

Vincent Corruble
Masayuki Takeda
Einoshin Suzuki (Eds.)

LNAI 4755

Discovery Science

10th International Conference, DS 2007
Sendai, Japan, October 2007
Proceedings



Springer

TP301-53
D611
2007

Vincent Corruble Masayuki Takeda
Einoshin Suzuki (Eds.)

Discovery Science

10th International Conference, DS 2007
Sendai, Japan, October 1-4, 2007
Proceedings



Springer



E2007003595

Series Editors

Jaime G. Carbonell, Carnegie Mellon University, Pittsburgh, PA, USA
Jörg Siekmann, University of Saarland, Saarbrücken, Germany

Volume Editors

Vincent Corruble

Université Pierre et Marie Curie (Paris 6)

Laboratoire d'Informatique de Paris 6

104 avenue du Président Kennedy, 75016 Paris, France

E-mail: Vincent.Corruble@lip6.fr

Masayuki Takeda

Einoshin Suzuki

Kyushu University

Department of Informatics

744 Motooka, Nishi, Fukuoka 819-0395, Japan

E-mail: {takeda, suzuki}@i.kyushu-u.ac.jp

Library of Congress Control Number: 2007935939

CR Subject Classification (1998): I.2, H.2.8, H.3, J.1, J.2

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN 0302-9743

ISBN-10 3-540-75487-3 Springer Berlin Heidelberg New York

ISBN-13 978-3-540-75487-9 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 2007

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper SPIN: 12170210 06/3180 5 4 3 2 1 0

Lecture Notes in Artificial Intelligence 4755

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Lecture Notes in Artificial Intelligence (LNAI)

- Vol. 4755: V. Corruble, M. Takeda, E. Suzuki (Eds.), *Discovery Science*. XI, 298 pages. 2007.
- Vol. 4754: M. Hutter, R.A. Servedio, E. Takimoto (Eds.), *Algorithmic Learning Theory*. XI, 403 pages. 2007.
- Vol. 4737: B. Berendt, A. Hotho, D. Mladenic, G. Semeraro (Eds.), *From Web to Social Web: Discovering and Deploying User and Content Profiles*. XI, 161 pages. 2007.
- Vol. 4733: R. Basili, M.T. Pazzienza (Eds.), *AI*IA 2007: Artificial Intelligence and Human-Oriented Computing*. XVII, 858 pages. 2007.
- Vol. 4724: K. Mellouli (Ed.), *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*. XV, 914 pages. 2007.
- Vol. 4722: C. Pelachaud, J.-C. Martin, E. André, G. Chollet, K. Karpouzis, D. Pelé (Eds.), *Intelligent Virtual Agents*. XV, 425 pages. 2007.
- Vol. 4720: B. Konev, F. Wolter (Eds.), *Frontiers of Combining Systems*. X, 283 pages. 2007.
- Vol. 4702: J.N. Kok, J. Koronacki, R. Lopez de Mantaras, S. Matwin, D. Mladenić, A. Skowron (Eds.), *Knowledge Discovery in Databases: PKDD 2007*. XXIV, 640 pages. 2007.
- Vol. 4701: J.N. Kok, J. Koronacki, R.L.d. Mantaras, S. Matwin, D. Mladenić, A. Skowron (Eds.), *Machine Learning: ECML 2007*. XXII, 809 pages. 2007.
- Vol. 4696: H.-D. Burkhard, G. Lindemann, R. Verbrugge, L.Z. Varga (Eds.), *Multi-Agent Systems and Applications V*. XIII, 350 pages. 2007.
- Vol. 4694: B. Apolloni, R.J. Howlett, L. Jain (Eds.), *Knowledge-Based Intelligent Information and Engineering Systems, Part III*. XXIX, 1126 pages. 2007.
