

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

Alfred Kobsa

University of California, Irvine, CA, 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

TU Dortmund University, Germany

Madhu Sudan

Microsoft Research, Cambridge, MA, USA

Demetri Terzopoulos

University of California, Los Angeles, CA, USA

Doug Tygar

University of California, Berkeley, CA, USA

Gerhard Weikum

Max Planck Institute for Informatics, Saarbruecken, Germany

Liliana Calderón-Benavides
Cristina González-Caro
Edgar Chávez Nivio Ziviani (Eds.)

String Processing and Information Retrieval

19th International Symposium, SPIRE 2012
Cartagena de Indias, Colombia, October 21-25, 2012
Proceedings

Volume Editors

Liliana Calderón-Benavides
Cristina González-Caro
Universidad Autónoma de Bucaramanga
Information Technologies Research Group
Bucaramanga, Colombia
E-mail: {mcalderon, cgonzalc}@unab.edu.co

Edgar Chávez
Universidad Michoacana
School of Physics and Mathematics
Edificio B, Ciudad Universitaria
Morelia, México 58000, Mexico
E-mail: elchavez@umich.mx

Nivio Ziviani
Universidade Federal de Minas Gerais
Department of Computer Science
Av. Antonio Carlos 6627, Belo Horizonte 31270-010, MG, Brazil
E-mail: nivio@dcc.ufmg.br

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-34108-3 e-ISBN 978-3-642-34109-0
DOI 10.1007/978-3-642-34109-0
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: Applied for

CR Subject Classification (1998): H.3, J.3, H.2.8, I.5, I.2.7, H.4

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

© Springer-Verlag Berlin Heidelberg 2012

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.

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This volume contains the papers presented at the 19th International Symposium on String Processing and Information Retrieval (SPIRE), held in Cartagena de Indias, Colombia, during October 21–25, 2012. SPIRE 2012 was organized in tandem with the 8th Latin American Web Congress (LA-WEB), with both conferences sharing a common day in Web Retrieval.

In the 19 years since its first edition back in 1993, SPIRE has become the reference meeting for the interdisciplinary community of researchers whose activity lies at the crossroads of string processing and information retrieval. The first four editions of this event concentrated mainly on string processing, and were held in South America under the title “South American Workshop on String Processing” (WSP) in 1993 (Belo Horizonte, Brazil), 1995 (Valparaiso, Chile), 1996 (Recife, Brazil), and 1997 (Valparaiso, Chile). WSP was renamed as SPIRE in 1998 (Santa Cruz, Bolivia) when the scope of the event was broadened to include also information retrieval. The change was motivated by the increasing relevance of information retrieval and its close interrelationship with the general area of string processing. From 1999 to 2007, the venue of SPIRE alternated between South / Latin America (odd years) and Europe (even years), with Cancun, Mexico, in 1999, A Coruña, Spain, in 2000, Laguna de San Rafael, Chile, in 2001, Lisbon, Portugal, in 2002, Manaus, Brazil, in 2003, Padova, Italy, in 2004, Buenos Aires, Argentina, in 2005, Glasgow, UK, in 2006, and Santiago, Chile, in 2007. This pattern was broken when SPIRE 2008 was held in Melbourne, Australia, but it was restarted in 2009 when the venue was Saariselk, Finland, Los Cabos, Mexico, in 2010, followed by Pisa, Italy, in 2011.

The SPIRE 2012 call for papers resulted in the submission of 81 papers. Each submitted paper was reviewed by at least three of the 40 members of the Program Committee (PC), which eventually engaged in discussions coordinated by the two PC Chairmen in case of lack of consensus. We believe this resulted in a very accurate selection of the truly best submitted papers. As a result, 26 long papers and 13 short papers were accepted, and are published in these proceedings.

The program of SPIRE 2012 started on October 21 with three tutorials providing in-depth coverage of topics in string processing (“Space-Efficient Data Structures,” by Francisco Claude and Gonzalo Navarro) and information retrieval (“Evaluation Metrics for Information Access,” by Enrique Amig and Julio Gonzalo and “Information Dissemination in Social Networks,” by Aristides Gionis). On October 22, SPIRE 2012 hosted two workshops, i.e., the 7th Workshop on Compression, Text, and Algorithms (WCTA) and the Workshop on Algorithmic Analysis of Biological Data (WAABD). On the following three days the main conference featured keynote speeches by Amihood Amir, Ricardo Baeza-Yates, and Ian H. Witten, plus the presentations of the 26 full papers and

13 short papers. A Best Paper Award and a Best Student Paper Award were also assigned.

