Lecture Notes in Computer Science 11147

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

Editorial Board

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

Travis Gagie · Alistair Moffat
Gonzalo Navarro · Ernesto Cuadros-Vargas (Eds.)

String Processing and Information Retrieval

25th International Symposium, SPIRE 2018
Lima, Peru, October 9–11, 2018
Proceedings

Springer

Editors
Travis Gagie
Diego Portales University
Santiago
Chile

Gonzalo Navarro
University of Chile
Santiago
Chile

Alistair Moffat
The University of Melbourne
Melbourne, VIC
Australia

Ernesto Cuadros-Vargas
Universidad de Ingeniería y Tecnología
Lima
Peru

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-030-00478-1 ISBN 978-3-030-00479-8 (eBook)
https://doi.org/10.1007/978-3-030-00479-8

Library of Congress Control Number: 2018954071

LNCS Sublibrary: SL1 – Theoretical Computer Science and General Issues

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This volume contains the papers presented at the 25th International Symposium on String Processing and Information Retrieval (SPIRE), held in Lima, Peru, October 9–11, 2018. The annual SPIRE symposium provides an opportunity for researchers to present original contributions in the three complementary areas of string processing, information retrieval, and computational biology. 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 growing emphasis on information processing. The first 24 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), Padua (Italy, 2004), Buenos Aires (Argentina, 2005), Glasgow (UK, 2006), Santiago (Chile, 2007), Melbourne (Australia, 2008), Saariselkä (Finland, 2009), Los Cabos (Mexico, 2010), Pisa (Italy, 2011), Cartagena de Indias (Colombia, 2012), Jerusalem (Israel, 2013), Ouro Preto (Brazil, 2014), London (UK, 2015), Beppu (Japan, 2016), and Palermo (Italy, 2017).

The 28 papers accepted for presentation at SPIRE 2018 were selected from 51 submissions received in response to the call for papers. Each submission was reviewed by at least three referees. After discussion, 22 full papers were accepted, as well as a further 6 short papers. The program also included three talks by invited speakers: Philip Bille, from the Technical University of Denmark; Nataša Pržulj, from University College London; and Rossano Venturini, from the Università di Pisa.

While many people helped make this conference possible, we particularly thank the members of the Program Committee and the additional reviewers who worked diligently to ensure the timely review of all submitted manuscripts. We are also grateful to the conference sponsors: Google and eBay, who each donated 5000 USD, which recompensed two of the invited speakers and sponsored ten 500 USD student travel grants; the Chilean Centro de Biotecnología y Bioingeniería (CeBiB), who contributed 2500 USD for the third invited speaker; Springer, who sponsored the 1000-euro best-paper award; and the Bioinformatics and Information Retrieval Data Structures Analysis and Design (BIRDS) project, who sponsored the colocated 13th Workshop on Compression, Text and Algorithms (WCTA) with funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sk lodowska-Curie grant agreement No. 690941. Submissions were managed and the proceedings produced using the EasyChair conference system.

August 2018

Travis Gagie
Alistair Moffat
Gonzalo Navarro
Ernesto Cuadros-Vargas

Organization

Program Committee

Diego Arroyuelo	Universidad Técnica Federico Santa María, Chile
Ricardo Baeza-Yates	NTENT, USA; Northeastern University, USA
Hideo Bannai	Kyushu University, Japan
Ilaria Bordino	UniCredit R&D, Italy
Christina Boucher	University of Florida, USA
Broňa Brejová	Comenius University in Bratislava, Slovakia
Nieves R. Brisaboa	Universidade da Coruña, Spain
Ruey-Cheng Chen	SEEK, Australia
Shane Culpepper	RMIT University, Australia
Fabio Cunial	MPI Molecular Cell Biology and Genetics, Germany
Antonio Fariña	Universidade da Coruña, Spain
David Fernández-Baca	Iowa State University, USA
Allyx Fontaine	Université de Guyane, France
Travis Gagie (Chair)	Universidad Diego Portales, Chile
Simon Gog	eBay, USA; KIT, Germany
Roberto Grossi	Università di Pisa, Italy
Inge Li Gørtz	Technical University of Denmark, Denmark
Cecilia Hernandez	Universidad de Concepción, Chile
Wing-Kai Hon	National Tsing Hua University, Taiwan
Tomohiro I	Kyushu Institute of Technology, Japan
Katharina Jahn	ETH Zürich, Switzerland
Dominik Kempa	University of Helsinki, Finland
Roberto Konow	eBay, USA
Gregory Kucherov	Université Paris-Est Marne-la-Vallée, France
Susana Ladra	Universidade da Coruña, Spain
Gad M. Landau	University of Haifa, Israel; NYU, USA
Yiqun Liu	Tsinghua University, China
Veli Mäkinen	University of Helsinki, Finland
Alistair Moffat (Chair)	The University of Melbourne, Australia
Gonzalo Navarro (Chair)	Universidad de Chile, Chile
Matthias Petri	The University of Melbourne, Australia
Cinzia Pizzi	Università di Padova, Italy
Giovanna Rosone	Università di Pisa, Italy
Leena Salmela	University of Helsinki, Finland
Diego Seco	Universidad de Concepción, Chile
Julian Shun	Massachusetts Institute of Technology, USA
Jouni Sirén	University of California, Santa Cruz, USA
Wing-Kin Sung	National University of Singapore, Singapore

