8-25-1995

Negotiation in Database Schema Integration

Gillian Hall
McGill University

Negotiation in Database Schema Integration

Gillian Hall, Faculty of Management, McGill University, Montreal, PQ, Canada

1 Introduction

Databases are playing an increasingly important role in organizations. Timely, accurate access to information has become a critical component of gaining competitive advantage. Data availability is commonly perceived as a critical success factor for an organizationÕs long-term survival, and day-to-day operations can be crippled by failure of the database system to satisfy user requirements. However, a number of emerging issues complicate organizationsÕ ability to provide comprehensive and reliable access to disparate information resources. Further, data accessibility is often compromised due to the typically high cost associated with addressing these issues in practice.

Examples of such issues which have emerged in the past decade include the proliferation and investment in autonomous databases within organizations, heterogeneity among data models and database management systems employed, the increasingly important role of distributed systems, and the increasing complexity and knowledge-intensive nature of integrating database schemas. All these factors contribute to the increasing importance of developing feasible options for providing interoperability among existing databases, and therefore, of pursuing research in the area of database schema integration. Indeed, this research focuses specifically on knowledge requirement problems involved in integrating the schema of existing databases in order to provide interoperability and transparent access to disparate information resources without the investment involved in complete systems redesign.

2 Schema Integration

As database design methodologies emerged in the 1970s, one of the fundamental motivations for using the database approach over the traditional Òdata-processing-using-filesÓ approach was the assertion that database management systems would make it possible to define an integrated schema of relevant data for all applications, thereby eliminating duplication, avoiding problems of multiple updates, and minimizing inconsistencies across applications (Batini, 1986). These important advantages have motivated research in the area of schema integration over the past two decades.

The general objective of schema integration is to integrate an organizationÕs different proposed or existing database systems and user perceptions of the world thereby facilitating global access to an integrated organizational information resource. However, schema integration research has been specialized into two areas: 1) View integration which addresses ÔproposedÕ databases, and 2) Database integration which addresses ÔexistingÕ databases. View integration is used as a bottom-up database design tool and produces a global conceptual description of a proposed database by merging different data requirements or user ÒviewsÓ. On the other hand, database integration is used to produce a global schema representing a collection of related databases throughout an

organization. This global schema is a virtual view of all databases taken together in a distributed database environment. While database and view integration differ contextually, they can both be described as the activities of integrating the schemas of existing or proposed databases into a global, unified schema (Batini, 1986) which satisfies constraints imposed by all component schemas.

3 Causes of Schema Diversity

It is the concurrent satisfaction of all component schema constraints which is the achilles heel of schema integration research. Such constraint satisfaction, as already mentioned, can be complicated by data model heterogeneity. For example, in an Entity-Relationship (ER) model, a generalization hierarchy may be represented using Òis aÓ relationships, while in an extended ER model, the same construct might be modeled using generalization relationships, and in the relational model, there is no construct specifically for modeling abstractions. A number of other data models have been introduced in the literature as well, each with its own constructs for representing relationships between data. Because the integration of schemas subject to different data model constraints quickly becomes unmanageably complex, some authors have advocated the translation of all component schemas into the same data model prior to attempting integration. Such translation, however, has the potential to violate component schema constraints and the autonomy of component schemas. The existence of data model constraint heterogeneity, therefore, is a significant issue in schema integration.

A more pervasive and universal complication in ensuring the satisfaction of component schema constraints in schema integration is that different user groups and designers adopt their own perspectives or ÒviewsÓ of the data depending upon their specific requirements and the relevance of data items within the context of their work processes. The result is differing representations of the data or the use of different modeling constructs to represent semantically equivalent real-world objects. Even when using the same data model, differing representations of the same real-world objects may easily occur because most commonly used data models are expressive enough to represent the same application domain equivalently using a number of different constructs. Representational differences among related database schemas result in conflicts during schema integration, a major complication in any schema integration process. In fact, when schemas are developed by different user groups or designers, while the reality being modeled may be equivalent, some constructs in the resulting schemas may actually be incompatible, and either the conflicting construct or knowledge regarding that constructÕs relationship to other schemas must be modified before integration may take place.

To complicate matters further, while equivalent concepts may be expressed differently in different schemas, other semantic relationships may also exist between constructs in component schemas -- and as with equivalency relationships, other types of relationships may also be modeled with different constructs in different schemas. Such semantic relationships are generally classified as set theoretic relationships such as exclusion, inclusion, and intersection. Far more than equivalence relationships, these other types of

interschema relationships significantly complicate the schema integration task, especially when heterogeneous data models are used to represent component schemas.

4 Distributing the Knowledge Burden

Many different approaches to schema integration have been presented over the past two decades, each addressing the problems of inconsistencies between local schemas and reconciling those inconsistencies in different ways. These schema integration approaches have evolved to reflect new data models and new technologies. Currently, a number of issues for schema integration research are emerging from technological advances and changes in organizational information systems. As already introduced, these issues include data model and DBMS heterogeneity and autonomy of existing databases.

