Advances in Database Technology
– EDBT 2010

13th International Conference on Extending Database Technology
Lausanne, Switzerland, March 22–26, 2010
Proceedings

Editors:
Ioana Manolescu (INRIA, France)
Stefano Spaccapietra (EPFL, Switzerland)
Jens Teubner (ETH Zurich, Switzerland)
Masaru Kitsuregawa (Tokyo University, Japan)
Alain Léger (Orange – France Telecom R&D, France)
Felix Naumann (Hasso Plattner Institute, Germany)
Anastasia Ailamaki (EPFL, Switzerland)
Fatma Özcan (IBM Almaden Research Center, USA)
Table of Contents

Foreword ................................................................. ix
Program Committee Members ........................................ x–xi

Invited Papers
Provenance for Database Transformations
   *Val Tannen* ............................................................. 1
Scalable Ontology-Based Information Systems
   *Ian Horrocks* .......................................................... 2

Research Sessions
XML and Semi-Structured Data
Feedback-driven Result Ranking and Query Refinement for Exploring Semi-structured Data Collections
   *Huiping Cao, Yan Qi, K. Selcuk Candan, and Maria Luisa Sapino* ................. 3–14
Beyond Pages: Supporting Efficient, Scalable Entity Search with Dual-Inversion Index
   *Tao Cheng and Kevin Chang* .............................................. 15–26
Processing XPath queries with forward and downward axes over XML Streams
   *Makoto Onizuka* .......................................................... 27–38

OLAP and Decision Support
Correlation Aware Synchronization for Near Real Time Decision Support Systems
   *Ying Yan and Wen-Syan Li* ................................................ 39–50
Turbo-Charging Hidden Database Samplers with Overflowing Queries and Skew Reduction
   *Arjun Dasgupta, Nan Zhang, and Gautam Das* ........................................ 51–62
Region-based Online Promotion Analysis
   *Tianyi Wu, Yizhou Sun, Cuiping Li, and Jiawei Han* ................................. 63–74

Distributed Databases
The Data Cyclotron Query Processing Scheme
   *Romulo Goncalves and Martin Kersten* .............................................. 75–86
Gossiping Personalized Queries
   *Xiao Bai, Marin Bertier, Rachid Guerraoui, Anne-Marie Kermarrec, and Vincent Leroy* ...................................................................................................................... 87–98
Optimizing Joins in a Map-Reduce Environment
   *Foto Afrati and Jeffrey Ullman* ..................................................... 99–110
### Security and Privacy 1

- **Private Record Matching Using Differential Privacy**  
  *Ali Inan, Murat Kantarcioglu, Gabriel Ghinita, and Elisa Bertino*  
  111–122

- **The Hardness and Approximation Algorithms for L-Diversity**  
  *Xiaokui Xiao, Ke Yi, and Yufei Tao*  
  123–134

- **K-Symmetry Model for Identity Anonymization in Social Networks**  
  *Wentao Wu, Yanghua Xiao, Wei Wang, Zhengying He, and Zhikui Wang*  
  135–146

### XPath and XQuery Optimization

- **Let SQL Drive the XQuery Workhorse**  
  *Torsten Grust, Manuel Mayr, and Jan Rittinger*  
  147–158

- **Efficient Physical Operators for cost-based XPath Execution**  
  *Haris Georgiadis, Minas Charalambides, and Vasilis Vassalos*  
  159–170

- **Statistics-based Parallelization of XPath Queries in Shared Memory Systems**  
  *Rajesh Bordawekar, Lipyeow Lim, Anastasios Kementsietsidis, and Bryant Wei-Lun Kok*  
  171–182

### Query Processing and Optimization 1

- **Adaptive Join Processing in Pipelined Plans**  
  *Kwanchai Eurviriyankul, Norman W. Paton, Alvaro A. A. Fernandes, and Steven Lynden*  
  183–194

- **BSkyTree: Scalable Skyline Computation Using A Balanced Pivot Selection**  
  *Jongwuk Lee and Seung-won Hwang*  
  195–206

