A mediation framework for a transparent access to biological data sources

The MEDIAGRID project [1,2,3] - http://www-lsr.imag.fr/mediagrid
Contact: Christine.Collet@imag.fr
[1] Laboratoire LSR IMAG - UMR 5526 BP 72, 38402 Saint-Martin d'Hères Cedex
[2] Laboratoire PRiSM - UMR 8636 Universite' de Versailles St-Quentin 78035 Versailles Cedex
[3] Laboratoire LaMI - UMR 8042 Univ. d'Evry-Val-d'Essone, Genopole Evry, 91000 Evry

keywords: Mediation systems, Query evaluation, Gene expression

Data sources mediation in biology

Nowadays, data are distributed over the net and are stored in data sources with different formats and data models. All this heterogeneity makes the querying task more difficult because users have to cultivate their knowledge in terms of formats, query languages and data models. Instead of using different tools for querying, users need tools allowing data formats and access interfaces homogenisation. To resolve these problems, *mediation systems* have been proposed [1].

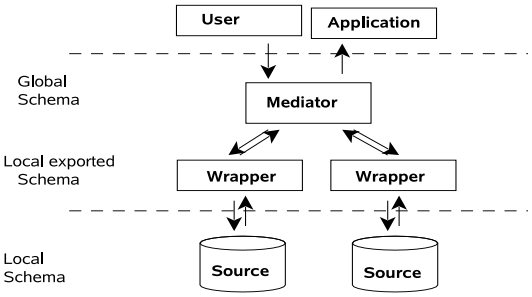

Figure 1. Mediation system

Figure 1 shows a classical mediation system architecture. Users and applications access data contained in local sources through a *mediator*. A mediator provides a global schema and translates global expressions into local ones according to the characteristics of concerned data sources. Local expressions are then sent to local sources for their evaluation. As local sources are heterogeneous, a mediator is connected to local sources through *wrappers*. A wrapper is a piece of software which transforms local expressions from the mediator into queries understandable for the local source's. Local sources export their query capabilities to the system by providing an *exported local schema*, and necessary metadata for mapping it to the local schema.

Biological data sources are highly heteregeneous in terms of content types (sequences, structures, analysis results, bibliography, *etc.*) and data storage models. Several mediation systems for biology sources have been developped. The most important ones are: TINet/OPM [4], DiscoveyLink [5], K2/Kleisli [3] and TAMBIS [6]. All existing biological mediation systems provide a minimal infrastructure for mediating heterogeneous and distributed sources, but none of them is able to represent local sources capabilities and availability. The main cause for this lack of flexibility is attributed to their unefficient metadata management. Existing mediation systems are not able to represent characteristics of sources such as functional dependencies or semantic equivalence, nor to optimise queries dynamically. All these aspects are essential to react in case of local sources unavailability and to provide alternative or semantically-equivalent query plans. The MEDIAGRID project has as objective to better take up these challenges.

MEDIAGRID project

MEDIAGRID is a multidisciplinary projet supported by French Minister of Research (ACI-GRID). Its general objective is to provide a mediation framework for a transparent access to heterogenous data sources. It adopts a classical architecture based on wrappers and mediators and considers a Global as View approach [2]. Mediation systems built from such a framework are able to: consider sources containing weakly structured data which are generated by applications and stored as HTML or XML files; authorise partial results for queries in case of data sources unavailability; and be efficient even if queries are very complex and/or net traffic slows down.

To achieve these objectives, we focus our research topics in: (i) mediation queries generation, (ii) iterative and dynamic query evaluation, and (iii) validating our approach in biological data sources mediation.

Mediation queries generation

Mediation queries are generated by using metainformation, which is stored in a metadata server. This server collects all metainformation on local sources and is accessed by mediators for retrieving useful metainformation for generating queries. Examples of necessary metadata are *intra-schema assertions*, such as functional dependencies, referential constraints and constraints on the values, and *inter-schema assertions* such as compatible domains, semantic equivalence between attribues and instances of key attributes. Using metadata, our mediation system is able to select relevant sources and to identify the candidate operations between sources for generating the optimal query

Iterative and dynamic query evaluation

After being generated, queries must be optimised and executed on distributed sources. To do this, meta-information play a very important role. Indeed, a query evaluator must take into account sources capabilities in order to delegate tasks to them and avoid a huge data transfer over the net. Estimation of response time is very difficult in distributed environments. Because of the random response time to access data, it is not possible to optimise queries a priori. For this reason we think that query plan should be modified dynamically according to changes produced within the execution context. Moreover, the choice of the best query must be the result of an interactive and iterative process in which the user may interact, *i.e.* the user may look at the first obtained results and give more information to refine the query. Building such an architecture means developing a query evaluator authorising *partial results*.

Biological data sources mediation

The biological context chosen for validating our approach consists of correlating expression levels of each gene with its genomic location and observing its evolution. Performing a such task requires the capability of selecting relevant sources (e.g. cartographic and gene expression information), integrating them for discovering interesting correlations. Evaluating the pertinence and accuracy of generated queries and end/partial results must be done by biologists.

References

[1] Gio Wiederhold. Mediators in the architecture of future information systems. *Computer*, 25(3):38–49, 1992.

[2] D. Calvanese, D. Lembo, M. Lenzerini *Survey on methods for query rewriting and query answering using views*, Technical Report D2I (Integration, Warehousing and Mining of Heterogeneous Data Sources) Project - Report D1.R5, 2001.

[3] S. B. Davidson, J. Crabtree, B. Brunk, J. Schug, V. Tannen, C. Overton, and C. Stoeckert. K2kleisli and gus: Experiments in integrated access to genomic data sources, 2001.

[4] B. A. Eckman, A. S. Kosky, and L. L. A. Extending traditional query-based integration approaches for functional characterization of post-genomic data. *Bioinformatics*, 2001.

[5] L. Haas, J. Rice, P. Schwarz, W. Swope, P. Kodali, and E. Kotlar. Discoverylink : A system for integrated access to life sciences data sources. *Deep computing for the life sciences*, 40(2), 2001.

[6] R. Stevens, P. Baker, S. Bechhofer, G. Ng, A. Jacoby, N. Paton, C. Goble, and A. Brass. TAMBIS : Transparent Access to Multiple Bioinformatics Information Sources. *Bioinformatics*, 2000.