Each of these issues is closely related to the knowledge intensive nature of the schema integration task, the cause of the primary limitation in schema integration research to date. To address the knowledge requirements of the schema integration task, the assumption of a human interactor with globally complete and correct knowledge regarding all component database schemas is at the foundation of most schema integration methodologies. This assumption is becoming increasingly unrealistic in the face of issues such as autonomy and heterogeneity. In addition, as databases become larger, more complex, and less centralized, and as organizations grow, change, and decentralize, the likelihood of having one individual with adequate global knowledge to make intelligent schema integration decisions becomes increasingly remote.

Therefore, while database schema integration methodologies must begin to address issues such as autonomy and heterogeneity (as in Spaccapietra and Parent, 1994), schema integration research must also begin to explore ways to redress traditional reliance upon a single human for global knowledge and expertise regarding integration strategy, data models, and the semantics and relationships of and between component database schemas. Because the databases involved in integration efforts are generally decentralized, there will be a number of people with expertise about those component databases. When the burden of knowledge in the schema integration process is reduced to a local component schema, and decision-making regarding constraint satisfaction is reduced from the global to the local scale, the likelihood of one person or user group having adequate knowledge to make sound decisions is far greater than when complete global knowledge is required.

5 A Role for Negotiation

The natural next step in database schema integration is a distributed technique which decentralizes the knowledge used in managing a decentralized information resource. Cooperative problem solving or negotiation between cooperating experts offers such decentralization. In the context of federated database systems, the dialogue between the administrators of component databases or between the administrator of a component database and the administrator has been called negotiation (Sheth and Larson, 1990). The purpose of such dialogue may be to reach agreement about allowable access and

operations to component schemas, to agree upon the semantic relationships between structures in component schemas, or to satisfy constraints inherent to component database schemas. Negotiation protocols, while addressed by a number of researchers in the area of federated and multidatabase systems (Sheth and Larson, 1990; Elmasri, Larson and Navathe, 1986; Litwin, 1990; Dayal et. al., 1984), is generally not addressed or accommodated within the framework of schema integration methodologies. Instead, negotiation is usually associated with the preintegration process or with the maintenance of a multidatabase or federated database system (e.g., adjusting for structural changes to component schemas). However, the concept of negotiation offers significant advantages throughout the process of database schema integration. Foremost of these is the potential for alleviating reliance upon an individual for global knowledge regarding all component schemas, and instead distributing reliance upon human expertise among local experts on component database schemas.

Such negotiation and the coordination of multiple Ôknowledge sourcesÕ offers an important next step in schema integration research. The purpose of negotiation in this context is to allow schema integration and conflict resolution decisions to be made based upon localized goal, constraint, integration strategy, and interschema correspondence knowledge which need not be globally known. In so doing, the assumption of global knowledge may be discarded. In its place, a more realistic assumption can be made that, given a local database schema, there is an individual with complete knowledge regarding that local schema. Thereafter, the coordination of these local ÔexpertsÕ through a process of ÔnegotiatedÕ schema integration becomes the next challenge.

Polat, et. al. (1993) presents a model for distributed conflict resolution among cooperating expert systems for the design of an office. This model in which individual intelligent agents ÒnegotiateÓ conflict resolution through the use of a partitioned, shared blackboard is generically applicable to any number of design tasks. The model presented is based on the idea that each design agent has its own conflict resolution expertise separate from its domain-level design expertise, and that in the context of particular conflicts, this expertise can be instantiated into specific advice for resolving these conflicts. The model allows a new problem-solver to be added or an existing one to be removed without requiring any modification to the rest of the system, thereby taking advantage of the perks offered by open-systems architectures, the very architectures on which distributed interoperable database systems are being constructed.

In the context of schema integration, specifically the approach presented by Spaccapietra and Parent (1994), each local database within an organization may be represented in a knowledge-based system in which a local conceptual schema, data model constraints, semantic constraints, and interschema correspondences are captured. Using interschema correspondence assertion declarations and integration algorithms presented in Spaccapietra and Parent (1994), an integrated conceptual schema may be proposed by an agent when another local agent presents a new schema for integration. Other agents, representing other local databases, may then examine the proposed integrated schema, critique it, and suggest modifications when conflicts with their own local constraints are detected. When a proposed integrated schema satisfies the constraints of all applicable

agents, it is accepted. Following a negotiation protocol proposed by Polat et. al. (1993), the agents, each with its own localized perspective and constraints, attempt to reach consensus. With the addition of a stopping rule for ÒunresolvableÓ conflicts, this model is well-suited to the task of database integration. In addition, it fully supports the notion of autonomy, allowing the structural preservation of component schemas. Further, because each agent in the system uses local knowledge, integration rules, and conflict resolution strategies, the Polat model supports the notion of data model and constraint heterogeneity.

