

*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*

Gonzalo Navarro Vladimir Pestov (Eds.)

# Similarity Search and Applications

5th International Conference, SISAP 2012  
Toronto, ON, Canada, August 9-10, 2012  
Proceedings

 Springer

## Volume Editors

Gonzalo Navarro  
Universidad de Chile  
Departamento de Ciencias de la Computación  
Blanco Encalada 2120, Santiago, Chile  
E-mail: gnavarro@dcc.uchile.cl

Vladimir Pestov  
University of Ottawa  
Department of Mathematics and Statistics  
585 King Edward Avenue, Ottawa, ON, Canada, K1N 6N5  
E-mail: vpest283@uottawa.ca

ISSN 0302-9743  
ISBN 978-3-642-32152-8  
DOI 10.1007/978-3-642-32153-5  
Springer Heidelberg Dordrecht London New York

e-ISSN 1611-3349  
e-ISBN 978-3-642-32153-5

Library of Congress Control Number: 2012942592

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 2012

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

The use of general descriptive names, registered names, trademarks, 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.

*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 5th International Conference on Similarity Search and Applications (SISAP 2012), which took place during August 9–10, 2012, at the Fields Institute for Research in Mathematical Sciences, Toronto, Ontario, Canada.

SISAP is a conference devoted to similarity searching, with emphasis on metric space searching. It aims to fill in the gap left by the various scientific venues devoted to similarity searching in spaces with coordinates, by providing a common forum for theoreticians and practitioners around the problem of similarity searching in general spaces (metric and non-metric) or using distance-based (as opposed to coordinate-based) techniques in general. Four types of contributions are welcome: (1) fundamental techniques to handle general similarity search problems, (2) applied techniques to solve particular similarity search problems of wide interest, (3) new similarity search problems, where their features and challenges are studied, and (4) actual systems for similarity search, in the form of demos. SISAP is seen as a forum for not only exchanging new indexing techniques and real-world applications, but also common testbeds and benchmarks, and source code. Authors are expected to use the testbeds and code from the SISAP website ([www.sisap.org](http://www.sisap.org)) for comparing new applications, databases, indexes, and algorithms.

This year we received 19 full-paper and two demo submissions, from Argentina, Chile, Czech Republic, France, Japan, Mexico, Norway, Russia, Spain, Switzerland, UK, and USA. Each submission was assigned, in double-blind mode, to three Program Committee (PC) members, who reviewed them themselves and/or supervised subreviews. Submissions received two to five reviews (3.14 on average). Then the PC Chairs and involved members discussed the articles where no obvious agreement had been reached. The final decisions of acceptance or rejection were made by the PC Chairs. Finally, 14 full papers and the two demos were selected to be presented at the conference and to appear in the proceedings.

Of the full papers accepted, nine refer to techniques to handle general similarity search problems, improving upon the state of the art on topics like parallelism, dynamism, secondary memory, approximation techniques, optimized construction, combinations of data structures, and novel scenarios, such as streams of related searches and inferring factual space properties from the data. Further, two accepted papers refer to applied techniques, to similarity searching in string dictionaries and in images. The other three papers study the properties of specific spaces such as sequences under time-warping distance and factorized tensors, and propose and study new distances for vector spaces based on entropy correlations. Of the two demos, one presents an image meta-search engine, and

the other introduces a tool for identifying protein and peptide sequences from tandem mass spectra.

Overall, the articles formed an extremely stimulating set of contributions to many of the most relevant aspects of similarity searching. Two invited presentations and papers from prominent researchers further enriched this year's SISAP. The first one, "Effective Principal Component Analysis," by Santosh Vempala, is about the success and challenges around this technique, of wide relevance in similarity search and various other fields. The second one, "Future Trends in Similarity Searching," by Pavel Zezula, is a revealing survey and analysis of where the discipline is expected to head in the forthcoming years.

This year the proceedings of SISAP were published by Springer-Verlag, in the *Lecture Notes in Computer Science* series. A selection of the best papers was recommended for inclusion in a special issue of the *Information Systems* journal dedicated to this conference. These were chosen by the PC Chairs based on the original reviews of the articles and their oral presentation during the conference, as well as appropriateness to the journal.

The subject matter of the SISAP conferences, although primarily a computer science topic, uses a great deal of advanced mathematical methods, such as those of geometric functional analysis and statistical machine learning. The conference is a perfect platform for interactions between computer scientists and mathematicians, and the stimulating research ambiance of the Fields Institute gave fresh impetus to such interactions. We thank the Fields Institute for the hosting of SISAP 2012 conference.

Last, but not least, we acknowledge the generous financial support from (again) the Fields Institute for Research in Mathematical Sciences, Canada; the Canadian Network of Excellence in Mathematics of Information Technology and Complex Systems (MITACS); and the Natural Sciences and Engineering Research Council of Canada (NSERC) research grant "New Set-Theoretic Tools for Statistical Learning." All the submission, reviewing, and proceedings generation processes were handled through the EasyChair platform.

