Lecture Notes in Computer Science 14899

The series Lecture Notes in Computer Science (LNCS), including its subseries Lecture Notes in Artificial Intelligence (LNAI) and Lecture Notes in Bioinformatics (LNBI), has established itself as a medium for the publication of new developments in computer science and information technology research, teaching, and education.

LNCS enjoys close cooperation with the computer science R & D community, the series counts many renowned academics among its volume editors and paper authors, and collaborates with prestigious societies. Its mission is to serve this international community by providing an invaluable service, mainly focused on the publication of conference and workshop proceedings and postproceedings. LNCS commenced publication in 1973.

Zsuzsanna Lipták · Edleno Moura ·
Karina Figueroa · Ricardo Baeza-Yates
Editors

String Processing and Information Retrieval

31st International Symposium, SPIRE 2024
Puerto Vallarta, Mexico, September 23–25, 2024
Proceedings

Springer

Editors
Zsuzsanna Lipták
University of Verona
Verona, Italy

Karina Figueroa
Michoacan University of Saint Nicholas
of Hidalgo
Morelia, Mexico

Edleno Moura
Federal University of Amazonas
Manaus, Brazil

Ricardo Baeza-Yates
Northeastern University
Boston, MA, USA

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-031-72199-1 ISBN 978-3-031-72200-4 (eBook)
https://doi.org/10.1007/978-3-031-72200-4

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, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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

If disposing of this product, please recycle the paper.

Preface

The 31st International Symposium on String Processing and Information Retrieval (SPIRE) was held on September 23–25, 2024, in Puerto Vallarta (Mexico), followed by the 18th Workshop on Compression, Text, and Algorithms (WCTA) held on September 26, 2024.

SPIRE started in 1993 as the South American Workshop on String Processing. It was held in Latin America until 2000. Then, SPIRE moved to Europe, and from then on, it has been held in Australia, Japan, the UK, Spain, Italy, Finland, Portugal, Israel, Brazil, Chile, Colombia, Mexico, Argentina, Bolivia, Peru, the USA, and France. SPIRE continues the long and well-established tradition of encouraging high-quality research at the broad nexus of string processing, information retrieval, and computational biology.

This volume contains the accepted papers presented at SPIRE 2024. SPIRE 2024 received a total of 41 submissions, 34 full papers and 7 short papers. Each submission received at least three single-blind reviews. After the discussion phase, the Scientific Program Committee accepted 22 full papers and 4 short papers. We thank all the authors for their valuable contributions and presentations at the conference and thank the Program Committee members and additional reviewers for their valuable work during the review and discussion phases. We also thank the members of the Local Organizing Committee for their support in organizing SPIRE.

We appreciate the high-quality talks included in the scientific program from three renowned researchers: Juliana Freire (New York University, USA), Marinella Sciortino (University of Palermo, Italy), and Gerardo Sierra (National Autonomous University of Mexico, Mexico). This edition also had a Best Paper Award, sponsored by Springer. The award was announced during the conference.

We thank our sponsors: ACM SIGIR, Web4Good, Springer, Dipartimento di Informatica of Università di Verona, and Universidad Michoacana de San Nicolás de Hidalgo. Their generous support has been instrumental in making this conference a reality, fostering academic excellence and enabling us to bring together a diverse group of researchers and students. Finally, we thank Springer for publishing the proceedings of SPIRE 2024 in the LNCS series.

August 2024

Zsuzsanna Lipták
Edleno Moura
Karina Figueroa
Ricardo Baeza-Yates

Organization

General Chairs

Karina Figueroa-Mora	Universidad Michoacana, Mexico
Ricardo Baeza-Yates	Northeastern University, USA, & University of Chile, Chile

Program Committee Chairs

Zsuzsanna Lipták	University of Verona, Italy
Edleno S. de Moura	Federal University of Amazonas and Jusbrasil, Brazil

Steering Committee

Diego Arroyuelo	Pontificia Universidad Católica de Chile and Millennium Institute for Foundational Research on Data, Chile
Ricardo Baeza-Yates	Northeastern University, USA & University of Chile, Chile
Thierry Lecroq	University of Rouen Normandy, France
Franco Maria Nardini	ISTI-CNR, Pisa, Italy
Nadia Pisanti	University of Pisa, Italy
Barbara Poblete	University of Chile and Amazon, Chile
Berthier Ribeiro-Neto	Google Inc. and Federal University of Minas Gerais, Brazil
Hélène Touzet	CNRS Lille, France
Rossano Venturini	University of Pisa, Italy
Nivio Ziviani	Universidade Federal Minas Gerais, Brazil