We would like to take the opportunity to thank Google, Yahoo! Research, and Universidad Autónoma de Bucaramanga. We would also like to thank everybody involved in making SPIRE 2012 such an exciting event. Specifically, we would like to thank all conference, tutorial, and workshop participants and presenters, who provided a fascinating one-week program of high-quality presentations and intensive discussions. Thanks also to all the members of the PC and to the additional reviewers, who went to great lengths to ensure the high quality of this conference. We are specially grateful to EasyChair for saving us a lot of work and for providing timely support.

Furthermore, we would like to thank all the members of the local organizing team at the Universidad Autónoma de Bucaramanga (UNAB). Particularly, we would like to thank Fabrizio Silvestri, who acted as Tutorials Chair, Gonzalo Navarro, who acted as Workshops Chair, Luz Emilia Jimenez, who gave us support in local arrangements, the IT team from UNAB who designed the official image and website of the symposium, and to all the student volunteers. They all made tremendous efforts to make sure that this event became an exciting and enjoyable one. It is due to them that the organization of SPIRE 2012 was a pleasure.

October 2012

Liliana Calderón-Benavides
Edgar Chávez
Cristina González-Caro
Nivio Ziviani

Organization

Program Committee

Giambattista Amati	Fondazione Ugo Bordoni, Italy
Amihood Amir	Bar Ilan University and Johns Hopkins University, Israel/USA
Ricardo Baeza-Yates	Yahoo! Research
Paolo Boldi	Università degli Studi di Milano, Italy
Liliana Calderón-Benavides	Universidad Autónoma de Bucaramanga, Colombia
Jamie Callan	Carnegie Mellon University, USA
Edgar Chavez	Universidad Michoacana, Mexico
Francisco Claude	University of Waterloo, Canada
Cesar A. Collazos	Universidad del Cauca, Colombia
Fabio Crestani	University of Lugano, Switzerland
Marco Cristo	Universidade Federal do Amazonas, Brazil
Maxime Crochemore	Kings College London and Université Paris-Est, UK/France
Bruce Croft	University of Massachusetts Amherst, USA
Edleno Silva De Moura	Universidade Federal do Amazonas, Brazil
Marcos Goncalves	Universidade Federal do Minas Gerais, Brazil
Cristina González-Caro	Universidad Autónoma de Bucaramanga, Colombia
Concettina Guerra	University of Padova and Georgia Tech, Italy/USA
Jan Holub	Czech Technical University in Prague, Czech Republic
Lucian Ilie	University of Western Ontario, Canada
Costas Iliopoulos	King's College London, UK
Shen Jialie	Singapore Management University, Singapore
Gregory Kucherov	CNRS/LIGM, France
Alberto Laender	Universidade Federal de Minas Gerais, Brazil
Mounia Lalmas	Yahoo Research
Moshe Lewenstein	Bar Ilan University, Israel
Alistair Moffat	University of Melbourne, Australia
Veli Mäkinen	University of Helsinki, Finland
Gonzalo Navarro	University of Chile, Chile
Laxmi Parida	IBM T.J. Watson Research Center, USA

VIII Organization

Kunsoo Park	Seoul National University, Korea
Marco Pellegrini	Institute for Informatics and Telematics of C.N.R., Italy
Yoan Pinzon	National University of Colombia - ALGOS UN, Colombia
Simon Puglisi	Royal Melbourne Institute of Technology, Australia
Berthier Ribeiro-Neto	Google and UFMG, Brazil
Luis M. S. Russo	IST / INESC-ID, Portugal
Rahul Shah	Louisiana State University, USA
Torsten Suel	Yahoo! Research
Esko Ukkonen	University of Helsinki, Finland
Adriano Veloso	UFMG, Brazil
Jeff Vitter	University of Kansas, USA
Nivio Ziviani	Universidade Federal de Minas Gerais, Brazil