- **Stream Schema: Providing and Exploiting Static Metadata for Data Stream Processing**  
  *Peter M. Fischer, Kyumars Sheykh Esmaili, and Renée J. Miller*  
  207–218

### Optimization for Modern Hardware

- **Warm Cache Costing – A Feedback Optimization Technique for Buffer Pool Aware Costing**  
  *Ramanujam Halasipuram S and Edwin Seputis*  
  219–227

- **Position List Word Aligned Hybrid: Optimizing Space and Performance for Compressed Bitmaps**  
  *François Deliègue and Torben Bach Pedersen*  
  228–239

- **Suffix Tree Construction Algorithms on Modern Hardware**  
  *Dimitris Tsirigiotis and Nick Koudas*  
  240–251

### Scientific Databases and OLAP

- **Reducing Metadata Complexity for Faster Table Summarization**  
  *K. Selcuk Candan, Mario Cataldi, and Maria Luisa Sapino*  
  252–263

- **Splash: Ad-Hoc Querying of Data and Statistical Models**  
  *Lujun Fang and Kristen LeFevre*  
  264–275

- **Anchoring Millions of Distinct Reads on the Human Genome within Seconds**  
  *Tien Huynh, Michail Vlachos, and Isidore Rigoutsos*  
  276–286
Data Provenance

Techniques for Efficiently Querying Scientific Workflow Provenance Graphs

Manish Anand, Shawn Bowers, and Bertram Ludaescher ........................................... 287–298

Fine-grained and efficient lineage querying of collection-based workflow provenance

Paolo Missier, Norman W. Paton, and Khalid Belhajjame ........................................ 299–310

Lost Source Provenance

Jing Zhang and H.V. Jagadish ................................................................. 311–322

Probabilistic and Spatial Databases

Bridging the Gap Between Intensional and Extensional Query Evaluation in
Probabilistic Databases

Abhay Jha, Dan Olteanu, and Dan Suciu ...................................................... 323–334

Probabilistic Path Queries in Road Networks: Traffic Uncertainty Aware Path Selection

Ming Hua and Jian Pei ................................................................. 335–346

Probabilistic Threshold k Nearest Neighbor Queries over Moving Objects in Symbolic
Indoor Space

Bin Yang, Hua Lu, and Christian S. Jensen ........................................... 347–358

Query Processing and Optimization 2

A Simple (yet Powerful) Algebra for Pervasive Environments

Yann Gripay, Frédérique Laforest, and Jean-Marc Petit ..................................... 359–370

Self-selecting, self-tuning, incrementally optimized indexes

Goetz Graefe and Harumi Kuno ...................................................... 371–381

Minimizing Database Repros using Language Grammars

Nicolas Bruno ................................................................. 382–393

Spatial Databases

Efficient and Scalable Multi-Geography Route Planning

Vidhya Balasubramanian, Dmitri Kalashnikov, Sharad Mehrotra, and Nalini
Venkatasubramanian ................................................................. 394–405

Querying Trajectories Using Flexible Patterns

Marcos R Vieira, Petko Bakalov, and Vassilis J. Tsotras ...................................... 406–417

Querying Spatial Patterns

Vishwakarma Singh, Arnab Bhattacharya, and Ambuj K. Singh ................................ 418–429

Technologies for the Web

Indexing Relations on the Web

Sergio Mergen, Juliana Freire, and Carlos Heuser ........................................... 430–440

An Execution Environment for C-SPARQL Queries

Davide Francesco Barbieri, Daniele Braga, Stefano Ceri, and Michael Grossniklaus 441–452

Rewrite Techniques for Performance Optimization of Schema Matching Processes

Eric Peukert, Henrik Berthold, and Erhard Rahm ........................................... 453–464
Ranking and Nearest Neighbor

Fast Computation of SimRank for Static and Dynamic Information Networks
Cuiping Li, Jiawei Han, Guoming He, Xin Jin, Yizhou Sun, Yintao Yu, and Tianyi Wu .......................................................... 465–476

