

*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*

Nivio Ziviani Ricardo Baeza-Yates (Eds.)

# String Processing and Information Retrieval

14th International Symposium, SPIRE 2007  
Santiago, Chile, October 29-31, 2007  
Proceedings

## Volume Editors

Nivio Ziviani

Federal University of Minas Gerais

Department of Computer Science

Av. Antônio Carlos 6627, 31270-010 Belo Horizonte, MG, Brazil

E-mail: nivio@dcc.ufmg.br

Ricardo Baeza-Yates

Yahoo! Research Latin America

Blanco Encalada 2120, Santiago 6511224, Chile

E-mail: ricardo@baeza.cl

Library of Congress Control Number: 2007937296

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 3-540-75529-2 Springer Berlin Heidelberg New York

ISBN-13 978-3-540-75529-6 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: 12171385 06/3180 5 4 3 2 1 0

# Preface

This volume contains the papers presented at the 14th International Symposium on String Processing and Information Retrieval (SPIRE), held in Santiago, Chile, on October 29–31, 2007. SPIRE 2007 was organized in tandem with the 5th Latin American Web Congress (LA-WEB), with both conferences sharing a common day on Web Retrieval.

The papers in this volume were selected from 77 papers submitted from 25 different countries in response to the Call for Papers. Due to the high quality of the submissions, a total of 27 papers were accepted as full papers, yielding an acceptance rate of about 35%. SPIRE 2007 also featured three talks by invited speakers: Andrew Tomkins (Yahoo! Research, USA), Nivio Ziviani (Federal University of Minas Gerais, Brazil) and Justin Zobel (NICTA, Melbourne, Australia).

The SPIRE annual symposium provides an opportunity for researchers to present original contributions on 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* (IR modeling, indexing, ranking and filtering, interface design, visualization, cross-lingual IR systems, multimedia IR, digital libraries, collaborative retrieval, and Web related applications), *interaction of biology and computation* (DNA sequencing and applications in molecular biology, evolution and phylogenetics, recognition of genes and regulatory elements, and sequence driven protein structure prediction), and *information retrieval languages and applications* (XML, SGML, information retrieval from semi-structured data, text mining, and generation of structured data from text).

Special thanks are due to the members of the Program Committee and the additional reviewers who worked very hard to ensure the timely review of all submitted manuscripts. Thanks are due to Fabiano Cupertino Botelho, a Ph.D. student volunteer who ran the OpenConf system during the reviewing process and helped with the editorial work for this volume. We also thank the local organizers for their support and organization of SPIRE, in particular Javier Velasco, Christian Middleton, and Sara Quiñones, as well as the local team of student volunteers, whose efforts ensured the smooth organization and running of the event.

We would like to thank the sponsoring institutions, the Millennium Nucleus Center for Web Research of the Dept. of Computer Science of the University of Chile, the Dept. of Computer Science of the Federal University of Minas Gerais and Yahoo! Research Latin America.

October 2007

Nivio Ziviani  
Ricardo Baeza-Yates

# **SPIRE 2007 Organization**

## **General Chair**

Ricardo Baeza-Yates	Yahoo! Research (Spain & Chile) and CWR/DCC, Universidad de Chile (Chile)
---------------------	--

## **Program Committee Chair**

Nivio Ziviani	Universidade Federal de Minas Gerais (Brazil)
---------------	---

## **Local Organization**

Fabiano C. Botelho	Universidade Federal de Minas Gerais (Brazil)
Christian Middleton	Universitat Pompeu Fabra (Spain)
Sara Quiñones	Yahoo! Research Latin America (Chile)
Javier Velasco	CWR/DCC, Universidad de Chile (Chile)

## **Steering Committee**

Alberto Apostolico	Università di Padova (Italy) and Georgia Tech (USA)
Ricardo Baeza-Yates	Yahoo! Research (Spain & Chile) and CWR/DCC, Universidad de Chile (Chile)
Mariano Consens	University of Toronto (Canada)
Fabio Crestani	Università della Svizzera Italiana (Switzerland)
Paolo Ferragina	Università di Pisa (Italy)
Massimo Melucci	Università di Padova (Italy)
Gonzalo Navarro	Universidad de Chile (Chile)
Berthier Ribeiro-Neto	Universidade Federal de Minas Gerais (Brazil)
Mark Sanderson	University of Sheffield (UK)
Nivio Ziviani	Universidade Federal de Minas Gerais (Brazil)

