

Proceedings

Seventeenth International Conference on

DATA ENGINEERING

2-6 April 2001
Heidelberg, Germany

Sponsored by
IEEE Computer Society Technical Committee on Data Engineering
EML, IBM, Hewlett-Packard, SAS, Microsoft, ABB, Software AG, sd&m


IEEE
COMPUTER
SOCIETY



Proceedings

17th International Conference on Data Engineering

2-6 April 2001

Heidelberg, Germany

Sponsored by

IEEE Computer Society Technical Committee on Data Engineering (TCDE)

ABB



invent

IBM

Microsoft

SOFTWARE AG
THE XML COMPANY

sas
e-Intelligence

s | d | m
software | design & | management

IEEE 
**COMPUTER
SOCIETY**



Los Alamitos, California

Washington □ Brussels □ Tokyo

Copyright © 2001 by The Institute of Electrical and Electronics Engineers, Inc.
All rights reserved

Copyright and Reprint Permissions: Abstracting is permitted with credit to the source. Libraries may photocopy beyond the limits of US copyright law, for private use of patrons, those articles in this volume that carry a code at the bottom of the first page, provided that the per-copy fee indicated in the code is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923.

Other copying, reprint, or republication requests should be addressed to: IEEE Copyrights Manager, IEEE Service Center, 445 Hoes Lane, P.O. Box 133, Piscataway, NJ 08855-1331.

The papers in this book comprise the proceedings of the meeting mentioned on the cover and title page. They reflect the authors' opinions and, in the interests of timely dissemination, are published as presented and without change. Their inclusion in this publication does not necessarily constitute endorsement by the editors, the IEEE Computer Society, or the Institute of Electrical and Electronics Engineers, Inc.

IEEE Computer Society Order Number PR01001

ISBN 0-7695-1001-9

ISBN 0-7695-1002-7 (case)

ISBN 0-7695-1003-5 (microfiche)

ISSN 1063-6382

Additional copies may be ordered from:

IEEE Computer Society
Customer Service Center
10662 Los Vaqueros Circle
P.O. Box 3014
Los Alamitos, CA 90720-1314
Tel: + 1 714 821 8380
Fax: + 1 714 821 4641
<http://computer.org/>
csbooks@computer.org

IEEE Service Center
445 Hoes Lane
P.O. Box 1331
Piscataway, NJ 08855-1331
Tel: + 1 732 981 0060
Fax: + 1 732 981 9667
<http://shop.ieee.org/store/>
customer-service@ieee.org

IEEE Computer Society
Asia/Pacific Office
Watanabe Bldg., 1-4-2
Minami-Aoyama
Minato-ku, Tokyo 107-0062
JAPAN
Tel: + 81 3 3408 3118
Fax: + 81 3 3408 3553
tokyo.ofc@computer.org

Editorial production by Danielle C. Young

Cover art production by Joe Daigle/Studio Productions

Printed in the United States of America by The Printing House

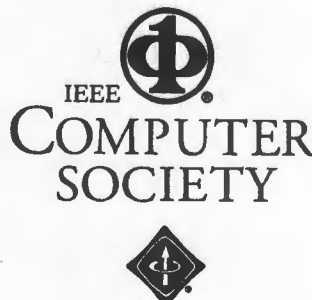


Table of Contents

17th International Conference on Data Engineering (ICDE 2001)

Message from the Program Co-Chairs	xii
Conference Officers	xiii
Program Committee	xiv
Tutorials Program	xvii
Demos Program	xx
External Referees	xxi

Session 1: Middleware

Exactly-Once Semantics in a Replicated Messaging System	3
<i>Y. Huang and H. Garcia-Molina</i>	
CORBA Notification Service: Design Challenges and Scalable Solutions	13
<i>R. Gruber, B. Krishnamurthy, and E. Panagos</i>	
Measuring and Optimizing a System for Persistent Database Sessions	21
<i>R. Barga and D. Lomet</i>	

Session 2: Temporal Databases

A Temporal Algebra for an ER-Based Temporal Data Model	33
<i>J. Lee and R. Elmasri</i>	
A Split Operator for Now-Relative Bitemporal Databases	41
<i>M. Agesen, M. Böhlen, L. Poulsen, and K. Torp</i>	
Incremental Computation and Maintenance of Temporal Aggregates	51
<i>J. Yang and J. Widom</i>	

Industry Session 1: E-Commerce

The Importance of Extensible Database Systems for e-Commerce	63
<i>S. DeFazio, R. Krishnan, J. Srinivasan, and S. Zeldin</i>	
E-Business Applications for Supply Chain Management: Challenges and Solutions	71
<i>F. Casati, U. Dayal, and M.-C. Shan</i>	
Delivering E-Business Faster with Database Technology	
<i>P. Shum</i>	

