Research Fellows and Research Projects (2003/2004)

CS programme

Dr Cui Bin
Expertise: Database management systems, multi/high-dimensional databases, main memory indexing, concurrency control techniques, location based services

Indexing High-Dimensional Data for Efficient In-Memory Similarity Search

Project Advisor (Singapore)	:	Assoc Prof Tan Kian Lee
Duration	:	March 2004 to February 2005

Project Abstract :
In main memory systems, the L2 cache typically employs cache line sizes of 32-128 bytes. These values are relatively small compared to high-dimensional data, e.g. > 32D. The consequence is that existing techniques (on low-dimensional data) that minimize cache misses are no longer effective. We propose a novel index structure, called Delta-tree, to speed up the high-dimensional query in main memory environment. The Delta-tree is a multi-level structure where each level represents the data space at different dimensionalities: the number of dimensions increases towards the leaf level which contains the data at their full dimensions. The remaining dimensions are obtained using Principal Component Analysis, which has the desirable property that the first few dimensions capture most of the information in the dataset. Each level of the tree serves to prune the search space more efficiently as the lower dimensions can reduce the distance computation and better exploit the small cache line size. Additionally, the top-down clustering scheme can capture the feature of the dataset, and hence reduces the search space. We also propose an extension, called Delta+-tree, that globally clusters the data space and then further partitions clusters into small regions. The Delta+-tree can further reduce the computational cost and cache misses. We conducted extensive experiments to evaluate the proposed structures against existing techniques on different kinds of datasets.

Optimizing Main Memory Utilization for Moving Object Indexing

Project Advisor (Singapore)	:	Assoc Prof Tan Kian Lee
Duration	:	March 2004 to February 2005

Project Abstract :
With the rapid advancement in wireless communications and positioning techniques, it is now feasible to track the positions of moving objects. However, existing indexes and associated algorithms which are usually disk-based are unable to keep up with the high update rate while providing speedy retrieval at the same time. Since main memory is much faster than disk, efficient management of moving-object database can be achieved through aggressive use of main memory. We propose an Integrated Memory Partitioning and Activity Conscious Twin-index (IMPACT) framework where the moving object database is indexed by a pair of indexes based on the properties of the objects' movement - a main-memory structure manages active objects while a disk-based index handles inactive objects. As objects become active (or inactive), they dynamically migrate from one structure to the other. Moreover, the main memory is also organized into two partitions - one for the main memory index, and the other as buffers for the frequently accessed nodes of the disk-based index. In this way, active objects that generally incur higher update rates can be processed more efficiently in memory; while there will be less update activities occurring at the disk. To realize the framework, we employ grid structures for the main memory index and disk-based structure, and adapt the OLRU buffering strategy as the buffer management scheme. We also devise a memory partitioning scheme to optimally allocate space for the main memory index and buffers.

Dr Fang Bin
Expertise: Computer vision, pattern recognition, image processing in biometrics, document processing, and medical image processing

Vascular Structures Identification and Registration in Retinal Images

Project Advisor (Singapore)	:	Assoc Prof Wynne Hsu
Duration	:	October 2002 to September 2003

Project Abstract :
Temporal registration of retinal images is fundamental in tracking evolution of eye-related diseases and providing important information for physicians to decide further treatments. In this project, we propose an elastic matching-based registration algorithm for fundus images using vascular structure features. A two-stage process is first presented to identify and extract vascular structure. Vessels are enhanced by mathematical morphology transformation with respect to their spatial properties and are differentiated from background patterns by curvature evaluation and linear filtering. Morphology reconstruction is performed using dynamic local region growing to recover the complete vascular structure. With the recovered vascular structure, we perform the registration using an elastic matching algorithm. Prior to registration, the extracted vessels are thinned and approximated using short line segments of equal length. Two sets of images over different time periods are used to obtain their respective extracted vessels: the first set corresponds to the template elements, while the second set corresponds to the input elements. The template is elastically deformed to optimally match the input. We have tested the algorithm on 98 pairs of temporal fundus images with 100% success in registration.

