2013 IEEE 29th International Conference on Data Engineering

(ICDE 2013)

Brisbane, Australia
8 - 12 April 2013

Pages 697-1379
Research Session 20: Trees and XML

Ontology-Based Subgraph Querying  697
(Yinghui Wu, Shengqi Yang, Xifeng Yan)

Stratification Driven Placement of Complex Data: A Framework for Distributed Data Analytics  709
(Ye Wang, Srinivasan Parthasarathy, P. Sadayappan)

Optimizing Approximations of DNF Query Lineage in Probabilistic XML  721
(Asma Souihli, Pierre Senellart)
Research Session 21: Security and Privacy

Secure Nearest Neighbor Revisited 733
(Bin Yao, Feifei Li, Xiaokui Xiao)

Accurate and Efficient Private Release of Datacubes and Contingency Tables 745
(Grigory Yaroslavtsev, Graham Cormode, Cecilia M. Procopiuc, Divesh Srivastava)

Differentially Private Grids for Geospatial Data 757
(Wahbeh Qardaji, Weining Yang, Ninghui Li)

Research Session 22: Randomized Algorithms for Graphs

Faster Random Walks by Rewiring Online Social Networks On-the-Fly 769
(Zhuojie Zhou, Nan Zhang, Zhiguo Gong, Gautam Das)

Sampling Node Pairs Over Large Graphs 781
(Pinghui Wang, Junzhou Zhao, John C.S. Lui, Don Towsley, Xiaohong Guan)

Link Prediction Across Networks by Biased Cross-Network Sampling 793
(Guo-Jun Qi, Charu C. Aggarwal, Thomas Huang)

Research Session 23: Distributed Data Processing

Interval Indexing and Querying on Key-Value Cloud Stores 805
(George Sfakianakis, Ioannis Patlakas, Nikos Ntarmos, Peter Triantafillou)

Robust Distributed Stream Processing 817
(Chuan Lei, Elke A. Rundensteiner, Joshua D. Guttman)
Research Session 24: Data Mining II

Learning to Rank from Distant Supervision: Exploiting Noisy Redundancy for Relational Entity Search 829
(Mianwei Zhou, Hongning Wang, Kevin Chen-Chuan Chang)

Affinity: Efficiently Querying Statistical Measures on Time-Series Data 841
(Saket Sathe, Karl Aberer)

Forecasting the Data Cube: A Model Configuration Advisor for Multi-Dimensional Data Sets 853
(Ulrike Fischer, Christopher Schildt, Claudio Hartmann, Wolfgang Lehner)

Research Session 25: Lineage and Provenance

SubZero: A Fine-Grained Lineage System for Scientific Databases 865
(Eugene Wu, Samuel Madden, Michael Stonebraker)

Logical Provenance in Data-Oriented Workflows 877
(Robert Ikeda, Akash Das Sarma, Jennifer Widom)

Revision Provenance in Text Documents of Asynchronous Collaboration 889
(Jing Zhang, H.V. Jagadish)

Research Session 26: Similarity Search

Inverted Linear Quadtree: Efficient Top K Spatial Keyword Search 901
(Chengyuan Zhang, Ying Zhang, Wenjie Zhang, Xuemin Lin)

Similarity Query Processing for Probabilistic Sets 913
(Ming Gao, Cheqing Jin, Wei Wang, Xuemin Lin, Aoying Zhou)

