

*Commenced Publication in 1973*

Founding and Former Series Editors:

Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

## Editorial Board

David Hutchison

*Lancaster University, Lancaster, 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, Zürich, Switzerland*

John C. Mitchell

*Stanford University, Stanford, CA, USA*

Moni Naor

*Weizmann Institute of Science, Rehovot, Israel*

C. Pandu Rangan

*Indian Institute of Technology, Madras, India*

Bernhard Steffen

*TU Dortmund University, Dortmund, Germany*

Demetri Terzopoulos

*University of California, Los Angeles, CA, USA*

Doug Tygar

*University of California, Berkeley, CA, USA*

Gerhard Weikum

*Max Planck Institute for Informatics, Saarbrücken, Germany*

More information about this series at <http://www.springer.com/series/7407>

Costas S. Iliopoulos · Simon J. Puglisi  
Emine Yilmaz (Eds.)

# String Processing and Information Retrieval

22nd International Symposium, SPIRE 2015  
London, UK, September 1–4, 2015  
Proceedings

*Editors*

Costas S. Iliopoulos  
King's College London  
London  
UK

Emine Yilmaz  
University College London  
London  
UK

Simon J. Puglisi  
University of Helsinki  
Helsinki  
Finland

ISSN 0302-9743                      ISSN 1611-3349 (electronic)  
Lecture Notes in Computer Science  
ISBN 978-3-319-23825-8              ISBN 978-3-319-23826-5 (eBook)  
DOI 10.1007/978-3-319-23826-5

Library of Congress Control Number: 2015947399

LNCS Sublibrary: SL1 – Theoretical Computer Science and General Issues

Springer Cham Heidelberg New York Dordrecht London  
© Springer International Publishing Switzerland 2015

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, 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.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media  
([www.springer.com](http://www.springer.com))

## Preface

From humble beginnings as a regional meeting focused on string algorithms (under the auspices of WSP: South American Workshop on String Processing), the International Symposium on String Processing and Information Retrieval (SPIRE) has, in the last two decades, developed into a vibrant conference at the broad nexus of algorithms and data structures for sequences and graphs, data compression, databases, data mining, and information retrieval.

This volume contains the papers presented at SPIRE 2015, the 22nd International Symposium on String Processing and Information Retrieval, held from August 31 to September 2, 2015 in London, UK, in the Great Hall of King's College London's Strand Campus. There were 90 submissions. Each submission was reviewed by at least 3, and on average 3.1, Program Committee members. The committee decided to accept 34 papers. The program also included 3 invited talks.

The main conference, which spanned the three days during 1–3 September, featured keynote talks by Aristides Gionis (Aalto University, Finland), Mounia Lalmas (Yahoo! Labs London, UK), and Rajeev Raman (University of Leicester, UK), together with presentations by authors of the 33 accepted papers. The 10th Workshop on Compression, Text, and Algorithms (WCTA 2015) was then held on September 4, the day immediately after the main conference, as has become a recent tradition. WCTA was coordinated this year by Travis Gagie and Tatiana Starikovskaya, and featured a keynote talk from Richard Durbin of the Wellcome Trust Sanger Institute, UK.

We take this opportunity to thank King's College London for its generous sponsorship of SPIRE this year. Our deep thanks also go to all the members of this year's Program Committee and additional reviewers, for the prompt, thorough reviewing and vibrant discussion that made our job as chairs easy. We thank the SPIRE Steering Committee, for giving us the opportunity to host this wonderful community of researchers in London, and finally, the Local Organizing Committee (led by Solon Pissis), for their efforts to ensure that the whole week ran smoothly, and that a relaxed and inspiring time was had by all.

