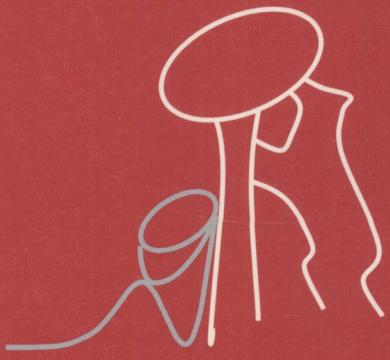
A. Fazel Famili Joost N. Kok
José M. Peña Arno Siebes
Ad Feelders (Eds.)

LNCS 3646

Advances in Intelligent Data Analysis VI

6th International Symposium on
Intelligent Data Analysis, IDA 2005
Madrid, Spain, September 2005, Proceedings



Springer

A. Fazel Famili Joost N. Kok
José M. Peña Arno Siebes
Ad Feelders (Eds.)

Advances in Intelligent Data Analysis VI

6th International Symposium on
Intelligent Data Analysis, IDA 2005
Madrid, Spain, September 8-10, 2005
Proceedings

Springer

Volume Editors

A. Fazel Famili
IIT/ITI - National Research Council Canada, Ottawa University
School of Information Technology and Engineering
1200 Montreal Rd, M-50, Ottawa, ON K1A 0R6, Canada
E-mail: fazel.famili@nrc-cnrc.gc.ca

Joost N. Kok
Leiden University, Leiden Institute of Advanced Computer Science
Niels Bohrweg 1, 2333 CA Leiden, The Netherlands
E-mail: joost@liacs.nl

José M. Peña
Universidad Politécnica de Madrid, DATSI - Facultad de Informática
Campus de Montegancedo S/N, Boadilla del Monte, 28660 Madrid, Spain
E-mail: jmpena@fi.upm.es

Arno Siebes
Ad Feelders
Utrecht University, Department of Information and Computing Sciences
PO Box 80.089, 3508 TB Utrecht, The Netherlands
E-mail: {arno, ad}@cs.uu.nl

Library of Congress Control Number: 2005931595

CR Subject Classification (1998): H.3, I.2, G.3, I.5.1, I.4.5, J.2, J.1, J.3

ISSN 0302-9743
ISBN-10 3-540-28795-7 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-28795-7 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11552253 06/3142 5 4 3 2 1 0

Lecture Notes in Computer Science 3646

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

Preface

One of the superb characteristics of intelligent data analysis (IDA) is that it is an interdisciplinary field in which researchers and practitioners from a number of areas are involved in a typical project. This also creates a challenge in which the success of a team depends on the participation of users and domain experts who need to interact with researchers and developers of any IDA system. All this is usually reflected in successful projects and of course in the papers that were evaluated by this year's Program Committee from which the final program has been developed.

In our call for papers, we solicited papers on (i) applications and tools, (ii) theory and general principles, and (iii) algorithms and techniques. We received a total of 184 papers, reviewing these was a major challenge. Each paper was assigned to three reviewers. In the end 46 papers were accepted, all of which were included in the proceedings and presented at the conference.

This year's papers reflect the results of applied and theoretical research from a number of disciplines all of which are related to the field of intelligent data analysis. To have the best combination of theoretical and applied research and also provide the best focus, we divided this year's IDA program into tutorials, invited talks, panel discussions and technical sessions.

We managed to organize two excellent tutorials on the first day by Luc De Raedt and Kristian Kersting, entitled *Probabilistic Inductive Logic Programming*, and by Bruno Apolloni, Dario Malchiodi and Sabrina Gaito, entitled *Statistical Bases of Machine Learning*. Our invited speakers were Prof. Ivan Bratko from the Jozef Stefan Institute in Slovenia, and Prof. Alex Freitas from the University of Kent.

We wish to express our sincere thanks to many people who worked hard for the IDA conference to happen in Madrid. Special thanks to tutorial, publicity, local organization, and panel chairs who were in charge of a large portion of our responsibilities. We would also like to thank Xiaohui Liu and Michael Berthold who worked as advisors to this conference, and the members of the Local Organizing Committee for their hard work. Finally, we are grateful to the members of our Program Committee; without their help it would have been impossible to put together such a valuable program.