Top-k String Similarity Search with Edit-Distance Constraints 925
(Dong Deng, Guoliang Li, Jianhua Feng, Wen-Syan Li)
<table>
<thead>
<tr>
<th>Research Session 27: Shortest and Direct Query</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>On Shortest Unique Substring Queries</strong>  937</td>
</tr>
<tr>
<td>Jian Pei, Wush Chi-Hsuan Wu, Mi-Yen Yeh</td>
</tr>
<tr>
<td><strong>Engineering Generalized Shortest Path Queries</strong>  949</td>
</tr>
<tr>
<td>Michael N. Rice, Vassilis J. Tsotras</td>
</tr>
<tr>
<td><strong>Efficient Direct Search on Compressed Genomic Data</strong>  961</td>
</tr>
<tr>
<td>Xiao Chun Yang, Bin Wang, Chen Li, Jiaying Wang, Xiaohui Xie</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Research Session 28: Skyline and Snapshot Query</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>On Answering Why-Not Questions in Reverse Skyline Queries</strong>  973</td>
</tr>
<tr>
<td>Md. Saiful Islam, Rui Zhou, Chengfei Liu</td>
</tr>
<tr>
<td><strong>Layered Processing of Skyline-Window-Join (SWJ) Queries Using Iteration-Fabric</strong>  985</td>
</tr>
<tr>
<td>Mithila Nagendra, K. Selçuk Candan</td>
</tr>
<tr>
<td><strong>Efficient Snapshot Retrieval Over Historical Graph Data</strong>  997</td>
</tr>
<tr>
<td>Udayan Khurana, Amol Deshpande</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Research Session 29: Large Graph Indexing</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>FERRARI: Flexible and Efficient Reachability Range Assignment for Graph Indexing</strong>  1009</td>
</tr>
<tr>
<td>Stephan Seufert, Avishek Anand, Srikanta Bedathur, Gerhard Weikum</td>
</tr>
<tr>
<td><strong>glIceberg: Towards Iceberg Analysis in Large Graphs</strong>  1021</td>
</tr>
<tr>
<td>Nan Li, Ziyu Guan, Lijie Ren, Jian Wu, Jiawei Han, Xifeng Yan</td>
</tr>
<tr>
<td><strong>Top-k Graph Pattern Matching Over Large Graphs</strong>  1033</td>
</tr>
<tr>
<td>Jiefeng Cheng, Xianggang Zeng, Jeffrey Xu Yu</td>
</tr>
</tbody>
</table>
ICDE Conference 2013 Table of Contents

Research Session 30: Web Data

**Breaking the Top-k Barrier of Hidden Web Databases** 1045  
*Saravanan Thirumuruganathan, Nan Zhang, Gautam Das*

**Automatic Extraction of Top-k Lists from the Web** 1057  
*Zhixian Zhang, Kenny Q. Zhu, Habun Wang, Hongsong Li*

**Finding Interesting Correlations with Conditional Heavy Hitters** 1069  
*Katsiaryna Mirylenka, Themis Palpanas, Graham Cormode, Divesh Srivastava*

Research Session 31: Query Optimization

**Predicting Query Execution Time: Are Optimizer Cost Models Really Unusable?** 1081  
*Wentao Wu, Yun Chi, Shenghuo Zhu, Junichi Tatemura, Hakan Hacigümüs, Jeffrey F. Naughton*

**Query Optimization for Differentially Private Data Management Systems** 1093  
*Shangfu Peng, Yin Yang, Zhenjie Zhang, Marianne Winslett, Yong Yu*

**Top Down Plan Generation: From Theory to Practice** 1105  
*Pit Fender, Guido Moerkotte*

Research Session 32: Data Storage

**TBF: A Memory-Efficient Replacement Policy for Flash-Based Caches** 1117  
*Cristian Ungureanu, Biplob Debnath, Stephen Rago, Akshat Aranya*

**Fast Peak-to-Peak Behavior with SSD Buffer Pool** 1129  
*Jaeyoung Do, Donghui Zhang, Jignesh M. Patel, David J. DeWitt*

**SELECT Triggers for Data Auditing** 1141  
*Daniel Fabbri, Ravi Ramamurthy, Raghav Kaushik*
ICDE Conference 2013 Table of Contents

Industry Session 1

**Data Services for E-Tailers: Leveraging Web Search Engine Assets** 1153  
(Tao Cheng, Kaushik Chakrabarti, Surajit Chaudhuri, Vivek Narasayya, Manoj Syamala)

