December 2002

A CRITICAL EVALUATION OF SCHEMA INTEGRATION METHODOLOGIES AND TRENDS

Marianne Murphy
Northeastern University

Ramesh Venkataraman
Indiana University

Uday Kulkarni
Arizona State University

A CRITICAL EVALUATION OF SCHEMA INTEGRATION METHODOLOGIES AND TRENDS

Marianne Murphy
Northeastern University
ma.murphy@neu.edu

Ramesh Venkataraman
Indiana University
venkat@indiana.edu

Uday Kulkarni
Arizona State University
uday.kulkarni@asu.edu

Abstract

As companies grow, become more complex and attempt to maintain or increase their competitive advantage, either through combining diverse operations or through mergers and acquisitions, so does the company's need to access information across diverse, often heterogeneous databases. Database schema integration research is motivated by the need for companies to provide seamless access across diverse databases. The purpose of our research is to critically evaluate the past 15 years of integration research to determine the progress and direction of future research in this area.

Introduction

Information is becoming an increasingly vital resource in the knowledge economy. The way in which we access and retrieve this information is becoming a key issue to many organizations. Companies are increasingly dependent upon the reliability of their databases for accurate information as well as their ability to access this information in a timely and efficient manner.

Much research has been done in database design and integration to develop efficient databases that meet the expectations and needs of user groups. Database technology has progressed to the point where many organizations are using databases for day-to-day operations, strategies and managerial applications. Additionally, distributed databases are becoming sufficiently understood, and it is becoming increasingly expected that many organizations will adopt distributed architectures by integrating their current diverse databases (Spaccapietra and Parent, 1994).

The increasing issues faced by organizations to use and access data more efficiently has motivated the area of schema integration research. Batini, Lenzerini and Navathe (BNL86) [1986] examined the integration methodologies proposed in the literature, up to 1986, which comprised ten years of research in this area. Since that time the research has taken many directions and it is unclear if progress has been made. Although the context has clearly changed in many proposals, the progress of such proposals is less clear.

The purpose of this research is to develop a framework for appraising database schema integration methodologies in the last fifteen years and determine the progress and direction of such proposals. There have been nearly 200 papers published in the past fifteen years. The approaches generally fall into one of four categories: schema integration, view integration, multidatabase distributed integration and wrappers. Additionally, the context of these approaches is heterogeneous database systems, data warehouses, database design and web based integration.

Framework for Analysis

In completing this research, we will categorize all methodologies proposed in the last fifteen years and critically evaluate the contributions. We have collected nearly 200 articles in the areas of schema integration, view integration, multidatabase,

distributed integration and wrappers. The context of these papers is generally heterogeneous database systems, data warehouses, database design and web based integration. Approximately 80 articles have currently been reviewed and contributions noted. In some of the articles it is difficult to assess the real or incremental contributions. However, when all articles are completed, the evaluation will include not only an assessment of the contribution purported but also a comparison to all work that is done in each category. A comparison is the best way to assess the true contribution.

Batini, Lenzerini and Navathe [1986]

Batini et al. [1986] performed a structured evaluation of ten years of integration research. The framework described by BLN86 analyzed and discussed the methodologies proposed up to 1986 on four major issues. These four issues are:

- data model
- inputs and outputs
- timing of the actual integration step
- the phases of integration
    1. preintegration
    2. schema comparison
    3. schema conforming
    4. merging and restructuring

They also evaluated the future direction of schema integration. The set of methodologies examined by Batini et al. [1986] only address the knowledge needed to integrate in a very limited way. Generally, this is only a description of the inputs and not necessarily how these inputs are represented nor specific procedures or processes to obtain the knowledge. Knowledge can be expressed in terms of assertions, productions and inference rules. The assertions are used as inputs in many integration methodologies. Building assertions from the knowledge contained in the databases and integration of rules is one area of limited coverage in the methodologies examined by Batini et al [1986].

Process integration refers to the transforming of the processes related to each component schema into a set of processes for the integrated schema. Again, like integration of the knowledge bases, process integration has received little attention in the methodologies examined by BLN86. Finally, Batini et al. [1986] suggested the continued investigation of expert systems to database design and in particular to schema integration. Some research to date has investigated the use of expert systems in database design (Bouzeghoub, 1991). Until 1986, very few methodologies had been implemented and those few only to a very limited extent. The use of expert systems, according to Batini et al. [1986] would progress the implementation of such methodologies.

Our Framework

The main interest of our research is what progress has been made and what should be the direction of future research. Batini's framework is hardly adequate to evaluate the current integration research. At first an extension of this framework was considered. However, it became clear that technological and other advances made Batini's framework obsolete. Much work has been done in the areas of expert systems and other decision support systems. In addition the typical business models have changed. Organizations can store large amounts of data relatively cheaply, and are looking for ways to manage this information through large depositories such as Data Warehouses. In addition, more and more business is conducted on the web and researchers have investigated ways to access and integrate this information.