Session 3: Integration and Query Processing in Distributed Environments

Model-Based Mediation with Domain Maps.....	81
<i>B. Ludäscher, A. Gupta, and M. Martone</i>	
Processing Queries with Expensive Functions and Large Objects in Distributed Mediator Systems	91
<i>L. Bouganim, F. Fabret, F. Porto, and P. Valduriez</i>	
Tuning an SQL-Based PDM System in a Worldwide Client/Server Environment	99
<i>E. Müller, P. Dadam, J. Enderle, and M. Feltes</i>	

Session 4: Issues in Information Processing

Bundles in Captivity: An Application of Superimposed Information	111
<i>L. Delcambre, D. Maier, S. Bowers, M. Weaver, L. Deng, P. Gorman, J. Ash, M. Lavelle, and J. Lyman</i>	
High-Level Parallelisation in a Database Cluster: A Feasibility Study Using Document Services.....	121
<i>T. Grabs, K. Böhm, and H.-J. Schek</i>	
Efficient Sequenced Temporal Integrity Checking.....	131
<i>W. Li, R. Snodgrass, S. Deng, V. Gattu, and A. Kasthurirangan</i>	

Industry Session 2: XML

XML Data and Object Databases: The Perfect Couple?	143
<i>A. Renner</i>	
Tamino — A DBMS Designed for XML	149
<i>H. Schöning</i>	
The Nimble XML Data Integration System	155
<i>D. Draper, A. HaLevy, and D. Weld</i>	

Session 5: Database Engines

High-Performance, Space-Efficient, Automated Object Locking.....	163
<i>L. Daynès and G. Czajkowski</i>	
Differential Logging: A Commutative and Associative Logging Scheme for Highly Parallel Main Memory Database	173
<i>J. Lee, K. Kim, and S. Cha</i>	
Efficient Bulk Deletes in Relational Databases	183
<i>A. Gärtner, A. Kemper, D. Kossmann, and B. Zeller</i>	

Session 6: Data Mining from Patterns

On Dual Mining: From Patterns to Circumstances, and Back	195
<i>G. Grahne, L. Lakshmanan, X. Wang, and M. Xie</i>	
Mining Partially Periodic Event Patterns with Unknown Periods.....	205
<i>S. Ma and J. Hellerstein</i>	
PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth	215
<i>J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M.-C. Hsu</i>	

Industry Session 3: Mobility and Pervasive Computing

Mobile Data Management: Challenges of Wireless and Offline Data Access	227
<i>E. Giguère</i>	
Microsoft Server Technology for Mobile and Wireless Applications	229
<i>P. Seshadri</i>	
IBM DB2 Everyplace: A Small Footprint Relational Database System	230
<i>J. Karlsson, A. Lal, C. Leung, and T. Pham</i>	

Session 7: Workflow Management

Dependable Computing in Virtual Laboratories	235
<i>G. Alonso, W. Bausch, C. Pautasso, A. Kahn, and M. Hallett</i>	
Workflow and Process Synchronization with Interaction Expressions and Graphs	243
<i>C. Heinlein</i>	
Inter-Enterprise Collaborative Business Process Management	253
<i>Q. Chen and M. Hsu</i>	

Session 8: Time Series

Duality-Based Subsequence Matching in Time-Series Databases.....	263
<i>Y.-S. Moon, K.-Y. Whang, and W.-K. Loh</i>	
Variable Length Queries for Time Series Data	273
<i>T. Kahveci and A. Singh</i>	
TAR: Temporal Association Rules on Evolving Numerical Attributes.....	283
<i>W. Wang, J. Yang, and R. Muntz</i>	

Industry Session 4: E-Commerce/Internet Applications

Database Performance for Next Generation Telecommunications	295
<i>M. Cochinwala</i>	
SpinCircuit: A Collaborative Portal Powered by E-Speak	661
<i>R. Pathak</i>	
Data Management Support of Web Applications	299
<i>D. Fishman</i>	

Session 9: XML

An Automated Change-Detection Algorithm for HTML Documents Based on Semantic Hierarchies	303
<i>S.-J. Lim and Y.-K. Ng</i>	
An XML Indexing Structure with Relative Region Coordinate	313
<i>D. Kha, M. Yoshikawa, and S. Uemura</i>	
Querying XML Documents Made Easy: Nearest Concept Queries	321
<i>A. Schmidt, M. Kersten, and M. Windhouwer</i>	
A Graph-Based Approach for Extracting Terminological Properties of Elements of XML Documents	330
<i>L. Palopoli, G. Terracina, and D. Ursino</i>	

Session 10: Indexing and Clustering

B ⁺ -Tree Indexes with Hybrid Row Identifiers in Oracle8i	341
<i>E. Chong, S. Das, C. Freiwald, J. Srinivasan, A. Yalamanchi, M. Jagannath, A.-T. Tran, and R. Krishnan</i>	
B-Tree Indexes and CPU Caches	349
<i>G. Graefe and P.-Å. Larson</i>	
Spatial Clustering in the Presence of Obstacles	359
<i>A. Tung, J. Hou, and J. Han</i>	
Selectivity Estimation for Spatial Joins	368
<i>N. An, Z.-Y. Yang, and A. Sivasubramaniam</i>	