Program Committee

Omar Alonso	Amazon, USA
Diego Arroyuelo	Pontificia Universidad Católica de Chile and Millennium Institute for Foundational Research on Data, Chile

Golnaz Badkobeh	University of London, UK
Djamal Belazzougui	CERIST (Research Centre for Scientific and Technical Information), Algeria
Giulia Bernardini	Università di Trieste, Italy
Marilia Braga	Bielefeld University, Germany
Laurent Bulteau	CNRS - Université Gustave Eiffel, France
Edgar Chavez	CICESE, Mexico
Manuel Cáceres	Aalto University, Finland
Edleno S. De Moura (Co-chair)	Federal University of Amazonas and Jusbrasil, Brazil
Nadia El-Mabrouk	University of Montreal, Canada
Jonas Ellert	ENS - PSL, France
Antonio Fariña	University of A Coruña, Spain
Paweł Gawrychowski	University of Wrocław, Poland
Daniel Gibney	University of Texas at Dallas, USA
Inge Li Gørtz	Technical University of Denmark, Denmark
Meng He	Dalhousie University, Canada
Katharina T. Huber	University of East Anglia, UK
Tomohiro I	Kyushu Institute of Technology, Japan
Tomász Kociumaka	Max Planck Institute for Informatics, Germany
M. Oguzhan Kulekci	Indiana University Bloomington, USA
Dominik Köppl	University of Yamanashi, Japan
Susana Ladra	University of A Coruña, Spain
Moshe Lewenstein	Bar Ilan University, Israel
Zsuzsanna Lipták (Co-chair)	University of Verona, Italy
Felipe A. Louza	Universidade Federal de Uberlândia, Brazil
Camille Marchet	CNRS - Université de Lille, France
Viviane P. Moreira	Instituto de Informatica - UFRGS, Brazil
Nadia Pisanti	University of Pisa, Italy
Cinzia Pizzi	University of Padua, Italy
Svetlana Puzynina	Saint Petersburg State University, Russia
Narad Rampersad	University of Winnipeg, Canada
Kunihiko Sadakane	University of Tokyo, Japan
Leena Salmela	University of Helsinki, Finland
Blerina Sinaimeri	LUISS University of Rome, Italy
Jouni Sirén	University of California, Santa Cruz, USA
Tatiana Starikovskaya	École Normale Supérieure, France
Rossano Venturini	Università di Pisa, Italy
Aaron Williams	Williams College, USA
Michal Ziv-Ukelson	Ben-Gurion University of the Negev, Israel

Additional Reviewers

Aksenov, Vitaly
Amadini, Roberto
Ascone, Rocco
Benslimane, Seddik
Carmel, Amir
Cenzato, Davide
Chakraborty, Sankardeep
Cordeiro, Lucas
Cotumaccio, Nicola
Deharbe, David
Delabre, Mattéo
Fici, Gabriele
Ganardi, Moses
Ganguly, Arnab
Gascon, Mathieu
Ghazawi, Samah
Gómez-Brandón, Adrián
Hossen, Md Helal
Maity, Anuran
Martayan, Igor
Mhaskar, Neerja
Mäkinen, Veli
Nishimoto, Takaaki
Olbrich, Jannik
Parmigiani, Luca
Pirola, Yuri
Prezza, Nicola
Radoszewski, Jakub
Saad, Daniel
Steiner, Teresa Anna
Stoye, Jens
Tiskin, Alexander
Waleń, Tomasz
Wu, Kaiyu
Zuba, Wiktor

Abstracts of Invited Talks

Dataset Search for Data Discovery, Augmentation, and Explanation

Juliana Freire

New York University
juliana.freire@nyu.edu

Abstract. In recent years, we have witnessed an explosion in our capacity to collect and catalog vast amounts of data about our environment, society, and populace. Moreover, with the push towards transparency and open data, scientists, governments, and organizations are increasingly making structured data available on the Web and in various repositories and data lakes. Combined with advances in analytics and machine learning, the availability of such data should, in theory, allow us to make progress on many of our most important scientific and societal questions.

