Preface

Lecture Notes in Computer Science 8821

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

Agma Juci Machado Traina · Caetano Traina Jr.
Robson Leonardo Ferreira Cordeiro (Eds.)

Similarity Search and Applications

7th International Conference, SISAP 2014, Los Cabos, October, 29–31, 2014 Proceedings

Springer

Editors
Agma Juci Machado Traina
Computer Science Department - ICMC
University of São Paulo at São Carlos
São Carlos
Brazil

Robson Leonardo Ferreira Cordeiro
Computer Science Department - ICMC
University of São Paulo at São Carlos
São Carlos
Brazil

Caetano Traina Jr.
Computer Science Department - ICMC
University of São Paulo at São Carlos
São Carlos
Brazil

ISSN 0302-9743
ISBN 978-3-319-11987-8
DOI 10.1007/978-3-319-11988-5

ISSN 1611-3349 (electronic)
ISBN 978-3-319-11988-5 (eBook)

Library of Congress Control Number: 2014950507

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

Springer Cham Heidelberg New York Dordrecht London

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 seventh International Conference on Similarity Search and Applications (SISAP 2014), held at Los Cabos, Mexico, during October 29–31, 2014.

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 focuses on 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 distance-based searching, but in general the conference concerns both the effectiveness and efficiency aspects of any similarity search approach, welcoming contributions that range from theoretical aspects to innovative developments for which similarity search plays the central role.

The call for papers welcomed research papers (full or short papers) presenting previously unpublished research contributions, as well as case studies and application papers (short papers) describing existing applications of similarity search in real scenarios.

We received 45 complete submissions. The Program Committee (PC) comprised 53 researchers from 18 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 consensus had been reached. The final selection of papers was made by the PC chairs based on the reviews received for each submission. Finally, the conference program includes 21 full papers and 6 short papers, which results in a 46.66% acceptance ratio.

The conference program and the proceedings are organized into five parts. The first part comprises papers proposing improvements to different methods and techniques for similarity search. A second part is devoted to papers dealing with efficient indexing solutions for similarity search and their application in real settings. The third part focuses on particular metrics and their effectiveness. The fourth part of the conference program includes papers dealing with new scenarios or presenting new approaches to similarity search. Finally, the last part comprises those papers devoted to solutions for similarity search in specific application domains, such as in streaming time series, image and audio retrieval and analysis, systems with CPU- and GPU-based processing, astroinformatics, computational neuroscience, and in particular types of recommender systems and search engines.

The conference program also includes two invited talks from outstanding scholars in the field. The first one, "Scalable Retrieval and Analysis of Simulation and Observation Data Sets" by Prof. K. Selçuk Candan, introduces and

presents solutions to computational challenges that arise from the need to process, index, search, and analyze, in a scalable manner, large volumes of temporal data resulting from data-intensive simulations. The second one, "Visual Analytics for Interactive Subspace Similarity Search" by Prof. Daniel Keim, presents novel techniques that combine automated and visual methods to improve subspace search in high-dimensional data.

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 thank all the authors who submitted papers to SISAP 2014. We would also like to thank all members of the PC and the external reviewers, for the enormous amount of work they have done. We would like to acknowledge the generous collaboration and financial support from Centro de Investigación Científica y de Educación Superior de Ensenada, B.C. (CICESE); the host institution, and from the Consejo Nacional de Ciencia y Tecnología (CONACyT); the Mexican public research agency. 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 2014

Agma Juci Machado Traina
Caetano Traina Jr.
Robson Leonardo Ferreira Cordeiro

Organization

Program Committee Chairs

Agma Juci Machado Traina	University of São Paulo, Brazil
Caetano Traina Jr.	University of São Paulo, Brazil

Program Committee Members

Agma Juci Machado Traina	University of São Paulo, Brazil
Ahmet Sacan	Drexel University, USA
Alberto Del Bimbo	Università degli Studi di Firenze, Italy
Alberto Laender	Federal University of Minas Gerais, Brazil
Altigran S. da Silva	Federal University of Amazonas, Brazil
Ambuj Singh	University of California at Santa Barbara, USA
Andre Balan	Federal University of ABC, Brazil
Andreas Zuefle	Ludwig-Maximilians-Universität München, Germany
Apostolos N. Papadopoulos	Aristotle University of Thessaloniki, Greece
Benjamin Bustos	University of Chile, Chile
Bjorn Thor Jonsson	Reykjavik University, Iceland
Caetano Traina Jr.	University of São Paulo, Brazil
Claudio Gennaro	ISTI-CNR, Italy
Christian Böhm	Ludwig-Maximilians-Universität München, Germany
Daniel Keim	University of Konstanz, Germany
Dimitrios Gunopulos	University of Athens, Greece
Divesh Srivastava	AT&T Labs-Research, USA
Dong Deng	Tsinghua University, China
Eamonn Keogh	University of California at Riverside, USA
Edgar Chavez	Universidad Nacional Autónoma de México, Mexico
Eduardo Valle	University of Campinas, Brazil
Elaine Parros Machado de Sousa	University of São Paulo, Brazil
Fabrizio Falchi	ISTI-CNR, Italy
Giuseppe Amato	ISTI-CNR, Italy
Gonzalo Navarro	University of Chile, Chile
Hanghang Tong	City College, CUNY, USA
Henning Müller	HES-SO, Switzerland
Jimeng Sun	Georgia Institute of Technology, USA

