

Nora Reyes · Richard Connor ·
Nils Kriege · Daniyal Kazempour ·
Ilaria Bartolini · Erich Schubert ·
Jian-Jia Chen (Eds.)

LNCS 13058

Similarity Search and Applications

14th International Conference, SISAP 2021
Dortmund, Germany, September 29 – October 1, 2021
Proceedings



 Springer

Founding Editors

Gerhard Goos

Karlsruhe Institute of Technology, Karlsruhe, Germany

Juris Hartmanis

Cornell University, Ithaca, NY, USA


Editorial Board Members

Elisa Bertino

Purdue University, West Lafayette, IN, USA

Wen Gao

Peking University, Beijing, China

Bernhard Steffen 

TU Dortmund University, Dortmund, Germany

Gerhard Woeginger 

RWTH Aachen, Aachen, Germany

Moti Yung

Columbia University, New York, NY, USA

More information about this subseries at <http://www.springer.com/series/7409>

Nora Reyes · Richard Connor · Nils Kriege ·
Daniyal Kazempour · Iaria Bartolini ·
Erich Schubert · Jian-Jia Chen (Eds.)

Similarity Search and Applications

14th International Conference, SISAP 2021
Dortmund, Germany, September 29 – October 1, 2021
Proceedings

Editors


Nora Reyes 
National University of San Luis
San Luis, Argentina


Richard Connor 
University of St Andrews
St Andrews, UK

Nils Kriege 
University of Vienna
Vienna, Austria

Daniyal Kazempour 
Kiel University
Kiel, Germany

Ilaria Bartolini 
University of Bologna
Bologna, Italy

Erich Schubert 
TU Dortmund University
Dortmund, Germany

Jian-Jia Chen 
TU Dortmund University
Dortmund, Germany

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-030-89656-0 ISBN 978-3-030-89657-7 (eBook)
<https://doi.org/10.1007/978-3-030-89657-7>

LNCS Sublibrary: SL3 – Information Systems and Applications, incl. Internet/Web, and HCI

© Springer Nature Switzerland AG 2021

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.

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.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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

Preface

This volume contains the papers presented at the 14th International Conference on Similarity Search and Applications (SISAP 2021) held between September 29 and October 1, 2021. The conference was hosted by TU Dortmund, Germany. Due to the COVID-19 pandemic and international travel restrictions around the globe, SISAP 2021 was planned as a “hybrid or virtual” event, and in August 2021 it was decided that it would be held as an online conference only due to rapidly increasing incidences in Germany despite a good vaccination rate.

SISAP is an annual forum for researchers and application developers in the area of similarity data management. It focuses on the technological problems shared by numerous application domains, such as data mining, information retrieval, multimedia, computer vision, pattern recognition, computational biology, geography, biometrics, machine learning, and many others that make use of similarity search as a necessary supporting service.

From its roots as a regional workshop in metric indexing, SISAP has expanded to become the only international conference entirely devoted to the issues surrounding the theory, design, analysis, practice, and application of content-based and feature-based similarity search. The SISAP initiative has also created a repository¹ serving the similarity search community, for the exchange of examples of real-world applications, source code for similarity indexes, and experimental testbeds and benchmark data sets.

SISAP 2021 continued the two-year tradition of the SISAP Doctoral Symposium, for which a technical program was assembled to give PhD students an opportunity to present their research ideas in an international research venue. The Doctoral Symposium provides a forum that facilitates interaction among PhD students and stimulates feedback from more experienced researchers. This year’s SISAP also included a single special session, on the topic of search in graph-structured data. Again in keeping with previous years, the reviewing process for the special session was integrated with the main conference program to ensure the same quality of acceptance.

The call for papers welcomed full research papers and short research papers, as well as position and demonstration papers, with all manuscripts presenting previously unpublished research contributions.

We received 44 submissions from authors based in 16 different countries. The Program Committee (PC) was composed of 55 members from 20 countries. Each submission received at least four reviews, and the papers and reviews were thoroughly discussed by the chairs and PC members. Based on the reviews and discussions, the PC chairs accepted 23 full papers and 5 short papers, resulting in an acceptance rate of 52% for the full papers and a cumulative acceptance rate of 64% for full and short papers. These rates are a little higher than usual, however the PC chairs are confident that this does not reflect a drop in standards, but rather is an artifact of the context of the COVID-19 pandemic. After a separate review by the Doctoral Symposium Program

¹ <https://www.sisap.org/>.

Committee members, three Doctoral Symposium papers, giving a clear sample of emerging topics in similarity search and applications, were accepted for presentation and included in the program and proceedings.

The proceedings of SISAP are published by Springer as a volume in the Lecture Notes in Computer Science (LNCS) series. For SISAP 2021, as in previous years, extended versions of selected excellent papers were invited for publication in a special issue of the journal Information Systems. The conference also conferred a Best Paper Award, a Best Student Paper Award, and a Best Doctoral Symposium Paper Award, as judged by the PC chairs and the Steering Committee.

