DATA AND KNOWLEDGE BASE RESEARCH AT HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

P. Drew, B. Hamidzadeh, K. Karlapalem, A. Kean,
D. Lee, Q. Li, F. Lochovsky, C.D. Shum, B. Wuthrich
Department of Computer Science,
Hong Kong University of Science and Technology,
Clear Water Bay, Hong Kong
email: {*pam, hamidzad, kamal, kean,*
dlee, qing, fred, shum, beat}@cs.ust.hk

1 Introduction

This report presents a brief description of the current database research activities of the Data and Knowledge Base Group of the Department of Computer Science of the Hong Kong University of Science and Technology. The department had its first intake of students four years ago, in October, 1991. It currently has about 40 faculty members, 500 undergraduate and 100 postgraduate students. About a quarter of the faculty members have research interests related to database systems.

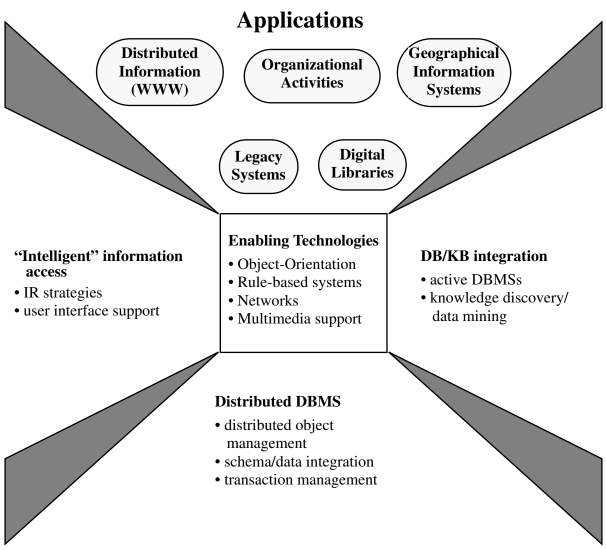

Figure 1: Data and Knowledge Base Research Areas

Our research focus falls into the main areas shown in Figure 1; we use the same categories to roughly organize the project summaries in this report. Abbreviations given for the funding sources are as follows: Research Grants Council Earmarked Grant for Research (RGC/ERG), University Grants Council Research Infrastructure Grant (UGC/RIG), University Grants Council Direct Allocation Grant (UGC/DAG), Sino Software Research Center (SSRC), and Hong Kong Industry Technology and Development Council (ITDC). More information and references about the research projects of our group can be found at our web site [http://www.cs.ust.hk/].

2 Indexing and Searching on the World Wide Web

Investigator: D. Lee

Funding: SSRC

A World Wide Web Index and Search Engine (WISE), accessible at http://www.cs.ust.hk/cgi-bin/IndexServer, was developed. We evaluated the retrieval effectiveness of four ranking algorithms and chose to implement the vector-space model with tfx-idf term weighting in WISE [2]. WISE has a graphical interface that allows the user to navigate forward starting from the retrieved items. It also performs group ranking based on the connections between the retrieved pages so that related items tend to be ranked higher and supports relevance feedback [1]. WISE currently indexes most, if not all, of the WWW servers in Hong Kong. We are working on extending WISE to handle Chinese documents, and to use a distributed indexing scheme so that so that servers can communicate with each other and cooperate in answering a user query.

[1] B. Yuwono, S.L.Y. Lam, J.H. Ying and D.L. Lee, "A World Wide Web Resource Discovery System", *Proceedings of the 4th International World Wide Web Conference*, Boston, MA, Dec 1995.
[2] B. Yuwono and D.L. Lee, "Search and Ranking Algorithms for Locating Resources on the World Wide

Web", *Proceedings of the 12th International Conference on Data Engineering*, New Orleans, LA, Feb 1996.

3 Finding and Managing Information in Distributed, Digital Information Sources

Investigators: F. Lochovsky, P. Drew, Q. Li, and B. Wuthrich

Funding: RGC/ERG, Pending

This research project is investigating techniques and developing facilities for finding and managing information in widely-distributed, digital information sources. There are two main foci to the project. First, we are investigating the use of metadata to represent high-level knowledge of the distributed information sources. In this task, we are exploring the use of "intelligent" software agents that can be sent out to extract the "essence" of an information source (i.e., its metadata). The second focus of the project is on the investigation and development of query formulation (e.g. by example, visual) and multimedia data presentation facilities. Additionally, techniques for effectively and efficiently extracting data using the metadata discussed above are being investigated.

