Lecture Notes in Computer Science **9954**

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

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

Shunsuke Inenaga · Kunihiko Sadakane
Tetsuya Sakai (Eds.)

String Processing and Information Retrieval

23rd International Symposium, SPIRE 2016
Beppu, Japan, October 18–20, 2016
Proceedings

Springer

Editors
Shunsuke Inenaga
Informatics
Kyushu University
Fukuoka
Japan

Tetsuya Sakai
Computer Science and Engineering
Waseda University
Tokyo
Japan

Kunihiko Sadakane
Mathematical Informatics
University of Tokyo
Tokyo
Japan

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-319-46048-2 ISBN 978-3-319-46049-9 (eBook)
DOI 10.1007/978-3-319-46049-9

Library of Congress Control Number: 2016950414

LNCS Sublibrary: SL1 – Theoretical Computer Science and General Issues

Printed on acid-free paper

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

Preface

This volume contains the papers presented at SPIRE 2016, the 23rd International Symposium on String Processing and Information Retrieval, held October 18–20, 2016 in Beppu, Japan. Following the tradition from previous years, the focus of SPIRE this year was on fundamental studies on string processing and information retrieval, as well as application areas such as bioinformatics, Web mining, and so on.

The call for papers resulted in 46 submissions. Each submitted paper was reviewed by at least three Program Committee members. Based on the thorough reviews and discussions by the Program Committee members and additional subreviewers, the Program Committee decided to accept 25 papers.

The main conference featured three keynote speeches by Kunsoo Park (Seoul National University), Koji Tsuda (University of Tokyo), and David Hawking (Microsoft & Australian National University), together with presentations by authors of the 25 accepted papers. Prior to the main conference, two satellite workshops were held: String Masters in Fukuoka, held October 12–14, 2016 in Fukuoka, and the 11th Workshop on Compression, Text, and Algorithms (WCTA 2016), held on October 17, 2016 in Beppu. String Masters was coordinated by Hideo Bannai, and WCTA was coordinated by Simon J. Puglisi and Yasuo Tabei. WCTA this year featured two keynote speeches by Juha Kärkkäinen (University of Helsinki) and Yoshitaka Yamamoto (University of Yamanashi).

We would like to thank the SPIRE Steering Committee for giving us the opportunity to host this wonderful event. Also, many thanks go to the Program Committee members and the additional subreviewers, for their valuable contribution ensuring the high quality of this conference. We appreciate Springer for their professional publishing work and for sponsoring the Best Paper Award for SPIRE 2016. We finally thank the Local Organizing Team (led by Hideo Bannai) for their effort to run the event smoothly.

October 2016

Shunsuke Inenaga
Kunihiko Sadakane
Tetsuya Sakai

Organization

Program Committee

Leif Azzopardi	University of Glasgow, UK
Philip Bille	Technical University of Denmark, Denmark
Praveen Chandar	University of Delware, USA
Raphael Clifford	University of Bristol, UK
Shane Culpepper	RMIT University, Australia
Zhicheng Dou	Renmin University of China, China
Hui Fang	University of Delaware, USA
Simone Faro	University of Catania, Italy
Johannes Fischer	TU Dortmund, Germany
Sumio Fujita	Yahoo! Japan Research, Japan
Travis Gagie	University of Helsinki, Finland
Pawel Gawrychowski	University of Wroclaw, Poland and University of Haifa, Israel
Simon Gog	Karslruhe Institute of Technology, Germany
Roberto Grossi	Università di Pisa, Italy
Ankur Gupta	Butler University, USA
Wing-Kai Hon	National Tsing Hua University, Taiwan
Shunsuke Inenaga	Kyushu University, Japan
Makoto P. Kato	Kyoto University, Japan
Gregory Kucherov	CNRS/LIGM, France
Moshe Lewenstein	Bar Ilan University, Israel
Yiqun Liu	Tsinghua University, China
Mihai Lupu	Vienna University of Technology, Austria
Florin Manea	Christian-Albrechts-Universität zu Kiel, Germany
Gonzalo Navarro	University of Chile, Chile
Yakov Nekrich	University of Waterloo, Canada
Tadashi Nomoto	National Institute of Japanese Literature, Japan
Iadh Ounis	University of Glasgow, UK
Simon Puglisi	University of Helsinki, Finland
Kunihiko Sadakane	University of Tokyo, Japan
Tetsuya Sakai	Waseda University, Japan
Hiroshi Sakamoto	Kyushu Institute of Technology, Japan
Leena Salmela	University of Helsinki, Finland
Srinivasa Rao Satti	Seoul National University, South Korea
Ruihua Song	Microsoft Research Asia, China
Young-In Song	Wider Planet, South Korea
Kazunari Sugiyama	National University of Singapore, Singapore