Dr Hu Yanhong
Expertise: Modeling and simulation, automation, peer-to-peer system

Efficient Algorithms for Load Shuffling in Split Platform AS/RS

Project Advisor (Singapore)	:	Assoc Prof Hsu Wen Jing
Duration	:	July 2003 to July 2005

Project Abstract :

In conventional Automated Storage and Retrieval Systems (AS/RS), stacker cranes are used to access the AS/RS racks. However, the stacker cranes are inadequate for heavy loads. To handle extra heavy loads at high performance, in this project, we presented a novel design of a split platform AS/AS, or SP-AS/RS for short. With the SP-AS/RS, the horizontal movement and vertical movement of a load are carried out by separate devices, namely the horizontal platforms and the vertical platform. These platforms can operate independently and concurrently. Theoretical analysis and experimental results all show that SP-AS/RS can provide better performance. To further expedite the load retrievals, we addressed the issue of shuffling loads in SP-AS/RS. 2D SP-AS/RS for shuffling was specially designed to achieve the shuffling efficiently. The corresponding shuffling algorithms were developed. The response time of retrieval, the lower and upper bounds of energy consumption were also derived. Results of the analysis and numerical experiments showed that the shuffling algorithms are quite efficient indeed. We also extended our results to 3D AS/RS.

Adaptive Location Services over Peers Networks

Project Advisor (Singapore)	:	Assoc Prof Hsu Wen Jing
Duration	:	January 2004 to July 2005

Project Abstract :

Naming and lookup services are essential in any distributed systems for locating resources. Conventional systems use either centralized or hierarchic schemes which often create hot spots in a system. Recent developments in peer-to-peer systems offer new scalable alternatives. Peer-to-Peer systems, or P2P systems for short, are distributed, self-organizing systems without any central authority. Nodes or peers in P2P systems are of equal roles and capacities in sharing resources (including data, computing power etc.), and they can communicate directly with each other. To evaluate different options for the overlay network (i.e. the communication links between the peers), the cost and performance of a design are measured mainly in terms of the node degrees of the overlay network, latency (number of hops) associated with a query and overhead due to the topology-maintenance protocols. Compared with unstructured P2P systems, structured systems offer better performance and are more scalable, but they are less robust. The goal of this project is to design structured P2P systems for efficient lookup service that adapt to dynamic fluctuations in the overlay networks due to nodes' joining, departure, and failures. We will design and analyze suitable overlay topologies and protocols for supporting various forms of queries.

Dr Li Xiaoli

Expertise: Data mining, machine learning, artificial intelligence, text retrieval and bioinformatics

Extracting Interesting Rules from Medical Database

Project Advisor (Singapore)	:	Assoc Prof Leong Tze Yun
Duration	:	March 2003 to December 2003

Project Abstract :

An important problem in bioinformatics is how to get knowledge from the vast database. Association rules are very useful in practical applications. However, it tends to produce a huge number of rules, most of which are of no interest to the user. Due to the large number of rules, it is very difficult for the user to analyze them manually in order to identify those truly interesting ones. In this project, we propose a new approach to finding interesting rules. Based on the analysis the false examples which is usually misclassified by a particular classifier, we can automatically get the interesting rules which can help the doctors to make decisions when a patient's data is difficult to do judgement. This kind of rules is very important since they are dedicated to hard cases and they are similar to the knowledge of experienced doctors.

Automated Information Extraction to Support Biomedical Decision Model Construction

Project Advisor (Singapore)	:	Assoc Prof Leong Tze Yun
Duration	:	October 2003 to June 2004

Project Abstract :

Decision analysis aids decision-making under uncertainty by systematically representing, analyzing, and solving complex decision models. With the rapid advancement of biomedical knowledge, a large quantity of new findings, methodologies, and insights are published and made available online. Decision model construction in biomedical decision analysis can be greatly facilitated by automatically deriving the relevant semantic knowledge from online biomedical resources. Our proposed framework first classifies text-based documents from a large biomedical literature repository, e.g., MEDLINE, into predefined categories, e.g., diagnosis, screening, symptoms, etc. Then we try to find those authoritative and typical documents within a category to perform further knowledge extraction. For each category, for example, “colorectal cancer screening”, we find those documents which describe the important and comprehensive screening methods and steps. Templates for each category, in the form of Bayesian networks, can then be constructed with the help of the domain experts. Data mining and information extraction techniques are then applied to extract the semantic knowledge for filling in the templates to construct the final decision models.

Dr Ng Wee Siong
Expertise: Distributed computing, peer-to-peer data management, grid computing, database query processing and mobile agents

Peer-to-Peer Data Sharing and Management

Project Advisor (Singapore)	:	Prof Ooi Beng Chin
Duration	:	July 2003 to June 2005

Project Abstract :
In a distributed P2P system, nodes of equivalent capabilities and responsibilities pool their resources together in order to share information and services. Such systems are inexpensive, easy to use, highly scalable and do not require central administration. However, many of the existing P2P systems are limited in several ways. First, they provide only file level sharing (coarse granularity) and lack object/data management capabilities and support for content-based search. Second, there is no predetermined global schema shared among nodes. As a result, the query is largely based on keywords. Third, they are limited in extensibility and flexibility. Finally, a node's peers are typically statically defined. Based on the above observation, there is a great demand for much research on data sharing and query processing in the presence of dynamic peers and heterogeneous data sources.