- Vol. 4693: B. Apolloni, R.J. Howlett, L. Jain (Eds.), *Knowledge-Based Intelligent Information and Engineering Systems, Part II*. XXXII, 1380 pages. 2007.
- Vol. 4692: B. Apolloni, R.J. Howlett, L. Jain (Eds.), *Knowledge-Based Intelligent Information and Engineering Systems, Part I*. LV, 882 pages. 2007.
- Vol. 4687: P. Petta, J.P. Müller, M. Klusch, M. Georgeff (Eds.), *Multiagent System Technologies*. X, 207 pages. 2007.
- Vol. 4682: D.-S. Huang, L. Heutte, M. Loog (Eds.), *Advanced Intelligent Computing Theories and Applications*. XXVII, 1373 pages. 2007.
- Vol. 4676: M. Klusch, K.V. Hindriks, M.P. Papazoglou, L. Sterling (Eds.), *Cooperative Information Agents XI*. XI, 361 pages. 2007.
- Vol. 4667: J. Hertzberg, M. Beetz, R. Englert (Eds.), *KI 2007: Advances in Artificial Intelligence*. IX, 516 pages. 2007.
- Vol. 4660: S. Džeroski, J. Todorovski (Eds.), *Computational Discovery of Scientific Knowledge*. X, 327 pages. 2007.
- Vol. 4659: V. Mařík, V. Vyatkin, A.W. Colombo (Eds.), *Holonic and Multi-Agent Systems for Manufacturing*. VIII, 456 pages. 2007.
- Vol. 4651: F. Azevedo, P. Barahona, F. Fages, F. Rossi (Eds.), *Recent Advances in Constraints*. VIII, 185 pages. 2007.
- Vol. 4648: F. Almeida e Costa, L.M. Rocha, E. Costa, I. Harvey, A. Coutinho (Eds.), *Advances in Artificial Life*. XVIII, 1215 pages. 2007.
- Vol. 4635: B. Kocinov, D.C. Richardson, T.R. Roth-Berghofer, L. Vieu (Eds.), *Modeling and Using Context*. XIV, 574 pages. 2007.
- Vol. 4632: R. Alhajj, H. Gao, X. Li, J. Li, O.R. Zaiane (Eds.), *Advanced Data Mining and Applications*. XV, 634 pages. 2007.
- Vol. 4629: V. Matoušek, P. Mautner (Eds.), *Text, Speech and Dialogue*. XVII, 663 pages. 2007.
- Vol. 4626: R.O. Weber, M.M. Richter (Eds.), *Case-Based Reasoning Research and Development*. XIII, 534 pages. 2007.
- Vol. 4617: V. Torra, Y. Narukawa, Y. Yoshida (Eds.), *Modeling Decisions for Artificial Intelligence*. XII, 502 pages. 2007.
- Vol. 4612: I. Miguel, W. Ruml (Eds.), *Abstraction, Reformulation, and Approximation*. XI, 418 pages. 2007.
- Vol. 4604: U. Priss, S. Polovina, R. Hill (Eds.), *Conceptual Structures: Knowledge Architectures for Smart Applications*. XII, 514 pages. 2007.
- Vol. 4603: F. Pfenning (Ed.), *Automated Deduction – CADE-21*. XII, 522 pages. 2007.
- Vol. 4597: P. Perner (Ed.), *Advances in Data Mining*. XI, 353 pages. 2007.
- Vol. 4594: R. Bellazzi, A. Abu-Hanna, J. Hunter (Eds.), *Artificial Intelligence in Medicine*. XVI, 509 pages. 2007.
- Vol. 4585: M. Kryszkiewicz, J.F. Peters, H. Rybinski, A. Skowron (Eds.), *Rough Sets and Intelligent Systems Paradigms*. XIX, 836 pages. 2007.
- Vol. 4578: F. Masulli, S. Mitra, G. Pasi (Eds.), *Applications of Fuzzy Sets Theory*. XVIII, 693 pages. 2007.
- Vol. 4573: M. Kauers, M. Kerber, R. Miner, W. Windsteiger (Eds.), *Towards Mechanized Mathematical Assistants*. XIII, 407 pages. 2007.
- Vol. 4571: P. Perner (Ed.), *Machine Learning and Data Mining in Pattern Recognition*. XIV, 913 pages. 2007.