4 Distributed Geographic Information Exchange

Investigator: P. Drew, D. McInnis (Civil Eng.)

Funding: UGC/RIG, SSRC

Focussing on a specific application for Digital Libraries research, the goal of this project is to build a system that allows transparent information exchange between existing, third-party geographic information systems (GIS) via networks, such as the Internet. Our research agenda includes the definition of an appropriate system architecture and services; the representation, storage, and use of metadata; intuitive information discovery and extraction tools; and appropriate system support and algorithms for automated extracts so that users can populate new GIS applications on the fly. Currently, we are developing a prototype of the proposed facility, called GeoChange, which allows translation between various geodatabases provided by the Hong Kong Government Works Branch and some of the local private utility companies.

[1] P. Drew and D. McInnis, "A Heterogeneous Geographic Information Architecture for Hong Kong Infrastructure Systems", *Proceedings of IEEE Workshop on Metadata for Scientific and Technical Data Management*, National Archives II, Washington, D.C., May 1994.
[2] P. Drew and J.Ying, "MetaData Management for Geographic Information Exchange", *Working Paper*, 1995.

5 Indexing and Query Optimization for Document Retrieval

Investigator: D. Lee

Funding: RGC/ERG

In this project, we continue our long-term investigation on efficient indexing and query processing methods for text. We have studied partitioning methods for the signature file to support efficient document ranking. We are exploring further optimization techniques by combining partial ranking with partitioning inverted files. Partial ranking produces a ranking of the documents without processing all of the queries terms or accessing the entire index. The goal is to produce a ranking as accurate as possible (with respect to full ranking which processes all of the query terms and makes use of all information in the index) while maximizing saving in processing cost. In addition, we have examined the optimal generation of signatures to minimize false drop probability.

[1] D.L. Lee and L. Ren, "Document Ranking on Weight-Partitioned Signature Files", *Accepted for publication in ACM Transactions on Information Systems.*
[2] C.W. Leng and D.L. Lee, "Optimal Weight Assignment for Signature Generation", *ACM Transactions on Database Systems*, Vol. 17, No. 2, June, 1992, pp. 346-373.

6 Information Access Methods on Wireless Channels

Investigator: D. Lee

Funding: RGC/ERG, Pending

We investigate techniques for the efficient indexing of information broadcast on wireless channels. Both filtering and caching based on the signature file technique were studied. The performance was evaluated based on access time, tune-in time and probe time. The tune-in time in particular is directly related to the battery consumption of the mobile unit [1]. We are studying access methods and channel management in a dual-channel environment where both broadcast and point-to-point channels are available. The availability of both types of channels allows information to be returned to the user through either type of channels with different cost factors.

[1] W.-C. Lee and D.L. Lee, "Information Filtering in Wireless and Mobile Environments", *Submitted for Publication*, April 1995.

7 Forecasting Currency Exchange Rates

Investigator: B. Wuthrich

Funding: RGC/ERG, UGC/DAG

The objective of this project is to develop decision-support systems for currency exchange traders. Based on input provided by currency traders, the system generates, in the morning before the financial markets open in Hong Kong, a forecasting of the bid rate of various currencies against the US dollar (some of the techniques we use are reported in [1,2]). The predictions include a forecasting of whether the bid rate as expected at 4 pm Hong Kong time will be higher, steady or lower against the bid rate closings over night on Wall Street. Also, we generate forecastings of the exact highest and lowest bid rate seen between the market opening in Sydney and the closing of the Asian markets in Singapore and Hong Kong.

[1] B. Wuthrich, "Probabilistic Knowledge Bases", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 7, No. 5, Oct 1995.
[2] B. Wuthrich, Tong Wing Chung and Krishnan Sankaran, "A Temporal and Probabilistic, Deductive and Object-Oriented Query Language", *Workshop on Temporal Reasoning in Deductive and Object-Oriented Databases*, Singapore, Dec 1995.

8 Automated Stock Picking

Investigator: B. Wuthrich

Funding: RGC/ERG, UGC/DAG

The objective of this project is to develop decision-support systems for mutual fund and portfolio managers. Due to the tremendous amount of possible shares which could be included into a particular portfolio, fund managers usually cannot make an assessment of all the possible shares. The objective is to build systems which will preselect for fund and portfolio managers the potentially promising shares. The fund manager would then only make assessments of the shares already suggested by the system. A preliminary system is running (based on techniques reported in [1]) and we are currently programming a nice user-interface for the system.

