Lecture Notes in Computer Science 11811

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

Nieves R. Brisaboa · Simon J. Puglisi (Eds.)

String Processing and Information Retrieval

26th International Symposium, SPIRE 2019
Segovia, Spain, October 7–9, 2019
Proceedings

Springer

Editors
Nieves R. Brisaboa
University of A Coruña
A Coruña, Spain

Simon J. Puglisi
University of Helsinki
Helsinki, Finland

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-030-32685-2 ISBN 978-3-030-32686-9 (eBook)
https://doi.org/10.1007/978-3-030-32686-9

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

SPIRE 2019, held October 7–9, 2019, in Segovia, Spain, was the 26th International Symposium on String Processing and Information Retrieval. SPIRE started in 1993 as the South American Workshop on String Processing, therefore it was held in Latin America until 2000, when SPIRE traveled to Europe. From then on, SPIRE meetings have been held in Australia, Japan, UK, Spain, Italy, Finland, Portugal, Israel, Brazil, Chile, Colombia, Mexico, Argentina, Bolivia, and Peru.

In this edition, again in Spain, we continued the long and well-established tradition of encouraging high-quality research at the broad nexus of algorithms and data structures for sequences and graphs, data compression, databases, data mining, information retrieval, and computational biology. As usual, SPIRE 2019 continues to provide an opportunity to bring together specialists and young researchers working in these areas.

This volume contains the 36 papers, out of a total of 59 submissions accepted to be presented in SPIRE 2019. Each submission received at least three reviews. Authors of accepted papers come from 17 countries, across five continents (Africa, Asia, Europe, North America, South America). We thank all authors who submitted their work for consideration to SPIRE 2019 and we especially thank the Program Committee and the external reviewers, whose many thorough reviews helped us select the papers presented. The success of the scientific program is due to their hard work.

Besides the 36 accepted papers, the scientific program included three invited lectures, given by:

- Veli Mäkinen on "When Stringology Meets Graphs"
- Alistair Moffat on "User-Based Evaluation in Information Retrieval"
- Gonzalo Navarro on "Repetitiveness and Indexability"

We thank the invited speakers for accepting our invitation and for their excellent presentations at the conference.

To complete the event, this year for the fourth year running, SPIRE 2019 had a Best Paper Award, sponsored by Springer that was announced during the conference. Besides Springer, we thank the EU project BIRDS (H2020-MSCA-RISE-2015 GA No 690941) for its financial support and the ICT Research Center CITIC at the University of A Coruña and the Segovia Campus of the University of Valladolid whose administrative and financial support we gratefully acknowledge.

October 2019

Nieves R. Brisaboa
Simon J. Puglisi

Organization

Program Committee

Amihood Amir	Bar-Ilan University, Israel and Johns Hopkins University, USA
Ricardo Baeza-Yates	Northeastern University, USA and Pompeu Fabra University, Spain
Hideo Bannai	Kyushu University, Japan
Christina Boucher	University of Florida, USA
Nieves Brisaboa	University of A Coruña, Spain
Antonio Fariña	University of A Coruña, Spain
Johannes Fischer	TU Dortmund, Germany
Jose Fuentes	University of Chile, Chile
Travis Gagie	Dalhousie University, Canada
Pawel Gawrychowski	University of Wroclaw, Poland
Simon Gog	eBay Inc., USA
Inge Li Gørtz	Technical University of Denmark, Denmark
Susana Ladra	University of A Coruña, Spain
Zsuzsanna Liptak	University of Verona, Italy
Miguel A. Martinez-Prieto	University of Valladolid, Spain
Jose R. Parama	University of A Coruña, Spain
Kunsoo Park	Seoul National University, Republic of Korea
Matthias Petri	The University of Melbourne, Australia
Solon Pissis	CWI, The Netherlands
Simon Puglisi	University of Helsinki
Marinella Sciortino	University of Palermo, Italy
Diego Seco	University of Concepción, Chile
Jouni Sirén	University of California at Santa Cruz, USA
Yasuo Tabei	RIKEN Center for Advanced Intelligence Project, Japan
Rossano Venturini	University of Pisa, Italy
Nivio Ziviani	Federal University of Minas Gerais, Brazil