Industry Session 5: Data Warehousing and Data Mining

Integrating Data Mining with SQL Databases: OLE DB for Data Mining	379
<i>A. Netz, S. Chaudhuri, U. Fayyad, and J. Bernhardt</i>	

SAP Business Information Warehouse — From Data Warehousing to an E-Business Platform	388
<i>T. Zurek and K. Kreplin</i>	
fAST Refresh Using Mass Query Optimization	391
<i>W. Lehner, B. Cochrane, H. Pirahesh, and M. Zaharioudakis</i>	
Session 11: Similarity Processing and Special Operators	
Quality-Aware and Load-Sensitive Planning of Image Similarity Queries	401
<i>K. Böhm, M. Mlivoncic, and R. Weber</i>	
A Cost Model and Index Architecture for the Similarity Join	411
<i>C. Böhm and H.-P. Kriegel</i>	
The Skyline Operator	421
<i>S. Börzsönyi, D. Kossmann, and K. Stocker</i>	
Session 12: Data Mining/Frequent Item Sets	
Mining Frequent Itemsets with Convertible Constraints	433
<i>J. Pei, J. Han, and L. Lakshmanan</i>	
MAFIA: A Maximal Frequent Itemset Algorithm for Transactional Databases	443
<i>D. Burdick, M. Calimlim, and J. Gehrke</i>	
An Efficient Approximation Scheme for Data Mining Tasks	453
<i>G. Kollios, D. Gunopoulos, N. Koudas, and S. Berchtold</i>	
Industry Session 6: Applications	
Bringing the Internet to Your Database: Using SQL Server 2000 and XML to Build Loosely-Coupled Systems	465
<i>M. Rys</i>	
Infrastructure for Web-Based Application Integration	473
<i>D. Gawlick</i>	
Developing Web Services	477
<i>A. Bosworth</i>	
Session 13: Nearest Neighbor Queries	
An Index Structure for Efficient Reverse Nearest Neighbor Queries	485
<i>C. Yang and K.-I. Lin</i>	

Distinctiveness-Sensitive Nearest-Neighbor Search for Efficient Similarity Retrieval of Multimedia Information	493
<i>N. Katayama and S. Satoh</i>	

Approximate Nearest Neighbor Searching in Multimedia Databases	503
<i>H. Ferhatosmanoglu, E. Tuncel, D. Agrawal, and A. El Abbadi</i>	

Session 14: OLAP

Rewriting OLAP Queries Using Materialized Views and Dimension Hierarchies in Data Warehouses	515
<i>C.-S. Park, M. Kim, and Y.-J. Lee</i>	

The MD-Join: An Operator for Complex OLAP	524
<i>D. Chatziantoniou, M. Akinde, T. Johnson, and S. Kim</i>	

Overcoming Limitations of Sampling for Aggregation Queries.....	534
<i>S. Chaudhuri, G. Das, M. Datar, R. Motwani, and V. Narasayya</i>	

Industry Session 7: DBMS Industrial Issues

Pseudo Column Level Locking.....	545
<i>N. Ponnekanti</i>	

Discovery and Application of Check Constraints in DB2.....	551
<i>J. Gryz, B. Schiefer, J. Zheng, and C. Zuzarte</i>	

Database Managed External File Update	557
<i>N. Mittal and H.-I. Hsiao</i>	

Session 15: Query Processing

Block Oriented Processing of Relational Database Operations in Modern Computer Architectures	567
<i>S. Padmanabhan, T. Malkemus, R. Agarwal, and A. Jhingran</i>	

Integrating Semi-Join-Reducers into State-of-the-Art Query Processors.....	575
<i>K. Stocker, D. Kossmann, R. Braumandl, and A. Kemper</i>	

Using EELs: A Practical Approach to Outerjoin and Antijoin Reordering.....	585
<i>J. Rao, B. Lindsay, G. Lohman, H. Pirahesh, and D. Simmen</i>	

Counting Twig Matches in a Tree	595
<i>Z. Chen, H. Jagadish, F. Korn, N. Koudas, S. Muthukrishnan, R. Ng, and D. Srivastava</i>	

Session 16: Similarity Searches

An Index-Based Approach for Similarity Search Supporting Time Warping in Large Sequence Databases.....	607
<i>S.-W. Kim, S. Park, and W. Chu</i>	
High Dimensional Similarity Search with Space Filling Curves	615
<i>S. Liao, M. Lopez, and S. Leutenegger</i>	
Similarity Search without Tears: The OMNI-Family of All-Purpose Access Methods	623
<i>R. Filho, A. Traina, C. Traina, Jr., and C. Faloutsos</i>	

Session 17: Caching

Cache-On-Demand: Recycling with Certainty.....	633
<i>K.-L. Tan, S.-T. Goh, and B. Ooi</i>	
Cache-Aware Query Routing in a Cluster of Databases	641
<i>U. Röhm, K. Böhm, and H.-J. Schek</i>	
Prefetching Based on the Type-Level Access Pattern in Object-Relational DBMSs.....	651
<i>W.-S. Han, K.-Y. Whang, Y.-S. Moon, and I.-Y. Song</i>	
Author Index	665