## **Program Committee Members**

James Allan	University of Massachusetts Amherst (USA)
Amihood Amir	Bar-Ilan University (Israel)
Alberto Apostolico	University of Padova (Italy) and Georgia Tech (USA)
Chris Buckley	Sabir Research (USA)

Pável Pereira Calado	Instituto Superior Técnico/INESC-ID (Portugal)
Maxime Crochemore	University of Marne-la-Vallée (France)
Bruce Croft	University of Massachusetts Amherst (USA)
Martin Farach-Colton	Rutgers University (USA)
Edward Fox	Virginia Tech (USA)
Kimmo Fredriksson	University of Joensuu (Finland)
Raffaele Giancarlo	University of Palermo (Italy)
Marcos André Gonçalves	Federal University of Minas Gerais (Brazil)
Roberto Grossi	University of Pisa (Italy)
Heikki Hyyrö	University of Tampere (Finland)
Lucian Ilie	University of Western Ontario (Canada)
Costas Iliopoulos	University of London (UK)
Juha Kärkkäinen	University of Helsinki (Finland)
Mounia Lalmas	University of London (UK)
Tak-Wah Lam	University of Hong Kong (Hong Kong)
Gad Landau	University of Haifa (Israel)
Thierry Lecroq	University of Rouen (France)
Andrew MacFarlane	City University London (UK)
Veli Mäkinen	University of Helsinki (Finland)
Giovanni Manzini	University of Piemonte Orientale (Italy)
Massimo Melucci	University of Padua (Italy)
Alistair Moffat	University of Melbourne (Australia)
Edleno Silva de Moura	Federal University of Amazonas (Brazil)
Ian Munro	University of Waterloo (Canada)
Gonzalo Navarro	University of Chile (Chile)
Arlindo Oliveira	Inst. Superior Técnico/INESC-ID/IST (Portugal)
Sándor Pongor	Intl. Centre for Genetic Eng. and Biotechnology (Italy)
Bruno Póssas	Google Inc.(Brazil)
Mathieu Raffinot	CNRS (France)
Kunihiko Sadakane	Kyushu University (Japan)
Marie-France Sagot	INRIA and University Claude Bernard, Lyon I (France)
João Setubal	Virginia Tech (USA)
Rahul Shah	Purdue University (USA)
Altigran Soares da Silva	Federal University of Amazonas (Brazil)
Fabrizio Silvestri	ISTI - CNR (Italy)
Wing-Kin Sung	National University of Singapore (Singapore)
Masayuki Takeda	Kyushu University (Japan)
Jorma Tarhio	Helsinki University of Technology (Finland)
Gabriel Valiente	Technical University of Catalonia (Spain)
Hugo Zaragoza	Yahoo! Research (Spain)
Justin Zobel	RMIT University (Australia)

## Additional Reviewers

José Ramón Pérez Agüera	Diego Arroyuelo
Guilherme Assis	Claudine Badue
Marie-Pierre Béal	Mathieu Constant
Ido Dagan	Gianna Del Corso
Chiara Epifanio	Ankur Gupta
Iman Hajirasouliha	Jan Holub
Wing-Kai Hon	Petri Kalsi
Tomi Klein	Alberto H. F. Laender
Jorma Laurikkala	Sabrina Mantaci
Turkka Näppilä	Hannu Peltola
Simon Puglisi	Cenk Sahinalp
Leena Salmela	Borkur Sigurbjornsson
Tuomas Talvensaari	Andrew Turpin
Jarkko Toivonen	Shuly Wintner
Sebastiano Vigna	

## Previous Venues of SPIRE

The first four editions focused primarily on *string processing* and were held in Brazil and Chile. At that 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 the latter's increasing relevance and its inter-relationship with the area of string processing, and the name of the workshop was changed to the current one. In addition, since 2000, the symposium has been held alternately in Europe and Latin America, and has so far been held in Mexico, Spain, Chile, Portugal, Brazil, Italy, Argentina and the UK.

