BIRN-M: A Semantic Mediator for Solving Real-World Neuroscience Problems *

Amarnath Gupta
Univ. of California San Diego
La Jolla, CA, USA
gupta@sdsc.edu

Bertram Ludäscher
Univ. of California San Diego
La Jolla, CA, USA
ludaesch@sdsc.edu

Maryann E. Martone
Univ. of California San Diego
La Jolla, CA, USA
mmartone@ucsd.edu

1. INTRODUCTION

A goal of the Biomedical Informatics Research Network (BIRN) project is to develop a multi-institution information management system for Neurosciences to gain a deeper understanding of several neurological disorders. Each institution specializes in a different subdiscipline and produces a database of its experimental or computationally derived data; a mediator module performs semantic integration over the databases to enable neuroscientists to perform analyses that could not be done from any single institution's data. The overall system architecture of the BIRN system is that of a wrapper-mediator system. The information sources are various relational sources including Oracle 9i having user-defined packages, Oracle 8i with the Spatial Data Cartridge, and databases made available over the web. Sources also include *computational resources* that need to be "run" to produce data. A source can be either accessed directly by the mediator, or through a middleware called the *Storage Resource Broker* (SRB) to shield the mediator from keeping low-level information like data location and user authentication information, and from large object handling. Wrappers handle source registration, schema mapping, and translation from the mediator's logic queries to the corresponding SQL or web request. BIRN-M also admits *knowledge sources*, which host ontologies and domain-specific general knowledge in the form of logic rules. BIRN-M maintains a registry with semantic schemas from all sources, integrated views defined in a GAV fashion, and source query capabilities as binding patterns and special predicates or functions admissible by it. The registry also records the types of data a client can handle, and has an API for interactive custom-built clients.

2. THE DEMONSTRATION

The demonstration features the entire operational system and contains at least the following components:

A Variety of Clients: The demonstration will show three different clients developed for the user community. The first is a specialized SQL interface adapted to redirect the output of a query into any other specialized client and to chain multiple queries. The second client has an interface that allows the user to navigate large ontology graphs visually, and launch queries. The third is the Atlas interface which allows brain mappers to specify queries with spatial operations.

A Variety of Remote Neuroscience Sources: We have developed specialized data models to represent a number of ADTs specific to neuroscience data. For example, we have created a data model to represent brain surfaces as a triangular mesh where each node point has a relational tuple for structured data and an XML object for variable-structured information. Similarly, for the graph-structured vocabulary of the Unified Medical Language System (UMLS) path and subgraph extraction queries can be executed.

The Query Rewriting Module: BIRN-M exports mediated views that are defined by *integrated view definitions* Given a user query q against the integrated views v, the mediator has to rewrite q into a (maximally contained) q' over the registered sources s. This distributed query planning includes an unfolding procedure LAV several query rewriting steps that take into account semantic constraints and the limited query capabilities of sources. BIRN-M includes novel techniques for handling plans that are partially infeasible due to binding pattern restrictions: At *compile-time*, a procedure for scheduling goals determines whether all partial plans are feasible and reports the results. Infeasible partial plans are split into their feasible and infeasible components.

Semantic Registration: A wrapper engineer can map data types of a source to data types of the mediator. A registration tool extracts catalog information (including indexed fields, key and check constraints, user-defined views and functions etc.). The wrapper engineer can define an export schema over the extracted fields. Additional semantic information can be provided, including the semantic types mediator have of the data, a class hierarchy, range-restrictions on attributes, inter-class constraints etc.

Capability-based Translation: The mediator-side of the capabilities based translation involves using the appropriate set of built-in functions produced from the user-defined functions registered by the wrapper. The wrapper-side of the translation consists of a query listener, a Prolog to SQL compiler and a query executor. Source-specific queries are generated by recognizing special goals or terms in Prolog queries, e.g., *descendant()* is a special function that translates to Oracle hierarchical query having a CONNECT BY expression.

Additional authors: Haiyun He, Xufei Qian, Arcot Rajasekar, Edward Ross, Simone Santini and Ilya Zaslavsky, UCSD

*This work is supported by NIH BIRN-CC Award No. 8P41 RR08605-08S1, NSF NPACI Neuroscience Thrust Award No. ACI 9619020, NIH Human Brain Project Award No. 5RO1DC03192

SIGMOD 2003, June 9-12, 2003, San Diego, CA.