INFORMATION SYSTEMS RESEARCH AT GEORGE MASON UNIVERSITY

Sushil Jajodia, Daniel Barbará, Alex Brodsky, Larry Kerschberg, Ami Motro, Edgar Sibley, and X. Sean Wang

Department of Information and Software Systems Engineering
George Mason University, Fairfax, VA 22030-4444, USA
`www.isse.gmu.edu`

Overview

George Mason University began as an independent state university in 1972. Its development has been marked by rapid growth and innovative planning, resulting in an enrollment of more than 24,000 students in 1997. It is located in Fairfax, Virginia—about fifteen miles southwest of Washington, DC—near many governmental agencies and industrial firms specializing in information-intensive products and services.

Information and Software Systems Engineering (ISSE) is one of six departments in GMU's School of Information Technology and Engineering (SITE). Established in 1985, SITE has approximately 90 faculty and ISSE has 13 full time faculty. ISSE is a rapidly growing department with wide-ranging teaching and research interests. The department offers no undergraduate degree programs and Master of Science degrees in Information Systems (MSIS) and Software Engineering (SWSE). MSIS has about 800 students and the SWSE has approximately 400 students enrolled. The MSIS program graduates about 120 students and the SWSE program awards 40 degrees per year. ISSE faculty participate in the SITE doctoral program in Information Technology. ISSE Faculty chair the committees of more than one third of the doctoral students in the SITE program, which currently graduates about 30 PhDs per year.

Two research centers are associated with the department: The Center for Secure Information Systems (Sushil Jajodia, Director) and the Center for Information Systems Integration and Evolution (Larry Kerschberg, Director).

Departmental research in information systems is supported by grants and contracts from several sources. The following awards have been received so far for the academic year 1997–1998 and beyond:

[1] Knowledge Rovers: A Family of Intelligent Software Agents for Logistics for the Warrior. Defense Advanced Research Projects Agency (co-PIs: Kerschberg, Gomaa, Jajodia, Motro)

[2] Electronic Commerce for Logistics, Teaming Agreement with American Management Systems for DARPA BAA 95-25 Logistics Research and Development (co-PIs: Kerschberg, Gomaa, Jajodia, Motro)

[3] Linear Constraint Databases, NSF Research Initiation Award (PI: Brodsky)

[4] Linear Constraint Programming, ONR (co-PI: Brodsky with late Kannelakis (PI), Van Hentenryck, and Lassez)

[5] Towards Expressive and Efficient Queries on Sequenced Data, NSF Research Initiation Award (PI: Wang)

[6] Supporting Multiple time granularities in Query Evaluation and Data Mining, NSF (co-PIs: Jajodia, Wang)

[7] Fine Granularity Access Controls in World Wide Web, NSA (PI: Jajodia)

[8] Information Flow Control in Object-Oriented Systems NSA (PI: Jajodia)

[9] Exploring Steganography: Seeing the Unseen, NSA (PI: Jajodia)

[10] Trusted Recovery from Information Attacks, Rome Laboratory (co-PIs: Jajodia, Ammann)

[11] A Unified Framework for Supporting Multiple Access Control Policies, DARPA (PI: Jajodia)

The remainder of this article provides a brief overview of our research followed by a selected list of publications. More detailed information is available at `www.isse.gmu.edu`.

Knowledge Rovers: A Family of Configurable Software Agents for Logistics

Participating Faculty: Kerschberg, Gomaa, Jajodia, Motro
Visiting Scholars: Len Seligman, Jong Pil Yoon
URL: `nomad.krl.gmu.edu`

Knowledge rovers represent a family of cooperating intelligent agents that may be configured to support enterprise tasks, scenarios, and decision-makers. These rovers play specific roles within an enterprise information architecture, supporting users, maintaining active views, mediating between users and heterogeneous data sources, refining data into knowledge, and roaming the Global Information Infrastructure seeking, locating, negotiating for and retrieving data and knowledge specific to their mission. The concept of Knowledge Rovers serves as a metaphor for the family of cooperating intelligent agents that support an enterprise's information architecture. The goal is to configure rovers automatically with appropriate knowledge bases (ontologies), task-specific information, negotiation and communication protocols for specific scenarios.