September 2005

A. Fazel Famili,
José Maria S. Peña,
Joost Kok,
Arno Siebes,
Ad Feelders

Organization

Conference Organization

General Chair

A. Fazel Famili
National Research Council
Ottawa, Canada

Program Chairs

José M. Peña
Universidad Politécnica de Madrid
Madrid, Spain

Arno Siebes
Utrecht University
Utrecht, The Netherlands

Joost Kok
Leiden University
Leiden, The Netherlands

Tutorial Chair

Pedro Larrañaga
EHU-Universidad del País Vasco
San Sebastián, Spain

Publication Chair

Ad Feelders
Utrecht University
Utrecht, The Netherlands

Publicity Chairs

Jorge Muruzábal
Universidad Rey Juan Carlos
Madrid, Spain

Julián Sánchez
Quinao S.L.
Madrid, Spain

Local Organization Chair

Víctor Robles
Universidad Politécnica de Madrid
Madrid, Spain

Panel Chair

Sofian Maabout
LaBRI-Université Bordeaux
Bordeaux, France

Local Committee

Universidad Politécnica de Madrid
Madrid, Spain

María S. Pérez
Vanessa Herves
Francisco Rosales
Antonio García
Óscar Cubo
Pilar Herrero
Antonio LaTorre
Alberto Sánchez

Universidad Rey Juan Carlos
Madrid, Spain

Susana Vegas
Andrés L. Martinez

Program Committee

Niall Adams, Imperial College London, UK
Riccardo Bellazzi, University of Pavia, Italy
Bettina Berendt, Humboldt University of Berlin, Germany
Michael Berthold, University of Konstanz, Germany
Hans-Georg Beyer, Vorarlberg University of Applied Sciences, Austria
Jean-François Boulicaut, INSA Lyon, France
Christian Borgelt, Otto-von-Guericke-Universität Magdeburg, Germany
Hans-Dieter Burkhard, Humboldt Universität Berlin, Germany
Luis M. de Campos, Universidad de Granada, Spain
Fazel Famili, Institute for Information Technology, NRC, Canada
Giuseppe Di Fatta, University of Konstanz, Germany
Fridtjof Feldbusch, University of Karlsruhe, Germany
Ingrid Fischer, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
Douglas Fisher, Vanderbilt University, USA
Peter Flach, University of Bristol, UK
Eibe Frank, University of Waikato, New Zealand
Karl A. Fröschl, ec3 – eCommerce Competence Center, Vienna DC, Austria
Gabriela Guimaraes, CENTRIA UNL, Portugal
Lawrence O. Hall, University of South Florida, USA
Pilar Herrero, Universidad Politécnica de Madrid, Spain
Tom Heskes, Radboud University Nijmegen, The Netherlands
Alexander Hinneburg, University of Halle, Germany
Frank Hoeppner, University of Wolfenbuettel, Germany
Adele Howe, Colorado State University, USA
Klaus-Peter Huber, SAS Institute, Germany
Anthony Hunter, University College London, UK
Alfred Inselberg, Tel Aviv University, Israel
Bert Kappen, Radboud University Nijmegen, The Netherlands
Frank Klawonn, University of Wolfenbuettel, Germany
Joost N. Kok, Leiden University, The Netherlands
Walter Kosters, Leiden University, The Netherlands
Rudolf Kruse, Otto-von-Guericke-Universität Magdeburg, Germany
Pedro Larrañaga, Universidad del País Vasco, Spain
Hans-Joachim Lenz, Freie Universität Berlin, Germany
Xiaohui Liu, Brunel University, UK
Sofian Maabout, LaBRI-Université Bordeaux, France
Rainer Malaka, European Media Laboratory, Heidelberg, Germany
Jorge Muruzábal, Universidad Rey Juan Carlos, Spain
Susana Nascimento, CENTRIA-Universidade Nova de Lisboa, Portugal
Detlef Nauck, BTexact Technologies, UK
Tim Oates, University of Maryland Baltimore County, USA
Simon Parsons, Brooklyn College, City University of New York, USA
José M. Peña, Universidad Politécnica de Madrid, Spain