**SAP HANA Distributed In-Memory Database System: Transaction, Session, and Metadata Management** 1165  

Industry Session 2

**HEMS: Managing the Lifecycle and Complexity of Hybrid Analytic Data Flows** 1174  
(Alkis Simitsis, Kevin Wilkinson, Umeshwar Dayal, Meichun Hsu)

**KuaFu: Closing the Parallelism Gap in Database Replication** 1186  
(Chuntao Hong, Dong Zhou, Mao Yang, Carbo Kuo, Lintao Zhang, Lidong Zhou)

**Materialization Strategies in the Vertica Analytic Database: Lessons Learned** 1196  
(Lakshmikant Shrintvas, Sreenath Bodagala, Ramakrishna Varadarajan, Ariel Cary, Vivek Bharathan, Chuck Bear)

Industry Session 3

**Pipe Failure Prediction: A Data Mining Method** 1208  
(Rui Wang, Weishan Dong, Yu Wang, Ke Tang, Xin Yao)

**SASH: Enabling Continuous Incremental Analytic Workflows on Hadoop** 1219  
(Manish Sethi, Narendran Sachindran, Sriram Raghavan)

**Automating Pattern Discovery for Rule Based Data Standardization Systems** 1231  
(Snigdha Chaturvedi, Hima Prasad K., Tanveer A. Faruque, Bhupesh S. Chawda, L. Venkata Subramaniam, Raghu Krishnapuram)
Seminars

Machine Learning on Big Data  1242
(Tyson Condie, Paul Mineiro, Neoklis Polyzotis, Markus Weimer)

Big Data Integration  1245
(Xin Luna Dong, Divesh Srivastava)

Workload Management for Big Data Analytics  1249
(Ashraf Aboulnaga, Shivnath Babu)

Knowledge Harvesting from Text and Web Sources  1250
(Fabian Suchanek, Gerhard Weikum)

Sorting in Space: Multidimensional, Spatial, and Metric Data Structures for Applications in Spatial Databases, Geographic Information Systems (GIS), and Location-Based Services  1254
(Hanan Samet)

Triples in the Clouds  1258
(Zoi Kaoudi, Ioana Manolescu)

Querying Encrypted Data  1262
(Arvind Arasu, Ken Eguro, Raghav Kaushik, Ravi Ramamurthy)

Shallow Information Extraction for the Knowledge Web  1264
(Denilson Barbosa, Haixun Wang, Cong Yu)

Secure and Privacy-Preserving Database Services in the Cloud  1268
(Divyakant Agrawal, Amr El Abbadi, Shiyuan Wang)
ICDE Conference 2013 Table of Contents

Demo Groups 1 and 2

Twitter+: Build Personalized Newspaper for Twitter 1272
(Chen Liu, Anthony K.H. Tung)

A Generic Database Benchmarking Service 1276
(Martin Kaufmann, Peter M. Fischer, Donald Kossmann, Norman May)

Aeolus: An Optimizer for Distributed Intra-Node-Parallel Streaming Systems 1280
(Matthias J. Sax, Malu Castellanos, Qiming Chen, Meichun Hsu)

Crowd-Answering System via Microblogging 1284
(Xianke Zhou, Ke Chen, Sai Wu, Bingbing Zhang)

With a Little Help from My Friends 1288
(Arnab Nandi, Stelios Paparizos, John C. Shafer, Rakesh Agrawal)

Peeking into the Optimization of Data Flow Programs with MapReduce-Style UDFs 1292
(Fabian Hueske, Mathias Peters, Aljoscha Krettek, Matthias Ringwald, Kostas Tzoumas, Volker Markl, Johann-Christoph Freytag)

Very Fast Estimation for Result and Accuracy of Big Data Analytics: The ERL System 1296
(Nikolay Laptev, Kai Zeng, Carlo Zaniolo)