July 2015

Costas S. Iliopoulos  
Simon J. Puglisi  
Emine Yilmaz

# Organization

## Program Committee

Sengor Altingovde	
Amihood Amir	Bar-Ilan University, Israel
Leif Azzopardi	University of Glasgow, UK
Golnaz Badkobeh	University of Sheffield, UK
Hideo Bannai	Kyushu University, Japan
Philip Bille	Technical University of Denmark, Denmark
Christina Boucher	Colorado State University, CO, USA
Ben Carterette	University of Delaware, DE, USA
Charles Clarke	University of Waterloo, Canada
Gianluca Demartini	University of Sheffield, UK
Johannes Fischer	TU Dortmund, Germany
Travis Gagie	University of Helsinki, Finland
Paweł Gawrychowski	University of Warsaw, Poland
Simon Gog	Karlsruhe Institute of Technology, Germany
Danny Hermelin	Ben-Gurion University, Israel
Djoerd Hiemstra	University of Twente, The Netherlands
Katja Hofmann	Microsoft Research Cambridge, UK
Costas S. Iliopoulos	King's College London, UK
Jaap Kamps	University of Amsterdam, The Netherlands
Evangelos Kanoulas	University of Amsterdam, The Netherlands
Gabriella Kazai	Lumi, Semion Ltd, UK
Juha Kärkkäinen	University of Helsinki, Finland
Susana Ladra	University of A Coruña, Spain
Gad Landau	University of Haifa, Israel
Zsuzsanna Lipták	University of Verona, Italy
Gonzalo Navarro	University of Chile, Chile
Kunsoo Park	Seoul National University, South Korea
Nadia Pisanti	University of Pisa, Italy
Solon Pissis	King's College London, UK
Simon J. Puglisi	University of Helsinki, Finland
Jakub Radoszewski	University of Warsaw, Poland
Falk Scholer	RMIT, Australia
Marinella Sciortino	University of Palermo, Italy
Jouni Sirén	Wellcome Trust Sanger Institute, UK
Tatiana Starikovskaya	University of Bristol, UK
Torsten Suel	NYU Poly, NY, USA
Yasuo Tabei	Japan Science and Technology Agency, Japan

Rossano Venturini	University of Pisa, Italy
Grace Yang	Georgetown University, DC, USA
Emine Yilmaz	University College London, UK

## Additional Reviewers

Abouelhoda, Mohamed	Inenaga, Shunsuke	Rahman, M. Sohel
Amit, Mika	Karimi, Sarvnaz	Raymond, Rob
Belazzougui, Djamel	Kayaaslan, Enver	Reynier, Pierre-Alain
Bilò, Davide	Kempa, Dominik	Rodriguez, Juan
Bingmann, Timo	Keogh, Eamonn	Rosone, Giovanna
Brown, C. Titus	Korkin, Dmitry	Rozenberg, Liat
Christiansen, Anders Roy	Kurpicz, Florian	Sacomoto, Gustavo
Cunial, Fabio	Köppl, Dominik	Salmela, Leena
Della Vedova, Gianluca	Larsson, N. Jesper	Seco, Diego
Doerr, Benjamin	Levy, Avivit	Shangsong, Liang
Eisenberg, Estrella	Lewenstein, Moshe	Song, Xuemeng
Epifanio, Chiara	Mantaci, Sabrina	Straszak, Damian
Farruggia, Andrea	Marino, Andrea	Sugimoto, Shiho
Fici, Gabriele	Markov, Ilya	Tischler, German
Ganguly, Debasis	Metke, Alejandro	Tomescu, Alexandru I.
Gasieniec, Leszek	Micale, Giovanni	Turpin, Andrew
Giaquinta, Emanuele	Na, Joong Chae	Valenzuela, Daniel
Grossi, Roberto	Nadalin, Francesca	Välimäki, Niko
Harrison, Thomas	Ozcan, Rifat	Walen, Tomasz
I, Tomohiro	Piatkowski, Marcin	Weimann, Oren
Ilie, Lucian	Pulvirenti, Alfredo	

# **Invited Talks**

# Computational Problems in Mining Urban Data

Aristides Gionis

Department of Computer Science, Aalto University

With the fast growth of smart devices and sensor networks, large amounts of data are collected recording location, activity, and mobility of people living in urban environments. Additionally, data generated on location-aware social media provide rich information about places where people spend their time (shopping malls, caf  s, parks, etc). The availability of this type of data provides novel opportunities for developing methods for extracting interesting patterns, detecting trends, modelling people's behaviour, and eventually building intelligent systems that improve the interaction of citizens with their cities and help them to utilize better the available resources. In this talk we will review recent work in the area of mining urban data. We formulate and discuss computational problems motivated by applications in detecting events, mining trajectories, finding similar neighbourhoods, and recommending locations.

