Fabio Crestani
Paolo Ferragina
Mark Sanderson (Eds.)

# String Processing and Information Retrieval

LNCS 4209

13th International Conference, SPIRE 2006
Glasgow, UK, October 2006
Proceedings

Springer

Fabio Crestani Paolo Ferragina
Mark Sanderson (Eds.)

# String Processing and Information Retrieval

13th International Conference, SPIRE 2006
Glasgow, UK, October 11-13, 2006
Proceedings

江苏工业学院图书馆
藏书章

Springer

Volume Editors

Fabio Crestani
University of Strathclyde
Department of Computer and Information Sciences
16 Richmond Street, Glasgow G12 0NX, UK
E-mail: f.crestani@cis.strath.ac.uk

Paolo Ferragina
University of Pisa
Department of Computer Science
Largo Bruno Pontecorvo 3, 56127 Pisa, Italy
E-mail: ferragina@di.unipi.it

Mark Sanderson
University of Sheffield
Department of Information Studies
Regent Court, 211 Portobello St, Sheffield, S1 4DP, UK
E-mail: m.sanderson@shef.ac.uk

Library of Congress Control Number: 2006932965

CR Subject Classification (1998): H.3, H.2.8, I.2, E.1, E.5, F.2.2

LNCS Sublibrary: SL 1 – Theoretical Computer Science and General Issues

ISSN 0302-9743
ISBN-10 3-540-45774-7 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-45774-9 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 2006
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11880561 06/3142 5 4 3 2 1 0

# Preface

This volume contains the papers presented at the 13th International Symposium on String Processing and Information Retrieval (SPIRE), held October 11-13, 2006, in Glasgow, Scotland.

The SPIRE annual symposium provides an opportunity for both new and established researchers to present original contributions to areas such as string processing (dictionary algorithms, text searching, pattern matching, text compression, text mining, natural language processing, and automata-based string processing); information retrieval languages, applications, and evaluation (IR modelling, indexing, ranking and filtering, interface design, visualization, cross-lingual IR systems, multimedia IR, digital libraries, collaborative retrieval, Web-related applications, XML, information retrieval from semi-structured data, text mining, and generation of structured data from text); and interaction of biology and computation (sequencing and applications in molecular biology, evolution and phylogenetics, recognition of genes and regulatory elements, and sequence-driven protein structure prediction).

The papers in this volume were selected from 102 papers submitted from over 20 different countries in response to the Call for Papers. A total of 26 submissions were accepted as full papers, yielding an acceptance rate of about 25%. In view of the large number of good-quality submissions the Program Committee decided to accept 5 short papers, that have also been included in the proceedings. SPIRE 2006 also featured two talks by invited speakers: Jamie Callan (Carnegie Mellon University, USA) and Martin Farach-Colton (Rutgers University, USA).

The Organizing Committee would like to thank all the authors who submitted their work for consideration and the participants of SPIRE 2006 for making the event a great success.

Special thanks are due to the members of the Program Committee who worked very hard to ensure the timely review of all the submitted manuscripts, and to the invited speakers, Jamie Callan and Martin Farach-Colton, for their inspiring presentations. We also would like to thank the sponsoring institutions, EPSRC (Engineering and Physical Sciences Research Council) , Yahoo! Research, the Kelvin Institute, the BCS-IRSG (British Computer Society - Information Retrieval Specialist Group), and the University of Strathclyde, for their generous financial and institutional support, and Glasgow City Council for civic hospitality.

Thanks are due to the editorial staff at Springer for their agreement to publish the colloquium proceedings as part of the *Lecture Notes in Computer Science* series.

Thanks are also due to the local team of student volunteers (in particular Mark Baillie, Murat Yakici and Emma Nicol), the secretaries (Carol-Ann Seath and Linda Hunter), and the information officer (Paul Smith), whose efforts ensured the smooth organization and running of the event.

Finally, we would like to thank Ricardo Baeza-Yates, who, on behalf of the Steering Committee, invited us to organize SPIRE 2006 and supported us at every step of the way.

October 2006

Fabio Crestani
Paolo Ferragina
Mark Sanderson

试读结束，需要全本PDF请购买 www.ertongbook.com

# SPIRE 2006 Organization

## Organizing Institution

SPIRE 2006 was organized by the Department of Computer and Information Sciences of the University of Strathclyde and held at the Teacher Building in Glasgow, Scotland, UK.

