

*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

*New York University, NY, 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*

Mariano Consens Gonzalo Navarro (Eds.)

# String Processing and Information Retrieval

12th International Conference, SPIRE 2005  
Buenos Aires, Argentina, November 2-4, 2005  
Proceedings



Springer

Volume Editors

Mariano Consens  
University of Toronto  
Department of Mechanical and Industrial Engineering  
Department of Computer Science  
Toronto, Canada  
E-mail: consens@cs.toronto.edu

Gonzalo Navarro  
University of Chile  
Center for Web Research, Dept. of Computer Science, Chile  
E-mail: gnavarro@dcc.uchile.cl

Library of Congress Control Number: 2005934415

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

ISSN 0302-9743  
ISBN-10 3-540-29740-5 Springer Berlin Heidelberg New York  
ISBN-13 978-3-540-29740-6 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

[springeronline.com](http://springeronline.com)

© Springer-Verlag Berlin Heidelberg 2005  
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India  
Printed on acid-free paper SPIN: 11575832 06/3142 5 4 3 2 1 0

## Preface

The papers contained in this volume were presented at the 12th edition of the International Symposium on String Processing and Information Retrieval (SPIRE), held November 2–4, 2005, in Buenos Aires, Argentina. They were selected from 102 papers submitted from 25 countries in response to the Call for Papers. A total of 27 submissions were accepted as full papers, yielding an acceptance rate of about 26%. In view of the large number of good-quality submissions the conference program also included 17 short papers that also appear in the proceedings. In addition, the Steering Committee invited the following speakers: Prabhakar Raghavan (Yahoo! Research, USA), Paolo Ferragina (University of Pisa, Italy), and Gonzalo Navarro (University of Chile, Chile).

Papers solicited for SPIRE 2005 were meant to constitute 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 modeling, 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).

SPIRE has its origins in the South American Workshop on String Processing (WSP). Since 1998 the focus of the conference was broadened to include information retrieval. Starting in 2000, Europe has been the conference venue on even years. The first 11 meetings were held in Belo Horizonte (Brazil, 1993), Valparaíso (Chile, 1995), Recife (Brazil, 1996), Valparaíso (Chile, 1997), Santa Cruz (Bolivia, 1998), Cancún (Mexico, 1999), A Coruña (Spain, 2000), Laguna San Rafael (Chile, 2001), Lisboa (Portugal, 2002), Manaus (Brazil, 2003), and Padova (Italy, 2004).

SPIRE 2005 was held in tandem with LA-WEB 2005, the 3rd Latin American Web Congress, with both conferences sharing a common day in Web Retrieval.

SPIRE 2005 was sponsored by Centro Latinoamericano de Estudios en Informática (CLEI), Programa Iberoamericano de Ciencia y Tecnología para el Desarrollo (CYTED), Center for Web Research (CWR, University of Chile), and Sociedad Argentina de Informática e Investigación Operativa (SADIO).

We thank the local organizers for their support in the organization of SPIRE and the members of the Program Committee and the additional reviewers for providing timely and detailed reviews of the submitted papers and for their active participation in the email discussions that took place before we could assemble

the final program. Finally, we would like to thank Ricardo Baeza-Yates, who, on behalf of the Steering Committee, invited us to chair the Program Committee.

November 2005

Mariano P. Consens,  
Gonzalo Navarro

# SPIRE 2005 Organization

## Steering Committee

Ricardo Baeza-Yates (Chair)	ICREA-Universitat Pompeu Fabra (Spain) and Universidad de Chile (Chile)
Alberto Apostolico	Università di Padova (Italy) and Georgia Tech (USA)
Alberto Laender	Universidade Federal de Minas Gerais (Brazil)
Massimo Melucci	Università di Padova (Italy)
Edleno de Moura	Universidade Federal do Amazonas (Brazil)
Mario Nascimento	University of Alberta (Canada)
Arlindo Oliveira	INESC (Portugal)
Berthier Ribeiro-Neto	Universidade Federal de Minas Gerais (Brazil)
Nivio Ziviani	Universidade Federal de Minas Gerais (Brazil)

## Program Committee Chairs

Mariano Consens	Dept. of Mechanical and Industrial Engineering Dept. of Computer Science University of Toronto, Canada
Gonzalo Navarro	Center for Web Research Dept. of Computer Science Universidad de Chile, Chile

## Program Committee Members

Amihood Amir	Bar-Ilan University (Israel)
Alberto Apostolico	Università di Padova (Italy) and Georgia Tech (USA)
Ricardo Baeza-Yates	ICREA-Universitat Pompeu Fabra (Spain) and Universidad de Chile (Chile)
Nieves R. Brisaboa	Universidade da Coruña (Spain)
Edgar Chávez	Universidad Michoacana (Mexico)
Charles Clarke	University of Waterloo (Canada)
Bruce Croft	University of Massachusetts (USA)
Paolo Ferragina	Università di Pisa (Italy)
Norbert Fuhr	Universität Duisburg-Essen (Germany)
Raffaele Giancarlo	Università di Palermo (Italy)
Roberto Grossi	Università di Pisa (Italy)
Carlos Heuser	Universidade Federal de Rio Grande do Sul (Brazil)

