COMAD 2008

Proceedings of the 14th International Conference on Management of Data

December 17-19, 2008
Indian Institute of Technology, Bombay, India.

Editors

Gautam Das
University of Texas Arlington, USA

N.L.Sarda
Indian Institute of Technology, Bombay, India

P.Krishna Reddy
International Institute of Information Technology, Hyderabad, India

ALLIED PUBLISHERS PVT. LTD.

New Delhi • Mumbai • Kolkata • Lucknow • Chennai
Nagpur • Bangalore • Hyderabad • Ahmedabad

Allied Publishers Private Limited
Regd. Off. : 15 J.N. Heredia Marg, Ballard Estate, Mumbai - 400001
12 Prem Nagar, Ashok Marg, Opp. Indira Bhawan, Lucknow - 226001
Prarthna Flats (2nd Floor), Navrangpura, Ahmedabad - 380009
3-2-844/6 & 7, Kachiguda Station Road, Hyderabad - 500027
5th Main Road, Gandhinagar, Bangalore - 560009
1/13-14 Asaf Ali Road, New Delhi - 110002
17 Chittaranjan Avenue, Kolkata - 700072
81 Hill Road, Ramnagar, Nagpur - 440010
751 Anna Salai, Chennai - 600002

The opinion expressed and figures provided in the COMAD-2008 proceedings are the sole responsibility of the authors. The publishers and the editors bear no responsibility in this regard.

ISBN : 978-81-8424-370-3

Published by Sunil Sachdev and printed by Ravi Sachdev at Allied Publishers Pvt. Ltd., (Printing Division),

COMAD 2008

14th International Conference on Management of Data

December 17-19, 2008, Bombay, India.

Organization Committee

General Chair
- Anand Deshpande (Persistent Systems)

Organizing Chair
- S. Sudarshan (IIT Bombay)

Program Committee Co-Chairs
- N. L. Sarda (IIT Bombay) [Database Track]
- Gautam Das (Univ. Texas, Arlington) [Web, IR and Mining Track]

Tutorial Co-Chairs
- Sreenivasa Kumar (IIT Madras)
- Divesh Srivastava (AT&T Labs)

Application/Industrial Program Co-Chairs
- Govindarajan V.R.
- S.C. Gupta (National Informatics Centre)

Panels and Student Papers
- Srinath Srinivasa (IIIT Bangalore)
- Arvind Hulgeri (Persistent Systems)

Publications Chair
- P. Krishna Reddy (IIIT Hyderabad)

Demonstrations Chair
- Ravindra Guravannavar (IIT Bombay)

Program Committee

- A. K. Majumdar (IIT Kharagpur)
- Alan Fekete (University of Sydney)
- Amol Deshpande (University of Maryland, USA)
- Anthony Tung (NUS, Singapore)
- Aristidis Gionis (Yahoo! Research Barcelona, Spain)
- Arnab Bhattacharya (IIT Kanpur)
- Christian Konig (Microsoft Research, USA)
- Dimitrios Gunopulos (University of California, Riverside)
- Dimitris Papadias (HKUST, Hong Kong)
- George Kollios (Boston University, USA)
- Gopal Mulagund (Oracle, Bangalore)
- Harrick M. Vin (Tata Research, Development and Design Center (TRDDC))
- Heikki Mannila (University of Helsinki, Finland)
- Jayant Haritsa (IISc Bangalore)
- Jayavel Shanmugasundaram (Yahoo Research)
- Kamalakar Karlapalem (IIIT Hyderabad)
- Krithi Ramamritham(IIT Bombay)
- Mukesh Mohania (IBM India Research Laboratory)
- Nick Koudas (University of Toronto)
- Sharma Chakravarthy(University of Texas at Arlington)
- S. K. Gupta (IIT Delhi)
- Srikanta Bedathur (Max-Planck Institut fur Informtik)
- Srini Parthasarathy (Ohio State University, USA)
- Subbarao Kambhampati (Arizona State University, USA)
- Sunita Sarawagi (IIT Bombay)
- Vagelis Hristidis (Florida International University)
- Vikram Pudi (IIIT Hyderabad)
- Vipin Kumar (University of Minnesota)

