Lecture Notes in Computer Science 9939

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

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

Laurent Amsaleg · Michael E. Houle
Erich Schubert (Eds.)

Similarity Search and Applications

9th International Conference, SISAP 2016
Tokyo, Japan, October 24–26, 2016
Proceedings

Springer

Editors
Laurent Amsaleg
CNRS–IRISA
Rennes
France

Michael E. Houle
National Institute of Informatics
Tokyo
Japan

Erich Schubert
Ludwig-Maximilians-Universität München
Munich
Germany

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-319-46758-0 ISBN 978-3-319-46759-7 (eBook)
DOI 10.1007/978-3-319-46759-7

Library of Congress Control Number: 2016954121

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

Printed on acid-free paper

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

Preface

This volume contains the papers presented at the 9th International Conference on Similarity Search and Applications (SISAP 2016) held in Tokyo, Japan, during October 24–26, 2016. 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 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papers welcomed full papers, short papers, as well as demonstration papers, with all manuscripts presenting previously unpublished research contributions. At SISAP 2016, all contributions were presented both orally and in a poster session, which facilitated fruitful exchanges between the participants.

We received 47 submissions, 32 full papers and 15 short papers, from authors based in 21 different countries. The Program Committee (PC) was composed of 62 members from 26 countries. Reviews were thoroughly discussed by the chairs and PC members: each submission received at least three to five reviews, with additional reviews sometimes being sought in order to achieve a consensus. The PC was assisted by 23 external reviewers.

The final selection of papers was made by the PC chairs based on the reviews received for each submission as well as the subsequent discussions among PC members. The final conference program consisted of 18 full papers and seven short papers, resulting in an acceptance rate of 38 % for full papers and 53 % cumulative for full and short papers.

The proceedings of SISAP are published by Springer as a volume in the Lecture Notes in Computer Science (LNCS) series. For SISAP 2016, as in previous years, extended versions of five 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.

The conference program and the proceedings are organized in several parts. As a first part, the program includes three keynote presentations from exceptionally skilled scientists: Alexandr Andoni, from Columbia University, USA, on the topic of "Data-Dependent Hashing for Similarity Search"; Takashi Washio, from the University of Osaka, Japan, on "Defying the Gravity of Learning Curves: Are More Samples

Better for Nearest Neighbor Anomaly Detectors?"; and Zhi-Hua Zhou, from Nanjing University, China, on "Partial Similarity Match with Multi-instance Multi-label Learning".

The program then carries on with the presentations of the papers, grouped in eight categories: graphs and networks; metric and permutation-based indexing; multimedia; text and document similarity; comparisons and benchmarks; hashing techniques; time-evolving data; and scalable similarity search.

We would like to thank all the authors who submitted papers to SISAP 2016. 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 express our gratitude to the members of the Organizing Committee for the enormous amount of work they have done.

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

August 2016

Laurent Amsaleg
Michael E. Houle
Erich Schubert

Organization

Program Committee Chairs

Laurent Amsaleg	CNRS-IRISA, France
Michael E. Houle	National Institute of Informatics, Japan

Program Committee Members

Giuseppe Amato	ISTI-CNR, Italy
Laurent Amsaleg	CNRS-IRISA, France
Hiroki Arimura	Hokkaido University, Japan
Ira Assent	Aarhus University, Denmark
James Bailey	University of Melbourne, Australia
Christian Beecks	RWTH Aachen University, Germany
Panagiotis Bouros	Aarhus University, Denmark
Leonid Boytsov	Carnegie Mellon University, USA
Benjamin Bustos	University of Chile, Chile
K. Selçuk Candan	Arizona State University, USA
Guang-Ho Cha	Seoul National University of Science and Technology, Korea
Edgar Chávez	CICESE, Mexico
Paolo Ciaccia	University of Bologna, Italy
Richard Connor	University of Strathclyde, UK
Michel Crucianu	CNAM, France
Bin Cui	Peking University, China
Vlad Estivill-Castro	Griffith University, Australia
Andrea Esuli	ISTI-CNR, Italy
Fabrizio Falchi	ISTI-CNR, Italy
Claudio Gennaro	ISTI-CNR, Italy
Magnus Lie Hetland	NTNU, Norway
Michael E. Houle	National Institute of Informatics, Japan
Yoshiharu Ishikawa	Nagoya University, Japan
Björn Þór Jónsson	Reykjavik University, Iceland
Ata Kabán	University of Birmingham, UK
Ken-ichi Kawarabayashi	National Institute of Informatics, Japan
Daniel Keim	University of Konstanz, Germany
Yiannis Kompatsiaris	CERTH – ITI, Greece
Peer Kröger	Ludwig-Maximilians-Universität München, Germany
Guoliang Li	Tsinghua University, China
Jakub Lokoč	Charles University in Prague, Czech Republic