## VIII Organization

Carlos Hurtado	Universidad de Chile (Chile)
Lucian Ilie	University of Western Ontario (Canada)
Panagiotis Ipeirotis	New York University (USA)
Juha Kärkkäinen	University of Helsinki (Finland)
Nick Koudas	University of Toronto (Canada)
Mounia Lalmas	Queen Mary University of London (UK)
Gad Landau	University of Haifa (Israel) and Polytechnic University (NY, USA)
Stefano Lonardi	University of California at Riverside (USA)
Yoelle Maarek	IBM Haifa Research Lab (Israel)
Veli Mäkinen	Bielefeld University (Germany)
Mauricio Marín	Universidad de Magallanes (Chile)
João Meidanis	UNICAMP (Brazil)
Massimo Melucci	Università di Padova (Italy)
Edleno de Moura	Universidade Federal do Amazonas (Brazil)
Ian Munro	University of Waterloo (Canada)
Arlindo Oliveira	INESC (Portugal)
Kunsoo Park	Seoul National University (Korea)
Prabhakar Raghavan	Yahoo Inc. (USA)
Berthier Ribeiro-Neto	Universidade Federal de Minas Gerais (Brazil)
Kunihiko Sadakane	Kyushu University (Japan)
Marie-France Sagot	INRIA (France)
João Setubal	Virginia Tech (USA)
Jayavel Shanmugasundaram	Cornell University (USA)
Ayumi Shinohara	Tohoku University (Japan)
Jorma Tarhio	Helsinki University of Technology (Finland)
Jeffrey Vitter	Purdue University (USA)
Hugh Williams	Microsoft Corporation (USA)
Hugo Zaragoza	Microsoft Research (UK)
Nivio Ziviani	Universidade Federal de Minas Gerais (Brazil)
Justin Zobel	RMIT (Australia)

## External Reviewers

Jussara Almeida	Michela Bacchin
Ramurti Barbosa	Bodo Billerbeck
Sebastian Böcker	Michael Cameron
David Carmel	Luis Coelho
Marco Cristo	Giorgio Maria Di Nunzio
Alair Pereira do Lago	Shiri Dori
Celia Francisca dos Santos	Fan Yang
Feng Shao	Nicola Ferro
Kimmo Fredriksson	Gudrun Fisher
Paulo B. Golgher	Alejandro Hevia
Jie Zheng	Carmel Kent

Shahar Keret	Tsvi Kopelowitz
Sascha Kriewel	Michael Laszlo
Nicholas Lester	Saadia Malik
Julia Mixtacki	Viviane Moreira Orengo
Henrik Nottelmann	Nicola Orio
Rodrigo Paredes	Laxmi Parida
Hannu Peltola	Patrícia Peres
Nadia Pisanti	Benjamin Piwowski
Bruno Possas	Jussi Rautio
Davi de Castro Reis	Nora Reyes
Luis Russo	Klaus-Bernd Schürmann
Marinella Sciortino	Rahul Shah
Darren Shakib	Riva Shalom
S.M.M. (Saied) Tahaghoghi	Eric Tannier
Andrew Turpin	Rodrigo Verschae
Ying Zhang	

## Local Organization

SADIO (Argentine Society for Informatics and Operations Research)

SADIO President	Gabriel Baum
Local Arrangements Chair	Héctor Monteverde
Steering Committee Liaison	Ricardo Baeza-Yates
Administrative Manager	Alejandra Villa



# Table of Contents

## String Processing and Information Retrieval 2005

Enhanced Byte Codes with Restricted Prefix Properties <i>J. Shane Culpepper, Alistair Moffat</i> .....	1
Experimental Analysis of a Fast Intersection Algorithm for Sorted Sequences <i>Ricardo Baeza-Yates, Alejandro Salinger</i> .....	13
Compressed Perfect Embedded Skip Lists for Quick Inverted-Index Lookups <i>Paolo Boldi, Sebastiano Vigna</i> .....	25
XML Retrieval with a Natural Language Interface <i>Xavier Tannier, Shlomo Geva</i> .....	29
Recommending Better Queries from Click-Through Data <i>Georges Dupret, Marcelo Mendoza</i> .....	41
A Bilingual Linking Service for the Web <i>Alessandra Alaniz Macedo, José Antonio Camacho-Guerrero, Maria da Graça Campos Pimentel</i> .....	45
Evaluating Hierarchical Clustering of Search Results <i>Juan M. Cigarran, Anselmo Peñas, Julio Gonzalo, Felisa Verdejo</i> . . . .	49
Counting Suffix Arrays and Strings <i>Klaus-Bernd Schürmann, Jens Stoye</i> .....	55
Towards Real-Time Suffix Tree Construction <i>Amihood Amir, Tsvi Kopelowitz, Moshe Lewenstein, Noa Lewenstein</i> .....	67
Rank-Sensitive Data Structures <i>Iwona Bialynicka-Birula, Roberto Grossi</i> .....	79
Cache-Conscious Collision Resolution in String Hash Tables <i>Nikolas Askitis, Justin Zobel</i> .....	91
Measuring the Difficulty of Distance-Based Indexing <i>Matthew Skala</i> .....	103