Sharma Thankachan	University of Central Florida, USA
Andrew Trotman	University of Otago, New Zealand
Przemysław Uznański	ETH Zürich, Switzerland
Michal Ziv-Ukelson	Ben Gurion University of the Negev, Israel
Guido Zuccon	Queensland University of Technology, Australia

Steering Committee

Ricardo Baeza-Yates	NTENT, USA; Northeastern University, USA
Gabriele Fici	Università di Palermo, Italy
Costas Iliopoulos	King's College London, UK
Shunsuke Inenaga	Kyushu University, Japan
Simon J. Puglisi	University of Helsinki, Finland
Berthier Ribeiro-Neto	Google Inc., Brazil; Universidade Federal Minas Gerais, Brazil
Kunihiko Sadakane	University of Tokyo, Japan
Tetsuya Sakai	Waseda University, Japan
Marinella Sciortino	Università di Palermo, Italy
Rossano Venturini	Università di Pisa, Italy
Emine Yilmaz	University College London, UK
Nivio Ziviani	Universidade Federal Minas Gerais, Brazil

Additional Reviewers

Paniz Abedin	Avivit Levy
Jarno Alanko	Noa Lewenstein
Amir Carmel	Felipe A. Louza
Bastien Cazaux	Shima Moghtasedi
Panagiotis Charalampopoulos	Yuto Nakashima
Sriram Chockalingam	Takaaki Nishimoto
Francisco Claude	Alberto Ordóñez Pereira
Laxman Dhulipala	José Ramón Paramá
Gabriele Fici	Solon P. Pissis
Samah Ghazawi	Utkarsh Porwal
Adrián Gómez-Brandón	Nicola Prezza
Sahar Hooshmand	Dina Sokol
Shunsuke Inenaga	Dina Svetlitsky
Dmitry Kosolobov	Balaji Venkatachalam
Alan Kuhnle	Tomáš Vinař

Abstracts of Invited Talks

Techniques for Grammar-Based Compression

Philip Bille

Technical University of Denmark

Abstract. Grammar-based compression, where one replaces a long string by a small context-free grammar that generates the string, is a classic, simple, and powerful paradigm that captures many popular compression schemes with little or no reduction in compression rate. One of the most basic problems for grammar-based compression is to compactly represent the grammar while supporting efficient access to any character or substring without decompressing the string. The access problem naturally appears as a computational primitive in wide range of other problems for grammar-based compression such as indexing and pattern matching. Despite several recent breakthroughs and significant interest in the area many important open questions remain. In this talk we give an overview of the main techniques and results for the access problem and its variants. The talk is targeted to an audience with a general algorithmic background and we highlight the main general techniques, connections to other areas (e.g. graph decompositions and data structures), and a selection of open problems.

Mining the Integrated Connectedness of Biomedical Systems

Nataša Pržulj

University College London

Abstract. We are faced with a flood of molecular and clinical data. Various bio-molecules interact in a cell to perform biological function, forming large, complex systems. Large-scale patient-specific omics datasets are increasingly becoming available, providing heterogeneous, but complementary information about cells, tissues and diseases. The challenge is how to mine these interacting, complex, complementary data systems to answer fundamental biological and medical questions. Dealing with them is nontrivial, because many questions we ask to answer from them fall into the category of computationally intractable problems, necessitating the development of heuristic methods for finding approximate solutions.

