Program Overview

1st Day	8:30 AM – 9:00 AM	Opening and Award Ceremony	
	9:00 AM – 10:30 AM	Keynote Speech 1	
	10:30 AM – 11:00 AM		
	11:00 AM – 12:30 PM	Research Session 1	Tutorial 1
	12:30 PM – 2:00 PM		
	2:00 PM – 3:30 PM	Tutorial 2a	Tutorial 3a
	3:30 PM – 4:00 PM		
	4:00 PM – 5:30 PM	Tutorial 2b	Tutorial 3b
	5:30:00 PM - 6:30 PM	Poster Preview	Demo Session
	6:30 PM onwards	Reception + Poster Session + Demo Session	
2nd Day	9:00 AM – 10:30 AM	Keynote Speech 2	
	10:30 AM – 11:00 AM		
	11:00 AM – 12:30 PM	Research Session 2	Sponsor Session
	12:30 PM – 2:00 PM		
	2:00 PM – 3:30 PM	Tutorial 4a	Research Session 3
	3:30 PM – 4:00 PM		
	4:00 PM – 5:30 PM	Tutorial 4b	Industrial Session
	6:30 PM onwards	Banquet	
3rd Day	9:00 AM – 10:30 AM	Keynote Speech 3	
	10:30 AM – 11:00 AM		
	11:00 AM – 1:00 PM	Tutorial 5	Tutorial 6
	1:00 PM – 2:30 PM		
	2:30 PM – 4:30 PM	Research Session 4	Panel(TBD)
		End of Conference	

Keynote Speech 1: “Information Search in Peer-to-Peer Systems”

Abstract

The peer-to-peer (P2P) computing paradigm has been very successful in the proliferation of global applications like file sharing in Internet-wide communities (e.g., Gnutella, BitTorrent) or IP telephony (e.g., Skype). P2P systems should be completely decentralized and should work without any centralized components that could become bottlenecks in terms of performance, availability, or vulnerability to attacks. They should be scalable without any limitations, by being able to grow from a few nodes to many millions of computers. They emphasize the autonomy of the underlying computers and should tolerate frequent node failures, high dynamics in terms of rapidly changing data and load characteristics, and high churn by allowing nodes to join and leave the network without prior notice. They should even be robust to misbehaving peers which may span egoistic, cheating, and malicious peers. None of these salient properties should require any global planning, administration, or control. So P2P systems should be completely self-organizing. In addition, P2P systems should have a software architecture that is much simpler than that of big monolithic systems, to enable scalability and self-organization.

Despite the impressive success of P2P file-sharing applications, the question is still valid and, in my opinion, widely open whether the outlined P2P utopia will become practically viable also for more advanced forms of global data management and information search with more sophisticated functionality. This talk discusses this question, aiming to identify design principles and building blocks towards the P2P dream of simple, scalable, and self-organizing data management on an Internet scale. Advanced applications that drive the discussion include Google-style Web search implemented in a P2P manner, decentralized Web archiving with time-travel query support, and P2P publish-subscribe functionality for scholarly information (e.g., publications, projects, conference sites and reports, etc.) and similar social communities.

Speaker

Gerhard Weikum is a Scientific Director at the Max-Planck Institute for Informatics in Saarbruecken, Germany, where he is leading the research group on databases and information systems. Earlier he held positions at Saarland University in Germany, ETH Zurich in Switzerland, MCC in Austin, Texas, and he was a visiting senior researcher at Microsoft Research in Redmond, Washington. His recent working areas include implementation, optimization, and self-organization aspects of distributed information systems such as peer-to-peer systems, and intelligent search and organization of semi-structured data on the Web and in digital libraries. Dr. Weikum has received several best paper awards including the VLDB 2002 ten-year award, and he is an ACM Fellow. He has served on the editorial boards of various journals and book series, including ACM TODS, IEEE CS TKDE, and the Springer LNCS series, and as program committee chair for international conferences like ICDE 2000 and ACM SIGMOD 2004. He is currently the president of the VLDB Endowment.

