Lecture Notes in Computer Science 5721

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
University of Dortmund, 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 of Computer Science, Saarbruecken, Germany

Jussi Karlgren Jorma Tarhio
Heikki Hyyrö (Eds.)

String Processing and Information Retrieval

16th International Symposium, SPIRE 2009
Saariselkä, Finland, August 25-27, 2009
Proceedings

Springer

Volume Editors

Jussi Karlgren
Swedish Institute of Computer Science
Kista, Sweden
E-mail: jussi@sics.se

Jorma Tarhio
Helsinki University of Technology
Espoo, Finland
E-mail: tarhio@cs.hut.fi

Heikki Hyyrö
University of Tampere
Tampere, Finland
E-mail: heikki.hyyro@cs.uta.fi

Library of Congress Control Number: 2009932181

CR Subject Classification (1998): E.5, F.2.2, H.3.3, E.1, E.2, H.2.8, I.5

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

ISSN 0302-9743
ISBN-10 3-642-03783-6 Springer Berlin Heidelberg New York
ISBN-13 978-3-642-03783-2 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.com

© Springer-Verlag Berlin Heidelberg 2009
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12738502 06/3180 5 4 3 2 1 0

Preface

This volume contains the papers presented at the 16th International Symposium on String Processing and Information Retrieval (SPIRE 2009), held during August 25–27, 2009 in Saariselkä, Finland.

The annual SPIRE conference provides researchers within fields related to string processing and/or information retrieval a possibility to present their original contributions and to meet and discuss with other researchers with similar interests. The Call for Papers invited submissions related to string processing (dictionary algorithms; text searching; pattern matching; text and sequence compression; automata-based string processing), information retrieval (information retrieval models; indexing; ranking and filtering; interface design; visualization; benchmarking), natural language processing (text analysis; text mining; machine learning; information extraction; language models; knowledge representation), search applications and usage (cross-lingual information access systems; multimedia information access; digital libraries; collaborative retrieval and Web-related applications; semi-structured data retrieval; evaluation), and interaction of biology and computation (DNA sequencing and applications in molecular biology; evolution and phylogenetics; recognition of genes and regulatory elements; sequence driven protein structure prediction).

The papers presented at the symposium were selected from 84 submissions written by authors from 28 different countries. Each submission was reviewed by at least two and on average 2.9 reviewers. The Committee accepted 34 papers ($\approx 40\%$): 22 papers for 25-minute and 12 papers for 15-minute presentations. In addition to these, SPIRE 2009 also featured invited talks by Mehryar Mohri (New York University, USA) and Kalervo Järvelin (University of Tampere, Finland).

The Program Committee voted to give the Best Paper Award to Amihood Amir and Haim Parienty for their paper "Towards a Theory of Patches."

We are especially thankful to the members of the Program Committee, who provided us with thorough and timely reviews despite the fact that the reviewing schedule was unusually tight due to the early date of the 2009 symposium.

Additional thanks go to the following people for their generous helpfulness and support throughout the process of planning and organizing SPIRE 2009: SPIRE Steering Committee coordinator Ricardo Baeza-Yates, the editorial office staff at Springer, the office personnel at the University of Tampere, and the staff at the conference hotel Riekonlinna in Saariselkä.

We are grateful for the financial support provided by the Department of Computer Sciences of University of Tampere, the Federation of Finnish Learned Societies, Yahoo! Research, the Department of Computer Science of University of Helsinki, and the Swedish Institute of Computer Science.

June 2009

Jussi Karlgren
Jorma Tarhio
Heikki Hyyrö

SPIRE 2009 Organization

Organizing Institution

Department of Computer Sciences, University of Tampere, Finland

Organizing Committee

Local Organization Chair

Heikki Hyyrö	University of Tampere, Finland

Program Committee Chairs

Jussi Karlgren	Swedish Institute of Computer Science, Sweden
Jorma Tarhio	Helsinki University of Technology, Finland

Program Committee

Tatsuya Akutsu	Kyoto University, Japan
Srinivas Aluru	Iowa State University, USA
Amihood Amir	Bar-Ilan University, Israel, and Johns Hopkins University, USA
Alberto Apostolico	Georgia Institute of Technology, USA, and University of Padova, Italy
Ricardo Baeza-Yates	Yahoo! Research, Spain and Chile, and CWR/DCC, University of Chile, Chile
Roberto Basili	University of Rome, Tor Vergata, Italy
Carlos Castillo	Yahoo! Research, Spain
Edgar Chávez	Universidad Michoacana, Mexico
Maxime Crochemore	King's College London, UK, and Paris-Est University, France
Paolo Ferragina	University of Pisa, Italy
Raffaele Giancarlo	University of Palermo, Italy
Costas Iliopoulos	King's College London, UK
Noriko Kando	National Institute of Informatics, Japan
Kimmo Kettunen	Kymenlaakso University of Applied Sciences, Finland
Dong Kyue Kim	Hanyang University, Korea
Mounia Lalmas	University of Glasgow, UK

