

Proceedings

**18th International
Conference on Data Engineering**

Industrial Sponsors



Proceedings

**18th International
Conference on Data Engineering**

26 February–1 March 2002 • San Jose, California

Edited by

Rakesh Agrawal, Klaus Dittrich, and Anne H.H. Ngu

Sponsored by

IEEE Computer Society



Los Alamitos, California

Washington • Brussels • Tokyo

Copyright © 2002 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 PR01531
ISBN 0-7695-1531-2
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](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](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 A. Denise Williams and Stephanie Kawada

Cover art production by Joseph Daigle/Studio Productions

Printed in the United States of America by The Printing House


IEEE
COMPUTER
SOCIETY

 **IEEE**

Table of Contents

18th International Conference on Data Engineering

Message from the General Chair.....	xiii
Message from the Program Co-Chairs	xiv
Conference Officers.....	xv
Program Committee.....	xvi
External Reviewers	xix

Keynote Address 1

HP—Inventing the Future of Storage.....	1
<i>Nora Denzel, Hewlett-Packard</i>	

Session 1: Database Applications and Experiences

DBXplorer: A System for Keyword-Based Search over Relational Databases	5
<i>S. Agrawal, S. Chaudhuri, and G. Das</i>	
TAILOR: A Record Linkage Tool Box	17
<i>M.G. Elfeky, V.S. Verykios, and A.K. Elmagarmid</i>	
Providing Database as a Service	29
<i>H. Hacıgümüş, B. Iyer, and S. Mehrotra</i>	

Session 2: Semistructured Data, Metadata and XML 1

Detecting Changes in XML Documents	41
<i>G. Cobéna, S. Abiteboul, and A. Marian</i>	
Reverse Engineering for Web Data: From Visual to Semantic Structures.....	53
<i>C.Y. Chung, M. Gertz, and N. Sundaresan</i>	
From XML Schema to Relations: A Cost-Based Approach to XML Storage	64
<i>P. Bohannon, J. Freire, P. Roy, and J. Siméon</i>	

Industry Session 1: Web Services Infrastructures

Web Services Framework	77
<i>B. Reinwald and S. Weerawarana</i>	
BizTalk—Web Services for Businesses.....	77
<i>J. Klein</i>	

Session 3: Scientific, Engineering, Statistical, Temporal, Spatial, and Geographical Databases 1

Sequenced Subset Operators: Definition and Implementation.....	81
<i>J. Dunn, S. Davey, A. Descour, and R.T. Snodgrass</i>	
Exploring Spatial Datasets with Histograms	93
<i>C. Sun, D. Agrawal, and A. El Abbadi</i>	
Efficient Temporal Join Processing Using Indices	103
<i>D. Zhang, V.J. Tsotras, and B. Seeger</i>	

Session 4: Semistructured Data, Metadata and XML 2

Similarity Flooding: A Versatile Graph Matching Algorithm and Its Application to Schema Matching.....	117
<i>S. Melnik, H. Garcia-Molina, and E. Rahm</i>	
Exploiting Local Similarity for Indexing Paths in Graph-Structured Data.....	129
<i>R. Kaushik, P. Shenoy, P. Bohannon, and E. Gudes</i>	
Structural Joins: A Primitive for Efficient XML Query Pattern Matching.....	141
<i>S. Al-Khalifa, H.V. Jagadish, N. Koudas, J.M. Patel, D. Srivastava, and Y. Wu</i>	

Session 5: Data Warehousing and OLAP

Condensed Cube: An Efficient Approach to Reducing Data Cube Size.....	155
<i>W. Wang, J. Feng, H. Lu, and J.X. Lu</i>	
Indexing Spatio-Temporal Data Warehouses.....	166
<i>D. Papadias, Y. Tao, P. Kalnis, and J. Zhang</i>	
Processing Reporting Function Views in a Data Warehouse Environment.....	176
<i>W. Lehner, W. Hümmer, and L. Schlesinger</i>	

