

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 series at <http://www.springer.com/series/7409>

Giuseppe Amato · Claudio Gennaro ·
Vincent Oria · Miloš Radovanović (Eds.)

Similarity Search and Applications


12th International Conference, SISAP 2019
Newark, NJ, USA, October 2–4, 2019
Proceedings

Editors

Giuseppe Amato 
ISTI-CNR
Pisa, Italy

Vincent Oria
New Jersey Institute of Technology
Newark, NJ, USA

Claudio Gennaro 
ISTI-CNR
Pisa, Italy

Miloš Radovanović 
University of Novi Sad
Novi Sad, Serbia

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-030-32046-1 ISBN 978-3-030-32047-8 (eBook)
<https://doi.org/10.1007/978-3-030-32047-8>

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

© Springer Nature Switzerland AG 2019

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 12th International Conference on Similarity Search and Applications (SISAP 2019) held in Newark, NJ, USA, during October 2–4, 2019.

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 repository (<http://www.sisap.org/>) serving the similarity search community, for the exchange of examples of real-world applications, source code for similarity indexes, and experimental test beds and benchmark data sets.

The call for papers welcomed full papers, short papers, as well as demonstration papers, with all manuscripts presenting previously unpublished research contributions. All contributions were presented both orally and in a poster session, which facilitated fruitful exchanges between the participants. In addition, SISAP 2019 featured a doctoral consortium, accepting papers describing doctoral research and work in progress, providing students valuable feedback from experienced researchers in similarity search and related fields.

We received 42 submissions from authors based in 17 different countries. The Program Committee (PC) was composed of 63 international members. The papers and reviews were thoroughly discussed by the chairs and PC members: Each submission received three reviews. Based on these reviews and discussions among PC members, the PC chairs accepted 12 full papers to be included in the conference program and proceedings, resulting in an acceptance rate of 28% for the full papers. In addition, 18 short papers were accepted, and after separate review by SISAP chairs, two doctoral consortium papers were included in the program and proceedings as well.

The proceedings of SISAP are published by Springer as a volume in the *Lecture Notes in Computer Science* (LNCS) series. For SISAP 2019, 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, as judged by the PC co-chairs and Steering Committee.

Besides the presentations of the accepted papers, the conference program featured three keynote talks by exceptional researchers: Fabrizio Silvestri from Facebook, UK, Divesh Srivastava from AT&T Labs-Research, USA, and Prof. Alexander Tuzhilin from New York University, USA.

We would like to thank all the authors who submitted papers to SISAP 2019, as well as 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 invested in making the SISAP series of conferences possible.

We also thank our sponsors and supporters for their generosity. All the submission, reviewing, and proceedings generation processes were made much easier through the EasyChair platform.

August 2019

Giuseppe Amato
Claudio Gennaro
Vincent Oria
Miloš Radovanović

Organization

Program Committee Chairs

Vincent Oria	NJIT, USA
Giuseppe Amato	ISTI-CNR, Italy
Miloš Radovanović	University of Novi Sad, Serbia

Steering Committee

Laurent Amsaleg	CNRS-IRISA, France
Edgar Chavez	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
Martin Aumüller	IT University of Copenhagen, Denmark
Ilaria Bartolini	University of Bologna, Italy
Virendra Bhavsar	University of New Brunswick, Canada
Panagiotis Bouras	Johannes Gutenberg University Mainz, Germany
Benjamin Bustos	University of Chile, Chile
K. Selcuk Candan	Arizona State University, USA
Guang-Ho Cha	Seoul National University of Science and Technology, South Korea
Aniket Chakrabarti	Microsoft, USA
Edgar Chavez	CICESE, Mexico
Richard Chbeir	University of Pau and Pays Adour, UPPA/E2S, LIUPPA Anglet, France
Paolo Ciaccia	University of Bologna, Italy
Richard Connor	University of Stirling, UK
Robson Cordeiro	ICMC-USP, Brazil
Michel Crucianu	CNAM, France
Petros Daras	Information Technologies Institute, Greece
Alberto Del Bimbo	Università degli Studi di Firenze, Italy
Dong Deng	Rutgers University, USA
Vlad Estivill-Castro	Griffith University, Australia
Andrea Esuli	ISTI-CNR, Italy
Fabrizio Falchi	ISTI-CNR, Italy
Joao Eduardo Ferreira	University of São Paulo, Brazil
Karina Figueroa	Universidad Michoacana, Mexico

