

Opher Etzion
Tsvi Kuflik
Amihai Motro (Eds.)

LNCS 4032

Next Generation Information Technologies and Systems

6th International Conference, NGITS 2006
Kibbutz Shefayim, Israel, July 2006
Proceedings

TP3-53
N576
200.6
Opher Etzion Tsvi Kuflik
Amihai Motro (Eds.)

Next Generation Information Technologies and Systems

6th International Conference, NGITS 2006
Kibbutz Shefayim, Israel, July 4-6, 2006
Proceedings



Springer



E200603631

Volume Editors

Opher Etzion
IBM Haifa Labs
Haifa University Campus, Haifa 31905, Israel
E-mail: opher@il.ibm.com

Tsvi Kuflik
University of Haifa
MIS Department
Mount Carmel, Haifa, 31905, Israel
E-mail: tsvikak@is.haifa.ac.il

Amihai Motro
George Mason University
ISE Department
Fairfax, VA 22015, USA
E-mail: ami@gmu.edu

Library of Congress Control Number: 2006928068

CR Subject Classification (1998): H.4, H.3, H.5, H.2, D.2.12, C.2.4

LNCS Sublibrary: SL 3 – Information Systems and Application, incl. Internet/Web and HCI

ISSN 0302-9743
ISBN-10 3-540-35472-7 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-35472-7 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media
springer.com

© Springer-Verlag Berlin Heidelberg 2006
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11780991 06/3142 5 4 3 2 1 0

Commenced Publication in 1973

Founding and Former Series Editors:

Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

David Hutchison

Lancaster University, UK

Takeo Kanade

Carnegie Mellon University, Pittsburgh, PA, USA

Josef Kittler

University of Surrey, Guildford, UK

Jon M. Kleinberg

Cornell University, Ithaca, NY, USA

Friedemann Mattern

ETH Zurich, Switzerland

John C. Mitchell

Stanford University, CA, USA

Moni Naor

Weizmann Institute of Science, Rehovot, Israel

Oscar Nierstrasz

University of Bern, Switzerland

C. Pandu Rangan

Indian Institute of Technology, Madras, India

Bernhard Steffen

University of Dortmund, Germany

Madhu Sudan

Massachusetts Institute of Technology, MA, USA

Demetri Terzopoulos

University of California, Los Angeles, CA, USA

Doug Tygar

University of California, Berkeley, CA, USA

Moshe Y. Vardi

Rice University, Houston, TX, USA

Gerhard Weikum

Max-Planck Institute of Computer Science, Saarbruecken, Germany

Lecture Notes in Computer Science

For information about Vols. 1–3970

please contact your bookseller or Springer

Vol. 4067: D. Thomas (Ed.), ECOOP 2006 – Object-Oriented Programming. XIV, 527 pages. 2006.

Vol. 4063: I. Gorton, G.T. Heineman, I. Crnkovic, H.W. Schmidt, J.A. Stafford, C.A. Szyperski, K. Wallnau (Eds.), Component-Based Software Engineering. XI, 394 pages. 2006.

Vol. 4060: K. Futatsugi, J.-P. Jouannaud, J. Meseguer (Eds.), Algebra, Meaning and Computation. XXXVIII, 643 pages. 2006.

Vol. 4059: L. Arge, R. Freivalds (Eds.), Algorithm Theory – SWAT 2006. XII, 436 pages. 2006.

Vol. 4058: L.M. Batten, R. Safavi-Naini (Eds.), Information Security and Privacy. XII, 446 pages. 2006.

Vol. 4057: J.P.W. Pluim, B. Likar, F.A. Gerritsen (Eds.), Biomedical Image Registration. XII, 324 pages. 2006.

Vol. 4056: P. Flocchini, L. Gasieniec (Eds.), Structural Information and Communication Complexity. X, 357 pages. 2006.

Vol. 4055: J. Lee, J. Shim, S.-g. Lee, C. Bussler, S. Shim (Eds.), Data Engineering Issues in E-Commerce and Services. IX, 290 pages. 2006.

Vol. 4054: A. Horváth, M. Telek (Eds.), Formal Methods and Stochastic Models for Performance Evaluation. VIII, 239 pages. 2006.

Vol. 4053: M. Ikeda, K.D. Ashley, T.-W. Chan (Eds.), Intelligent Tutoring Systems. XXVI, 821 pages. 2006.