Additional Reviewers

Alatabi, Ali	Jiang, Wei
Badkobeh, Golnaz	Kazi, Serizhan
Baier, Jan	Kim, Yubin
Barros, Evandrino	Kolpakov, Roman
Barton, Carl	Konow, Roberto
Berlt, Klessius	Kubica, Marcin
Brandão, Michele	Kulkarni, Anagha
Carvalho, André	Ladra, Susana
Cheng, Zhiyong	Landau, Gad
Christou, Michalis	Liptak, Zsuzsanna
Cortez, Eli	Lonardi, Stefano
Dimopoulos, Constantinos	Mahdabi, Parvaz
Fariña, Antonio	Markov, Ilya
Fernandes, David	Miranda, Eulanda dos Santos
Fernandes, Francisco	Mostafa, Keikha
Ferreira, Anderson	Nekrich, Yakov
Francisco, Alexandre P.	Nepomnyachiy, Sergey
Fredriksson, Kimmo	Nicolas, Francois
Frousios, Kimon	Parama, Jose R.
Gagie, Travis	Petri, Matthias
Gerani, Shima	Pisanti, Nadia
Gottesman-Gelley, Bluma	Pissis, Solon
Gupta, Varun	Polishchuk, Valentin
Haiminen, Niina	Porat, Ely
He, Dan	Rodrigues, Kaio
Inches, Giacomo	Roma, Nuno

Salles, Thiago
Salmela, Leena
Sheng, Cheng
Sirén, Jouni
Souza, Jucimar
Thankachan, Sharma
Tsur, Dekel

Tyczynski, Wojciech
Utro, Filippo
Vahabi, Hossein
Välimäki, Niko
Wagner Rodrigues, Kaio
Will, Sebastian
Zhao, Le

Table of Contents

Approximate Period Detection and Correction	1
<i>Amihood Amir and Avivit Levy</i>	
Usage Data in Web Search: Benefits and Limitations	16
<i>Ricardo Baeza-Yates and Yoelle Maarek</i>	
Semantic Document Representation: Do It with Wikification	17
<i>Ian Witten</i>	
Clustering Heterogeneous Data with Mutual Semi-supervision	18
<i>Artur Abdullin and Olfa Nasraoui</i>	
Compressed Suffix Trees for Repetitive Texts	30
<i>Andrés Abeliuk and Gonzalo Navarro</i>	
Configurations and Minority in the String Consensus Problem	42
<i>Amihood Amir, Haim Paryenty, and Liam Roditty</i>	
A Study on Novelty Evaluation in Biomedical Information Retrieval	54
<i>Xiangdong An, Nick Cercone, Hai Wang, and Zheng Ye</i>	
Computing the Maximal-Exponent Repeats of an Overlap-Free String in Linear Time	61
<i>Golnaz Badkobeh, Maxime Crochemore, and Chalita Toopsuwan</i>	
Collection Ranking and Selection for Federated Entity Search	73
<i>Krisztian Balog, Robert Neumayer, and Kjetil Nørsvåg</i>	
Efficient LZ78 Factorization of Grammar Compressed Text	86
<i>Hideo Bannai, Shunsuke Inenaga, and Masayuki Takeda</i>	
Space-Efficient Computation of Maximal and Supermaximal Repeats in Genome Sequences	99
<i>Timo Beller, Katharina Berger, and Enno Ohlebusch</i>	
Active Microbloggers: Identifying Influencers, Leaders and Discussers in Microblogging Networks	111
<i>Lamjed Ben Jabeur, Lynda Tamine, and Mohand Boughanem</i>	
Efficient Bubble Enumeration in Directed Graphs	118
<i>Etienne Birmelé, Pierluigi Crescenzi, Rui Ferreira, Roberto Grossi, Vincent Lacroix, Andrea Marino, Nadia Pisanti, Gustavo Sacomoto, and Marie-France Sagot</i>	

