

Commenced Publication in 1973

Founding and Former Series Editors:

Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

David Hutchison

Lancaster University, UK

Takeo Kanade

Carnegie Mellon University, Pittsburgh, PA, USA

Josef Kittler

University of Surrey, Guildford, UK

Jon M. Kleinberg

Cornell University, Ithaca, NY, USA

Alfred Kobsa

University of California, Irvine, CA, USA

Friedemann Mattern

ETH Zurich, Switzerland

John C. Mitchell

Stanford University, CA, USA

Moni Naor

Weizmann Institute of Science, Rehovot, Israel

Oscar Nierstrasz

University of Bern, Switzerland

C. Pandu Rangan

Indian Institute of Technology, Madras, India

Bernhard Steffen

TU Dortmund University, Germany

Madhu Sudan

Microsoft Research, Cambridge, MA, USA

Demetri Terzopoulos

University of California, Los Angeles, CA, USA

Doug Tygar

University of California, Berkeley, CA, USA

Gerhard Weikum

Max Planck Institute for Informatics, Saarbruecken, Germany

Nieves Brisaboa Oscar Pedreira
Pavel Zezula (Eds.)

Similarity Search and Applications

6th International Conference, SISAP 2013
A Coruña, Spain, October 2-4, 2013
Proceedings



Springer

Volume Editors

Nieves Brisaboa
Oscar Pedreira
Universidade da Coruña
Department of Computer Science, A Coruña, Spain
E-mail: {brisaboa, opedreira}@udc.es

Pavel Zezula
Masaryk University
Department of Computer Systems and Communications
Brno, Czech Republic
E-mail: zezula@fi.muni.cz

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-41061-1 e-ISBN 978-3-642-41062-8
DOI 10.1007/978-3-642-41062-8
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013948103

CR Subject Classification (1998): H.3.1, I.5.3, E.1, H.3.3, H.2.4, H.2.8, F.2.2, G.1.2-3

LNCS Sublibrary: SL 3 – Information Systems and Application, incl. Internet/Web and HCI

© Springer-Verlag Berlin Heidelberg 2013

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

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.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This volume contains the papers presented at the 6th International Conference on Similarity Search and Applications (SISAP 2013), held at A Coruña, Spain, during October 2–4, 2013.

The International Conference on Similarity Search and Applications (SISAP) is an annual forum for researchers and application developers in the area of similarity data management. It aims at the technological problems shared by many application domains, such as data mining, information retrieval, computer vision, pattern recognition, computational biology, geography, biometrics, machine learning, and many others that need similarity searching as a necessary supporting service.

Traditionally, SISAP conferences have put emphasis on the distance-based searching, but in general the conference concerns both the effectiveness and efficiency aspects of any similarity search approach.

In this edition, we had the ambition of widening SISAP's scope to cover even more aspects related to similarity. As a result, we have attracted a higher number of applications covering a wider range of approaches and application domains than in previous editions. In this way, we have achieved a more competitive conference, attractive to a larger part of the computer science community. In the future, we will keep that intent wishing SISAP to become a relevant conference on similarity search, bringing together a wide community of researchers and practitioners and welcoming contributions that range from theoretical aspects to innovative developments for which similarity search plays the central role.

The call for papers welcomed three types of contributions: *(i)* research papers (full or short papers) presenting previously unpublished research contributions, *(ii)* case studies and application papers (short papers) describing existing applications of similarity search in real scenarios, and *(iii)* demo papers describing real or prototype systems for which similarity search technology is a core component, presented at the conference with a system demonstration.

We received 44 submissions, from which 31 were full papers, 11 were short papers, and 2 were demo papers. The Program Committee (PC) comprised 29 researchers from 17 different countries. Each submission was assigned to at least three PC members. Reviews were discussed by the chairs and PC members when the reviews diverged and no sound decision had been reached. The final selection of papers was made by the PC chairs based on the reviews received by each submission. Finally, the conference program includes 19 full papers, 6 short papers, and 2 demo papers, which results in a 61,29% acceptance ratio for full papers and a 54,54% acceptance ration for short papers.

The conference program and the proceedings are organized into five parts. The first one comprises papers dealing with new scenarios or presenting new approaches to similarity search. A second part is devoted to papers proposing

improvements to different methods and techniques for similarity search. The third part focuses on particular metrics and their effectiveness. The fourth part of the conference program includes papers devoted to solutions for similarity search in specific application domains, such as recommender systems, search engines, computational biology, and image and video retrieval. Finally, the last part comprises those papers dealing with efficient implementation and engineering solutions for similarity search in real settings.

The conference program also includes two invited talks from well-known researchers in the field. The first one, “Similarity in Web Search”, by Ricardo Baeza-Yates, surveys different aspects of Web search in which similarity search plays an important role, considering the variety of objects that need to be compared, and the nature and features of the metrics involved in each case. The second one, “Large Scale Visual Object Retrieval”, by Jiri Matas, presents the state of the art in visual retrieval of specific objects, and describes two new methods for large scale object retrieval.