Vol. 4048: L. Goble, J.-J.C. Meyer (Eds.), Deontic Logic and Artificial Normative Systems. X, 273 pages. 2006. (Sublibrary LNAI).

Vol. 4046: S.M. Astley, M. Brady, C. Rose, R. Zwiggelaar (Eds.), Digital Mammography. XVI, 654 pages. 2006.

Vol. 4045: D. Barker-Plummer, R. Cox, N. Swoboda (Eds.), Diagrammatic Representation and Inference. XII, 301 pages. 2006. (Sublibrary LNAI).

Vol. 4044: P. Abrahamsson, M. Marchesi, G. Succi (Eds.), Extreme Programming and Agile Processes in Software Engineering. XII, 230 pages. 2006.

Vol. 4043: A.S. Atzeni, A. Lioy (Eds.), Public Key Infrastructure. XI, 261 pages. 2006.

Vol. 4041: S.-W. Cheng, C.K. Poon (Eds.), Algorithmic Aspects in Information and Management. XI, 395 pages. 2006.

Vol. 4040: R. Reulke, U. Eckardt, B. Flach, U. Knauer, K. Polthier (Eds.), Combinatorial Image Analysis. XII, 482 pages. 2006.

Vol. 4039: M. Morisio (Ed.), Reuse of Off-the-Shelf Components. XIII, 444 pages. 2006.

Vol. 4038: P. Ciancarini, H. Wiklicky (Eds.), Coordination Models and Languages. VIII, 299 pages. 2006.

Vol. 4037: R. Gorrieri, H. Wehrheim (Eds.), Formal Methods for Open Object-Based Distributed Systems. XVII, 474 pages. 2006.

Vol. 4036: O. H. Ibarra, Z. Dang (Eds.), Developments in Language Theory. XII, 456 pages. 2006.

Vol. 4035: H.-P. Seidel, T. Nishita, Q. Peng (Eds.), Advances in Computer Graphics. XX, 771 pages. 2006.

Vol. 4034: J. Münch, M. Vierimaa (Eds.), Product-Focused Software Process Improvement. XVII, 474 pages. 2006.

Vol. 4033: B. Stiller, P. Reichl, B. Tuffin (Eds.), Performability Has its Price. X, 103 pages. 2006.

Vol. 4032: O. Etzion, T. Kuflik, A. Motro (Eds.), Next Generation Information Technologies and Systems. XIII, 366 pages. 2006.

Vol. 4031: M. Ali, R. Dapoigny (Eds.), Advances in Applied Artificial Intelligence. XXIII, 1353 pages. 2006. (Sublibrary LNAI).

Vol. 4027: H.L. Larsen, G. Pasi, D. Ortiz-Arroyo, T. Andreassen, H. Christiansen (Eds.), Flexible Query Answering Systems. XVIII, 714 pages. 2006. (Sublibrary LNAI).

Vol. 4026: P.B. Gibbons, T. Abdelzaher, J. Aspnes, R. Rao (Eds.), Distributed Computing in Sensor Systems. XIV, 566 pages. 2006.

Vol. 4025: F. Eliassen, A. Montresor (Eds.), Distributed Applications and Interoperable Systems. XI, 355 pages. 2006.

Vol. 4024: S. Donatelli, P. S. Thiagarajan (Eds.), Petri Nets and Other Models of Concurrency - ICATPN 2006. XI, 441 pages. 2006.

Vol. 4021: E. André, L. Dybkjær, W. Minker, H. Neumann, M. Weber (Eds.), Perception and Interactive Technologies. XI, 217 pages. 2006. (Sublibrary LNAI).

Vol. 4020: A. Bredendfeld, A. Jacoff, I. Noda, Y. Takahashi (Eds.), RoboCup 2005: Robot Soccer World Cup IX. XVII, 727 pages. 2006. (Sublibrary LNAI).

Vol. 4019: M. Johnson, V. Vene (Eds.), Algebraic Methodology and Software Technology. XI, 389 pages. 2006.

Vol. 4018: V. Wade, H. Ashman, B. Smyth (Eds.), Adaptive Hypermedia and Adaptive Web-Based Systems. XVI, 474 pages. 2006.

Vol. 4016: J.X. Yu, M. Kitsuregawa, H.V. Leong (Eds.), Advances in Web-Age Information Management. XVII, 606 pages. 2006.

