Database Research at UT Arlington (ITLab@CSE.UTA)

Sharma Chakravarthy, Alp Aslandogan, Ramez Elmasri, Leonidas Fegaras, and JungHwan Oh
Computer Science and Engineering Department
The University of Texas at Arlington, Arlington, TX 76019
{Sharma, alp, elmasri, fegaras, oh}@cse.uta.edu

The **Information Technology Laboratory (or ITLab**) at the Computer Science and Engineering Department at The University of Texas at Arlington was established by Sharma Chakravarthy in Spring 2000. The mission of the ITLab is to conduct research and development on all aspects of information technology. Some of the topics currently being investigated are: Data Warehousing/Information Integration, Data Mining/Knowledge Discovery, Stream Data Processing, Web Search, Image Databases, Active/Push Technology, Video Streaming, and Object-Oriented, Temporal & Heterogeneous Databases.

The vision of the group is to carry out both fundamental and practically applicable research and development to enable the use of information technology for diverse applications. The group views interaction with industry as critical for identifying fundamental problems in the areas of databases and information technology usage and functionality. Our aim is to identify problems based on the current and future needs, develop abstractions, and provide formal solutions as well as implement proof-of-principle systems. This approach provides a viable migration path for new techniques/solutions to be integrated into real-life systems/applications.

Sharma Chakravarthy is currently leading an effort for establishing a research infrastructure for High-performance Distributed and Parallel Computing Center (DPCC) with 150+ processors and ~20 Terabytes of storage at UTA (funded by NSF) for interdisciplinary collaborative research.

This short report describes the information technology research activities currently underway in our department. Additional information on these projects can be found at the web site http://itlab.uta.edu (or send email to itlab@cse.uta.edu). Currently, the department has a "top 25" initiative to increase research activity and to strengthen the PhD program at the CSE Department.

1 Data Warehousing

Faculty: Sharma Chakravarthy

This project addresses the problem of how to make quality decisions on the choice of data warehouse maintenance policies. For this work, the warehouse is a set of views on autonomous, heterogeneous and distributed sources.

As part of this work, the DW maintenance problem has been characterized and shown to be complex and relevant. For characterizing the problem, three main components have been identified: policies, QoS criteria, and source capabilities. Dependencies between policies, source capabilities and evaluation criteria have been modeled using a cost-based approach for both single and multiple (heterogeneous) data sources. The cost model has been incorporated into an analysis tool for systematically analyzing the search space for policy selection. This approach has turned out to be invaluable for understanding the dependencies between the components, and for deriving a set of heuristics for maintenance policy selection.

As a complement to the analytical results based on the cost model, a testbed (for both single source and multi-source DW) has been implemented for a typical maintenance scenario using Interbase and an XML server as sources. Experiments using the testbed have enabled a validation of the analytical results and an evaluation of the quality of policy decisions based on heuristics.

In addition to the above, work on automatic mapping between source and data warehouse schemas is ongoing.

[1] H. Engstrom, S. Chakravarthy, B. Lings, "A Heuristic or Refresh Policy Selection in Heterogeneous Environments", to appear in Proc. ICDE 2003 (Bangalore, India), March2003 (poster paper).

[2] K. Jagannathan, "An Approach to Schema Mapping Generation for Data Warehouses", MS Thesis, UTA, Fall 2002, http://itlab.uta.edu/ITLABWEB/Students/sharma/theses/karthik.pdf

[3] H. Engstrom, S. Chakravarthy and B. Lings, "Implementation and Comparative Evaluation of Maintenance Policies in a Data Warehouse Environment", in the Proc. of British National Database Conference (BNCOD), Sheffield, July 2002.

[4] H. Engstrom, S. Chakravarthy and B. Lings, "A Systematic Approach to Selecting Maintenance Policies in a Data Warehouse Environment", in the Proc. of Extended Database Technology (EDBT) Conference, Prague, March 2002

[5] H. Engstrom, S. Chakravarthy, B. Lings, "A User-centric View of Data Warehouse Issues", in Proc. of BNCOD International Conference, Exeter, July 2000.

