Communications in Computer and Information Science 1577

More information about this series at https://link.springer.com/bookseries/7899

Juan Antonio Lossio-Ventura ·
Jorge Valverde-Rebaza · Eduardo Díaz ·
Denisse Muñante · Carlos Gavidia-Calderon ·
Alan Demétrius Baria Valejo ·
Hugo Alatrista-Salas (Eds.)

Information Management and Big Data

8th Annual International Conference, SIMBig 2021
Virtual Event, December 1–3, 2021
Proceedings

Springer

Editors
Juan Antonio Lossio-Ventura
National Institutes of Health
Bethesda, MD, USA

Jorge Valverde-Rebaza
Visibilia
São Paulo, Brazil

Eduardo Díaz
Universidad Peruana de Ciencias Aplicadas
Lima, Peru

Denisse Muñante
ENSIIE and SAMOVAR
Evry, France

Carlos Gavidia-Calderon
The Open University
Milton Keynes, UK

Alan Demétrius Baria Valejo
Federal University of São Carlos
São Carlos, Brazil

Hugo Alatrista-Salas
University of Engineering and Technology
UTEC
Lima, Peru

ISSN 1865-0929 ISSN 1865-0937 (electronic)
Communications in Computer and Information Science
ISBN 978-3-031-04446-5 ISBN 978-3-031-04447-2 (eBook)
https://doi.org/10.1007/978-3-031-04447-2

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

Preface

SIMBig 2021[1], the 8th edition of the International Conference on Information Management and Big Data, presented novel methods for the analysis and management of large volumes of data, covering fields such as Artificial Intelligence (AI), Data Science, Machine Learning, Natural Language Processing, Semantic Web, Data-driven Software Engineering, Health Informatics, among others. The SIMBig conference series goal is to promote cooperation between national and international researchers to improve data-driven decision making by using new technologies dedicated to analyzing data.

SIMBig 2021 also encouraged proposals on the COVID-19 pandemic, with the goal of informing Latin American policy makers (specially in South America). SIMBig is a convivial place where participants present their scientific contributions in the form of full and short papers. This book contains the entire proceedings of the 8th edition of SIMBig, a full-virtual conference held between December 1 and December 3, 2021. For this edition, 23 long papers and four short papers were selected for publication, giving an acceptance rate of 40.3%.

Keynote Speakers' Resumes

SIMBig 2021 featured eight keynote speakers. Dr. Marinka Zitnik, from Harvard University, USA, presented her research group's work for infusing structure and knowledge into biomedical AI. Grand challenges in biology and medicine often lack annotated examples and require generalization to entirely new scenarios not seen during training. However, standard supervised learning is incredibly limited in scenarios such as designing novel medicines, modeling emerging pathogens, and treating rare diseases. In this talk, Dr. Zitnik presented their efforts to overcome these obstacles by infusing structure and knowledge into learning algorithms. First, Dr. Zitnik presented general-purpose and scalable algorithms for few-shot learning on graphs. At the core is the notion of local subgraphs that transfer knowledge from one task to another, even when only a handful of labeled examples are available. This principle is theoretically justified as the evidence for predictions can be found in subgraphs surrounding the targets. Dr. Zitnik concluded with applications in drug development and precision medicine where the algorithmic predictions were validated in human cells and led to the discovery of a new class of drugs.

Dr. Natasha Noy, from Google, USA, talked about the project Google Dataset Search that seeks to build an open ecosystem for dataset discovery. There are thousands of data repositories on the Web, providing access to millions of datasets. National and regional governments, scientific publishers and consortia, commercial data providers, and others publish data for fields ranging from social and life sciences through high-energy physics to climate science and more. Access to this data is critical to facilitating reproducibility of research results, enabling scientists to build on others' work, and providing data

[1] https://simbig.org/SIMBig2021/.

journalists easier access to information and its provenance. This talk gave an overview of Dataset Search by Google, which provides search capabilities over potentially all dataset repositories on the Web, and Dr. Noy discussed the open ecosystem for describing datasets that they hope to encourage.

Dr. Andrei Broder, from Google, USA, presented a point of view about the Web Advertising Ecosystem. The World Wide Web is arguably an engineering artifact and social environment that defines our era. A large part of it is made possible by money generated via advertising. The goal of Dr. Broder's talk was to give an introduction to the web advertising ecosystem and illuminate the complex relations between consumers, publishers, and advertisers.