Dr Qin Shengchao
Expertise: Formal specification and verification, hardware/software co-design, embedded systems, type systems, program analyses

Real-Time Java and its Implementation

Project Advisor (Singapore)	:	Assoc Prof Chin Wei Ngan
Duration	:	July 2003 to July 2005

Project Abstract :
Java is a relatively new and popular programming language. It provides a safe, garbage-collected memory model and enjoys broad support in industry. The goal of the Real Time Specification for Java is to extend Java to support key features required for writing real-time programs. These features include support for real-time scheduling and predictable memory management. As a short-term goal, the project addresses critical analyses and implementation issues on Real-Time Java memory management extensions by introducing sized regions into the memory model, so as to preserve the safety of the base Java memory model while giving the real-time programmers the additional control that they need to develop programs with predictable memory system behavior. Recently we have worked out a region type inference system for a significant subset of Java language, through which region annotations can be automatically inserted into programs to achieve region-based memory management. The extension of this work to the whole language and the experiment on larger benchmarks are on-going. We shall also address memory space issues for Java programs. We aim to use type-based static analysis to identify a class of memory conscious programs that are critical for important applications, like embedded software with limited memory footprint. Our system should be able to predict the memory requirement ahead of execution of such programs. As a mid-term goal, the project will also investigate the real-time scheduling for Real-Time Java.

Dr Stefan Andrei
Expertise: Compiler construction and formal languages, logic and functional programming, networking (Java), real-time systems

Incremental Satisfiability Counting for Debugging Real-Time Systems

Project Advisor (Singapore)	:	Assoc Prof Chin Wei Ngan
Project Advisor (MIT)	:	Assoc Prof Martin Rinard
Duration	:	June 2003 to July 2005

Project Abstract :
Testing constraints for real-time systems are usually verified through the satisfiability of propositional formulae. In this paper, we propose an alternative where the verification of timing constraints can be done by counting the number of truth assignments instead of boolean satisfiability. This number can also tell us how "far away" is a given specification from satisfying its safety assertion. Furthermore, specifications and safety assertions are often modified in an incremental fashion, where problematic bugs are fixed one at a time. To support this development, we propose an incremental algorithm for counting satisfiability. Our proposed incremental algorithm is optimal as no unnecessary nodes are created during each counting. This works for the class of path RTL ([JaM87, WaM94]). To illustrate this application, we show how incremental satisfiability counting can be applied to a well-known rail-road crossing example, particularly when its specification is still being refined.

Compiler Techniques (Scanning, Parsing, Error Recovery)

Project Advisor (Singapore)	:	Assoc Prof Chin Wei Ngan
Project Advisors (MIT)	:	Assoc Prof Martin Rinard & Assoc Prof Saman P. Amarashinghe
Duration	:	August 2002 to June 2004

Project Abstract :
Language recognition has received considerable attention as a fundamental problem in many diverse fields. General context-free languages are of particular importance since they are of sufficient complexity to model interesting real-world phenomena yet are recognizable by efficient algorithms. For example, CFL's are used effectively for syntactic

pattern recognition, programming language compiling and natural language processing. An LR parser will detect an error when it consults the parsing action table and finds an error entry ([ASU88]). An error recovery parser is able to provide all the errors in the program, not just the first error ([App98]). There exist recent classes of grammars which offer satisfactorily syntax error. In general, it is undecidable if an arbitrary context-free grammar has a regular solution. Past work has focused on special cases, such as one-letter grammars, non self-embedded grammars and the finite-language grammars, for which regular counterparts have been proven to exist. However, little is known about grammars with the self-embedded property. Using systems of equations, we highlight a number of subclasses of grammars, with self-embeddedness terms, such as XaX and gXg that can still have regular languages as solutions. Constructive proofs that allow these subclasses of context-free grammars to be transformed to regular expressions are provided. We also point out a subclass of context-free grammars that is inherently non-regular. Our latest results can help demarcate more precisely the known boundaries between the regular and non-regular languages, within the context-free domain.

Dr Wang Xianbing

Expertise: Distributed fault-tolerant computing, grid computing, mobile agents, and peer-to-peer

Fault-Tolerant Distributed Computing

Project Advisor (Singapore)	:	Assoc Prof Teo Yong Meng
Duration	:	January 2004 to December 2005

Project Abstract :

Consensus is a fundamental problem of fault-tolerant distributed computing for processes to reach a common decision despite failures. Most consensus algorithms for both synchronous and asynchronous distributed systems are based on the notion of round, and achieve consensus by exchanging messages during each round. Time complexity and message complexity are used to evaluate the efficiency of a consensus algorithm. There are various existing results on lower bound for consensus problem. But most of those results achieved by backward induction therefore are complicated and difficult to follow. Thus, the first objective of this project is to provide simpler and more intuitive proofs for some existing results, including:

(1) A simple bivalency proof of the lower bound for synchronous uniform consensus protocols.