August 2012

Gonzalo Navarro  
Vladimir Pestov

# Organization

## Committees

### Steering Committee

Edgar Chávez	Universidad Michoacana, Mexico
Gonzalo Navarro	Universidad de Chile, Chile

### Program Committee Chairs

Gonzalo Navarro	Universidad de Chile, Chile
Vladimir Pestov	Université d'Ottawa, Canada

### Program Committee Members

Edgar Chávez	Universidad Michoacana, Mexico
Paolo Ciaccia	Università di Bologna, Italy
Alfredo Ferro	Università di Catania, Italy
Daniel Keim	Universität Konstanz, Germany
Daniel Miranker	University of Texas at Austin, USA
Marco Patella	Università di Bologna, Italy
Hanan Samet	University of Maryland, USA
Tomáš Skopal	Charles University in Prague, Czech Republic
Aleksandar Stojmirović	NCBI/NLM/NIH, USA
Agma Traina	Universidade de São Paulo – São Carlos, Brazil
Pavel Zezula	Masaryk University, Czech Republic

### Organization Chair

Vladimir Pestov	Université d'Ottawa, Canada
-----------------	-----------------------------

### Publicity Chair

Tomáš Skopal	Charles University in Prague, Czech Republic
--------------	--

## Additional Reviewers

Marco Adelfio	David Hoksza	Luís M. Silveira Russo
Gelio Alves	Eamonn Keogh	Jan Sedmidubsky
Michal Batko	Jakub Lokoč	John Spouge
Petra Budikova	Jiří Novák	Eric Sadit Téllez Avila
Benjamin Bustos	Sarana Nutanong	Lee Thompson
Carlos Castillo	Ives Pola	German Tischler
Vlastislav Dohnal	Mônica R. Porto Ferreira	Kesheng Wu
Magnus Lie Hetland	Nora Reyes	

## **Sponsoring Institutions**

Fields Institute for Research in Mathematical Sciences, Canada

Canadian Network of Excellence in Mathematics of Information Technology and  
Complex Systems (MITACS)

Natural Sciences and Engineering Research Council of Canada (NSERC)

# Table of Contents

## Invited Papers

Effective Principal Component Analysis . . . . .	1
<i>Santosh S. Vempala</i>	
Future Trends in Similarity Searching . . . . .	8
<i>Pavel Zezula</i>	

## New Scenarios and Approaches

Snake Table: A Dynamic Pivot Table for Streams of k-NN Searches . . . .	25
<i>Juan Manuel Barrios, Benjamin Bustos, and Tomáš Skopal</i>	
Algorithmic Exploration of Axiom Spaces for Efficient Similarity Search at Large Scale . . . . .	40
<i>Tomáš Skopal and Tomáš Bartoš</i>	
Polyphasic Metric Index: Reaching the Practical Limits of Proximity Searching . . . . .	54
<i>Eric Sadit Tellez, Edgar Chavez, and Karina Figueroa</i>	

## Improving Metric Data Structures

Efficient Similarity Search in Metric Spaces with Cluster Reduction . . . .	70
<i>Luis G. Ares, Nieves R. Brisaboa, Alberto Ordóñez Pereira, and Oscar Pedreira</i>	
Cut-Region: A Compact Building Block for Hierarchical Metric Indexing . . . . .	85
<i>Jakub Lokoč, Přemysl Čech, Jiří Novák, and Tomáš Skopal</i>	
Static-to-Dynamic Transformation for Metric Indexing Structures . . . . .	101
<i>Bilegsaikhan Naidan and Magnus Lie Hetland</i>	

## Facing Scalability Issues

DSACL+-tree: A Dynamic Data Structure for Similarity Search in Secondary Memory . . . . .	116
<i>Luis Britos, A. Marcela Printista, and Nora Reyes</i>	



Scalable Distributed Algorithm for Approximate Nearest Neighbor Search Problem in High Dimensional General Metric Spaces ..... 132  
*Yury Malkov, Alexander Ponomarenko, Andrey Logvinov, and Vladimir Krylov*

Parallel Approaches to Permutation-Based Indexing Using Inverted Files ..... 148  
*Hisham Mohamed and Stéphane Marchand-Maillet*

**Searching in Specific Spaces**

Super-Linear Indices for Approximate Dictionary Searching ..... 162  
*Leonid Boytsov*

Visual Image Search: Feature Signatures or/and Global Descriptors .... 177  
*Jakub Lokoč, David Novák, Michal Batko, and Tomáš Skopal*

**New Similarity Spaces**

Revisiting Techniques for Lowerbounding the Dynamic Time Warping Distance ..... 192  
*Tomáš Bartoš and Tomáš Skopal*

A Multivariate Correlation Distance for Vector Spaces ..... 209  
*Richard Connor and Robert Moss*

Fast Similarity Computation in Factorized Tensors ..... 226  
*Michael E. Houle, Hisashi Kashima, and Michael Nett*

**Demo Papers**

SIR: The Smart Image Retrieval Engine ..... 240  
*Jakub Lokoč, Tomáš Grošup, and Tomáš Skopal*

SimTandem: Similarity Search in Tandem Mass Spectra..... 242  
*Jiří Novák, Jakub Galgonek, David Hoksza, and Tomáš Skopal*

**Erratum**

Parallel Approaches to Permutation-Based Indexing Using Inverted Files ..... E1  
*Hisham Mohamed and Stéphane Marchand-Maillet*

**Author Index** ..... 245