We would like to thank all the authors who submitted papers to SISAP 2021. We would also like to thank all members of the PC and the external reviewers for their effort and contribution to the conference. We want to extend our gratitude to the members of the Organizing Committee for the enormous amount of work they have done, and our sponsors and supporters for their generosity. Finally, we thank all the participants in the online event, who make up the thriving SISAP community.

September 2021

Nora Reyes
Richard Connor
Nils Kriege
Daniyal Kazempour
Ilaria Bartolini
Erich Schubert
Jian-Jia Chen

Organization

General Chairs

Erich Schubert TU Dortmund University, Germany
Jian-Jia Chen TU Dortmund University, Germany

Program Committee Chairs

Richard Connor University of St Andrews, UK
Nora Reyes Universidad Nacional de San Luis, Argentina

Doctoral Symposium Chair

Ilaria Bartolini University of Bologna, Italy

Publication Chair

Daniyal Kazempour Christian-Albrechts-Universität zu Kiel, Germany

Publicity Chair

Peer Kröger Christian-Albrechts-Universität zu Kiel, Germany

Steering Committee

Laurent Amsaleg CNRS-IRISA, France
Edgar Chávez CICESE, Mexico
Michael E. Houle National Institute of Informatics, Japan
Pavel Zezula Masaryk University, Czech Republic

Program Committee

Giuseppe Amato ISTI-CNR, Italy
Laurent Amsaleg CNRS-IRISA, France
Fabrizio Angiulli University of Calabria, Italy
Ilaria Bartolini University of Bologna, Italy
Christian Beecks University of Münster, Germany
Panagiotis Bouros Johannes Gutenberg University Mainz, Germany
Benjamin Bustos University of Chile, Chile

K. Selcuk Candan	Arizona State University, USA
Edgar Chavez	CICESE, Mexico
Alan Dearle	University of St Andrews, UK
Vlastislav Dohnal	Masaryk University, Czech Republic
Vladimir Estivill-Castro	Universitat Pompeu Fabra, Spain
Rolf Fagerberg	University of Southern Denmark, Denmark
Fabrizio Falchi	ISTI-CNR, Italy
Karina Figueroa	Universidad Michoacana de San Nicolas de Hidalgo, Mexico
Claudio Gennaro	ISTI-CNR, Italy
Magnus Lie Hetland	Norwegian University of Science and Technology, Norway
Thi Thao Nguyen Ho	Aalborg University, Denmark
Michael E. Houle	National Institute of Informatics, Japan
Daniyal Kazempour	Christian-Albrechts-Universität zu Kiel, Germany
Nils Kriege	University of Vienna, Austria
Peer Kröger	Christian-Albrechts-Universität zu Kiel, Germany
Yusuke Matsui	University of Tokyo, Japan
Vladimir Mic	Masaryk University, Czech Republic
Luisa Micó	University of Alicante, Spain
Lia Morra	Politecnico di Torino, Italy
Henning Müller	HES-SO, Switzerland
Deepak P.	Queen's University Belfast, UK
Rodrigo Paredes	Universidad de Talca, Chile
Marco Patella	University of Bologna, Italy
Oscar Pedreira	Universidade da Coruna, Spain
Miloš Radovanović	University of Novi Sad, Serbia
Marcela Ribeiro	Federal University of São Carlos, Brazil
Kunihiko Sadakane	University of Tokyo, Japan
Maria Luisa Sapino	Universita' di Torino, Italy
Erich Schubert	TU Dortmund University, Germany
Tetsuo Shibuya	University of Tokyo, Japan
Tomas Skopal	Charles University in Prague, Czech Republic
Nenad Tomasev	Google DeepMind, UK
Caetano Traina	University of São Paulo, Brazil
Goce Trajcevski	Iowa State University, USA
Lucia Vadicamo	ISTI-CNR, Italy
Takashi Washio	Osaka University, Japan
Pascal Welke	University of Bonn, Germany
Kaoru Yoshida	Sony Computer Science Laboratories, Inc., Japan
Pavel Zezula	Masaryk University, Czech Republic
Kaiping Zheng	National University of Singapore, Singapore
Arthur Zimek	University of Southern Denmark, Denmark
Andreas Züfle	George Mason University, USA

Additional Reviewers

Franka Bause	University of Vienna, Austria
Andre Droschinsky	TU Dortmund University, Germany
Erik Thordsen	TU Dortmund University, Germany
Florian Kurpicz	Karlsruhe Institute of Technology, Germany
Lukas Miklautz	University of Vienna, Austria
Lutz Oettershagen	University of Bonn, Germany
Till Schulz	University of Bonn, Germany