Probabilistic Ranking over Relations
Lijun Chang, Jeffrey Xu Yu, Lu Qin, and Xuemin Lin ........................ 477–488

Privacy Preserving Group Nearest Neighbor Queries
Tanzima Hashem, Lars Kulik, and Rui Zhang .............................. 489–500

Data Cleaning and Curation

Subsumption and Complementation as Data Fusion Operators
Jens Bleiholder, Sascha Szott, Melanie Herschel, Frank Kaufer, and Felix Naumann ............................................................... 501–512

HARRA: Fast Iterative Hashed Record Linkage for Large-Scale Data Collections
Hung-sik Kim and Dongwon Lee ................................................ 513–524

Finding Misplaced Items in Retail by Clustering RFID Data
Leonardo Weiss Ferreira Chaves, Erik Bachmann, and Klemens Böhm .......................... 525–536

XML Keyword Search

Keyword Search for Data-Centric XML Collections with Long Text Fields
Arash Termehchy and Marianne Winslett ...................................... 537–548

Fast ELCA Computation for Keyword Queries on XML Data
Rui Zhou, Chengfei Liu, and Jianxin Li ....................................... 549–560

Suggestion of Promising Result Types for XML Keyword Search
Jianxin Li, Chengfei Liu, Rui Zhou, and Wei Wang .......................... 561–572

Personalization and Preferences

Feedback-Based Annotation, Selection and Refinement of Schema Mappings for Dataspaces
Khalid Belhajjame, Norman W. Paton, Suzanne Embury, Alvaro A. A. Fernandes, and Cornelia Hedeler ........................................... 573–584

PerK: Personalized Keyword Search in Relational Databases through Preferences
Kostas Stefanidis, Marina Drosou, and Evaggelia Pitoura .................. 585–596

Efficient Computation of Trade-Off Skylines
Christoph Lofi, Ulrich Güntzer, and Wolf-Tilo Balke ........................ 597–608

Security and Privacy 2

How to Authenticate Graphs Without Leaking
Ashish Kundu and Elisa Bertino ................................................. 609–620

Trustworthy Vacuuming and Litigation Holds in Long-term High-integrity Records Retention
Ragib Hasan and Marianne Winslett ........................................ 621–632

Algorithm-safe Privacy-Preserving Data Publishing
Xin Jin, Nan Zhang, and Gautam Das ........................................... 633–644
Industrial Sessions
Transactions and Distribution

BronzeGate: Real-time Transactional Data Obfuscation for GoldenGate
  Shenoda Guirguis and Alok Pareek ........................................ 645–650

Logging Last Resource Optimization for Distributed Transactions in Oracle WebLogic Server
  Tom Barnes, Adam Messinger, Paul Parkinson, Amit Ganesh, German Shegalov, Saraswathy Narayan, and Srinivas Kareenhalli ........................................ 651–656

DEDUCE: At the Intersection of MapReduce and Stream Processing
  Vibhore Kumar, Henrique Andrade, Bugra Gedik, and Kun-Lung Wu ............ 657–662

New Applications

An Experimental Study of Time-Constrained Aggregate Queries
  Ying Hu, Wen-Chi Hou, Seema Sundara, and Jagannathan Srinivasan ............ 663–668

Xbase: Cloud-enabled Information Appliance for Healthcare
  Wen-Syan Li, Jianfeng Yan, Ying Yan, and Jin Zhang ............................ 669–674

Data Warehousing and Analytics

Aggregation of asynchronous electric power consumption time series knowing the integral
  Raja Chiky, Laurent Decreusefond, and Georges Hebrail ............................ 675–680

A Plan for OLAP
  Bernhard Jaecksch, Franz Faerber, and Wolfgang Lehner .......................... 681–686

Augmenting OLAP Exploration with Dynamic Advanced Analytics
  Benjamin Leonhardi, Bernhard Mitschang, Ruben Pulido de los Reyes, Christoph Sieb, and Michael Wurst ........................................ 687–692