External Referees

Aditya Telang (University of Texas at Arlington)

Aravind Kalavagattu (Arizona State University/Yahoo)

Fernando Farfan (Florida International University)

Ganesh Ramakrishnan (IIT, Bombay)

Gang Fang (University of Minnesota)

Gaurav Pandey (University of Minnesota)

Garrett Wolf (Arizona State University)

Gowtham Atluri (University of Minnesota)

Lini Thomas (IIIT, Hyderabad)

Nan Zhang (University of Texas at Arlington)

Pabitra Mitra (IIT, Kharagpur)

RadhaKrishna P (Infosys)

Raghav Gautam (Oracle)

Ramakrishna Varadarajan (Florida International University)

Raju Balakrishnan (Arizona State University)

Ritesh Tiwari (IIIT, Hyderabad)

Roochi Mishra (University of Texas at Arlington)

Rohit Gupta (University of Minnesota)

Satyanarayana Valluri (IIIT, Hyderabad)

Soujanya Vadapalli (IIIT, Hyderabad)

Sharat Chandran (IIT Bombay)

S. Sudarshan (IIT Bombay)

Shyam Boriah (University of Minnesota)

Varun Chandola (University of Minnesota)

Blank

PREFACE

The International Conference on Management of Data (COMAD) seeks to provide to researchers, practitioners, developers and users of database and data management technologies a forum to present and discuss problems, innovations, experiences and emerging trends.

Modeled after ACM SIGMOD, COMAD has emerged as the premier database conference hosted in India. COMAD was first held in 1989 and the most recent conference was held in December 2006 in New Delhi, India. COMAD 2008 is the 14th in the COMAD series. COMAD 2008 is being held from December 17-19, 2008 in Mumbai, the largest metropolis and the commercial center of India.

Similar to previous years, COMAD 2008 ‘s scope includes not only traditional database areas but also added emphasis on Web, Information Retrieval and Data Mining. To this end, COMAD 2008 has two PC chairs with backgrounds covering these areas, and a program committee that reflects the expanded scope.

COMAD 2008 attracted 60 research, 12 student, and 5 industry submissions from 11 countries. The program committee, consisting of renowned data management experts from around the world, rigorously reviewed each paper. Twenty papers were selected out of forty eight submissions in the research track, retaining COMAD’s reputation as a selective conference that publishes high quality research. Three papers were selected out of twelve submissions for the student paper track, and 2 papers were selected out of 5 submissions for the industry track, including several invited from leading industrial research and development groups.

The submission and acceptance statistics this year continue COMAD’s tradition of significant international participation, with about 35% of the papers coming from over Europe, USA, East/South-East Asia, Australia and New Zealand. In order to ensure wide visibility of papers published at the conference, we are making arrangements with ACM SIGMOD for including the proceedings of the conference in the SIGMOD on-line and CD-ROM archives.

We are delighted to feature three keynote talks from Dr. Surajit Chaudhuri (Microsoft Research, USA), Dr. Raghu Ramakrishnan (Yahoo! Research, USA), and Dr. S. Seshadri (Kosmix, USA), covering the areas of decision support. web information extraction, and cloud computing.

COMAD 2008 also features an excellent tutorial program covering uncertain clustering and graph mining topics. In addition, there are several vendor presentation, demonstration, and panel sessions.

We are grateful for the support and generosity of a large number of people and organizations. Conference sponsors include Microsoft, Sybase, Yahoo!, and IBM. Microsoft Corporation provided the Conference Management Toolkit. The Indian Institute of Technology, Bombay, generously hosted the event.

The members of the COMAD Organizing Committee have worked extremely hard to make this conference a success. The members of the program committee, despite their extremely tight schedules, have invested their valuable time and expertise to ensure a high-quality program. Allied Publishers have done a great job with conference publications. Sayali Kulkarni has designed the cover page.