- **Approach and Context**
  First we look at the basic approach of the research contribution. Most of the research uses schema integration either federated or global or view integration such as those used in design methodologies. These approaches may have different context and application. The generalized contexts can be schema integration only, heterogeneous database integration, data warehouse, web based or database design. The application context might be manufacturing, business, scientific, medical, etc.

- **Emphasis**
  We also will look at and evaluate the emphasis of the research. Most research usually only deals with a small aspect of the integration process. Some methodologies look at the schema transformations and/or schema evolution. Some look at conflict identification and/or conflict resolution. Finally others may address mapping generation or query transformation.

- **Data Model**
  Batini et al. [1986] determined that the methodologies that existed in 1986 were divided into two schools with regard to the data model used. One school used the relational or functional model and the other school used more of a semantic model. The conclusion they reached is that using a semantic model provided a richer set and allowed more flexibility in terms of naming and compatible and incompatible design perspectives. Data has become more complex including graphic, movie and music. The research direction for most methodologies is clearly towards a more complex model such as Extended ER and Object Oriented. In addition, we evaluate the common data model and whether it is translation or mapping. In translation, each schema is translated into a common data model whereas in mapping, each schema remains intact and is not "restructured" but rather is mapped to a common data model. The outcome is transparent to the user.

- **Assertions**
  In general assertions are statements of how a piece of one database matches a piece of another database. These assertions are then used as inputs to the integration process. We evaluate whether these assertions are declarative or procedural and whether they deal with equivalencies only or other aspects such as inclusion, intersection or subsets. Finally, we evaluate whether these assertions are gathered through a manual process strictly or whether some sort of automation process is developed.

- **Output**
  We evaluate whether the output is transparent to the users or whether some restructuring was necessary. In addition, we look at post mapping. We evaluate whether the methodology proposed query mapping, application mapping or both.

- **Strategy**
  The overall strategy of the approach is evaluated. One strategy is to integrate all component schemas at once; other approaches are two at a time. We also evaluate the impact of this approach on the overall contribution of the methodology.

- **Research Focus and Contribution**
  We critically evaluate all integration methodologies proposed based on the contributions the authors have claimed as well as their contribution to the literature in general. In some cases it is difficult to assess what the contribution may be. In addition, we evaluate whether the focus is purely conceptual or theoretical as well as whether a system is developed or if there are any practical applications to the proposal.

Contributions

The main contribution of our paper is a concise and organized view of what progress has been made in the schema integration research area in the last 15 years and a determination of what direction the research should be. Clearly, when the focus of business changes, the research changes as well. Therefore methodologies have been proposed that address web based systems and data warehouse issues. Organizations continually strive to cut costs, provide marketable services and products and stay competitive. For most organizations, these goals are challenging in a technologically advanced environment. The growing trend towards mergers and acquisitions as well as organizations ability to access information across diverse business units makes the integration research vital and important. However, has the integration research kept pace with technological advances and the changing business environments? Have real contributions been made in this area? Our research will evaluate these questions.

Selected References (For a complete list of references, please contact the authors.)

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

Bertino, E. (1991). "Integration of Heterogeneous Data Repositories by using Object-Oriented Views." Proceedings of IMS'91-The First International Workshop on Interoperability in Multidatabase Systems: 22-39.

Biskup, J., and Convent, B. (1986). A Formal View Integration Method. Proceedings of the International Conference on Management of Data, Washington, D. C., ACM.

Bouguettaya, A. and S. Milliner (1995). "Co-database Approach to Database Interoperability." IEICE transactions on information and systems 78(11): 1388.

Bouguettaya, A., B. Benatallah, et al. (1999). "WebFindIt: An Architecture and System for Querying Web Databases." Ieee internet computing 3(4): 30.

Bouzeghoub, M. and Comyn-Wattiau, I., "View Integration by Semantic Unification and Transformation of Data Structures", Entity-Relationship Approach: The Core of Conceptual Modeling, North Holland, 1991, pp. 381-398.

Bright, M. W., A. R. Hurson, et al. (1994). "Automated Resolution of Semantic Heterogeneity in Multidatabases." Acm transactions on database systems 19(2): 212.

Casanova, M., and Vidal, V. (1983). Towards a Sound View Integration Methodology. Proceedings of the 2nd ACM Conference on Principles of Database Systems, Atlanta, ACM.

Chen, P. P. (1976). "The Entity-Relationship Model: Toward a United View of Data." ACM Transactions on Database Systems 1(1): 9-36.

Chen, J., O. Bukhres, et al. (1991). "The Implementation of Cooperative Mechanisms among System Components in a Heterogeneous Multidatabase Environment." Computing systems: the journal of the USENIX A 6(3): 207.