Advanced Technology Seminar 1

P2P Information Systems.....	187
<i>K. Aberer and M. Hauswirth</i>	

Session 6: Scientific Engineering, Statistical, Temporal, Spatial, and Geographical Databases 2

Cost Models for Overlapping and Multi-Version B-Trees.....	191
<i>Y. Tao, D. Papadias, and J. Zhang</i>	
GADT: A Probability Space ADT for Representing and Querying the Physical World.....	201
<i>A. Faradjian, J. Gehrke, and P. Bonnet</i>	
Similarity Search Over Time-Series Data Using Wavelets.....	212
<i>I. Popivanov and R.J. Miller</i>	

Session 7: Semistructured Data, Metadata and XML 3

XGRIND: A Query-Friendly XML Compressor.....	225
<i>P.M. Tolani and J.R. Haritsa</i>	
Efficient Filtering of XML Documents with XPath Expressions.....	235
<i>C-Y. Chan, P. Felber, M. Garofalakis, and R. Rastogi</i>	
Mixing Querying and Navigation in MIX.....	245
<i>P. Mukhopadhyay and Y. Papakonstantinou</i>	

Industry Session 2: Data Mining Applications for CRM and Personalization

HP Decision Optimizer for Customer Relationship Management.....	255
<i>V. Singh</i>	
A System for Personalization of Web News Feeds.....	255
<i>C.C. Aggarwal and P.S. Yu</i>	

Poster Flash Session

An Intuitive Framework for Understanding Changes in Evolving Data Streams.....	261
<i>C.C. Aggarwal</i>	

Efficient OLAP Query Processing in Distributed Data Warehouses	262
<i>M. Akinde, M. Böhlen, T. Johnson, L.V.S. Lakshmanan, and D. Srivastava</i>	
FAST: A New Sampling-Based Algorithm for Discovering Association Rules.....	263
<i>B. Chen, P.J. Haas, and P. Scheuermann</i>	
A Graphical XML Query Language.....	264
<i>S. Flesca, F. Furfaro, and S. Greco</i>	
Out From Under the Trees	265
<i>C. Jermaine, E. Omiecinski, and W.G. Yee</i>	
An Efficient Index Structure for Shift and Scale Invariant Search of Multi-Attribute Time Sequences	266
<i>T. Kahveci, A. Singh, and A. Gürel</i>	
NeT & CoT: Inferring XML Schemas from Relational World	267
<i>D. Lee, M. Mani, F. Chiu, and W.W. Chu</i>	
Multivariate Time Series Prediction via Temporal Classification	268
<i>B. Liu and J. Liu</i>	
Data Cleaning and XML: The DBLP Experience.....	269
<i>W.L. Low, W.H. Tok, M.L. Lee, and T.W. Ling</i>	
Using Smodels (Declarative Logic Programming) to Verify Correctness of Certain Active Rules	270
<i>M. Nakamura and R. Elmasri</i>	
Attribute Classification Using Feature Analysis	271
<i>F. Naumann, C-T. Ho, X. Tian, L. Haas, and N. Megiddo</i>	
BestPeer: A Self-Configurable Peer-to-Peer System	272
<i>W.S. Ng, B.C. Ooi, and K-L. Tan</i>	
Efficient Algorithm for Projected Clustering	273
<i>E. Ng Ka Ka and A.W. Fu</i>	
Multiple Query Optimization by Cache-Aware Middleware Using Query Teamwork	274
<i>K. O’Gorman, D. Agrawal, and A. El Abbadi</i>	
Runtime Data Declustering over SAN-Connected PC Cluster System.....	275
<i>M. Oguchi and M. Kitsuregawa</i>	
How Good Are Association-Rule Mining Algorithms?.....	276
<i>V. Pudi and J.R. Haritsa</i>	
Extensible and Similarity-Based Grouping for Data Integration	277
<i>E. Schallehn, K-U. Sattler, and G. Saake</i>	
Specification-Based Data Reduction in Dimensional Data Warehouses	278
<i>J. Skyt, C.S. Jensen, and T.B. Pedersen</i>	
Exploiting Punctuation Semantics in Data Streams	279
<i>P. Tucker and D. Maier</i>	
The ATLaS System and Its Powerful Database Language Based on Simple Extensions of SQL	280
<i>H. Wang and C. Zaniolo</i>	
A Framework Towards Efficient and Effective Sequence Clustering	282
<i>W. Wang and J. Yang</i>	