Vol. 4014: T. Uustalu (Ed.), Mathematics of Program Construction. X, 455 pages. 2006.

- Vol. 4013: L. Lamontagne, M. Marchand (Eds.), *Advances in Artificial Intelligence*. XIII, 564 pages. 2006. (Sublibrary LNAI).
- Vol. 4012: T. Washio, A. Sakurai, K. Nakajima, H. Takeda, S. Tojo, M. Yokoo (Eds.), *New Frontiers in Artificial Intelligence*. XIII, 484 pages. 2006. (Sublibrary LNAI).
- Vol. 4011: Y. Sure, J. Domingue (Eds.), *The Semantic Web: Research and Applications*. XIX, 726 pages. 2006.
- Vol. 4010: S. Dunne, B. Stoddart (Eds.), *Unifying Theories of Programming*. VIII, 257 pages. 2006.
- Vol. 4009: M. Lewenstein, G. Valiente (Eds.), *Combinatorial Pattern Matching*. XII, 414 pages. 2006.
- Vol. 4007: C. Álvarez, M. Serna (Eds.), *Experimental Algorithms*. XI, 329 pages. 2006.
- Vol. 4006: L.M. Pinho, M. González Harbour (Eds.), *Reliable Software Technologies – Ada-Europe 2006*. XII, 241 pages. 2006.
- Vol. 4005: G. Lugosi, H.U. Simon (Eds.), *Learning Theory*. XI, 656 pages. 2006. (Sublibrary LNAI).
- Vol. 4004: S. Vaudenay (Ed.), *Advances in Cryptology – EUROCRYPT 2006*. XIV, 613 pages. 2006.
- Vol. 4003: Y. Koucheryavy, J. Harju, V.B. Iversen (Eds.), *Next Generation Teletraffic and Wired/Wireless Advanced Networking*. XVI, 582 pages. 2006.
- Vol. 4001: E. Dubois, K. Pohl (Eds.), *Advanced Information Systems Engineering*. XVI, 560 pages. 2006.
- Vol. 3999: C. Kop, G. Fliedl, H.C. Mayr, E. Métais (Eds.), *Natural Language Processing and Information Systems*. XIII, 227 pages. 2006.
- Vol. 3998: T. Calamoneri, I. Finocchi, G.F. Italiano (Eds.), *Algorithms and Complexity*. XII, 394 pages. 2006.
- Vol. 3997: W. Grieskamp, C. Weise (Eds.), *Formal Approaches to Software Testing*. XII, 219 pages. 2006.
- Vol. 3996: A. Keller, J.-P. Martin-Flatin (Eds.), *Self-Managed Networks, Systems, and Services*. X, 185 pages. 2006.
- Vol. 3995: G. Müller (Ed.), *Emerging Trends in Information and Communication Security*. XX, 524 pages. 2006.
- Vol. 3994: V.N. Alexandrov, G.D. van Albada, P.M.A. Sloot, J. Dongarra (Eds.), *Computational Science – ICCS 2006, Part IV*. XXXV, 1096 pages. 2006.
- Vol. 3993: V.N. Alexandrov, G.D. van Albada, P.M.A. Sloot, J. Dongarra (Eds.), *Computational Science – ICCS 2006, Part III*. XXXVI, 1136 pages. 2006.
- Vol. 3992: V.N. Alexandrov, G.D. van Albada, P.M.A. Sloot, J. Dongarra (Eds.), *Computational Science – ICCS 2006, Part II*. XXXV, 1122 pages. 2006.
- Vol. 3991: V.N. Alexandrov, G.D. van Albada, P.M.A. Sloot, J. Dongarra (Eds.), *Computational Science – ICCS 2006, Part I*. LXXXI, 1096 pages. 2006.
- Vol. 3990: J. C. Beck, B.M. Smith (Eds.), *Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems*. X, 301 pages. 2006.