<i>N</i> -Gram Similarity and Distance <i>Grzegorz Kondrak</i> .....	115
Using the <i>k</i> -Nearest Neighbor Graph for Proximity Searching in Metric Spaces <i>Rodrigo Paredes, Edgar Chávez</i> .....	127
Classifying Sentences Using Induced Structure <i>Menno van Zaanen, Luiz Augusto Pizzato, Diego Mollá</i> .....	139
Counting Lumps in Word Space: Density as a Measure of Corpus Homogeneity <i>Magnus Sahlgren, Jussi Karlgren</i> .....	151
Multi-label Text Categorization Using K-Nearest Neighbor Approach with M-Similarity <i>Yi Feng, Zhaohui Wu, Zhongmei Zhou</i> .....	155
Lydia: A System for Large-Scale News Analysis <i>Levon Lloyd, Dimitrios Kechagias, Steven Skiena</i> .....	161
Composite Pattern Discovery for PCR Application <i>Stanislav Angelov, Shunsuke Inenaga</i> .....	167
Lossless Filter for Finding Long Multiple Approximate Repetitions Using a New Data Structure, the Bi-factor Array <i>Pierre Peterlongo, Nadia Pisanti, Frederic Boyer, Marie-France Sagot</i> .....	179
Linear Time Algorithm for the Generalised Longest Common Repeat Problem <i>Inbok Lee, Yoan José Pinzón Ardila</i> .....	191
Application of Clustering Technique in Multiple Sequence Alignment <i>Patrícia Silva Peres, Edleno Silva de Moura</i> .....	202
Stemming Arabic Conjunctions and Prepositions <i>Abdusalam F.A. Nwesri, S.M.M. Tahaghoghi, Falk Scholer</i> .....	206
XML Multimedia Retrieval <i>Zhigang Kong, Mounia Lalmas</i> .....	218
Retrieval Status Values in Information Retrieval Evaluation <i>Amélie Imafouo, Xavier Tannier</i> .....	224

A Generalization of the Method for Evaluation of Stemming Algorithms Based on Error Counting <i>Ricardo Sánchez de Madariaga, José Raúl Fernández del Castillo, José Ramón Hilera</i> . . . . .	228
Necklace Swap Problem for Rhythmic Similarity Measures <i>Yoan José Pinzón Ardila, Raphaël Clifford, Manal Mohamed</i> . . . . .	234
Faster Generation of Super Condensed Neighbourhoods Using Finite Automata <i>Luís M.S. Russo, Arlindo L. Oliveira</i> . . . . .	246
Restricted Transposition Invariant Approximate String Matching Under Edit Distance <i>Heikki Hyvrö</i> . . . . .	256
Fast Plagiarism Detection System <i>Maxim Mozgovoy, Kimmo Fredriksson, Daniel White, Mike Joy, Erkki Sutinen</i> . . . . .	267
A Model for Information Retrieval Based on Possibilistic Networks <i>Asma H. Brini, Mohand Boughanem, Didier Dubois</i> . . . . .	271
Comparison of Representations of Multiple Evidence Using a Functional Framework for IR <i>Ilmério R. Silva, João N. Souza, Luciene C. Oliveira</i> . . . . .	283
Deriving TF-IDF as a Fisher Kernel <i>Charles Elkan</i> . . . . .	295
Utilizing Dynamically Updated Estimates in Solving the Longest Common Subsequence Problem <i>Lasse Bergroth</i> . . . . .	301
Computing Similarity of Run-Length Encoded Strings with Affine Gap Penalty <i>Jin Wook Kim, Amihoud Amir, Gad M. Landau, Kunsoo Park</i> . . . . .	315
$L_1$ Pattern Matching Lower Bound <i>Ohad Lipsky, Ely Porat</i> . . . . .	327
Approximate Matching in the $L_\infty$ Metric <i>Ohad Lipsky, Ely Porat</i> . . . . .	331
An Edit Distance Between RNA Stem-Loops <i>Valentin Guignon, Cedric Chauve, Sylvie Hamel</i> . . . . .	335

A Multiple Graph Layers Model with Application to RNA Secondary Structures Comparison <i>Julien Allali, Marie-France Sagot</i> .....	348
Normalized Similarity of RNA Sequences <i>Rolf Backofen, Danny Hermelin, Gad M. Landau, Oren Weimann</i> .....	360
A Fast Algorithmic Technique for Comparing Large Phylogenetic Trees <i>Gabriel Valiente</i> .....	370
Practical and Optimal String Matching <i>Kimmo Fredriksson, Szymon Grabowski</i> .....	376
A Bit-Parallel Tree Matching Algorithm for Patterns with Horizontal VLDC's <i>Hisashi Tsuji, Akira Ishino, Masayuki Takeda</i> .....	388
A Partition-Based Efficient Algorithm for Large Scale Multiple-Strings Matching <i>Ping Liu, Yan-bing Liu, Jian-long Tan</i> .....	399
<b>Author Index</b> .....	405