Contents

Similarity Search and Retrieval

Organizing Similarity Spaces Using Metric Hulls	3
<i>Miriama Jánošová, David Procházka, and Vlastislav Dohnal</i>	
Scaling Up Set Similarity Joins Using a Cost-Based Distributed-Parallel Framework	17
<i>Fabian Fier and Johann-Christoph Freytag</i>	
A Triangle Inequality for Cosine Similarity	32
<i>Erich Schubert</i>	
A Cost Model for Reverse Nearest Neighbor Query Processing on R-Trees Using Self Pruning	45
<i>Felix Borutta, Peer Kröger, and Matthias Renz</i>	
How Many Neighbours for Known-Item Search?	54
<i>Jakub Lokoč and Tomáš Souček</i>	
On Generalizing Permutation-Based Representations for Approximate Search	66
<i>Lucia Vadicamo, Claudio Gennaro, and Giuseppe Amato</i>	
Data-Driven Learned Metric Index: An Unsupervised Approach	81
<i>Terézia Slanínáková, Matej Antol, Jaroslav Ořha, Vojtěch Kaňá, and Vlastislav Dohnal</i>	
Towards a Learned Index Structure for Approximate Nearest Neighbor Search Query Processing	95
<i>Maximilian Hünemörder, Peer Kröger, and Matthias Renz</i>	
Similarity vs. Relevance: From Simple Searches to Complex Discovery	104
<i>Tomáš Skopal, David Bernhauer, Petr Škoda, Jakub Klímek, and Martin Nečaský</i>	
Non-parametric Semi-supervised Learning by Bayesian Label Distribution Propagation	118
<i>Jonatan Møller Nuutinen Gøttcke, Arthur Zimek, and Ricardo J. G. B. Campello</i>	

Optimizing Fair Approximate Nearest Neighbor Searches Using Threaded B+-Trees	133
<i>Omid Jafari, Preeti Maurya, Khandker Mushfiqul Islam, and Parth Nagarkar</i>	
Fairest Neighbors: Tradeoffs Between Metric Queries	148
<i>Magnus Lie Hetland and Halvard Hummel</i>	
Intrinsic Dimensionality	
Local Intrinsic Dimensionality and Graphs: Towards LID-aware Graph Embedding Algorithms	159
<i>Miloš Savić, Vladimir Kurbalija, and Miloš Radovanović</i>	
Structural Intrinsic Dimensionality	173
<i>Stephane Marchand-Maillet, Oscar Pedreira, and Edgar Chávez</i>	
Relationships Between Local Intrinsic Dimensionality and Tail Entropy	186
<i>James Bailey, Michael E. Houle, and Xingjun Ma</i>	
The Effect of Random Projection on Local Intrinsic Dimensionality	201
<i>Michael E. Houle and Ken-ichi Kawarabayashi</i>	
Clustering and Classification	
Accelerating Spherical k -Means	217
<i>Erich Schubert, Andreas Lang, and Gloria Feher</i>	
MESS: Manifold Embedding Motivated Super Sampling	232
<i>Erik Thordsen and Erich Schubert</i>	
Handling Class Imbalance in k -Nearest Neighbor Classification by Balancing Prior Probabilities	247
<i>Jonatan Møller Nuutinen Gøttcke and Arthur Zimek</i>	
Applications of Similarity Search	
Similarity Search for an Extreme Application: Experience and Implementation	265
<i>Vladimir Mic, Tomáš Raček, Aleš Křenek, and Pavel Zezula</i>	
What Makes a Good Movie Recommendation? Feature Selection for Content-Based Filtering	280
<i>Maciej Gawinecki, Wojciech Szmyd, Urszula Żuchowicz, and Marcin Walas</i>	

Indexed Polygon Matching Under Similarities	295
<i>Fernando Luque-Suarez, J. L. López-López, and Edgar Chavez</i>	
Clustering Adverse Events of COVID-19 Vaccines Across the United States ...	307
<i>Ahmed Askar and Andreas Züfle</i>	
Similarity Search in Graph-Structured Data	
Metric Indexing for Graph Similarity Search	323
<i>Franka Bause, David B. Blumenthal, Erich Schubert, and Nils M. Kriege</i>	
The Minimum Edit Arborescence Problem and Its Use in Compressing Graph Collections	337
<i>Lucas Gnecco, Nicolas Boria, Sébastien Bougleux, Florian Yger, and David B. Blumenthal</i>	
Graph Embedding in Vector Spaces Using Matching-Graphs	352
<i>Mathias Fuchs and Kaspar Riesen</i>	
An A*-algorithm for the Unordered Tree Edit Distance with Custom Costs	364
<i>Benjamin Paafßen</i>	
FIMSIM: Discovering Communities by Frequent Item-Set Mining and Similarity Search	372
<i>Jakub Peschel, Michal Batko, Jakub Valcik, Jan Sedmidubsky, and Pavel Zezula</i>	
Doctoral Symposium	
Towards an Italian Healthcare Knowledge Graph	387
<i>Marco Postiglione</i>	
Progressive Query-Driven Entity Resolution	395
<i>Luca Zecchini</i>	
Discovering Latent Information from Noisy Sources in the Cultural Heritage Domain	402
<i>Fabrizio Scarrone</i>	
Author Index	409