Additional Reviewers

Abedin, Paniz
Ahanonu, Eze
Arroyuelo, Diego
Ayad, Lorraine
Bernardini, Giulia
Bille, Philip
Boneh, Itai
Brandão, Wladmir
Calvo-Zaragoza, Jorge
Cazaux, Bastien
Charalampopoulos, Panagiotis
Clifford, Raphael
De Sensi, Daniele
Epifanio, Chiara
Fici, Gabriele
Freire Castro, Borja
Fujishige, Yuta
Galaktionov, Daniil
Gańczorz, Michał
Glowacka, Dorota
Goto, Keisuke
Gómez-Brandón, Adrián
Holland, William
Inenaga, Shunsuke
Janczewski, Wojciech
Kanda, Shunsuke
Kida, Takuya
Klein, Shmuel Tomi
Kolesnikov, Vladimir
Komusiewicz, Christian
Kondratovsky, Eitan
Köppl, Dominik
Levy, Avivit
MacKenzie, Joel
Mieno, Takuya
Moffat, Alistair
Mukherjee, Kingshuk
Nakashima, Yuto
Nekrich, Yakov
Nishimoto, Takaaki
Ochoa, Carlos
Pibiri, Giulio Ermanno
Piatkowski, Marcin
Pokorski, Karol
Prezza, Nicola
Radoszewski, Jakub
Raman, Rajeev
Rossi, Massimiliano
Salmela, Leena
Schmid, Markus L.
Shalom, Riva
Silvestre Vilches, Jorge
Thankachan, Sharma V.
Trotman, Andrew
Turpin, Andrew
Valle, Dan
Zhukova, Bella
Zobel, Justin
Zuin, Gianlucca

Abstracts of Invited Talks

Repetitiveness and Indexability

Gonzalo Navarro

CeBiB—Center for Biotechnology and Bioengineering,
IMFD—Millennium Institute for Foundational Research on Data,
Department of Computer Science, University of Chile, Chile
gnavarro@dcc.uchile.cl

Abstract. Compressed indexes for highly repetitive text collections can reduce the data size by orders of magnitude while still supporting efficient searches. Compression of this kind of data requires dictionary-based methods, because statistical compression fails to capture repetitiveness. Unlike statistical compression, where the state of the art is mature and indexes reaching entropy size are already several years old, there is not even a clear concept of entropy for highly repetitive collections. There is a wealth of measures, some more ad-hoc and some more principled. Some relations are known between them, other relations are unknown. It is known that no compressor can reach some measures, it is known how to reach others, and for some it is unknown whether this is possible. From the reachable ones, some allow random access to the compressed text, for others it is unknown how to do it. Finally, some admit indexed searches, for others we do not know if this is possible. In this talk I will survey this zoo of measures, show their properties and known relations, show what is known and unknown about them, and point out several open questions that relate repetitiveness with indexability.

Keywords: Repetitive text collections · Compressed text indexing · Entropy

Partially supported by Basal Funds B0001, Conicyt, Chile and by the Millennium Institute for Foundational Research on Data, Mideplan, Chile.

C/W/L Spells "Cool": User-Based Evaluation in Information Retrieval

Alistair Moffat

The University of Melbourne, Australia

Abstract. The Information Retrieval community pride themselves on the strength of their evaluation protocols: working with large test collections; executing dozens or hundreds of queries taken to be representative of typical information requirements; and, in many cases, employing expert assessors to form relevance judgments. System scores using these resources are then computed using an effectiveness metric such as precision at depth k, expected reciprocal rank, or average precision; and champion-versus-challenger evaluations are carried out by considering the two system means through the lens of a statistical significance test.

This presentation focuses on the effectiveness metrics that are at the heart of this batch evaluation pipeline. After describing a range of traditional approaches to measuring effectiveness, the "C/W/L" framework [2, 3] is motivated and defined, and a range of implications of this approach to IR evaluation then explored. Notable in the C/W/L structure is the explicit correspondence between metrics and user models. This relationship makes it possible for metrics to be evaluated and compared in terms of their suitability for different types of search task, based on the extent to which the user model associated with each candidate metric correlates with observed user behavior when performing that task [1, 4, 5]. Measurement accuracy is also considered for C/W/L metrics, together with the implications that certain types of user behavior then have on experimental design.