The Longest Common Subsequence Problem with Crossing-Free Arc-Annotated Sequences	130
<i>Guillaume Blin, Minghui Jiang, and Stéphane Vialette</i>	
A Zipf-Like Distant Supervision Approach for Multi-document Summarization Using Wikinews Articles	143
<i>Felipe Bravo-Marquez and Manuel Manriquez</i>	
Ranked Document Retrieval in (Almost) No Space	155
<i>Nieves R. Brisaboa, Ana Cerdeira-Pena, Gonzalo Navarro, and Óscar Pedreira</i>	
Impact of Regionalization on Performance of Web Search Engine Result Caches	161
<i>B. Barla Cambazoglu and Ismail Sengor Altıngövdü</i>	
The Wavelet Matrix	167
<i>Francisco Claude and Gonzalo Navarro</i>	
Improved Grammar-Based Compressed Indexes	180
<i>Francisco Claude and Gonzalo Navarro</i>	
Experiments on Pseudo Relevance Feedback Using Graph Random Walks	193
<i>Clément de Groc and Xavier Tannier</i>	
Temporal Web Image Retrieval	199
<i>Gaël Dias, José G. Moreno, Adam Jatowt, and Ricardo Campos</i>	
Improved Address-Calculation Coding of Integer Arrays	205
<i>Amr Elmasry, Jyrki Katajainen, and Jukka Teuhola</i>	
Fast Multiple String Matching Using Streaming SIMD Extensions Technology	217
<i>Simone Faro and M. Oğuzhan Külekci</i>	
Faster Algorithm for Computing the Edit Distance between SLP-Compressed Strings	229
<i>Paweł Gawrychowski</i>	
Basic Word Completion and Prediction for Hebrew	237
<i>Yaakov HaCohen-Kerner and Izek Greenfield</i>	
Eager XPath Evaluation over XML Streams	245
<i>Kazuhito Hagio, Takashi Ohgami, Hideo Bannai, and Masayuki Takeda</i>	
Position-Aligned Translation Model for Citation Recommendation	251
<i>Jing He, Jian-Yun Nie, Yang Lu, and Wayne Xin Zhao</i>	

Compressed Representation of Web and Social Networks via Dense Subgraphs	264
<i>Cecilia Hernández and Gonzalo Navarro</i>	
Method of Mining Subtopics Using Dependency Structure and Anchor Texts	277
<i>Se-Jong Kim and Jong-Hyeok Lee</i>	
Efficient Data Structures for the Factor Periodicity Problem	284
<i>Tomasz Kociumaka, Jakub Radoszewski, Wojciech Rytter, and Tomasz Walen</i>	
Dual-Sorted Inverted Lists in Practice	295
<i>Roberto Konow and Gonzalo Navarro</i>	
Computing Discriminating and Generic Words	307
<i>Gregory Kucherov, Yakov Nekrich, and Tatiana Starikovskaya</i>	
Computing Maximum Number of Runs in Strings	318
<i>Kazuhiko Kusano, Kazuyuki Narisawa, and Ayumi Shinohara</i>	
Grammar Precompression Speeds Up Burrows–Wheeler Compression . . .	330
<i>Juha Kärkkäinen, Pekka Mikkola, and Dominik Kempa</i>	
Parikh Matching in the Streaming Model	336
<i>Lap-Kei Lee, Moshe Lewenstein, and Qin Zhang</i>	
Relevance Feedback Method Based on Vector Space Basis Change	342
<i>Rabeb Mbarek and Mohamed Tmar</i>	
Approximate Function Matching under δ - and γ - Distances	348
<i>Juan Mendivelso, Inbok Lee, and Yoan J. Pinzón</i>	
The Position Heap of a Trie	360
<i>Yuto Nakashima, Tomohiro I, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda</i>	
Smaller Self-indexes for Natural Language	372
<i>Nieves R. Brisaboa, Gonzalo Navarro, and Alberto Ordóñez</i>	
Parallel Suffix Array Construction for Shared Memory Architectures . . .	379
<i>Vitaly Osipov</i>	
Characterization and Extraction of Irredundant Tandem Motifs	385
<i>Laxmi Parida, Cinzia Pizzi, and Simona E. Rombó</i>	
Variable-Length Codes for Space-Efficient Grammar-Based Compression	398
<i>Yoshimasa Takabatake, Yasuo Tabei, and Hiroshi Sakamoto</i>	
Author Index	411