## Sponsoring Institutions

Engineering and Physical Sciences Research Council, UK.
Yahoo! Research, Barcellona, Spain.
Kelvin Institute, Glasgow, Scotland, UK.
British Computer Society - Information Retrieval Specialist Group, UK.
Glasgow City Council, Glasgow, Scotland, UK.

## Organizing Committee

General Chair: Fabio Crestani (University of Strathclyde, UK)

Program Committee Chairs: Paolo Ferragina (University of Pisa, Italy) and Mark Sanderson (University of Sheffield, UK)

## Program Committee

Gianni Amati (Fondazione Ugo Bordoni, Italy)
Amihood Amir (University Bar-Ilan, Israel and Georgia Tech, USA)
Alberto Apostolico (Georgia Tech, USA and University of Padua, Italy)
Ricardo Baeza-Yates (Yahoo! Research, Spain and Chile)
Michael Bender (Stony Brook University, USA)
Mohand Boughanem (University of Tolouse, France)
Giorgio Brajnik (University of Udine, Italy)
Gerth S. Brodal (University of Aarhus, Denmark)
Paul Browne (Imperial College, UK)
Chris Buckley (Sabir Research, USA)
Mariano Consens (University of Toronto, Canada)
Nick Craswell (Microsoft Research, UK)
Maxime Crochemore (University of Marne-la-Vallée, France)
Bruce Croft (University of Massachusetts at Amherst, USA)
Erik Demaine (MIT, USA)
Martin Farach-Colton (Rutgers University, USA)
Edward Fox (Virginia Tech, USA)

Norbert Fuhr (University of Duisburg-Essen, Germany )
Eric Gaussier (Xerox-RCE, France)
Raffaele Giancarlo (University of Palermo, Italy)
Mark Girolami (University of Glasgow, UK)
Nazli Goharian (IIT, USA)
Enrique Herrera-Viedma (University of Granada, Spain)
Costas Iliopoulos (King's College London, UK)
Joemon Jose (University of Glasgow, UK)
Juha Kärkkäinen (University of Helsinki, Finland)
Jussi Karlgren (SICS, Sweden)
Mounia Lalmas (Queen Mary, University of London, UK)
Gadi Landau (University of Haifa, Israel and Polytechnic University, NY, USA)
Hans-Peter Lenhof (University of Saarbrücken, Germany)
Moshe Lewenstein (University Bar-Ilan, Israel)
Stefano Lonardi (University of California Riverside, USA)
David Losada (University of Santiago de Compostela, Spain)
Andrew MacFarlane (City University, London, UK)
Veli Mäkinen (University of Helsinki, Finland)
Giovanni Manzini (University of Piemonte Orientale, Italy)
Paul McNamee (JHU, USA )
Massimo Melucci (University of Padova, Italy)
Alistair Moffat (University of Melbourne, Australia)
Gonzalo Navarro (University of Chile, Chile)
Paul Ogilvie (CMU, USA)
Arlindo Oliveira (INESC-ID/Technical University of Lisbon, Portugal)
Pietro Pala (University of Firenze, Italy)
Gabriella Pasi (Università degli Studi di Milano Bicocca, Italy)
Mathieu Raffinot (CNRS, France)
Rajeev Raman (Leicester University, UK)
Andreas Rauber (Technical University of Vienna, Austria)
Crawford Revie (University of Strathclyde, UK)
Keith van Rijsbergen (University of Glasgow, UK)
Ian Ruthven (University of Strathclyde, UK)
Kunihiko Sadakane (Kyushu University, Japan)
Marie-France Sagot (INRIA Rhone-Alpes, France)
Falk Scholer (RMIT, Australia)
Steven Skiena (Stony Brook University, USA)
Ian Soboroff (NIST, USA )
Jens Stoye (University of Bielefeld, Germany)
Tassos Tombros (Queen Mary, University of London, UK)
Andrew Trotman (Otago, New Zealand)
Andrew Turpin (RMIT, Australia)
Sebastiano Vigna (Università degli Studi di Milano, Italy)
Arjen P. de Vries (CWI, The Netherlands)
Peter Widmayer (ETH Zurich, Switzerland )

Yi Zhang (University of California Santa Cruz, USA)
Nivio Ziviani (Federal University of Minas Gerais, Brazil)
Roelof van Zwol (Utrecht University, Netherlands)

