SIMS: Retrieving and Integrating Information From Multiple Sources*

Yigal Arens & Craig Knoblock
USC/Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292
U.S.A.
+1 310 822-1511
arens@isi.edu, knoblock@isi.edu

Introduction

Most tasks performed by users of complex information systems involve interaction with multiple knowledge- and data- bases. Examples can be found in the areas of analysis, resource planning and briefing applications. Retrieval of desired information dispersed in multiple sources requires general familiarity with their contents and structure, with their query languages, with their location on existing networks, and more. The user must break down a given retrieval task into a sequence of actual queries to databases and/or knowledge bases, and must handle the temporary storing and possible transformation of intermediate results — all this while satisfying constraints on reliability of the results and the cost of the retrieval process.

With a large number of information sources, it is difficult to find individuals who possess the required knowledge, and automation becomes a necessity.

There is an elegant solution to the problem described above: the creation of a knowledge server that will form the interface between information sources and applications in need of that information. A user, or an application, will query the knowledge server in a manner that is independent of the distribution of information over various sources, independent of the various query languages, the location of data- and knowledge- bases, etc. It will be the task of the knowledge server to determine how to obtain the desired information, which data sources to use, how and where to temporarily store and manipulate intermediate data, and how to maintain a satisfactory level of efficiency in performing its task.

SIMS[1] is such a server.

The SIMS Knowledge Server

The SIMS project has set itself the aim of solving research issues involved in creating such a knowledge server, and constructing a prototype dealing specifically with queries in the transporation planning domain — organizing the movement of personnel and materiel from one geographic location to another using available transportation facilities and vehicles.

SIMS accepts queries in the form of a description of a class of objects about which information is desired. This description is composed of statements in the Loom knowledge representation language. The user is not presumed to know how information is distributed over the data- and knowledge bases to which SIMS has access — but he/she is assumed to be familiar with the application domain, and to use standard terminology to compose the Loom query. SIMS' interface enables the user to inspect the domain model as an aid to composing queries.

SIMS proceeds to reformulate the user's query as a collection of more elementary statements that refer to data stored in available information sources. SIMS then creates a plan for retrieving the desired information, establishing the order and content of the various plan steps/subqueries. Using modeled knowledge about the contents and structure of information sources, SIMS reformulates the plan to minimize its expected execution time.

The resulting plan is then executed by performing local data manipulation and/or passing subqueries to the

*The research reported here was supported by Rome Laboratory of the Air Force Systems Command and the Defense Advanced Research Projects Agency under contract no. F30602-91-C-0081. Views and conclusions contained in this report are the authors' and should not be interpreted as representing the official opinion or policy of DARPA, RL, the U.S. Government, or any person or agency connected with them.

SIGMOD /5/93/Washington, DC,USA
© 1993 ACM 0-89791-592-5/93/0005/0562...$1.50

[1] Services and Information Management for decision Systems.

LIM system (developed by Paramax Corp.) which generates the final translation into database queries in the appropriate language(s). A graphical user interface enables the user to inspect the plan in its various stages and to supervise its execution.

While there has been some work on this problem in the database community (for example, [1, 2]), our approach differs in that a complete semantic model of the application domain is created and used in order to provide a collection of terms with which to describe the contents of (i.e., to create a semantic model of) available information sources. In contrast to previous work, the model is not specific to a particular set of information sources and there is not necessarily a direct mapping from the concepts in the model to the objects in the information source. This approach supports a much more flexible and easily extensible interface to a collection of information sources.

SIMS applies and extends a variety of AI techniques and systems to build an intelligent interface to information sources. SIMS builds on the following research:

First, the Loom knowledge representation language is used to describe the domain about which information is stored in the information sources, as well the structure and contents of the information sources themselves. The domain model is a declarative description of the objects and activities possible in the application domain as seen by a typical user. The model of each information source should indicates the data-model used, query language, network location, size estimates, update frequency, etc., and describe the contents of its fields in terms of the domain model. The user formulates queries using terms from the application domain, without needing to know anything about specific information sources. SIMS' models of different information sources are completely independent, greatly easing the processes of enlarging the system.

Second, the Prodigy planning system is used to construct a sequence of queries to individual information sources that will satisfy the user's query. A planner is used in an initial reformulation step that selects the information sources to be used in answering a query. It is also used to order the queries to the individual information sources, select the location for processing the intermediate data, and determine which queries can be executed in parallel.

Third, the Loom Interface Manager (LIM) is used to provide direct access to Oracle databases. LIM translates a Loom query that SIMS has determined accesses only a single database into the appropriate set of SQL statements for retrieving the data. After retrieving the data, it builds an appropriate set of Loom objects, which SIMS can then manipulate and combine with data from other sources.

Current Status

An initial prototype knowledge server has been built and applied to the domain of transportation planning. The system currently has access to nine Oracle databases and a Loom knowledge base with information about ships, ports, locations, relevant activities, etc. SIMS provides a graphical user interface for monitoring the planning and execution of queries and for constructing queries and models. It is written in Common Lisp and uses CLIM for its graphics.

The SIMS system provides an elegant and extensible solution to the problem of accessing data that is being stored and maintained at distributed remote sites. It provides an efficient and flexible interface for accessing this data by both users and application programs.

Acknowledgments

Thanks to Chin Chee and Chun-Nan Hsu, other members of the SIMS project, for their substantial contributions to the effort described in this paper. Thanks to Manuela Veloso for providing us with her code for generating partial orders. Thanks also to the LIM project for providing us with the databases and the database models used for our work, as well many sample queries.

References

[1] Umeshwar Dayal. Query processing in a multidatabase system. In *Query Processing in Database Systems*, pages 81–108. Springer Verlag, New York, 1985.

[2] R. Hull and R. King. Semantic database modeling: Survey, applications, and research issues. *ACM Computing Surveys*, 19(3):201–260, 1987.