Renato Fileto	UFSC, Brazil
Teddy Furon	Inria, France
Renata Galante	UFRGS, Brazil
Claudio Gennaro	ISTI-CNR, Italy
Michael E. Houle	National Institute of Informatics, Japan
Ichiro Ide	Nagoya University, Japan
Mirjana Ivanović	University of Novi Sad, Serbia
Vladimir Kurbalija	University of Novi Sad, Serbia
Jakub Lokoc	Charles University in Prague, Czech Republic
S. Marchand-Maillet	University of Geneva, Switzerland
Luisa Mico	University of Alicante, Spain
David Mount	University of Maryland, USA
Henning Müller	HES-SO, Switzerland
Vincent Oria	NJIT, USA
Deepak Padmanabhan	Queen's University Belfast, UK
A. N. Papadopoulos	Aristotle University of Thessaloniki, Greece
Rodrigo Paredes	Universidad de Talca, Chile
Marco Patella	University of Bologna, Italy
Oscar Pedreira	Universidade da Coruna, Spain
Raffaele Perego	ISTI-CNR, Italy
Andreas Rauber	Vienna University of Technology, Austria
Nora Reyes	Universidad Nacional de San Luis, Argentina
Marcela Ribeiro	Federal University of São Carlos, UFSCar, Brazil
Kunihiko Sadakane	The University of Tokyo, Japan
Maria Luisa Sapino	Università di Torino, Italy
Miloš Savić	University of Novi Sad, Serbia
Erich Schubert	Technische Universität Dortmund, Germany
Thomas Seidl	Ludwig-Maximilians-Universität München (LMU Munich), Germany
Tetsuo Shibuya	Human Genome Center, Institute of Medical Science, The University of Tokyo, Japan
Fabrizio Silvestri	Facebook, UK
Tomas Skopal	Charles University in Prague, Czech Republic
Elaine Sousa	University of Sao Paulo, ICMC/USP, Brazil
Peter Stanchev	Kettering University, USA
Miloš Radovanović	University of Novi Sad, Serbia
Yasuo Tabei	RIKEN Center for Advanced Intelligence Project, Japan
Joe Tekli	Lebanese American University, Lebanon
Nenad Tomašev	DeepMind, UK
Agma Traina	University of São Paulo, Brazil
Caetano Traina	University of São Paulo, Brazil
Takashi Washio	ISIR, Osaka University, Japan

Marcel Worring
Kaoru Yoshida
Pavel Zezula
Arthur Zimek

University of Amsterdam, The Netherlands
Sony Computer Science Laboratories, Inc., Japan
Masaryk University, Czech Republic
University of Southern Denmark, Denmark

Additional Reviewers

Vladimir Mic
Joe Tekli

Keynote Abstracts

Applications of Similarity Search to Socially Relevant Problems

Fabrizio Silvestri

Facebook, UK

Abstract. The Facebook AI team in London deals with applying artificial intelligence techniques to address societal problems such as the spread of online misinformation, or the integrity of election processes around the world. To do so, we have developed throughout the last years a set of tools that exploit similarity search technologies to efficiently and effectively run a very high number of classification tasks on a massive set of data.

In this talk, we are going to review some of the problems we have studied in the last year and we are going to show some of the solutions we have adopted in order to make the system run efficiently. We are also going to showcase some details of an internal project that uses similarity search as a core operation to allow efficient and effective inference operations.

Repairing Noisy Graphs

Divesh Srivastava

AT&T Labs-Research, USA

Abstract. Graphs are a flexible way to represent data in a variety of applications, with nodes representing domain-specific entities (e.g., records in record linkage, products and types in an ontology) and edges capturing a variety of relationships between these entities (e.g., an equivalence relationship between records in record linkage, a type-subtype relationship between types in an ontology). Often, the edges in this graph are inferred based on similarities between nodes and are noisy, in that some edges are missing (i.e., real-world relationships that do not have corresponding edges in the graph) and some edges are spurious (i.e., edges in the graph that do not have corresponding real-world relationships). Directly analyzing such graphs can lead to undesirable outcomes, making it important to repair noisy graphs. In this talk, we describe an approach that takes advantage of properties of real-world relationships and their estimated probabilities to ask oracle queries (an abstraction of crowdsourcing) to efficiently repair the noisy graphs. We illustrate this approach for the case of graphs that are unions of cliques (which is the case for record linkage) and graphs that are trees (which is the case for ontologies), and present theoretical and empirical results for these cases.

On Similarity Measures in Recommender Systems

Alexander Tuzhilin

New York University, USA

Abstract. Measures of similarity between users and between items to be recommended to the users lie at the core of many recommendation algorithms, and numerous metrics have been proposed in the recommender systems field since its inception. This talk will explore evolution of various similarity-based measures from the initial class of rating-based measures to the more recently proposed latent metrics and the metric learning methods. We will also explore possible future research directions and novel applications of similarity measures in recommender systems.

