Database Technologies on the Web

J.F. Aldana Montes
Universidad de Málaga, Spain

A.C. Gómez Lora
Universidad de Málaga, Spain

N. Moreno Vergara
Universidad de Málaga, Spain

I. Navas Delgado
Universidad de Málaga, Spain

M.M. Roldán García
Universidad de Málaga, Spain

INTRODUCTION

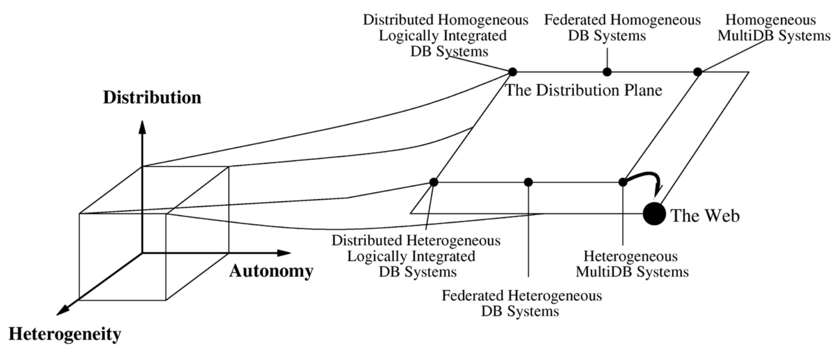
Database community has been seriously disturbed with the Web technologies expansion. Particularly, two reports have produced a special commotion in database field. The first one, the Asilomar report (Bernstein et al., 1998), postulates the new directives in databases tendencies, previewing the Web impact in this field. The second one, Breaking out the Box (Silberschatz & Zdonik, 1996), proposes how database community must transfer its technology to be introduced into Web technology. In this sense, the database box must be broken out into its autonomous functional components, and they must be used to reach a solution for the problem of heterogeneous data sources integration.

BACKGROUND

Thinking about the Web as a huge, highly distributed database, we may consider different dimensions to conceptually describe it. Tamer Özsu and Valduriez (1999) define a classification of database systems with respect to: 1) their distribution; 2) the autonomy of local systems; and 3) the heterogeneity of database systems. The autonomy concept is considered as the distribution of control, not of data. This indicates the degree to which individual DBMSs can operate independently. Whereas autonomy refers to the distribution of control, the distribution dimension deals with the physical distribution of data over multiple sites. With respect to heterogeneity, this can range from hardware heterogeneity, differences in networking protocols, variations in DBMSs, and so forth, to the data model or the policy for managing integrity on the database.

Obviously, the Web is in the distribution plane, and, as shown in Figure 1, we think that "it falls out" of the cube because it presents the highest degree of distribution, heterogeneity, and autonomy. Therefore, traditional distributed database techniques must be further extended to deal with this new environment in order to face Web environment intrinsic problems, such as the management of semi-structured information.

Figure 1. Extending the cube



A lot of work has still to be carried out in the database community to resolve all of the issues related to such a kind of distributed and heterogeneous database, which is what the Web actually is.

Due to space limitations, we only review some related and open issues, such as the integrity problem, query optimisation problem, integration issues in both Web and Semantic Web. But the reader must note that this coverage is not complete. There are some interesting issues uncovered, such as data models, security, transaction processing, and so forth.

THE INTEGRITY PROBLEM IN THE WEB

Some important factors related to this issue include the structure of the underlying database upon which the constraints are imposed, the nature of the imposed constraints, and the method adopted for their evaluation. Traditionally, two restrictions are focused in databases: domain and structural restrictions. So they must be considered in a Web data model.

From a database perspective, XML Schema (2001) provides a new technological standard which enables us to represent data semantics like a database does. We can find a discussion about schema languages in Lee and Chu (2000).

A domain restriction defines the set of values that an attribute may have. XML-Schema provides enhanced data types and user-defined data types. New data types can be created from base data types specifying values for one or more facets for the base data type. Moreover, on XML-Schema we can define subclasses and superclasses of types. We can also disable the restriction or extension of a particular type.

With respect to structural restrictions, in this schema language, we can represent:

1. Uniqueness for attribute. Furthermore, XML Schemas specify uniqueness not only for attributes but also for arbitrary elements or even composite objects.
2. Key for attribute.
3. Foreign key for attribute; using the *key/keyref* mechanism, which complements the *id/idref* one of previous versions of XML Schema and solves its associated problems.

QUERY OPTIMISATION IN THE WEB

Optimisation on regular path queries (Grahne & Thomo, 2000, 2001) and indexing techniques over semi-structured information (McHugh & Widom, 1999) have already been studied. One of the most important studies on XML and semi-structured data optimisation (McHugh & Widom, 1999) has been developed for the LOREL system (Abiteboul, Quass, McHugh, Widom & Wiener, 1997). LOREL not only supports the traditional value index, but also label, edge, and path indexes.

However, other relevant aspects, such as composition reordering, restriction propagations, and more complex and sophisticated techniques, magic rewriting for example (Bancilhon, Maier, Sagiv & Ullman, 1986), are well known in database systems and have yet to be tested in the XML context.

From the XML perspective, constraint propagation cannot be used exclusively for query optimisation, but it can and must be used during new type derivation or inference (Fan, Kuper & Siméon 2001). This is generically called subtyping. One of the main characteristics of XML is that it allows new type extensions and restrictions through inheritance. The computation of this new set of constraints is fundamental in object-oriented systems, and very important in preventing integrity constraints violations in our schema.

Constraint derivation in subtyping is useful with domain and column constraints, but it is also useful with entity and referential constraints (Fan, Kuper, & Siméon, 2001). The existence of physical index associated with keys and references in order to accelerate relationship composition is common.

INTEGRATION ISSUES IN THE WEB

Wrappers were the first building block on 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. 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. Table 1 summarizes the most popular wrapper-mediator systems. The next level of abstraction on Web integration corresponds to ontology-based systems. From the data perspective, ontologies enrich the semantic of the schema, resolving synonymy and polysemy problems (Heflin, 2001). The reader can find an excellent review in ontology engineering in (Corcho, Fernández-López & Gómez-Pérez, 2001). Table 2 summarizes the most popular ontology-based systems.