- Vol. 3989: J. Zhou, M. Yung, F. Bao, *Applied Cryptography and Network Security*. XIV, 488 pages. 2006.
- Vol. 3988: A. Beckmann, U. Berger, B. Löwe, J.V. Tucker (Eds.), *Logical Approaches to Computational Barriers*. XV, 608 pages. 2006.
- Vol. 3987: M. Hazas, J. Krumm, T. Strang (Eds.), *Location- and Context-Awareness*. X, 289 pages. 2006.
- Vol. 3986: K. Stølen, W.H. Winsborough, F. Martinelli, F. Massacci (Eds.), *Trust Management*. XIV, 474 pages. 2006.
- Vol. 3984: M. Gavrilova, O. Gervasi, V. Kumar, C.J. K. Tan, D. Taniar, A. Laganà, Y. Mun, H. Choo (Eds.), *Computational Science and Its Applications – ICCSA 2006, Part V*. XXV, 1045 pages. 2006.
- Vol. 3983: M. Gavrilova, O. Gervasi, V. Kumar, C.J. K. Tan, D. Taniar, A. Laganà, Y. Mun, H. Choo (Eds.), *Computational Science and Its Applications – ICCSA 2006, Part IV*. XXVI, 1191 pages. 2006.
- Vol. 3982: M. Gavrilova, O. Gervasi, V. Kumar, C.J. K. Tan, D. Taniar, A. Laganà, Y. Mun, H. Choo (Eds.), *Computational Science and Its Applications – ICCSA 2006, Part III*. XXV, 1243 pages. 2006.
- Vol. 3981: M. Gavrilova, O. Gervasi, V. Kumar, C.J. K. Tan, D. Taniar, A. Laganà, Y. Mun, H. Choo (Eds.), *Computational Science and Its Applications – ICCSA 2006, Part II*. XXVI, 1255 pages. 2006.
- Vol. 3980: M. Gavrilova, O. Gervasi, V. Kumar, C.J. K. Tan, D. Taniar, A. Laganà, Y. Mun, H. Choo (Eds.), *Computational Science and Its Applications – ICCSA 2006, Part I*. LXXV, 1199 pages. 2006.
- Vol. 3979: T.S. Huang, N. Sebe, M.S. Lew, V. Pavlović, M. Kölsch, A. Galata, B. Kisačanin (Eds.), *Computer Vision in Human-Computer Interaction*. XII, 121 pages. 2006.
- Vol. 3978: B. Hnich, M. Carlsson, F. Fages, F. Rossi (Eds.), *Recent Advances in Constraints*. VIII, 179 pages. 2006. (Sublibrary LNAI).
- Vol. 3977: N. Fuhr, M. Lalmas, S. Malik, G. Kazai (Eds.), *Advances in XML Information Retrieval and Evaluation*. XII, 556 pages. 2006.
- Vol. 3976: F. Boavida, T. Plagemann, B. Stiller, C. Westphal, E. Monteiro (Eds.), *NETWORKING 2006. Networking Technologies, Services, and Protocols; Performance of Computer and Communication Networks; Mobile and Wireless Communications Systems*. XXVI, 1276 pages. 2006.
- Vol. 3975: S. Mehrotra, D.D. Zeng, H. Chen, B. Thuraisingham, F.-Y. Wang (Eds.), *Intelligence and Security Informatics*. XXII, 772 pages. 2006.
- Vol. 3973: J. Wang, Z. Yi, J.M. Zurada, B.-L. Lu, H. Yin (Eds.), *Advances in Neural Networks – ISNN 2006, Part III*. XXIX, 1402 pages. 2006.
- Vol. 3972: J. Wang, Z. Yi, J.M. Zurada, B.-L. Lu, H. Yin (Eds.), *Advances in Neural Networks – ISNN 2006, Part II*. XXVII, 1444 pages. 2006.
- Vol. 3971: J. Wang, Z. Yi, J.M. Zurada, B.-L. Lu, H. Yin (Eds.), *Advances in Neural Networks – ISNN 2006, Part I*. LXVII, 1442 pages. 2006.