Message from the Program Co-Chairs

These proceedings contain the research and industrial papers presented at the Seventeenth International Conference on Data Engineering held in Heidelberg, Germany. Keeping with the tradition of ICDE, this years' conference emphasizes the engineering aspects of data management but also considers data engineering in a wider sense that encompasses middleware, distributed systems and workflow. The areas in which papers were requested are: XML, metadata and semistructured data; database engines and engineering; query processing; data warehouses, data mining and knowledge discovery; advanced information systems middleware; scientific and engineering databases; extreme databases; e-commerce and e-services; workflow and process-oriented systems; emerging trends; and system applications and experience.

A total of 296 research papers were submitted to the conference from 34 countries. After a thorough review process the program committee accepted 54 research papers organized in 17 research sessions. 7 industrial sessions make up the industrial track focusing on e-commerce, data warehousing, and mobility. The plenary panel expands on the theme of mobility and will explore the issues involved in managing billions of devices. The program is highlighted by 3 invited speakers who address key application domains of data engineering: Stuart Feldman (IBM) addresses e-business, Peter Zencke (SAP) discusses data engineering issues in standard business software while Gerhard Barth (Dresdner Bank) presents the new challenges in the financial sector.

Special thanks are due to Peter Lockemann and Tamer Özsu for organizing the industrial track. The program is flanked by interesting tutorials organized by Guido Moerkotte and Eric Simon, a variety of demos of interesting research prototypes organized by Andreas Eberhart and site-visits with presentations at some of the leading software development centers of the Rhine-Main-Neckar region.

The quality of the technical program depends on the high number of superior papers submitted and the effort of the reviewers. Therefore, we would like to thank all the authors who submitted their work and the members of the program committee, the vice-chairs and the many referees who invested so much time and effort in assembling an interesting program. Their names are listed separately.

Finally we would like to express our appreciation to Andreas Reuter and David Lomet, the general chairs who have driven the organization of ICDE 2001. The local arrangements and countless other tasks were tirelessly handled by Isabel Rojas.

We hope that the conference will be stimulating and enjoyable to all who attend it, and that the papers in these proceedings will be a source of relevant information for the data engineering community.