Welcome to the COMAD 2008 conference in Mumbai!

Gautam Das, University of Texas, Arlington, USA
N.L.Sarda, Indian Institute of Technology, Bombay, India December 2008
P.Krishna Reddy, IIIT, Hyderabad, India
S.Sudarshan, Indian Institute of Technology, Bombay, India

CONTENTS

Preface..... iv

Keynote Addresses 1

- Decision Support Queries: A solved problem?..... 3
  Surajit Chaudhuri (Head, Data Management Research Microsoft Research, Redmond WA, USA)
- Finding Information on the Web - Past, Present and Future,.....4
  S. Seshadri (CTO, Kosmix Inc.)
- Cloud Computing at Yahoo! Research,.....5
  Raghu Ramakrishnan (Chief Scientist, Yahoo! Research)

Research Sessions –Contributed Papers.....7

Web Search/IR

- *Kshitij: A Search and Page Recommendation System for Wikipedia*.....9
  Phanikumar Bhamidipati (IIIT Hyderabad), Kamalakar Karlapalem (IIIT Hyderabad)
- *Query Heartbeat: A Strange Property of Keyword Queries on the Web*.....21
  Karthik R (IIITB), Aditya Rachakonda (IIITB), Srinath Srinivasa (Indian Institute of Information Technology)
- *Personalized Web-page Rendering System*.....30
  Swapna Raj Prabakara Raj (IIT Madras), Balaraman Ravindran (IIT Madras)

Data Mining I

- *Disk-Based Sampling for Outlier Detection in High Dimensional Data*.....40
  Timothy De Vries (The University of Sydney), Sanjay Chawla, Pei Sun, Gia vinh Anh Pham
- *CUM: An Efficient Framework for Mining Concept Units*.....51
  Santhi Thilagam (NITK Surathkal)
- *Discovering Interesting Subsets Using Statistical Analysis* 60
  Maitreya Natu (TRDDC), Girish Palshikar (TRDDC)

Data Mining II

- *REBMEC: Repeat Based Maximum Entropy Classifier for Biological Sequences*.....71
  Pratibha Rani (IIIT, Hyderabad), Vikram Pudi (IIIT, Hyderabad)

- *An Incremental Summary Generation System*.....83
  Ravindranath Chowdary (IIT Madras), Sreenivasa Kumar P (IIT Madras)
- *Topic Distillation using Support Vector Data Description*.....93
  Vijaya Saradhi (TRDDC, Pune), Harish Karnick (Dept. of CSE, IIT Kanpur), Pabitra Mitra (Dept. of CSE, IIT Kharagpur)

Query Processing/Optimization

- *Runtime Optimization of Continuous Queries*.....104
  Balakumar Kendai, Sharma Chakravarthy (University of Texas Arlington)
- *Towards the Preservation of Keys in XML Data Transformation for Integration*.....116
  Md. Sumon Shahriar (University of south australia)
- *Exploiting Asynchronous IO using the Asynchronous Iterator Model*.....127
  Suresh Iyengar (IIT Bombay), S Sudarshan (IIT Bombay), Santosh Kumar (IIT Bombay), Raja Agrawal (IIT Bombay)

Security/Integrity

- *The Efficient Maintenance of Access Roles with Role Hiding*..... 139
  Chaoyi Pang (CSIRO), Xiuzhen Zhang (RMIT University), Yanchun Zhang, Ramamohanarao Kotagiri (University of Melbourne)
- *Ambiguity: Hide the Presence of Individuals and Their Privacy with Low information Loss* 150
  Hui Wang (Stevens Institute of Technology)