# **A Journey into Evaluation: From Retrieval Effectiveness to User Engagement**

Mounia Lalmas

Yahoo! Labs London

Building retrieval systems that return results to users that satisfy their information need is one thing; Information Retrieval has a long history in evaluating how effective retrieval systems are. Building a retrieval system that not only returns good results to users, but does so in a way that users will want to use that system again is something more challenging; a positive search experience has been shown to lead to users engaging long-term with the retrieval system. In this talk, I will review state-of-the-art evaluation approaches for search, with respect to retrieval effectiveness but also user satisfaction. I will then focus on those approaches aiming at evaluating user engagement, and describe current works in this area within and outside the search realm. The talk will end with the proposal of a framework incorporating effectiveness evaluation into user engagement in search. An important component of this framework is to consider both within- and across-search session measurement.

# **Encodings = (Data Structures) - (Data)**

Rajeev Raman

Department of Computer Science, University of Leicester

Driven by the increasing need to analyze and search for complex patterns in very large data sets, the area of compressed and succinct data structures has grown rapidly in the last 10-15 years. Such data structures have very low memory requirements, allowing them to fit into the main memory of a computer, which in turn avoids expensive computation on hard disks.

This talk will focus on a sub-topic that has become popular recently: encoding “the data structure” itself. Some data structuring problems involve supporting queries on data, but the queries that need to be supported do not allow the original data to be deduced from the queries. This presents opportunities to obtain space savings even when the data is incompressible, by pre-processing the data, extracting only the information needed to answer the queries, and then deleting the data. The minimum information needed to answer the queries is called the effective entropy of the problem: precisely determining the effective entropy can involve interesting combinatorics.

# Contents

Faster Exact Search Using Document Clustering . . . . .	1
<i>Jonathan Dimond and Peter Sanders</i>	
Fast Online Lempel-Ziv Factorization in Compressed Space . . . . .	13
<i>Alberto Policriti and Nicola Prezza</i>	
Adaptive Computation of the Swap-Insert Correction Distance . . . . .	21
<i>Jérémy Barbay and Pablo Pérez-Lantero</i>	
Transforming XML Streams with References . . . . .	33
<i>Sebastian Maneth, Alberto Ordóñez, and Helmut Seidl</i>	
Efficient Term Set Prediction Using the Bell-Wigner Inequality . . . . .	46
<i>Massimo Melucci</i>	
On Prefix/Suffix-Square Free Words . . . . .	54
<i>Marius Dumitran, Florin Manea, and Dirk Nowotka</i>	
Temporal Analysis of CHAVE Collection . . . . .	67
<i>Olga Craveiro, Joaquim Macedo, and Henrique Madeira</i>	
DeShaTo: Describing the Shape of Cumulative Topic Distributions to Rank Retrieval Systems Without Relevance Judgments . . . . .	75
<i>Radu Tudor Ionescu, Adrian-Gabriel Chifu, and Josiane Mothe</i>	
Induced Sorting Suffixes in External Memory with Better Design and Less Space . . . . .	83
<i>Wei Jun Liu, Ge Nong, Wai Hong Chan, and Yi Wu</i>	
Efficient Algorithms for Longest Closed Factor Array . . . . .	95
<i>Hideo Bannai, Shunsuke Inenaga, Tomasz Kociumaka, Arnaud Lefebvre, Jakub Radoszewski, Wojciech Rytter, Shiho Sugimoto, and Tomasz Waleń</i>	
A Compact RDF Store Using Suffix Arrays . . . . .	103
<i>Nieves R. Brisaboa, Ana Cerdeira-Pena, Antonio Fariña, and Gonzalo Navarro</i>	
Chaining Fragments in Sequences: To Sweep or Not (Extended Abstract) . . .	116
<i>Julien Allali, Cedric Chauve, and Laetitia Bourgeade</i>	