Dr. Jiawei Han, from the University of Illinois Urbana-Champaign, USA, discussed their work to convert unstructured text data to structured knowledge. The real-world big data are largely dynamic, interconnected, and unstructured text. It is highly desirable to transform such massive unstructured data into structured knowledge. Many researchers rely on labor-intensive labeling and curation to extract knowledge from such data. Such approaches, however, are not scalable. We envision that massive text data itself may disclose a large body of hidden structures and knowledge. Equipped with pretrained language models and text embedding methods, it is promising to transform unstructured data into structured knowledge. In this talk, Dr. Han introduced a set of methods developed recently in his group for such an exploration, including joint spherical text embedding, discriminative topic mining, taxonomy construction, text classification, and taxonomy-guided text analysis. Dr. Han showed that a data-driven approach could be promising at transforming massive text data into structured knowledge.

Dr. Vipin Kumar, from the University of Minnesota, USA, focused on the opportunities and challenges for machine learning on topics related to big data in water. Water resources worldwide are coming under stress due to increasing demand from a growing population, increasing pollution, and depleting or uncertain supplies due to changing climate in which drought and floods have both become more frequent. As domains associated with water continue to experience tremendous data growth from models, sensors, and satellites, there is an unprecedented opportunity for machine learning to help address urgent water challenges facing humanity. Dr. Kumar's talk examined the role that big data and machine learning can play in advancing water science, challenges faced by traditional machine learning methods in addressing the domain of water, and some early successes.

Dr. Jian Pei, from Simon Fraser University, Canada, talked about relevant aspects such as interpretability, fairness, and marketplaces towards trustworthy data science. We believe data science and AI will change the world. No matter how smart and powerful an AI model we can build, the ultimate testimony of the success of data science and AI is users' trust. How can we build trustworthy data science? At the level of user-model interaction, how can we convince users that a data analytic result is trustworthy? At the level of group-wise collaboration for data science and AI, how can we ensure that the parties and their contributions are recognized fairly, and establish trust between the outcome (e.g., a model) of the group collaboration and the external users? At the level of data science participant eco-systems, how can we effectively and efficiently connect many participants of various roles and facilitate the connection between supplies and demands of data and models? Dr. Pei's talk brainstormed possible directions to the above questions in the context of an end-to-end data science pipeline. To strengthen trustworthy interactions between models and users, Dr. Pei advocated exact and consistent interpretation of machine learning models. His recent results showed that exact and consistent interpretations are not just theoretically feasible, but also practical even for API-based AI services. To build trust in collaboration among multiple participants in coalition, Dr. Pei reviewed some progress in ensuring fairness in federated learning, including fair assessment of contributions and fairness enforcement in collaboration outcomes. Last, to address the need of trustworthy data science eco-systems, Dr. Pei reviewed some latest efforts in building data and model marketplaces and preserving fairness and privacy. Through reflection Dr. Pei discussed some challenges and opportunities in building trustworthy data science for possible future work.

Dr. Francisco Pereira, from the National Institutes of Health (NIH), USA, presented his research aiming at revealing interpretable object representations from human behavior. Objects can be characterized according to a vast number of possible criteria (e.g. animacy, shape, color, function), but some dimensions are more useful than others for making sense of the objects around us. Dr. Pereira described an ongoing effort by his collaborators to collect a behavioral dataset of millions of odd-one-out similarity judgements on thousands of objects, and a new approach to identify the "core dimensions" of object representations used in those judgements. His approach models each object as a sparse, non-negative embedding, and judgements as a function of the similarity of those embeddings. The resulting model predicts subject behaviour on test data, as well as the fine-grained structure of object similarity. The dimensions of the embedding space are coherently interpretable by test subjects, and reflect degrees of taxonomic membership, functionality, and perceptual or structural attributes, among other characteristics. Further, naive subjects can accurately rate objects along these dimension, without training. Collectively, these results demonstrate that human similarity judgments can be captured by a fairly low-dimensional, interpretable embedding that generalizes to external behavior.

Dr. Jean Vanderdonckt, from the Université catholique de Louvain, Belgium, provided an overview of dimension reduction by model-based approaches and its application to gesture recognition. Machine learning algorithms used for 2D/3D gesture recognition typically require a large training set of templates having many dimensions, depending on the sensor used. Instead of applying classical methods for reducing the dimensionality of these templates, Dr. Vanderdonckt proposed relying on a model-based approach where the problem is first mathematically described and then submitted to machine learning algorithms.

