration, to represent and store contextual information [Souza et al. 2008]. CODI is an ontology for representing context according to some Data Integration (DI) and PDMS issues. In our work, we consider that Contextual Elements (CEs) are used to characterize a given entity. Therefore, we identified six main domain entities around which we consider the CEs: user, environment, data, procedure, association, and application. We have already used CODI during 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 and Salgado 2010].

4.5 Main Results

Several results were obtained related to our research on data management in P2P systems. We have developed a SPEED prototype to allow the experimentation and validation of our proposals. In this context two PhD, two MSc and several undergraduate students have concluded their work. The main results were published in international and Brazilian conferences. In this research area, one project had financial support from CNPq from 2008 to 2011. Another one is currently being supported by CNPq.

5. IMPACT OF OUR RESEARCH ON SOCIETY

Considering the last fifteen years of research activities in data integration and related subjects, the number of people that concluded their studies in several levels are: six PhD and seventeen MSc students, and twenty two undergraduate students. Also, fourteen undergraduate students had participated in the group research activities implementing most part of the systems'modules. The students came from several parts of Brazil, most of them from the Northeast region. The ones that concluded their PhD have now positions in Brazilian universities and have their own research groups in related subjects. Currently, there are two PhD, seven MSc and ten undergraduate students participating in our research group.

The main obtained results were published in more than thirty papers in international and Brazilian forums (journals, conferences or workshops). Furthermore, the group has participated in cooperative projects with institutions in France (Université de Versailles Saint-Quentin en Yvelines and Université Paul Cézanne Aix-Marseille 3), in Uruguay (Universidad de la República) and in United Kingdom (University of Manchester) resulting in some joint publications.

6. CURRENT AND OPEN CHALLENGES

Hopefully, there are still a lot of work to be done in data integration and related areas. In this section, we present the work we are currently working on, and also what we intend to do in the future.

6.1 Current Challenges

In this section, we present a summary of the works that are being developed by the group to face some of the major challenges in the area of data integration in dynamic environments.

Some work has been developed to propose solutions for more general P2P data management problems, including: query routing and cluster balancing. Regarding query routing, semantics issues (e.g., contextual information) are being used as a way to enhance the selection of relevant semantic neighbours and their ranking. We are also analyzing quality criteria (related to peers, schemas, data, mappings) to improve query routing process. Moreover, the focus is to preserve query semantics as better as possible throughout the query reformulation process. Concerning cluster balancing, an approach is being defined to periodically balance current clusters and still maintain the semantic organization of peers in the overlay network. Clearly, some important maintenance tasks include: (i) the redistribution of data peers between the semantic neighbours of an overloaded cluster; and (ii) the merging of two clusters or the split of an existing cluster into two new clusters. Both solutions are being developed in the setting of the SPEED project.

We have also investigated some issues related to the use of ontologies as uniform conceptual representations of peer schemas. More specifically, we have been implementing a tool that automatically extracts semantics from data sources and builds peer ontologies. Meanwhile, we are working with geographic data sources in order to instantiate SPEED. When dealing with geographic data, specific problems with representation and usage occur. Thereby, we are implementing some new extraction rules for the spatial entities and relations. To this end, we have been working on extensions to the tool developed to build peer ontologies.

Also regarding ontologies, we are interested on the ontology evolution problem, which includes not only the maintenance of the ontology consistency but also the maintenance of the corresponding schema mappings. Particularly in SPEED, the cluster ontology will evolve frequently because of the dynamicity of peer connectivity and this will provoke the frequent evolution of mappings between neighbour clusters, requiring robust change management algorithms.

Other interesting ongoing work is related with the SemRef approach and the context ontology (CODI) that was developed to represent and store context. We have extended CODI in order to provide more contextual elements related to the user. We are also developing a CODI Data Service, which will be responsible for storage and retrieval of these contextual elements. This service will be coupled to the SPEED query system. Moreover, the current interface for the SPEED query module is being extended in order to provide users with a high-level interface, in such a way that both novel and experienced users can formulate their queries. Thus, query formulation by using the concepts visually provided in the ontology is being developed as well as other kinds of query formulation interfaces. We are also interested in visualization techniques in a way to show the provenance of the query results.

