01-IR

INFRASTRUCTURE 2001

NSF CISE/ELA RI and MII PI'S Workshop

Ballston, Virginia
July 23 - 24, 2001

A workshop for the principal investigators of NSF / CISE infrastructure awards, sponsored by the CISE Research Infrastructure Programs, supported by the Division of Experimental and Integrative Activities (EIA) in the Directorate Computer and Information Science and Engineering (CISE) of the National Science Foundation.

Infrastructure for Research and Training on High-Performance Heterogeneous Distributed Database Management

High Performance Database Research Center
School of Computer Science, Florida International University
Miami, FL 33199; rishen@fiu.edu; http://hpdrc.cs.fiu.edu

PI: Naphtali Rishe
Co-PIs: W. Sun, S.-C. Chen, M. Chekmasov

Introduction

Florida International University (FIU) is one of the largest majority-minority doctoral-granting universities in the United States. Nearly 70% of our students are minorities. The University has the largest contingent of Hispanic students of any doctoral-granting university in the country, and graduates the most Hispanic computer science and engineering students in the Nation. The High Performance Database Research Center (HPDRC) was founded in 1994, and is associated with the School of Computer Science at Florida International University. HPDRC conducts research on database management systems and various applications, leading to the development of new types of database systems and refinement of existing database systems.

The general goals of the project, now in its fourth year, are to provide an infrastructure that will enable FIU's HPDRC to perform heterogeneous database research and to better recruit and retain minority students through their M.S. and Ph.D. degrees. Students participate in in-depth research and training in heterogeneous database integration.

FIU is an urban university whose surrounding community base is substantially comprised of under-represented minorities: 88% of the students of Miami-Dade County Public Schools are members of under-represented minorities, including 54% Hispanic, 32% Black non-Hispanic, and 2% Other. One goal of this project is to establish a regional outreach program to attract talented local minority students to FIU. Without the support of this project, those students would otherwise not be able to take advantage of the career and educational opportunities, or would have to attend an out-of-state university (a non-favorable choice of many of the local minority students).

The infrastructure being assembled provides the students with a networked computing environment on which the research work will be conducted. The ultimate research goal is to develop a heterogeneous database management system, using semantic modeling to integrate and reconcile information from multiple, disparate data sources. Of particular interest are the methodologies to integrate geo-spatial and Web data sources. Geo-spatial data are vital to environmental research and studies (e.g. the Global Warming effect) but are often collected and stored in independently operated organizations. Web data generate new issues in data integration because, unlike traditional databases or data repositories, Web data are usually made available through form-filling interfaces, without divulging the data model behind the scenes. Specific research issues include: heterogeneous data model integration using semantic modeling, specification of Web data sources, geospatial data integration, reconciliation, and fusion (e.g. overlapping raster and vector data), rapid integration methodologies, query processing and optimization, and exploration of mobile agent technology.

HPDRC maintains a WWW page describing its projects and staff at http://hpdrc.cs.fiu.edu.

Fourth Year Accomplishments

Goals, Objectives, and Targeted Activities

The goals of our MII (Minority Institute Infrastructure) grant are to provide an infrastructure that will enable FIU's HPDRC to better recruit minority faculty members, better recruit and retain graduate students through the Ph.D., and to perform more in-depth research and training in database management. Since the grant's inception in the Fall of 1997, we have been striving to achieve these goals. The activities we have been engaged in are described in the following sections.

Recruiting Minority Faculty: Last year the School of Computer Science successfully hired a female Hispanic professor, Marie Roch. We are collabarating with her on research. Maria Martinez, an FIU Computer Science PhD graduate and a visiting professor of Electrical and Computer Engineering at FIU, is an HPDRC-affiliated faculty member who continues to serve as a role model to our students. Minority SCS faculty member Joslyn Smith is taking part in the NSF-sponsored HPDRC activities with students.