- Vol. 4570: H.G. Okuno, M. Ali (Eds.), *New Trends in Applied Artificial Intelligence*. XXI, 1194 pages. 2007.
- Vol. 4565: D.D. Schmorow, L.M. Reeves (Eds.), *Foundations of Augmented Cognition*. XIX, 450 pages. 2007.
- Vol. 4562: D. Harris (Ed.), *Engineering Psychology and Cognitive Ergonomics*. XXIII, 879 pages. 2007.
- Vol. 4548: N. Olivetti (Ed.), *Automated Reasoning with Analytic Tableaux and Related Methods*. X, 245 pages. 2007.
- Vol. 4539: N.H. Bshouty, C. Gentile (Eds.), *Learning Theory*. XII, 634 pages. 2007.
- Vol. 4529: P. Melin, O. Castillo, L.T. Aguilar, J. Kacprzyk, W. Pedrycz (Eds.), *Foundations of Fuzzy Logic and Soft Computing*. XIX, 830 pages. 2007.
- Vol. 4520: M.V. Butz, O. Sigaud, G. Pezzulo, G. Baldassarre (Eds.), *Anticipatory Behavior in Adaptive Learning Systems*. X, 379 pages. 2007.
- Vol. 4511: C. Conati, K. McCoy, G. Paliouras (Eds.), *User Modeling 2007*. XVI, 487 pages. 2007.
- Vol. 4509: Z. Kobti, D. Wu (Eds.), *Advances in Artificial Intelligence*. XII, 552 pages. 2007.
- Vol. 4496: N.T. Nguyen, A. Grzech, R.J. Howlett, L.C. Jain (Eds.), *Agent and Multi-Agent Systems: Technologies and Applications*. XXI, 1046 pages. 2007.
- Vol. 4483: C. Baral, G. Brewka, J. Schlipf (Eds.), *Logic Programming and Nonmonotonic Reasoning*. IX, 327 pages. 2007.
- Vol. 4482: A. An, J. Stefanowski, S. Ramanna, C.J. Butz, W. Pedrycz, G. Wang (Eds.), *Rough Sets, Fuzzy Sets, Data Mining and Granular Computing*. XIV, 585 pages. 2007.
- Vol. 4481: J. Yao, P. Lingras, W.-Z. Wu, M. Szczuka, N.J. Cercone, D. Ślęzak (Eds.), *Rough Sets and Knowledge Technology*. XIV, 576 pages. 2007.
- Vol. 4476: V. Gorodetsky, C. Zhang, V.A. Skormin, L. Cao (Eds.), *Autonomous Intelligent Systems: Multi-Agents and Data Mining*. XIII, 323 pages. 2007.
- Vol. 4456: Y. Wang, Y.-m. Cheung, H. Liu (Eds.), *Computational Intelligence and Security*. XXIII, 1118 pages. 2007.
- Vol. 4455: S. Muggleton, R. Otero, A. Tamaddoni-Nezhad (Eds.), *Inductive Logic Programming*. XII, 456 pages. 2007.
- Vol. 4452: M. Fasli, O. Shehory (Eds.), *Agent-Mediated Electronic Commerce*. VIII, 249 pages. 2007.
- Vol. 4451: T.S. Huang, A. Nijholt, M. Pantic, A. Pentland (Eds.), *Artificial Intelligence for Human Computing*. XVI, 359 pages. 2007.
- Vol. 4441: C. Müller (Ed.), *Speaker Classification*. X, 309 pages. 2007.
- Vol. 4438: L. Maicher, A. Sigel, L.M. Garshol (Eds.), *Leveraging the Semantics of Topic Maps*. X, 257 pages. 2007.
- Vol. 4434: G. Lakemeyer, E. Sklar, D.G. Sorrenti, T. Takahashi (Eds.), *RoboCup 2006: Robot Soccer World Cup X*. XIII, 566 pages. 2007.
- Vol. 4429: R. Lu, J.H. Siekmann, C. Ullrich (Eds.), *Cognitive Systems*. X, 161 pages. 2007.
- Vol. 4428: S. Edelkamp, A. Lomuscio (Eds.), *Model Checking and Artificial Intelligence*. IX, 185 pages. 2007.
- Vol. 4426: Z.-H. Zhou, H. Li, Q. Yang (Eds.), *Advances in Knowledge Discovery and Data Mining*. XXV, 1161 pages. 2007.
- Vol. 4411: R.H. Bordini, M. Dastani, J. Dix, A.E.F. Seghrouchni (Eds.), *Programming Multi-Agent Systems*. XIV, 249 pages. 2007.
- Vol. 4410: A. Branco (Ed.), *Anaphora: Analysis, Algorithms and Applications*. X, 191 pages. 2007.
- Vol. 4399: T. Kovacs, X. Llorà, K. Takadama, P.L. Lanzi, W. Stolzmann, S.W. Wilson (Eds.), *Learning Classifier Systems*. XII, 345 pages. 2007.
- Vol. 4390: S.O. Kuznetsov, S. Schmidt (Eds.), *Formal Concept Analysis*. X, 329 pages. 2007.
- Vol. 4389: D. Weyns, H. Van Dyke Parunak, F. Michel (Eds.), *Environments for Multi-Agent Systems III*. X, 273 pages. 2007.
- Vol. 4386: P. Noriega, J. Vázquez-Salceda, G. Boella, O. Boissier, V. Dignum, N. Fornara, E. Matson (Eds.), *Coordination, Organizations, Institutions, and Norms in Agent Systems II*. XI, 373 pages. 2007.
- Vol. 4384: T. Washio, K. Satoh, H. Takeda, A. Inokuchi (Eds.), *New Frontiers in Artificial Intelligence*. IX, 401 pages. 2007.
- Vol. 4371: K. Inoue, K. Satoh, F. Toni (Eds.), *Computational Logic in Multi-Agent Systems*. X, 315 pages. 2007.
- Vol. 4369: M. Umeda, A. Wolf, O. Bartenstein, U. Geske, D. Seipel, O. Takata (Eds.), *Declarative Programming for Knowledge Management*. X, 229 pages. 2006.
- Vol. 4363: B.D. ten Cate, H.W. Zeevat (Eds.), *Logic, Language, and Computation*. XII, 281 pages. 2007.
- Vol. 4343: C. Müller (Ed.), *Speaker Classification I*. X, 355 pages. 2007.
- Vol. 4342: H. de Swart, E. Orłowska, G. Schmidt, M. Roubens (Eds.), *Theory and Applications of Relational Structures as Knowledge Instruments II*. X, 373 pages. 2006.
- Vol. 4335: S.A. Brueckner, S. Hassas, M. Jelasity, D. Yamins (Eds.), *Engineering Self-Organising Systems*. XII, 212 pages. 2007.
- Vol. 4334: B. Beckert, R. Hähnle, P.H. Schmitt (Eds.), *Verification of Object-Oriented Software*. XXIX, 658 pages. 2007.
- Vol. 4333: U. Reimer, D. Karagiannis (Eds.), *Practical Aspects of Knowledge Management*. XII, 338 pages. 2006.
- Vol. 4327: M. Baldoni, U. Endriss (Eds.), *Declarative Agent Languages and Technologies IV*. VIII, 257 pages. 2006.
- Vol. 4314: C. Freksa, M. Kohlhase, K. Schill (Eds.), *KI 2006: Advances in Artificial Intelligence*. XII, 458 pages. 2007.
- Vol. 4304: A. Sattar, B.-h. Kang (Eds.), *AI 2006: Advances in Artificial Intelligence*. XXVII, 1303 pages. 2006.

