Data Integration and Interrogation

Jules Verso*
INRIA-Rocquencourt
BP 105, 78153, Le Chesnay Cedex, France
verso@inria.fr

Abstract

One major concern of the Verso group at Inria is the development of technology for data integration and interrogation, especially for non-traditional data formats such as structured text. This demonstration of aspects of Verso's technology is partly sponsored by the European Community (AQUARELLE project, Esprit IV projects OPAL and WIRE). Related publications can be found on Verso's web page1.

1. Introduction

The emergence of the Web has placed new demands on database technology and has posed new and significant problems for database research. In particular, this medium requires new ways of querying information that is to be found on the Web such as structured or semi-structured documents (SGML, HTML).

Existing tools for finding Web documents are based on information retrieval techniques. They provide sophisticated tools for pattern matching and browsing, but they are limited in their ability to deal with the structure of documents, the presence of non-textual data, and the larger structure of the Web. On the other hand, database research has provided declarative languages for dealing with structured data. We believe that database techniques will greatly enhance the way we make use of the Web and similar media, but significant extensions are needed to database technology before it can be used for interrogating documents and integrating them with other forms of data.

2. Interrogation

Database query languages generally lack features that are essential for documents on the Web. People wanting to browse such documents typically have little a priori knowledge of the HTML or SGML structure, and there is no direct way of representing the variety of document structures in a simple uniform database schema. Therefore we need query languages that do not require exact knowledge of a schema before they can be used. In fact we need a model and language that blurs the existing distinction between data and schema. To this end we have designed and implemented the Path Object Query Language (POQL) that allows schema and data to be queried in a homogeneous fashion. We will demonstrate the power of this language and our progress in extending it with Web interfaces and full-text indexes.

3. Integration

Database queries combining information from multiple sources require the understanding of the mapping between these sources and the database. A first step towards this integration effort was the design and implementation of a tool for mapping SGML documents into ODMG databases. However, this tool is restricted to a single mapping onto a single format. We are currently working on developing an interactive tool for specifying mappings between various formats. Clearly, the scope of this tool goes beyond file to database mappings. For example, it can be used for generating HTML documents from other data formats. As another example of data integration, we want to transport relational data to an object-oriented database. The main challenge here is to maintain efficiency in both the transformation process and in the resulting object-oriented implementation.

*Serge Abiteboul, Bernd Amann, Sihem Amer-Yahia, Elia Carsenat, Vassilis Christophides, Sophie Cluet, Claude Delobel, Michel Scholl, Jérôme Siméon, Fanny Wattez
1http://www-rocq.inria.fr/verso/