However, this opportunity is often unrealized due to a central technical barrier: it is remains nearly impossible for domain experts to sift through the overwhelming amount of available information to discover datasets they need for their specific applications. While search engines have addressed the discovery problem for Web documents, supporting the discovery of structured data presents new challenges. These include crawling the Web in search of datasets, indexing datasets and supporting dataset-oriented queries, and creating new techniques to rank and display results.

In this talk, I will discuss these challenges and present our recent work in this area. Specifically, I will describe strategies for finding relevant datasets on the web and deriving metadata to be indexed. Additionally, I will introduce a new class of data-relationship queries and outline a collection of methods that efficiently support various types of relationships, demonstrating how they can be used for data explanation and augmentation. Finally, I will showcase Auctus, an open-source dataset search engine that we have developed at the NYU Visualization, Imaging, and Data Analysis (VIDA) Center. I will conclude by highlighting open problems and suggesting directions for future research.

Exploring Repetitiveness in Texts: From BWT to Morphisms

Marinella Sciortino

Dipartimento di Matematica e Informatica, University of Palermo, Italy
marinella.sciortino@unipa.it

Abstract. The notion of repetitiveness plays a fundamental role in processing very large collections of texts. In many applications, massive and highly repetitive data need to be stored, analyzed, and queried. Therefore, having good measures capable of capturing repetitiveness implies having effective parameters to evaluate the performance of compressed indexing data structures for such types of data.

Many repetitiveness measures are defined using compression schemes. One of these measures, denoted *r*, is the number of maximal equal-letter runs in the output produced by the Burrows-Wheeler Transform (BWT), a transformation which permutes the characters of a text to boost the effects of run-length encoding. Besides having a crucial role in the definition of recent compressed indexing data structures, such as the *r*-index, the measure *r* has attracted attention in Combinatorics on Words because it has allowed for defining and recognizing properties of repetitive strings. A pioneering result is the characterization of finite Sturmian words as the binary strings for which *r* assumes its minimum value.

From a complementary perspective, morphisms are classic tools in Combinatorics on Words for generating collections of repetitive texts. Injective morphisms, known as codes, are widely used in Information Theory. Recently, morphisms, combined with copy-paste mechanisms, have been used to define new repetitiveness measures and compressors, called NU-systems.

In this talk, I will explore our recent results on the properties of the measure *r* that allow analysis of the combinatorial characteristics of input texts. I will then show very recent interesting findings on the identification of collections of generic highly repetitive strings using the measure *r*. Next, we will see recent results on the evaluation of some compression-based repetitiveness measures for collections of strings generated by morphisms. I will close with our latest research on the close correlations between morphisms and the measure *r*, with exciting implications in the theory of codes.

Preservation and Accessibility of Documentary Heritage

Gerardo Sierra

National Autonomous University of Mexico (UNAM), Mexico

Abstract. Preservation and accessibility of documentary heritage are essential for maintaining and disseminating the cultural and historical wealth of a society. These concepts encompass a set of actions and strategies aimed at conserving historical documents and ensuring their availability for future generations, fostering research and knowledge across various disciplines.

National libraries play a crucial role as the primary reservoir of a country's documentary heritage. They store and protect a vast collection of documents, both printed and digital, that reflect a nation's cultural diversity and legacy.

Printed documents include codices, manuscripts, documents in indigenous languages, and multimodal texts. Each type presents unique preservation challenges due to its fragility, rarity, and linguistic and material diversity. The preservation of printed documents faces several challenges, such as the need for specialized techniques for physical conservation, the digitization of multimodal texts, and the translation and cataloging of documents in indigenous languages. These tasks require an interdisciplinary approach and advanced technologies to ensure the integrity and accessibility of these materials.

Natural language processing (NLP) and artificial intelligence (AI) offer powerful tools to address these challenges. These technologies can support, among others: Metadata extraction, cataloging and classification, and summary generation.

The use of NLP and AI not only enhances preservation but also increases the accessibility of documentary heritage. These technologies enable the creation of digital access platforms, vectorized databases, and advanced search tools, which are essential for research in digital humanities, stylometry, literary studies, and more.

Contents