Juan Antonio Lossio-Ventura
Jorge Valverde-Rebaza
Jorge Díaz
Denisse Muñante
Carlos Gavidia-Calderon
Alan Demétrius Baria Valejo
Hugo Alatrista-Salas

Organization

Organizing Committee

General Chairs

Juan Antonio Lossio-Ventura	National Institutes of Health (NIH), USA
Hugo Alatrista-Salas	Universidad de Ingeniería y Tecnología (UTEC), Peru

SNMAM Track Organizers

Jorge Valverde-Rebaza	Visibilia, Brazil
Alan Demétrius Baria Valejo	Federal University of São Carlos, Brazil
Thiago de Paulo Faleiros	University of Brasília, Brazil

DISE Track Organizers

Denisse Muñante	ENSIIE and SAMOVAR, France
Carlos Gavidia-Calderon	The Open University, UK
Jorge Díaz	Universidad Peruana de Ciencias Aplicadas, Peru

BIOMEDS Track Organizers

Pablo Fonseca	Universidad Peruana Cayetano Heredia, Peru
José Ferrer	Universidad Peruana Cayetano Heredia, Peru
Mabel K. Raza	Universidad Peruana Cayetano Heredia, Peru
Joseph Pinto	Auna, Peru

Local Organizers

Jesús Bellido-Angulo	Universidad de Ingeniería y Tecnología, Peru
Yamilet Serrano-Llerena	Universidad de Ingeniería y Tecnología, Peru
Cristhian Ganvini Valcarcel	Universidad Andina del Cusco, Peru
Armando Fermín Pérez	Universidad Nacional Mayor de San Marcos, Peru
José A. Herrera-Quispe	Universidad Nacional Mayor de San Marcos, Peru

Program Committee

SIMBig Program Committee

Asma Ben Abacha	National Institutes of Health, USA
Nathalie Abadie	COGIT, IGN, France
Amine Abdaoui	Huawei Research, France

Pedro Marco Achanccaray Diaz	Pontifical Catholic University of Rio de Janeiro, Brazil
Hugo Alatrista-Salas	Universidad de Ingeniería y Tecnología, Peru
Marco Alvarez	University of Rhode Island, USA
Yuan An	Drexel University, USA
Sophia Ananiadou	University of Manchester, UK
Erick Antezana	Norwegian University of Science and Technology, Norway
Ghislain Atemezing	Mondeca, France
John Atkinson	Universidad Adolfo Ibañez, Chile
Jérôme Azé	University of Montpellier, France
Imon Banerjee	Mayo Clinic, USA
Patrice Bellot	Aix-Marseille Université, France
Cesar Beltrán	Pontificia Universidad Católica del Perú, Peru
Farah Benamara	Université Paul Sabatier, France
Elena Beretta	Vrije Universiteit Amsterdam, The Netherlands
Giacomo Bergami	Newcastle University, UK
Jose David Bermudez Castro	Pontifical Catholic University of Rio de Janeiro, Brazil
Jiang Bian	University of Florida, USA
Selen Bozkurt	Flatiron Health, USA
Carmen Brando	École des hautes études en sciences sociales, France
Sandra Bringay	Paul Valéry University, France
Andrei Broder	Google, USA
Jean-Paul Calbimonte Pérez	University of Applied Sciences and Arts Western Switzerland, Switzerland
Hugo David Calderon Vilca	Universidad Nacional Mayor de San Marcos, Peru
Guillermo Calderón-Ruiz	Universidad Católica de Santa María, Peru
Mete Celik	Erciyes Üniversitesi, Turkey
Qingyu Chen	National Institutes of Health, USA
Davide Chicco	Peter Munk Cardiac Centre, Canada
Diego Collarana Vargas	Fraunhofer IAIS, Germany
Nelly Condori-Fernandez	Universidade da Coruña, Spain
Adrien Coulet	Inria, Paris, France
Bruno Cremilleux	Université de Caen Normandie, France
Fabio Crestani	University of Lugano, Switzerland
Gabriela Csurka	NAVER LABS Europe, France
Nuno Datia	ISEL, Portugal
Jessica de Faria Dafflon	National Institutes of Health, USA
Jos De Roo	Ghent University, Belgium
Dina Demner-Fushman	National Institutes of Health, USA

