Lecture Notes in Computer Science 4209

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

Fabio Crestani Paolo Ferragina
Mark Sanderson (Eds.)

String Processing and Information Retrieval

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

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

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

Table of Contents

Web Clustering and Text Categorization

MP-Boost: A Multiple-Pivot Boosting Algorithm and Its Application to Text Categorization 1
Andrea Esuli, Tiziano Fagni, Fabrizio Sebastiani

TreeBoost.MH: A Boosting Algorithm for Multi-label Hierarchical Text Categorization 13
Andrea Esuli, Tiziano Fagni, Fabrizio Sebastiani

Cluster Generation and Cluster Labelling for Web Snippets 25
Filippo Geraci, Marco Pellegrini, Marco Maggini, Fabrizio Sebastiani

Principal Components for Automatic Term Hierarchy Building 37
Georges Dupret, Benjamin Piwowarski

Strings

Computing the Minimum Approximate λ-Cover of a String 49
Qing Guo, Hui Zhang, Costas S. Iliopoulos

Sparse Directed Acyclic Word Graphs 61
Shunsuke Inenaga, Masayuki Takeda

On-Line Repetition Detection 74
Jin-Ju Hong, Gen-Huey Chen

User Behavior

Analyzing User Behavior to Rank Desktop Items 86
Paul-Alexandru Chirita, Wolfgang Nejdl

The Intention Behind Web Queries 98
Ricardo Baeza-Yates, Liliana Calderón-Benavides, Cristina González-Caro

Web Search Algorithms

Compact Features for Detection of Near-Duplicates in Distributed Retrieval 110
Yaniv Bernstein, Milad Shokouhi, Justin Zobel

Inverted Files Versus Suffix Arrays for Locating Patterns in Primary Memory 122
Simon J. Puglisi, W.F. Smyth, Andrew Turpin

Efficient Lazy Algorithms for Minimal-Interval Semantics 134
Paolo Boldi, Sebastiano Vigna

Output-Sensitive Autocompletion Search 150
Holger Bast, Christian W. Mortensen, Ingmar Weber

Compression

A Compressed Self-index Using a Ziv-Lempel Dictionary 163
Luís M.S. Russo, Arlindo L. Oliveira

Mapping Words into Codewords on PPM 181
Joaquín Adiego, Pablo de la Fuente

Correction

Improving Usability Through Password-Corrective Hashing 193
Andrew Mehler, Steven Skiena

Word-Based Correction for Retrieval of Arabic OCR Degraded Documents 205
Walid Magdy, Kareem Darwish

Information Retrieval Applications

A Statistical Model of Query Log Generation 217
Georges Dupret, Benjamin Piwowarski, Carlos Hurtado, Marcelo Mendoza

Using String Comparison in Context for Improved Relevance Feedback in Different Text Media 229
Adenike M. Lam-Adesina, Gareth J.F. Jones

A Multiple Criteria Approach for Information Retrieval 242
Mohamed Farah, Daniel Vanderpooten

English to Persian Transliteration 255
Sarvnaz Karimi, Andrew Turpin, Falk Scholer

Bio Informatics

Efficient Algorithms for Pattern Matching with General Gaps and Character Classes 267
Kimmo Fredriksson, Szymon Grabowski

Matrix Tightness: A Linear-Algebraic Framework for Sorting by Transpositions 279
Tzvika Hartman, Elad Verbin

How to Compare Arc-Annotated Sequences: The Alignment Hierarchy 291
Guillaume Blin, Hélène Touzet

Web Search Engines

Structured Index Organizations for High-Throughput Text Querying 304
Vo Ngoc Anh, Alistair Moffat

Adaptive Query-Based Sampling of Distributed Collections 316
Mark Baillie, Leif Azzopardi, Fabio Crestani

Short Papers

Dotted Suffix Trees A Structure for Approximate Text Indexing 329
Luís Pedro Coelho, Arlindo L. Oliveira

Phrase-Based Pattern Matching in Compressed Text 337
J. Shane Culpepper, Alistair Moffat

Discovering Context-Topic Rules in Search Engine Logs 346
Carlos A. Hurtado, Mark Levene

Incremental Aggregation of Latent Semantics Using a Graph-Based Energy Model 354
Aditya Ramana Rachakonda, Srinath Srinivasa

A New Algorithm for Fast All-Against-All Substring Matching 360
Marina Barsky, Ulrike Stege, Alex Thomo, Chris Upton

Author Index 367