Retaining Graduate Students Through the Ph.D.: Rosany Rodriguez, a minority PhD student, is nearing the completion of her dissertation. Melinda Whiley, a graduate of Florida A&M University, joined FIU's HPDRC as a graduate research assistant in March 2000 as a result of our recruiting efforts. Debra Davis received her MS degree and will be continuing on in our PhD program. Alejandro Mendoza, who received his BS degree in the Summer of 2000, and Daniel Mendez, who received his BS degree in the Spring of 2001 have entered our graduate program. We continue to recruit promising students to take advantage of the funds provided by our MII grant and continue to fund current graduate students via MII as new students are identified. Jai Navlakha (FIU School of Computer Science) and Andrei Kirienko (FIU High Performance Database Research Center) continue to research methods to enhance the retention of graduate students through the PhD. We have stepped up our recruiting efforts by sending out a letter containing information on FIU's PhD program in Computer Science to current graduate students. We have also sent information on FIU's graduate programs in Computer Science to 800 graduates of FIU's CS undergraduate program who have qualifying GPAs. Presentations on FIU's CS graduate programs were made at meetings of the local ACM student chapter and followup discussions were held. We are continuing recruitment efforts directly with Florida A&M University, Florida Memorial College, Miami-Dade Community College, and the Miami-Dade County Public Schools. To this end, Martha Gutierrez, an NSF-supported PhD student, made a presentation on "Career Opportunities in Computer Science" at "The Exact and Natural Sciences Conference 2000," which was held at Miami Dade Community College's InterAmerican Campus. She stressed the need for computer scientists with four year and advanced degrees and the opportunities available to students who achieve these degrees. Nagarajan Prabakar, a Senior Investigator, gave presentations of our computer science research and recruited students at Miami Sunset Senior High and Miami-Dade County's Marine and Science Technology magnet high school. We have involved high school students in our research through Miami-Dade County Public School's Advanced Academic Internship Program. Their involvement provides us a pipeline of researchers from high school through the PhD.

Affinity Groups: Four Affinity Groups modeled after those at the University of Texas, El Paso are in place at HPDRC. The Affinity Groups are made up of faculty members, postdoctoral associates, and graduate and undergraduate students. The following groups continue to pursue research tracks at HPDRC:

Semantic Database Engine Group – devoted to designing and developing semantic database technology; *Applications Group* – devoted to investigating spatial data technology and applications and GIS; *Heterogeneous Database Group* – devoted to deepening research in distributed heterogeneous databases; and *Semantic-Relational Systems Group* – devoted to making the semantic database technology available to all database users.

Grants Awarded: FIU HPDRC has been awarded an additional three years of funding for our NASA Institutional Research Award. FIU HPDRC has been awarded additional funds to continue its work in developing better database applications for researchers at Everglades National Park. The United States Geological Survey (USGS) is planning to enter into a Cooperative Research and Development Agreement (CRADA) with the High Performance Database Research Center and the NASA Regional Applications Center at Florida International University. Among the plethora of benefits that Florida International University would receive as a result of this CRADA would be the delivery of USGS data, including 1-meter resolution Aerial Photography over the United States and Landsat-7 30-meter resolution satellite imagery over much of populated global land mass. This would comprise more than 18 terabytes of data over the period of the CRADA, and would enable the HPDRC and the NASA RAC to create potentially one of the largest databases on the Web.

Outreach Program to Schools: FIU's High Performance Database Center hosted seven students from the Miami-Dade Public Schools under the school system's Advanced Academic Internship Program (AAIP) during the 1998-1999 school year, seven additional students during the 1999-2000 school year, and four additional students during the 2000-2001 school year. These students worked alongside researchers at HPDRC and contributed to the research and development goals of the Center. After serving as an intern, Roy Duque de Estrada graduated from Hialeah-Miami Lakes High and has enrolled as a freshman at FIU. He is presently volunteering at HPDRC to stay involved in cutting-edge computer science research; we hope to hire him to a paying position in the near future. The involvement of these students provides us a pipeline of researchers from high school through the PhD.

We have continued to develop an outreach program that will ultimately consist of both visits to FIU and a traveling 'show' that includes a presentation geared to the appropriate audience at schools. The presentation is followed by a hands-on demonstration of interesting database projects to which the students can relate, such as advanced 'virtual reality' demonstrations and the like. One aspect of this show is viewing a South Florida Landsat image through which it is possible to 'fly' by updating the image in real-time from the semantic database in which the Landsat data is stored. Nagarajan Prabakar took this show to Miami Sunset Senior high and to Miami-Dade County's Maritime and Science Technology magnet high school during the 2000-2001 school year.

Course Development: The courses proposed in our response to the site visit have been approved as experimental by the School of Computer Science's curriculum committee. Dr. Shu-Ching Chen, will teach his a new course featuring some of the research conducted under this award, "Distributed Multimedia Database Systems and Information Systems," this Fall or next Spring. He is presently using the new course's curriculum in an "Advanced Topics in Information Processing" course. Dr. Chen and Dr. Prabakar are developing a course on Computer Networks to be taught at the 5000 level. This course should prove useful to students performing distributed database research.