The family of rovers supports the data and information infrastructure by providing specialized information mediation services such as: (1) approximate consistency services which monitor deviations of cached objects and the underlying databases and take actions based on user-specified consistency conditions, (2) object replication services that ensure object availability, reliability, performance, and survivability, and (3) information repository services consisting of ontology services, object location services, and domain servers that integrate heterogeneous types of data obtained from diverse heterogeneous sources including the Internet.

[1] S. Jajodia and L. Kerschberg, eds., *Advanced Transaction Models and Architectures*, Kluwer Academic Publishers, 1997, 381 pages.

[2] L. Kerschberg, "Knowledge Rovers: Cooperative intelligent agent support for enterprise information architectures," *Springer-Verlag LNCS, Vol. 1202*, 1997, pp. 79–100.

[3] L. Kerschberg, "The role of intelligent software agents in advanced information systems," *Springer-Verlag LNCS, Vol. 1271*, 1997, pp. 1–22.

[4] O. Wolfson, S. Jajodia, and Y. Huang, "An adaptive data replication algorithm," *ACM TODS*, 22(2)1997, pp. 255–314.

Temporal Databases with Multiple Granularities

Participating Faculty: Jajodia, Wang
Visiting Scholar: Claudio Bettini
URL: `www.isse.gmu.edu/~csis/tdb`

There is a rich variety of time granularities, and users as well as applications often require the flexibility of viewing the same temporal data in terms of different time granularities. The increased awareness of this requirement by the database community is evident from the growing research literature on this subject and from the inclusion of constructs for handling multiple granularities in SQL-92, the relational query language standard, and TSQL2, a temporal extension of SQL-92. In spite of this, the fact remains that whenever there is a difference between the way the information is stored in the database and the way it is required by the users, it is up to the users to understand these differences and specify appropriate operations in their queries to reconcile them.

The objective of this project is to provide a flexible framework for supporting multiple time granularities. The targeted use of the framework is for automatic evaluation of user queries and for discovering temporal patterns (i.e., data mining) in an environment where either the user queries or temporal patterns involve granularities that do not match the granularity of the stored data. The basic idea is to add the necessary functionalities so that the database system is able to understand and reason about information involving multiple time granularities. Algorithms for efficiently evaluating user queries and discovering temporal patterns are also being investigated, and an experimental prototype is being built.

We are also investigating networks having temporal constraints on the distances among event occurrences. When multiple granularities are used in the distance specification, new techniques and algorithms for consistency checking and for deriving solutions are required. We are considering applying the results that we have obtained in this area to database integrity constraint checking and to trigger condition evaluation in active databases.

[1] X. S. Wang, C. Bettini, A. Brodsky, and S. Jajodia, "Logical design for temporal databases with multiple granularities," *ACM TODS*, Vol. 22, No. 2, June 1997, pages 115–170.

[2] C. Bettini, X. S. Wang, S. Jajodia, and J. Lin, "Discovering temporal relationships with multiple granularities in time sequences," *IEEE TKDE*, To appear. A preliminary version appeared as "Testing complex temporal relationships involving multiple granularities and its application to data mining," *Proc. ACM PODS*, 1996, pp. 68–78.

[3] C. Bettini, X. S. Wang, and S. Jajodia, "Temporal semantic assumptions and their use in database query evaluation," *IEEE TKDE*, To appear. A preliminary version appeared in *Proc. ACM SIGMOD*, 1995, pp. 257–268.

[4] X. S. Wang, S. Jajodia, V. S. Subrahmanian, "Temporal Modules: An Approach Toward Federated Temporal Databases," *Information Sciences*, (82)1995, pp. 103–128. A preliminary version appeared in *Proc. ACM SIGMOD*, 1993, pp. 227–236.

