Abstract. We propose an extension of a mediator architecture. This extension is oriented to ontology-driven data integration. In our architecture ontologies are not managed by an external component or service, but are integrated in the mediation layer. This approach implies rethinking the mediator design, but at the same time provides advantages from a database perspective. Some of these advantages include the application of optimization and evaluation techniques that use and combine information from all abstraction levels (physical schema, logical schema and semantic information defined by ontology).

1. Introduction

Although the Web is probably the richest information repository in human history, users cannot specify what they want from it. Two major problems that arise in current search engines (Heflin, 2001) are: a) polysemy, when the same word is used with different meanings; b) synonymy, when two different words have the same meaning. Polysemy causes irrelevant information retrieval. On the other hand, synonymy produces lost of useful documents. The lack of a capability to understand the context of the words and the relationships among required terms, explains many of the lost and false results produced by search engines.

The Semantic Web will bring structure to the meaningful content of Web pages, giving semantic relationships among terms and possibly avoiding the previous problems. Various proposals have appeared for meta-data representation and communication standards, and other services and tools that may eventually merge into the global Semantic Web (Bemers-Lee, 2001). Hopefully, in the next few years we will see the universal adoption of open standards for representation and sharing of meta-information. In this environment, software agents roaming from page to page can readily carry out sophisticated tasks for users (Berners-Lee, 2001).

In this context, ontologies can be seen as metadata that represent semantic of data; providing a knowledge domain standard vocabulary, like DTDs and XML Schema do. If its pages were so structured, the Web could be seen as a heterogeneous collection of autonomous databases. This suggests that techniques developed in the Database area could be useful. Database research mainly deals with efficient storage and retrieval and with powerful query languages.

The database community has been seriously disturbed with the recent Web technologies expansion. Particularly, two reports have produced special commotion in database field. The first one, the Asilomar report (Bernstein, 1998), postulated the new directives in databases tendencies, previewing the Web impact in this field. The second one, Breaking out the Box (Silberschatz, 1996, 1997), proposes how database community must transfer its technology to be introduced into Web technology. In this context we have broken out the database box and into its
autonomous functional components, and we are using these to reach a solution for the problem of heterogeneous service integration.

In recent years, mediators are responsible of integrating databases, using wrappers to translate the different data source information to a common model. They require a single schema for query processing. This means that possible local data source schema extension will not be available until the mediator and its wrappers are updated. Ontology-based systems use the semantics of data for data source integration. They must supply ontology query processing. So they do not need a single schema, giving dynamism to query processing.

2. Background

Wrappers were the first building block for Web integration. They act as interfaces to each data source, providing (semi) structure to non-structured sources or mapping the original data source structure to a common one. Unfortunately, it is very difficult to switch unstructured data into a specific schema. Issues related to wrapper design and implementation can be found in (Roth, 1997), (Sahuguet, 1999a; 1999b).

The knowledge about evaluating a query over multiple wrappers is encapsulated by mediators. The wrapper-mediator approach provides an interface to a group of (semi) structured data sources, combining their local schemas in a global one and integrating the information of local sources. So the views of the data that mediators offer are coherent, performing semantic reconciliation of the common data model representations carried out by the wrappers. One of the main problems in data integration is related to the maintaining of the integrity constraints. As this problem is not yet solved, mediators need to deal with the problem of evaluating consistent queries over possibly inconsistent sources with respect to the common schema constraints. Some good examples of the wrapper-mediator systems are AMOS (Fahl, 1993), TSIMMIS (Garcia, 1995), DISCO (Tomasic, 1995), GARLIC (Roth, 1996;), (Hass, 1997; 1999). Recently, many of these approaches have moved toward XML standard, like AMOS and TSIMMIS. On the other hand, MIX (Baru, 1999) (the successor to the TSIMMIS project) and MOCHA (Rodriguez, 2000) projects are initially XML-based.

The next level of abstraction on Web integration corresponds to ontology-based systems. Its main advantage over the mediators is their capacity for managing a priori unknown schemas. This is achieved by means of a mechanism that allows contents and query capabilities of the data source to be described declaratively. From the data perspective, Ontologies enrich the semantics of the schema, resolving synonymy and polysemy problems, viewed at the beginning of the paper.

The reader can find an excellent review of the state of the art in ontology engineering in (Corcho, 2001). Well-known environments for building ontologies are WebODE (Arpirez, 2001) and WebOnto (Domingue, 1998). The first one provides an API for ontology access (based on Prolog) and import/export utilities from/to diverse markup and ontology languages. WebOnto is a powerful collaborative environment focused on the ontology creation and navigation.

From the integration point of view, many studies have been and are still being developed using ontologies. At the moment this paper is being written we can remark on two main projects: Ariadne (Arens, 1996), (Baris, 2000) and Observer (Mena, 2000). Ariadne aims at the development of technologies and tools for
rapidly constructing intelligent agents to extract, query, and integrate data from Web sources. Observer uses different ontologies to represent information data sources. The user explicitly selects the ontology that will be used for query evaluation. The existence of mapping among ontologies allows the user to change the ontology initially selected.

The intended domain of our framework is closed to Observer. We designed an architectural framework for the integration of both heterogeneous online data sources into Web services and heterogeneous online services (Aldana, 2001b). This architectural framework enables the capability of an easy development and deployment of small and autonomous cooperating services. Furthermore, meta-information described in the services allows the cooperation and integration among them. In (Aldana, 2001a) we presented three different integration models: query delegation, service cooperation and semantic integration. Now our objective is to extend the framework meta-data component to support new concepts that emerge with the Semantic Web.