Preface

This volume contains the papers presented at DS-2007: The Tenth International Conference on Discovery Science held in Sendai, Japan, October 1–4, 2007.

The main objective of the Discovery Science (DS) conference series is to provide an open forum for intensive discussions and the exchange of new ideas and information among researchers working in the area of automating scientific discovery or working on tools for supporting the human process of discovery in science. It has been a successful arrangement in the past to co-locate the DS conference with the International Conference on Algorithmic Learning Theory (ALT). This combination of ALT and DS allows for a comprehensive treatment of the whole range, from theoretical investigations to practical applications. Continuing this tradition, DS 2007 was co-located with the 18th ALT conference (ALT 2007). The proceedings of ALT 2007 were published as a twin volume 4754 of the LNCS series.

The International Steering Committee of the Discovery Science conference series provided important advice on a number of issues during the planning of Discovery Science 2007. The members of the Steering Committee are Einoshin Suzuki (Kyushu University, Chair), Achim G. Hoffmann (University of New South Wales, Vice Chair), Setsuo Arikawa (Kyushu University), Hiroshi Motoda (Osaka University), Masahiko Sato (Kyoto University), Satoru Miyano (University of Tokyo), Thomas Zeugmann (Hokkaido University), Ayumi Shinohara (Tohoku University), Alberto Apostolico (Georgia Institute of Technology and University of Padova), Massimo Melucci (University of Padova), Tobias Scheffer (Max Planck Institute for Computer Science), Ken Satoh (National Institute of Informatics), Nada Lavrac (Jozef Stefan Institute), Ljupco Todorovski (University of Ljubljana), and Hiroki Arimura (Hokkaido University).

In response to the call for papers 55 manuscripts were submitted. The Program Committee selected for publication 17 submissions as long papers and 10 submissions as regular papers. Each submission was reviewed by at least two members of the Program Committee, which consisted of international experts in the field. The selection was made after careful evaluation of each paper based on clarity, significance, technical quality, and originality, as well as relevance to the field of discovery science. This volume consists of three parts. The first part contains the papers/abstracts of the invited talks, the second part contains the accepted long papers, and the third part contains the accepted regular papers.

We are deeply indebted to the Program Committee members as well as their subreferees who played the critically important role of reviewing the submitted papers and contributing to the intense discussions which resulted in the selection of the papers published in this volume. Without their enormous effort, ensuring the high quality of the work presented at Discovery Science 2007 would not have been possible. Furthermore, we would like to thank all individuals and

institutions who contributed to the success of the conference: the authors for submitting papers, the invited speakers for their acceptance of the invitation and their stimulating contributions to the conference, the Steering Committee, and the sponsors for their support. In particular, we acknowledge the generous financial support from the Air Force Office of Scientific Research (AFOSR), Asian Office of Aerospace Research and Development (AOARD)¹; the Graduate School of Information Sciences (GSIS), Tohoku University for providing secretarial assistance and equipment; the Research Institute of Electrical Communication (RIEC), Tohoku University; New Horizons in Computing, MEXT Grant-in-Aid for Scientific Research on Priority Areas; and the Semi-Structured Data Mining Project, MEXT Grant-in-Aid for Specially Promoted Research.

July 2007

Vincent Corruble
Masayuki Takeda
Einoshin Suzuki

¹ AFOSR/AOARD support is not intended to express or imply endorsement by the U. S. Federal Government.

