CEREALAB database: data integration with the MOMIS System

ID - 195

Sala Antonio[1], Bergamaschi Sonia[1]

[1]Dipartimento di Ingegneria dell'Informazione, Università di Modena e Reggio Emilia, Modena

Motivation

Biological informations are frequently widespread over the Web and retrieving knowledge in this domain often requires to navigate through several websites. Moreover, data sources are usually heterogeneuos and present different structures and interfaces. Mediator systems can be used to perform integration of such databases in order to have integrated view of multiple information sources and to query them.

The MOMIS system (Mediator envirOnment for Multiple Information Sources) is a framework developed by the Database Group of the University of Modena and Reggio Emilia (www.dbgroup.unimo.it) to perform intelligent information integration from both structured and semistructured data sources.

The result of the integration process is a Global Virtual View (GVV) of the underlying sources which is a conceptualization of the underlying domain and then may be thought of as an ontology describing the involved sources. Moreover, queries can be posed over the GVV regardless of the structure of the local sources in a transparent way for the user.

Methods

MOMIS performs information extraction and integration from both structured and semistructured data sources. Information integration is performed in a semi-automatic way, by exploiting the knowledge in a Common Thesaurus (defined by the framework and based on Wordnet) and descriptions of source schemas with a combination of clustering techniques and Description Logics inferences. Mapping rules are then specified to handle heterogeneity. In particular, we have experimented the MOMIS System for the realization of the CEREALAB database. CEREALAB is a research project of technology transfer for applying Marker Assisted Selection (MAS) techniques to Cereal Breeding in Italian Seed Companies. It regards four different species of cereals: durum and bread wheat, maize, rice and barley.

Genotypic information coming from two existing databases available on the web, Gramene (http://www.gramene.org/) and Graingenes (http://wheat.pw.usda.gov/), have been integrated with phenotypic data extracted from the American Germplasm Resources Information Network (GRIN, http://www.ars-grin.gov/) and with data coming from the CEREALAB research project. CEREALAB database performs both the tasks of providing a valid support for the research activity, supplying information from the existing databases, and being a knowledge base to store the data obtained by researchers.

Results

The result of the integration process is a global virtual schema (GVV) of the underlying sources which can be thought as an ontology regarding both phenotypic and genotypic information about cereals. Users pose queries over the GVV and the MOMIS query manager transparently execute them. Query execution is performed by: decomposing a global query into local queries to be executed locally by the data sources; local answers are then fused and reconciled to return the answer to the user.

A graphical interface has been developed for supporting visual query formulation; visual query are automatically translated into SQL queries supported by the MOMIS Query Manager. This interface is useful for users, like biologists, who do not have specific information technology skills. In this way users can have access to data coming from different data source through a single interface without taking care about the structure or the query interface of each single database.

Furthermore, the GVV can be exported in OWL thus guaranteeing interoperability with other external applications/ontologies or external users.

The project is conducted in collaboration with the Agrarian Faculty of the University of Modena and Reggio Emilia and funded by the Regional Government of Emilia Romagna.

Email: antonio.sala@unimore.it