Joao Eduardo Ferreira	University of São Paulo, Brazil
Joe Tekli	Lebanese American University, Lebanon
Jose Oncina	University of Alicante, Spain
Luisa Mico	University of Alicante, Spain
Marcela Ribeiro	Federal University of São Carlos – UFSCar, Brazil
Marco Patella	University of Bologna, Italy
Nieves R. Brisaboa	University of A Coruña, Spain
Panagiotis Bouros	Humboldt-Universität zu Berlin, Germany
Paolo Ciaccia	University of Bologna, Italy
Pavel Zezula	Masaryk University, Czech Republic
Oscar Pedreira	University of A Coruña, Spain
Renata Galante	Federal University of Rio Grande do Sul, Brazil
Renato Fileto	Federal University of Santa Catarina, Brazil
Richard Connor	University of Strathclyde, UK
Richard Chbeir	IUT de Bayonne et du Pays Basque, France
Robson Leonardo Ferreira Cordeiro	University of São Paulo, Brazil
Rui Zhang	University of Melbourne, Australia
Simone Santini	Universidad Autómoma de Madrid, Spain
Thomas Seidl	RWTH Aachen University, Germany
Tomas Skopal	Charles University in Prague, Czech Republic
Vassilis Tsotras	University of California at Riverside, USA
Vincent Oria	NJIT, USA
Vladimir Pestov	University of Ottawa, Canada
Yasin Silva	Arizona State University, USA
Yoshiharu Ishikawa	Nagoya University, Japan

Additional Reviewers

Amelkin, Victor
Araujo, Samur
Bartoli, Federico
Bartolini, Ilaria
Calvo-Zaragoza, Jorge
Ercoli, Simone
Hoang, Minh
Huang, Jin
Ma, Xiguo
Marvulle, Valdecir
Prati, Ronaldo
Qi, Jianzhong
Sun, Jichao
Taddesse, Fekade Getahun
Tellez, Eric Sadit
Turchini, Francesco

Invited Talks (Abstracts)

Scalable Retrieval and Analysis of Simulation and Observation Data Sets*

K. Selçuk Candan

Professor of Computer Science and Engineering
Arizona State University

Abstract. Data- and model-driven computer simulations for under- standing spatio-temporal dynamics of emerging phenomena are increasingly critical in various application domains, from predicting geo-temporal evolution of epidemics to helping reduce energy footprints of buildings leading to more sustainable building systems and architectural designs. These simulations track 10s or 100s of inter-dependent parameters, spanning multiple information layers and spatio-temporal frames, affected by complex dynamic processes operating at different resolutions. Consequently, the key characteristics of data sets and models relevant to these data-intensive simulations often include the following: (a) voluminous, (b) multi-variate, (c) multi-resolution, (d) spatio-temporal, and (e) inter-dependent. While very powerful and highly modular and flexible simulation software exists, because of the volume and complexity of the simulation data, the varying spatial and temporal scales at which the key transmission processes operate and relevant observations are made, today experts lack the means to adequately and systematically interpret observations, understand the underlying processes, and re-use of existing simulation results in new settings. In this talk, I will introduce computational challenges that arise from the need to process, index, search, and analyze, in a scalable manner, large volumes of temporal data resulting from data-intensive simulations and present some solutions.

Keywords: Time series, simulations, feature extration, analysis, indexing

* This work is partially funded by NSF grants #1339835 ("E-SDMS: Energy Simulation Data Management System Software"), #1318788 ("Data Management for Real-Time Data Driven Epidemic Spread Simulations"), #116394 ("RanKloud: Data Partitioning and Resource Allocation Strategies for Scalable Multimedia and Social Media Analysis"), #1016921 ("One Size Does Not Fit All: Empowering the User with User-Driven Integration"), and #1430144 ("Fraud Detection via Visual Analytics: An Infrastructure to Support Complex Financial Patterns (CFP)-based Real-Time Services Delivery"). This work is also supported in part by the NSF I/UCRC Center for Embedded Systems established through the NSF grant #0856090 in partnership with Johnson Controls Inc.