Components and Materials Required and Indications of Success

Infrastructure Additions: The infrastructure acquired is being used every day by the student and faculty researchers. Additional acquisitions are planned before the end of the current award period.

REU Supplement: HPDRC requested an REU supplement to our MII grant. The students supported, in part, by the REU supplement are detailed in the Immediate Impact section.

Students Supported directly by MII: Sixteen graduate and undergraduate students have been directly supported, in part, by our MII grant during the past year; fifty-three have been supported since the grant's inception. The majority of these have been members of under-represented groups. Students supported by out MII grant are detailed in the Immediate Impact section.

Publications: During the past year, we published 15 items under the support of this grant.

Evaluation

Degree of Success

Toward the goal of better recruiting and retaining under-represented minority students in our graduate programs, we have successfully increased supported student enrollment as a result of the support from our MII grant. Several Affinity Groups and an Outreach Program are in place to enhance our teaching environment and graduate recruitment. We have worked progressively towards the design and analysis of the heterogeneous database system. The effort led to the publication of several technical papers. Appropriate facilities have been acquired and are being constantly enhanced to support students and the research. These activities and achievements evidence a great success in fulfilling the grant's goals and a continuing improvement over the first three years' results.

Outcome

Our research is based on Sem-ODB database technology under development at FIU's High Performance Database Research Center (HPDRC). Sem-ODB is a general-purpose DBMS based on the Semantic Binary Object Data Model (Sem-ODM). It supports a wide spectrum of applications ranging from transaction-oriented to decision-support applications, with specific unique features making it particularly ergonomic and efficient for scientific applications. Among such features are the variable-length numeric fields, elimination of restrictions on precision and magnitude of numbers, ergonomic handling of multi-valued attributes, application of the industry-standard SQL language and interfaces to non-conventional data structures. A Multi-user Parallel Semantic Database Engine is operational. A main goal of our work has been to achieve the quality that would make the Sem-ODB server viable as a commercial product. We have used Sem-ODB to model current and historical observations made at the Everglades National Park; the resulting model contained over two thousand categories and relations. Sem-ODB is also being used as the unifying database for TerraFly, a system for browsing geospatial data, which will soon manage over 13 TB of data. The inherent benefit of the semantic model is its ease of use and understandability; the resulting schemas directly model the real world. The level of abstraction inherent in such models allows Sem-ODB to be much more efficient than the typical relational database, which requires the user to create a schema using tables and indirect relations using foreign keys.

Scientific data sometimes needs to be distributed across the network. Each site might maintain an individual database with high autonomy. Nevertheless, data from several component databases may need to be combined in order to fulfill the requirements of an application. It is unavoidable that heterogeneity, which is partially caused by different data models or different query interfaces, will exist among component databases. Even if component databases implement the same data model, there is usually a need for us to deal with schematic or semantic conflicts. Such conflicts include structural conflicts, naming conflicts, abstraction conflicts, domain conflicts, etc. In order to meet all these requirements, we are building a heterogeneous database system (HDB) that extends our semantic binary database.

In our research, the database schemas of component databases are transformed into a canonical data model and exported to the global schema/knowledge reconciliation site for the creation of global/federated schemas and views for querying. The schema transformation phase resolves conflicts caused by the different data models of component databases. Also, a canonical data model provides a uniform query facility for each component site. This allows for less complex query processing techniques at the Integration/Knowledge Reconciliation site.

Recently, wrappers have been developed for performing schema transforming and query translation tasks at the component sites. At the Integration/Knowledge Reconciliation site, the heterogeneities due to the different user perspectives in a set of multiple database schemas are resolved providing users with global/federated views and schemas. The resolving of conflicts at component and integration sites requires acquiring and managing knowledge and meta-data. Our framework proposes the use of knowledge bases for this purpose. The types of meta-data and knowledge acquired are different from the component sites and the integration site. Thus, knowledge bases at these sites focus on different meta-data management and knowledge acquiring techniques.

At the component site, schemas and relevant meta-data are imported and transformed into the canonical data model by the Schema Loader & Transformer component. These meta-data and schema mapping information are stored in the Knowledge Base. The KDBTool/Semantic Enrichment component interacts with the DBA for advanced knowledge acquiring and conflict resolution processing. The transformed schemas including the relevant meta-data are exported to the Integration/Knowledge Reconciliation site. Query Translator component translates queries posed on the transformed schema of the canonical data model into semantically equivalent queries of the component database schema. A wrapper developed using the architecture described above is the Semantic Wrapper. The heterogeneities that occur due to multiple data models have been resolved using the wrapper at the component site. The Integration/Reconciliation site addresses the heterogeneities that occur due to a multitude of homogeneous database schemas exported from the component sites. Our research includes specific techniques for accessing data in relational databases, XML and other Web data, as well as unstructured data.