We are also working on problems related to dataspaces, which also follow a dynamic data integration approach. One of the distinguishing issues of such approach is that data integration is performed in a pay-as-you-go fashion [Vaz Salles et al. 2007], requiring the development of more flexible solutions. Particularly, we are interested on the evaluation and improvement of IQ criteria for some dataspace processes as, for example, schema integration, schema mappings, user queries and user feedback, considering the dynamicity and flexibility of such environment. Moreover, we are investigating solutions to help evaluating the impact of adding new data sources to a given dataspace. Our idea consists in using previously submitted queries and user feedback already gathered to annotate the new data

source with relevant information quality. Such information will be used to identify if the new data source has a positive impact on the dataspace and, therefore, should be included in the system.

PDMS simulators need to associate a local database schema to each peer in the simulated overlay network. To make sense, the local schemas should belong to the same application domain. Depending on the domain that is considered, it can be easy to find a small number of related schemas. However, in order to simulate a realistic and large-scale PDMS environment, one often needs a very large number of schemas. Moreover, these schemas must belong to a common domain and exhibit sufficiently large overlap (so as to allow the definition of mappings between them). On the other hand, there must also be some kind of heterogeneity among the schemas in order to better test the strengths and weaknesses of the simulated approaches. Clearly, one possible solution is to generate these schemas manually, which is time consuming and error-prone, and thus does not scale. Instead, we are working on an automatic approach to generate multiple synthetic database schemas to be used in PDMS simulations.

More specific subjects of our current research are: (i) an approach to identify semantic relations between tags in a folksonomy using similarity measures such as co-occurrence, cosine and folkrank; and (ii) the development of tools to help publishing datasets as linked data.

6.2 Open Challenges

There is a growing need for data integration in several environments: Web Information Systems (WISs), Peer Data Management Systems (PDMSs), Enterprise Information Integration (EII), Dataspaces, to say some. In this sense, even if a lot of work has already been done, it is important to highlight some future (or ongoing) trends in data integration systems (in all mentioned environments):

—data sources'description: are metadata enough?
—entity resolution: what data objects refer to the same real-world entity?
—'automatic'generation and evolution of schema mappings or correspondences: where and when must the user intervene?
—reformulation and query answering: what are the query plans?
—quality criteria: how to measure data and information quality?
—scalability and performance: what about the dynamic environments ?
—reasoning: how reasoning mechanisms provided by contextual information may enhance data integration processes?
—schema mapping: how to incrementally generate mappings between two or more schemas?
—mediation schema or global schema: how to choose the best mediation schema at query execution time?
—data sources: how to choose the 'best 'data sources to answer a given query?
—user feedback: how to gather use feedback? how to use it to improve the data integration envioronment?

It is clear that we need more semantics to handle information mainly in heterogeneous environments where we have to share data and thus to reconcile terminologies. Actually, data sources in general do not have enough metadata to allow this task and the use of ontologies as a common conceptualization is essential. Existing ontology matchers must consider in their algorithms all the necessary semantic information to measure similarity.

In the last years, we saw the growth of context-based or context-sensitive systems. The concept of context is larger than space and time information usually employed in ubiquitous systems. In fact, contextual information about users, applications, environments, data and relationships, may also be employed to improve data management processes. The open issues are how to automatically acquire and manage contextual information in data management systems.

To deal with semantic issues a huge interaction with Artificial Intelligence (AI) discipline is crucial. We need to use techniques of knowledge representation (including ontologies and contextual information), logics (causal, temporal and probabilistic), rules and inference, description logics, machine learning, only to refer some of them.

All the challenging data integration problems related to data uncertainty, inconsistency, incompleteness, provenance and trust are related to data quality. Once identified the relevant data quality criteria, it is necessary to identify metrics to evaluate them. Some works have been proposed but as a multidimensional concept it is not so simple to define, assess and apply usable data quality metrics.