Semantic-based Transaction Processing

Participating Faculty: Ammann, Jajodia

The traditional correctness criteria of serializability forces database designers into tradeoffs among design objectives. For example, in multidatabases, the designer balances the objectives of local design and execution autonomy, decentralized management of global transactions, maintenance of global integrity constraints, and execution history correctness. The last objective is typically assessed with respect to some variant of conflict serializability. Switching to a semantics-based perspective of correctness can greatly reduce the conflict between the remaining objectives. In the case of multidatabases, the conflict can be entirely avoided for certain applications.

We are utilizing the semantics-based perspective in three distinct application areas: multidatabases, secure multilevel databases, and long duration transactions. Additionally, the method holds promise for such areas as database recovery and survivability. The cost of the semantics-based approach is additional off-line transaction analysis early in the lifecycle of a system; however, this up-front cost is amortized over the numerous transaction invocations during the system's lifetime.

[1] P. Ammann, S. Jajodia, and I. Ray, "Applying formal methods to semantic-based decomposition of transactions," *ACM TODS*, 22(2) 1997, pp. 215–254.
[2] P. Ammann, S. Jajodia, and I. Ray, "Ensuring atomicity of multilevel transactions," *Proc. IEEE Symp. Security and Privacy*, 1996, pp. 74–84.

OLAP

Participating Faculty: Barbará, Wang

OLAP Data Model We formalize a multidimensional data (MDD) model for OLAP, and develops an algebraic query language called *grouping algebra*. The basic component of the MDD model is a multidimensional cube, a popular abstraction for multidimensional data. A cube is simply a multidimensional structure that contains at each cell an aggregate value, i.e., the result of applying an aggregate function to an underlying relation. In order to express user queries, relational algebra expressions are then extended to those on basic groupings for obtaining complex groupings, including order-oriented groupings (for expressing, e.g., cumulative sum). We then consider the environment where the multidimensional cubes are materialized views derived from base data situated at remote sites. A multidimensional cube algebra is introduced in order to facilitate the data derivation. We also studied optimization issues.

Quasi-Cubes: Exploiting Approximations Even though vendors have been selling products to support data cubes for a while it is accepted that OLAP products do not scale to large datasets or high dimensions.

There are two obstacles that make scaling difficult. First, there is the issue of database explosion: even though the multidimensional cube is usually sparse, materializing every cell is very often prohibitive. Secondly, the demands on query performance in OLAP are strict (analysts need the answers quickly, so they can figure out the next question to ask to the system). So, even if all cells are materialized, there is a need to support a large variety of queries efficiently.

In this project we investigate techniques for efficiently scaling cubes. These techniques are based in a variety of statistical tools and aim to provide approximations to query answers, trading off errors for better space management or query response. The main idea is to describe regions of the cube by statistical models that can be represented succinctly. In doing so, one relies on the models to reconstruct some of the cells in the cube, incurring in errors in the process. To keep the errors under control, some cell values (namely, the outliers for the models) must be retained. We call these approximated cubes *Quasi-Cubes*.. Our preliminary results show that this technique is feasible and provides with an excellent way of reducing the storage needs for the cube. Even if the usage of approximations is not possible for a given application, the modeling techniques enable the implementation of systems which provide answers that are progressively polished on-line (until the correct answer is given), eliminating the traditional latency that users experience when they pose queries. This is possible since the models provide the designer with a good classification method for the cells of the cube. Each cell can be put in an error bin, according to the error one would incur if the cell value were to be estimated by the model. When answering a query, the cells in higher error bins are retrieved first while the other cell values are estimated by the models. If one wants to refine the answer, some of the estimated cell values can be replaced by the real ones by retrieving cells in the next error bin, and so on.

[1] D. Barbará and M. Sullivan, "Quasi-Cubes: A space-efficient way to support approximate multidimensional databases," Technical Report, ISSE Dept., September 1997.
[2] C. Li and X. S. Wang, "A Data Model for Supporting On-Line Analytical Processing", *Proc. CIKM Conf.*, 1996.
[3] C. Li and X. S. Wang, "Optimizing Statistical Queries by Exploiting Orthogonality and Interval Properties of Grouping Relations," *Proc. Int'l. Conf.*

Scientific and Statistical Database Management, June 1996.