Demonstrations

Advanced Knowledge Discovery on Movement Data with the GeoPKDD system
  Mirco Nanni, Roberto Trasarti, Chiara Renso, Fosca Giannotti, and Dino Pedreschi ........................................ 693–696

Timely YAGO: Harvesting, Querying, and Visualizing Temporal Knowledge from Wikipedia
  Yafang Wang, mingjie zhu, Lizhen Qu, Marc Spaniol, and Gerhard Weikum ....... 697–700

PARINDA: An Interactive Physical Designer for PostgreSQL
  Cristina Maier, Debabrata Dash, Ioannis Alagiannis, Anastasia Ailamaki, and Thomas Heinis ........................................ 701–704

BIAEditor - Matching Process and Operational Data for a Business Impact Analysis
  Sylvia Radeschütz, Florian Niedermann, and Wolfgang Bischoff .............. 705–708

Pattern Detector: Fast Detection of Suspicious Stream Patterns for Immediate Reaction
  Ira Assent, Hardy Kremer, Stephan Günemann, and Thomas Seidl ............... 709–712

BP-Ex – A uniform query engine for Business Process Execution traces
  Eran Balan, Tova Milo, and Tal Sterenzy ........................................ 713–716
B-Fabric: The Swiss Army Knife for Life Sciences
Can Türker, Fuat Akal, Dieter Joho, Ralph Schlaphbach, Christian Panse, Simon
Barkow-Oesterreicher, and Hubert Rehrauer .......................................................... 717–720

Tutorials

FPGAs: A New Point in the Database Design Space
Rene Mueller and Jens Teubner ................................................................. 721–723

Querying the Deep Web
Andrea Calì and Davide Martinenghi ...................................................... 724–727

Author Index ................................................................. 728–730
Foreword

The papers in this volume were presented at the 13th International Conference on Extending Database Technology (EDBT 2010), held in Lausanne, Switzerland, March 22–26, 2010. Starting last year, EDBT is held jointly with the ICDT (International Conference on Database Theory) conference.

This year, the conference research track received a total of 307 submissions, out of which 54 were accepted. This amounts to an acceptance rate of about 18%, confirming the reputation of the EDBT conference as being quite selective! The program committee consisted of 75 members, leading to 12–13 articles to be reviewed by each PC member. While such load is not per se inconsistent with previous editions of EDBT or other prestigious conferences such as, e.g., SIGMOD or VLDB, some reviewers found this quite a heavy load, and significant care had to be taken to ensure appropriate external reviewers or load re-balancing. The submissions were thoroughly discussed, typically as soon as three reviews were available for the paper, even if this was before the review submission deadline; overall, 444 comments were exchanged through the electronic submission tool (EasyChair). The chair is particularly grateful to the PC members which have taken the time to carefully consider papers and debate each other’s arguments. Given the high number of interesting submissions, the PC has worked with the chair to select the best submissions, and did not consider issues such as the overall balance of various database research areas in the conference program. Thus, the proceedings you are holding now reflect the best and most interesting works submitted to our conference, and you can consider the way they split over various topics as a snapshot of the state of our research community as of September 2009.

The conference program is beautifully complemented by the other tracks. Ian Horrocks and Val Tannen have gracefully accepted our invitation to deliver keynote talks at the conference. Alain Léger and Masaru Kitsuregawa have put together an exciting industrial program. Fatma Özcan has chaired the demonstration track, while Felix Naumann has organized the selection of two interesting tutorials. Natassa Ailamaki has helped add to the conference a set of workshops on timely issues. The chairs all join together in acknowledging the careful work of their program committees and external reviewers, and thanking them for their effort.

Standing quite proud of our selection, we hope you will enjoy the program as much as we enjoyed putting it together, and are looking forward to an exciting conference in Lausanne!

