

Lecture Notes in Computer Science

2857

Edited by G. Goos, J. Hartmanis, and J. van Leeuwen

Springer

Berlin

Heidelberg

New York

Hong Kong

London

Milan

Paris

Tokyo

Mario A. Nascimento Edleno S. de Moura
Arlindo L. Oliveira (Eds.)

String Processing and Information Retrieval

10th International Symposium, SPIRE 2003
Manaus, Brazil, October 8-10, 2003
Proceedings



Springer

Series Editors

Gerhard Goos, Karlsruhe University, Germany
Juris Hartmanis, Cornell University, NY, USA
Jan van Leeuwen, Utrecht University, The Netherlands

Volume Editors

Mario A. Nascimento
University of Alberta
Department of Computing Science
Edmonton, Alberta T6G 2E8, Canada
E-mail: mn@cs.ualberta.ca

Edleno S. de Moura
Universidade Federal do Amazonas
Departamento de Ciência da Computação
Av. Gal. Octavio Jordão Ramos, 3000, 69077-000 Manaus, Brazil
E-mail: edleno@dcc.fua.br

Arlindo L. Oliveira
Instituto Superior Técnico, INESC-ID
Avenida Duque d'Avila, 9, 1000-138 Lisboa, Portugal
E-mail: aml@inesc-id.pt

Cataloging-in-Publication Data applied for

A catalog record for this book is available from the Library of Congress.

Bibliographic information published by Die Deutsche Bibliothek
Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie;
detailed bibliographic data is available in the Internet at <<http://dnb.ddb.de>>.

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

ISSN 0302-9743

ISBN 3-540-20177-7 Springer-Verlag 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-Verlag. Violations are liable for prosecution under the German Copyright Law.

Springer-Verlag Berlin Heidelberg New York
a member of BertelsmannSpringer Science+Business Media GmbH

<http://www.springer.de>

© Springer-Verlag Berlin Heidelberg 2003
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Boller Mediendesign
Printed on acid-free paper SPIN: 10961155 06/3142 5 4 3 2 1 0

Preface

This volume of the Lecture Notes in Computer Science series provides a comprehensive, state-of-the-art survey of recent advances in string processing and information retrieval. It includes invited and research papers presented at the 10th International Symposium on String Processing and Information Retrieval, SPIRE 2003, held in Manaus, Brazil.

SPIRE 2003 received 54 full submissions from 17 countries, namely: Argentina (2), Australia (2), Brazil (9), Canada (1), Chile (4), Colombia (2), Czech Republic (1), Finland (10), France (1), Japan (2), Korea (5), Malaysia (1), Portugal (2), Spain (6), Turkey (1), UK (1), USA (4) – the numbers in parentheses indicate the number of submissions from that country. In the nontrivial task of selecting the papers to be published in these proceedings we were fortunate to count on a very international program committee with 43 members, representing all continents but one. These people, in turn, used the help of 40 external referees. During the review process all but a few papers had four reviews instead of the usual three, and at the end 21 submissions were accepted to be published as full papers, yielding an acceptance rate of about 38%. An additional set of six short papers was also accepted. The technical program spans over the two well-defined scopes of SPIRE (string processing and information retrieval) with a number of papers also focusing on important application domains such as bioinformatics.

SPIRE 2003 also features two invited speakers: Krishna Bharat (Google, Inc.) and João Meidanis (State Univ. of Campinas and Scylla Bioinformatics). We appreciate their willingness to help with SPIRE's technical program, and also their kindness in providing an invited paper covering the topics of their talks.

On behalf of SPIRE's program and steering committee we thank all authors, reviewers and attendees of this year's symposium. The local arrangements team, headed by Altigran Soares da Silva, also deserves a special acknowledgment. The program committee chair, Mario A. Nascimento, wishes to thank Alberto Laender and Ricardo Baeza-Yates for timely and insightful discussions, as well as for inviting him, on behalf of SPIRE's steering committee, to chair this year's program committee.

As we say in Portuguese: *Bem vindos a Manaus!* (Welcome to Manaus!)