[1] B. Wuthrich, "Knowledge Discovery in Databases", *Technical Report, HKUST-CS95-4*, Department of Computer Science, HKUST, 1995.

9 Abductive Inference Engine

Investigator: A. Kean

Deduction and inductive generalization are some of the widely known and used inference techniques. Abduction, another technique, is a logical inference used in explanation finding and a variety of consequence finding. The class of tools - reason maintenance systems, are domain independent tools that provide abductive inference as a sub-problem solver to application systems. One particular such reason maintenance system is the Assumption-based Clause Management System, which has been extended to include approximation in abduction. It has also been extended to provide formal methods of revising knowledge bases based on the theory of belief revision.

[2] A. Kean and G. Tsiknis, "A Corrigendum for the Optimized IPIA", *Journal of Symbolic Computation*, Vol. 17, 1994, pp. 181-187.
[3] A. Kean, "The Approximation of Implicates and Explanations", *International Journal of Approximate Reasoning*, Vol. 9, No. 2, 1993, pp. 97-128.

10 A Capability Based and Event Driven Activity Management System

Investigators: K. Karlapalem and F. Lochovsky

Funding: RGC/ERG, Pending

The objectives of this project are i) to architect a framework for supporting database-centric capability-based and event-driven activity management system [1,2], ii) to implement a centralized CapBasED-AMS [1], and iii) to design and implement a knowledge-based distributed CapBasED-AMS. An Activity Management System is a software system that facilitates the specification, maintenance, and execution of activities. An activity consists of one or more tasks (atomic activities) which can be executed by one or more Problem Solving Agent (PSA). We have already implemented [1] the database-centric centralized CapBasED-AMS, and are currently working on distributed activity management issues.

[1] K. Karlapalem, H. P. Yeung, and P. C. K. Hung, "CapBasED-AMS: A Framework for Capability Based and Event Driven Activity Management System", *Proceedings of 3rd International Conference on Cooperative Information Systems: CoopIS'95*, Vienna, 1995, pp. 205-219.
[2] S. Chakravarthy, K. Karlapalem, S. B. Navathe and A. Tanaka, "Database Supported Cooperative Problem Solving", *International Journal of Intelligent and Cooperative Information Systems*, Sept. 1993, Vol. 2, No. 3, pp. 248-287.

11 Development and Applications of an Advanced Object Modelling Environment

Investigators: Q. Li and F. Lochovsky

Funding: RGC/ERG

We have been developing an active expert DBMS based on advanced object-oriented and rule base facilities. The system is called ADOME (ADvanced Object Modelling Environment) which integrates an existing OODBMS and a rule base system through a versatile bridging mechanism. A prototype of ADOME has so far been constructed, and it is currently being applied to an organizational information and activity management environment.

[1] Q. Li and F. H. Lochovsky, "Roles: Extending Object Behavior to Support Knowledge Semantics", *Proceedings of the Int'l Symp. on Advanced Database Technologies and Their Integration*, Nara, Japan, Oct. 1994, pp. 314-322.
[2] Q. Li and F.H. Lochovsky, "Advanced Database Support Facilities for Groupware Systems", *Proceedings of 4th Workshop on Information Technologies and Systems: WITS'94*, Vancouver, Dec. 1994, pp. 292-301.

12 Selectivity Estimation and Access Methods for Real-Time Database Systems

Investigators: B. Hamidzadeh and K. Karlapalem

Real-time databases are those in which a transaction or a query needs to be processed within a deadline. A major aspect of such databases is to provide predictability about the timeliness of the responses they produce. The objective of this project is to be able to efficiently access the data items required for answering a real-time query. Another objective is to obtain an accurate estimate of the time it will take to execute a query. In this project, we have designed and implemented a set of index and access mechanisms that provide us with efficient access to the data in a real-time database. The index and access mechanisms also provide statistical information regarding the selectivity estimate of a query which is very useful for accurately estimating the execution time of that query.

[1] B. Hamidzadeh, K. Karlapalem and H. P. Yeung, "Memory Access and Selectivity Estimation Methods for Real-Time Main-Memory Databases", *Technical Report, HKUST-CS94-17*, Department of Computer Science, HKUST, 1994.