Conference Organization

Conference Chair

Ayumi Shinohara

Tohoku University, Japan

Program Committee

Vincent Corruble (Co-chair)	Université Pierre et Marie Curie, Paris, France
Masayuki Takeda (Co-chair)	Kyushu University, Japan
Jean-Francois Boulicaut	INSA Lyon, France
Will Bridewell	CSLI, Stanford, USA
Simon Colton	Imperial College London, UK
Antoine Cornuejols	Université Paris-Sud, France
Andreas Dress	Shanghai Institutes for Biological Sciences, China
Saso Dzeroski	Jozef Stefan Institute, Slovenia
Tapio Elomaa	Tampere University of Technology, Finland
Johannes Fuernkranz	Technical University of Darmstadt, Germany
Dragan Gamberger	Rudjer Boskovic Institute, Hungary
Ricard Gavaldà	Technical University of Catalonia, Spain
Gunter Grieser	Technical University of Darmstadt, Germany
Fabrice Guillet	Ecole Polytechnique of the University of Nantes, France
Mohand-Said Hacid	Université Lyon 1, France
Udo Hahn	Jena University, Germany
Makoto Haraguchi	Hokkaido University, Japan
Tomoyuki Higuchi	The Institute of Statistical Mathematics, Japan
Kouichi Hirata	Kyushu Institute of Technology, Japan
Tu Bao Ho	Japan Advanced Institute of Science and Technology, Japan
Achim Hoffmann	University of New South Wales, Australia
Tamás Horváth	Fraunhofer Institute for Intelligent Analysis and Information Systems, Germany
Daisuke Ikeda	Kyushu University, Japan
Kentaro Inui	Nara Institute of Science and Technology, Japan
Szymon Jaroszewicz	National Institute of Telecommunications, Poland
Hisashi Kashima	IBM Research, Tokyo Research Laboratory, Japan
Kristian Kersting	Universitaet Freiburg, Germany
Ross King	University of Wales, UK
Andras Kocsor	University of Szeged, Hungary
Kevin Korb	Monash University, Australia
Stefan Kramer	Technical University of Munich, Germany
Nicolas Lachiche	Université de Strasbourg, France

Table of Contents

Invited Papers

Challenge for Info-plosion	1
<i>Masaru Kitsuregawa</i>	
Machine Learning in Ecosystem Informatics	9
<i>Thomas G. Dietterich</i>	
Simple Algorithmic Principles of Discovery, Subjective Beauty, Selective Attention, Curiosity & Creativity	26
<i>Jürgen Schmidhuber</i>	
A Theory of Similarity Functions for Learning and Clustering	39
<i>Avrim Blum</i>	
A Hilbert Space Embedding for Distributions	40
<i>Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf</i>	

Long Papers

Time and Space Efficient Discovery of Maximal Geometric Graphs	42
<i>Hiroki Arimura, Takeaki Uno, and Shinichi Shimozono</i>	
Iterative Reordering of Rules for Building Ensembles Without Relearning	56
<i>Paulo J. Azevedo and Alípio M. Jorge</i>	
On Approximating Minimum Infrequent and Maximum Frequent Sets	68
<i>Mario Boley</i>	
A Partially Dynamic Clustering Algorithm for Data Insertion and Removal	78
<i>Haytham Elghazel, Hamamache Kheddouci, Véronique Deslandres, and Alain Dussauchoy</i>	
Positivism Against Constructivism: A Network Game to Learn Epistemology	91
<i>Hélène Hagège, Christopher Dartnell, and Jean Sallantin</i>	
Learning Locally Weighted C4.4 for Class Probability Estimation	104
<i>Liangxiao Jiang, Harry Zhang, Dianhong Wang, and Zhihua Cai</i>	

User Preference Modeling from Positive Contents for Personalized Recommendation	116
<i>Heung-Nam Kim, Inay Ha, Jin-Guk Jung, and Geun-Sik Jo</i>	
Reducing Trials by Thinning-Out in Skill Discovery	127
<i>Hayato Kobayashi, Kohei Hatano, Akira Ishino, and Ayumi Shinohara</i>	
A Theoretical Study on Variable Ordering of Zero-Suppressed BDDs for Representing Frequent Itemsets	139
<i>Shin-ichi Minato</i>	
Fast NML Computation for Naive Bayes Models	151
<i>Tommi Mononen and Petri Myllymäki</i>	
Unsupervised Spam Detection Based on String Alieness Measures	161
<i>Kazuyuki Narisawa, Hideo Bannai, Kohei Hatano, and Masayuki Takeda</i>	
A Consequence Finding Approach for Full Clausal Abduction	173
<i>Oliver Ray and Katsumi Inoue</i>	
Literature-Based Discovery by an Enhanced Information Retrieval Model	185
<i>Kazuhiro Seki and Javed Mostafa</i>	
Discovering Mentorship Information from Author Collaboration Networks	197
<i>V. Suresh, Narayanan Raghupathy, B. Shekar, and C.E. Veni Madhavan</i>	
Active Contours as Knowledge Discovery Methods	209
<i>Arkadiusz Tomczyk, Piotr S. Szczepaniak, and Michal Pryczek</i>	
An Efficient Polynomial Delay Algorithm for Pseudo Frequent Itemset Mining	219
<i>Takeaki Uno and Hiroki Arimura</i>	
Discovering Implicit Feedbacks from Search Engine Log Files	231
<i>Ashok Veilumuthu and Parthasarathy Ramachandran</i>	

Regular Papers

Pharmacophore Knowledge Refinement Method in the Chemical Structure Space	243
<i>Satoshi Fujishima, Yoshimasa Takahashi, and Takashi Okada</i>	
An Attempt to Rebuild C. Bernard's Scientific Steps	248
<i>Jean-Gabriel Ganascia and Bassel Habib</i>	

