Lecture Notes in Computer Science 5280

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Amihood Amir Andrew Turpin
Alistair Moffat (Eds.)

String Processing and Information Retrieval

15th International Symposium, SPIRE 2008
Melbourne, Australia, November 10-12, 2008
Proceedings

Springer

Volume Editors

Amihood Amir
Bar-Ilan University
Ramat-Gan, Israel
E-mail: amir@cs.biu.ac.il

Andrew Turpin
RMIT University
Melbourne, Australia
E-mail: aht@cs.rmit.edu.au

Alistair Moffat
The University of Melbourne
Carlton, Australia
E-mail: alistair@csse.unimelb.edu.au

Library of Congress Control Number: 2008938187

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 3-540-89096-3 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-89096-6 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12557561 06/3180 5 4 3 2 1 0

Preface

This volume contains the papers presented at the 15th String Processing and Information Retrieval Symposium (SPIRE), held in Melbourne, Australia, during November 10–12, 2008.

The papers presented at the symposium were selected from 54 papers submitted in response to the Call For Papers. Each submission was reviewed by a minimum of two, and usually three, Program Committee members, who are experts drawn from around the globe. The committee accepted 25 papers (46%), with the successful authors also covering a broad range of continents. The paper "An Efficient Linear Space Algorithm for Consecutive Suffix Alignment Under Edit Distance" by Heikki Hyyrö was selected for the Best Paper Award, while Dina Sokol was awarded the Best Reviewer Award for excellent contributions to the reviewing process. The program also included two invited talks: David Hawking, chief scientist at the Internet and enterprise search company Funnelback Pty. Ltd. based in Australia; and Gad Landau, from the Department of Computer Science at Haifa University, Israel.

SPIRE has its origins in the South American Workshop on String Processing which was first held in 1993. Starting in 1998, the focus of the symposium was broadened to include the area of information retrieval due to the common emphasis on information processing. The first 14 meetings were held in Belo Horizonte, Brazil (1993); Valparaiso, Chile (1995); Recife, Brazil (1996); Valparaiso, Chile (1997); Santa Cruz, Bolivia (1998); Cancun, Mexico (1999); A Coruña, Spain (2000); Laguna San Rafael, Chile (2001); Lisbon, Portugal (2002); Manaus, Brazil (2003); Padova, Italy (2004); Buenos Aires, Argentina (2005); Glasgow, UK (2006); and Santiago, Chile (2007).

The annual SPIRE conference provides an opportunity for researchers to present original contributions on areas such as string processing (the searching, compression and mining of text, pattern matching, natural language processing, automata based string processing); information retrieval (indexing, ranking, filtering, cross-lingual IR systems, multimedia IR, digital libraries, collaborative retrieval, Web-related applications); interaction of biology and computation particularly related to string processing and retrieval; and information retrieval languages and applications (XML, SGML, information retrieval from semi-structured data, generation of structured data from text).

While many people have helped to make this conference possible, we particularly thank the members of the Program Committee and the additional reviewers who worked hard to ensure the timely review of all submitted manuscripts. We also thank William Webber, who compiled the proceedings, and Shane Culpepper, who maintained the website for the conference. We are grateful to

Yahoo! Research for providing funding in support of student attendees. Submissions were managed using the EasyChair conference system.

August 2008

Amihood Amir
Andrew Turpin
Alistair Moffat

Organization

Conference Chair

Alistair Moffat	University of Melbourne, Australia

Program Chairs

Amihood Amir	Bar-Ilan University, Israel and Johns Hopkins University, USA
Andrew Turpin	RMIT University, Australia

Program Committee

Amihood Amir	Bar-Ilan University, Israel
Mike Atallah	Purdue University, USA
Gary Benson	Boston University, USA
Bodo Billerbeck	Microsoft Research, Cambridge
Carlos Castillo	Yahoo! Research, Spain
Charlie Clarke	University of Waterloo, Canada
Bruce Croft	University of Massachusetts, Amherst, USA
J. Shane Culpepper	University of Melbourne, Australia
Paolo Ferragina	University of Pisa, Italy
Edward Fox	Virginia Tech, USA
Jan Holub	Czech Technical University, Czech Republic
Shunsuke Inenaga	Kyushu University, Japan
Rao Kosaraju	Johns Hopkins University, USA
Avivit Levy	Shenkar College, Israel
Moshe Lewenstein	Bar Ilan University, Israel
Noa Lewenstein	Netanya College, Israel
Giovanni Manzini	University of East Piedmont, Italy
Massimo Melucci	University of Padua, Italy
Laurent Mouchard	University of Rouen, France and King's College, London, UK
Gonzalo Navarro	University of Chile, Chile
Igor Nor	University of Bristol, UK
Heejin Park	Hanyang University, Korea
Kunsoo Park	Seoul National University, Korea
Ron Y. Pinter	Technion, Israel
Ely Porat	Bar Ilan University, Israel
Simon Puglisi	RMIT University, Australia
Mathieu Raffinot	CNRS, France

