Journal of Discrete Algorithms 18 (2013) 1–2

Contents lists available at SciVerse ScienceDirect

Journal of Discrete Algorithms

www.elsevier.com/locate/jda

Editorial

In the 18 years since its first edition back in 1993 the International Symposium on String Processing and Information Retrieval (SPIRE) has become the reference meeting for an interdisciplinary community of researchers whose activity lies at the intersection between string processing and information retrieval. SPIRE 2011, the 18th conference in this series, was held in Pisa, Italy, on October 17–21, 2011. The event has been held under this title annually since 1998. The four first events concentrated mainly on string processing (SP), and were held in South America under the title South American Workshop on String Processing (WSP), in 1993 (Belo Horizonte, Brazil); 1995 (Valparaiso, Chile); 1996 (Recife, Brazil); and 1997 (Valparaiso, Chile). WSP was transformed into SPIRE in 1998 (Santa Cruz, Bolivia), when the scope of the event was broadened to include also information retrieval (IR). The change was motivated by the increasing relevance of information retrieval and its close interrelationship with the general area of string processing. From 1999 to 2007, the venue of SPIRE has been alternating between South/Latin America (odd years) and Europe (even years), with Cancun, Mexico in 1999; A Coruña, Spain in 2000; Laguna de San Rafael, Chile in 2001; Lisbon, Portugal in 2002; Manaus, Brazil in 2003; Padova, Italy in 2004; Buenos Aires, Argentina in 2005; Glasgow, UK in 2006; and Santiago, Chile in 2007. This pattern was broken when SPIRE 2008 was held in Melbourne, Australia, but it was probably restarted in 2009 when the venue was Saariselkä, Finland. In 2010 SPIRE took place in Los Cabos, Mexico.

This Special Issue brings together extended versions of some among the best papers of SPIRE 2011. The authors of the 13 papers which had been considered the best by the SPIRE 2011 Program Committee were invited to submit substantially extended and revised versions of their articles. Each among the papers that were submitted as a result was reviewed by three to four reviewers, each of whom assessed the overall quality of the submitted work and the degree to which it was indeed a substantial extension with respect to the original SPIRE 2011 paper.

As a result, 10 articles were selected for inclusion in this Special Issue. These papers range on a variety of topics, from string matching and text retrieval to document indexing and data compression, to bioinformatics, and still others.

The paper “Improved compressed indexes for full-text document retrieval”, by **Djamal Belazzougui**, **Gonzalo Navarro**, and **Daniel Valenzuela**, presents new space/time tradeoffs for compressed indexes that answer document retrieval queries on general sequences using monotone minimal perfect hash functions and succinct data structures, providing experimental results that show relevant practical tradeoffs for document listing with frequencies.

The paper “Approximate regular expression matching with multi-strings”, by **Djamal Belazzougui** and **Mathieu Raffinot**, addresses the problem of finding all the occurrences of a given regular expression in a text by allowing at most k errors, where an error consists in deleting, inserting, or substituting a character. New time bounds are proposed, improving over the best known ones when k is not too large and the regular expression contains much fewer occurrences of $\cup$ and $*$ than of $(\cdot)$.

The paper “Computing the longest common prefix array based on the Burrows–Wheeler transform”, by **Timo Beller**, **Simon Gog**, **Enno Ohlebusch**, and **Thomas Schnattinger**, presents the first algorithm that computes the longest common prefix array directly on the wavelet tree of the Burrows–Wheeler transformed string. Its running time is linear, and it can be implemented using a working memory of approximately extra 2.2 bytes per character.

The paper “Near real-time suffix tree construction via the fringe marked ancestor problem”, by **Dany Breslauer** and **Giuseppe F. Italiano**, received the Best Paper Award. It describes a further step towards the plausible real-time construction of suffix trees for text of length n by presenting an on-line algorithm that requires only $O(\log\log n)$ time for processing each input symbol, and $O(n\log\log n)$ time in total. It improves over previous work by adapting Weiner’s suffix tree construction algorithm to use a new data structure for the fringe marked ancestor problem, a special case of the nearest marked ancestor problem, which may be of independent interest.

