VLDB2011
37th International Conference on Verly Large Data Bases, Seattle, WA

Proceedings of the VLDB Endowment

Volume 4, No. 7 – April 2011

Proceedings of the 37th International Conference on Very Large Data Bases, Seattle, WA

Editor-in-Chief:

H. V. Jagadish

Guest Editors:

José Blakeley, Joseph M. Hellerstein, Nick Koudas, Wolfgang Lehner, Sunita Sarawagi, Uwe Röhm

PVLDB – Proceedings of the VLDB Endowment

Volume 4, No. 7, April 2011.

The 37th International Conference on Very Large Data Bases, Seattle, WA.

Copyright 2011 VLDB Endowment

Volume 4, Number 7: VLDB 2011 Research Track Papers

Pages ii – vi and 409 – 469

ISSN 2150-8097, April 2011.

Additional copies only online at: portal.acm.org, arxiv.org/corr, and www.vldb.org

TABLE OF CONTENTS

Front Matter

Letters

Research Track Papers

PVLDB REVIEW BOARD

VLDB 2011 General PC Co-Chairs

José Blakeley, Microsoft

Joe Hellerstein, University of California – Berkeley

VLDB 2011 Research Track Co-Chairs

Nick Koudas, University of Toronto and Sysomos Inc.

Wolfgang Lehner, Dresden University of Technology

Sunita Sarawagi, IIT Bombay

Reviewer

Ashraf Aboulnaga (University of Waterloo)

Sibel Adali (Rensselaer Polytechnic Institute)

Charu Aggarwal (IBM Watson Research Center)

Divyakant Agrawal (Univ. California, Santa Barbara)

Anastasia Ailamaki (EPFL Lausanne)

Gustavo Alonso (ETH Zurich)

Shivnath Babu (Duke University)

Roberto Bayardo (Google)

Elisa Bertino (Purdue University)

Peter Boncz (CWI, Netherlands)

Angela Bonifati (Icar-CNR)

Christof Bornhoevd (SAP Palo Alto)

Mike Cafarella (University of Washington)

K. Selcuk Candan (Arizona State University)

Malu Castellanos (HP Labs)

Tiziana Catarci (University of Rome)

Chee-Yong Chan (National University of Singapore)

Kevin Chang (University of Illinois, Urbana-Champaign)

Surajit Chaudhuri (Microsoft Research)

Rada Chirkova (North Carolina State University)

Jan Chomicki (University at Buffalo)

Chin-Wan Chung (Korea Advanced Institute of SaT)

Chris Clifton (Purdue University)

Christine Collet (Grenoble Institute of Technology)

Graham Cormode (AT&T Labs)

Gautam Das (University of Texas, Arlington)

Anish Das Sarma (Yahoo! Research)

Amol Deshpande (University of Maryland)

AnHai Doan (University of Wisconsin)

Xin Dong (AT&T Labs)

Alexandre Evfimievski (IBM Research)

Wenfei Fan (University of Edinburgh & Bell Labs)

Johann-Christoph Freytag (Humboldt-Universität Berlin)

Johannes Gehrke (Cornell University)

Rainer Gemulla (IBM Almaden Research Center)

Aristides Gionis (Yahoo! Research)

Goetz Graefe (HP Labs)

Torsten Grust (Universität Tübingen, Germany)

Giovanna Guerrini (University of Genova)

Dimitris Gunopulos (University of Athens, Greece)

Theo Haerder (University of Kaiserslautern)

Alon Halevy (Google)

Vagelis Hristidis (Florida International University)

Meichun Hsu (HP Labs, Palo Alto)

Ihab Ilyas (University of Waterloo)

Zachary Ives (University of Pennsylvania)

Dean Jacobs (SAP)

Christian Jensen (Aalborg University)

Chris Jermaine (University of Florida)

Raghav Kaushik (Microsoft Research)