We develop methods for extracting new biomedical knowledge from the wiring patterns of systems-level, heterogeneous, networked biomedical data. Our methods link the patterns in molecular networks and the multi-scale network organization with biological function. In this way, we translate the information hidden in the wiring patterns into domain-specific knowledge. In addition, we introduce a versatile data fusion (integration) framework that can effectively integrate the information obtained from mining molecular networks with patient-specific somatic mutation data and drug chemical data to address key challenges in precision medicine: stratification of patients, prediction of driver genes in cancer, and re-purposing of approved drugs to particular patients and patient groups. Our new methods stem from novel network science approaches coupled with graph-regularized non-negative matrix tri-factorization, a machine learning technique for dimensionality reduction and co-clustering of heterogeneous datasets. We utilize our new framework to develop methodologies for performing other related tasks, including disease re-classification from modern, heterogeneous molecular level data, inferring new Gene Ontology relationships, and aligning multiple molecular networks.

Data Compression: The Whole is Larger than the Sum of Its Parts

Rossano Venturini

Department of Computer Science, University of Pisa

Abstract. More than 70 years of research in data compression led to the design of several effective classes of compressors to deal with sequences of different types and with different characteristics. Their use in practice is widespread as encoding data to save space is of utmost importance to enable the effective exploitation of the very large datasets managed by today's systems.

Only recently, however, it has been investigated the possibility of boosting the performance of a given compressor by partitioning its input sequence. Indeed, as data compressors are very sensitive to changes of characteristics in the underlying sequence, we can achieve better results by partitioning the input sequence into homogeneous parts and compressing them separately rather than compressing the entire sequence at once.

Consider the following toy example to appreciate the benefits of this approach. We are given a sequence of n zeros followed by n ones to be compressed with arithmetic coding, the most effective entropy encoder. Encoding the whole sequence gives no compression at all as the output has size $2n$ bits. Instead, partitioning it in two halves and compressing them independently gives a compressed size of $\Theta(\log n)$ bits. An exponential improvement!

Among all the possible partitions, we are looking for an optimal one, i.e., a partition that minimizes the compressed size. Several optimization algorithms have been introduced in order to compute an optimal partition for the most important classes of compressors, e.g., zero-th and k-th order encoders [4], Burrows-Wheeler Transform-based compressors [3, 6], Lempel-Ziv '77 and '78 [1, 2, 5, 7], Elias-Fano representation [8], and so on. In this talk we will present those solutions and we will introduce the most important open problems.

References

1. Buchsbaum, A.L., Caldwell, D.F., Church, K.W., Fowler, G.S., Muthukrishnan, S.: Engineering the compression of massive tables: an experimental approach. In: Proceedings of the 11th ACM-SIAM Symposium on Discrete Algorithms, SODA 2000, pp. 175–184 (2000)
2. Buchsbaum, A., Fowler, G., Giancarlo, R.: Improving table compression with combinatorial optimization. J. ACM **50**(6), 825–851 (2003)
3. Ferragina, P., Giancarlo, R., Manzini, G., Sciortino, M.: Boosting textual compression in optimal linear time. J. ACM **52**, 688–713 (2005)

4. Ferragina, P., Nitto, I., Venturini, R.: On optimally partitioning a text to improve its compression. Algorithmica **61**(1), 51–74 (2011)
5. Ferragina, P., Nitto, I., Venturini, R.: On the bit-complexity of Lempel-Ziv compression. SIAM J.Comput. (SICOMP) **42**, 1521–1541 (2013)
6. Kärkkäinen, J., Puglisi, S.J.: Fixed block compression boosting in FM-indexes. In: Proceedings of the 18th International Symposium on String Processing and Information Retrieval, SPIRE 2011, pp. 174–184 (2011)
7. Matias, Y., Sahinalp, S.C.: On the optimality of parsing in dynamic dictionary based data compression. In: Proceedings of the Tenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 1999, pp. 943–944 (1999)
8. Ottaviano, G., Venturini, R.: Partitioned Elias-Fano indexes. In: Proceedings of the 37th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2014, pp. 273–282 (2014), (Best Paper Award)

Contents

Recoloring the Colored de Bruijn Graph 1
Bahar Alipanahi, Alan Kuhnle, and Christina Boucher