## Additional Reviewers

José Augusto Amgarten Quitzao, Vo Ngoc Anh, Diego Arroyuelo, Elham Ashoori, Claudine Badue, Bodo Billerbeck, Guillaume Blin, Ciccio Bozza, Pavel Calado, Ana Cardoso-Cachopo, Carlos Castillo , Jean-Marc Champarnaud, Massi Ciaramita, Raphael Clifford, Luís Coelho, Roberto Cornacchia, Thierson Couto Rosa, Marco Antonio Cristo, J. Shane Culpepper, Fabiano Cupertino Botelho, Shiri Dori, Gudrun Fischer, Matthias Fitzi, Ingo Frommholz, Lilia Greenenko, Sàndor Héman, MohammadTaghi Hajiaghayi, Danny Hermelin, Andreas Hildebrandt, Jan Holub, Sarvnaz Karimi, Shahar Keret, Tsvi Kopelowits, Adrian Kosowski, Thierry Lecroq, Liat Leventhal, Kan Liu, Sabrina Mantaci, Rudolf Mayer, Laurent Mouchard, Joong Chae Na, Nitsan Oz, Andreas Pesenhofer, Georg Poelzlbauer, Simon Puglisi, James F. Reid, Eric Rivals, Luís Russo, Klaus-Bernd Schürmann, Marinella Sciortino, Edleno Silva de Moura, Lynda Tamine, Theodora Tsikrika, Alexandra Uitdenbogerd, Marion Videau, Newton Jose Vieira, Jun Wang, Oren Weimann, YongHui Wu.

## Previous Venues of SPIRE

The first four editions focused primarily on *string processing* and were held in South America. At the time SPIRE was called WSP (South American Workshop on String Processing). Starting in 1998, the focus of the workshop was broadened to include the area of *information retrieval* due to its increasing relevance and its inter-relationship with the area of string processing, changing to its current name. In addition, since 2000, the symposium started to alternate between Europe and Latin America, being held in Spain, Chile, Portugal, Brazil, and Italy in the last years. This is the first time that SPIRE was held in the United Kingdom.

2005: Buenos Aires, Argentina
2004: Padova, Italy
2003: Manaus, Brazil
2002: Lisboa, Portugal
2001: Laguna San Rafael, Chile
2000: A Coruna, Spain
1999: Cancun, Mexico
1998: Santa Cruz, Bolivia
1997: Valparaso, Chile
1996: Recife, Brazil
1995: Valparaso, Chile
1993: Belo Horizonte, Brazil

# Lecture Notes in Computer Science

For information about Vols. 1–4132

please contact your bookseller or Springer

Vol. 4238: Y.-T. Kim, M. Takano (Eds.), Management of Convergence Networks and Services. XVIII, 604 pages. 2006.

Vol. 4228: D.E. Lightfoot, C.A. Szyperski (Eds.), Modular Programming Languages. X, 415 pages. 2006.

Vol. 4227: W. Nejdl, K. Tochtermann (Eds.), Innovative Approaches for Learning and Knowledge Sharing. XVII, 721 pages. 2006.

Vol. 4224: E. Corchado, H. Yin, V. Botti, C. Fyfe (Eds.), Intelligent, Data Engineering and, Automated Learning – IDEAL 2006. XXVII, 1447 pages. 2006.

Vol. 4219: D. Zamboni, C. Kruegel (Eds.), Recent Advances in Intrusion Detection. XII, 331 pages. 2006.

Vol. 4217: P. Cuenca, L. Orozco-Barbosa (Eds.), Personal Wireless Communications. XV, 532 pages. 2006.

Vol. 4216: M.R. Berthold, R. Glen, I. Fischer (Eds.), Computational Life Sciences. XIII, 269 pages. 2006. (Sublibrary LNBI).

Vol. 4213: J. Fürnkranz, T. Scheffer, M. Spiliopoulou (Eds.), Knowledge Discovery in Databases: PKDD 2006. XXII, 660 pages. 2006. (Sublibrary LNAI).

Vol. 4212: J. Fürnkranz, T. Scheffer, M. Spiliopoulou (Eds.), Machine Learning: ECML 2006. XXIII, 851 pages. 2006. (Sublibrary LNAI).