María S. Pérez, Universidad Politécnica de Madrid, Spain
Bhanu Prasad, Florida A&M University, USA
Víctor Robles, Universidad Politécnica de Madrid, Spain
Lorenza Saitta, Università del Piemonte Orientale, Italy
Paola Sebastiani, Boston University School of Public Health, USA
Arno Siebes, Universiteit Utrecht, The Netherlands
Maarten van Someren, University of Amsterdam, The Netherlands
Myra Spiliopoulou, Otto-von-Guericke-Universität Magdeburg, Germany
Martin Spott, BTexact Technologies, UK
Reinhard Viertl, Vienna University of Technology, Austria
Richard Weber, University of Chile, Chile
Stefan Wrobel, Fraunhofer AIS & University of Bonn, Germany
Mohammed Zaki, Rensselaer Polytechnic Institute, USA

Referees

Silvia Acid
David Auber
Roland Barriot
Concha Bielza
Bouchra Bouqata
Kai Broszat
Andres Cano
Javier G. Castellano
Nicolas Cebron
Víctor Uc Cetina
T.K. Cocx
Nuno Correia
Óscar Cubo
Santiago Eibe
Lukas C. Faulstich
Juan M. Fernández-Luna
Fulvia Ferrazzi
Manuel Gómez
Daniel Goehring
Edgar de Graaf
J.M. de Graaf
Jose A. Gámez
Mark Hall
Alexander Hinneburg
Susanne Hoche
Geoff Holmes
Rainer Holve
Tamás Horváth
Juan F. Huete
Fabien Jourdan
Florian Kaiser
Joerg Kindermann
Christine Koerner
Antonio LaTorre
Marie-Jeanne Lesot
Andres L. Martinez
Michael Mayo
Thorsten Meinl
Ernestina Menasalvas
Dagmar Monett
Serafín Moral
Siegfried Nijssen
Juan A. Fernández del Pozo
Simon Price
Jose M. Puerta
Simon Rawles
Frank Rügheimer
Lucia Sacchi
Alberto Sánchez
Karlton Sequeira
Zujun Shentu
David James Sherman
Hendrik Stange
Micheal Syrjakow
Xiaomeng Wang
Bernd Wiswedel
Marta Elena Zorrilla

Lecture Notes in Computer Science

For information about Vols. 1–3569

please contact your bookseller or Springer

Vol. 3697: W. Duch, J. Kacprzyk, E. Oja, S. Zadrożny (Eds.), Artificial Neural Networks: Formal Models and Their Applications - ICANN 2005, Part II. XXXII, 1045 pages. 2005.

Vol. 3696: W. Duch, J. Kacprzyk, E. Oja, S. Zadrożny (Eds.), Artificial Neural Networks: Biological Inspirations - ICANN 2005, Part I. XXXI, 703 pages. 2005.

Vol. 3687: S. Singh, M. Singh, C. Apte, P. Perner (Eds.), Pattern Recognition and Image Analysis, Part II. XXV, 809 pages. 2005.

Vol. 3686: S. Singh, M. Singh, C. Apte, P. Perner (Eds.), Pattern Recognition and Data Mining, Part I. XXVI, 689 pages. 2005.

Vol. 3674: W. Jonker, M. Petković (Eds.), Secure Data Management. X, 241 pages. 2005.

Vol. 3672: C. Hankin, I. Siveroni (Eds.), Static Analysis. X, 369 pages. 2005.

Vol. 3671: S. Bressan, S. Ceri, E. Hunt, Z.G. Ives, Z. Bellahsène, M. Rys, R. Unland (Eds.), Database and XML Technologies. X, 239 pages. 2005.

Vol. 3670: M. Bravetti, L. Kloul, G. Zavattaro (Eds.), Formal Techniques for Computer Systems and Business Processes. XIII, 349 pages. 2005.

Vol. 3664: C. Türker, M. Agosti, H.-J. Schek (Eds.), Peer-to-Peer, Grid, and Service-Orientation in Digital Library Architectures. X, 261 pages. 2005.

Vol. 3663: W. Kropatsch, R. Sablatnig, A. Hanbury (Eds.), Pattern Recognition. XIV, 512 pages. 2005.

Vol. 3662: C. Baral, G. Greco, N. Leone, G. Terracina (Eds.), Logic Programming and Nonmonotonic Reasoning. XIII, 454 pages. 2005. (Subseries LNAI).

Vol. 3660: M. Beigl, S. Intille, J. Rekimoto, H. Tokuda (Eds.), UbiComp 2005: Ubiquitous Computing. XVII, 394 pages. 2005.