Aixin Sun	Nanyang Technological University, Singapore
Wing-Kin Sung	National University of Singapore, Singapore
Julián Urbano	University Carlos III of Madrid, Spain
Sebastiano Vigna	Università degli Studi di Milano, Italy
Takehiro Yamamoto	Kyoto University, Japan

Additional Reviewers

Bingmann, Timo
Bouvel, Mathilde
Chikhi, Rayan
Cicalese, Ferdinando
Conte, Alessio
Farach-Colton, Martin
Fici, Gabriele
Fontaine, Allyx
Frith, Martin
Ganguly, Arnab
I, Tomohiro
Jo, Seungbum
Kempa, Dominik
Kosolobov, Dmitry
Lee, Joo-Young
Liu, Xitong
Mercas, Robert
Ordóñez Pereira, Alberto
Pisanti, Nadia
Rosone, Giovanna
Schmid, Markus L.
Starikovskaya, Tatiana
Thankachan, Sharma V.
Välimäki, Niko

Keynote Speeches

Indexes for Highly Similar Sequences

Kunsoo Park

Department of Computer Science and Engineering, Seoul National University,
Seoul, South Korea
kpark@theory.snu.ac.kr

The 1000 Genomes Project aims at building a database of a thousand individual human genome sequences using a cheap and fast sequencing, called next generation sequencing, and the sequencing of 1092 genomes was announced in 2012. To sequence an individual genome using the next generation sequencing, the individual genome is divided into short segments called reads and they are aligned to the human reference genome. This is possible because an individual genome is more than 99 % identical to the reference genome. This similarity also enables us to store individual genome sequences efficiently.

Recently many indexes have been developed which not only store highly similar sequences efficiently but also support efficient pattern search. To exploit the similarity of the given sequences, most of these indexes use classical compression schemes such as run-length encoding and Lempel-Ziv compression.

We introduce a new index for highly similar sequences, called FM index of alignment. We start by finding common regions and non-common regions of highly similar sequences. We need not find a multiple alignment of non-common regions. Finding common and non-common regions is much easier and simpler than finding a multiple alignment, especially in the next generation sequencing. Then we make a transformed alignment of the given sequences, where gaps in a non-common region are put together into one gap. We define a suffix array of alignment on the transformed alignment, and the FM index of alignment is an FM index of this suffix array of alignment. The FM index of alignment supports the LF mapping and backward search, the key functionalities of the FM index. The FM index of alignment takes less space than other indexes and its pattern search is also fast.

This research was supported by the Bio & Medical Technology Development Program of the NRF funded by the Korean government, MSIP (NRF-2014M3C9A3063541).

Simulation in Information Retrieval: With Particular Reference to Simulation of Test Collections

David Hawking

Microsoft, Canberra, Australia
david.hawking@acm.org

Keywords: Information retrieval · Simulation · Modeling

Simulation has a long history in the field of Information Retrieval. More than 50 years ago, contractors for the US Office of Naval Research (ONR) were working on simulating information storage and retrieval systems.[1]

The purpose of simulation is to predict the behaviour of a system over time, or under conditions in which a real system can't easily be observed. My talk will review four general areas of simulation activity. First is the simulation of entire information retrieval systems, as for example exemplified by Blunt (1965):

> A general time-flow model has been developed that enables a systems engineer to simulate the interactions among personnel, equipment and data at each step in an information processing effort.

and later by Cahoon and McKinley (1996).