Ioana Manolescu (Program Chair)  
Stefano Spaccapietra (General Chair)  
Jens Teubner (Proceedings Chair)  
Masaru Kitsuregawa (Industr. & Appl. Co-Chair)  

Alain Léger (Industr. & Appl. Co-Chair)  
Felix Naumann (Tutorial Chair)  
Anastasia Ailamaki (Workshop Chair)  
Fatma Özcan (Demonstrations Chair)
### Program Committee Members

#### Research

<table>
<thead>
<tr>
<th>Names</th>
<th>Names</th>
<th>Names</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amr El Abbadi</td>
<td>Torsten Grust</td>
<td>Jaroslav Pokorny</td>
</tr>
<tr>
<td>Gustavo Alonso</td>
<td>Vagelis Hristidis</td>
<td>Ravishankar Ramamurthy</td>
</tr>
<tr>
<td>Denilson Barbosa</td>
<td>Wynne Hsu</td>
<td>Philippe Rigaux</td>
</tr>
<tr>
<td>Klemens Boehm</td>
<td>Ihab Ilyas</td>
<td>Tore Risch</td>
</tr>
<tr>
<td>Angela Bonifati</td>
<td>Yaron Kanza</td>
<td>Domenico Sacca</td>
</tr>
<tr>
<td>Philippe Bonnet</td>
<td>Alfon Kemper</td>
<td>Mohand Said-Hacid</td>
</tr>
<tr>
<td>Luc Bouganim</td>
<td>Eamonn Keogh</td>
<td>Pierangela Samaratati</td>
</tr>
<tr>
<td>Loreto Bravo</td>
<td>Martin Kersten</td>
<td>Thomas Seidl</td>
</tr>
<tr>
<td>Luca Cabibbo</td>
<td>Alexandros Labrinidis</td>
<td>Kyuseok Shim</td>
</tr>
<tr>
<td>Diego Calvanese</td>
<td>Philippe Lamarre</td>
<td>Altigian Da Silva</td>
</tr>
<tr>
<td>Alessandro Campi</td>
<td>Paul Larson</td>
<td>Divesh Srivastava</td>
</tr>
<tr>
<td>Malu Castellanos</td>
<td>Bertram Ludascher</td>
<td>Dan Suciu</td>
</tr>
<tr>
<td>Dario Colazzo</td>
<td>Stefan Manegold</td>
<td>Jens Teubner</td>
</tr>
<tr>
<td>Emirar Curtmola</td>
<td>Ioana Manolescu (chair)</td>
<td>Peter Triantafilou</td>
</tr>
<tr>
<td>AnHai Doan</td>
<td>Amélie Marian</td>
<td>Patrick Valduriez</td>
</tr>
<tr>
<td>Anne Doucet</td>
<td>Volker Markl</td>
<td>Maurice Van Keulen</td>
</tr>
<tr>
<td>Cedric DuMouza</td>
<td>Giansalvatore Mecca</td>
<td>Vassilis Vassalos</td>
</tr>
<tr>
<td>Piero Fraternali</td>
<td>Paolo Missier</td>
<td>Michalis Vazirgiannis</td>
</tr>
<tr>
<td>Christine Froidevaux</td>
<td>Mirella Moro</td>
<td>Jef Wijsen</td>
</tr>
<tr>
<td>Irini Fundulaki</td>
<td>Wolfgang Nejdl</td>
<td>Haruo Yokota</td>
</tr>
<tr>
<td>Ariel Fuxman</td>
<td>Raymond Ng</td>
<td>Masatoshi Yoshikawa</td>
</tr>
<tr>
<td>Helena Galhardas</td>
<td>Matthias Nicola</td>
<td>Jeffrey Yu Xu</td>
</tr>
<tr>
<td>Floris Geerts</td>
<td>Boris Novikov</td>
<td>Aoying Zhou</td>
</tr>
<tr>
<td>Aristides Gionis</td>
<td>Nicola Onose</td>
<td></td>
</tr>
<tr>
<td>Lukasz Golab</td>
<td>Michalis Petropoulos</td>
<td></td>
</tr>
<tr>
<td>Daniela Grigori</td>
<td>Neoklis Polyzotis</td>
<td></td>
</tr>
</tbody>
</table>