Dimitrios Georgakopoulos

Alex Buchmann



Conference Officers

General Chairs

Andreas Reuter, *EML and International University, Germany*

David Lomet, *Microsoft Research, USA*

Program Co-Chairs

Alex Buchmann, *Darmstadt University, Germany*

Dimitrios Georgakopoulos, *Telcordia Technologies, USA*

Panel Program Chair

Erich Neuhold, *GMD-IPSI, Germany*

Tutorial Program Chair

Guido Moerkotte, *Mannheim University, Germany*

Eric Simon, *INRIA, France*

Industrial Program Co-Chairs

Peter Lockemann, *Karlsruhe University, Germany*

Tamer Özsu, *University of Waterloo, Canada*

Steering Committee Liaison

Erich Neuhold, *GMD-IPSI, Germany*

Marek Rusinkiewicz, *MCC, USA*

Organizing Chair

Isabel Rojas, *European Media Laboratory, Germany*

Demos & Exhibits

Wolfgang Becker, *International University, Germany*

Andreas Eberhart, *International University, Germany*

Program Committee

Program Committee Vice-Chairs

XML, METADATA, and SEMISTRUCTURED DATA

Dan Suciu, *University of Washington, USA*

DATABASE ENGINES & ENGINEERING

Bruce Lindsay, *IBM Almaden, USA*

QUERY PROCESSING

Joseph Hellerstein, *University of California at Berkeley, USA*

DATA WAREHOUSES, DATA MINING, AND KNOWLEDGE DISCOVERY

Rakesh Agrawal, *IBM Almaden, USA*

ADVANCED IS MIDDLEWARE

Kriithi Ramamritham, *University of Massachusetts, USA and IIT Bombay, India*

SCIENTIFIC AND ENGINEERING DATABASES

Theo Haerder, *University of Kaiserslautern, Germany*

EXTREME DATABASES

Martin Kersten, *CWI, Netherlands*

E-COMMERCE and E-SERVICES

Umesh Dayal, *Hewlett-Packard Laboratories, USA*

WORKFLOW and PROCESS-ORIENTED SYSTEMS

Gustavo Alonso, *ETH Zentrum, Switzerland*

SYSTEM APPLICATIONS AND EXPERIENCE

José Blakeley, *Microsoft, USA*

Program Committee Members

Amr El Abbadi, *University of California at Santa Barbara, USA*

Karl Aberer, *EPFL-DSC, Switzerland*

Brad Adelberg, *Northwestern University, USA*

Rafael Alonso, *Sarnoff, USA*

Paolo Atzeni, *University of Rome, Italy*

Jean Bacon, *University of Cambridge Computer Laboratory, UK*

Ricardo Baeza-Yates, *University of Chile, Chile*

Daniel Barbará, *George Mason University, USA*

Claudia Bauzer Medeiros, *UNICAMP, Brazil*

Philip Bernstein, *Microsoft, USA*
Elisa Bertino, *University of Milano, Italy*
Wojciech Cellary, *Poznan University of Economics, Poland*
Sharma Chakravarthy, *University of Texas at Arlington, USA*
Surajit Chaudhuri, *Microsoft Research, USA*
Panos Chrysanthis, *University of Pittsburgh, USA*
Isabel Cruz, *Worcester Polytechnic Institute, USA*
Peter Dadam, *Ulm University, Germany*
Anindya Datta, *Georgia Institute of Technology, USA*
Oscar Diaz, *University of the Basque Country, Spain*
Asuman Dogac, *Middle East Technical University, Turkey*
Pamela Drew, *Boeing, USA*
Ahmed Elmagarmid, *Purdue University, USA*
Opher Etzion, *IBM, Israel*
Mary Fernandez, *AT&T Labs - Research, USA*
Daniela Florescu, *INRIA, France*
Juliana Freire, *Bell Laboratories - Lucent, USA*
Johannes Gehrke, *Cornell University, USA*
Claude Godart, *LORIA, France*
Luis Gravano, *Columbia University, USA*
Paul Grefen, *University of Twente, Netherlands*
Ashish Gupta, *Amazon.com, USA*
Howard Ho, *IBM Almaden, USA*
Matthias Jarke, *RWTH Aachen, Germany*
Leonid Kalinichenko, *Russian Academy of Science, Russia*
Yahiko Kambayashi, *Kyoto University, Japan*
George Karabatis, *Telcordia, USA*
Alfons Kemper, *Passau University, Germany*
Masaru Kitsuregawa, *University of Tokyo, Japan*
Hank Korth, *Lucent, USA*
Paul Larson, *Microsoft, USA*
Chiang Lee, *National Cheng-Kung University, Taiwan*
Frank Leymann, *IBM, Germany*
Josephine Micallef, *Telcordia, USA*
Bernhard Mitschang, *Stuttgart University, Germany*
Lory Molesky, *Oracle, USA*
Marian Nodine, *MCC, USA*
Maria Orłowska, *University of Queensland, Australia*
Euthimios Panagos, *AT&T Labs-Research, USA*
Michael Papazoglou, *Tilburg University, Netherlands*
Norman Paton, *Manchester University, UK*
Jaroslav Pokorný, *Charles University, Czech Republic*
Calton Pu, *Georgia Tech, USA*

Tore Risch, *Uppsala University, Sweden*
Marek Rusinkiewicz, *MCC, USA*
George Samaras, *University of Cyprus, Cyprus*
Sunita Sarawagi, *IIT Bombay, India*
Peter Scheuermann, *Northwestern University, USA*
Hans Schuster, *MCC, USA*
Timos Sellis, *National Technical University of Athens, Greece*
Ming-Chein Shan, *Hewlett-Packard Laboratories, USA*
Wang Shan, *Renmin University of China, China*
Jerome Simeon, *Bell Laboratories - Lucent, USA*
Kazimierz Subieta, *Institute of Computer Science PAS, Poland*
S. Sudarshan, *IIT Bombay, India*
Aphrodite Tsalgatidou, *University of Athens, Greece*
Vassilis Tsotras, *University of California at Riverside, USA*
Ozgur Ulusoy, *Bilkent University, Turkey*
Susan Urban, *Arizona State University, USA*
Jari Veijalainen, *University of Jyväskylä, Finland*
Viacheslav Wolfengagen, *Institute for Contemporary Education, Russia*
Antoni Wolski, *VTT, Finland*

Industrial Program Committee Members

Anthony Tomasic, *Digital Integrity Inc.*
Swarup Acharya, *Bell Labs*
Rudolf Munz, *SAP AG*
Michael Carey, *Propel*
Harald Schoening, *Software AG*
Jim Gray, *Microsoft Corp*
Paul Larson, *Microsoft Corp*
Alex Biliris, *AT&T Research*
Jiawei Han, *Simon Fraser University*
Glen Paulley, *iAnywhere Solutions, A Sybase Company*
Rüdiger Wirth, *Daimler-Chrysler*
Klaus Kuespert, *Friedrich-Schiller University, Jena*
Atul Adya, *Microsoft Corp.*
Kaladhar Voruganti, *IBM Almaden*