Gad M. Landau	University of Haifa, Israel, and Polytechnic Institute of NYU, USA
Edgar Meij	University of Amsterdam, The Netherlands
Alistair Moffat	University of Melbourne, Australia
Veli Mäkinen	University of Helsinki, Finland
Gonzalo Navarro	University of Chile, Chile
Fredrik Olsson	Swedish Institute of Computer Science, Sweden
Marco Pennacchiotti	Yahoo! Inc., USA
Yoan J. Pinzón	National University of Colombia, Colombia
Mathieu Raffinot	LIAFA, CNRS-University Denis Diderot (Paris-7), France
Ian Ruthven	University of Strathclyde, UK
Wojciech Rytter	University of Warsaw, Poland
Marie-France Sagot	INRIA, France
Tetsuya Sakai	Microsoft Research Asia, China
Ayumi Shinohara	Tohoku University, Japan
Borkur Sigurbjornsson	Yahoo! Research, Spain
William F. Smyth	McMaster University, Canada, and Curtin University, Australia
Ian Soboroff	National Institute of Standards and Technology, USA
Wing-Kin Sung	National University of Singapore, Singapore
Masayuki Takeda	Kyushu University, Japan
Jeffrey Vitter	Texas A&M University, USA
Nivio Ziviani	Federal University of Minas Gerais, Brazil

Additional Reviewers

Cyril Allauzen
Humberto Almeida
Pavlos Antoniou
Diego Arroyuelo
Claudine Badue
Hideo Bannai
Guillaume Blin
Fabiano C. Botelho
Wladmir Brandao
Marc Bron
Manolis Christodoulakis
Krzysztof Ciebiera
Francisco Claude
Chiara Epifanio
Roberto Grossi
Jean-Loup Guillaume
Danny Hermelin
Bouke Huurnink
Shunsuke Inenaga
Takuya Kida
Shmuel Klein
Juha Kärkkäinen
Anisio Lacerda
Vincent Lacroix
Shir Landau
Avivit Levy
Sabrina Mantaci
Juan Mendivelso
Guilherme Vale Menezes
Spiros Michalakopoulos
Mirella M. Moro
Kazuyuki Narisawa

Laurence Park
Marcin Piatkowski
Solon Pissis
Benjamin Piwowarski
Simon Puglisi
Jakub Radoszewski
Stéphane Raux
Hiroshi Sakamoto
Riva Shalom
Jouni Sirén
Wilson Soto
Torsten Suel
German Tischler
Manos Tsagkias
Dekel Tsur
Oscar Täckström
Rossano Venturini
Anh Vo
Niko Välimäki
William Webber
Oren Weimann
Xiao Yang

SPIRE Steering Committee

Amihood Amir	Bar-Ilan University, Israel, and Johns Hopkins University, USA
Ricardo Baeza-Yates (Coordinator)	Yahoo! Research, Spain and Chile, and CWR/DCC, University of Chile, Chile
Fabio Crestani	University of Lugano, Switzerland
Paolo Ferragina	University of Pisa, Italy
Alistair Moffat	University of Melbourne, Australia
Berthier Ribeiro-Neto	Federal University of Minas Gerais, Brazil
Mark Sanderson	University of Sheffield, UK
Andrew Turpin	RMIT University, Australia
Nivio Ziviani	Federal University of Minas Gerais, Brazil

SPIRE Venues

SPIRE 2009 was the 16th edition of the Symposium on String Processing and Information Retrieval. The four first events concentrated mainly on string processing and were held in South America under the title South American Workshop on String Processing (WSP) in 1993, 1995, 1996 and 1997. WSP was transformed into SPIRE in 1998, when the scope of the conference was broadened to include also the area of information retrieval.

1993	Belo Horizonte, Brazil	2002	Lisbon, Portugal
1995	Valparaíso, Chile	2003	Manaus, Brazil
1996	Recife, Brazil	2004	Padova, Italy
1997	Valparaíso, Chile	2005	Buenos Aires, Argentina
1998	Santa Cruz de la Sierra, Bolivia	2006	Glasgow, UK
1999	Cancún, Mexico	2007	Santiago, Chile
2000	A Coruña, Spain	2008	Melbourne, Australia
2001	Laguna de San Rafael, Chile	2009	Saariselkä, Finland

Table of Contents

Algorithms on Trees

Range Quantile Queries: Another Virtue of Wavelet Trees 1
Travis Gagie, Simon J. Puglisi, and Andrew Turpin