Vol. 4211: P. Vogt, Y. Sugita, E. Tuci, C. Nehaniv (Eds.), Symbol Grounding and Beyond. VIII, 237 pages. 2006. (Sublibrary LNAI).

Vol. 4209: F. Crestani, P. Ferragina, M. Sanderson (Eds.), String Processing and Information Retrieval. XIV, 367 pages. 2006.

Vol. 4208: M. Gerndt, D. Kranzlmüller (Eds.), High Performance Computing and Communications. XXII, 938 pages. 2006.

Vol. 4207: Z. Ésik (Ed.), Computer Science Logic. XII, 627 pages. 2006.

Vol. 4206: P. Dourish, A. Friday (Eds.), UbiComp 2006: Ubiquitous Computing. XIX, 526 pages. 2006.

Vol. 4205: G. Bourque, N. El-Mabrouk (Eds.), Comparative Genomics. X, 231 pages. 2006. (Sublibrary LNBI).

Vol. 4203: F. Esposito, Z.W. Ras, D. Malerba, G. Semeraro (Eds.), Foundations of Intelligent Systems. XVIII, 767 pages. 2006. (Sublibrary LNAI).

Vol. 4202: E. Asarin, P. Bouyer (Eds.), Formal Modeling and Analysis of Timed Systems. XI, 369 pages. 2006.

Vol. 4201: Y. Sakakibara, S. Kobayashi, K. Sato, T. Nishino, E. Tomita (Eds.), Grammatical Inference: Algorithms and Applications. XII, 359 pages. 2006. (Sublibrary LNAI).

Vol. 4199: O. Nierstrasz, J. Whittle, D. Harel, G. Reggio (Eds.), Model Driven Engineering Languages and Systems. XVI, 798 pages. 2006. (Sublibrary LNBI).

Vol. 4197: M. Raubal, H.J. Miller, A.U. Frank, M.F. Goodchild (Eds.), Geographic, Information Science. XIII, 419 pages. 2006.

Vol. 4196: K. Fischer, I.J. Timm, E. André, N. Zhong (Eds.), Multiagent System Technologies. X, 185 pages. 2006. (Sublibrary LNAI).

Vol. 4195: D. Gaiti, G. Pujolle, E. Al-Shaer, K. Calvert, S. Dobson, G. Leduc, O. Martikainen (Eds.), Autonomic Networking. IX, 316 pages. 2006.

Vol. 4194: V.G. Ganzha, E.W. Mayr, E.V. Vorozhtsov (Eds.), Computer Algebra in Scientific Computing. XI, 313 pages. 2006.

Vol. 4193: T.P. Runarsson, H.-G. Beyer, E. Burke, J.J. Merelo-Guervós, L. D. Whitley, X. Yao (Eds.), Parallel Problem Solving from Nature - PPSN IX. XIX, 1061 pages. 2006.

Vol. 4192: B. Mohr, J.L. Träff, J. Worringen, J. Dongarra (Eds.), Recent Advances in Parallel Virtual Machine and Message Passing Interface. XVI, 414 pages. 2006.

Vol. 4191: R. Larsen, M. Nielsen, J. Sporring (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006, Part II. XXXVIII, 981 pages. 2006.

Vol. 4190: R. Larsen, M. Nielsen, J. Sporring (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006, Part I. XXXVVIII, 949 pages. 2006.

Vol. 4189: D. Gollmann, J. Meier, A. Sabelfeld (Eds.), Computer Security – ESORICS 2006. XI, 548 pages. 2006.

Vol. 4188: P. Sojka, I. Kopeček, K. Pala (Eds.), Text, Speech and Dialogue. XIV, 721 pages. 2006. (Sublibrary LNAI).

Vol. 4187: J.J. Alferes, J. Bailey, W. May, U. Schwertel (Eds.), Principles and Practice of Semantic Web Reasoning. XI, 277 pages. 2006.

Vol. 4186: C. Jesshope, C. Egan (Eds.), Advances in Computer Systems Architecture. XIV, 605 pages. 2006.

Vol. 4185: R. Mizoguchi, Z. Shi, F. Giunchiglia (Eds.), The Semantic Web – ASWC 2006. XX, 778 pages. 2006.

Vol. 4184: M. Bravetti, M. Núñez, G. Zavattaro (Eds.), Web Services and Formal Methods. X, 289 pages. 2006.