Kunihiko Sadakane	Kyushu University, Japan
Falk Scholer	RMIT University, Australia
Ayumi Shinohara	Tohoku University, Japan
Fabrizio Silvestri	CNR, Italy
Steven Skiena	Stony Brook University, USA
Bill Smyth	McMaster University, Canada and Curtin University, Australia
Dina Sokol	Brooklyn College, New York, USA
Wing-Kin Sung	NUS, Singapore
Andrew Turpin	RMIT University, Australia
Alexandra Uitdenbogerd	RMIT University, Australia
Esko Ukkonen	University of Helsinki, Finland
Anh Vo	University of Melbourne, Australia

External Reviewers

Miroslav Balik	Juha Kärkkäinen	Justin Tojeira
Hideo Bannai	Dong Kyue Kim	Raymond Wan
Peter Bruza	Joong Chae Na	Mingfang Wu
Gabriele Capannini	Kazuyuki Narisawa	
Ben Carterette	Raffaele Perego	
Michael Harris	Royi Ronen	

Local Organization

Shane Culpepper	University of Melbourne, Australia
Alistair Moffat	University of Melbourne, Australia
Simon Puglisi	RMIT University, Australia
William Webber	University of Melbourne, Australia
Justin Zobel	NICTA and University of Melbourne, Australia

Table of Contents

"Search Is a Solved Problem" and Other Annoying Fallacies (Invited Talk) 1
David Hawking

Approximate Runs – Revisited (Invited Talk) 2
Gad M. Landau

Engineering Radix Sort for Strings 3
Juha Kärkkäinen and Tommi Rantala

Faster Text Fingerprinting 15
Roman Kolpakov and Mathieu Raffinot

Context-Sensitive Grammar Transform: Compression and Pattern Matching 27
Shirou Maruyama, Yohei Tanaka, Hiroshi Sakamoto, and Masayuki Takeda

Improved Variable-to-Fixed Length Codes 39
Shmuel T. Klein and Dana Shapira

Term Impacts as Normalized Term Frequencies for BM25 Similarity Scoring 51
Vo Ngoc Anh, Raymond Wan, and Alistair Moffat

The Effect of Weighted Term Frequencies on Probabilistic Latent Semantic Term Relationships 63
Laurence A.F. Park and Kotagiri Ramamohanarao

Comparison of s-gram Proximity Measures in Out-of-Vocabulary Word Translation 75
Anni Järvelin and Antti Järvelin

Speeding Up Pattern Matching by Text Sampling 87
Francisco Claude, Gonzalo Navarro, Hannu Peltola, Leena Salmela, and Jorma Tarhio

Mismatch Sampling 99
Raphaël Clifford, Klim Efremenko, Benny Porat, Ely Porat, and Amir Rothschild

Sliding CDAWG Perfection 109
Martin Senft and Tomáš Dvořák

Self-indexing Natural Language 121
Nieves R. Brisaboa, Antonio Fariña, Gonzalo Navarro, Angeles S. Places, and Eduardo Rodríguez

New Perspectives on the Prefix Array 133
W.F. Smyth and Shu Wang

Indexed Hierarchical Approximate String Matching 144
Luís M.S. Russo, Gonzalo Navarro, and Arlindo L. Oliveira

An Efficient Linear Space Algorithm for Consecutive Suffix Alignment under Edit Distance (*Short Preliminary Paper*) 155
Heikki Hyyrö

Run-Length Compressed Indexes Are Superior for Highly Repetitive Sequence Collections 164
Jouni Sirén, Niko Välimäki, Veli Mäkinen, and Gonzalo Navarro

Practical Rank/Select Queries over Arbitrary Sequences 176
Francisco Claude and Gonzalo Navarro

Clique Analysis of Query Log Graphs 188
Alexandre P. Francisco, Ricardo Baeza-Yates, and Arlindo L. Oliveira

Out of the Box Phrase Indexing 200
Frederik Transier and Peter Sanders

Approximated Pattern Matching with the L_1, L_2 and L_∞ Metrics 212
Ohad Lipsky and Ely Porat

Interchange Rearrangement: The Element-Cost Model 224
Oren Kapah, Gad M. Landau, Avivit Levy, and Nitsan Oz

$\delta\gamma$ – Parameterized Matching 236
Inbok Lee, Juan Mendivelso, and Yoan J. Pinzón

Pattern Matching with Pair Correlation Distance 249
Benny Porat, Ely Porat, and Asaf Zur

Some Approximations for Shortest Common Nonsubsequences and Supersequences 257
Vadim G. Timkovsky

On the Structure of Small Motif Recognition Instances 269
Christina Boucher, Daniel G. Brown, and Stephane Durocher

Exact Distribution of a Spaced Seed Statistic for DNA Homology Detection 282
Gary Benson and Denise Y.F. Mak

Author Index 295