July 2003

Mario A. Nascimento
Program Committee Chair

Edleno S. de Moura
General Chair

Arlindo L. Oliveira
Publications Chair

SPIRE 2003 Organization

General Chair

Edleno S. de Moura, Universidade Federal do Amazonas, Brazil

Program Committee Chair

Mario A. Nascimento, University of Alberta, Canada

Publications Chair

Arlindo L. Oliveira, Instituto Superior Técnico/INESC-ID, Portugal

Local Arrangements

Altigran Soares da Silva, Universidade Federal do Amazonas, Brazil

Steering Committee

Ricardo Baeza-Yates, Universidad de Chile, Chile

Berthier Ribeiro-Neto, Universidade Federal de Minas Gerais, Brazil

Nivio Ziviani, Universidade Federal de Minas Gerais, Brazil

Arlindo L. Oliveira, Instituto Superior Técnico/INESC-ID, Portugal

Alberto Laender, Universidade Federal de Minas Gerais, Brazil

Program Committee

Alberto Apostolico (Purdue Univ., USA)

Ricardo Baeza-Yates (Univ. de Chile, Chile)

Michael Benedikt (Bell Labs, USA)

Elisa Bertino (Univ. of Milan, Italy)

Nieves Brisaboa (Universidad de A Coruña, Spain)

Edgar Chavez (Universidad Michoacana, Mexico)

Roger Chiang (Univ. of Cincinnati, USA)

Maxime Crochemore (Université de Marne-la-Vallée, France)

Bruce Croft (Univ. of Massachusetts, USA)

Edward Fox (Virginia Tech, USA)

Juliana Freire (Oregon Graduate Institute, USA)

Ophir Frieder (Illinois Institute of Technology, USA)

Pablo de la Fuente (Univ. of Valladolid, Spain)

Norbert Fuhr (Univ. of Duisburg, Germany)

David Grossman (Illinois Institute of Technology, USA)

David Hawking (Australian National Univ., Australia)
 Carlos Alberto Heuser (UFRGS, Brazil)
 Thomas Roelleke (Queen Mary Univ. of London, UK)
 Costas Iliopoulos (King's College London, UK)
 Alberto Laender (Federal Univ. of Minas Gerais, Brazil)
 Ee-Peng Lim (Nanyang Technological Univ., Singapore)
 Dekang Lin (Univ. of Alberta, Canada)
 Joel Martin (National Research Council, Canada)
 João Meidanis (Univ. of Campinas, Brazil)
 Massimo Melucci (Univ. of Padova, Italy)
 Alistair Moffat (Univ. of Melbourne, Australia)
 Gonzalo Navarro (Universidad de Chile, Chile)
 Charles Nicholas (Univ. Maryland, Baltimore County, USA)
 Jian-Yun Nie (Univ. of Montreal, Canada)
 Arlindo L. Oliveira, (Instituto Superior Técnico/INESC-ID, Portugal)
 Gabriella Pasi (CNR, Italy)
 Berthier Ribeiro-Neto (Federal Univ. of Minas Gerais, Brazil)
 Altigran Silva (Federal Univ. of Amazonia, Brazil)
 Marie-France Sagot (INRIA Rhône-Alpes, France)
 Fabrizio Sebastiani (CNR, Italy)
 Ayumi Shinohara (Kyushu University, Japan)
 Amit Singhal (Google, USA)
 Dan Suciu (Univ. of Washington, USA)
 Jorma Tarhio (Helsinki Univ. of Technology, Finland)
 Ulrich Thiel (GMD-IPSI, Germany)
 Frank Tompa (Univ. of Waterloo, Canada)
 Nivio Ziviani (Federal Univ. of Minas Gerais, Brazil)
 Justin Zobel (RMIT, Australia)

External Referees

Alan Watt	Heikki Hyyrö
Alex Lopez-Ortiz	Henrik Nottelmann
Ana Cardoso Cachopo	Hugh E. Williams
Andrew Turpin	Joyce Christina de Paiva Carvalho
Bruno Pôssas	Juliano Palmieri Lage
Carina Friederich Dorneles	Jussara Marques de Almeida
Carlos Castillo	Kai Großjohann
Catalina Luiza Antonie	Kimmo Fredriksson
Cecil Eng Huang Chua	Juha Karkkainen
Claudine Santos Badue	Kjell Lemström
Claus-Peter Klas	Laurent Mouchard
Gudrun Fischer	Mara Abel
Guohui Lin	Marco Antonio Pinheiro de Cristo
Hannu Peltola	Miguel R. Penabad

VIII SPIRE 2003 Organization

Pável Calado
Reem K. Al-Halimi
Renato Ferreira
Robert Warren
Roberto Grossi
Ronaldo dos Santos Mello
Saied Tahaghoghi

Tomasz Radzik
Takuya Kida
Wagner Meira Jr.
Wong Hao Chi
Yoan J. Pinzon
Zanoni Dias

Table of Contents

Invited Papers

Patterns on the Web	1
<i>Krishna Bharat</i>	
Current Challenges in Bioinformatics	16
<i>João Meidanis</i>	