Vol. 4183: J. Euzenat, J. Domingue (Eds.), Artificial Intelligence: Methodology, Systems, and Applications. XIII, 291 pages. 2006. (Sublibrary LNAI).

Vol. 4182: H.T. Ng, M.-K. Leong, M.-Y. Kan, D. Ji (Eds.), Information Retrieval Technology. XVI, 684 pages. 2006.

Vol. 4180: M. Kohlhase, OMDoc – An Open Markup Format for Mathematical Documents [version 1.2]. XIX, 428 pages. 2006. (Sublibrary LNAI).

Vol. 4179: J. Blanc-Talon, W. Philips, D. Popescu, P. Scheunders (Eds.), Advanced Concepts for Intelligent Vision Systems. XXIV, 1224 pages. 2006.

Vol. 4178: A. Corradini, H. Ehrig, U. Montanari, L. Ribeiro, G. Rozenberg (Eds.), Graph Transformations. XII, 473 pages. 2006.

Vol. 4176: S.K. Katsikas, J. Lopez, M. Backes, S. Gritzalis, B. Preneel (Eds.), Information Security. XIV, 548 pages. 2006.

Vol. 4175: P. Bücher, B.M.E. Moret (Eds.), Algorithms in Bioinformatics. XII, 402 pages. 2006. (Sublibrary LNBI).

Vol. 4174: K. Franke, K.-R. Müller, B. Nickolay, R. Schäfer (Eds.), Pattern Recognition. XX, 773 pages. 2006.

Vol. 4173: S. El Yacoubi, B. Chopard, S. Bandini (Eds.), Cellular Automata. XV, 734 pages. 2006.

Vol. 4172: J. Gonzalo, C. Thanos, M. F. Verdejo, R.C. Carrasco (Eds.), Research and Advanced Technology for Digital Libraries. XVII, 569 pages. 2006.

Vol. 4169: H.L. Bodlaender, M.A. Langston (Eds.), Parameterized and Exact Computation. XI, 279 pages. 2006.

Vol. 4168: Y. Azar, T. Erlebach (Eds.), Algorithms – ESA 2006. XVIII, 843 pages. 2006.

Vol. 4167: S. Dolev (Ed.), Distributed Computing. XV, 576 pages. 2006.

Vol. 4166: J. Górski (Ed.), Computer Safety, Reliability, and Security. XIV, 440 pages. 2006.

Vol. 4165: W. Jonker, M. Petković (Eds.), Secure, Data Management. X, 185 pages. 2006.

Vol. 4163: H. Bersini, J. Carneiro (Eds.), Artificial Immune Systems. XII, 460 pages. 2006.

Vol. 4162: R. Královič, P. Urzyczyn (Eds.), Mathematical Foundations of Computer Science 2006. XV, 814 pages. 2006.

Vol. 4161: R. Harper, M. Rauterberg, M. Combetto (Eds.), Entertainment Computing - ICEC 2006. XXVII, 417 pages. 2006.

Vol. 4160: M. Fisher, W.v.d. Hoek, B. Konev, A. Lisitsa (Eds.), Logics in Artificial Intelligence. XII, 516 pages. 2006. (Sublibrary LNAI).

Vol. 4159: J. Ma, H. Jin, L.T. Yang, J.J.-P. Tsai (Eds.), Ubiquitous Intelligence and Computing. XXII, 1190 pages. 2006.

Vol. 4158: L.T. Yang, H. Jin, J. Ma, T. Ungerer (Eds.), Autonomic and Trusted Computing. XIV, 613 pages. 2006.

Vol. 4156: S. Amer-Yahia, Z. Bellahsène, E. Hunt, R. Unland, J.X. Yu (Eds.), Database and XML Technologies. IX, 123 pages. 2006.

Vol. 4155: O. Stock, M. Schaerf (Eds.), Reasoning, Action and Interaction in AI Theories and Systems. XVIII, 343 pages. 2006. (Sublibrary LNAI).

Vol. 4154: Y.A. Dimitriadis, I. Zigurs, E. Gómez-Sánchez (Eds.), Groupware: Design, Implementation, and Use. XIV, 438 pages. 2006.

Vol. 4153: N. Zheng, X. Jiang, X. Lan (Eds.), Advances in Machine Vision, Image Processing, and Pattern Analysis. XIII, 506 pages. 2006.