Semantic Annotation of Data Tables Using a Domain Ontology	253
<i>Gaëlle Hignette, Patrice Buche, Juliette Dibia-Barthélemy, and Olivier Haemmerlé</i>	
Model Selection and Estimation Via Subjective User Preferences	259
<i>Jaakko Hollmén</i>	
Detecting Concept Drift Using Statistical Testing	264
<i>Kyosuke Nishida and Koichiro Yamauchi</i>	
Towards Future Technology Projection: A Method for Extracting Capability Phrases from Documents	270
<i>Risa Nishiyama, Hironori Takeuchi, and Hideo Watanabe</i>	
Efficient Incremental Mining of Top-K Frequent Closed Itemsets	275
<i>Andrea Pietracaprina and Fabio Vandin</i>	
An Intentional Kernel Function for RNA Classification	281
<i>Hiroshi Sankoh, Koichiro Doi, and Akihiro Yamamoto</i>	
Mining Subtrees with Frequent Occurrence of Similar Subtrees	286
<i>Hisashi Tosaka, Atsuyoshi Nakamura, and Mineichi Kudo</i>	
Semantic Based Real-Time Clustering for PubMed Literatures	291
<i>Ruey-Ling Yeh, Ching Liu, Ben-Chang Shia, I-Jen Chiang, Wen-Wen Yang, and Hsiang-Chun Tsai</i>	
Author Index	297

Challenge for Info-plosion

Masaru Kitsuregawa

Institute of Industrial Science, The University of Tokyo
4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan
kitsure@tkl.iis.u-tokyo.ac.jp

Abstract. Information created by people has increased rapidly since the year 2000, and now we are in a time which we could call the “information-explosion era.” The project “Cyber Infrastructure for the Information-explosion Era” is a six-year project from 2005 to 2010 supported by Grant-in-Aid for Scientific Research on Priority Areas from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan. The project aims to establish the following fundamental technologies in this information-explosion era: novel technologies for efficient and trustable information retrieval from explosively growing and heterogeneous information resources; stable, secure, and scalable information systems for managing rapid information growth; and information utilization by harmonized human-system interaction. It also aims to design a social system that cooperates with these technologies. Moreover, it maintains the synergy of cutting-edge technologies in informatics.

1 New IT Infrastructure for the Information Explosion Era

The volume of information generated by mankind has increased exponentially, i.e., “exploded” since 2000. The purpose of our research project, “Cyber Infrastructure for the Info-plosion Era” in the Ministry of Education, Culture, Sports, Science and Technology (MEXT) Grand-in-Aid for Scientific Research on Priority Area, is to build advanced IT infrastructure technologies for this information explosion era [1]. According to the research by the University of California at Berkeley, the volume of information created by human is explosively increasing [2, 3]. Huge volume of data is also created by sensors and machines. We have considered that the most important theme for researchers in the field of computer science is the research on new IT infrastructure for the Information-explosion Era.

The project has three major research components (Research Groups) to achieve this goal: build technologies to search for needed information efficiently, without bias and without being at risk from the rapidly growing volume of information (A01); build new and sustainable technologies that can operate large-scale information systems managing enormous amounts of information safely and securely (A02), and build human-friendly technologies to enable flexible dialogue between men and machines and enable everyone to utilize information (A03).

Underlying these three components is the research into new social systems that can facilitate the use of advanced, information-based IT services (B01). In addition, Large-scale Info-plosion Platform (LIP) is implemented as the shared platforms used by all Research Groups. The project is interdisciplinary in its structure, bringing together advanced research methods in information-related areas. (Project leader: Masaru Kitsuregawa)

2 Infrastructure for Information Management, Fusion and Utilization in the Information Explosion Era (A01)

This Group focuses on the shortfalls of present internet searches and looks into new search methods, including better ranking systems (where minority opinions are not overlooked), interactive searches, reliability assessments and time-space searches. Currently, for knowledge workers, about 30% of their time on intellectual activities is spent just for retrieval information [4]. The Group will attempt to create a system of search platforms to search a massive volume of web page contents.

Another important issue for information retrieval is the dangers associated with information rankings. Search engines are extensively used in the web world. When a general word is given, it may hit millions of pages, and only about 10