(2) A simple bivalency proof of the lower bound for early-stopping synchronous consensus protocols.

The second objective is to obtain some new consensus protocols which are time or message efficient, including:

(1) Investigate the consensus with orderly crash failures.

Traditionally, the consensus problem is considered in a fully connected network. In practice, most of the network topologies may not be fully connected, due to the cost of the fully connected network is huge. Thus, the third objective of this project is to study consensus in un-fully connected network:

(1) Especially on chordal rings, in which non-round based consensus protocol will be used to reduce the message complexity.

Grid Computing and Peer-to-Peer

Project Advisor (Singapore)	:	Assoc Prof Teo Yong Meng
Duration	:	August 2003 to August 2005

Project Abstract :

Grid computing is an emerging technology that enables the utilization of shared resources distributed across multiple administrative domains, thereby providing dependable, consistent, pervasive, and inexpensive access to high-end computational capabilities in a collaborative environment. These resources can include supercomputers, storage systems, data sources and special classes of devices. Because grid environments are characteristically dynamic, and the grid must adapt to the aggregation and dislodging of resources any time in operation, P2P will be used for resource discovery and look-up. And P2P is also considered to help scheduling in computational grid. To achieve dynamic load balancing, process migration will be investigated in grid scheduling. Because most applications are inherently sequential and difficult to parallelize effectively, there is a need to provide middleware and feasible higher-level programming models to facilitate application development in a distributed environment. Thus, the research work is on ALiCE, which is a grid computing core middleware developed in NUS for secure, reliable and efficient execution of distributed applications on any Java-compatible platform.

Dr Xiong Xuejian

Expertise: Bioinformatics, machine learning, data mining, pattern recognition, information retrieval

Analysis of Gene Expression Data

Project Advisor (Singapore)	:	Assoc Prof Tan Kian Lee
Duration	:	July 2002 to July 2004

Project Abstract :

DNA microarrays offer the ability to measure the levels of expression of thousands of genes simultaneously. These arrays consist of large numbers of specific oligonucleotides or cDNA sequences, each corresponding to a different gene, affixed to a solid surface at very precise location. Typically gene expression data sets have high dimensionality and a lot of varieties, and the data often contain `technical' noise that can be introduced at a number of different stages. Analysis of DNA gene microarray expression data is a fast growing research area that interfaces various disciplines such as biology, biochemistry, computer science and statistics.

It is concluded that clustering and classification techniques can be successfully employed to group genes based on the similarity of their expression patterns. However, classification on the basis of microarray data presents some algorithmic challenges. In this project, the machine learning methods are explored for learning and classifying multiple gene classes gene microarray expression data.

Dr Zhang De
Expertise: Machine learning, information retrieval and data mining

Web Taxonomy Integration

Project Advisor (Singapore)	:	Assoc Prof Lee Wee Sun
Duration	:	September 2003 to January 2004

Project Abstract :
We address the problem of integrating objects from a source taxonomy into a master taxonomy. This problem is not only currently pervasive on the web, but also important to the emerging semantic web. A straightforward approach to automating this process would be to learn a classifier that can classify objects from the source taxonomy into categories of the master taxonomy. Our key insight is that the availability of the source taxonomy data could be helpful to build better classifiers for the master taxonomy if their categorizations have some semantic overlap. In this project, we propose and evaluate new approaches to taxonomy integration that can enhance the classification by exploiting such implicit knowledge. Our experiments with real-world web data show substantial improvements in the performance of taxonomy integration.

Biomedical Information Retrieval and Digging

Project Advisor (Singapore)	:	Assoc Prof Lee Wee Sun
Duration	:	January 2004 to January 2006

Project Abstract :
The Biomedical Information Retrieval & Digging (BIRD) project aims at creating computer programs that can automatically read and analyze the tremendous amount of biomedical information on the Internet. We think the key feature of biomedical information is the rich data with rich annotations and linkages. The biomedical data resources include: text literature (e.g., journal/conference papers, book reviews), bibliographic databases (e.g., MEDLINE), genome databases (e.g. mouse, yeast), gene/protein function databases (e.g., GeneOntology, LocusLink, GeneRIF), and so on. Such a data richness poses great challenges and opportunities to computer scientists.