2 WebVigiL/Active Technology

Faculty: Sharma Chakravarthy
Funding Sources: NSF, AF/SPAWAR

Until now, active capability has been mostly investigated in the context of databases and some work has been done for distributed event detection and rule execution. Research has shown that active capability can be effectively used for a variety of applications including information filtering, workflow management, and self-monitoring using which a system can adapt to various kinds of changes. This project investigates research issues in making the push paradigm effective in large network-centric environments such as web and distributed heterogeneous information systems.

The objectives are to investigate the specification, management, and propagation of changes as requested by a user in a timely manner and meeting the quality of service requirements. Our approach allows the users to specify (through the browser) various types of changes that are of interest to them on (web) the documents at different levels of granularity. They can also specify how they need to be notified when the requested information becomes available. Quality of service (QoS) information such as timing constraints, aggregated vs. individual changes will also be a part of the user specification. Based on the user requirements, the techniques developed in this project will determine how these changes are monitored, collected and propagated. User specifications will be translated into a set of rules that are used for monitoring changes and propagating relevant information.

We are extensively drawing upon our previous research on Sentinel/Snoop and our experience in building prototypes for centralized, agent-based, and distributed active capability. The ability to selectively monitor and be informed of changes will augment the current strategy of pulling information periodically and checking for interesting changes.

[1] N. Pandragi, J. Jacob, A. Sanka, and S. Chakravarthy, "WebVigiL: User Profile-Based Change Detection for HTML/XML documents", January 2003. http://itlab.uta.edu/sharma/Projects/WebVigil/files/DiffAlgos.pdf

[2] S. Chakravarthy, J. Jacob, N. Pandragi, and A. Sanka, "WebVigiL: Architecture and Functionality of a Web Monitoring System", October 2002. http://itlab.uta.edu/sharma/Projects/WebVigil/files/WVFetch.pdf

[3] R. Adaikkalavan, "Snoop Event Specification: Formalization, Algorithms, and implementation using Interval-Based Semantics." MS Thesis August 2002, http://itlab.uta.edu/ITLABWEB/Students/sharma/theses/raman.pdf

[4] S. Chakravarthy, J. Jacob, N. Pandragi, and A. Sanka, "WebVigiL: An Approach to Just-In-Time (JIT) Information Propagation in Large Network-Centric Environments", in

the Proc. of Web Dynamics Workshop, Hawaii, May 2002.

[5] S. Chakravarthy and S. Yang, "Architecture and Implementation of an Interactive Tool for the Visualization and Debugging of Active Capability", in the Proc. of Visual Database Conference, Brisbane, Australia, May 2002.

[6] S. Chakravarthy and H. Liao, "Asynchronous Monitoring of Events for cooperative Distributed Environments", in the Proc. of CODAS 2001, Beijing, China, April 2001.

3 Stream Processing/MavHome

Faculty: Sharma Chakravarthy, Diane Cook, Sajal Das, and Larry Holder
Funding Sources: NSF

Research on traditional database management systems (DBMSs) has concentrated on data that has been collected and stored. However, a wide variety of applications – network management, financial applications, and sensor-based system – generate massive amounts of data streams. Their processing requirements are quite different from those used in traditional DBMS applications. The new processing requirements of these and other applications are forcing a re-examination of the approaches and techniques used in a traditional database management system due to its inability to manage and operate on streaming data and provide near real-time response. The data streams generated by these applications need to be processed in a single pass and in real-time, and the computation needs (need for multi-processors and parallel algorithms are critical) to keep up with the data flow.

Currently, we are investigating some of the above problems using the MavHome project (a smart home viewed as an intelligent agent that perceives its environment through the use of sensors, and can act upon the environment through the use of actuators) as an example of a system that generates stream data. We are also investigating the problem analytically by using queuing models to estimate the memory requirements and response time of queries over stream data. The following problems are being investigated in this project:

- A system for processing queries over streams.
- Modeling and estimating memory and response-time (resources in general) requirements of continuous queries using queueing models.
- Mining the stream database to predict usage patterns to optimize the overall system functionality and performance.
- Management and processing of stream data in conjunction with conventional databases.