Vol. 3659: J.R. Rao, B. Sunar (Eds.), Cryptographic Hardware and Embedded Systems – CHES 2005. XIV, 458 pages. 2005.

Vol. 3658: V. Matoušek, P. Mautner, T. Pavelka (Eds.), Text, Speech and Dialogue. XV, 460 pages. 2005. (Subseries LNAI).

Vol. 3654: S. Jajodia, D. Wijesekera (Eds.), Data and Applications Security XIX. X, 353 pages. 2005.

Vol. 3653: M. Abadi, L.d. Alfaro (Eds.), CONCUR 2005 – Concurrency Theory. XIV, 578 pages. 2005.

Vol. 3649: W.M.P. van der Aalst, B. Benatallah, F. Casati, F. Curbera (Eds.), Business Process Management. XII, 472 pages. 2005.

Vol. 3648: J.C. Cunha, P.D. Medeiros (Eds.), Euro-Par 2005 Parallel Processing. XXXVI, 1299 pages. 2005.

Vol. 3646: A. F. Famili, J.N. Kok, J.M. Peña, A. Siebes, A. Feelders (Eds.), Advances in Intelligent Data Analysis VI. XIV, 522 pages. 2005.

Vol. 3645: D.-S. Huang, X.-P. Zhang, G.-B. Huang (Eds.), Advances in Intelligent Computing, Part II. XIII, 1010 pages. 2005.

Vol. 3644: D.-S. Huang, X.-P. Zhang, G.-B. Huang (Eds.), Advances in Intelligent Computing, Part I. XXVII, 1101 pages. 2005.

Vol. 3642: D. Ślezak, J. Yao, J.F. Peters, W. Ziarko, X. Hu (Eds.), Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, Part II. XXIV, 738 pages. 2005. (Subseries LNAI).

Vol. 3641: D. Ślezak, G. Wang, M.S. Szczuka, I. Düntsch, Y. Yao (Eds.), Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, Part I. XXIV, 742 pages. 2005. (Subseries LNAI).

Vol. 3639: P. Godefroid (Ed.), Model Checking Software. XI, 289 pages. 2005.

Vol. 3638: A. Butz, B. Fisher, A. Krüger, P. Olivier (Eds.), Smart Graphics. XI, 269 pages. 2005.

Vol. 3637: J. M. Moreno, J. Madrenas, J. Cosp (Eds.), Evolvable Systems: From Biology to Hardware. XI, 227 pages. 2005.

Vol. 3636: M.J. Blesa, C. Blum, A. Roli, M. Sampels (Eds.), Hybrid Metaheuristics. XII, 155 pages. 2005.

Vol. 3634: L. Ong (Ed.), Computer Science Logic. XI, 567 pages. 2005.

Vol. 3633: C. Bauzer Medeiros, M. Egenhofer, E. Bertino (Eds.), Advances in Spatial and Temporal Databases. XIII, 433 pages. 2005.

Vol. 3632: R. Nieuwenhuis (Ed.), Automated Deduction – CADE-20. XIII, 459 pages. 2005. (Subseries LNAI).

Vol. 3629: J.L. Fiadeiro, N. Harman, M. Roggenbach, J. Rutten (Eds.), Algebra and Coalgebra in Computer Science. XI, 457 pages. 2005.

Vol. 3628: T. Gschwind, U. Aßmann, O. Nierstrasz (Eds.), Software Composition. X, 199 pages. 2005.

Vol. 3627: C. Jacob, M.L. Pilat, P.J. Bentley, J. Timmis (Eds.), Artificial Immune Systems. XII, 500 pages. 2005.

Vol. 3626: B. Ganter, G. Stumme, R. Wille (Eds.), Formal Concept Analysis. X, 349 pages. 2005. (Subseries LNAI).

Vol. 3625: S. Kramer, B. Pfahringer (Eds.), Inductive Logic Programming. XIII, 427 pages. 2005. (Subseries LNAI).

Vol. 3624: C. Chekuri, K. Jansen, J.D.P. Rolim, L. Trevisan (Eds.), Approximation, Randomization and Combinatorial Optimization. XI, 495 pages. 2005.