Constraint Databases

Participating faculty: Brodsky

Constraints provide a flexible and uniform way to represent and manipulate diverse data capturing spatio-temporal behavior, complex modeling requirements, partial and incomplete information, etc. They have been used in a wide variety of application domains. Constraint databases (CDBs) have recently emerged as a tool for deep integration of heterogeneous data captured by constraints in databases.

This project involves research in the incorporation of successful constraint technology (including such aspects as arithmetic constraints over reals, interval constraint propagation, and combinatorial optimization over finite domains) with database technology, in the framework of CDB, and aimed especially at two broad application areas:

1. Spatial and temporal, in which there is need to represent such data as complex objects in a low-dimensional space (typically, up to 4-5), movement of objects in 3D-space, transformations among various (possibly polar) coordinate systems, and patterns of behavior in space over time. In these applications low-dimensionality and domain-specific properties of the data can be exploited in developing efficient data structures and algorithms.
2. Applications requiring mathematical optimization, such as linear programming, in presence of large amounts of data. Mathematical optimization techniques are used to facilitate query evaluation, and, even more importantly, database set-at-a-time processing, indexing, and the ability to keep very large intermediate results can be exploited to facilitate mathematical (combinatorial) optimization.

Examples of spatio-temporal applications include CAD/CAM systems, GIS and environmental systems, command and control systems, such as maneuver planning and data fusion and sensor management. Examples of applications requiring combinatorial optimization and search include manufacturing and warehouse support systems, financial systems such as electronic trade, and many traditional mathematical programming problems involving large amounts of data. We strongly believe that the CDB technology will have a significant impact on these application areas. Moreover, it has the potential to become an integral part of a new generation of DBMS.

The major aspects of this project are (1) constraint modeling, canonical forms and algebras, (2) data models and query languages, (3) indexing and approximation-based filtering, and (4) constraint algebra algorithms and global optimization, and, most importantly, (5) building a system and demonstrating the feasibility of the CDB technology by means of case studies. The more theoretical work on aspects (1)-(4) led to the large-scale development of CCUBE, the first object-oriented database system, developed at GMU. The challenge involves achieving both declarative and efficient querying of large data sets involving constraints. A successful integration of constraint programming techniques with object-oriented or relational database systems is possible, given the current programming and database state of the art, but this is also challenging, given the demands for high level specification and efficiency.

[1] A. Brodsky and Y. Kornatzky, "The $\mathcal{L}yri\mathcal{C}$ language: Querying constraint objects,". *Proc. ACM SIGMOD.* 1995.
[2] A. Brodsky, V.E. Segal, J. Chen and P.A. Exarkhopoulo, "The CCUBE constraint object-oriented database system," *Constraints, An Int'l. Journal*, To appear.
[3] A. Brodsky, C. Lassez, J.-L. Lassez, M.J. Maher, "Separability of polyhedra for optimal filtering of spatial and constraint data," *Proc.ACM PODS*, 1995.
[4] A. Brodsky, J. Jaffar, M. Maher, "Toward practical constraint databases," *Constraints, An Int'l. Journal*, To appear. A preliminary version also appeared in *Proc. VLDB Conf.*, 1993.
[5] A. Brodsky, X.S. Wang: On Approximation-based Query Evaluation, "Expensive Predicates and Constraint Objects," *Proc. Workshop on Constraints, Databases, and Logic Programming*, December 1995.

Integrating Heterogeneous and Inconsistent Information

Participating Faculty: Motro

The integration of information from multiple databases has been an enduring subject of research for almost 20 years, and many different solutions have been attempted or proposed. The major goals of this project, called Multiplex, are to (1) define a *formal model* of multidatabases; (2) provide simple, rich and flexible support for *heterogeneity*; and (3) in situations where single, authoritative answers are not feasible, either because there is "too little information" (e.g., an information source went off-line) or there is "too much information" (e.g., there are multiple, mutually inconsistent answers), provide *approximative answers*. The present version of Multiplex is available on the Internet as an *integration server*: after defining a new database scheme, users need only specify links to sources that deliver *views* of that scheme. Queries (in SQL) submitted to the server are answered transparently from the available sources. The present approach to the resolution of inconsistencies is based on majority votes; our current work is to strengthen this capability, to allow different schemes for resolving inconsistencies, and flexible user control over these schemes.