[1] Q. Jiang and S. Chakravarthy, "Queueing Analysis of SPJ Queries over Continuous Data Streams", http://itlab.uta.edu/sharma/Projects/MavHome/files/QA-SPJQueries.pdf

3.1 Query Processing of Streamed XML Data

Faculty: Leonidas Fegaras and David Levine

This research work addresses the efficient processing of XQueries over continuous streams in which a server broadcasts XML data to multiple clients concurrently and a client may tune-in to multiple streams at a time. The goal of this project is to develop a framework that maximizes the throughput of queries of all clients, taking full advantage of their limited resources. Under this framework, the unit of transmission in an XML stream is an XML fragment, which corresponds to one XML element from the transmitted document. A server may choose to disseminate XML fragments from multiple documents in the same stream, can repeat some fragments when they are critical or in high demand, can replace them when they change by sending delta changes, and can introduce new fragments or delete invalid ones. One difficult problem in processing queries against continuous data streams is the presence of blocking operations, which require the processing of the entire stream before generating the first result, and the unbounded stateful operations, which require caching all stream data in memory.

Our work makes a number of improvements to main-memory evaluation algorithms for these operators by making effective use of condensed

summaries of data to produce results earlier and to flush buffers faster than conventional methods. Servers broadcast these data summaries to the clients along with the data. In this framework, XQueries over continuous XML streams are translated into a novel XML algebra, then optimized, and mapped into evaluation plans based on the improved main-memory algorithms.

[1] L. Fegaras, D. Levine, S. Bose, and V. Chaluvadi. "Query Processing of Streamed XML Data". 11th International Conference on Information and Knowledge Management (CIKM'2002). McLean, VA, November 2002, pages 126-133.

4 Data Mining

Faculty: Sharma Chakravarthy, Diane Cook, Larry. Holder
Funding Sources: NSF

Subdue data-mining system supports mining on complex, structured data. Subdue searches for patterns, or substructures, in a graph-based structural representation of data, which is a much richer representation than used by most other data-mining systems.

We are currently investigating efficient persistence, indexing, and retrieval techniques to improve Subdue's scalability and performance. Scalability is being addressed by translating Subdue algorithms into SQL and SQL-OR extensions on RDBMSs.

We are also investigating the architecture, algorithms, optimization and scalability issues to enable mining directly on data stored in (multiple) databases. We have already developed a layered architecture and have translated several mining algorithms (for association rules) to the relational context using commercial DBMSs. We have successfully implemented K-way join, Query/Sub-query, and 2-GroupBy approaches in SQL92, and tested them for performance and scalability. We have studied several additional optimizations for the K-way join approach and SQL-OR based approaches (VerticalTid, Gather Join and Gather Count) and evaluated them using DB2 and Oracle. We have experimentally evaluated these approaches and their optimizations on large data sets. The larger goal of this work is to feed these results into a mining optimizer that chooses the specific strategy for mining the input dataset based on its characteristics.

[1] S. Chakravarthy and H. Zhang, "Visualization of association Rules over RDBMSs", in the proceedings of ACM SAC 2003, Melbourne, FL, March 2003 (Multi-media and Visualization Track).
[2] P. Mishra, "Performance Evaluation and Analysis of SQL-Based Approaches for Association Rule Mining", MS Thesis, Fall 2002, http://itlab.uta.edu/ITLABWEB/Students/sharma/theses/Pratyush_ThesisReport.pdf
[3] S. Thomas and S. Chakravarthy, "Incremental Mining of Constrained Associations", in Proc. of High Performance Computing (HiPC), Bangalore, India, Dec. 2000 Ph.D. Thesis, http://itlab.uta.edu/sharma/People/ThesisWeb/md1_thesis.pdf
[4] M. Dudgikar, "A layered Optimizer for Mining Association Rules Over Relational Database Management Systems", Summer 2000
[5] S. Thomas, "Architectures and Optimizations for Integrating Data Mining Algorithms with Database Systems", Ph.D. Thesis, Fall 1998, http://itlab.uta.edu/sharma/People/ThesisWeb/sthomas_thesis.pdf