3. Integration Component

Traditionally, heterogeneous data source integration systems have been developed according to onion architecture. Following the nomenclature in (Stuckenschmidt, 2002), there are three layers: Syntax, Structural, and Semantic. Wrappers are in the syntax layer, which is the onion core. Mediators added a new layer, the structural one, covering wrappers. Currently, the most external layer is the semantic one that fits with ontology issues. Many of the developed systems match this architecture, or have been designed following it.

In Aldana (2001b) we presented a framework for the construction of services as Web component, mediators, which only reach the structural level. This framework follows some W3C consortium recommendations and latest Web technologies (Abitebould, 2000). XML is an appropriated standard for data interchange and nowadays is one of the most generalized ones. Our framework uses XML-Schemas and RDF-Schemas to represent schemas and specific information.

Our Framework helps the system developer to support them in the construction of services that integrate heterogeneous data sources, building a global schema for them (conceptual, logical and physical schemas). To construct such schema the developer has to know the data sources schemas and has to integrate them manually. The query processor accesses the data sources using the wrappers and combines their answer to present the final result to the user. This process is hidden to the user, who views the system as a single data source. The Framework architecture is shown in figure 1.

We replace the Framework Metadata Component (Aldana, 2001b) with a new one called Integration Component (IC) to support new Semantic Web emerging concepts. This new component achieves query processing over meta-information stored in the schemas. Besides, it allows the ontologies inclusion. Ontologies establish a joint terminology between members of a community of interest. These members can be human or automated agents. To represent a conceptualization, a representation language is need. However, for applications on the web it is important to have a language with a standardized syntax. Because XML emerges as the standard language for data interchange on the web, it is also
desirable to exchange ontologies using XML syntax. The integration component architecture is shown in figure 3.

The IC has the same metadata information as the metadata component plus the ontology information. This architecture allows performing two levels of query processing. The first one, the IC-processing, generates a complex schema from several heterogeneous data sources, using ontologies. The second one, the Query Processor Component Processing, performs specific application domain tasks. This two-level structure separates knowledge organization and the computing issues. Under this structural design changes in schema complexity, the use of alternative ontology algorithms, changes in user views, could be performed without affecting applications, being unnecessary modifications in their programming code.

Among other uses, we are especially interested in the ontology as a powerful mechanism for schema integration and schema enhancement. Given that the initial framework design was very close to a database organization, we have proposed an approximation to ontology-driven integration being consistent with our approach. On the other hand, ontology enriches the base model providing a set of valid models by applying axioms and rules. Since our interest is focused on how ontologies interact with the schema, ontology facilities will cover schema definitions, instead of being a new layer over the whole mediation service. Doing so we need to redesign the Metadata component of our architecture, but we gain some advantages.

We are more interested in the construction of mediators rather than in a specific mediation system, and in the study, development and trials of new
algorithms for Semantic Web Integration. Particularly we pretend to apply distributed deductive database evaluation and optimization techniques.

Extensions of Datalog (Ceri, 1990), a deductive database language, have been used as a formalism for Ontology definition. We think it could be very interesting to adapt many Datalog specific techniques to ontology-driven integration. Especially when mediation and ontology reside both into an integrated system, because it allows full access to mediator schema and ontology. For example, this allows the use of semantic query optimization techniques that use ontological constraints, schema constraints and physical level information (like access time statistics, the presence of indexation, etc.).

One of the most relevant aspects that the Web imposes in a database is the distribution aspect. Information distribution not only has effects on the efficiency of a database, but it affects the whole database management system, from data consistency and integrity issues to query processing. From a theoretical point of view, the Web favors the existence of distributed autonomous schemas. Under these circumstances, and considering ontologies as a specification model for advanced schemas, we are interested in data integration when ontologies, as a schema, are also distributed.

Our database research line was initially focused on deductive databases, with a special interest in distribution aspects. Recursion management in a distributed system is not an easy issue, and is more complicated in an environment like Internet. Questions related with cycle detection, evaluation synchronization and termination detection must be considered and studied carefully.

Finally, we think that an important factor for the growth of the Semantic Web is related with the users capability to query multiple ontologies that have been related in a decentralized way, using references or single relationships among its terms. This does not exclude but complements specifically constructed global ontologies. Currently, mediators can interoperate in a similar way as a distributed database does. So, we believe that is justified enough to fit ontologies together with the schema of the mediator, instead of locating them on top of it.

4. Conclusions and Future Work

We propose an extension of a mediator architecture dealing with ontology-driven data integration. In our architecture, ontology is not managed by an external component or service, but is integrated in the mediation layer.

The Framework Metadata Component (2001b) is replaced by a new component called the Integration Component. This new component achieves query processing over meta-information stored in the schemas. Besides, it allows ontology processing.

This approach implies rethinking the mediator design, but at the same time provides advantages from a database perspective. Particularly, it enables the application of optimization and evaluation techniques that use and combine information from all abstraction levels (physical schema, logical schema and semantic information defined by ontology), which is the focus of our current research.
References
Baru, Gupta, Ludaescher, Marciano, Papakonstantinou & Velikhov. (1999). “XML-Based Information Mediation with MIX”. In Exhibitions Program of ACM-SIGMOD.