Keynote Address 2

The Evolution of eBusiness Integration—from Data to Process.....	283
<i>Dale Skeen, Vitria Technology, Inc.</i>	

Session 8: Middleware and Metadata

Integrating Workflow Management Systems with Business-to-Business Interaction Standards.....	287
<i>M. Sayal, F. Casati, U. Dayal, and M-C. Shan</i>	
Declarative Composition and Peer-to-Peer Provisioning of Dynamic Web Services.....	297
<i>B. Benatallah, M. Dumas, Q.Z. Sheng, and A.H.H. Ngu</i>	
A Publish & Subscribe Architecture for Distributed Metadata Management.....	309
<i>M. Keidl, A. Kreutz, A. Kemper, and D. Kossmann</i>	

Industry Session 3: Web Caching

Personalization, Volatility, Performance, and Consistency in Web Caching.....	321
<i>X. Liu, Z. Zeng, and L. Jacobs</i>	
Alternative Approaches to Middle Tier Caching in WWW Server Side Infrastructures.....	321
<i>A. Datta</i>	

Advanced Technology Seminar 2

Techniques for Storing XML.....	323
<i>M. Fernandez and S. Amer-Yahia</i>	

Demo & Industrial Exhibits Session 1

Predator-Miner: Ad hoc Mining of Associations Rules within a Database Management System.....	327
<i>W.H. Tok, T.H. Ong, W.L. Low, I. Atmosukarto, and S. Bressan</i>	
Using Unity to Semi-Automatically Integrate Relational Schema.....	329
<i>R. Lawrence and K. Barker</i>	
SG-WRAP: A Schema-Guided Wrapper Generator.....	331
<i>X. Meng, H. Lu, H. Wang, and M. Gu</i>	
Demonstration: Active Asynchronous Transaction Management in High-Autonomy Federated Environment Using Data Agents: Global Change Master Directory v8.0.....	333
<i>O. Bukhres, S. Sikkupparbathyam, K. Nagendra, Z.B. Miled, M. Areal, L. Olsen, C. Gokey, D. Kendig, R. Cordova, G. Major, and J. Savage</i>	
XParent: An Efficient RDBMS-Based XML Database System.....	335
<i>H. Jiang, H. Lu, W. Wang, and J.X. Yu</i>	
The BINGO! Focused Crawler: From Bookmarks to Archetypes.....	337
<i>S. Sizov, S. Siersdorfer, M. Theobald, and G. Weikum</i>	
An Authorization System for Temporal Data.....	339
<i>A. Gal, V. Atluri, and G. Xu</i>	
YFilter: Efficient and Scalable Filtering of XML Documents.....	341
<i>Y. Diao, P. Fischer, M.J. Franklin, and R. To</i>	

Session 9: WWW and Databases

Design and Evaluation of Alternative Selection Placement Strategies in Optimizing Continuous Queries.....	345
<i>J. Chen, D.J. DeWitt, and J.F. Naughton</i>	