Preface

Information technology is a rapidly changing field in which researchers and developers must continuously set their vision on the next generation of technologies and the systems that they enable. Suitably, Next-Generation Information Technologies and Systems (NGITS) is an ongoing series of workshops that provides a forum for presenting and discussing the latest advances in information technology. NGITS took are international events held in Israel; previous workshops took place in 1993, 1995, 1997, 1999 and 2002.

The call for papers for the 2006 workshop was answered by 138 papers on a very diverse range of subjects, thus presenting a considerable challenge to the Program Committee. Each of these papers received professional reviewing from at least three experts, and eventually 32 papers (less than 25%) were selected for presentation at the workshop and inclusion in these proceedings. Of these, 28 are full-length papers and 4 are short papers. In addition, several research works were selected to display “poster” presentations, and a number of research projects exhibited demonstrations. Finally, the workshop featured three keynote lectures by notable experts. The selected papers may be classified roughly in ten broad areas:

- Information systems development
- Distributed systems
- Semi-structured data
- Data mining and agent-oriented computing
- User-oriented design
- Frameworks, models and taxonomies
- Simulation and incremental computing
- Information integration
- Security and privacy
- Next-generation applications

This event is the culmination of efforts by many talented and dedicated individuals. We are pleased to extend our thanks to the authors of all submitted papers, the members of the Steering Committee, the members of the Program Committee, and the external reviewers. Special thanks are due to Dan Berry, for his discrete handling of papers that presented conflicts of interest. Many thanks are also due to Nilly Schnapp for local organization and logistics, and to Lior Shmueli for managing the website and all other things technical. Finally, we are pleased to acknowledge the support of our institutional sponsors: The Faculty of Social Sciences and the MIS Department at the University of Haifa, and the IBM Research Lab in Haifa.

July 2006

Tsvi Kuflik
General Chair
Opher Etzion and Amihai Motro
Program Co-Chairs

Organization

General Chair

Tsvi Kuflik

Steering Committee

Opher Etzion
Amihai Motro
Peretz Shoval

Avigdor Gal
Ron Pinter
Shalom Tsur

Alon Halevy
Avi Silberschatz

Program Committee Chairs

Opher Etzion

Amihai Motro

Program Committee Vice Chair

Dan Berry

Program Committee

Serge Abiteboul
Eugene Agichtein
Iris Berger
Andrei Broder
Vijay Dialani
Carlotta Domeniconi
Ehud Gudes
David Konopnicki
Tova Milo
George Papadopoulos
Naphtali Rishé
Yuval Shahar
Pnina Soffer
Yair Wand

Nabil Adam
Irit Askira-Gelman
Elisa Bertino
Jen-Yao Chung
AnHai Doan
Ophir Frieder
Alfons Kemper
Manolis Koubarakis
Mukesh Mohania
Norman Paton
Doron Rotem
Oded Shmueli
Carlo Strapparava
Ouri Wolfson

Hamideh Afsarmanesh
Catriel Beeri
Martin Bichler
Alessandro D'Atri
Asuman Dogac
Paolo Giorgini
Larry Kerschberg
Yossi Matias
Felix Naumann
Francesco Ricci
Steve Schach
Charles A. Shoniregun
Bernhard Thalheim
Carlo Zaniolo

Additional Reviewers

Aybar C. Acar	Benni Arazi	Yijian Bai
Khalid Belhajjame	Shlomo Berkovsky	Alexander Bilke
Jens Bleiholder	Vlaidmir Braverman	Neslihan Bulut
Nunzio Casalino	Hu Cao	Bogdan Cautis
Dario Cavada	Kuo-Ming Chao	Mariangela
Contenti	Bonaventura Coppola	Luiz Marcio Cysneiros
Marco De Marco	Pedro DeRose	Chrysanne DiMarco
Nurit Galoz	Ozgur Gulderen	Jeff Heard
Yildiray Kabak	Zoi Kaoudi	Gokce Banu Laleci
Yoonkyong Lee	Yinsheng Li	Jessica Lin
Juhong Liu	Sean Luke	Steven Lynden
Robert McCann	Iris Miliaraki	Oleg Missikoff
Simon Msanjila	Tuncay Namli	Alper Okcan
Mehmet Olduz	Frank Olken	Jeffery Orchard
Umut Orhan	Ekow Otoo	Christoforos
Michael Pantazoglou	Nguyen Nhat Quang	Andrea Resca
Boris Rozenberg	Mayssam Sayyadian	Tayfun Sen
Pierre Senallart	Warren Shen	Paolo Spagnoletti
Arnon Sturm	Ibrahim Tasyurt	Hetal Thakkar
Aimilia Tzanavar	Ozgul Unal	Marcos Vieira
Ronen Waisenberg	Wensheng Wu	Bo Xu
Bojun Yan	Huabei Yin	Mustafa Yuksel
Xin Zhou		