Contents

Similarity Search and Retrieval

Fast Locality-Sensitive Hashing Frameworks for Approximate Near Neighbor Search	3
<i>Tobias Christiani</i>	
Storing Data Once in M-tree and PM-tree	18
<i>Humberto Razente and Maria Camila Nardini Barioni</i>	
Index Maintenance Strategy and Cost Model for Extended Cluster Pruning	32
<i>Anders Munck Højsgaard, Björn Þór Jónsson, and Philippe Bonnet</i>	
SPLX-Perm: A Novel Permutation-Based Representation for Approximate Metric Search	40
<i>Lucia Vadicamo, Richard Connor, Fabrizio Falchi, Claudio Gennaro, and Fausto Rabitti</i>	
Fast and Exact Nearest Neighbor Search in Hamming Space on Full-Text Search Engines	49
<i>Cun (Matthew) Mu, Jun (Raymond) Zhao, Guang Yang, Binwei Yang, and Zheng (John) Yan</i>	
k-Distance Approximation for Memory-Efficient RkNN Retrieval	57
<i>Max Berrendorf, Felix Borutta, and Peer Kröger</i>	
Pruning Algorithms for Low-Dimensional Non-metric k-NN Search: A Case Study	72
<i>Leonid Boytsov and Eric Nyberg</i>	
Non-metric Similarity Search Using Genetic TriGen	86
<i>David Bernhauer and Tomáš Skopal</i>	
Privacy-Preserving Text Similarity via Non-Prefix-Free Codes	94
<i>M. Oğuzhan Külekci, Ismail Habib, and Amir Aghabaiglou</i>	
Explainable Similarity of Datasets Using Knowledge Graph	103
<i>Petr Škoda, Jakub Klímek, Martin Nečaský, and Tomáš Skopal</i>	

The Curse of Dimensionality

The Role of Local Intrinsic Dimensionality in Benchmarking Nearest Neighbor Search	113
<i>Martin Aumüller and Matteo Ceccarelo</i>	

Accurate and Fast Retrieval for Complex Non-metric Data via Neighborhood Graphs 128
Leonid Boytsov and Eric Nyberg

Indexability-Based Dataset Partitioning 143
Angello Hoyos, Ubaldo Ruiz, Stephane Marchand-Maillet, and Edgar Chávez

Permutation’s Signatures for Proximity Searching in Metric Spaces 151
Karina Figueroa and Nora Reyes

A k -Skyband Approach for Feature Selection 160
Marcos Bedo, Paolo Ciaccia, Davide Martinenghi, and Daniel de Oliveira

Clustering and Outlier Detection

Faster k -Medoids Clustering: Improving the PAM, CLARA, and CLARANS Algorithms 171
Erich Schubert and Peter J. Rousseeuw

MORE++: k -Means Based Outlier Removal on High-Dimensional Data 188
Anna Beer, Jennifer Lauterbach, and Thomas Seidl

A Generic Summary Structure for Arbitrarily Oriented Subspace Clustering in Data Streams 203
Felix Borutta, Peer Kröger, and Thomas Hubauer

Similarity Grouping in Big Data Systems 212
Yasin N. Silva, Manuel Sandoval, Diana Prado, Xavier Wallace, and Chuitian Rong

SIDEKICK: Linear Correlation Clustering with Supervised Background Knowledge 221
Maximilian Archimedes Xaver Hünemörder, Daniyal Kazempour, Peer Kröger, and Thomas Seidl

Subspaces and Embeddings

Query Filtering with Low-Dimensional Local Embeddings 233
Edgar Chávez, Richard Connor, and Lucia Vadicamo

Characteristics of Local Intrinsic Dimensionality (LID) in Subspaces: Local Neighbourhood Analysis 247
Tahrira Hashem, Lida Rashidi, James Bailey, and Lars Kulik

Metric Embedding into the Hamming Space with the n-Simplex Projection	265
<i>Lucia Vadicamo, Vladimir Mic, Fabrizio Falchi, and Pavel Zezula</i>	
On coMADs and Principal Component Analysis	273
<i>Daniyal Kazempour, M. A. X. Hünemörder, and Thomas Seidl</i>	
Subspace Determination Through Local Intrinsic Dimensional Decomposition	281
<i>Ruben Becker, Imane Hafnaoui, Michael E. Houle, Pan Li, and Arthur Zimek</i>	
Applications	
Leveraging Feature Similarity for Earlier Detection of Unwanted Feature Interactions in Evolving Software Product Lines	293
<i>Seyedehzahra Khoshmanesh and Robyn R. Lutz</i>	
Protein Complex Similarity Based on Weisfeiler-Lehman Labeling	308
<i>Bianca K. Stöcker, Till Schäfer, Petra Mutzel, Johannes Köster, Nils Kriege, and Sven Rahmann</i>	
Multiple Instance Classification in the Image Domain	323
<i>Ilaria Bartolini, Pietro Pascarella, and Marco Patella</i>	
An Image Retrieval System for Video	332
<i>Paolo Bolettieri, Fabio Carrara, Franca Debole, Fabrizio Falchi, Claudio Gennaro, Lucia Vadicamo, and Claudio Vairo</i>	
Towards Automatic Configuration of Interactive Known-Item Search Systems	340
<i>Ladislav Peška, Gregor Kovalčík, and Jakub Lokoč</i>	
Doctoral Symposium Papers	
ADAMiSS: Advanced Data Analysis, Mining and Search, System	351
<i>Jakub Peschel and Pavel Zezula</i>	
Feature Similarity: A Method to Detect Unwanted Feature Interactions Earlier in Software Product Lines	356
<i>Seyedehzahra Khoshmanesh and Robyn R. Lutz</i>	
Author Index	363