Vol. 3623: M. Liśkiewicz, R. Reischuk (Eds.), Fundamentals of Computation Theory. XV, 576 pages. 2005.

Vol. 3621: V. Shoup (Ed.), Advances in Cryptology – CRYPTO 2005. XI, 568 pages. 2005.

Vol. 3620: H. Muñoz-Avila, F. Ricci (Eds.), Case-Based Reasoning Research and Development. XV, 654 pages. 2005. (Subseries LNAI).

Vol. 3619: X. Lu, W. Zhao (Eds.), Networking and Mobile Computing. XXIV, 1299 pages. 2005.

Vol. 3618: J. Jedrzejowicz, A. Szepietowski (Eds.), Mathematical Foundations of Computer Science 2005. XVI, 814 pages. 2005.

Vol. 3617: F. Roli, S. Vitulano (Eds.), Image Analysis and Processing – ICIAP 2005. XXIV, 1219 pages. 2005.

Vol. 3615: B. Ludäscher, L. Raschid (Eds.), Data Integration in the Life Sciences. XII, 344 pages. 2005. (Subseries LNBI).

Vol. 3614: L. Wang, Y. Jin (Eds.), Fuzzy Systems and Knowledge Discovery, Part II. XLI, 1314 pages. 2005. (Subseries LNAI).

Vol. 3613: L. Wang, Y. Jin (Eds.), Fuzzy Systems and Knowledge Discovery, Part I. XLI, 1334 pages. 2005. (Subseries LNAI).

Vol. 3612: L. Wang, K. Chen, Y. S. Ong (Eds.), Advances in Natural Computation, Part III. LXI, 1326 pages. 2005.

Vol. 3611: L. Wang, K. Chen, Y. S. Ong (Eds.), Advances in Natural Computation, Part II. LXI, 1292 pages. 2005.

Vol. 3610: L. Wang, K. Chen, Y. S. Ong (Eds.), Advances in Natural Computation, Part I. LXI, 1302 pages. 2005.

Vol. 3608: F. Dehne, A. López-Ortiz, J.-R. Sack (Eds.), Algorithms and Data Structures. XIV, 446 pages. 2005.

Vol. 3607: J.-D. Zucker, L. Saitta (Eds.), Abstraction, Reformulation and Approximation. XII, 376 pages. 2005. (Subseries LNAI).

Vol. 3606: V. Malyshkin (Ed.), Parallel Computing Technologies. XII, 470 pages. 2005.

Vol. 3604: R. Martin, H. Bez, M. Sabin (Eds.), Mathematics of Surfaces XI. IX, 473 pages. 2005.

Vol. 3603: J. Hurd, T. Melham (Eds.), Theorem Proving in Higher Order Logics. IX, 409 pages. 2005.

Vol. 3602: R. Eigenmann, Z. Li, S.P. Midkiff (Eds.), Languages and Compilers for High Performance Computing. IX, 486 pages. 2005.

Vol. 3599: U. Aßmann, M. Aksit, A. Rensink (Eds.), Model Driven Architecture. X, 235 pages. 2005.

Vol. 3598: H. Murakami, H. Nakashima, H. Tokuda, M. Yasumura, Ubiquitous Computing Systems. XIII, 275 pages. 2005.

Vol. 3597: S. Shimojo, S. Ichii, T.W. Ling, K.-H. Song (Eds.), Web and Communication Technologies and Internet-Related Social Issues - HSI 2005. XIX, 368 pages. 2005.

Vol. 3596: F. Dau, M.-L. Mugnier, G. Stumme (Eds.), Conceptual Structures: Common Semantics for Sharing Knowledge. XI, 467 pages. 2005. (Subseries LNAI).

Vol. 3595: L. Wang (Ed.), Computing and Combinatorics. XVI, 995 pages. 2005.

Vol. 3594: J.C. Setubal, S. Verjovski-Almeida (Eds.), Advances in Bioinformatics and Computational Biology. XIV, 258 pages. 2005. (Subseries LNBI).

Vol. 3593: V. Mařík, R. W. Brennan, M. Pěchouček (Eds.), Holonic and Multi-Agent Systems for Manufacturing. XI, 269 pages. 2005. (Subseries LNAI).

Vol. 3592: S. Katsikas, J. Lopez, G. Pernul (Eds.), Trust, Privacy and Security in Digital Business. XII, 332 pages. 2005.