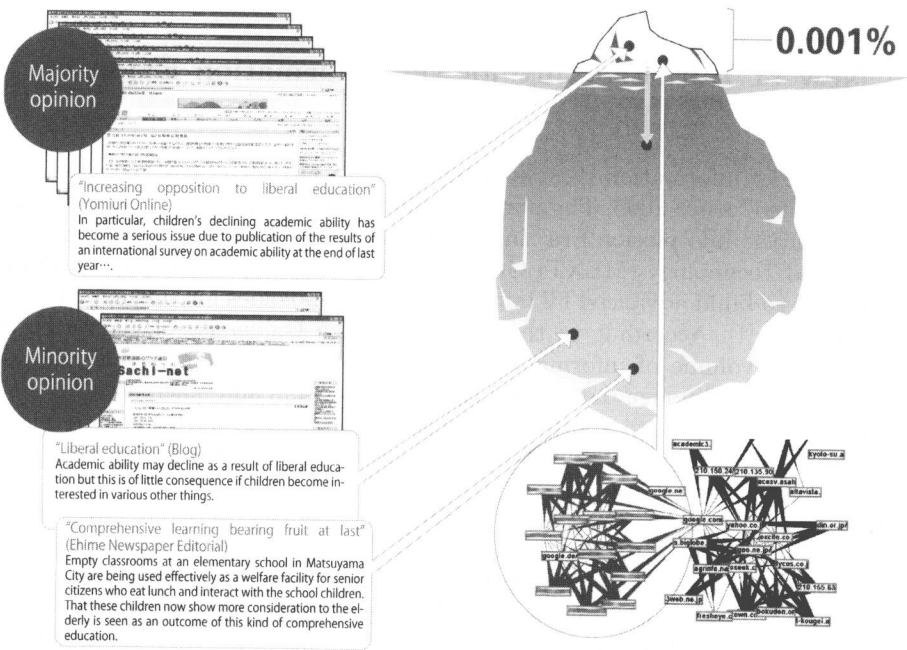


Fig. 1. Dangers Associated with Information Rankings

probable candidates are listed as top rankers on the first page of the search result. Now, is this ranking really reasonable? Who guarantee the correctness of it? Actually, this is controlled by just a private company. There should be a possibility that the ranking is sold and bought. In addition, it is possible to control the ranking deliberately. We have found, as shown in the Fig.1, that the red island of links suddenly appear on February 2004, while only the black island of links is found in 2003. This is an example of a trick to raise their own rank by linking them to a well-known site.

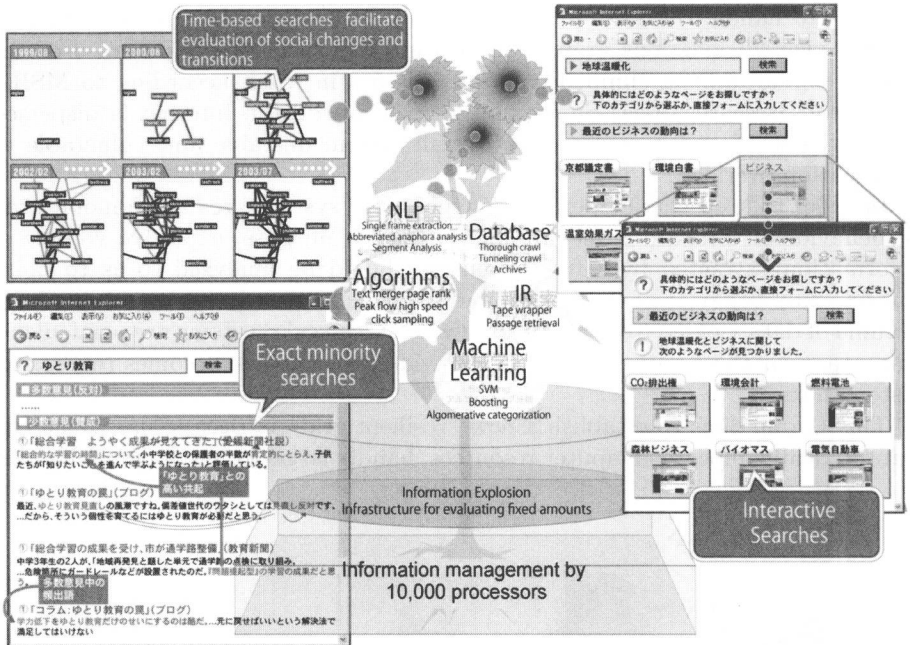


Fig. 2. Next generation searches

Also, in the current ranking system, majority opinions are highlighted while minority opinions are buried. For example, when “Yutori Kyouiku” (liberal education) is searched, majority opinions are found easily like “increasing opposition to liberal education”. Although there are minority opinions such as “comprehensive learning bearing fruit at last”, they could be completely neglected.

Various technologies, including information searching, natural language processing, machine learning, artificial intelligence (AI) and database technology, are integrated into this system to enable quantitative analysis to be performed (Fig.2). We hope to enable a remarkable level of interdisciplinary synergy among different fields. It is better for human to understand by presenting relational information also than information itself. Several innovative researches are expected in this area, for example, presenting comparative information [5] and information

on time sequence [6]. Research on the methods of management, integration and processing of “exploding” real-world information, including cyber information and information obtained from remote sensors, will also be conducted. (Group leader: Masaru Kitsuregawa, University of Tokyo)

3 Infrastructure for Information-Explosion-Proof IT Systems (A02)

The exponential increase in the amount of information requires far larger IT systems to handle the volume. According to [7], Google has now over 450,000 servers over 25 locations around the world. The ratio of computers used for search engines among shipped computers is 5% in total, according to MSRA (Microsoft Research Asia) Summit in 2006. Data on the Internet is dispersed over millions of nodes, making the overall system unstable and vulnerable to information overload.