4.1 Data Mining on Raw Video Stream

Faculty: JungHwan Oh
Funding sources: UTA startup funds

Multimedia data mining has been performed for different types of multimedia data; image, audio and video. An example of video and audio data mining can be found in *Mining Cinematic Knowledge* project [1] which creates a movie mining system by examining the suitability of existing concepts in data mining to multimedia, where the semantic content is time sensitive and constructed by fusing data obtained from component streams. We also can find some multimedia data mining frameworks [2] for traffic monitoring system. There have been some efforts about video data mining for

movies, and traffic videos as mentioned above. Among them, the developments of complex video surveillance systems [3] and traffic monitoring systems [4] have recently captured the interest of both research and industrial worlds due to the growing availability of sensors and processors at reasonable costs, and the increasing safety and security concerns. The common approach in these works is that the objects (i.e., person, car, airplane, etc.) are extracted from video sequences, and modeled by the specific domain knowledge, then, the behavior of those objects are monitored (tracked) to find any abnormal situations. What are missing in these efforts are first, how to index and cluster these unstructured and enormous video data for real-time processing, and second, how to mine them, in other words, how to extract previously unknown knowledge and detect interesting patterns.

In this project, we investigate a general framework for real time video data mining to be applied to the raw videos (traffic videos, surveillance videos, etc.). The first step of our framework for mining raw video data is grouping input frames to a set of basic units, which are relevant to the structure of the video. We call this unit as *segment.* These segments are classified into a number of categories. This is one of the most important tasks since it is the step to construct the building blocks for video database and video data mining. The second step is characterizing each segment. To do this, we extract some features (motion, object, colors, etc.) from it. In our framework, we focus on *motion* as a feature, and study how to compute and represent it for further processes. The third step of our framework is to cluster the decomposed segments into a number of groups of similar segments to discover unknown knowledge and to detect interesting patterns. In our clustering, we employ a multi-level hierarchical clustering approach to group segments using *category* and *motion*. Finally, we investigate an algorithm to find whether a segment has normal or abnormal events based on clustering and modeling of normal events, which occur mostly. In addition to deciding normal or abnormal, the algorithm computes *Degree of Abnormality* of a segment, which represents to what extent a segment, is distant from the existing segments in relation with normal events.

[1] I. Pavlidis and V. Morellas and P. Tsiamyrtzis and S. Harp. "Urban Surveillance systems: From the Laboratory to the Commercial World". In *Proc. of IEEE*. 89(10): 1478-1497. October. 2001.

[2] S. Chen and M. Shyu and C. Zhang and J. Strickrott. "Multimedia Data Mining for Traffic Video Sequences". In *Proc. of International Workshop on Multimedia Data Mining (MDM/KDD'2001)* pages 78-86. San Francisco, CA, August 2001.

[3] D. Wijesekera and D. Barbara. "Mining Cinematic Knowledge: Work in Progress". In *Proc. of International Workshop on Multimedia Data Mining (MDM/KDD'2000*) pages 98-103. Boston, MA, August 2000.

[4] R. Cucchiara and M. Piccardi and P. Mello. "Image Analysis and Rule-based Reasoning for a Traffic Monitoring System". *IEEE Transactions on Intelligent Transportation Systems*. 1(2): 119-130. June 2000.

4.2 DESCARTES: Medical Data Mining

Faculty: Alp Aslandogan, Sharma Chakravarthy, Diane Cook, Farhad Kamangar.

In this project we develop data mining techniques and tools that help patients, health-care providers and educators, who screen for dangerous skin lesions, and explore the properties of skin lesion images in a large database. Some of these services are made available via web interfaces over the Internet or Internet 2. We develop data mining services such as visual example-based querying with statistical summarization, feature association mining and classification. For classification, we focus on evidence combination and decision fusion methods. We combine a number of modules such as Bayesian, kNN, neural network, and decision tree based classifiers in a formal evidence combination framework for boosting classification accuracy. The evidence combination methods that we employ include the Dempster-Shafer approach, general evidence processing and the Bayesian approach. With the help of a web service, remote users such as rural

patients, primary care physicians and school nurses are able to access the image database, upload their lesion images and compare with the database of pre-diagnosed images to assess cancer risk. We are also exploring the cognitive aspects of the user interface to customize it as an educational tool for educators such as medical school professors, dermatologists or school nurses. Finally, we provide time series analysis to provide lesion-monitoring (change detection) services to patients with large number of lesions. Efforts are currently underway to incorporate a multi-camera monitoring system into the MavHome intelligent home project.