6 Conclusions and Future Work

While a schema integration method which addresses some of the important issues such as heterogeneity and autonomy is a critical component of furthering research in this area, automating such a method within an architecture appropriate to the task is also an important component of such research efforts. The increasing complexity and importance of the schema integration task make automated tools an inevitable component of schema integration research to come.

In the context of database schema integration, negotiation is an inevitable component in reaching consensus with regard to interschema correspondences, constraints, conflicts, access privileges, etc. The investigation of the effectiveness of a negotiating experts architecture for database schema integration through the development and validation of a negotiating experts testbed system based upon the Polat et. al. (1993) architecture will provide a contribution both to the growing body of schema integration research and to the successful management of increasingly distributed and disparate database systems within organizations.

References

Batini, C. & Lenzerini, M. A Methodology for Data Schema Integration in the Entity-Relationship Model, IEEE Transactions on Software Engineering, November, 1984, SE-10(6), 650-664.

Batini, C., Lenzerini, M. & Navathe, S.B. A Comparative Analysis of Methodologies for Database Schema Integration, ACM Computing Surveys, December, 1986, 18(4), 323-364.

Biskup, J. & Convent, B. A Formal View Integration Method, SIGMOD Record, June, 1986, 15(2), 398-407.

Breitbart, Y., Olson, P.L. & Thompson, G.R. Database Integration in a Distributed Heterogeneous Database System, Proceedings of the 2nd International Conference on Data Engineering, Los Angeles, CA, 1986, 301-310.

Buneman, P., Davidson, S. & Kosky, A. Theoretical Aspects of Schema Merging, Advances in Database Technology - EDBT Ô92 , Berlin, Springer-Verlag, 1992, 152-167.

Casanova, M.A. & Vidal, V.M.P. Towards a Sound View Integration Methodology, Proceedings of the 2nd ACM SIGMOD/SIGACT Symposium on the Principles of Database Systems, Atlanta, GA, 1983, 36-47.

Chatterjee, A. & Segev, A. Data Manipulation in Heterogeneous Databases, SIGMOD Record, December, 1991, 20(4), 64-68.

Dayal, U. & Hwang, H.-Y. View Definition and Generalization for Database Integration in a Multidatabase System, IEEE Transactions on Software Engineering, November, 1984, SE-10(6), 628-645.

Elmasri, R., Larson, J. & Navathe, S.B. Schema Integration Algorithms for Federated Databases and Logical Database Design, Honeywell Corporate Research Center (Technical Report), 1986.

Jardine, D. & Yazid, S. Integration of Information Submodels, Information Systems Concepts: An In-Depth Analysis, 1989, 247-267.

Litwin, W., Mark, L. & Roussopoulos, N. Interoperability of Multiple Autonomous Databases, ACM Computing Surveys, September, 1990, 22(3), 267-293.

Mannino, M.V. Matching Techniques in Global Schema Design, Proceedings of the International Conference on Data Engineering, 1984, 418-425.

Miller, R.J., Ioannidis, Y.E. & Ramakrishnan, R. Schema Equivalence in Heterogeneous Systems: Bridging Theory and Practice (Extended Abstract), Advances in Database Technology - EDBT Ô94, Cambridge, UK, 1994, 73-80.

Motro, A. Superviews: Virtual Integration of Multiple Databases, IEEE Transactions on Software Engineering, July, 1987, SE-13(7), 785-798.

Navathe, S.B. & Gadgil, S.G. A Methodolgy for View Integration in Logical Data Base Design, Proceedings of the 8th International Conference on Very Large Databases, Saratoga, CA, VLDB Endowment, 1982.

Navathe, S., Elmasri, R. & Larson, J. Integrating User View in Database Design, Computer, January, 1986, 50-62.

Polat, F., S. Shekhar, and H.A. Guvenir, ÒDistributed Conflict Resolution Among Cooperating Expert Systems,Ó Expert Systems, 10, 4 (1993), 227-236.

Sheth, A.P. & Larson, J.A. Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases, ACM Computing Surveys, September, 1990, 22(3), 183-236.

Spaccapietra, S. & Parent, C. Conflicts and Correspondence Assertions in Interoperable Databases, SIGMOD Record, December, 1991, 20(4), 49-54.

Spaccapietra, S. & Parent, C. View Integration: A Step Forward in Solving Structural Conflicts, IEEE Transactions on Knowledge and Data Engineering, April, 1994, 6(2), 258-274.

Urban, S.D. & Wu, J. Resolving Semantic Heterogeneity Through the Explicit Representation of Data Model Semantics, SIGMOD Record, December, 1991, 20(4), 55-58.