Keynote Speech 2: “StrangerDB: Safe Data Management with Untrusted Servers”

Abstract

Imagine that you and your friends want to share information in a database because you want concurrency control, recovery, and query processing, but you don't trust the database administrator. You want to protect data from being observed (privacy). You want to make unauthorized modifications evident (a form safety). You want to force the server to deliver a consistent picture to all honest users or be discovered (a form of liveness). Encryption and signatures make the first two possible. Liveness is another matter since the database administrator could "fork" the database into several copies, keeping some of your friends ignorant of your latest updates and you ignorant of theirs. In joint work with David Mazieres and some great students, we have worked out how to achieve these properties for file systems. This talk presents a design for database systems that integrates these goals with query processing, concurrency control, and recovery.

Speaker

Dennis Shasha is a professor of computer science at the Courant Institute of New York University where he works with biologists on pattern discovery for microarrays, combinatorial design, and network inference; with physicists, musicians, and financial people on algorithms for time series; and on database applications in untrusted environments. Other areas of interest include database tuning as well as tree and graph matching. Because he likes to type, he has written five books of puzzles, a biography about great computer scientists, and technical books about database tuning, biological pattern recognition and time series. He has co-authored fifty journal papers, sixty conference papers, and seven patents. For fun, he writes the puzzle column for Scientific American. Until July of 2007, he is at INRIA, Rocquencourt (near Paris, France) with the group of Philippe Pucheral.

Keynote Speech 3: “Taming the dynamics of Distributed Data”

Abstract
Data gathered from (wireless) sensor networks and those delivered or streamed today via the internet reflect rapid and unpredictable changes in the world around us. Clearly, the Quality of Service needs for such delivery are much more stringent than for static data. This talk will examine the nature of dynamics of distributed data, study the suitability of the current infrastructure for disseminating time varying information, and discuss fresh approaches to maintain the temporal coherency of dynamic data and of queries over such data. We argue that executing user queries over dynamic data calls for the careful design of techniques for change dissemination, dynamic and cooperative caching, in-network filtering and aggregated query processing. We show how exploiting the characteristics of
data and the correctness requirements associated with query results leads to effective and efficient solutions that improve scalability and reduce overheads.

Speaker

Krithi Ramamritham received the Ph.D. in Computer Science from the University of Utah and then joined the University of Massachusetts. He did his B.Tech. in Electrical Engineering and M.Tech. in Computer Science, both from the Indian Institute of Technology Madras. He is currently at the Indian Institute of Technology Bombay as the Vijay and Sita Vashee Chair Professor in the Department of Computer Science and Engineering.

His areas of interest include database systems, real-time systems and internet computing. He has co-authored two IEEE tutorial texts on real-time systems, a text on advances in database transaction processing, and a text on scheduling in real-time systems. He is an Editor-in-Chief of Springer's Real-Time Systems Journal. His other editorial board contributions include IEEE Transactions on Mobile Computing, IEEE Transactions on Knowledge and Data Engineering, IEEE Transactions on Parallel and Distributed Systems, IEEE Internet Computing, the WWW Journal, the Distributed and Parallel Databases journal, and the VLDB Journal. Prof. Ramamritham is a Fellow of the IEEE, a Fellow of the ACM, and a Fellow of the Indian National Academy of Engineering. He is a recipient of the Distinguished Alumnus Award from IIT Madras.

Research Session

Reseach Session 1(XML Processing)

A Concise Labeling Scheme for XML Data
Risi Thonangi(SETLabs, Infosys Technologies Ltd.)

Statistical Analysis of Real XML Data Collections
Irena Mlynkova(MFF UK), Kamil Toman(MFF UK), Jaroslav Pokorny(MFF UK)

A Hybrid Approach for XML Similarity
Richard Chbeir(LE2I-CNRS)

Reseach Session 2(Indexing and Similarity Search)

Using Relations to Index Biological Document Repositories for Efficient Searching
Rohit Goyal(IIT Delhi), Lipika Dey(IIT Delhi)

Efficient Similarity Retrieval In Music Databases
Maria M. Ruxanda(Aalborg University), Christian S. Jensen(Aalborg University)

Supporting Approximate Similarity Queries with Quality Guarantees in P2P Systems
Qi Zhong(Microsoft), Iosif Lazaridis(UC Irvine), Mayur Deshpande(UC Irvine), Chen Li(UC Irvine), Sharad Mehrotra(UC Irvine), Hal Stern(UC Irvine)