Bettina Kemme (McGill University)
Eamonn Keogh (University of California, Riverside)
Martin Kersten (CWI)
Christoph Koch (Cornell University)
Flip Korn (AT&T Labs)
Donald Kossmann (ETH Zurich)
Alberto Laender (Federal University of Minas Gerais)
Dongwon Lee (Penn State University)
Kristen Lefevre (University of Michigan)
Chen Li (University of California, Irvine)
Bin Liu (University of Michigan)
David Lomet (Microsoft Research)
Samuel Madden (MIT)
Nikos Mamoulis (University of Hong Kong)
Ioana Manolescu (INRIA)
Claudia Medeiros (University of Campinas)
Sergey Melnik (Google)
Marco Mesiti (Universita degli Studi di Milano)
Chaitanya Mishra (Facebook Inc.)
Felix Naumann (University of Potsdam)
Raymond Ng (University of British Columbia)
Christopher Olston (Yahoo! Research)
Themis Palpanas (University of Trento)
Dimitris Papadias (Hong Kong University of SaT)
Stavros Papadopoulos (Chinese University of Hong Kong)
Stefano Paraboschi (University of Bergamo)
Jian Pei (Simon Fraser University)
Rachel Pottinger (University of British Columbia)
Vijayshankar Raman (IBM Almaden Research Centre)
Prakash Ramanan (Wichita State University)
Louiqa Raschid (University of Maryland)
Kenneth Ross (Columbia University)
Elke Rundensteiner (Worcester Polytechnic Institute)
Yehoshua Sagiv (Hebrew University, Jerusalem)
Ken Salem (University of Waterloo)
Kai-Uwe Sattler (Ilmenau University of Technology)
Bernhard Seeger (University of Marburg)
Jayavel Shanmugasundaram (Yahoo! Research)
Kyuseok Shim (Seoul National University)
Divesh Srivastava (AT&T Labs)
Dan Suciu (University of Washington)
S. Sudarshan (IIT Bombay)
Kian-Lee Tan (National University of Singapore)
Val Tannen (University of Pennsylvania)
Jens Teubner (ETH Zurich)
Martin Theobald (Max-Planck-Institut für Informatik)
Frank Tompa (University of Waterloo)
Anthony Tung (National University of Singapore)
Patrick Valduriez (INRIA)
Wie Wang (University of North Carolina)
Gerhard Weikum (Max Planck Institute, Germany)
Yuqing Wu (Indiana University)
Fei Xu (Microsoft Search)
Sihem Yahia (Yahoo! Research)
Jun Yang (Duke University)
Cong Yu (Yahoo! Research)
Jefferey Yu (Chinese University of Hong Kong)
Ting Yu (North Carolina State University)
Xiaohui Yu (York University)
Justin Zobel (University of Melbourne)

PVLDB Information Director

Gerald Weber (University of Auckland)

VLDB 2011 Proceedings Chair

Uwe Röhm (University of Sydney)

Steering Committee

Serge Abiteboul, Peter Apers, Philip Bernstein, Elisa Bertino, Peter Buneman, Martin Kersten, Z. Meral Ozsoyuglu

LETTER FROM THE VLDB 2011 PROCEEDINGS CHAIR

It is my pleasure to introduce the April issue of the Proceedings of the VLDB Endowment (PVLDB). As you know, PVLDB has now a monthly publication cycle with a rolling monthly deadline. Establishing this new continuous reviewing and publication process is an interesting experience for everyone involved – and at this occasion, many thanks to all colleagues who help making this happen. The goal is to provide a venue for high-quality database research papers that allows publishing research results as early as possible, while still getting the opportunity to present them at VLDB. The six papers papers contained in this issue will be presented at the 37th International Conference on Very Large Data Bases (VLDB 2011) to take place in Seattle later this year. In the best tradition of VLDB, the papers are of excellent quality and cover a wide range of topics including schema matching techniques, optimizations for MapReduce, data privacy in recommendations systems, and implementation techniques for logical database recovery.

The first paper, "Synthesizing Products for Online Catalogs" by Hoa Nguyen et al, presents a new schema reconciliation approach to automatically synthesize the product entries of online product search engines. Using sample data from a commercial online shopping engine, the authors compare the effectiveness of their approach to state-of-the-art schema matching techniques with regard to the precision and recall of search results. The next two papers focus on performance optimizations for modern data processing infrastructures: "Column-Oriented Storage Techniques for MapReduce", by Avrilia Floratou et al, investigates how one can incorporate column-oriented storage techniques into a MapReduce framework. The work discusses several implementation details and shows that it can significantly improve the performance of the map phase. In "Implementing Performance Competitive Logical Recovery", David Lomet at al study the implications of a de-coupled database architecture – with strictly separated transactional and data management functionalities – for the recovery component. The authors discuss how logical database recovery can be implemented for such architectures, and present performance optimizations that allow achieving comparable performance to physical database recovery.

Next is "Personalized Social Recommendations - Accurate or Private?" by Ashwin Machanavajjhala et al. This paper studies the tradeoff between accuracy and privacy-preservation when making social recommendations. The authors show how existing recommendation algorithms can be modified to provide differential privacy, but argue that accurate recommendations are limited to small subsets of a social network or to choosing more lenient privacy settings. Another important topic for today's web environment is investigated in "Efficient Diversification of Web Search Results" by Gabriele Capannini et al: This paper presents several new methods for search result diversification, some of which are shown to offer up-to two-orders of magnitude faster performance than state-of-the-art approaches. We stay in the context of today's Internet services with the last paper of this issue too: "Social Content Matching in MapReduce" by Gianmarco De Francisci Morales et al, investigates how to efficiently solve content matching problems using a MapReduce framework. The authors propose two corresponding matching algorithms and explore the scalability and the effectiveness of their approaches using datasets extracted from real-world social-media web sites (Flickr and Yahoo! Answers).

I hope that you enjoy these papers as much as I did, and I look forward to the presentations and discussions of these research results at this year's VLDB 2011 in Seattle.

Uwe Röhm, University of Sydney
VLDB 2011 Proceedings Chair