Rui Mao	Shenzhen University, China
Stéphane Marchand-Maillet	Viper Group - University of Geneva, Switzerland
Henning Müller	HES-SO, Switzerland
Gonzalo Navarro	University of Chile, Chile
Chong-Wah Ngo	City University of Hong Kong, SAR China
Beng Chin Ooi	National University of Singapore, Singapore
Vincent Oria	New Jersey Institute of Technology, USA
M. Tamer Özsu	University of Waterloo, Canada
Deepak P	IBM Research, India
Apostolos N. Papadopoulos	Aristotle University of Thessaloniki, Greece
Marco Patella	DEIS – University of Bologna, Italy
Oscar Pedreira	Universidade da Coruña, Spain
Miloš Radovanović	University of Novi Sad, Serbia
Kunihiko Sadakane	The University of Tokyo, Japan
Shin'ichi Satoh	National Institute of Informatics, Japan
Erich Schubert	Ludwig-Maximilians-Universität München, Germany
Tetsuo Shibuya	Human Genome Center, Institute of Medical Science, The University of Tokyo, Japan
Yasin Silva	Arizona State University, USA
Matthew Skala	IT University of Copenhagen, Denmark
John Smith	IBM T.J. Watson Research Center, USA
Nenad Tomašev	Google, UK
Agma Traina	University of São Paulo at São Carlos, Brazil
Takeaki Uno	National Institute of Informatics, Japan
Michel Verleysen	Université Catholique de Louvain, Belgium
Takashi Washio	ISIR, Osaka University, Japan
Marcel Worring	University of Amsterdam, The Netherlands
Pavel Zezula	Masaryk University, Czech Republic
De-Chuan Zhan	Nanjing University, China
Zhi-Hua Zhou	Nanjing University, China
Arthur Zimek	Ludwig-Maximilians-Universität München, Germany
Andreas Züfle	George Mason University, USA

Additional Reviewers

Tetsuya Araki
Konstantinos Avgerinakis
Nicolas Basset
Michal Batko
Jessica Beltran
Hei Chan
Elisavet Chatzilari
Anh Dinh
Alceu Ferraz Costa
Karina Figueroa
David Novak
Ninh Pham
Nora Reyes
José Fernando Rodrigues Jr
Ubaldo Ruiz
Manos Schinas
Pascal Schweitzer
Diego Seco
Francesco Silvestri
Eleftherios Spyromitros-Xioufis
Eric S. Tellez
Xiaofei Zhang
Yue Zhu

Keynotes

Data-Dependent Hashing for Similarity Search

Alexandr Andoni

Columbia University, New York, USA

The quest for efficient similarity search algorithms has lead to a number of ideas that proved successful in both theory and practice. Yet, the last decade or so has seen a growing gap between the theoretical and practical approaches. On the one hand, most successful theoretical methods rely on data-indepependent hashing, such as the classic Locality Sensitive Hashing scheme. These methods have provable guarantees on correctness and performance. On the other hand, in practice, methods that adapt to the given datasets, such as the PCA-tree, often outperform the former, but provide no guarantees on performance or correctness.

This talk will survey the recent efforts to bridge this gap between theoretical and practical methods for similarity search. We will see that data-dependent methods are provably better than data-independent methods, giving, for instance, the first improvements over the Locality Sensitive Hashing schemes for the Hamming and Euclidean spaces.

Defying the Gravity of Learning Curves: Are More Samples Better for Nearest Neighbor Anomaly Detectors?

Takashi Washio

Osaka University, Suita, Japan

Machine learning algorithms are conventionally considered to provide higher accuracy when more data are used for their training. We call this behavior of their learning curves “the gravity”, and it is believed that no learning algorithms are “gravity-defiant”. A few scholars recently suggested that some unsupervised anomaly detector ensembles follow the gravity defiant learning curves. One explained this behavior in terms of the sensitivity of the expected k-nearest neighbor distances to the data density. Another discussed the former's incorrect reasoning, and demonstrated the possibilities of both gravity-compliance and gravity-defiant behaviors by applying the statistical bias-variance analysis. However, the bias-variance analysis for density estimation error is not an appropriate tool for anomaly detection error. In this talk, we argue that the analysis must be based on the anomaly detection error, and clarify the mechanism of the gravity-defiant learning curves of the nearest neighbor anomaly detectors by applying analysis based on computational geometry to the anomaly detection error. This talk is based on collaborative work with Kai Ming Ting, Jonathan R. Wells, and Sunil Aryal from Federation University, Australia.