As in previous editions, the proceedings are published by Springer-Verlag, in the *Lecture Notes in Computer Science* series. A selection of the best papers presented at the conference were recommended for publication in the journal *Information Systems*. The selection of best papers was made by the PC, based on the reviews received by each paper, and on the discussion during the conference.

SISAP conferences are organized by the SISAP initiative (www.sisap.org), which aims to become a forum to exchange real-world, challenging and innovative examples of applications, new indexing techniques, common test-beds and benchmarks, source code and up-to-date literature through its web page, serving the similarity search community.

We would like to acknowledge the generous collaboration and financial support from University of A Coruña, Spain (hosting institution), the Fields Institute for Research in Mathematical Sciences, Canada, and the Center for Research and Development in Information Technologies (CITIC) of University of A Coruña. We want to express our gratitude to the PC members for their effort and contribution to the conference. All the submission, reviewing, and proceedings generation processes were carried out through the EasyChair platform.

October 2013

Nieves Brisaboa
Oscar Pedreira
Pavel Zezula

Organization

Program Committee Chairs

Nieves R. Brisaboa	Universidade da Coruña, Spain
Pavel Zezula	Masaryk University, Czech Republic

Program Committee Members

Giuseppe Amato	Istituto di Scienza e Tecnologie dell'Informazione (CNR), Italy
Laurent Amsaleg	Institut de Recherche en Informatique et Systèmes Aléatoires, France
Nieves Brisaboa	Universidade da Coruña, Spain
Benjamin Bustos	Universidad de Chile, Chile
Edgar Chavez	Universidad Nacional Autónoma de México, Mexico
Paolo Ciaccia	University of Bologna, Italy
Richard Connor	University of Strathclyde, UK
Andrea Esuli	Istituto di Scienza e Tecnologie dell'Informazione (CNR), Italy
Rosalba Giugno	University of Catania, Italy
Michael Houle	National Institute of Informatics, Japan
Alexis Joly	Inria, France
Björn Jónsson	Reykjavík University, Iceland
Daniel Keim	Universität Konstanz, Germany
Eamonn Keogh	University of California at Riverside, USA
Magnus Lie Hetland	Norwegian University of Science and Technology (NTNU), Norway
Yannis Manolopoulos	Aristotle University of Thessaloniki, Greece
Rui Mao	Shenzhen University, China
Luisa Micó	Universidad de Alicante, Spain
Henning Müller	University of Applied Sciences Western Switzerland, Switzerland
Gonzalo Navarro	Universidad de Chile, Chile
Arlindo Oliveira	Lisbon Technical University, Portugal
José Oncina	Universidad de Alicante, Spain
Apostolos Papadopoulos	Aristotle University of Thessaloniki, Greece
Marco Patella	University of Bologna, Italy
Oscar Pedreira	Universidade da Coruña, Spain
Vladimir Pestov	University of Ottawa, Canada

Matthias Renz

Ludwig-Maximilians-Universität München,
Germany

Hanan Samet

University of Maryland, USA

Tomas Skopal

Charles University in Prague, Czech Republic

Pavel Zezula

Masaryk University, Czech Republic

Additional Reviewers

Bartolini, Ilaria

Pedreira, Oscar

Buisson, Olivier

Reyes, Nora

Falchi, Fabrizio

Spretke, David

Hoksza, David

Stoffel, Andreas

Krulis, Martin

Stoffel, Florian

Li, Hao

Symeonidis, Panagiotis

Lokoc, Jakub

Tellez, Eric Sadit

Moss, Robert

Wanner, Franz

Paredes, Rodrigo

Table of Contents

Invited Papers

Similarity in Web Search	1
<i>Ricardo Baeza-Yates</i>	
Image Retrieval for Online Browsing in Large Image Collections	3
<i>Andrej Mikulik, Ondřej Chum, and Jiří Matas</i>	

New Scenarios and Approaches

Rank Cover Trees for Nearest Neighbor Search	16
<i>Michael E. Houle and Michael Nett</i>	
A New Concept of Sets to Handle Similarity in Databases: The SimSets	30
<i>Ives R.V. Pola, Robson L.F. Cordeiro, Caetano Traina Jr., and Agma J.M. Traina</i>	
Similarity Search on Uncertain Spatio-temporal Data	43
<i>Johannes Niedermayer, Andreas Züfle, Tobias Emrich, Matthias Renz, Nikos Mamoulis, Lei Chen, and Hans-Peter Kriegel</i>	
List of Clustered Permutations for Proximity Searching	50
<i>Karina Figueroa and Rodrigo Paredes</i>	
Machine Learning for Image Classification and Clustering Using a Universal Distance Measure	59
<i>Uzi Chester and Joel Ratsaby</i>	
Faster Algorithms for Tree Similarity Based on Compressed Enumeration of Bounded-Sized Ordered Subtrees	73
<i>Kunihiro Wasa, Kouichi Hirata, Takeaki Uno, and Hiroki Arimura</i>	