2006: Glasgow, UK  
 2005: Buenos Aires, Argentina  
 2004: Padova, Italy  
 2003: Manaus, Brazil  
 2002: Lisboa, Portugal  
 2001: Laguna San Rafael, Chile  
 2000: A Coruña, Spain  
 1999: Cancun, Mexico  
 1998: Santa Cruz de la Sierra, Bolivia  
 1997: Valparaiso, Chile  
 1996: Recife, Brazil  
 1995: Viña del Mar, Chile  
 1993: Belo Horizonte, Brazil

# Table of Contents

A Chaining Algorithm for Mapping cDNA Sequences to Multiple Genomic Sequences . . . . .	1
<i>Mohamed Abouelhoda</i>	
Edge-Guided Natural Language Text Compression . . . . .	14
<i>Joaquín Adiego, Miguel A. Martínez-Prieto, and Pablo de la Fuente</i>	
Local Transpositions in Alignment of Polyphonic Musical Sequences . . . .	26
<i>Julien Allali, Pascal Ferraro, Pierre Hanna, and Costas Iliopoulos</i>	
Efficient Computations of $\ell_1$ and $\ell_\infty$ Rearrangement Distances . . . . .	39
<i>Amihoud Amir, Yonatan Aumann, Piotr Indyk, Avivit Levy, and Ely Porat</i>	
Generalized LCS . . . . .	50
<i>Amihoud Amir, Tzvi Hartman, Oren Kapah, B. Riva Shalom, and Dekel Tsur</i>	
Exploiting Genre in Focused Crawling . . . . .	62
<i>Guilherme T. de Assis, Alberto H.F. Laender, Marcos André Gonçalves, and Altigran S. da Silva</i>	
Admission Policies for Caches of Search Engine Results . . . . .	74
<i>Ricardo Baeza-Yates, Flavio Junqueira, Vassilis Plachouras, and Hans Friedrich Witschel</i>	
A Pocket Guide to Web History . . . . .	86
<i>Klaus Berberich, Srikantha Bedathur, and Gerhard Weikum</i>	
Jump-Matching with Errors . . . . .	98
<i>Ayelet Butman, Noa Lewenstein, Benny Porat, and Ely Porat</i>	
Estimating Number of Citations Using Author Reputation . . . . .	107
<i>Carlos Castillo, Debora Donato, and Aristides Gionis</i>	
A Fast and Compact Web Graph Representation . . . . .	118
<i>Francisco Claude and Gonzalo Navarro</i>	
A Filtering Algorithm for $k$ -Mismatch with Don't Cares . . . . .	130
<i>Raphaël Clifford and Ely Porat</i>	
Compact Set Representation for Information Retrieval . . . . .	137
<i>J. Shane Culpepper and Alistair Moffat</i>	



Approximate Swap and Mismatch Edit Distance .....	149
<i>Yair Dombb, Ohad Lipsky, Benny Porat, Ely Porat, and Asaf Tsur</i>	
Approximating Constrained LCS .....	164
<i>Zvi Gotthilf and Moshe Lewenstein</i>	
Tuning Approximate Boyer-Moore for Gene Sequences .....	173
<i>Petri Kalsi, Leena Salmela, and Jorma Tarhio</i>	
Optimal Self-adjusting Trees for Dynamic String Data in Secondary Storage .....	184
<i>Pang Ko and Srinivas Aluru</i>	
Indexing a Dictionary for Subset Matching Queries .....	195
<i>Gad M. Landau, Dekel Tsur, and Oren Weimann</i>	
Extending Weighting Models with a Term Quality Measure .....	205
<i>Christina Lioma and Iadh Ounis</i>	
Highly Frequent Terms and Sentence Retrieval .....	217
<i>David E. Losada and Ronald T. Fernández</i>	
Implicit Compression Boosting with Applications to Self-indexing .....	229
<i>Veli Mäkinen and Gonzalo Navarro</i>	
A Web-Page Usage Prediction Scheme Using Weighted Suffix Trees .....	242
<i>Christos Makris, Yannis Panagis, Evangelos Theodoridis, and Athanasios Tsakalidis</i>	
Enhancing Educational-Material Retrieval Using Authored-Lesson Metadata .....	254
<i>Olivier Motelet, Benjamin Piwowarski, Georges Dupret, Jose A. Pino, and Nelson Baloian</i>	
Approximate String Matching with Lempel-Ziv Compressed Indexes ....	264
<i>Luís M.S. Russo, Gonzalo Navarro, and Arlindo L. Oliveira</i>	
Algorithms for Weighted Matching .....	276
<i>Leena Salmela and Jorma Tarhio</i>	
Efficient Text Proximity Search .....	287
<i>Ralf Schenkel, Andreas Broschart, Seungwon Hwang, Martin Theobald, and Gerhard Weikum</i>	
Prefix-Shuffled Geometric Suffix Tree .....	300
<i>Tetsuo Shibuya</i>	
<b>Author Index</b> .....	311