Constant Factor Approximation of Edit Distance of Bounded Height Unordered Trees 7
Daiji Fukagawa, Tatsuya Akutsu, and Atsuhiro Takasu

k^2-Trees for Compact Web Graph Representation 18
Nieves R. Brisaboa, Susana Ladra, and Gonzalo Navarro

On-Line Construction of Parameterized Suffix Trees 31
Taehyung Lee, Joong Chae Na, and Kunsoo Park

Compressed Indexes

Succinct Text Indexing with Wildcards 39
Alan Tam, Edward Wu, Tak-Wah Lam, and Siu-Ming Yiu

A Compressed Enhanced Suffix Array Supporting Fast String Matching 51
Enno Ohlebusch and Simon Gog

Compressed Suffix Arrays for Massive Data 63
Jouni Sirén

On Entropy-Compressed Text Indexing in External Memory 75
Wing-Kai Hon, Rahul Shah, Sharma V. Thankachan, and Jeffrey Scott Vitter

Compression

A Linear-Time Burrows-Wheeler Transform Using Induced Sorting 90
Daisuke Okanohara and Kunihiko Sadakane

Novel and Generalized Sort-Based Transform for Lossless Data Compression 102
Kazumasa Inagaki, Yoshihiro Tomizawa, and Hidetoshi Yokoo

A Two-Level Structure for Compressing Aligned Bitexts 114
Joaquín Adiego, Nieves R. Brisaboa, Miguel A. Martínez-Prieto, and Felipe Sánchez-Martínez

Directly Addressable Variable-Length Codes 122
Nieves R. Brisaboa, Susana Ladra, and Gonzalo Navarro

Content Analysis

Identifying the Intent of a User Query Using Support Vector Machines 131
Marcelo Mendoza and Juan Zamora

Syntactic Query Models for Restatement Retrieval 143
Niranjan Balasubramanian and James Allan

Use of Co-occurrences for Temporal Expressions Annotation 156
Olga Craveiro, Joaquim Macedo, and Henrique Madeira

On-Demand Associative Cross-Language Information Retrieval 165
André Pinto Geraldo, Viviane P. Moreira, and Marcos A. Gonçalves

A Comparison of Data-Driven Automatic Syllabification Methods 174
Connie R. Adsett and Yannick Marchand

Indexing

Efficient Index for Retrieving Top-k Most Frequent Documents 182
Wing-Kai Hon, Rahul Shah, and Shih-Bin Wu

Fast Single-Pass Construction of a Half-Inverted Index 194
Marjan Celikik and Hannah Bast

Two-Dimensional Distributed Inverted Files 206
Esteban Feuerstein, Mauricio Marin, Michel Mizrahi, Veronica Gil-Costa, and Ricardo Baeza-Yates

Indexing Variable Length Substrings for Exact and Approximate Matching 214
Gonzalo Navarro and Leena Salmela

String Algorithms and Bioinformatics

Expectation of Strings with Mismatches under Markov Chain Distribution 222
Cinzia Pizzi and Mauro Bianco

Consensus Optimizing Both Distance Sum and Radius 234
Amihood Amir, Gad M. Landau, Joong Chae Na, Heejin Park, Kunsoo Park, and Jeong Seop Sim

Faster Algorithms for Sampling and Counting Biological Sequences 243
Christina Boucher

String Algorithms and Theory I

Towards a Theory of Patches 254
Amihood Amir and Haim Parienty

The Frequent Items Problem, under Polynomial Decay, in the Streaming Model 266
Guy Feigenblat, Ofra Itzhaki, and Ely Porat

Improved Approximation Results on the Shortest Common Supersequence Problem 277
Zvi Gotthilf and Moshe Lewenstein

String Algorithms and Theory II

Set Intersection and Sequence Matching 285
Ariel Shiftan and Ely Porat

Generalised Matching 295
Raphael Clifford, Aram W. Harrow, Alexandru Popa, and Benjamin Sach

Practical Algorithms for the Longest Common Extension Problem 302
Lucian Ilie and Liviu Tinta

Using and Understanding Usage

A Last-Resort Semantic Cache for Web Queries 310
Flavio Ferrarotti, Mauricio Marin, and Marcelo Mendoza

A Task-Based Evaluation of an Aggregated Search Interface 322
Shanu Sushmita, Hideo Joho, and Mounia Lalmas

Efficient Language-Independent Retrieval of Printed Documents without OCR 334
Walid Magdy, Kareem Darwish, and Motaz El-Saban

Sketching Algorithms for Approximating Rank Correlations in Collaborative Filtering Systems 344
Yoram Bachrach, Ralf Herbrich, and Ely Porat

Author Index 353