[1] A. Motro, "Multiplex: A formal model for multidatabases and its implementation," Technical Report, ISSE Dept., 1995.

Information Quality and Uncertainty

Participating Faculty: Motro

With more and more electronic information sources becoming widely available, the issue of the quality of these, often-competing, sources has become germane. Since 1993 we have been exploring this relatively neglected subject. We have been proposing a new standard for rating information sources with respect to their quality. This standard, based on the concepts of *soundness and completeness*, attempts to gauge the distance of the information in a database from the truth, and is implemented by combining manual verification with statistical methods. Once a source has been rated for quality, the quality of arbitrary queries is estimated with an appropriately-extended relational algebra. At the present, we are experimenting with this methodology. We plan to incorporate information quality considerations into Multiplex, as a strategy for resolving information inconsistencies. We also plan to address the issue of adjusting quality specifications to reflect changes in the information.

As models of the real world, databases are often permeated with various forms of uncertainty, including imprecision, incompleteness, vagueness, inconsistency and ambiguity. Ever since our work on the Vague database interface (1988), we have sustained continued interest in this area, and have been advocating the adaptation of various uncertainty theories that have been developed within the AI community to the needs of practical information systems. At the present, we are developing our *soundness and completeness* model of uncertainty, which is based on the proximity of a *stored instance* of a database and the *real-world instance* which it tries to approximate.

[1] A. Motro and I. Rakov, "Not all answers are equally good: Estimating the quality of database answers," In *Flexible Query-Answering Systems* (T. Andreasen et al., Editors). Kluwer Academic Publishers, 1997.
[2] A. Motro and P. Smets, Editors. *Uncertainty Management in Information Systems: from Needs to Solutions*, Kluwer Academic Publishers, 1996, 480 pages.

Cooperative Databases

Participating Faculty: Motro

This area of interest focuses on database retrieval methods that offer alternatives to formal querying (such as SQL). Research in this area has been going on for over 10 years, resulting in several new retrieval paradigms and user interfaces. Highlights include Baroque, an early browser for relational databases (1986); Vague, a user interface to relational databases that permits weakly specified queries (1988); Flex, a tolerant and cooperative query system that can be used satisfactorily by users with different levels of expertise (1990); ViewFinder, a graphical object-oriented database browser (1993, with D'Atri and Tarantino from the University of L'Aquila); and, most recently, Panorama, a database system that annotates its answers to queries with their properties (1996).

[1] A. Motro, "Intensional answers to database queries," *IEEE TKDE*, 6(3)1994, pp. 444–454.
[2] A. Motro, "Cooperative database systems," *Int'l. Jour. of Intelligent Systems*, 11(10)1996, pp. 717–732.
[3] A. Motro, "Panorama: A database system that annotates its answers to queries with their properties," *Jour. of Intelligent Information Systems*, 7(1)1996, pp. 51–73.

Information Systems Security

Participating Faculty: Jajodia, Ammann, Brodsky, Motro, Sandhu, Sibley, Wang
URL: `www.isse.gmu.edu/~csis`

The Center for Secure Information Systems provides a focal point for research in Information Security. CSIS has been created to provide a dedicated environment to encourage the development of expertise in both the theoretical and applied aspects of information systems security. This is an area of increasing importance in governmental, military and commercial arenas, and CSIS's emphasis on information security makes it unique among the institutions of higher learning in this country.

The scope of CSIS encompasses information secrecy, integrity, and availability problems in military, civil, and commercial sectors. Among topics of current interest are:

- Flexible access control models and mechanisms
- Securing the World Wide Web
- Survivability and information warfare
- Intrusion detection and prevention
- Cryptography and Steganography
- Inference and aggregation
- Design techniques for secure database systems
- Secure transaction processing

- Implementing security in distributed databases
- Auditing
- Security in object-oriented systems
- Integrity mechanisms and models

[1] P. Ammann, S. Jajodia, C. D. McCollum, and B. T. Blaustein, "Surviving information warfare attacks on databases," *Proc. IEEE Symp. Security and Privacy*, 1997, pp. 31–42.
[2] S. Jajodia et al., "A unified framework for enforcing multiple access control policies," *Proc. ACM SIGMOD*, 1997, pp. 474–485.
[3] S. Jajodia, P. Samarati, V. S. Subrahmanian, "A logical language for expressing authorizations," *Proc. IEEE Symp. Security and Privacy*, 1997, pp. 31–42.
[4] S. Jajodia, "Database security and privacy," *ACM Computing Surveys*, 50th anniversary commemorative issue, 28(1)1996, pp. 129–131.