Efficient Computation of Sequence Mappability 12
Mai Alzamel, Panagiotis Charalampopoulos, Costas S. Iliopoulos, Tomasz Kociumaka, Solon P. Pissis, Jakub Radoszewski, and Juliusz Straszyński

Longest Common Prefixes with k-Errors and Applications 27
Lorraine A. K. Ayad, Carl Barton, Panagiotis Charalampopoulos, Costas S. Iliopoulos, and Solon P. Pissis

Longest Property-Preserved Common Factor. 42
Lorraine A. K. Ayad, Giulia Bernardini, Roberto Grossi, Costas S. Iliopoulos, Nadia Pisanti, Solon P. Pissis, and Giovanna Rosone

Adaptive Computation of the Discrete Fréchet Distance 50
Jérémy Barbay

Indexed Dynamic Programming to Boost Edit Distance and LCSS Computation 61
Jérémy Barbay and Andrés Olivares

Compressed Communication Complexity of Longest Common Prefixes 74
Philip Bille, Mikko Berggreen Ettienne, Roberto Grossi, Inge Li Gørtz, and Eva Rotenberg

New Structures to Solve Aggregated Queries for Trips over Public Transportation Networks 88
Nieves R. Brisaboa, Antonio Fariña, Daniil Galaktionov, Tirso V. Rodeiro, and M. Andrea Rodríguez

3DGraCT: A Grammar-Based Compressed Representation of 3D Trajectories 102
Nieves R. Brisaboa, Adrián Gómez-Brandón, Miguel A. Martínez-Prieto, and José Ramón Paramá

Towards a Compact Representation of Temporal Rasters 117
Ana Cerdeira-Pena, Guillermo de Bernardo, Antonio Fariña, José Ramón Paramá, and Fernando Silva-Coira

On Extended Special Factors of a Word. 131
Panagiotis Charalampopoulos, Maxime Crochemore, and Solon P. Pissis

Truncated DAWGs and Their Application to Minimal Absent Word Problem 139
Yuta Fujishige, Takuya Takagi, and Diptarama Hendrian

The Colored Longest Common Prefix Array Computed via Sequential Scans 153
Fabio Garofalo, Giovanna Rosone, Marinella Sciortino, and Davide Verzotto

Early Commenting Features for Emotional Reactions Prediction 168
Anastasia Giachanou, Paolo Rosso, Ida Mele, and Fabio Crestani

Block Palindromes: A New Generalization of Palindromes. 183
Keisuke Goto, I Tomohiro, Hideo Bannai, and Shunsuke Inenaga

Maximal Motif Discovery in a Sliding Window 191
Costas S. Iliopoulos, Manal Mohamed, Solon P. Pissis, and Fatima Vayani

Compressed Range Minimum Queries 206
Seungbum Jo, Shay Mozes, and Oren Weimann

Fast Wavelet Tree Construction in Practice. 218
Yusaku Kaneta

Faster Recovery of Approximate Periods over Edit Distance. 233
Tomasz Kociumaka, Jakub Radoszewski, Wojciech Rytter, Juliusz Straszyński, Tomasz Waleń, and Wiktor Zuba

Searching for a Modified Pattern in a Changing Text 241
Amihood Amir and Eitan Kondratovsky

Recovering, Counting and Enumerating Strings from Forward and Backward Suffix Arrays. 254
Yuki Kuhara, Yuto Nakashima, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda

Optimal In-Place Suffix Sorting 268
Zhize Li, Jian Li, and Hongwei Huo

Computing Burrows-Wheeler Similarity Distributions for String Collections. 285
Felipe A. Louza, Guilherme P. Telles, Simon Gog, and Liang Zhao

Better Heuristic Algorithms for the Repetition Free LCS and Other Variants 297
Radu Stefan Mincu and Alexandru Popa

Linear-Time Online Algorithm Inferring the Shortest Path from a Walk 311
Shintaro Narisada, Diptarama Hendrian, Ryo Yoshinaka, and Ayumi Shinohara

Trickier XBWT Tricks 325
Enno Ohlebusch, Stefan Stauß, and Uwe Baier

Fast and Effective Neural Networks for Translating Natural Language into Denotations 334
Tiago Pimentel, Juliano Viana, Adriano Veloso, and Nivio Ziviani

Faster and Smaller Two-Level Index for Network-Based Trajectories 348
Rodrigo Rivera, M. Andrea Rodríguez, and Diego Seco

Author Index 363