Session 3(Potpourri)

Preserving Obliviousness Characteristic of Honeypot database
Anand Gupta(IIT Delhi), Netaji Subhas(IIT Delhi), Shyam Gupta(IIT Delhi), Renu Damor(IIT Delhi), Vikram Goyal(IIT Delhi), Sangeeta Sabharwal(IIT Delhi), Netaji Subhas(IIT Delhi)

Towards Kernel Density Estimation over Streaming Data
Christoph Heinz(University of Marburg), Bernhard Seeger(University of Marburg)

Genea: Schema-Aware Mapping of Ontologies into Relational Databases
Tim Kraska(University of Muenster), Uwe Roehm(University of Sydney)

Reseach Session 4(Web and Distributed Data)

Fault-Tolerant Queries over Sensor Data
Iosif Lazaridis(UC Irvine), Qi Han(Colorado School of Mines), Sharad Mehrotra(UC Irvine), Nalini Venkatasubramanian(UC Irvine)

EcoRep: An Economic Model for efficient dynamic replication in Mobile-P2P networks
Anirban Mondal(University of Tokyo), Sanjay Madria(University of Missouri-Rolla), Masaru Kitsuregawa(University of Tokyo)

Unsupervised Learning from URL Corpora
Deepak P(IBM India Research Lab-SIRC), Deepak Khemani(IIT Madras)

Measures of Ignorance on the Web
Siddhartha Reddy(IIIT Bangalore), Srinath Srinivasa(IIIT Bangalore), Mandar Mutalikdesai(IIIT Bangalore)

Research Poster Session

Scheduling and Caching in MultiQuery Optimization
Dilys Thomas (Stanford University, USA), A. A. Diwan (IIT Bombay, India), S. Sudarshan (IIT Bombay, India)

Efficient Algorithm for Hierarchical Online Mining of Association Rules
Kishore B. Kumar (MindTree Consulting Private Limited, India), Naresh Jotwani (DA-IICT, India)

A Scalable Replica Management Method in Peer-to-Peer Distributed Storage Systems
Jing Zhou (National U. of Defence Tech, China), Yijie Wang (National U. of Defence Tech, China), Sikun Li (National U. of Defence Tech, China)

Prefix Tree with Encryption of Data and Itemsets
Ramkishore Bhattacharyya (Jadavpur University, India)

Anomaly Detection In Labeled Data
Rohit Kelkar (Tata Research Development and Design Centre, India), Girish Palshikar (Tata Research Development and Design Centre, India)

Using Domain Ontologies for Efficient Information Retrieval
Sandhya Revuri (IIT Madras, India), Sujatha R Upadhyaya (IIT Madras, India), P Sreenivasa Kumar (IIT Madras, India)

Managing XML data with Evolving Schema
B. V. N. Prashant (IIT Madras, India), P. Sreenivasa Kumar (IIT Madras, India)

BasisGraph: Combining Storage and Structure Index for Similarity Search in Graph DBs
Mistry Harjinder Singh (IIIT Bangalore, India), Srinath Srinivasa (IIIT Bangalore, India)

Algebra-Based Optimization of XML-Extended OLAP Queries
Xuepeng Yin (Aalborg University, Denmark), Torben Bach Pedersen (Aalborg University, Denmark)

Symbiosis in the Intranet: How Document Retrieval Benefits from Database Information
Christoph Mangold (Universitaet Stuttgart, Germany), Holger Schwarz (Universitaet Stuttgart, Germany), Bernhard Mitschang (Universitaet Stuttgart, Germany)

Applications and Industrial Session

Efficient Detection of Distributed Constraint Violations
Shipra Agrawal (Stanford University, USA), Supratim Deb (Bell Labs Research, India), K. V. M. Naidu (Bell Labs Research, India), Rajeev Rastogi (Bell Labs Research, India)

On Engineering Web-based Enterprise Applications
Srinivasa Narayanan (Tavant Technologies, USA), Subbu N. Subramanian (Tavant Technologies, USA), Manish Arya (Tavant Technologies, USA)

Sponsor Session

Data Management Technologies for High-Demand Analytics
Speaker: Mr. Vinay Santurkar, Senior Manager, Business Intelligence Group (IQ), Sybase India

TBD
Speaker: From Google

TBA
Speaker: From Yahoo!