A Faster Algorithm for Computing Maximal $\alpha$ -gapped Repeats in a String. . .	124
<i>Yuka Tanimura, Yuta Fujishige, Tomohiro I, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda</i>	
Selective Labeling and Incomplete Label Mitigation for Low-Cost Evaluation . . . . .	137
<i>Kai Hui and Klaus Berberich</i>	
Relative Select . . . . .	149
<i>Christina Boucher, Alexander Bowe, Travis Gagie, Giovanni Manzini, and Jouni Sirén</i>	
Temporal Query Classification at Different Granularities . . . . .	156
<i>Dhruv Gupta and Klaus Berberich</i>	
Prefix and Suffix Reversals on Strings. . . . .	165
<i>Guillaume Fertin, Łoïc Jankowiak, and Géraldine Jean</i>	
Filtration Algorithms for Approximate Order-Preserving Matching . . . . .	177
<i>Tamanna Chhabra, Emanuele Giaquinta, and Jorma Tarhio</i>	
Fishing in Read Collections: Memory Efficient Indexing for Sequence Assembly . . . . .	188
<i>Vladimír Boža, Jakub Jursa, Broňa Brejová, and Tomáš Vinař</i>	
How Big is that Genome? Estimating Genome Size and Coverage from $k$ -mer Abundance Spectra. . . . .	199
<i>Michal Hozza, Tomáš Vinař, and Broňa Brejová</i>	
Assessing the Efficiency of Suffix Stripping Approaches for Portuguese Stemming . . . . .	210
<i>Wadson Gomes Ferreira, Willian Antônio dos Santos, Breno Macena Pereira de Souza, Tiago Matta Machado Zaidan, and Wladimir Cardoso Brandão</i>	
Space-Efficient Detection of Unusual Words . . . . .	222
<i>Djamal Belazzougui and Fabio Cunial</i>	
Parallel Construction of Succinct Representations of Suffix Tree Topologies . . . . .	234
<i>Uwe Baier, Timo Beller, and Enno Ohlebusch</i>	
Computing the Longest Unbordered Substring . . . . .	246
<i>Paweł Gawrychowski, Gregory Kucherov, Benjamin Sach, and Tatiana Starikovskaya</i>	
Online Self-Indexed Grammar Compression . . . . .	258
<i>Yoshimasa Takabatake, Yasuo Tabei, and Hiroshi Sakamoto</i>	

Tight Bound for the Number of Distinct Palindromes in a Tree . . . . . 270  
*Paweł Gawrychowski, Tomasz Kociumaka, Wojciech Rytter,  
and Tomasz Waleń*

Beyond the Runs Theorem . . . . . 277  
*Johannes Fischer, Štěpán Holub, Tomohiro I, and Moshe Lewenstein*

Sampling the Suffix Array with Minimizers . . . . . 287  
*Szymon Grabowski and Marcin Raniszewski*

Longest Common Prefix with Mismatches . . . . . 299  
*Giovanni Manzini*

Evaluating Geographical Knowledge Re-Ranking, Linguistic Processing  
and Query Expansion Techniques for Geographical  
Information Retrieval. . . . . 311  
*Daniel Ferrés and Horacio Rodríguez*

Improved Practical Compact Dynamic Tries . . . . . 324  
*Andreas Poyias and Rajeev Raman*

ShRkC: Shard Rank Cutoff Prediction for Selective Search . . . . . 337  
*Anagha Kulkarni*

Range LCP Queries Revisited. . . . . 350  
*Amihood Amir, Moshe Lewenstein, and Sharma V. Thankachan*

Feasibility of Word Difficulty Prediction . . . . . 362  
*Ricardo Baeza-Yates, Martí Mayo-Casademont, and Luz Rello*

**Author Index** . . . . . 375