13 Parallel Multimedia/Video Disk Server

Investigators: C.D. Shum and T.F. Ngai (Computer Science)

Funding: SSRC

The objective of this project is to provide concurrent, multi-user, real-time and non-interruptive access to video data on disks. We design and implement a scalable, parallel and distributed disk server prototype. The system bandwidth is equal to the aggregate bandwidth of individual disk. We addressed the real-time disk scheduling problem by providing a real-time operating system support and implementing a seek time optimized disk scheduling policy similar to CSCAN. We developed a video file system featuring buffer cache bypass and user-controlled data placement. The current two-node system is under experimentation and evaluation.

14 Dynamic Object Clustering with Spatial/Multimedia Applications

Investigator: Q. Li

Funding: RGC/ERG, SSRC

This research is to extend the conventional OODB models and systems with advanced features to accommodate the dynamic nature of many non-traditional applications. These new features will support the dynamic grouping/clustering of objects (with assigned roles) to form new "composite objects". A Dynamic Object Clustering (DOC) mechanism has been designed based on a conceptual clustering model, and a DOC prototype being implemented on top of a commercial OODBMS. Real-life applications in the areas of spatial and/or multimedia systems will be used to test and refine the DOC prototype. In a related project, we are developing a video database management in which collections of video objects can be dynamically created and manipulated [3].

[1] Q. Li and J.L. Smith, "A Conceptual Model for Dynamic Clustering in Object Databases", *Proceedings of the 18th Int'l Conference on VLDB*, Vancouver, Aug. 1992, pp. 457-468.

[2] Q. Li, "Advanced Functions for Conceptual Clustering in Object Databases", *Proceedings of the 1995 Int'l Conference on Applications of Databases: ADB'95*, Santa Clara, CA, Dec. 1995.

[3] Q. Li and C.M. Lee, "Dynamic Object Clustering for Video Database Manipulations", *Proceedings of IFIP 2.6 Third Working Conference on Visual Database Systems: VDB-3*, Lausanne, Switzerland, 1995, pp. 125-137.

15 Distributed Object Management

Investigator: D. Lee

Funding: RGC/ERG

This project investigates the problems of indexing and query processing in a distributed object environment. We proposed an access structure called the path dictionary index (PDI) for a centralized environment. The connection information between objects through object identifiers is stored in PDI separated from attribute indexes. The separation reduces storage overhead and update costs significantly without imposing any penalty on retrieval. The complete connection information kept in PDI supports a wide variety of query types, including combinations of top-down, bottom-up, target-predicate and predicate-target queries. We are extending PDI to a distributed environment and the construction of a prototype for the distributed environment based on PDI is planned.

[1] W.-C. Lee and D.L. Lee, "Combining Indexing Technique with Path Dictionary for Nested Object Queries", *Proceedings of the 4th International Conference on Database Systems for Advanced Applications: DASFAA '95)*, Singapore, Apr. 1995, pp. 107-114.

[2] W.-C. Lee and D.L. Lee, "On Processing Nested Queries in Distributed Object-oriented Database Systems", *Proceedings of the 5th International Workshop on Research Issues in Data Engineering - Distributed Object Management: RIDE-DOM '95*, Taipei, Taiwan, March 1995, pp. 10-17.

16 Class Partitioning Schemes in Object-Oriented Databases

Investigators: K. Karlapalem, Q. Li, S. Vieweg (Post-Doctorate)

Funding: RGC/ERG, Pending

We have been working on the issue of facilitating class partitioning schemes in the context of a distributed OODB system. Meaningfully partitioning a class of objects into smaller fragments allows more efficient accesses to objects by focusing the search space. In this proposed research, we investigate the different types of class partitioning schemes that can arise in an OODB system by studying the concepts, representations, and implementation approaches for partitioning OODB classes.

[1] K. Karlapalem and Q. Li, "Partitioning Schemes for Object Oriented Databases", *Proceedings of IEEE 5th Int'l Workshop on Research Issues in Data Engineering – Distributed Object Management: RIDE-DOM'95*, Taipei Taiwan, March 1995, pp. 42-49.

[2] K. Karlapalem, Q. Li and S. Vieweg, "Method Induced Partitioning Schemes in Object-Oriented Databases", *Technical Report HKUST-CS95-38*, Department of Computer Science, HKUST, Aug. 1995.

17 Incremental Homogenization of Legacy Database Systems

Investigators: K. Karlapalem, D. Lee, Q. Li, and C.D. Shum

Funding: ITDC, Pending

This project addresses the issue of legacy database migrations, by proposing a flexible architecture (called HODFA) that can accommodate incremental mirroring of legacy database data onto a homogenized database federation environment. The proposed architecture is currently being experimented, and we plan to work with local industry to test and refine the underlying concepts and techniques based upon the feedback we receive.