Impact

The exploding growth and use of the Internet and World Wide Web have enabled users to access huge volumes of data with unprecedented convenience and speed. However, the data sources often diverge in their data model (how the data are organized) and their retrieval interface (how the data can be queried). The deployment of a heterogeneous database will greatly benefit the users in translating isolated, multi-sourced data into integrative information. Focusing on reconciliation of text as well as geospatial data, our project will have a great impact on better facilitating earth scientists in collecting and integrating environmental data (images, maps, and texts) for analysis.

Immediate Impact

Students: The following undergraduates have been supported, in part, by our MII grant: Fabian Alcantara, Enrique Almendral, Abraham Anzardo, Michael Armentano, Jorge Besada, Joel Delgado, Jorge Du Quesne, Julie Fernandez*, Alejandro Gonzalez, Freddy Haayen, Robert Hazbun, Alexander Hernandez, Sergio Hernandez, Sheldon Himmelsbach, Jose Iglesias, Alexander Jelinek, Sheldon Jones, Ying Liu, Daniel Mendez*, Reinaldo Morejon, Luis Pachas, Wilbis Padron, Guido Pozo, Patrick Quinlivan, Dario Rivera, Gianina Rocha*, Freddy Rodriguez*, Alejandro Roque, Julio Ruano, and Roberto Valenti. All but one of these students are from under-represented groups; those marked with an * received their B.S. degrees during the past year. The following graduate students have been supported, in part, by our MII grant: Elma Alvarez, Juan Carlos Carrillo, Debra Davis-Chu*, Guillermo Fernandez, Dario Gonzalez, Scott Graham, Martha Gutierrez, Tin Ho, Guangyi Li, Daniel Mendez, Rebecca Miro, Khaled Naboulsi, Philippe Pardo, Michael Perez, Joseph Pontillo, Steve Rios, Gianina Rocha, and Rosany Rodriguez. All but four of these students are from under-represented groups; those marked with an * received their Master's degrees during the past year. The following undergraduates have been supported, in part, by the REU supplement to out MII grant: Abraham Anzardo, Juan Carlos Carrillo*, Luis Espinal, Luis Llanes, Daniel Mendez*, Michael Olivero, Jose Obando, Sebastian Ojanguren, Wilbis Padron, Oscar Parrales, and Rob Valenti. All of these students are from under-represented groups; those marked with an * received their Bachelor's degrees during the past year.

Publications: 15 publications this year acknowledge the support of our MII award, including:

N. Rishe, J. Yuan, R. Athauda, S.-C. Chen, X. Lu, X. Ma, A. Vaschillo, A. Shaposhnikov, D. Vasilevsky. "SemanticAccess: Semantic Interface for Querying Databases". Proceedings of the 26th International Conference on Very Large Databases, 2000. pp. 591-594.

N. Rishe, R. Athauda, J. Yuan, S.-C. Chen. "KnowledgeManagement for Database Interoperability". Proceedings of the ISCA 2nd International Conference on Information Reuse and Integration (IRI-2000). pp. 23-26.

S.-C. Chen, X. Wang, N. Rishe, M. Weiss. "A High-Performance Web-Based System Design for Spatial Data Accesses". Proceedings of the Eighth International Symposium of ACMGIS (ACM GIS 2000). pp. 33-38.

Project Outcome

The heterogeneous database research performed under FIU's MII support has enabled the development of TerraFly, a tool that allows users to fly via the Web over distributed geospatial data, such as satellite imagery, aerial photography, and geographic points of interest [www.terrafly.fiu.edu]. TerraFly is a step toward on-demand merging and visualization of such data using Next Generation Web techniques. In collaboration with US Government and major satellite data suppliers, the TerraFly database is expected to become one of the largest databases on the Web, delivering over 13 terabytes of data to the public, local governments, and industry. The heterogeneous database research enabled by the NSF MII award allows spatial data stored in a variety of formats in various databases and across the web under a unifying database schema that is then accessed by TerraFly. This allows the easy addition of new data types to the TerraFly database. The application of the heterogeneous database research also enables the user to combine data sets, point, and line data in any way that the user desires. TerraFly makes FIU's research immediately usable by a broad range of users and promotes the further use of geospatial data.