Transaction Management, Data Integration

- *Exploiting Semantics and Speculation for Improving the Performance of Read-only Transactions*.....162
  Ragunathan T (IIIT, Hyderabad), Krishna Reddy P (IIIT-Hyderabad)
- *Concurrency Control in Distributed MRA Index Structures*.....174
  Neha Singh (IIT Bombay), S Sudarshan (Indian Institute of Technology - Bombay)
- *Information Integration Across Heterogeneous Sources: Where Do We Stand and How to Proceed?*.....186
  Aditya Telang, Sharma Chakravarthy (University of Texas Arlington), Yan Huang

XML processing

- *Declaratively Producing Data Mash-ups* 198
  Sudarshan Murthy (Portland State University), David Maier (Portland State University)
- *On Inferring K Optimum Transformations of XML Document from Update Script to DTD* 210
  Nobutaka Suzuki (University of Tsukuba)
- *Efficient Evaluation of Forward XPath Axes over XML Streams*..... 222
  Abdul Nizar (IIT Madras), Sreenivasa Kumar (IIT Madras)

Research Session: Student Papers.....235

- *Modeling Uncertain and Imprecise Information in Process Modeling with UML* 237
  XIAO Jing (INSA-LATTIS), Pierre PINEL (INSA-LATTIS), Lei PI (IRIT), Vincent ARANEGA (IRIT), Claude BARON (INSA-LATTIS)
- *An Experiment with Distance Measures for Clustering* 241
  Ankita Vimal (IIIT, Hyderabad), Satyanarayana Valluri (IIIT-Hyderabad), Kamalakar Karlapalem (IIIT Hyderabad)
- *Querying for Information Integration: How to go from an Imprecise Intent to a Precise Query?*.....245
  Aditya Telang, Sharma Chakravarthy (University of Texas Arlington), Chengkai Li

Industrial/Application Session: Contributed Papers..... 249

- *Native Multidimensional Indexing in Relational Databases*..... 251
  David Hoksza (Charles University, Prague), Tomas Skopal (Charles University, Prague)
- *Forecasting Using Consistent Experts*.....261
  M Vijayalakshmi (IIT Bombay), Bernard Menezes (IIT Bombay), Venu Gopal (IIT Bombay)

Demonstrations.....271

- *The Orion Uncertain Data Management System*..... 273
  Sarvjeet Singh, Chris Mayfield, Sagar Mittal, Sunil Prabhakar, Susanne Hambrusch, Rahul Shah Department of Computer Science, Purdue University
- *Silverfish: A Contextual Knowledge Extraction and Aggregation System for Academics* 277
  Srinath Srinivasa and Aditya Ramana Rachakonda
  International Institute of Information Technology, Bangalore

Tutorials..... 281

- *Uncertain Clustering: Models, Methods and Applications*..... 283
  Zhenjie Zhang (National University of Singapore), and Dr. Anthony K. H. Tung (National University of Singapore)
- *Graph Mining Techniques and Their Applications*..... 284
  Prof. Sharma Chakravarty (University of Texas Arlington, USA)

Invited Industrial Talks..... 287

- *BI on Data and Content Together: What will you do with the derived insights?* 289
  Speaker: Mukesh Mohania, Senior Manager (Information Management), IBM India Research Lab
- *Internet Research: What's hot in Search, Advertizing, and Cloud Computing*..... 290
  Rajeev Rastogi, Vice President, Yahoo! Labs Bangalore
- *Sybase Appliance for Extreme Analytics*..... 291
  Shailesh Mungikar, Senior Engineer and Architect, Sybase Software (India) Pvt.Ltd.

Author Index.....293

Keynote Addresses

Keynote 1

Decision Support Queries: A solved problem?

Surajit Chaudhuri
(Head, Data Management Research,
Microsoft Research, Redmond WA, USA)

Abstract: Businesses rely on complex queries to enable data analysis. Yet, our understanding of how to handle complex queries is at best half-developed. Admittedly, this is a problem that is hardly new. Therefore, it is not surprising that the research world has moved on to newer problems. Yet, I argue that understanding the central technical problems and addressing them are foundational - as relevant for cloud data services for BI as they are for traditional BI software. In my talk, I will focus on a few of these challenges - resource governance and monitoring, physical design, and query optimization.