Design and Implementation of a High-Performance Distributed Web Crawler.....	357
<i>V. Shkapenyuk and T. Suel</i>	
Evaluating Top- <i>k</i> Queries over Web-Accessible Databases	369
<i>N. Bruno, L. Gravano, and A. Marian</i>	
Session 10: Heterogeneous, Distributed & Parallel Databases	
Exploring Aggregate Effect with Weighted Transcoding Graphs for Efficient Cache Replacement in Transcoding Proxies	383
<i>C-Y. Chang and M-S. Chen</i>	
Efficiently Ordering Query Plans for Data Integration	393
<i>A. Doan and A. Halevy</i>	
Active XQuery	403
<i>A. Bonifati, D. Braga, A. Campi, and S. Ceri</i>	
Industry Session 4: DBMS Extensions for Data Warehousing and Spatial Data	
Exploring SQL for ETML Program Execution.....	413
<i>J-C. Freytag, J. Huang, J. McPherson, and A. Bordia</i>	
Comparison of QuadTree and R-Tree Indexes in Oracle Spatial.....	413
<i>K.V. Ravi Kanth, S. Ravada, and D. Abugov</i>	
Session 11: WWW, Databases and Others	
A Fast Regular Expression Indexing Engine	419
<i>J. Cho and S. Rajagopalan</i>	
Keyword Searching and Browsing in Databases using BANKS	431
<i>G. Bhalotia, A. Hulgeri, C. Nakhe, S. Chakrabarti, and S. Sudarshan</i>	
Geometric-Similarity Retrieval in Large Image Bases	441
<i>I. Foudos, L. Palios, and E. Pitoura</i>	
Session 12: Database Access Structures and Techniques	
Efficient Indexing Structures for Mining Frequent Patterns	453
<i>B. Lan, B.C. Ooi, and K-L. Tan</i>	
Indexing of Moving Objects for Location-Based Services	463
<i>S. Šaltenis and C.S. Jensen</i>	
SCADDAR: An Efficient Randomized Technique to Reorganize Continuous Media Blocks.....	473
<i>A. Goel, C. Shahabi, S-Y. Yao, and R. Zimmermann</i>	
Advanced Technology Seminar 3	
Database Replication for the Mobile Era	483
<i>A. Wolski</i>	
Demo & Industrial Exhibits Session 2	
Advanced Process-Based Component Integration in Telcordia’s Cable OSS.....	485
<i>A.H.H. Ngu, D. Georgakopoulos, D. Baker, A. Cichocki, J. Desmarais, and P. Bates</i>	
OntoWebber: A Novel Approach for Managing Data on the Web	488
<i>Y. Jin, S. Xu, S. Decker, and G. Wiederhold</i>	

A Distributed Database Server for Continuous Media.....	490
<i>W.G. Aref, A.C. Catlin, A.K. Elmagarmid, J. Fan, J. Guo, M. Hammad, I.F. Ilyas, M.S. Marzouk, S. Prabhakar, A. Rezgui, S. Teoh, E. Terzi, Y. Tu, A. Vakali, and X.Q. Zhu</i>	
Managing Complex and Varied Data with the IndexFabric™.....	492
<i>N. Sample, B. Cooper, M. Franklin, G. Hjaltason, M. Shadmon, and L. Cohen</i>	
Content-Based Video Indexing for the Support of Digital Library Search.....	494
<i>M. Petković, R. van Zwol, H.E. Blok, W. Jonker, P.M.G. Apers, M. Windhouwer, and M. Kersten</i>	
NAPA: Nearest Available Parking Lot Application	496
<i>H.D. Chon, D. Agrawal, and A. El Abbadi</i>	
Mapping XML and Relational Schemas with Clio	498
<i>L. Popa, M.A. Hernández, Y. Velegrakis, R.J. Miller, F. Naumann, and H. Ho</i>	
StreamCorder: Fast Trial-and-Error Analysis in Scientific Databases.....	500
<i>E. Stolte and G. Alonso</i>	
Keynote Address 3	
Peer-to-Peer Data Management	503
<i>Hector Garcia-Molina, Stanford University</i>	
Session 13: Data, Text and Web Mining 1	
Data Mining Meets Performance Evaluation: Fast Algorithms for Modeling Bursty Traffic	507
<i>M. Wang, T. Madhyastha, N.H. Chan, S. Papadimitriou, and C. Faloutsos</i>	
δ -Clusters: Capturing Subspace Correlation in a Large Data Set.....	517
<i>J. Yang, W. Wang, H. Wang, and P. Yu</i>	
Efficient Evaluation of Queries with Mining Predicates.....	529
<i>S. Chaudhuri, V. Narasayya, and S. Sarawagi</i>	
Session 14: Database Engine Architecture and Implementation	
Recovery Guarantees for General Multi-Tier Applications	543
<i>R. Barga, D. Lomet, and G. Weikum</i>	
Fjording the Stream: An Architecture for Queries Over Streaming Sensor Data	555
<i>S. Madden and M.J. Franklin</i>	
Approximating a Data Stream for Querying and Estimation: Algorithms and Performance Evaluation	567
<i>S. Guha and N. Koudas</i>	
Industry Session 5: E-Business I	
From Marketplaces to Web Services.....	577
<i>M. Hsu</i>	
Engineering Challenges for Effective Supply Chain Control	577
<i>R. Krishnamurthy</i>	