Vol. 3591: M.A. Wimmer, R. Traunmüller, Å. Grönlund, K.V. Andersen (Eds.), Electronic Government. XIII, 317 pages. 2005.

Vol. 3590: K. Bauknecht, B. Pröll, H. Werthner (Eds.), E-Commerce and Web Technologies. XIV, 380 pages. 2005.

Vol. 3589: A M. Tjoa, J. Trujillo (Eds.), Data Warehousing and Knowledge Discovery. XVI, 538 pages. 2005.

Vol. 3588: K.V. Andersen, J. Debenham, R. Wagner (Eds.), Database and Expert Systems Applications. XX, 955 pages. 2005.

Vol. 3587: P. Perner, A. Imiya (Eds.), Machine Learning and Data Mining in Pattern Recognition. XVII, 695 pages. 2005. (Subseries LNAI).

Vol. 3586: A.P. Black (Ed.), ECOOP 2005 - Object-Oriented Programming. XVII, 631 pages. 2005.

Vol. 3584: X. Li, S. Wang, Z.Y. Dong (Eds.), Advanced Data Mining and Applications. XIX, 835 pages. 2005. (Subseries LNAI).

Vol. 3583: R.W. H. Lau, Q. Li, R. Cheung, W. Liu (Eds.), Advances in Web-Based Learning – ICWL 2005. XIV, 420 pages. 2005.

Vol. 3582: J. Fitzgerald, I.J. Hayes, A. Tarlecki (Eds.), FM 2005: Formal Methods. XIV, 558 pages. 2005.

Vol. 3581: S. Miksch, J. Hunter, E. Keravnou (Eds.), Artificial Intelligence in Medicine. XVII, 547 pages. 2005. (Subseries LNAI).

Vol. 3580: L. Caires, G.F. Italiano, L. Monteiro, C. Palamidessi, M. Yung (Eds.), Automata, Languages and Programming. XXV, 1477 pages. 2005.

Vol. 3579: D. Lowe, M. Gaedke (Eds.), Web Engineering. XXII, 633 pages. 2005.

Vol. 3578: M. Gallagher, J. Hogan, F. Maire (Eds.), Intelligent Data Engineering and Automated Learning - IDEAL 2005. XVI, 599 pages. 2005.

Vol. 3577: R. Falcone, S. Barber, J. Sabater-Mir, M.P. Singh (Eds.), Trusting Agents for Trusting Electronic Societies. VIII, 235 pages. 2005. (Subseries LNAI).

Vol. 3576: K. Etessami, S.K. Rajamani (Eds.), Computer Aided Verification. XV, 564 pages. 2005.

Vol. 3575: S. Wermter, G. Palm, M. Elshaw (Eds.), Biomimetic Neural Learning for Intelligent Robots. IX, 383 pages. 2005. (Subseries LNAI).

Vol. 3574: C. Boyd, J.M. González Nieto (Eds.), Information Security and Privacy. XIII, 586 pages. 2005.

Vol. 3573: S. Etalle (Ed.), Logic Based Program Synthesis and Transformation. VIII, 279 pages. 2005.

Vol. 3572: C. De Felice, A. Restivo (Eds.), Developments in Language Theory. XI, 409 pages. 2005.

Vol. 3571: L. Godo (Ed.), Symbolic and Quantitative Approaches to Reasoning with Uncertainty. XVI, 1028 pages. 2005. (Subseries LNAI).

Vol. 3570: A. S. Patrick, M. Yung (Eds.), Financial Cryptography and Data Security. XII, 376 pages. 2005.

Table of Contents

Probabilistic Latent Clustering of Device Usage

Jean-Marc Andreoli and Guillaume Bouchard

Xerox Research Centre Europe, Grenoble, France
FirstName.LastName@xrce.xerox.com

Abstract. We investigate an application of Probabilistic Latent Semantics to the problem of device usage analysis in an infrastructure in which multiple users have access to a shared pool of devices delivering different kinds of service and service levels. Each invocation of a service by a user, called a job, is assumed to be logged simply as a co-occurrence of the identifier of the user and that of the device used. The data is best modelled by assuming that multiple latent variables (instead of a single one as in traditional PLSA) satisfying different types of constraints explain the observed variables of a job. We discuss the application of our model to the printing infrastructure in an office environment.