Tutorials Program

Tutorial 1: Wavelets and Their Applications in Databases

Instructors: Daniel A. Keim and Martin Heczeko

The roots of wavelet theory reach back to the end of the 19th century. The so-called, developed in 1909 by A. Haar, still serves as the foundation of modern wavelet theory. It took a long time, however, until the wavelet-based hierarchical data decomposition found its widespread application in computer science. Wavelets are seen as the “(re)discovery of the last decade” in Computer Graphics and, in the meantime, they are used in a wide variety of applications including a number of diverse database applications. Examples are: similarity search, data compression, dimensionality reduction, time series analysis, and data clustering. The wavelet theory is well founded and of very high practical impact. The large number of advantages include the strict hierarchical and multiresolutional nature of the wavelet decomposition, the linear time and space complexity of the wavelet transformations, and the high flexibility of different wavelet functions, leading to considerably more effective and efficient solutions of well-known problems. The goal of the tutorial is to make the valuable knowledge about wavelets available to a broader portion of the database research community in order to increase the benefits, which can be gained from using wavelets. The tutorial gives an overview of recent database research projects, which already benefit from the advantages of wavelets. Among the numerous successful applications are: approximation and clustering techniques for large databases, similarity search in image and time series databases, and even standard database applications such as selectivity estimation. The tutorial is structured as follows: After a brief motivation of wavelets, we provide an application-oriented overview of the foundations of wavelet theory and discuss their general advantages. Next, we provide a brief overview of some interesting standard applications of wavelets. In the main portion of the tutorial, we then focus on the recent applications of wavelets in the database area, providing a detailed description and discussion of their main contributions. In concluding the tutorial, we discuss the impact of wavelets for the database area and outline potential future research directions and applications.

Tutorial 2: Similarity Join

Instructor: Christian Boehm

Larger and larger amounts of data are collected and stored in databases, increasing the need for efficient and effective analysis methods to make use of the information contained implicitly in the data. Innumerable approaches for the various data mining tasks such as association rule discovery, classification, clustering, regression, and outlier detection have been proposed from different research communities like statistics and machine learning. An important aspect of contributions from the database research is the scalability of algorithms when facing large data sets. The relational join is one of the most important and most powerful operators of a commercial database system. Both database vendors as well as academic researchers have made every possible effort to implement the join efficiently. Even the whole area of relational query optimization deals primarily with different aspects of joins such as optimizing the join order or selecting the optimal algorithm and parametrization for each join. Recently, it has been recognized that join operations are also a powerful database primitive to support data mining algorithms. Joins do not only provide an easy and universal means to tackle the scalability problem. Moreover, using highly optimized join operations can even accelerate existing mining algorithms by large factors. Of particular interest are mining algorithms, which are based on the notion of the point density. Examples include various clustering algorithms, outlier detection, time series analysis, spatial trend detection, etc. Such algorithms typically issue a large number of similarity queries (i.e. range queries or (k-) nearest neighbor queries) in a multidimensional or metric feature space. Since many queries can be executed simultaneously, the query set can be rewritten as a similarity join between the set of the original query points and the set of the database points. Some data mining algorithms even evaluate a similarity query for each database point. Substituting this massive query set by a single similarity self-join offers a particularly high optimization potential. Due to the high relevance of the similarity join, a large number of different algorithms have been proposed. Our tutorial reviews the state-of-the-art in this area of research. The structure of our tutorial is guided by the intention to bring together the experts of data mining and query processing. First, we will introduce several representative data mining algorithms and show how to rewrite them on top of a similarity join. Starting from this, we will categorize the different types of similarity joins such as distance range joins, k-nearest neighbor joins, etc. The major part of the tutorial is then dedicated to the various algorithms for evaluating the similarity join. Next, we will go into the details of cost modeling and parameter optimization. A perspective on future research directions will conclude the tutorial.

Tutorial 3: Data Warehouse Design

Instructors: Stefano Rizzi and Matteo Golfarelli

Building a *data warehouse* (DW) for an enterprise is a huge and complex task, which requires accurate planning aimed at devising satisfactory answers to organizational and architectural questions. Despite the pushing demand for working solutions coming from enterprises and the wide offer of advanced technologies from producers, few attempts toward devising a specific, structured methodology for data warehouse design have been made. On the other hand, the statistic reports related to DW project failures state that a major cause lies in the absence of a global view of the design process: in other terms, in the absence of a design methodology. The tutorial aims at introducing a methodological framework for design, addressing the main topics in conceptual, logical and physical design of the data marts, which, assembled in a bottom-up fashion, concur in creating the data warehouse. Among the conceptual models proposed in the literature, we will focus in particular on the Dimensional Fact Model (DFM) as a support for the whole design process.

Outline:

The tutorial aims at enabling the participants to understand the basics in data warehousing and the underlying design principles, and more specifically to introduce them to the most critical issues in conceptual, logical and physical design.

This will be achieved by dealing with the following topics:

1. Introduction to Data Warehousing: from operational databases to data warehouses; the multidimensional model; architectural issues; ROLAP and MOLAP solutions.
 2. Conceptual design of Data Warehouses: E/R-based models; the Dimensional Fact Model; conceptual design from the operational schemes.
 3. Workload-based logical design for ROLAP: defining the workload; star and snowflake schemes; view materialization and fragmentation.
 4. Indices for physical design: B-trees, bitmap indices, join indices; selecting the indices for the data mart.
- In order to increase the educational efficacy, topic 2 will be supported by a CASE tool designed by the authors.