5 Query Optimization Techniques for OODB Languages

Faculty: Leonidas Fegaras and David Maier
Funding Sources: NSF

Object-Oriented Databases (OODBs) provide powerful data abstractions and modeling facilities but they generally lack a suitable framework for query processing and optimization. One of the key factors for OODB systems to successfully compete with relational systems as well as to meet the performance requirements of many non-traditional applications is the development of an effective query optimizer. We have developed a new database calculus and an optimization framework for object-oriented databases, based on a novel calculus, called the *monoid comprehension calculus* [1][4]. The calculus is based on *monoids*, a general template for a data type that can capture most collection and aggregate operators currently in use for relational and object-oriented databases. Monoid comprehensions are a uniform way to express queries that simultaneously deal with more than one collection type and also naturally compose in a way that mirrors the allowable query nesting in recent database query languages, such as the ODMG OQL and SQL3. The monoid comprehension calculus has sufficient expressive power to represent most language constructs present in these languages and is amenable to effective query optimization. The calculus has been used as the basis for the development of a high performance query optimizer for ODMG OQL. Our query optimization framework was based on query unnesting (query decorrelation), an optimization that, even though improves performance considerably, is not treated properly by most database systems. A framework was developed that generalizes many unnesting techniques proposed in the literature and is capable of removing any form of query nesting using a simple and efficient algorithm [3].

We have already built a high-performance object-oriented database management system based on our theoretical framework for query optimization, called Λ-DB [2]. Our system can handle most ODMG ODL declarations and can process most ODMG OQL queries. It supports embedded OQL in C++, transactions, updates, indexes, and methods with OQL body. It also provides a visual query interface for constructing ad-hoc OQL queries, which can be evaluated by our query interpreter. Our query optimizer unnests all nested queries, materializes path expressions into pointer joins, performs some forms of semantic optimizations, uses a cost-based polynomial-time heuristic for join ordering, and uses a rule-based cost-driven optimizer to generate physical plans. Our query evaluation engine is stream-based and supports many evaluation algorithms, including block-nested loop, indexed nested loop, pointer join, and sort-merge join. The source programs of our system are available at http://lambda.uta.edu/ldb/doc/.

[1] L. Fegaras and D. Maier. "Optimizing Object Queries Using an Effective Calculus". ACM Transactions on Database Systems (TODS), Volume 25, No. 4, December 2000, pages 457-516.

[2] L. Fegaras, C. Srinivasan, A. Rajendran, D. Maier. "Lambda-DB: An ODMG-Based Object-Oriented DBMS". ACM SIGMOD Conference on Management of Data, Dallas, Texas, May 2000, page 583.

[3] L. Fegaras. "Query Unnesting in Object-Oriented Databases". ACM SIGMOD International Conference on Management of Data, Seattle, Washington, June 1998, pages 49-60.

[4] L. Fegaras and D. Maier. "Towards an Effective Calculus for Object Query Languages". ACM SIGMOD International

Conference on Management of Data, San Jose, California, May 1995, pages 47-58.

6 A Temporal Object Query Language and an Optimization Framework

Faculty: Leonidas Fegaras and Ramez Elmasri
Funding Sources: Texas ARP

Temporal databases provide a complete history of all changes to a database and include the times when changes occurred. Because temporal concepts require complex type support and advanced modeling concepts, object-oriented databases are excellent candidates for realization of temporal databases without requiring fundamental extensions to the basic data model. Our research work, supported by a Texas ARP grant, aimed at two directions. First, we designed a language extension to OQL, called TOQL, to accommodate time information. Second, we developed an optimization framework to evaluate TOQL efficiently using advanced temporal evaluation techniques. More specifically, we designed and implemented a translation scheme for converting any TOQL query into an algebraic form and a number of optimization methods to rewrite these forms into temporal joins.

[1] L. Fegaras and R. Elmasri. "A Temporal Object Query Language and an Optimization Framework". Fifth International Workshop on Temporal Representation and Reasoning (TIME-98), Sanibel Island, Florida, May 1998, pages 51-59.