Given the increasing number of data sources it becomes almost impossible to integrate the source schemas using a one-shot strategy, e.g., a strategy where several schemas are integrated in a single step integration schema. In order to deal with the scalability of the systems, solutions to perform the incremental definition of correspondences and mappings between schemas must be developed. In a similar way, its not coherent to have just one global schema (integrated view) which specifies the set of all user requirements. Strategies for modeling, generating and maintaining a set of global schemas must also be developed. In more flexible environments, where mappings and global schemas are identified incrementally it is also important to consider how user feedback can be used to improve such mappings or schemas [Khalid et al. 2010]. It is worth mentioning that such solutions must always taking into account that the environments are very dynamic (data sources may join or leave the system frequently) and the data sources may have both semi-structured and unstructured data sources, as the data available on the Web.

7. CONCLUSION

Data integration has become the focus of several researches in the last decades. However, despite such efforts numerous open problems remain unsolved. In this paper we presented our contributions to some relevant data integration problems. Particularly, we have presented two data integration environments, namely: (i) Integra, a mediation-based data integration system, which uses XML as a common model and combines features in order to support the execution of virtual and materialized queries; and (ii) SPEED, a PDMS which employs semantic knowledge by means of ontologies and contextual information to support peer clustering, considering load balancing problem, and query reformulation and routing.

This work has highlighted some of our main contributions, including: (i) the definition and maintenance of schema mappings, which are meaningful issues to data integration systems, moreover to the ones that deal with dynamic environments; (ii) an information quality approach with the specification of quality criteria assessment methods for the evaluation of integrated schemas with objectives of achieving better query execution time. We also proposed an algorithm to improve the schemas'minimality score; (iii) the use of semantics in a clustering-based approach to group schema element names in order to generate an integrated schema; (iv) the use of semantics surrounding data sources, data, schemas and schema mappings (or correspondences) to improve the query reformulation process; and (v) the use of semantics to help clustering similar data sources.

In a general way, this paper has discussed the benefits of using semantics to deal with data management issues in dynamic data integration environments. There are a number of on going research issues concerned with the use of semantics in the settings described. We intend to use semantics and quality criteria to enhance the selection of relevant semantic neighbours in order to improve query routing, ontologies, schemas and mappings definitions. Furthermore, we have been carrying out experiments to obtain feedback from users in order to improve query answering tasks. We are also investigating the impact of adding new data sources into an existing schema, considering both user queries previously defined the user feedback already gathered.

REFERENCES

Abiteboul, S., Buneman, P., and Suciu, D. *Data on the Web: From Relations to Semistructured Data and XML*. Morgan Kaufmann, 1999.

An, Y. and Topaloglou, T. Semantic web, ontologies and databases. Springer-Verlag, Berlin, Heidelberg, Maintaining Semantic Mappings between Database Schemas and Ontologies, pp. 138–152, 2008.

Androutsellis-Theotokis, S. and Spinellis, D. A survey of peer-to-peer content distribution technologies. *ACM Comput. Surv.* vol. 36, pp. 335–371, December, 2004.

Batista, M. C. M. *Schema Quality Analysis in a Data Integration System*. Ph.D. thesis, Federal University of Pernambuco, 2008.

Batista, M. C. M. and Salgado, A. C. Data integration schema analysis: An approach with information quality. In *ICIQ*. pp. 447–461, 2007a.

Batista, M. C. M. and Salgado, A. C. Data integration schema analysis: An approach with information quality. In *ICIQ*. pp. 447–461, 2007b.

Batista, M. C. M. and Salgado, A. C. Minimality quality criterion evaluation for integrated schemas. In *ICDIM*. pp. 436–441, 2007c.

Belian, R. B. *A Context-based Name Resolution Approach for Semantic Schema Integration*. Ph.D. thesis, Federal University of Pernambuco, 2008.

Belian, R. B. and Salgado, A. C. A context-based schema integration process applied to healthcare data sources. In *OTM Workshops*. pp. 100–109, 2010.

Berners-Lee, T., Lassila, O., and Hendler, J. The semantic web. *Scientific American*, May, 2001.

Bolchini, C., Curino, C. A., Orsi, G., Quintarelli, E., Rossato, R., Schreiber, F. A., and Tanca, L. And what can context do for data? *Communications of the ACM* 52 (11): 136–140, November, 2009.