Target audience and background: The tutorial is directed to enterprise analysts and designers, as well as to researchers wishing to get acquainted with data warehousing from the designer's point of view. A good background on the relational model and on the Entity/Relationship model is required.

Tutorial 4: Next Generation of Data Mining Tools, Using SDV and Fractals

Instructor: Christos Faloutsos

What patterns can we find in a bursty Web traffic? On the Web graph itself? How about the distributions of galaxies in the sky, or the distribution of a company's customers in geographical space? How long should we expect a nearest-neighbor search to take, when there are 100 attributes per patient or customer record? The traditional assumptions (uniformity, independence, Poisson arrivals, Gaussian distributions), often fail miserably. Should we give up trying to find patterns in such settings? This tutorial focuses on two powerful but less known tools, namely on the Singular Value Decomposition (SVD) and on Fractals. SVD is a provably optimal method for dimensionality reduction and feature selection; it is the engine-under-the hood for breakthrough concepts like the Latent Semantic Indexing (LSI), the Karhunen-Loeve transform and the Kleinberg algorithm for Web site importance ranking, to name a few. Fractals, self-similarity and power laws are extremely successful in describing real datasets (coast-lines, rivers basins, stock-prices, brain-surfaces, Web and disk traffic, to name a few). Although both tools are impressively general and useful, their introductory papers are typically not tailored toward a database audience, rendering them inaccessible. This tutorial exactly tries to remedy the situation. Specifically, it has two goals: (a) to introduce the most useful concepts from SVD and Fractals, emphasizing the intuition behind them, and avoiding the unnecessary mathematical intricacies and (b) to illustrate the usefulness of SVD and fractals for a variety of database and data mining applications.

Target Audience: Researchers working on spatial access methods, on query optimization, and on data mining.

Prerequisites: None.

Benefits to Participants: The participants will gain the intuition behind these powerful tools, and they will get exposed to numerous settings where SVD and fractals solved the data mining/data base problem at hand.

Tutorial 5: XML

Instructors: Dana Florescu and Jerome Simeon

XML is a document mark-up language designed for data exchange between Web applications. Developed and promoted by the World Wide Web Consortium (W3C), XML technology has attracted a lot of attention over the last two years, both from the industry and from the research community. But if the XML world was at first limited to a unique, simple, self-contained specification (XML 1.0), the sudden interest for XML has generated an incredible amount of activity. Now, with a multitude of inter-related standards, industry proposals, and research literature, finding its way in the XML maze has become a challenging enterprise. The objective of this tutorial is to draw a clear, simple and meaningful panorama of existing standards and research contributions related to XML. To decode the various XML activities, we will see them through database glasses: we will look at the development of XML technology as a data management problem. The tutorial will be organized in three parts: data models for XML, data definition languages for XML and data manipulation languages for XML. For each of these three aspects, we will introduce the standards and explain their relationship to current state of research. We will notably cover the following material from the W3C: XML 1.0, XML Query Data Model, XML Infoset, XML Schema, DTDs, XPath, XSLT, XML Query Algebra and XML Query Language.

Tutorial 6: Publish and Subscribe Systems

Instructors: Arno Jacobsen and Francois Llirbat

The publish and subscribe paradigm is a simple to use interaction model that consists of information providers, who publish events to the system, and of information consumers, who subscribe to events of interest within the system. The publish and subscribe system ensures the timely notification of subscribers upon event occurrence. The publish and subscribe paradigm has recently gained great interest in the database community as a solution methodology for information dissemination applications with which the classical request/reply-style communication model (a.k.a. client/server model) fails to cope adequately. Information dissemination applications include applications such as: stock, sports and news tickers, tourist, travel and traffic information systems, as well as emergency notification systems. Common to all of these applications is the need to continuously collect and integrate data distributed among a large set of users, sites, and applications. The application must filter and deliver relevant data to interested users and applications in a timely manner. The classical pull-based approach is not suited to implement these applications for two reasons. First, to approximate “real time” behavior a client would need to continuously increase its frequency of information requests leading to server resource and network overload and congestion. Second, a pure pull-based solution does not support a high volatility of information sources, since new sources can only be discovered by searching the network. This may be very demanding when the network is large and is impossible in mobile and wireless environments where a continuous network access may not always be possible. The objective of this tutorial is twofold. One, we aim to present a comprehensive survey of application domains, system design choices, and existing system implementations to understand scope and applicability of this paradigm. Two, we aim to discuss the strengths and weaknesses of these systems and evaluate what still needs to be done to make the publish and subscribe paradigm a practical solution for large-scale information dissemination applications. To achieve this, the tutorial is organized along four main axes: applications, publish and subscribe systems, algorithms deployed in these systems, and open research questions.