Keywords: Information retrieval evaluation · Web search · User model · Effectiveness metric

Acknowledgment. The work presented in this talk was carried out in collaboration with Peter Bailey, Falk Scholer, Paul Thomas, Alfan Wicaksono, and Justin Zobel. Their various contributions are gratefully acknowledged.

References

1. Azzopardi, L., Thomas, P., Craswell, N.: Measuring the utility of search engine result pages. In: Proceedings of SIGIR, pp. 605–614 (2018)
2. Moffat, A., Bailey, P., Scholer, F., Thomas, P.: Incorporating user expectations and behavior into the measurement of search effectiveness. ACM Trans. Inf. Sys. **35**(3), 24:1–24:38 (2017)

3. Moffat, A., Thomas, P., Scholer, F.: Users versus models: what observation tells us about effectiveness metrics. In: Proceedings of CIKM, pp. 659–668 (2013)
4. Wicaksono, A.F., Moffat, A.: Empirical evidence for search effectiveness models. In: Proceedings of CIKM, pp. 1571–1574 (2018)
5. Wicaksono, A.F., Moffat, A., Zobel, J.: Modeling user actions in job search. In: Azzopardi, L., Stein, B., Fuhr, N., Mayr, P., Hauff, C., Hiemstra, D. (eds.) Advances in Information Retrieval. ECIR 2019. LNCS, vol. 11437, pp. 652–664. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-15712-8_42

When Stringology Meets Graphs

Veli Mäkinen

Department of Computer Science, University of Helsinki, Finland
veli.makinen@helsinki.fi

Abstract. Consider a directed acyclic graph (DAG) G with nodes labelled with characters. We say that a string pattern P occurs in G if there is a path spelling P. When G is deterministic, that is, no node has two edges leading to nodes with the same character label, there is a trivial algorithm to locate P in G: Start at all places and check if path spelling P exists. This trivial algorithm turns out to be optimal under the Strong Exponential Time Hypothesis (SETH). The talk starts by explaining this result by Equi, Grossi, Mäkinen, and Tomescu (*ICALP* 2019).

Quadratic running time for matching pattern P against graph G can slightly be improved without violating SETH, by using bit-parallelism. The talk discusses extensions of Shift-And and Myers' algorithms for exact and approximate pattern matching on graphs as studied by Rautiainen, Mäkinen, and Marschall (*Bioinformatics*, to appear).

Sparse dynamic programming is another technique that can evade the quadratic bound, assuming a sub-quadratic size set of anchors is given as input to limit the alignment options. An anchor defines a plausible alignment of a substring against a subpath. An ordered subset of anchors forms a co-linear chain if the corresponding substrings are in linear order in P and the corresponding subpaths are in linear order in some path of G. Consider the problem of finding a co-linear chain that maximally covers P. This problem is studied by Mäkinen, Tomescu, Kuosmanen, Paavilainen, Gagie, and Chikhi (*ACM Transactions on Algorithms*, 2019), who give an algorithm whose running time depends on the number of paths needed to cover G; the algorithm is optimal once G is just a string. The talk covers the main insights of this algorithm.

The talk concludes with another alignment problem related to path covers. Consider two DAGs G_1 and G_2 each of which is coverable by at most two paths. Such DAGs can be seen as simplest extension of strings into graphs and are also representing diploid genomes. A covering alignment asks for path covers *(A, B)* and *(C, D)* of G_1 and G_2, respectively, that minimize the sum of edit distance between A and C and between B and D. Covering alignment turns out to be NP-hard as shown by Rizzi, Cairo, Mäkinen, Tomescu, and Valenzuela (*IEEE/ACM Transactions on Computational Biology and Bioinformatics*). The talk gives an overview of the reduction.

Keywords: String matching · Graphs · SETH · Bit-parallelism · Sparse dynamic programming · Covering alignment

Partially supported by the Academy of Finland (grant 309048).

Contents

Indexing and Compression

Compressed Data Structures