References

1. Candan, K.S., Rossini, R., Sapino, M.L., Wang, X.: SDTW: Computing DTW Distances using Locally Relevant Constraints based on Salient Feature Alignments. PVLDB 5(11), 1519–1530 (2012)
2. Candan, K.S., Rossini, R., Sapino, M.L., Wang, X.: STFMap: Query- and Feature-Driven Visualization of Large Time Series Data Sets. CIKM 2012, 2743–2745 (2012)
3. Chen, X., Candan, K.S.: LWI-SVD: Low-rank, Windowed, Incremental Singular Value Decompositions on Time-Evolving Data Sets. In: Accepted for Publication at the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD (2014)
4. Chen, X., Candan, K.S.: GI-NMF: Group Incremental Non-Negative Matrix Factorization on Data Streams. In: Accepted for Publication at the ACM International Conference on Conference on Information and Knowledge Management, CIKM (2014)
5. Huang, S., Li, X., Candan, K.S., Sapino, M.L.: Can you really trust that seed?": Reducing the Impact of Seed Noise in Personalized PageRank. In: Accepted for Publication at the International Conference on Advances in Social Network Analysis and Mining, ASONAM (2014)
6. Kim, M., Candan, K.S.: Efficient Static and Dynamic In-Database Tensor Decompositions on Chunk-Based Array Stores. In: Accepted for Publication at the ACM International Conference on Conference on Information and Knowledge Management, CIKM (2014)
7. Kim, M., Candan, K.S.: TensorDB: In-Database Tensor Manipulation with Tensor-Relational Query Plans. In: Accepted for Demonstration at the ACM International Conference on Conference on Information and Knowledge Management, CIKM (2014)
8. Kim, M., Selçuk Candan, K.: Pushing-down tensor decompositions over unions to promote reuse of materialized decompositions. In: Calders, T., Esposito, F., Hüllermeier, E., Meo, R. (eds.) ECML PKDD 2014, Part I. LNCS, vol. 8724, pp. 688–704. Springer, Heidelberg (2014)
9. Li, X., Huang, S., Candan, K.S., Sapino, M.L.: Focusing Decomposition Accuracy by Personalizing Tensor Decomposition (PTD). In: Accepted for Publication at the ACM International Conference on Conference on Information and Knowledge Management, CIKM (2014)
10. Nagendra, M., Candan, K.S.: SkySuite: A Framework of Skyline-Join Operators for Static and Stream Environments. In: Proceedings of the VLDB Endowment (PVLDB), vol. 6(12) (2013)
11. Nagendra, M., Candan, K.S.: Layered processing of skyline-window- join (SWJ) queries using iteration-fabric. In: IEEE International Conference on Data Engineering (ICDE), pp. 985–996 (2013)
12. Nagarkar, P., Candan, K.S.: HCS: Hierarchical Cut Selection for Efficiently Processing Queries on Data Columns using Hierarchical Bitmap Indices. In: International Conference on Extending Database Technology (EDBT), pp. 271–282 (2014)
13. Schifanella, C., Sapino, M.L., Candan, K.S.: On Context-Aware Co-Clustering with Metadata Support. J. Intell. Inf. Syst. 38(1), 209–239 (2012)
14. Wang, X., Candan, K.S., Sapino, M.L.: Leveraging Metadata for Identifying Local, Robust Multi-variate Temporal (RMT) Features. In: IEEE International Conference on Data Engineering (ICDE), pp. 388–399.

Visual Analytics for Interactive Subspace Similarity Search

Daniel Keim

Head of the Information Visualization and Data Analysis Research Group,
University of Konstanz, Germany

Abstract. In most similarity search applications, the data under consideration resides in high-dimensional data spaces, which often consist of combined features measuring different properties. In order to determine useful similarity measures, appropriate feature combinations (subspaces) of the data have to be taken into consideration, since they may show complementary, conjoint, or contradicting relations between the data items [3]. Which subspace is best in a given application context is difficult to determine by fully automatic methods, and therefore it is important to include the human in the process and combine the creativity and general knowledge of the human with the fast searching and analysis capabilities of the computer. Visual Analytics – the combination of automated and visual methods – can help to interactively determine the most relevant subspaces and define appropriate subspace similarity measures [4]. Subspace search algorithms guided by interestingness measures can be used to compute candidate sets of subspaces, which are then visualized to enable the user to compare and relate subspaces with respect to the involved dimensions and clusters of objects [1]. The approach helps the understanding of high-dimensional data from different perspectives and allows a flexible definition of subspace similarity measures [2].

Keywords: Visual Analytics, Interactive Similarity Search, Subspace Similarity, Interestingness Measures

References

1. Bertini, E., Tatu, A., Keim, D.: Quality metrics in high-dimensional data visualization:An overview and systematization. IEEE Transactions on Visualization and Computer Graphics 17(12), 2203–2212 (2011)
2. Tatu, A., Albuquerque, G., Eisemann, M., Bak, P., Theisel, H., Magnor, M., Keim, D.: Automated analytical methods to support visual exploration of high-dimensional data. IEEE Transactions on Visualization and Computer Graphics 17(5), 584–597 (2011)
3. Tatu, A., Maaß, F., Färber, I., Bertini, E., Schreck, T., Seidl, T., Keim, D.: Subspace search and visualization to make sense of alternative clusterings in high-dimensional data. In: 2012 IEEE Conference on Visual Analytics Science and Technology (VAST), pp. 63–72 (October 2012)

4. Tatu, A., Zhang, L., Bertini, E., Schreck, T., Keim, D., Bremm, S., von Landesberger, T.: ClustNails: Visual analysis of subspace clusters. Tsinghua Science and Technology 17(4), 419–428 (2012)

Contents

Improving Similarity Search Methods and Techniques

Indexing and Applications

Metrics and Evaluation

New Scenarios and Approaches

Applications and Specific Domains