Session 15: Data, Text and Web Mining 2

OSSM: A Segmentation Approach to Optimize Frequency Counting.....	583
<i>C.K-S. Leung, R.T. Ng, and H. Mannila</i>	
Towards Meaningful High-Dimensional Nearest Neighbor Search by Human-Computer Interaction.....	593
<i>C.C. Aggarwal</i>	
Fast Mining of Massive Tabular Data via Approximate Distance Computations.....	605
<i>G. Cormode, P. Indyk, N. Koudas, and S. Muthukrishnan</i>	

Session 16: Query Processing and Optimization 1

A Sampling-Based Estimator for Top- <i>k</i> Query.....	617
<i>C-M. Chen and Y. Ling</i>	
Improving Range Query Estimation on Histograms	628
<i>F. Buccafurri, L. Pontieri, D. Rosaci, and D. Saccà</i>	
Query Estimation by Adaptive Sampling.....	639
<i>Y-L. Wu, D. Agrawal, and A. El Abbadi</i>	

Advanced Technology Seminar 4

Bioinformatics Databases 1.....	649
<i>F. Olken</i>	

Industry Session 6: E-Business II

Enabling the Real-Time Enterprise	653
<i>A. Nori</i>	
E-Business Trends: Technology and Business Perspectives.....	653
<i>C. Kleissner</i>	

Panel Session 1

Processing Data Streams: Applications, Challenges and Approaches	655
<i>Panelists: David Maier, OGI, panel chair</i>	
<i>Michael Franklin, UC Berkeley</i>	
<i>Johannes Gehrke, Cornell</i>	
<i>Praveen Sheshadri, Microsoft</i>	
<i>Jennifer Widom, Stanford</i>	

Session 17: Data, Text and Web Mining 3

Lossy Reduction for Very High Dimensional Data	663
<i>C. Jermaine and E. Omiecinski</i>	
Discovering Similar Multidimensional Trajectories	673
<i>M. Vlachos, G. Kollios, and D. Gunopulos</i>	
Streaming-Data Algorithms for High-Quality Clustering.....	685
<i>L. O'Callaghan, N. Mishra, A. Meyerson, S. Guha, and R. Motwani</i>	

Session 18: Query Processing and Optimization 2

A Non-Blocking Parallel Spatial Join Algorithm.....	697
<i>G. Luo, J.F. Naughton, and C.J. Ellmann</i>	
Data Reduction by Partial Preaggregation	706
<i>P.-Å. Larson</i>	

Decoupled Query Optimization for Federated Database Systems	716
<i>A. Deshpande and J.M. Hellerstein</i>	
Industry Session 7: Performance-Oriented Architectures for Web Applications	
Clustered Application Servers.....	729
<i>D. Jacobs</i>	
TimesTen Caching Infrastructure and Tools.....	729
<i>M-A. Neimat</i>	
Panel Session 2	
Where are Our Promising Research Directions: Database Server, Middleware, or Applications?	731
<i>Panelists: Michael Carey, BEA systems</i>	
<i>Hector Garcia-Molina, Stanford University</i>	
<i>James Hamilton, Microsoft</i>	
<i>Hamid Pirahesh, IBM Almaden Research Center</i>	
<i>Bhavani Thuraisingham, NSF</i>	
<hr/>	
Author Index	733