A second area is the simulation of behaviour when a person interacts with an information retrieval service, with particular interest in multi-turn interactions. For example user simulation has been used to study implicit feedback systems (White et al., 2004), PubMed browsing strategies (Lin and Smucker, 2007), and query suggestion algorithms (Jiang and He, 2013).

A third area has been little studied – simulating an information retrieval service (in the manner of Kemelen's 1770 Automaton Chess Player) in order to study the behaviour of real users when confronted with a retrieval service which hasn't yet been built.

The final area is that of simulation of test collections. It is an area in which I have been working recently, with my colleagues Bodo Billerbeck, Paul Thomas and Nick Craswell. My talk will include some preliminary results.

As early as 1973, Michael Cooper published a method for generating artificial documents and queries in order to, "evaluate the effect of changes in characteristics of the query and document files on the quantity of material retrieved." More recently, Azzopardi and de Rijke (2006) have studied the automated creation of known-item test collections.

[1] "System" used in the Systems Theory sense.

Organizations like Microsoft have a need to develop, tune and experiment with information retrieval services using simulated versions of private or confidential data. Furthermore, there may be a need to predict the performance of a retrieval service when an existing data set is scaled up or altered in some way.

We have been studying how to simulate text corpora and query sets for such purposes. We have studied many different corpora with a wide range of different characteristics. Some of the corpora are readily available to other researchers; others we are unable to share. With accurate simulation models we may be able to share sufficient characteristics of those data sets to enable others to reproduce our results.

The models underpinning our simulations include:

1. Models of the distribution of document lengths.
2. Models of the distribution of word frequencies. (Revisiting Zipf's law.)
3. Models of term dependence.
4. Models of the representation of indexable words.
5. Models of how these change as the corpus grows. (e.g. revisiting the models due to Herdan and Heaps.)

We have implemented a document generator based on these models and software for estimating model parameters from a real corpus. We test the models by running the generator with extracted parameters and comparing various properties of the resulting corpus with those of the original. In addition, we test the growth model by extracting parameters from 1 % samples and simulating a corpus 100 times larger. In early experimentation we have found reasonable agreement between the properties of the real corpus and its scaled-up emulation.

The value gained from a simulation approach depends heavily on the accuracy of the system model, but a highly accurate model may be very complex and may be over-fitted to the extent that it doesn't generalise. We study what is required to achieve high fidelity but also discuss simpler forms of model which may be sufficiently accurate for less demanding requirements.

References

1. Blunt, C.R.: An information retrieval system model. Report of Contract Nonr. 3818(00), ONR (1965). http://www.dtic.mil/dtic/tr/fulltext/u2/623590.pdf
2. Cooper, M.D.: A simulation model of a retrieval system. Inf. Storage Retrieval **9**, 13–32 (1973)

Significant Pattern Mining: Efficient Algorithms and Biomedical Applications

Koji Tsuda

Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Japan

Pattern mining techniques such as itemset mining, sequence mining and graph mining have been applied to a wide range of datasets. To convince biomedical researchers, however, it is necessary to show statistical significance of obtained patterns to prove that the patterns are not likely to emerge from random data. The key concept of significance testing is family-wise error rate, i.e., the probability of at least one pattern is falsely discovered under null hypotheses. In the worst case, FWER grows linearly to the number of all possible patterns. We show that, in reality, FWER grows much slower than the worst case, and it is possible to find significant patterns in biomedical data. The following two properties are exploited to accurately bound FWER and compute small p-value correction factors. (1) Only closed patterns need to be counted. (2) Patterns of low support can be ignored, where the support threshold depends on the Tarone bound. We introduce efficient depth-first search algorithms for discovering all significant patterns and discuss about parallel implementations.

Contents