Bart Desmet	National Institutes of Health, USA
Damiano Distante	University of Rome Unitelma Sapienza, Italy
Marcos Aurélio Domingues	State University of Maringá, Brazil
Catherine Dominguès	IGN, France
Martín Ariel Domínguez	Universidad Nacional de Córdoba, Argentina
Alexandre Donizeti	UFABC, Brazil
Dejing Dou	University of Oregon, USA
Brett Drury	Liverpool Hope University, UK
Jocelyn Dunstan	University of Chile, Chile
Tome Eftimov	Jožef Stefan Institute, Slovenia
Okyaz Eminaga	Stanford University, USA
Almudena Espin Perez	Stanford Universiry, USA
Jacinto Estima	INESC-ID and Universidade Europeia, Portugal
Paula Estrella	Universidad Nacional de Córdoba, Argentina
Karina Figueroa Mora	Universidad Michoacana de San Nicolás de Hidalgo, Mexico
Frédéric Flouvat	University of New Caledonia, New Caledonia
Cidália Fonte	University of Coimbra and INESC Coimbra, Portugal
Markus Forsberg	University of Gothenburg, Sweden
Carlos Gavidia	Pontifica Universidad Católica del Perú, Peru
Andrew Gentles	Stanford University, USA
Olivier Gevaert	Stanford University, USA
Ayush Goyal	Texas A&M University-Kingsville, USA
Natalia Grabar	University of Lille, France
Irlán Grangel-González	Robert Bosch GmbH, Germany
Juan Gutierrez Alva	Universidad de Ingeniería y Tecnología, Peru
Adrien Guille	Université Lumière Lyon 2, France
Yi Guo	University of Florida, USA
Thomas Guyet	Agrocampus Ouest and IRISA, France
Zhe He	Florida State University, USA
Diana Inkpen	University of Ottawa, Canada
Clement Jonquet	University of Montpellier, France
Jennifer Joseph	Abbott, USA
Frank Dennis Julca-Aguilar	Algolux, Canada
Maulik Kamdar	Elsevier Inc., USA
Maria Keet	University of Cape Town, South Africa
Ravi Kumar	Google, USA
Ka Chun Lam	National Institutes of Health, USA
Juan Lazo Lazo	Universidad del Pacífico, Peru
Florence Le Ber	Université de Strasbourg and ENGEES, France
Cédric López	Emvista, France

Maysa Macedo	IBM Research, Brazil
Sanjay Kumar Madria	Missouri University of Science and Technology, USA
Sabrine Mallek	ICN Business School, France
Ricardo Marcacini	University of São Paulo, Brazil
Marcos Martínez-Romero	Stanford University, USA
Bruno Martins	INESC-ID, Portugal
Florent Masseglia	Inria, France
Patrick McClure	National Institutes of Health, USA
Rosario Medina Rodriguez	Pontificia Universidad Católica del Perú, Peru
Guillaume Metzler	Université Lumière Lyon 2, France
Claudio Miceli	Federal University of Rio de Janeiro, Brazil
Nandana Mihindukulasooriya	IBM Research, USA
Adam Miner	Stanford University, USA
André Miralles	Irstea, France
Antonio Miranda	Barcelona Supercomputing Center, Spain
Avdesh Mishra	Texas A&M University-Kingsville, USA
Pritam Mukherjee	Stanford University, USA
Denisse Muñante	ENSIIE and SAMOVAR, France
Mark Musen	Stanford University, USA
Behzad Naderalvojoud	Stanford University, USA
Nhung Nguyen	University of Manchester, UK
Mais Nijim	Texas A&M University-Kingsville, USA
Jordi Nin	Universitat Ramon Llull and ESADE, Spain
Miguel Nuñez-del-Prado-Cortéz	Universidad del Pacífico, Peru
José Eduardo Ochoa Luna	Universidad Católica San Pablo, Peru
Maciej Ogrodniczuk	Institute of Computer Science of the Polish Academy of Sciences, Poland
Matilde Pato	ISEL, Portugal
Jessica Pinaire	Kalya, France
Jorge Poco	Universidad Católica San Pablo, Peru
Pascal Poncelet	University of Montpellier, France
Mattia Prosperi	University of Florida, USA
Marcos Quiles	Federal University of São Paulo, Brazil
Julien Rabatel	Catholic University of Leuven, Belgium
Mabel Raza	Universidad Peruana Cayetano Heredia, Peru
José Fabián Reyes Román	Universitat Politècnica de València, Spain
Justine Reynaud	University of Normandie, France
Edgar Rios	Genentech, USA
Mathieu Roche	CIRAD, France
Nancy Rodriguez	University of Montpellier, France
Alejandro Rodríguez-Gonzales	Universidad Politécnica de Madrid, Spain