In order to keep stable operation of such huge systems, real-time monitoring of behavior of software is inevitable. For such a purpose, explosive volume of information extracted by software sensors should be analyzed so as to point out anomaly behavior of the system and stabilize it. Researches on mining huge volume of data yielded from monitoring very-large-scale systems are particularly important for the age in which a nation-wide cyber attack becomes reality like Estonian case.

This Group aims to establish a new “resilient grid” infrastructure which can automatically allocate computer resources, handle large-scale system faults over

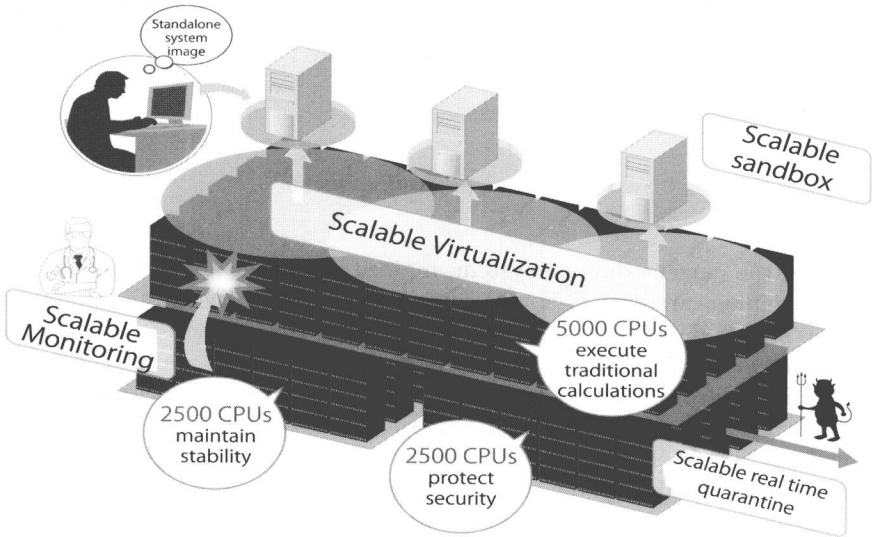


Fig. 3. Infrastructure for large scale system

the network without human intervention and without the modern-day concerns of security breaches and intrusion (Fig.3). This is partially based on Autonomic Computing [8] and interesting research works performed by UC Berkeley and Stanford University [9]. The resilient grid will allow a high-performance virtual computing environment to be configured autonomously, on which applications can be deployed safely and securely. (Group leader: Satoshi Matsuoka, Tokyo Institute of Technology)

4 Infrastructure for Human Communication in the Information Explosion Era (A03)

The information explosion has two aspects: qualitative (volume) and quantitative (complexity). This Group proposes studies of the advancement of human communication to address the issues related to complexity. The underlying concept is a mutually adaptable multi-modal interaction that can fill the communication gaps between people and information systems. This is the key to overcoming the complexity resulting from highly functional and multi-functional information systems, and establishing a secure and user-friendly interactive environment (Fig.4).

Searching information from explosively huge size of information space still requires advanced skills, since existing tools for such a purpose is not necessarily easy to use for naive users. Human-friendly interfaces as well as communications

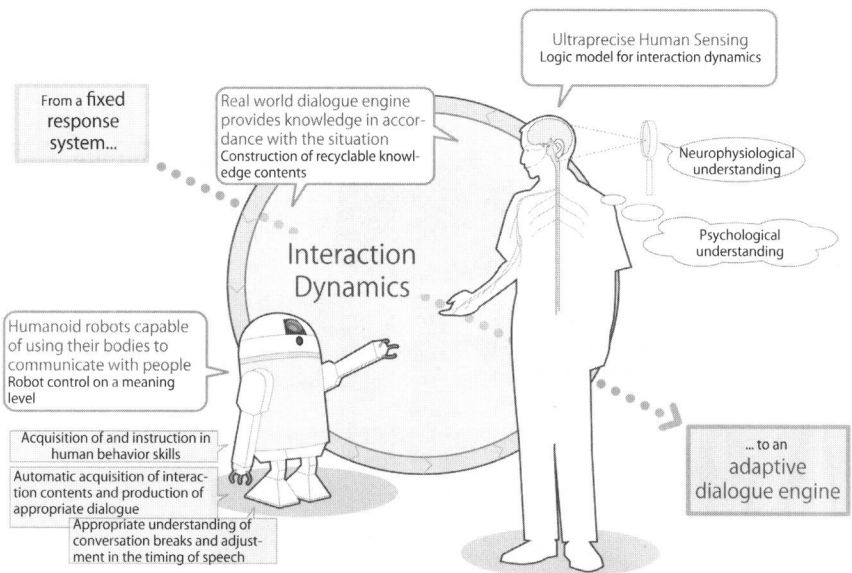


Fig. 4. Infrastructure for human