Local Arrangements Chair

Nilly Schnapp

Website Manager

Lior Shmueli

Table of Contents

Full Papers

Information Integration

Efficiently Updating Cost Repository Values for Query Optimization on Web Data Sources in a Mediator/Wrapper Environment <i>Justo Hidalgo, Alberto Pan, Manuel Álvarez, Jaime Guerrero</i>	1
TupleRank: Ranking Discovered Content in Virtual Databases <i>Jacob Berlin, Amihai Motro</i>	13

Next-Generations Applications

Analysis of Queries Reaching SHIL on the Web - An Information System Providing Citizen Information <i>Gilad Ravid, Judit Bar-Ilan, Sheizaf Rafaeli, Shifra Baruchson-Arbib</i>	26
On Mediated Search of the United States Holocaust Memorial Museum Data <i>Jefferson Heard, Jordan Wilberding, Gideon Frieder, Ophir Frieder, David Grossman, Larry Kane</i>	38
A Repository of Services for the Government to Businesses Relationship <i>Daniele Barone, Gianluigi Viscusi, Carlo Batini, Paolo Naggari</i>	47

Information Systems Development

Towards Automatic Integration of Persistency Requirements in Enterprise-Systems – The Persistent-to-Persistent Patterns <i>Mira Balaban, Lior Limonad</i>	59
Consistency of UML Class Diagrams with Hierarchy Constraints <i>Mira Balaban, Azzam Maraee</i>	71
A Framework-Based Design for Supporting Availability of Layered Distributed Applications <i>Heung Seok Chae, Jaegeol Park, Jinwook Park, Sungho Ha, Joon-Sang Lee</i>	83

Security and Privacy

Web Application Security Gateway with Java Non-blocking IO
Zhenxing Luo, Nuermaimaiti Heilili, Dawei Xu, Chen Zhao, Zuoquan Lin 96

Microaggregation for Database and Location Privacy
Josep Domingo-Ferrer 106

Efficient Caching Strategies for Gnutella-Like Systems to Achieve Anonymity in Unstructured P2P File Sharing
Byung Ryong Kim, Ki Chang Kim 117

Semi-structured Data

Conjunctive Queries over DAGs
Royi Ronen, Oded Shmueli 129

Incrementally Computing Ordered Answers of Acyclic Conjunctive Queries
Benny Kimelfeld, Yehoshua Sagiv 141

Count-Constraints for Generating XML
Sara Cohen 153

Frameworks, Models and Taxonomies

A Data Model for Music Information Retrieval
Tamar Berman 165

A Taxonomy and Representation of Sources of Uncertainty in Active Systems
Segev Wasserkrug, Avigdor Gal, Opher Etzion 174

Analyzing Object-Oriented Design Patterns from an Object-Process Viewpoint
Galia Shlezinger, Iris Reinhartz-Berger, Dov Dori..... 186

Simulation and Incremental Computing

Modeling and Simulation of Grid Information Service
Xia Xie, Hai Jin, Jin Huang, Qin Zhang 198

Simulations of Distributed Service-Based Content Adaptation for Network Optimization <i>Eric Y. Cheng, Shuo Hung Jian</i>	210
---	-----

Distributed Systems

Second Order Snapshot-Log Relations: Supporting Multi-directional Database Replication Using Asynchronous Snapshot Replication <i>Yochai Ben-Chaim, Avigdor Gal</i>	221
The Replica Management for Wide-Area Distributed Computing Environments <i>Jaechun No, Chang Won Park, Sung Soon Park</i>	237
Developing Parallel Cell-Based Filtering Scheme Under Shared-Nothing Cluster-Based Architecture <i>Jae-Woo Chang, Tae-Woong Wang</i>	249

User Oriented Design

How Deep Should It Be? On the Optimality of Hierarchical Architectures <i>Amihai Motro, Alessandro D'Atri, Eli Gafni</i>	260
A Spreadsheet Client for Web Applications <i>Dirk Draheim, Peter Thiemann, Gerald Weber</i>	274
Dynamic Construction of User Defined Virtual Cubes <i>Dehui Zhang, Shaohua Tan, Dongqing Yang, Shiwei Tang, Xiuli Ma, Lizheng Jiang</i>	287

Data Mining and Agent-Oriented Computing

Automatic Discovery of Regular Expression Patterns Representing Negated Findings in Medical Narrative Reports <i>Roni Romano, Lior Rokach, Oded Maimon</i>	300
A Hybrid Method for Speeding SVM Training <i>Zhi-Qiang Zeng, Ji Gao, Hang Guo</i>	312
Design of Service Management System in OSGi Based Ubiquitous Computing Environment <i>Seungkeun Lee, Jeonghyun Lee</i>	321