Road Network Mix-Zones for Anonymous Location Based Services 1300
(Balaji Palanisamy, Sindhuja Ravichandran, Ling Liu, Binh Han, Kisung Lee, Calton Pu)

Query Time Scaling of Attribute Values in Interval Timestamped Databases 1304
(Anton Dignös, Michael Böhlen, Johann Gamper)

Extracting Interesting Related Context-Dependent Concepts from Social Media Streams Using Temporal Distributions 1308
(Craig P. Sayers, Meichun Hsu)

Demo Groups 1 and 2 continues next page ...
Demo Groups 1 and 2 continued...

VERDICT: Privacy-Preserving Authentication of Range Queries in Location-Based Services 1312
(Haibo Hu, Qian Chen, Jianting Xu)

ExpFinder: Finding Experts by Graph Pattern Matching 1316
(Wenfei Fan, Xin Wang, Yinghui Wu)

Tajo: A Distributed Data Warehouse System on Large Clusters 1320
(Hyunsik Choi, Jihoon Son, Haemi Yang, Hyoseok Ryu, Byungnam Lim, Soohyung Kim, Yon Dohn Chung)

Demo Groups 3 and 4

ngora: An Integration System for Probabilistic Data 1324
(Dan Olteanu, Lampros Papageorgiou, Sebastiaan J. van Schaik)

Complex Pattern Matching in Complex Structures: The XSeq Approach 1328
(Kai Zeng, Mohan Yang, Barzan Mozafari, Carlo Zaniolo)

T-Music: A Melody Composer Based on Frequent Pattern Mining 1332
(Cheng Long, Raymond Chi-Wing Wong, Raymond Ka Wai Sze)

SHARE: Secure Information Sharing Framework for Emergency Management 1336
(Barbara Carminati, Elena Ferrari, Michele Guglielmi)

KORS: Keyword-Aware Optimal Route Search System 1340
(Xin Cao, List Chen, Gao Cong, Jihong Guan, Nhan-Tue Phan, Xiaokui Xiao)

CrowdPlanr: Planning Made Easy with Crowd 1344
(Ilia Lotosh, Tova Milo, Slava Novgorodov)

ASVTDECTOR: A Practical Near Duplicate Video Retrieval System 1348
(Xiangmin Zhou, Lei Chen)

Demo Groups 3 and 4 continues next page...
Demo Groups 3 and 4 continued ...

Yumilnt — A Deep Web Integration System for Local Search Engines for Geo-Referenced Objects 1352
(Eduard Dragut, B.P. Beirne, A. Neyestani, B. Atassi, Clement Yu, Bhaskar DasGupta, Weiyi Meng)

A Demonstration of the G* Graph Database System 1356
(Sean R. Spillane, Jeremy Birnbaum, Daniel Bokser, Daniel Kemp, Alan Labouseur, Paul W. Olsen Jr., Jayadevan Vijayan, Jeong-Hyon Hwang, Jun-Weon Yoon)

RECODS: Replica Consistency-On-Demand Store 1360
(Yuqing Zhu, Philip S. Yu, Jianmin Wang)

SODIT: An Innovative System for Outlier Detection Using Multiple Localized Thresholding and Interactive Feedback 1364
(Ji Zhang, Hua Wang, Xiaohui Tao, Lili Sun)

COLA: A Cloud-Based System for Online Aggregation 1368
(Yantao Gan, Xiaofeng Meng, Yingjie Shi)

RoadAlarm: A Spatial Alarm System on Road Networks 1372
(Kisung Lee, Emre Yigitoglu, Ling Liu, Binh Han, Balaji Palanisamy, Calton Pu)

A Real-Time Abnormality Detection System for Intensive Care Management 1376
(Guangyan Huang, Jing He, Jie Cao, Zhi Qiao, Michael Steyn, Kersi Taraporewalla)