Chawathe, S., Garcia-Molina, H., Hammer, J., Ireland, K., Papakonstantinou, Y., Ullman, J. D., and Widom, J. The TSIMMIS Project: Integration of heterogeneous information sources. In *Proc. of the 16th Meeting of the Information Processing Society of Japan*, 1994.

Colazzo, D. and Sartiani, C. Detection of corrupted schema mappings in xml data integration systems. *ACM Trans. Internet Technol.* vol. 9, pp. 14:1–14:53, October, 2009.

Dey, A. K. Understanding and using context. *Personal and Ubiquitous Computing* 5 (1): 4–7, 2001.

Duchateau, F. and Bellahsene, Z. Measuring the quality of an integrated schema. In *ER*. pp. 261–273, 2010.

Friedman, M., Levy, A., and Millstein, T. Navigational plans for data integration. In *In Proceedings of the National Conference on Artificial Intelligence (AAAI*. pp. 67–73, 1999.

Guarino, N. Formal ontology and information systems. In *FOIS98 - Conference on formal ontology in information systems*. pp. 3–15, 1998.

Halevy, A. Y. Theory of answering queries using views. *SIGMOD Record*, 2000.

Halevy, A. Y., Franklin, M. J., and Maier, D. Principles of dataspace systems. In *PODS*. pp. 1–9, 2006.

Halevy, A. Y., Ives, Z. G., Mork, P., and Tatarinov, I. Piazza: data management infrastructure for semantic web applications. In *Proceedings of the 12th international conference on World Wide Web*. WWW '03. New York, NY, USA, pp. 556–567, 2003.

Halevy, A. Y., Rajaraman, A., and Ordille, J. J. Data integration: The teenage years. In *VLDB*. pp. 9–16, 2006.

Hedeler, C., Belhajjame, K., Fernandes, A. A., Embury, S. M., and Paton, N. W. Dimensions of dataspaces. In *BNCOD 26: Proceedings of the 26th British National Conference on Databases*. Springer-Verlag, Berlin, Heidelberg, pp. 55–66, 2009.

Humboldt-University, F. N., Naumann, F., Leser, U., and Freytag, J. C. Quality-driven integration of heterogeneous information systems. In *In VLDB Conference*. pp. 447–458, 1999.

Kantere, V., Tsoumakos, D., Sellis, T., and Roussopoulos, N. Groupeer: Dynamic clustering of p2p databases. *Inf. Syst.* vol. 34, pp. 62–86, March, 2009.

Khalid, B., Paton, N. W., Embury, S. M., Fernandes, A. A. A., and Hedeler, C. Feedback-based annotation, selection and refinement of schema mappings for dataspaces. In *Proceedings of the 13th International Conference on Extending Database Technology*. EDBT '10. ACM, New York, NY, USA, pp. 573–584, 2010.

Kondylakis, H., Flouris, G., and Plexousakis, D. Ontology and schema evolution in data integration: Review and assessment. In *Proceedings of the Confederated International Conferences, CoopIS, DOA, IS, and ODBASE 2009 on On the Move to Meaningful Internet Systems: Part II*. OTM '09. Berlin, Heidelberg, pp. 932–947, 2009.

Lenzerini, M. Data integration: a theoretical perspective. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. PODS '02. New York, NY, USA, pp. 233–246, 2002.

Lima, C. R. d. A., Schramm, J. M. d. A., Coeli, C. M., and da Silva, M. E. M. Review of data quality dimensions and applied methods in the evaluation of health information systems. *Cad Saude Publica* 25 (10): 2095–109, 2009.

Lóscio, B. F. *Managing the evolution of XMl-based data integration systems*. Ph.D. thesis, Federal University of Pernambuco (UFPE), Brazil, 2003.

Lóscio, B. F., Costa, T., and Salgado, A. C. Query reformulation for an xml-based data integration system. In *SAC*. Dijon, France, pp. 498–502, 2006.

Lóscio, B. F. and Salgado, A. C. Generating mediation queries for xml-based data integration systems. In *SBBD*. Manaus, Brazil, pp. 99–113, 2003.

Lóscio, B. F., Salgado, A. C., and Rêgo Galvão, L. Conceptual modeling of xml schemas. In *Proceedings of the 5th ACM international workshop on Web information and data management*. WIDM '03. pp. 102–105, 2003.