7 Query Engines for Web-Accessible XML Data

Faculty: Leonidas Fegaras and Ramez Elmasri

XML has emerged as the leading textual language for representing and exchanging data on the web. Because of its popularity, many database researchers believe that XML will soon dominate the database area and will eventually replace relational databases. One of the active areas of research is utilizing the already established database technology in storing and querying XML data. We have proposed an effective framework for storing XML data in an object-oriented database and an optimization framework with a solid theoretical basis for translating XML queries into efficient algorithms. Since XML data are often scattered throughout the web as text files and sometimes in unknown locations, we have addressed the problem of querying the entire Internet to answer XML queries by using an indexing technique based on both structure and content information of the XML documents. Based on these indexes, we have developed algorithms for locating relevant documents that match an XML query with high degree of precision. This indexing technique generalizes the inverse indexing techniques used by current web search engines.

[1] L. Fegaras and R. Elmasri. "Query engines for Web-accessible XML data". Very Large Data Bases (VLDB) Conference, Roma, Italy, September 2001, pages 251-260.

8 DIOGENES: Distributed Multimedia Search Agents

Faculty: Alp Aslandogan

In this project we develop distributed multimedia search agents that integrate content analysis, context analysis and background domain knowledge for finding images, videos, sound bytes etc., about different kinds of objects and persons from the web [1]. In content analysis we use object detectors such as human face detectors, animal detectors, building detectors etc. These detectors might be using a neural network, wavelet-based shape analysis or distance-based clustering methods. For context analysis we use text and HTML features that provide semantic information about multimedia elements. These text/HTML features may include word frequency, word position, structural relationships, lexical properties and semantic properties [2]. We represent background domain knowledge in XML with an ontological knowledge representation scheme such as Topic Maps or CKML. The information obtained from the multimedia content and the text/HTML context are enhanced using the background knowledge. The end result of this process is a semantically-rich index that can be used to index the multimedia elements effectively.

[1] Y. Alp Aslandogan, Clement T. Yu, "Automatic Feedback for Content Based Image Retrieval on the Web." IEEE ICME 2002.

[2] Y. Alp Aslandogan, Clement T. Yu, "Evaluating Strategies and Systems for Indexing Person Images on the Web." ACM Multimedia 2000.

9 AVICENNA: An Educational Digital Library for Educators

Faculty: Alp Aslandogan

In this project we develop an educational multimedia library that is able to populate and refresh itself via the automatic search agents developed in project DIOGENES. This database, with the help of automatic presentation generation tools, helps educators generate multimedia presentations with minimal interaction [1]. The system consists of six modules: A presentation template database, a presentation generation engine, the multimedia search agent, the multimedia database, a domain knowledge base and web-based user interface as illustrated in Figure 1.

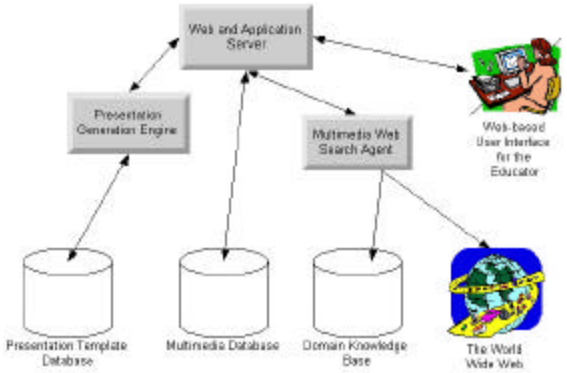

Figure 1: An Automatic Multimedia Presentation Generation Tool

The presentation template database is populated by artists using technologies that allow presentations to be designed with variable elements. Currently we use two such technologies: Macromedia Flash and the XML-based SVG/SMIL combination. The content objects are obtained from the multimedia "content" database. This database is populated automatically by the multimedia search agents. Named entities identified in background domain knowledge sources using information extraction methods are used to generate automatic queries. The search agents, starting with these queries, gather multimedia elements such as images, sound bytes, video clips, graphic illustrations or animations related to those named entities. At the time of presentation generation, the generic templates and the content objects are integrated by the presentation generation engine to produce a full presentation.

[1] Y. Alp Aslandogan, S. Vana, P. Boppana, V. Mariappan, L.Stvan. "Empowering Educators in Visual Learning: Information Extraction and Visualization for Automatic Generation of Multimedia Presentations." To appear in the Proceedings of SITE '03 (Society for Information Technology and teacher Education), Albuquerque, NM.