Improving Similarity Search Methods and Techniques

Principal Directions-Based Pivot Placement	85
<i>Fabrizio Angiulli and Fabio Fassetti</i>	
Pivot Selection Strategies for Permutation-Based Similarity Search	91
<i>Giuseppe Amato, Andrea Esuli, and Fabrizio Falchi</i>	
Quantized Ranking for Permutation-Based Indexing	103
<i>Hisham Mohamed and Stéphane Marchand-Maillet</i>	

Extreme Pivots for Faster Metric Indexes	115
<i>Guillermo Ruiz, Francisco Santoyo, Edgar Chávez, Karina Figueroa, and Eric Sadit Tellez</i>	
Quicker Similarity Joins in Metric Spaces	127
<i>Kimmo Fredriksson and Billy Braithwaite</i>	

Metrics and Evaluation

Evaluation of Different Metrics for Shape Based Image Retrieval Using a New Contour Points Descriptor	141
<i>María Teresa García-Ordás, Enrique Alegre, Oscar García-Olalla, and Diego García-Ordás</i>	
Evaluation of LBP Variants Using Several Metrics and kNN Classifiers	151
<i>Oscar García-Olalla, Enrique Alegre, María Teresa García-Ordás, and Laura Fernández-Robles</i>	
Evaluation of Jensen-Shannon Distance over Sparse Data	163
<i>Richard Connor, Franco Alberto Cardillo, Robert Moss, and Fausto Rabitti</i>	
A Multi-way Divergence Metric for Vector Spaces	169
<i>Robert Moss and Richard Connor</i>	
Optimal Distance Bounds for the Mahalanobis Distance	175
<i>Tobias Emrich, Gregor Jossé, Hans-Peter Kriegel, Markus Mauder, Johannes Niedermayer, Matthias Renz, Matthias Schubert, and Andreas Züfle</i>	

Applications and Specific Domains

Text Categorization via Similarity Search: An Efficient and Effective Novel Algorithm	182
<i>Hubert Haoyang Duan, Vladimir G. Pestov, and Varun Singla</i>	
Efficient Approximate Indexing in High-Dimensional Feature Spaces	194
<i>Simone Santini</i>	
Semi-supervised Tag Extraction in a Web Recommender System	206
<i>Vasily A. Leksin and Sergey I. Nikolenko</i>	
A Similarity Model for 3D Objects Based on Stable Sub-clouds	213
<i>Markus Mauder, Peer Kröger, and Karl-Ludwig Schinner</i>	
Accurate and Efficient Search Prediction Using Fuzzy Matching and Outcome Feedback	219
<i>Christopher Shaun Wagner, Sahra Sedigh, and Ali R. Hurson</i>	

Beyond Bag of Words for Concept Detection and Search of Cultural Heritage Archives	233
<i>Costantino Grana, Giuseppe Serra, Marco Manfredi, and Rita Cucchiara</i>	

Large Scale Image Retrieval Using Vector of Locally Aggregated Descriptors	245
<i>Giuseppe Amato, Paolo Bolettieri, Fabrizio Falchi, and Claudio Gennaro</i>	

Longest Common Subsequence in k Length Substrings	257
<i>Gary Benson, Avivit Levy, and B. Riva Shalom</i>	

Implementation and Engineering Solutions

Database Similarity Join for Metric Spaces	266
<i>Yasin N. Silva, Spencer S. Pearson, and Jason A. Cheney</i>	

Engineering Efficient and Effective Non-metric Space Library	280
<i>Leonid Boytsov and Bilegsaikhan Naidan</i>	

Designing Similarity Indexes with Parallel Genetic Programming	294
<i>Tomáš Bartoš and Tomáš Skopal</i>	

(Very) Fast (All) k -Nearest Neighbors in Metric and Non Metric Spaces without Indexing	300
<i>Natalia Miranda, Edgar Chávez, María Fabiana Piccoli, and Nora Reyes</i>	

On Scalable Approximate Search with the Signature Quadratic Form Distance	312
<i>Jakub Lokoč, Tomáš Grošup, and Tomáš Skopal</i>	

Demo Papers

PFMFind: A System for Discovery of Peptide Homology and Function	319
<i>Aleksandar Stojmirović, Peter Andreae, Mike Boland, Thomas William Jordan, and Vladimir G. Pestov</i>	

Retrieving Similar Movements in Motion Capture Data	325
<i>Jan Sedmidubsky and Jakub Valcik</i>	

Author Index	331
--------------------	-----