#### Industrial & Applications

<table>
<thead>
<tr>
<th>Names</th>
<th>Names</th>
<th>Names</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soren Auer</td>
<td>Alain Léger (chair)</td>
<td>Carsten Saathoff</td>
</tr>
<tr>
<td>Francois Bry</td>
<td>Wen-Syan Li</td>
<td>Tomas Vitvar</td>
</tr>
<tr>
<td>Orri Erling</td>
<td>Anne Monceaux</td>
<td>Alexander Wahler</td>
</tr>
<tr>
<td>Fabien Gandon</td>
<td>Miyuki Nakano</td>
<td>Haofen Wang</td>
</tr>
<tr>
<td>German Herrero</td>
<td>Massimo Paolucci</td>
<td>Kun-Lung Wu</td>
</tr>
<tr>
<td>Ruben Lara Hernandez</td>
<td>Dimitris Plexousakis</td>
<td>Yiannis Kompatsiariys</td>
</tr>
<tr>
<td>Masaru Kitsuregawa (chair)</td>
<td>Steffen Saab</td>
<td></td>
</tr>
</tbody>
</table>

#### Demonstrations

<table>
<thead>
<tr>
<th>Names</th>
<th>Names</th>
<th>Names</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laurent Amsaleg</td>
<td>Ohad Greenshpam</td>
<td>Stelios Paparizos</td>
</tr>
<tr>
<td>Christof Bornhoedv</td>
<td>Kristen LeFevre</td>
<td>Evaggelia Pitoura</td>
</tr>
<tr>
<td>Bjorn Tor Jonsson</td>
<td>Qiong Luo</td>
<td>Pinar Senkul</td>
</tr>
<tr>
<td>Hakan Hacigumus</td>
<td>Fatma Özcan (chair)</td>
<td>Alkis Simits</td>
</tr>
</tbody>
</table>
David Simmen
Nesime Tatbul

Zografoula Vagena
Yannis Velegrakis
Stratis Viglas

Cong Yu
Provenance for Database Transformations

Val Tannen
University of Pennsylvania, USA
val@cis.upenn.edu

ABSTRACT

Database transformations (queries, views, mappings) take apart, filter, and recombine source data in order to populate warehouses, materialize views, and provide inputs to analysis tools. As they do so, applications often need to track the relationship between parts and pieces of the sources and parts and pieces of the transformations’ output. This relationship is what we call database provenance.

This talk presents an approach to database provenance that is based on two observations. First, provenance is a kind of annotation, and we can develop a general approach to annotation propagation that also covers other applications, for example to uncertainty and access control. In fact, provenance turns out to be the most general kind of such annotation, in a precise and practically useful sense. Second, the propagation of annotation through a broad class of transformations relies on just two operations: one when annotations are jointly used and one when they are used alternatively. This leads to annotations forming a specific algebraic structure, a commutative semiring.

The semiring approach works for annotating tuples, field values and attributes in standard relations, in nested relations (complex values), and for annotating nodes in (unordered) XML. It works for transformations expressed in the positive fragment of relational algebra, nested relational calculus, unordered XQuery, as well as for Datalog, GLAV schema mappings, and tgd constraints. Specific semirings correspond to earlier approaches to provenance, while others correspond to forms of uncertainty, trust, cost, and access control.

This is joint work with J.N. Foster, T.J. Green, Z. Ives, and G. Karvounarakis, done in part within the frameworks of the Orchestra and pPOD projects.
ABSTRACT
Ontologies and ontology based systems are becoming increasingly important in meeting the demand for more powerful and flexible information systems. Requirements for such systems include the need to deal with incomplete and semi-structured information, to integrate information from heterogeneous sources, to employ richer and more flexible schemas, and for query answers to reflect both knowledge and data. Provision of such enhanced capabilities must, however, be in addition to, and not instead of, the well-established features of existing database systems, in particular their robust scalability. Achieving this is, of course, extremely challenging. In this talk I will present some recent research efforts that tackle this problem, including investigations of tractable fragments, new algorithmic techniques, new optimisations and the exploitation of relational database technology.