Biography: Surajit Chaudhuri is a Principal Researcher and a Research Area Manager at Microsoft Research, Redmond. Surajit has a PhD from Stanford University and is an ACM Fellow. He was awarded the ACM SIGMOD Contributions Award in 2004 and a 10 yearVLDB Best paper Award in 2007.

Keynote 2

Finding Information on the Web - Past, Present and Future.

S. Seshadri
(CTO, Kosmix Inc.)

Abstract: We will trace the evolution of how people have found information on the web in the past and outline how we think this will evolve in the future. We will dwell on the consumer pain points as well as the technology/business challenges of various approaches to finding information on the web.

Biography: Sesh is the CTO of Kosmix, a consumer internet company focused on organizing and connecting people to the information they want.

Sesh has several years of experience in academia, research labs, startups and Fortune 500 companies combining technology, products and business.Sesh was the Chief Technology Officer for Yahoo's R&D Center in India. He is also the founder of Strand Genomics, a biotechnology company and was a researcher at Bell Labs, Lucent Technologies after serving in the faculty of the Indian Institute of Technology, Mumbai. Sesh is a graduate of the Indian Institute of Technology, Chennai and the University of Wisconsin, Madison.

Keynote 3

Cloud Computing at Yahoo! Research

Raghu Ramakrishnan
(Chief Scientist, Yahoo! Research)

Title: Cloud Computing at Yahoo! Abstract: We are in the midst of a computing revolution. As the cost of provisioning hardware and software stacks grows, and the cost of securing and administering these complex systems grows even faster, we're seeing a shift towards computing clouds. Clouds are essentially services accessed over a network, and offer developers scalable, robust infrastructure on a "pay as you go" basis, with the ability to dynamically adjust the amount of "rented" resources, and thereby, the bill. Cloud services also raise the level of abstraction at which developers program, leading to shorter development cycles, and often enable previously unrealistic computational tasks at massive scale, leading to increased innovation. For cloud service providers, there is efficiency from amortizing costs and averaging usage peaks. Internet portals like Yahoo! have long offered application services, such as email for individuals and organizations. Companies are now offering services such as storage and compute cycles, enabling higher-level services to be built on top. In this talk, I will discuss Yahoo!'s vision of cloud computing, and describe some of the key initiatives.

Biography: Raghu Ramakrishnan is Chief Scientist for Audience and Cloud Computing at Yahoo!,and is a Research Fellow, heading the Community Systems area in Yahoo! Research.He is Professor of Computer Sciences at the University of Wisconsin-Madison (on leave), and was founder and CTO of QUIQ, a company that pioneered question-answering communities, powering Ask Jeeves' AnswerPoint as well as customer-support for companies such as Compaq.

Ramakrishnan's research is in the area of database systems, with a focus on data mining, query optimization, and web-scale data management, and has influenced query optimization in commercial database systems and the design of window functions in SQL:1999. His paper on the Birch clustering algorithm received the SIGMOD 10-Year Test-of-Time award, and he has written the widely-used text "Database Management Systems" (with Johannes Gehrke).

He is Chair of ACM SIGMOD, on the Board of Directors of ACM SIGKDD and the Board of Trustees of the VLDB Endowment, and has served as editor-in-chief of the Journal of Data Mining and Knowledge Discovery, associate editor of ACM Transactions on Database Systems, and the Database area editor of the Journal of Logic Programming. Ramakrishnan is a Fellow of the Association for Computing Machinery (ACM) and the Institute of Electrical and Electronics Engineers (IEEE), and has received several awards, including the ACM SIGKDD Innovations Award, the ACM SIGMOD Contributions Award, a Distinguished Alumnus Award from IIT Madras, a Packard Foundation Fellowship in Science and Engineering, and an NSF Presidential Young Investigator Award.

Research Sessions

Contributed Papers