Vol. 4152: Y. Manolopoulos, J. Pokorný, T. Sellis (Eds.), Advances in Databases and Information Systems. XV, 448 pages. 2006.

Vol. 4151: A. Iglesias, N. Takayama (Eds.), Mathematical Software - ICMS 2006. XVII, 452 pages. 2006.

Vol. 4150: M. Dorigo, L.M. Gambardella, M. Birattari, A. Martinoli, R. Poli, T. Stützle (Eds.), Ant Colony Optimization and Swarm Intelligence. XVI, 526 pages. 2006.

Vol. 4149: M. Klusch, M. Rovatsos, T.R. Payne (Eds.), Cooperative Information Agents X. XII, 477 pages. 2006. (Sublibrary LNAI).

Vol. 4148: J. Vounckx, N. Azemard, P. Maurine (Eds.), Integrated Circuit and System Design. XVI, 677 pages. 2006.

Vol. 4147: M. Broy, I.H. Krüger, M. Meisinger (Eds.), Automotive Software – Connected Services in Mobile Networks. XIV, 155 pages. 2006.

Vol. 4146: J.C. Rajapakse, L. Wong, R. Acharya (Eds.), Pattern Recognition in Bioinformatics. XIV, 186 pages. 2006. (Sublibrary LNBI).

Vol. 4144: T. Ball, R.B. Jones (Eds.), Computer Aided Verification. XV, 564 pages. 2006.

Vol. 4143: R. Lämmel, J. Saraiva, J. Visser (Eds.), Generative and Transformational Techniques in Software Engineering. X, 471 pages. 2006.

Vol. 4142: A. Campilho, M. Kamel (Eds.), Image Analysis and Recognition, Part II. XXVII, 923 pages. 2006.

Vol. 4141: A. Campilho, M. Kamel (Eds.), Image Analysis and Recognition, Part I. XXVIII, 939 pages. 2006.

Vol. 4139: T. Salakoski, F. Ginter, S. Pyysalo, T. Pahikkala, Advances in Natural Language Processing. XVI, 771 pages. 2006. (Sublibrary LNAI).

Vol. 4138: X. Cheng, W. Li, T. Znati (Eds.), Wireless Algorithms, Systems, and Applications. XVI, 709 pages. 2006.

Vol. 4137: C. Baier, H. Hermanns (Eds.), CONCUR 2006 – Concurrency Theory. XIII, 525 pages. 2006.

Vol. 4136: R.A. Schmidt (Ed.), Relations and Kleene Algebra in Computer Science. XI, 433 pages. 2006.

Vol. 4135: C.S. Calude, M.J. Dinneen, G. Păun, G. Rozenberg, S. Stepney (Eds.), Unconventional Computation. X, 267 pages. 2006.

Vol. 4134: K. Yi (Ed.), Static Analysis. XIII, 443 pages. 2006.

Vol. 4133: J. Gratch, M. Young, R. Aylett, D. Ballin, P. Olivier (Eds.), Intelligent Virtual Agents. XIV, 472 pages. 2006. (Sublibrary LNAI).

# Table of Contents

## Web Clustering and Text Categorization



## Strings



## User Behavior



试读结束，需要全本PDF请购买 www.ertongbook.com

## Web Search Algorithms



## Compression



## Correction



## Information Retrieval Applications

## Bio Informatics



## Web Search Engines



## Short Papers

# MP-Boost: A Multiple-Pivot Boosting Algorithm and Its Application to Text Categorization

Andrea Esuli, Tiziano Fagni, and Fabrizio Sebastiani

Istituto di Scienza e Tecnologia dell'Informazione
Consiglio Nazionale delle Ricerche
Via Giuseppe Moruzzi 1 – 56124 Pisa, Italy
{andrea.esuli, tiziano.fagni, fabrizio.sebastiani}@isti.cnr.it

**Abstract.** AdaBoost.MH is a popular supervised learning algorithm for building multi-label (aka $n$-of-$m$) text classifiers. AdaBoost.MH belongs to the family of "boosting" algorithms, and works by iteratively building a committee of "decision stump" classifiers, where each such classifier is trained to especially concentrate on the document-class pairs that previously generated classifiers have found harder to correctly classify. Each decision stump hinges on a specific "pivot term", checking its presence or absence in the test document in order to take its classification decision. In this paper we propose an improved version of AdaBoost.MH, called MP-Boost, obtained by selecting, at each iteration of the boosting process, not one but *several* pivot terms, one for each category. The rationale behind this choice is that this provides highly individualized treatment for each category, since each iteration thus generates, for each category, the best possible decision stump. We present the results of experiments showing that MP-Boost is much more effective than AdaBoost.MH. In particular, the improvement in effectiveness is spectacular when few boosting iterations are performed, and (only) high for many such iterations. The improvement is especially significant in the case of macroaveraged effectiveness, which shows that MP-Boost is especially good at working with hard, infrequent categories.