Ordered Data

Participating faculty: Wang
Visiting Scholar: Yunyao Qu

Operations on sequences are a basic component of database queries that extract information from sequenced data. We introduce a family of regular sequence operations (called rs-operations) to be used in such queries. The family is based on a simple pattern matching mechanism using regular expressions as its patterns, and includes most of the "natural" operations on sequences. Properties of the family are examined. In particular, operations in the family are characterized by a mechanical device called generic a-transducer, and the expressive power of the family is studied through an investigation of finite generators of the operations. We study the usage of the rs-operations in database queries through an extended relational data model. Two equivalent query languages, one algebraic and the other calculus, are given in the model. In these query languages, rs-operations are the only components used for dealing with sequences. We also study query languages on genetic sequences, especially RNA sequences, aimed at efficiently retrieving secondary structures.

We also study ordered-oriented queries in OLAP applications. These are order-oriented aggregation (e.g., calculating cumulative sums and moving averages) and order-oriented selection (e.g., obtaining top 5 and bottom 10 values along a sequence). In order to express such order-oriented queries in an intuitive yet powerful way, a ROLL clause of the form "`ROLL` attribute list `WITH BEGIN` num_1, `STEP` num_2, `LENGTH` num_3" is added into SQL. This clause uses the three numbers to group values along the sequence obtained by sorting the given attributes. The numbers num_1 and num_2 determine the starting positions of the groups, and num_3 decides the number of consecutive tuples to be included in each group. Depending on the parameters, some values may not be included in this process; order-oriented selection is achieved by dropping these values.

[1] S. Ginsburg and X. S. Wang, "Regular sequence operations and their use in database queries," To appear in *J. of Computer and System Sciences.*
[2] S. Ginsburg, M. Gribskov and X. S. Wang, "Structural queries on nucleic acid databases," *ACM Workshop on Information Retrieval and Genomics*, 1994.
[3] C. Li and X. S. Wang, "A data model for supporting on-line analytical processing," *Proc. CIKM Conf.*, 1996.

Large-Scale Scientific Database Systems

Participating faculty: Gomaa, Kerschberg, Wang, Menasce, Kafatos, Michaels
Visiting Scholar: Jong Pil Yoon

The Center for Information System Integration and Evolution (CISIE) participated in the NASA-sponsored Independent Architecture Study of the Earth Observing System Data and Information System. This study led to the development of the GMU Federated Client-Server Architecture which was based on the federated approach developed for DARPA's Intelligent Integration of Information Program (I*3).

Dr. Kerschberg and Dr. Michaels teach GMU's Scientific Database Course—one of very few such courses taught world-wide—and several research papers have appeared in the 1997 IEEE Statistical and Scientific Database Management Conference (SSDBM). The EOSDIS and SSDBM-related publications are listed below:

[1] L. Kerschberg, H. Gomaa, D. A. Menasce, J. P. Yoon, "Data and information architectures for large-scale distributed data intensive information systems," *Proc. IEEE Int'l. Conf. Scientific & Statistical Database Management*, 1996.
[2] M. Kafatos, X. S. Wang, et al., "The virtual domain application data center: Serving interdisciplinary earth scientists," *Proc. IEEE Int'l. Conf. Scientific & Statistical Database Management*, 1997.
[3] D. A. Menasce, H. Gomaa, L. Kerschberg, "A performance-oriented design methodology for large-scale distributed data intensive information systems," *Proc. 1st IEEE Int'l. Conf. on Engineering of Complex Computer Systems*, 1995.