Web Algorithms

What's Changed? Measuring Document Change in Web Crawling for Search Engines	28
<i>Halil Ali and Hugh E. Williams</i>	
Link Information as a Similarity Measure in Web Classification	43
<i>Marco Cristo, Pavel Calado, Edleno Silva de Moura, Nivio Ziviani, and Berthier Ribeiro-Neto</i>	
A Three Level Search Engine Index Based in Query Log Distribution	56
<i>Ricardo Baeza-Yates and Felipe Saint-Jean</i>	

Bit-Parallel Algorithms

Row-wise Tiling for the Myers' Bit-Parallel Approximate String Matching Algorithm	66
<i>Kimmo Fredriksson</i>	
Alternative Algorithms for Bit-Parallel String Matching	80
<i>Hannu Peltola and Jorma Tarhio</i>	
Bit-Parallel Approximate String Matching Algorithms with Transposition	95
<i>Heikki Hyyrö</i>	

Compression

Processing of Huffman Compressed Texts with a Super-Alphabet	108
<i>Kimmo Fredriksson and Jorma Tarhio</i>	
(S,C)-Dense Coding: An Optimized Compression Code for Natural Language Text Databases	122
<i>Nieves R. Brisaboa, Antonio Fariña, Gonzalo Navarro, and María F. Esteller</i>	

Linear-Time Off-Line Text Compression by Longest-First Substitution . . .	137
<i>Shunsuke Inenaga, Takashi Funamoto, Masayuki Takeda, and Ayumi Shinohara</i>	

SCM: Structural Contexts Model for Improving Compression in Semistructured Text Databases	153
<i>Joaquín Adiego, Gonzalo Navarro, and Pablo de la Fuente</i>	

Categorization and Ranking

Ranking Structured Documents Using Utility Theory in the Bayesian Network Retrieval Model	168
<i>Fabio Crestani, Luis M. de Campos, Juan M. Fernández-Luna, and Juan F. Huete</i>	

An Empirical Comparison of Text Categorization Methods	183
<i>Ana Cardoso-Cachopo and Arlindo L. Oliveira</i>	

Improving Text Retrieval in Medical Collections Through Automatic Categorization	197
<i>Rodrigo F. Vale, Berthier Ribeiro-Neto, Luciano R.S. de Lima, Alberto H.F. Laender, and Hermes R.F. Junior</i>	

Music Retrieval

A Bit-Parallel Suffix Automaton Approach for (δ, γ) -Matching in Music Retrieval	211
<i>Maxime Crochemore, Costas S. Iliopoulos, Gonzalo Navarro, and Yoan J. Pinzon</i>	

Flexible and Efficient Bit-Parallel Techniques for Transposition Invariant Approximate Matching in Music Retrieval	224
<i>Kjell Lemström and Gonzalo Navarro</i>	

Multilingual Information Retrieval

FindStem: Analysis and Evaluation of a Turkish Stemming Algorithm . . .	238
<i>Hayri Sever and Yülan Bitirim</i>	

Non-adjacent Digrams Improve Matching of Cross-Lingual Spelling Variants	252
<i>Heikki Keskustalo, Ari Pirkola, Kari Visala, Erkkä Leppänen, and Kalervo Järvelin</i>	

The Implementation and Evaluation of a Lexicon-Based Stemmer	266
<i>Gilberto Silva and Claudia Oliveira</i>	

French Noun Phrase Indexing and Mining for an Information Retrieval System	277
<i>Hatem Haddad</i>	

Subsequences and Distributed Algorithms

New Refinement Techniques for Longest Common Subsequence Algorithms	287
<i>Lasse Bergroth, Harri Hakonen, and Juri Väisänen</i>	
The Size of Subsequence Automaton	304
<i>Zdeněk Troníček and Ayumi Shinohara</i>	
Distributed Query Processing Using Suffix Arrays	311
<i>Mauricio Marín and Gonzalo Navarro</i>	

Algorithms on Strings and Trees

BFT: Bit Filtration Technique for Approximate String Join in Biological Databases	326
<i>S. Alireza Aghili, Divyakant Agrawal, and Amr El Abbadi</i>	
A Practical Index for Genome Searching	341
<i>Heikki Hyvrö and Gonzalo Navarro</i>	
Using WordNet for Word Sense Disambiguation to Support Concept Map Construction	350
<i>Alberto J. Cañas, Alejandro Valerio, Juan Lalinde-Pulido, Marco Carvalho, and Marco Arguedas</i>	
Memory-Adaptative Dynamic Spatial Approximation Trees	360
<i>Diego Arroyuelo, Francisca Muñoz, Gonzalo Navarro, and Nora Reyes</i>	
Large Edit Distance with Multiple Block Operations	369
<i>Dana Shapira and James A. Storer</i>	
Author Index	379