Partial Similarity Match with Multi-Instance Multi-Label Learning

Zhi-Hua Zhou

Nanjing University, Nanjing, China

In traditional supervised learning settings, a data object is usually represented by a single feature vector, called an instance. Such a formulation has achieved great success; however, its utility is limited when handling data objects with complex semantics where one object simultaneously belongs to multiple semantic categories. For example, an image showing a lion besides an elephant can be recognized simultaneously as an image of a lion, an elephant, "wild" or even "Africa"; the text document "Around the World in Eighty Days" can be classified simultaneously into multiple categories such as scientific novel, Jules Verne's writings or even books on traveling, etc. In many real tasks it is crucial to tackle such data objects, particularly when the labels are relevant to partial similarity match of input patterns. In this talk we will introduce the MIML (Multi-Instance Multi-Label learning) framework which has been shown to be useful for these scenarios.

Contents

Graphs and Networks

BFST_ED: A Novel Upper Bound Computation Framework for the Graph Edit Distance 3
Karam Gouda, Mona Arafa, and Toon Calders

Pruned Bi-directed K-nearest Neighbor Graph for Proximity Search 20
Masajiro Iwasaki

A Free Energy Foundation of Semantic Similarity in Automata and Languages 34
Cewei Cui and Zhe Dang

Metric and Permutation-Based Indexing

Supermetric Search with the Four-Point Property 51
Richard Connor, Lucia Vadicamo, Franco Alberto Cardillo, and Fausto Rabitti

Reference Point Hyperplane Trees. 65
Richard Connor

Quantifying the Invariance and Robustness of Permutation-Based Indexing Schemes 79
Stéphane Marchand-Maillet, Edgar Roman-Rangel, Hisham Mohamed, and Frank Nielsen

Deep Permutations: Deep Convolutional Neural Networks and Permutation-Based Indexing 93
Giuseppe Amato, Fabrizio Falchi, Claudio Gennaro, and Lucia Vadicamo

Multimedia

Patch Matching with Polynomial Exponential Families and Projective Divergences. 109
Frank Nielsen and Richard Nock

Known-Item Search in Video Databases with Textual Queries 117
Adam Blažek, David Kuboň, and Jakub Lokoč

Combustion Quality Estimation in Carbonization Furnace Using Flame Similarity Measure 125
Fredy Martínez, Angelica Rendón, and Pedro Guevara

Text and Document Similarity

Bit-Vector Search Filtering with Application to a Kanji Dictionary 137
Matthew Skala

Domain Graph for Sentence Similarity 151
Fumito Konaka and Takao Miura

Context Semantic Analysis: A Knowledge-Based Technique for Computing Inter-document Similarity 164
Fabio Benedetti, Domenico Beneventano, and Sonia Bergamaschi

Comparisons and Benchmarks

An Experimental Survey of MapReduce-Based Similarity Joins 181
Yasin N. Silva, Jason Reed, Kyle Brown, Adelbert Wadsworth, and Chuitian Rong

YFCC100M-HNfc6: A Large-Scale Deep Features Benchmark for Similarity Search 196
Giuseppe Amato, Fabrizio Falchi, Claudio Gennaro, and Fausto Rabitti

A Tale of Four Metrics 210
Richard Connor

Hashing Techniques

Fast Approximate Furthest Neighbors with Data-Dependent Candidate Selection 221
Ryan R. Curtin and Andrew B. Gardner

NearBucket-LSH: Efficient Similarity Search in P2P Networks 236
Naama Kraus, David Carmel, Idit Keidar, and Meni Orenbach

Speeding up Similarity Search by Sketches 250
Vladimir Mic, David Novak, and Pavel Zezula

Fast Hilbert Sort Algorithm Without Using Hilbert Indices 259
Yasunobu Imamura, Takeshi Shinohara, Kouichi Hirata, and Tetsuji Kuboyama

Time-Evolving Data

Similarity Searching in Long Sequences of Motion Capture Data 271
Jan Sedmidubsky, Petr Elias, and Pavel Zezula

Music Outlier Detection Using Multiple Sequence Alignment and Independent Ensembles 286
Dimitrios Bountouridis, Hendrik Vincent Koops, Frans Wiering, and Remco C. Veltkamp

Scalable Similarity Search in Seismology: A New Approach to Large-Scale Earthquake Detection 301
Karianne Bergen, Clara Yoon, and Gregory C. Beroza

Scalable Similarity Search

Feature Extraction and Malware Detection on Large HTTPS Data Using MapReduce 311
Přemysl Čech, Jan Kohout, Jakub Lokoč, Tomáš Komárek, Jakub Maroušek, and Tomáš Pevný

Similarity Search of Sparse Histograms on GPU Architecture 325
Hasmik Osipyan, Jakub Lokoč, and Stéphane Marchand-Maillet

Erratum to: Pruned Bi-directed K-nearest Neighbor Graph for Proximity Search E1
Masajiro Iwasaki

Author Index 339