1 Introduction

It is nowadays common that printing devices in an office or a workplace be accessed through the local network instead of being assigned and directly connected to individual desktops. As a result, a large amount of information can easily be collected about the actual use of the whole printing infrastructure, rather than individual devices. To be useful, this data needs to be analysed and presented in a synthetic way to the administrators of the infrastructure. We are interested here in analysing the correlation between users and devices in the data, ie. how the printing potential of users translates into actual use of the devices. We assume here that users are not strongly constrained in their use, the extreme case being when any user is allowed to print anything on any device in the infrastructure. The expected outcome of such an analysis may be diverse. For example, the administrator could discover communities of device usage, corresponding to different physical or virtual locations of the users at the time of the jobs, and, from these, form hypotheses on the actual behaviour of the users, both in the case of normal functioning of the infrastructure and in case of exceptions (device down or not working properly). This in turn could lead to more refined decisions as to the organisation of the infrastructure and to the instructions given to its users. It could also help work around failures of devices inside the infrastructure, by redirecting a job sent to a failing device toward a working one chosen in accordance with the community to which the job belongs.

A study on inhabitant-device interactions [6] shows that the recorded device usage can be mined to discover significant patterns, which in turn could be used to automate device interactions. To the authors knowledge, generic user-device interaction analysis in the presence of devices delivering possibly multiple services or levels of service has not been studied extensively.

A.F. Famili et al. (Eds.): IDA 2005, LNCS 3646, pp. 1–11, 2005.

Problem statement. Our overall goal is to analyse usage data in an infrastructure consisting of a set of independent devices offering services of different or identical classes, and operated by a set of independent users. An interaction between a user and a device is called a job. The usage data consists of a log of these jobs over a given period of time. More precisely, we make the following assumptions.

- Let N_U, N_D, N_K denote the number of, respectively, users, devices and service classes, assumed invariable over the analysed period. Each user, resp. device, resp. service class, can therefore be identified by a number $u \in \{1, \ldots, N_U\}$, resp. $d \in \{1, \ldots, N_D\}$, resp. $k \in \{1, \ldots, N_K\}$. Each user, device, service class also has a print name, for display and reference purpose.
- Each device offers services of one or more classes. This is captured in a boolean matrix f of dimension $N_K \times N_D$ where f_{kd} is 1 if device d offers the service class k and 0 otherwise. This matrix is assumed static over the analysed period.
- All the jobs are recorded over the analysed period. Let N be the number of recorded jobs. Each job can therefore be identified by an index $i \in \{1, \ldots, N\}$. Each job i contributes exactly one entry in the log, consisting of the pair (u_i, d_i) identifying the user and device involved in that job. Thus the data is entirely defined by the matrix n of dimension $N_U \times N_D$ where n_{ud} is the number of jobs by user u on device d.

A printing infrastructure in an office is a typical example where our method applies. In that case, a service class could be a particular type of printing. For simplification purpose, in the examples, we consider only two service classes: black&white ($k = 1$) and colour ($k = 2$). Note that a colour printer can always also perform black&white jobs, meaning that if $f_{2d} = 1$, then $f_{1d} = 1$.

Outline of the method. The purpose of our analysis is essentially to discover clusters in the usage data. Since the observed data correspond to co-occurrences of discrete variables, we have chosen an aspect model, which is an instance of latent class models [1], nowadays often referred to as Probabilistic Latent Semantics Analysis (PLSA) [5]. This model is particularly relevant here as its basic assumption has a straightforward interpretation in our context. Indeed, the PLSA assumption is that the data can be generated according to a process that first selects a (latent) cluster, then a user and a device, in such a way that, conditionally to the cluster, the choices of user and device are independent. There is a natural interpretation of such clusters as communities of usage which are associated to physical or virtual locations within the infrastructure. The PLSA assumption means that at a given location, users tend to choose devices in the same way, which is quite reasonable. For example, in an office infrastructure comprising multiple floors, each floor can correspond to a community, whose users share the same perception of the infrastructure and tend to choose printers in a similar fashion. PLSA clustering therefore offers a powerful tool to discover such communities of usage.

However, another important determining factor for the choice of device is the nature of the job to be performed. This information may not be directly available