McBrien, P. and Poulovassilis, A. Schema evolution in heterogeneous database architectures, a schema transformation approach. In *Proceedings of the 14th International Conference on Advanced Information Systems Engineering*. CAiSE '02. London, UK, UK, pp. 484–499, 2002.

Meilicke, C., Stuckenschmidt, H., and Tamilin, A. Reasoning support for mapping revision. *J. Log. and Comput.* vol. 19, pp. 807–829, October, 2009.

Miller, R. J., Hernández, M. A., Haas, L. M., Yan, L., Howard Ho, C. T., Fagin, R., and Popa, L. The clio project: managing heterogeneity. *SIGMOD Rec.* vol. 30, pp. 78–83, March, 2001.

Penzo, W., Lodi, S., Mandreoli, F., Martoglia, R., and Sassatelli, S. Semantic peer, here are the neighbors you want! In *Proceedings of the 11th international conference on Extending database technology: Advances in database technology*. EDBT '08. ACM, New York, NY, USA, pp. 26–37, 2008.

Pires, C., Sousa, P., Kedad, Z., and Salgado, A. Summarizing ontology-based schemas in pdms. In *Data Engineering Workshops (ICDEW), 2010 IEEE 26th International Conference on*. pp. 239 –244, 2010.

Pires, C., Souza, D., Pachêco, T., and Salgado, A. A semantic-based ontology matching process for pdms. In *Data Management in Grid and Peer-to-Peer Systems*, A. Hameurlain and A. Tjoa (Eds.). Lecture Notes in Computer Science, vol. 5697. Springer Berlin / Heidelberg, pp. 124–135, 2009.

Pires, C. E. S. *Ontology-based Clustering in a Peer Data Management System*. Ph.D. thesis, Federal University of Pernambuco, Recife, PE, Brazil, 2009.

Souza, D. *Using Semantics to Enhance Query Reformulation in Dynamic Distributed Environments*. Ph.D. thesis, CIn, Federal University of Pernambuco, Brazil, 2009.

Souza, D., Belian, R., Salgado, A. C., and Tedesco, P. A. Towards a context ontology to enhance data integration processes. In *ODBIS*. Auckland, New Zealand, pp. 49–56, 2008.

Spaccapietra, S. and Parent, C. View integration: A step forward in solving structural conflicts. *IEEE Trans. on Knowl. and Data Eng.* vol. 6, pp. 258–274, April, 1994.

Tayi, G. K. and Ballou, D. P. Examining data quality. *Commun. ACM* vol. 41, pp. 54–57, February, 1998.

Ullman, J. D. Information integration using logical views. In *Proceedings of the 6th International Conference on Database Theory*. London, UK, pp. 19–40, 1997.

Vaz Salles, M. A., Dittrich, J.-P., Karakashian, S. K., Girard, O. R., and Blunschi, L. itrails: pay-as-you-go information integration in dataspaces. In *Proceedings of the 33rd international conference on Very large data bases*. VLDB '07. VLDB Endowment, pp. 663–674, 2007.

Wang, J. A quality framework for data integration. In *BNCOD*, 2010.

Wang, R. Y. and Strong, D. M. Beyond accuracy: what data quality means to data consumers. *J. Manage. Inf. Syst.* vol. 12, pp. 5–33, March, 1996.

Widom, J. Research problems in data warehousing. In *International Conference on Information and Knowledge Management*. pp. 25–30, 1995.

Wiederhold, G. Mediators in the architecture of future information systems. *IEEE Computer* 25 (3): 38–49, 1992.

Zhou, H. and Ram, S. Clustering schema elements for semantic integration of heterogeneous data sources. *J. Database Manag.* 15 (4): 88–106, 2004.

Ziegler, P. and Dittrich, K. R. Three Decades of Data Integration - All Problems Solved? In *18th IFIP World Computer Congress (WCC 2004), Volume 12, Building the Information Society*. IFIP International Federation for Information Processing, vol. 156. Kluwer Academic Publishers, Toulouse, France, August 22-27, pp. 3–12, 2004.