## 1 Introduction

Given a set of textual documents $D$ and a predefined set of *categories* (aka *labels*) $C = \{c_1, \ldots, c_m\}$, *multi-label* (aka $n$-of-$m$) *text classification* is the task of approximating, or estimating, an unknown *target function* $\Phi : D \times C \rightarrow \{-1, +1\}$, that describes how documents ought to be classified, by means of a function $\hat{\Phi} : D \times C \rightarrow \{-1, +1\}$, called the *classifier*, such that $\Phi$ and $\hat{\Phi}$ "coincide as much as possible". Here, "multi-label" indicates that the same document can belong to zero, one, or several categories at the same time.

AdaBoost.MH [1] is a popular supervised learning algorithm for building multi-label text classifiers. AdaBoost.MH belongs to the family of "boosting" algorithms (see [2] for a review), which have enjoyed a wide popularity in the text categorization and filtering community because of their state-of-the-art

F. Crestani, P. Ferragina, and M. Sanderson (Eds.): SPIRE 2006, LNCS 4209, pp. 1–12, 2006.
© Springer-Verlag Berlin Heidelberg 2006

effectiveness and of the strong justifications they have received from computational learning theory. AdaBoost.MH works by iteratively building a committee of "decision stump" classifiers[1], where each such classifier is trained to especially concentrate on the document-category pairs that previously generated classifiers have found harder to correctly classify. Each decision stump hinges on a specific "pivot term", and takes its classification decision based on the presence or absence of the pivot term in the test document.

We here propose an improved version of AdaBoost.MH, called MP-Boost, obtained by selecting, at each iteration of the boosting process, not one but *several* pivot terms, one for each category. The rationale behind this choice is that this provides highly individualized treatment for each category, since each iteration generates, for each category, the best possible decision stump. The result of the learning process is thus not a single classifier committee, but a set of such committees, one for each category.

The paper is structured as follows. In Section 2 we concisely describe boosting and the AdaBoost.MH algorithm. Section 3 describes in detail our MP-Boost algorithm and the rationale behind it. In Section 4 we present experimental results comparing AdaBoost.MH and MP-Boost. Section 5 concludes.

## 2 An Introduction to Boosting and AdaBoost.MH

AdaBoost.MH [1] (see Figure 1) is a *boosting* algorithm, i.e. an algorithm that generates a highly accurate classifier (also called *final hypothesis*) by combining a set of moderately accurate classifiers (also called *weak hypotheses*). The input to the algorithm is a training set $Tr = \{\langle d_1, C_1\rangle, \ldots, \langle d_g, C_g\rangle\}$, where $C_i \subseteq C$ is the set of categories to each of which $d_i$ belongs.

AdaBoost.MH works by iteratively calling a *weak learner* to generate a sequence $\hat{\Phi}_1, \ldots, \hat{\Phi}_S$ of weak hypotheses; at the end of the iteration the final hypothesis $\hat{\Phi}$ is obtained as a sum $\hat{\Phi} = \sum_{s=1}^{S} \hat{\Phi}_s$ of these weak hypotheses. A weak hypothesis is a function $\hat{\Phi}_s : D \times C \rightarrow \mathbf{R}$. We interpret the sign of $\hat{\Phi}_s(d_i, c_j)$ as the prediction of $\hat{\Phi}_s$ on whether $d_i$ belongs to $c_j$, i.e. $\hat{\Phi}_s(d_i, c_j) > 0$ means that $d_i$ is believed to belong to $c_j$ while $\hat{\Phi}_s(d_i, c_j) < 0$ means it is believed not to belong to $c_j$. We instead interpret the absolute value of $\hat{\Phi}_s(d_i, c_j)$ (indicated by $|\hat{\Phi}_s(d_i, c_j)|$) as the strength of this belief.