The paper “Model based comparison of discounted cumulative gain and average precision”, by **Georges Dupret** and **Benjamin Piwowarski**, takes a detailed look at Discounted Cumulative Gain (DCG), an evaluation measure which has gained a lot of popularity in information retrieval in the last ten years. In this paper DCG is explained probabilistically, as the expected utility obtained by users who inspect the ranked result list working down from the top. The paper also introduces

1570-8667/$ – see front matter
http://dx.doi.org/10.1016/j.jda.2012.12.008

an interesting way to look at evaluation measures, distinguishing their *prognostic* value (i.e., their capability to predict future user behaviour) from their *diagnostic* value (i.e., their capability to explain past user behaviour).

The paper "A novel approach for leveraging co-occurrence to improve the false positive error in signature files", by **Pedram Ghodsnia**, **Kamran Tirdad**, **J. Ian Munro**, and **Alejandro López-Ortiz**, received the Best Student Paper Award for a paper coauthored by at least one student. It introduces COCA filters. In essence, COCA filters are Bloom filters in which the hash functions are selected to be min-wise independent permutations. The major interesting result of the paper is that they allow "small" signature files to achieve low false positive rates through the use of term co-occurrence information. The authors show experimentally that, by using COCA filters, the false positive error rate can be reduced by up to 21 times, for the same index size. The main conclusion is that COCA filters can be considered as a good replacement for Bloom filters whenever the co-occurrence of any two members of the universe is identifiable.

The paper "Modelling efficient novelty-based search result diversification in metric spaces", by **Veronica Gil-Costa**, **Rodrygo L.T. Santos**, **Craig Macdonald**, and **Iadh Ounis**, introduces a new approach for novelty-based search result diversification based on properties of metric spaces. The main result of the paper is the substantial reduction of the overhead incurred by document-document comparisons, by embedding the document space into a metric space. To this end, authors model the novelty promotion as a similarity search in a metric space, exploiting the properties of this space in order to efficiently identify novel documents. Through experiments on two TREC test collections for diversity evaluation, and on a large sample of the query stream of a commercial search engine, authors show that their approach performs at least as effectively as well-known novelty-based diversification approaches in the literature, while dramatically improving their efficiency.

The paper "Fast q-gram mining on SLP compressed strings", by **Keisuke Goto**, **Hideo Bannai**, **Shunsuke Inenaga**, and **Masayuki Takeda**, presents simple and efficient algorithms for calculating q-gram frequencies on strings represented in compressed form, namely, as a straight line program (SLP). Computational experiments show that the presented algorithms are practical for small q, actually running faster on various data mining and classification real datasets.

The paper "ESP-index: A compressed index based on edit-sensitive parsing", by **Shirou Maruyama**, **Masaya Nakahara**, **Naoya Kishiue**, and **Hiroshi Sakamoto**, proposes the ESP-index, a self-index based on edit-sensitive parsing where the text is represented succinctly by a restricted direct acyclic graph (DAG). Searching a pattern reduces to the problem of embedding its parse tree into the DAG of the compressed text.

The paper "Indexing hypertext", by **Chris Thachuk**, considers the case when a transcriptome can be modelled as a hypertext. This is a generalization of a linear text to a graph where nodes contain text and edges denote which nodes can be concatenated. Motivated by this application, the paper introduces the first succinct index for hypertext, describing a new exact pattern matching algorithm, capable of aligning a pattern to any path in the hypertext.

Many people have contributed to bringing this Special Issue to life. A special word of thank goes to the referees, whose job was instrumental in providing timely and high-quality feedback to the authors. It is thus a great pleasure to acknowledge the help of Achille Frigeri, Ben Carterette, Bojian Xu, Diego Ceccarelli, Emine Yilmaz, Fabio Vandin, Fabrizio Falchi, Francisco Claude, Franco Maria Nardini, German Tischler, Giovanni Manzini, Giuseppe Amato, Heikki Hyyrö, Kalervo Jarvelin, Laurent Mouchard, Lucian Ilie, Markus Lohrey, Meng He, Oren Weimann, Roberto Grossi, Roi Blanco, Rossano Venturini, Said Abdeddaim, Srinivasa Rao Satti, and Yasuo Tabei. We are also grateful to Dov Gabbay, Costas Iliopoulos, Michiel Smid, Eli Upfal, and Dorothea Wagner for encouraging us to produce this Special Issue, which we hope represents an important contribution to research on string processing and information retrieval.

Guest Editors
Roberto Grossi
Fabrizio Sebastiani
Fabrizio Silvestri