Demos Program

- “Pre-Aggregation for Irregular OLAP Hierarchies with the TreeScape System”
Torben Bach Pedersen
- “The Transbase Hypercube RDBMS: Multidimensional Indexing of Relational Tables in the ‘Real World’”
Volker Markl
- “WAND: A CASE Tool for Data Warehouse Design”
Mateo Golfarelli
- “SINGAPORE: A System for Querying Heterogeneous Data Sources”
Ruxandra Domenig
- “GeoNode: An End-to-End System from Research Components”
Chris Clifton
- “DIVE: Database Integration for Virtual Engineering”
Hans-Peter Kriegel, Andreas Müller, Marco Pötke, and Thomas Seidl
- “TERRAFLY: A High-Performance Web-Based Digital Library System for Spatial Data Access”
Naphtali Rishé
- “An Extensible Model-Based Mediator System with Domain Maps”
Amarnath Gupta, Bertram Ludaescher, and Maryann E. Martone
- “The Dynamic View System (DVS): Mobile Agents to Support Web Views”
Constantinos Spyrou

External Referees

Ruchi Agrawal	Yannis Dimopoulos	Christian Haul
Ismail S. Altingovde	Lyman Do	Gisli R. Hjaltason
Thomas Bauer	Alin Dobra	JoAnne Holliday
Roberto Bayardo	Mehmet E. Donderler	Arvind Hulgeri
Khalid Benali	Kaushik Dutta	Ihab Ilyas
Phil Bohannon	Mohamed G. Elfeky	Panagiotis Ipeirotis
Vinayak Borkar	Suzanne Embury	Hiromu Ishii
Christof Bornhoevd	Utku Erdogdu	Eric D. Jacobsen
Peter Bosch	Jianping Fan	David Jones
Nacer Boudjlida	Leonidas Fegaras	Yildiray Kabak
Paul Bradley	Hakan Ferhatosmanoglu	Cheng Kai
Nicolas Bruno	Avigdor Gal	Thalis A. Kalfigopoulos
Jihad Boulos	Helena Galhardas	Anne Kao
David Botzer	Pankaj K Garg	Daniel Kaster
Luca Cabibbo	Minos Garofalakis	Timour Katchaounov
Gerome Canals	Goetz Graefe	Dimitris Katsaros
Chee-Yong Chan	Mati Golani	Ahmad Kayed
Leena Chandran-Wadia	Jonathan Goldstein	Chih-Horng Ke
Francois Charoy	Gosta Grahne	Maurice van Keulen
Chao-Chun Chen	Tony Griffiths	Amit Khivesara
Andrzej Cichocki	Giovanna Guerrini	Jukka Kiviniemi
Mariano Cilia	Dimitris Gunopulos	Nick Kline
Chris Clifton	Peter Haas	George Kollios
Hae Don Chon	Mohand-Said Hacid	Birgitta Koenig-Ries
R M Colomb	Stathes Hadjiefthymiades	Evangelos Kotsakis
Gautam Das	Marios Hadjieleftheriou	Shinsuke Kuroda
Zoran Despotovic	Toshihiko Hamano	Laks V.S. Lakshmanan
Marios Dikaiakos	Moustafa Hammad	Gokce Banu Laleci

Johan Larson	Olivier Perrin	Hiroki Takakura
Daniel Lieuwen	Dieter Pfoser	Yannis Theodoridis
Heiko Ludwig	Aggelos Pikrakīs	Helen Thomas
Ioana Manolescu	Evaggelia Pitoura	Eleni Tousidou
Yannis Manolopoulos	Magdalena Punceva	Vassilis J. Tsotras
Nikos Manoulis	Rajeev Rastogi	Shin-Mu Tseng
Olivera Marjanovic	Manfred Reichert	Wil van der Aalst
Mirette Marzouk	Abdelmounaam Rezgui	Debra VanderMeer
Giansalvatore Mecca	Berthier Ribeiro-Neto	Costas Vassilakis
Paolo Merialdo	Mirek Riedewald	Michalis Vazyrgiannis
Isabella Merlo	Prasan Roy	Vassilis Verykios
Joris Mihaeli	Lee Ryong	Radek Vingralek
Jun Miyazaki	Shazia W.Sadiq	Jochem Vonk
Mohamed Mokbel	Sandra Sampaio	Arjen de Vries
Pascal Molli	Yucel Saygin	Krzysztof Walczak
Gero Muehl	Albrecht Schmidt	Changzhou Wang
Tohru Mukai	Karsten Schulz	Gerhard Weikum
Miyuki Nakano	Bernhard Seeger	Waldemar Wiczerzycki
Alex Nanopoulos	Juergen Sellentin	Ouri Wolfson
Gonzalo Navarro	Jayant Sharma	Ming-Chuan Wu
Anne H.H. Ngu	Hala Skaf	Yi-Leh Wu
Kevin O'Gorman	R. Srikant	Ming Xiong
Patrick O'Neal	Suryanarayana Sripada	Sunay Yaldiz
Aris M. Ouksel	Ioana Stanoi	Bin Zhang
Stavros Papastavrou	Bernhard Stegmaier	Donghui Zhang
Cris Pedregal-Martin	Konrad Stocker	Andreas Zeidler
Martin Peim	Emily Su	
Peter Peinl	ChengYu Sun	