[1] K. Karlapalem, Q. Li and C.D. Shum, "An Architecture for Homogenizing Federated Databases", *Proceedings of ISCA Int'l Conf. on Parallel and Distributed Computing Systems*, Poster Session, Oct. 1994, pp.318-319.

[2] K. Karlapalem, Q. Li and C.D. Shum, "HODFA: An Architectural Framework for Homogenizing Heterogeneous Legacy Databases", *ACM SIGMOD Record*, Vol. 24, No. 1, March 1995, pp.15-20.

18 Redesign of Distributed Relational Database

Investigator: K. Karlapalem

Funding: RGC/ERG

The objectives of this project are i) to develop methodologies to adaptively redesign distributed relational databases based on the changes in the distributed database environment, ii) to implement a distributed relational database design tool, and iii) to conduct case-studies and empirical performance evaluation to validate the redesign methodologies. The aim of a good distributed database design is to reduce irrelevant data access (fragmentation scheme) and data transfer (allocation scheme). But changes in distributed database environment and the application processing characteristics warrant constant redesign of the distributed database. As part of this project, efficient materialization algorithms have been developed, and are being implemented and tested.

[1] S. B. Navathe, K. Karlapalem and M. Ra, "A Mixed Fragmentation Methodology For Initial Distributed Database Design, *To Appear in Journal of Computer and Software Engineering.*

[2] K. Karlapalem and M. P. Ng, "Query Driven Data Allocation Algorithms for Distributed Database Systems", *Technical Report, HKUST-CS95-07*, Department of Computer Science, HKUST, 1995.

19 Global Integrity Constraints Across Multidatabase Views

Investigator: P. Drew

Funding: SSRC

This project investigates a specification language in which users declare how to maintain the integrity of inter-related data distributed across multidatabases with interdatabase integrity constraints and tailored user views. At run-time, a query translator and transaction processing system use the specification to execute the appropriate queries and updates to the backend systems, even if a user updates only part of the inter-related data. A prototype has been implemented and is being interfaced with the commercial, heterogeneous database manager, InterViso. A case study which analyzes the utility of the language in industrial information processing environments is underway.

[1] L. Do, P. Drew, and W. Tang, "A Heterogeneous Database Management Architecture to Support Tailored User Views over Interdatabase Constraints", *Invited Submission to Information Systems*, June 1995. Abstract in *Proceedings of the 4th International Workshop on Information Technologies and Systems: WITS'94*, Vancouver, December 1994, pp. 262-271.

[2] P. Drew and W. Tang, "The Composition of Updatable Views over Multidatabase Systems", *Submitted for Publication*, 1995.

20 Asynchronous Distributed Transaction Management

Investigator: P. Drew

Funding: SSRC

The objective of this project is to develop an execution framework in which transactions can update related information distributed across heterogeneous databases asynchronously, yet the consistency of the source information is maintained according to some specified correctness criteria. Our approach includes the development of the theoretical foundation for asynchronous transaction management, algorithms which implement the theory, and a systematic framework in which the algorithms can execute[1,2,3]. In a related project, we are collaborating with Prof. Calton Pu of the Oregon Graduate Institute on consistency restoration techniques for the general case of Epsilon serializability (ESR) in which transactions can import *and* export inconsistency [4]. We have also applied ESR to the management of inconsistency in multidatabase environments[2].

[1] L. Do and P. Drew, "Active Database Management of Global Integrity Constraints in Heterogeneous Database Environments", *Proceedings of IEEE 11th International Conference on Data Engineering*, Taipei, Taiwan, March 1995, pp. 99-108.

[2] L. Do and P. Drew, "The Management of Interdependent Asynchronous Transactions in Heterogeneous Database Environments", *Proceedings of the 4th International Conference on Database Systems for Advanced Applications: DASFAA'95*, Singapore, April 1995, pp. 16-25. Also *Invited Submission to Journal of Systems Integration*, 1995.

[3] L. Do and P. Drew, "A Formal Description of Local Asynchronous Updates in Multidatabase Environments", *Technical Report HKUST-CS95-25*, Department of Computer Science, Hong Kong University of Science and Technology, May 1995.

[4] P. Drew and C. Pu, "Asynchronous Consistency Restoration under Epsilon Serializability", *Proceedings of 28th Hawaii International Conference on System Sciences*, Maui, Jan. 1995, pp. 717-726.