Short Papers

Creating Consistent Diagnoses List for Developmental Disorders Using UMLS
Nuaman Asbeh, Mor Peleg, Mitchell Schertz, Tsvi Kuflik 333

Enhancing Domain Engineering with Aspect-Orientation
Iris Reinhartz-Berger, Alex Gold 337

An Agent-Oriented Approach to Semantic Web Services
In-Cheol Kim 341

Biometrics Authenticated Key Agreement Scheme
Eun-Jun Yoon, Kee-Young Yoo 345

Posters and Demonstrations

The Third Query Paradigm
Mikhail Gilula 350

τ -xSynopses – a System for Run-Time Management of XML Synopses
Natasha Drukh, Yariv Matia, Yossi Matias, Leon Portman 351

LTS: The List-Traversal Synopses System
Michael Furman, Yossi Matias, Ely Porat 353

The Promise and Challenge of Multidimensional Visualization
Alfred Inselberg 355

OPOSSUM: Bridging the Gap Between Web Services and the Semantic Web
Eran Toch, Iris Reinhartz-Berger, Avigdor Gal, Dov Dori 357

ProMo - A Scalable and Efficient Framework for Online Data Delivery
Haggai Roitman, Avigdor Gal, Louiqa Raschid 359

Invited Talks

Next Generation Enterprise IT Systems
Donald F. Ferguson 361

The Future of Web Search: From Information Retrieval to Information Supply
Andrei Broder 362

Is There Life After the Internet?	
<i>Yechiam Yemini</i>	363
Author Index	365

Efficiently Updating Cost Repository Values for Query Optimization on Web Data Sources in a Mediator/Wrapper Environment

Justo Hidalgo¹, Alberto Pan^{2,*}, Manuel Álvarez², and Jaime Guerrero³

¹ Denodo Technologies, Inc.
Madrid, Spain

jhidalgo@denodo.com

² Department of Information and Communications Technologies
University of A Coruña, Spain
{apan, mad}@udc.es

³ jaimeguerrero@wanadoo.es

Abstract. Optimizing accesses to sources in a mediator/wrapper environment is a critical need. Due to a variety of reasons, relational-based optimization techniques are of no use when having to handle HTTP-based web sources, so new approaches which take into account client/server communication costs must be devised. This paper describes a cost model that stores values from a complete set of web source-focused parameters obtained by the web wrappers, by using a novel updating technique that handles the values measured by the wrappers in previous query executions, and generates a new model instance in each new iteration with an efficient processing cost. This instance allows rapid value updates caused by changes of the server quality or bandwidth, so typical in this context. The results of these techniques are demonstrated both theoretically and by means of an implementation showing how performance improves in real-world web sources when compared to classical approaches.

1 Introduction

Virtual Databases, also called Mediators [14], allow the obtaining, unification and sampling of data residing in heterogeneous environments, as much because they are stored in different repositories, like their geographic location. Virtual Databases differ from standard databases because data are not really stored in the database, but remotely in several heterogeneous sources. Mediators form a middleware layer between the user and the sources to be integrated, providing a single point of access or data access layer to the underlying data. Each data source exposes its search capabilities through the set of parameters which can be used when querying it, along with the possible values and constraints for each parameter. At the time of executing a query in a mediator, a set of query plans is generated, each of which retains the same capabilities and restrictions, thus being able to respond to the query correctly, but with each of

* Alberto Pan's work was partially supported by the "Ramón y Cajal" programme of the Spanish Ministry of Education and Science.

them having different search methods of concrete sources, different sources, operator execution strategies, and so on. For example, imagine the case shown in Fig. 1, in which there are two electronic bookshop sources from which to search for works written by a particular author. Besides, there is a complementary set of web sources which store critics' reviews from books. We are interested on being able to query the system in order to retrieve all works written by an author, along with reviews made to that work from one of the sources. The basic query plan to be performed is therefore a relational algebra union operation between shop A and shop B, and a join between that union and the review site. But there can be different query plans depending, for example, in the type of join strategy to be performed, a hash join or a nested-loop join. Besides, the system could choose one review site or other depending on their reliability in a particular moment.