Edelweis Rohrer	Universidad de la República, Uruguay
Rafael Rossi	University of São Paulo, Brazil
Silvia Rueda Pascual	University of Valencia, Spain
Jose M. Saavedra	Orand, Chile
Arnaud Sallaberry	Paul Valéry University, France
Christian Sallaberry	UPPA, France
Shengtian Sang	Stanford University, USA
Rafael Santos	Instituto Nacional de Pesquisas Espaciais, Brazil
Edgar Sarmiento-Calisaya	Universidad Nacional de San Agustín, Peru
Lucile Sautot	AgroParisTech, France
José Segovia-Juárez	INIA, Peru
Nazha Selmaoui-Folcher	University of New Caledonia, New Caledonia
Matthew Shardlow	University of Manchester, UK
Pedro Nelson Shiguihara Juárez	Universidad San Ignacio de Loyola, Peru
Diego Silva	Universidade Federal de São Carlos, Brazil
João Simões	Cron.Studio, Portugal
Wenyu Song	Harvard University, USA
Aurea Rossy Soriano Vargas	University of Campinas, Brazil
Victor Stroele	Universidade Federal de Juiz de Fora, Brazil
Madhumita Sushil	University of California, San Francisco, USA
Silvia Lizeth Tapia Tarifa	University of Oslo, Norway
Maguelonne Teisseire	INRAE, France
Paul Thompson	University of Manchester, UK
Camilo Thorne	Elsevier, Germany
Sanju Tiwari	Universidad Autonoma de Tamaulipas, Mexico
Turki Turki	King Abdulaziz University, Saudi Arabia
Willy Ugarte	Universidad Peruana de Ciencias Aplicadas, Peru
Carlos Vázquez	École de Technologie Supérieure, Canada
Antonio Vetro	Politecnico di Torino, Italy
Maria-Esther Vidal	Universidad Simón Bolívar, Venezuela
Edwin Rafael Villanueva Talavera	Pontificia Universidad Católica del Perú, Peru
Boris Villazon-Terrazas	Fujitsu Laboratories of Europe, Spain
Sebastian Walter	Semalytix GmbH, Germany
Florence Wang	Telstra, Australia
Lana Yeganova	National Institutes of Health, USA
Sarah Zenasni	ICube and Strasbourg University, France
Chryssa Zerva	Instituto de Telecomunicações, Portugal
Yuan Zhao	National Institutes of Health, USA
Charles Zheng	National Institutes of Health, USA
Lu Zhou	Kansas State University, USA
Pierre Zweigenbaum	Université Paris-Saclay, CNRS, and LISN, France
Ivan Dimitry Zyrianoff	University of Bologna, Italy

SNMAM Program Committee

DISE Program Committee

Lenis Wong — Universidad Peruana de Ciencias Aplicadas, Peru
Glen Rodriguez — Universidad Nacional de Ingeniería, Peru
Jose Antonio Pow-Sang — Pontificia Universidad Catolica del Perú, Peru
Yudith Cardinale — Universidad Simón Bolívar, Peru
Manuel Munier — Université de Pau et Pays de l'Adour, France
Roxana Lisette Quintanilla Portugal — PUC-Rio, Brazil

Organizing Institutions and Sponsors

Organizing Institutions

Universidad de Ingeniería y Tecnología UTEC, Peru[2].
National Institutes of Health, USA[3].

Collaborating Institutions

Universidad Andina del Cusco, Peru[4].
Universidad Nacional Mayor de San Marcos, Peru[5].
Université de Montpellier, France[6].
Visibilia, Brazil[7].

[2] https://www.utec.edu.pe/en.
[3] https://www.nih.gov/.
[4] http://www.uandina.edu.pe/.
[5] http://www.unmsm.edu.pe/.
[6] https://www.umontpellier.fr/.
[7] http://visibilia.net.br.

Contents

Data-Driven Software Engineering

Health, NLP, and Social Media