A mediation framework for a transparent access to biological data sources *

The MediaGRID project[1,2,3] - http://www-lsr.imag.fr/mediagrid
Contact : Chrsitine.Collet@imag.fr

[1]LSR-IMAG Laboratory - UMR 5526 BP 72, 38402 Saint-Martin d'Hères Cedex, France
[2]PRiSM Laboratory - UMR 8636 Université de Versailles St-Quentin, 78035 Versailles Cedex, France
[3]LaMI Laboratory - UMR 8042 Univ. d'Evry-Val-d'Essone, Genopole Evry, 9100 Evry, France

Thanks to recent advents in information technology and the expansion of networks, in particular internet, access to geographically dispersed information is no longer an issue. Examples being search engines, digital libraries, and electronic store servers. Information of interest often requires accessing and combining data stored within sources using different data formats and models. Therefore, there is a need for tools and mechanisms that get, transform and combine such data. In this regard, mediation systems have been proposed. The idea behind is to provide clients with the illusion of dealing with a unique data source whereby handling data heterogeneity and abstracting them from internal functionalities, e.g., query processing. Generally speaking, a client query is performed in the following stages. Provided a client query, it is first rewritten and spited into local sub-queries according to a global mediation schema. Data sources then perform such queries and returns results. Finally, returned results are combined and visualized to the client. Though such an approach has proved its relevance, it suffers from some limitations. Query processing may demand a considerable period of time to end. In the meantime, clients have no mean for monitoring the query processing, modifying it, or even getting partial results. Data sources on the other hand may be modified regarding schema or data. New data sources may appear and be of interest to the client query and conversely other may be made unavailable. Our research addresses such issues by specifically proposing a mediation framework for building mediation systems able to generate mediation queries and to evaluate queries in an interactive and dynamic fashion. It considers a Global as View approach and XML as data exchange format.

Mediation queries generation Mediation queries [2] are queries computing the global schema from exported schemas of local sources. These queries are generated by using meta-information stored in a metadata server. Examples of necessary meta-information are descriptions of mediation schema and exported schemas of local sources, the correspondence between the mediation schema and local exported ones. Generating mediation queries are processed in three steps: (i)looking for pertinent portions of sources, (ii)identifying the candidate operations between sources (ii)generating the optimal query computing global schema from exported schemas. Generated mediation queries are used as input of unfolding algorithm to rewrite client queries into local sub-queries.

*This work is supported by the French Ministry of Research through the ACI-GRID program.

Iterative and dynamic query processing Client query concept is separated into two parts: *query core* defining what the result will be and *query context* determining constraints (user interests) on processing queries, e.g. number of output, period of execution time, preference. We design an evaluator driven by user interest. To deal with unpredictabilities of the environment, we use rules defining behaviors of evaluator according to network delays, unavailabilities of sources, user interaction for refining queries[4]. In order to augment the interactivity of query processing, the query evaluator must authorize *partial results* when data is missing. Looking at the first (incomplete) results, users can refine their long running queries. Additionally, capabilities of source are taken into consideration in order to delegate tasks to them and avoid a huge data transfer over the net.

Application Our framework [1] will be validated within biological context [3]. Particularly, it will be applied on sources giving information related to genes cartography and expression. The target is to provide biologists with means to correlate expression levels of a gene whose data are stored within different sources and observe their evolution. Performing such a task requires first selecting relevant data sources and then discovering correlations among them thereby being able to integrate data they give. During the processing, partial results have to be supported and given to biologists progressively so that they can intervene.

References

[1] Gennaro Bruno. Adems, an adaptable and extensible mediation service : application to biological sources. In *Proceedings of the 28th International conference on VLDB 2002, Ph.D. Doctoral Poster Session, Hong Kong, China*, 2002.

[2] Zoubida Kedad and Mokrane Bouzeghoub. Discovering view expressions from a multi-source information system. In *Proceedings of the Fourth IFCIS International Conference on Cooperative Information Systems, Edinburgh, Scotland*. IEEE Computer Society, 1999.

[3] MediaGrid Project. A mediation framework for biological data sources. In *European Conference on Computational Biology (ECCB), Paris, France*, 2003.

[4] Tuyet-Trinh Vu and Christine Collet. Query brokers for distributed and flexible query evaluation. In *Proceedings of the Conference RIVF, Hanoi, Vietnam*, 2003.