Figure 1 shows three web forms arranged horizontally. The first form is titled 'SHOP A' and contains two input fields: 'TITLE' and 'AUTHOR', each followed by an asterisk (*). Below these fields is a 'SUBMIT' button. The second form is titled 'SHOP B' and contains the same 'TITLE' and 'AUTHOR' fields with asterisks and a 'SUBMIT' button. The third form is titled 'REVIEW' and contains three input fields: 'TITLE', 'AUTHOR', and 'REVIEW', each followed by an asterisk (*). Below these fields is a 'SUBMIT' button. A note below the forms states '* ONE OR THE OTHER'.

Fig. 1. Electronic Bookshop Example

In a specific moment of time, only one of those plans will be optimal, that is to say, it will be the one that allows a faster and/or reliable execution, optimizing the resources to use. The importance of improving response times in web environments, where quality of service is not assured, has impelled this line of research in the last years.

The first approaches to query optimization were performed in the field of multidatabases [8], where cost estimation in relational databases uses statistics of the database to compute the cost of each operator in each plan. The authors propose a combination of a generic cost model with specific information of each wrapper. This concept is valuable when the information is very homogeneous. However, this approach is not directly applicable to virtual databases in the web environment, had mainly to the following reasons: (1) sources do not usually offer statistical information, and (2) costs of communications are not determined easily and can vary. Besides, remote communication costs are not considered as part of the model, which makes this idea nonviable in a web environment due to the lack of quality of service in the TCP/IP protocol which makes response times impossible to manage [15].

[9] treats this subject more extensively regarding the required group of parameters when taking into account communication costs among agents and remote sources. Nevertheless, cost propagation formulae provided are not complete enough for a web data source-enabled, industrial system.

Other investigations have offered different options. [3] focuses on coefficient estimation in a generic cost model; this approach is not useful enough with web data sources, since it relies on the remote databases to provide cost statistics.

A more interesting solution for our case is the one proposed in the Hermes project, specified in [1]; the cost model is evaluated from historical information, which can be used for later estimation for each execution plan cost. Hermes concept has been improved by other works such as [16], where a two-phase optimizer is depicted, or [7], where its research on dynamic adaptiveness starts from retrieved cost information.

Nonetheless, little research has been done neither on what kind of cost parameters should be stored in the cost repository from the wrappers when web sources are involved, nor on how these values retrieved in each measurement should be processed so as to better infer how the source will behave in the near future, and in a feasible way in terms of efficiency. Work by Rahal et al. [12], focused on multidatabases, propose a couple of approaches so that a cost model is able to get adjusted when the costs change; their first approach is interesting from a mediator point of view (the second one requires that changes are not abrupt so there is no need of a new model for each value measurement, which, in the case of web data sources characterization, can not be confirmed at all), even though cost information is obtained directly from the repositories (currently imposible to achieve in a web environment). Rahal defines a cost model as the independently-weighted sum of the characterization parameters, and describes a series of techniques to adjust them when the database behaviour changes. This is achieved by using an initial cost model, M_0 , which is updated after each sample query with an asymptotically tight bound of $\theta(n^2)$, where n is the number of parameters. Our criterion is that this approach is not the best when cost information can not be obtained directly from the database, such as in web sources, so a new model is shown with a simpler statistical method and is demonstrated how the update can be achieved with a bound of $\theta(1)$ on the number of sample queries.

The rest of the paper is organized as follows. Section 2 introduces the mediator/wrapper environment and the reasons why web sources optimization must be managed differently. Section 3 describes the cost repository structure and defines the model parameters. The fourth section defines how to obtain the best query plan by using this repository, and an efficient technique to update every parameter value in each query execution is defined and demonstrated in section 5. Section 6 shows our experimental results obtained by an implementation of the techniques described above and section 7 summarizes on conclusions.

2 Analysis of the Problem

Once received the query, the Mediator's Plan Generator produces the different possible plans of execution. Each plan will be made up of a set of subqueries over a set of remote sources; each one of these sources will return back a series of results that, after postprocessing and unification, allow to obtain the final result. The Plan Generator can generate a great amount of possible query plans for a single global user query; therefore, obtaining the optimal query plan for this specific moment is a task of the Logical Optimizer.

This paper will be facing the cost model for logical optimization, taking into account the cost information provided by the wrappers when accessing the web sources, and how these data are stored in the cost repository. Other issues must be taken into