At each iteration $s$ AdaBoost.MH tests the effectiveness of the newly generated weak hypothesis $\hat{\Phi}_s$ on the training set and uses the results to update a distribution $D_s$ of weights on the training pairs $\langle d_i, c_j\rangle$. The weight $D_{s+1}(d_i, c_j)$ is meant to capture how effective $\hat{\Phi}_1, \ldots, \hat{\Phi}_s$ have been in correctly predicting whether the training document $d_i$ belongs to category $c_j$ or not. By passing (together with the training set $Tr$) this distribution to the weak learner, AdaBoost.MH forces this latter to generate a new weak hypothesis $\hat{\Phi}_{s+1}$ that concentrates on the pairs with the highest weight, i.e. those that had proven harder to classify for the previous weak hypotheses.

[1] A *decision stump* is a decision tree of depth one, i.e. consisting of a root node and two or more leaf nodes.

*Input:* A training set $Tr = \{\langle d_1, C_1\rangle, \ldots, \langle d_g, C_g\rangle\}$
where $C_i \subseteq C = \{c_1, \ldots, c_m\}$ for all $i = 1, \ldots, g$.

*Body:* Let $D_1(d_i, c_j) = \dfrac{1}{gm}$ for all $i = 1, \ldots, g$ and for all $j = 1, \ldots, m$
For $s = 1, \ldots, S$ do:

- pass distribution $D_s(d_i, c_j)$ to the weak learner;
- get the weak hypothesis $\hat{\Phi}_s$ from the weak learner;
- set $D_{s+1}(d_i, c_j) = \dfrac{D_s(d_i, c_j)\exp(-\Phi(d_i, c_j)\cdot\hat{\Phi}_s(d_i, c_j))}{Z_s}$

where $Z_s = \sum_{i=1}^{g}\sum_{j=1}^{m} D_s(d_i, c_j)\exp(-\Phi(d_i, c_j)\cdot\hat{\Phi}_s(d_i, c_j))$

is a normalization factor chosen so that $\sum_{i=1}^{g}\sum_{j=1}^{m} D_{s+1}(d_i, c_j) = 1$

*Output:* A final hypothesis $\hat{\Phi}(d, c) = \sum_{s=1}^{S}\hat{\Phi}_s(d, c)$

**Fig. 1.** The AdaBoost.MH algorithm

The initial distribution $D_1$ is uniform. At each iteration $s$ all the weights $D_s(d_i, c_j)$ are updated to $D_{s+1}(d_i, c_j)$ according to the rule

$$D_{s+1}(d_i, c_j) = \frac{D_s(d_i, c_j)\exp(-\Phi(d_i, c_j)\cdot\hat{\Phi}_s(d_i, c_j))}{Z_s} \tag{1}$$

where

$$Z_s = \sum_{i=1}^{g}\sum_{j=1}^{m} D_s(d_i, c_j)\exp(-\Phi(d_i, c_j)\cdot\hat{\Phi}_s(d_i, c_j)) \tag{2}$$

is a normalization factor chosen so that $D_{s+1}$ is in fact a distribution, i.e. so that $\sum_{i=1}^{g}\sum_{j=1}^{m} D_{s+1}(d_i, c_j) = 1$. Equation 1 is such that the weight assigned to a pair $\langle d_i, c_j\rangle$ misclassified by $\hat{\Phi}_s$ is increased, as for such a pair $\Phi(d_i, c_j)$ and $\hat{\Phi}_s(d_i, c_j)$ have different signs and the factor $\Phi(d_i, c_j)\cdot\hat{\Phi}_s(d_i, c_j)$ is thus negative; likewise, the weight assigned to a pair correctly classified by $\hat{\Phi}_s$ is decreased.

### 2.1 Choosing the Weak Hypotheses

In AdaBoost.MH each document $d_i$ is represented as a vector $\langle w_{1i}, \ldots, w_{ri}\rangle$ of $r$ binary weights, where $w_{ki} = 1$ (resp. $w_{ki} = 0$) means that term $t_k$ occurs (resp. does not occur) in $d_i$; $T = \{t_1, \ldots, t_r\}$ is the set of terms that occur in at least one document in $Tr$.

In AdaBoost.MH the weak hypotheses generated by the weak learner at iteration $s$ are decision stumps of the form

试读结束，需要全本PDF请购买 www.ertongbook.com