Chiang, Roger H., Lim, Ed-Peng, Storey 2001, Veda, A Framework for Acquiring Domain Semantics and Knowledge for Database Integration, *The DATA BASE for Advances in Information Systems* 31(2), 46-64.

Comyn-Wattiau, I. and M. Bouzeghoub (1993). "Constraint Confrontation: An Important Step in View Integration." Proceedings of the 5th International Symposium on Advanced Information Systems Engineering, CAiSE'93: 507-523.

Czejdo, B. and M. Taylor (1991). "Integration of Database Systems using an Object-Oriented Approach." Proceedings of IMS'91-The First International Workshop on Interoperability in Multidatabase Systems: 30-37.

de Souza, J. M. (1986). SIS - A Schema Integration System. Proceedings of the Fifth British National Conference on Databases, Cambridge University Press.

El-Masri, R., J. Larson, et al. (1986). "Schema Integration Algorithms for Federated Databases and Logical Database Design." Technical Report, Honeywell Systems Development Division.

Elmasri, R., J. Larson, et al. (1987). "Integration Algorithms for Federated Databases and Logical Database Design." Technical Report; Honeywell Corporate Research Center.

Hayne, S. and S. Ram (1990). "Multi-User View Integration (MUVIS): An Expert System for View Integration." Proceedings of the Sixth International Conference on Data Engineering: 402-409.

Larson, J. A., S. B. Navathe, et al. (1989). "A Theory of Attribute Equivalence in Databases with Application to Schema Integration." IEEE Transaction on Software Engineering 15(4): 449-463.

Lim, E., Srivastava, J., Prabhakar, S., and Richardson, J. (1996). "Entity Identification in Database Integration." Information Sciences 89: 1-38.

Lim, E. P., J. Srivastava, et al. (1996). "An Evidential Reasoning Approach to Attribute Value Conflict Resolution in Database Integration." IEEE Transactions on Knowledge and Data Engineering 8(5).

Mannino, M. V. and W. E_lsberg (1984). "Matching Techniques in Global Schema Design." Proceedings of the First International Conference on Data Engineering: 418-425.

Martin, P., and Powley, W. (1993). Database Integration using Multidatabase Views. Proceedings of CASCON.

Miller, R. J., Y. E. Ioannidis, et al. (1994). "Schema Equivalences in Heterogeneous Systems: Bridging Theory and Practice." Information Systems **19**(1): 3-31.

Miller, L. L., V. Honavar, et al. (1997). "Warehousing Structured and Unstructured Data for Data Mining." Proceedings of the ASIS Annual Meeting 34: 215-224.

Naiman, C. and A. Ouksel (1995). "A Classification of Semantic Conflicts in Heterogeneous Database Systems." Journal of organizational computing 5(2): 167.

Put, F., and Liris, K. (1991). Schema Translation during Design and Integration of Databases. Entity-Relationship Approach: The Core of Conceptual Modeling. H. Kangassalo. Amsterdam, Elsevier Science: 399-421.

Ram, S. and V. Storey (1993). "Composite and Groupings in Semantic Modeling." Proceedings of the Hawaii International Conference on Systems and Sciences III: 80-90.

Ram, S., and Ramesh, V. (1995). "A Blackboard-Based Cooperative System for Schema Integration." IEEE Expert: 56-63.

Ram, S. and V. Ramesh (1998). "Collaborative Database Design: A Process Model and System." ACM Transactions on Information Systems.

Ramesh, V., and Ram, S. (1995). A Methodology for Interschema Relationship Identification in Heterogeneous Databases. Proceedings of the 28th Annual Hawaii International Conference on System Sciences, Hawaii.

Ramesh, V. and S. Ram (1997). "Integrity Constraint Integration in Heterogeneous Databases: An Enhanced Methodology for Schema Integration." Information Systems 22(8): 423-446.

Reddy, M., Prasad, B., Reddy, P., and Gupta, A. (1994). "A Methodology for Integration of Heterogeneous Databases." IEEE Transactions on Knowledge and Data Engineering 6(6): 920-933.

Santucci, G. (1998). "Semantic schema refinements for multilevel schema integration." Data & knowledge engineering 25(3): 301.

Sheth, A. P. and H. Marcus (1992). "Schema Analysis and Integration: Methodology, Techniques and Prototype Toolkit." Technical Memorandum, TM-STS-019981/1.

Spaccapietra, S., Parent, C., and Dupont, Y. (1992). "Model Independent Assertions for Integration of Heterogeneous Schemas." Very Large Database Journal 1(1): 81-126.

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

Thieme, C. and A. Siebes (1993). "Schema Integration in Object-Oriented Databases: Proceedings of the 5th International Symposium on Advanced Information Systems Engineering." CAiSE '93: 54-70.

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

Whang, W. K., S. B. Navathe, et al. (1991). Logic-Based Approach for Realizing a Federated Information System. Proceedings of IMS '91 - The First International Workshop on Interoperability in Multidatabase Systems.