Tutorials

Tutorial 1: Privacy preserving data publication: From Generalization to Anatomy

Abstract:
Companies and organizations often need to publish clients' information to institutions for research purposes. For example, a hospital periodically releases patients' diagnostic records so that medical scientists can study the correlation between diseases and various factors. Privacy preservation is an important topic in data publication. First, the publication should be fuzzy enough to disallow any adversary to figure out the exact medical history of any patient. On the other hand, the released data must be sufficiently precise to enable effective analysis. In this tutorial, we will review the existing techniques for striking an appropriate balance, in order to maximize the accuracy of data investigation, without breaching any patient's privacy.

Speaker's Profile: Yufei Tao is an Assistant Professor in the Department of Computer Science and Engineering, the Chinese University of Hong Kong. He holds a PhD from the Hong Kong University of Science and Technology, and did his post-doc at the Carnegie Mellon University, USA. Yufei received the Hong Kong Young Scientist Award in 2002. His current research interests include privacy preserving data publication, spatial databases, and uncertain databases.

Tutorial 2: Multilingual Database Systems

Abstract:
Efficient storage and query processing of data spanning multiple natural languages are of crucial importance in today's globalized world. A primary prerequisite to achieve this goal is that the defacto standard data repositories – relational database systems – should efficiently and seamlessly support multilingual data. In this tutorial, we will first present a detailed assessment of how good today's database systems (both commercial and public-domain) are with regard to the storage, management and processing of multilingual data. Our results will show that there are significant performance inefficiencies for languages based on scripts other than Latin (such as Devanagari, Kanji, Cyrillic, etc.). We will also outline techniques for alleviating these problems.

With regard to functionality, a major limitation of SQL is that it does not support querying of data across different natural languages, that is, cross-lingual queries. To address this lacuna, we will propose two new SQL operators that support phoneme-based matching of names, and ontology-based matching of concepts, in the multilingual world.

An algebra for integrating these new operators with relational systems will be defined as well as the associated cost models, selectivity estimators, and access methods. Our experience with a prototype implementation of these operators on PostgreSQL will be highlighted. In a nutshell, this tutorial will present practical approaches towards realizing the ultimate goal of natural language-"neutral" database engines.

Speaker's Profile: Jayant Haritsa is a Professor in the Supercomputer Education & Research Centre and in the Department of Computer Science & Automation at the Indian Institute of Science, Bangalore. He received the BTech degree in Electronics and Communications Engineering from the Indian Institute of Technology (Madras),and the MS and PhD degrees in Computer Science from the University of Wisconsin (Madison). His research interests are in database systems and real-time systems. He is a member of IEEE, ACM, and the Computer Society of India, and is an associate editor of the Real-Time Systems journal and the IEEE Data Engineering Bulletin.

Tutorial 3: Secure Data Outsourcing
Abstract

The networked and increasingly ubiquitous nature of today's data management services mandates assurances to detect and deter malicious or faulty behavior. This is particularly relevant for outsourced data frameworks in which clients place data management with specialized service providers. Clients are reluctant to place sensitive data under the control of a foreign party without assurances of confidentiality. Additionally, once outsourced, privacy and data access correctness (data integrity and query completeness) become paramount.

Today's solutions are fundamentally insecure and vulnerable to illicit behavior, because they do not handle these dimensions. In this tutorial we will discuss existing solutions and future designs for robust, efficient, and scalable data outsourcing mechanisms providing strong security assurances of (1) correctness, (2) confidentiality, and (3) data access privacy.

There exists a strong relationship between such assurances; for example, the lack of access pattern privacy usually allows for statistical attacks compromising data confidentiality. Confidentiality can be achieved by data encryption. However, to be practical, outsourced data services should allow expressive client queries (e.g., relational joins with arbitrary predicates) without compromising confidentiality. This is a hard problem because decryption keys cannot be directly provided to potentially untrusted servers. Moreover, if the remote server cannot be fully trusted, protocol correctness become essential. Therefore, solutions that do not address all three dimensions are incomplete and insecure.

It is important to design query mechanisms targeting outsourced relational data that (i) ensure queries have been executed with integrity and completeness over their respective target data sets, (ii) allow queries to be executed with confidentiality over encrypted data, (iii) guarantee the privacy of client queries and data access patterns. We will discuss protocols that adapt to the existence of trusted hardware – so critical functionality can be delegated securely from clients to servers. We will exemplify with practical protocols handling binary predicate JOINs with full privacy in outsourced scenarios.

Speaker's Profile: Radu Sion is an Assistant Professor in Computer Sciences at Stony Brook University. He is a member of the Network Security and Applied Cryptography (NSAC) Lab. His research interests are in Information Assurance, Applied Cryptography and Network Security. Instances are: wireless and sensor networks security, digital rights management, secure data outsourcing, queries over encrypted data, reputation systems, integrity proofs in sensor networks, secure storage in peer to peer and ad-hoc environments, data privacy and bounds on illicit inference over multiple data sources, security and policy management in computation/data grids.

Tutorial 4: High Performance Data Mining: Consolidation and Renewed Bearing
Abstract

Over the years the definition of high performance computing has taken on various forms as a function of the types of technical and creative uses and the underlying semantics of applications driving them. Traditional definitions often refer to the problem of using high end parallel computers to meet the need of applications. However in the modern context, high performance computing ranges from fast sequential algorithms that target memory and I/O performance on modern processors all the way to work on the computational and data grid.

In this tutorial, I will describe the impact of high performance computing, under this broad definition, on the field of data mining. I will review and consolidate some of the key algorithmic developments over the last decade as they pertain to high performance data mining. Specifically we will examine parallel and sequential performance of such algorithms on modern high performance systems. In the latter part of this tutorial, I will present an outlook towards the future

paying particular attention to recent technological advances (e.g. multi-core architectures, InfiniBand networks etc.) that I expect will have a bearing on this field of research. I will conclude by describing, why in light of these advances I expect that existing data mining algorithms will need to be re-architected and improved to realize performance commensurate with these technological advances. I also will describe why I believe they will lead to a renewed set of challenges for algorithm development and associated systems support spanning areas such as compilers, runtime, database, middleware and hardware systems.

Speaker's Profile: Srinivasan Parthasarathy Srinivasan Parthasarathy is currently an Associate professor at the Computer Science and Engineering Department at the Ohio State University (OSU). He heads up the data mining research laboratory and has a joint appointment in the Department of Biomedical Informatics at Ohio State. He received a B.E in Electrical Engineering from the University of Roorkee (now IIT-Roorkee) and an MS and PhD in Computer Science from the University of Rochester. His research interests include data mining, high performance computing & systems, scientific data analysis and bioinformatics. He is a recipient of the US National Science Foundation's CAREER award, the US Department of Energy's Early Career Principal Investigator Award, and an SBC/Ameritech Faculty fellowship. His work has received several awards including an IEEE Data Mining 2002 best paper, a SIAM Data Mining 2003 best paper, the VLDB 2005 best research paper and a "Best of SDM05" selection from SIAM Data Mining 2005.

Tutorial 5: Scalable Information Extraction and Integration
Abstract
Many applications over text require efficient methods for extracting and integrating structured data from large unstructured sources. This tutorial reviews the state of the art approaches for information extraction and duplicate elimination. First, we present an overview of the methods used, with particular emphasis on machine learning based approaches including sequential models like Conditional Random Fields and their generalizations. Second, we present scalable techniques for deploying these models on large unstructured text collections. We review key approaches for scaling up information extraction, including using general purpose search engines as well as indexing techniques specialized for information extraction applications. We also overview scalable techniques for integrating the extracted information using approximate join algorithms and fuzzy index lookups. We highlight research opportunities and challenges that remain.

Speaker's Profile: Sunita Sarawagi researches in the fields of databases, data mining, machine learning and statistics. She is associate professor at IIT Bombay. Prior to that she was a Research Staff Member at IBM Almaden Research Center. She got her PhD in databases from the University of California at Berkeley and a bachelors degree from IIT Kharagpur. She has several publications in databases and data mining including a best paper award at the 1998 ACM SIGMOD conference and several patents. She is on the editorial board of the ACM TODS and ACM KDD journals and was editor-in-chief of the ACM SIGKDD newsletter. She has served as program committee member for ACM SIGMOD, VLDB, ACM SIGKDD and IEEE ICDE, ICML conferences.

Tutorial 6: Data Grid Management
Abstract
If we analyze the history of computer science, most of the contributions including the DBMS were the result of a requirement that was pushing the limits of an existing technology. One such requirement today is to manage very large unstructured data that is distributed in multiple countries using traditional file systems. Some of the FORTUNE 500 companies face this problem today, due to out-sourcing and distributed global teams that collaborate with each other. Data Grid Management Systems (DGMS) manage collaborative global sharing of very large amounts of unstructured data amongst multiple teams. The core concepts of a DGMS are very similar to traditional RDBMS. A DGMS could be considered as a logical namespace (or a logical distributed

file system) of heterogeneous data storage resources from multiple sub-organizations. DGMS are powered by relational databases and provide both system and user defined schema to organize and query data. In this tutorial, we introduce DGMS concepts and explain with real use cases why such a system is needed in very large academic data centers and major companies. Novices and experts in distributed data management will have a chance to learn about this emerging technology, research problems and business opportunities.

Speaker's Profile: Arun Jagatheesan ("Arun") is a Dataflow/Data grid Specialist at the San Diego Supercomputer Center (SDSC) in University of California, San Diego. His research interests include Data Grid Management Systems (DGMS), peer-to-peer data management, and workflow management systems. Arun works on research, development and standardization of data grid technologies by collaborating with multiple academic and commercial organizations, as part of the SDSC Storage Resource Broker (SRB) Project. He is the founder and technical lead of the SRB Matrix Project on Gridflow Management Systems. He is currently involved in many data grid projects at SDSC including the new LUSciD collaboration, a joint effort by the University of California and Lawrence Livermore National Laboratory (LLNL) exploring the software requirements for managing very large amount of data. Arun also plays an active role in the LSST project that will manage hundreds of petabytes of data. Arun was previously an OPS faculty member at the University of Florida. He has published many papers and provided multiple invited talks or tutorials on data grids at multiple technical conferences.

Demonstration Session

Visual Clue Based Extraction of Web Data from Flat and Nested Data Records
Siddu P Algur (SDM College of Engg, Dharwad, India), P S Hiremath (Gulbarga University, India)

A Query Interface for Ubiquitous Access to Database Resources
Subhash Bhalla (University of Aizu, Japan), Masaki Hasegawa (University of Aizu, Japan)

UnURL: Unsupervised Learning from URLs
Deepak P (IBM IRL, India), Deepak Khemani (IIT Madras, India)

OCHD: Preserving Obliviousness Characteristic of Honeypot Database
S. K. Gupta (IIT Delhi, India), Renu Damor (IIT Delhi, India), Anand Gupta (NSIT, Delhi, India), Vikram Goyal (IIT Delhi, India)

Stream Mining via Density Estimators: A Concrete Application
Christoph Heinz (University of Marburg, Germany), Bernhard Seeger (University of Marburg, Germany)

Documents meet Databases: A System for Intranet Search
Christoph Mangold (University of Stuttgart, Germany), Holger Schwarz (University of Stuttgart, Germany)

Discovery Services-Enabling RFID Traceability in EPCglobal Networks
Steve Beier (IBM SVL, USA), Tyrone Grandison (IBM Almaden, USA), Karin Kailing (IBM Almaden, USA), Ralf Rantzau (IBM SVL, USA)

Semantic and Structure Based XML Similarity: The XS3 Prototype
Joe Tekli (University of Bourgogne, France), Richard Chbeir (University of Bourgogne, France), Kokou Yetongnon (University of Bourgogne, France)

The OLAP-XML Federation System
Xuepeng Yin (Aalborg University, Denmark), Torben Bach Pedersen (Aalborg University, Denmark)

BlogHarvest: Blog Mining and Search Framework
Mukul Joshi (Great Software Lab, India), Nikhil Belsare (Tech Mahindra Ltd., India)