

Communications in Computer and Information Science

656

Commenced Publication in 2007

Founding and Former Series Editors:

Alfredo Cuzzocrea, Dominik Ślęzak, and Xiaokang Yang

Editorial Board

Simone Diniz Junqueira Barbosa

*Pontifical Catholic University of Rio de Janeiro (PUC-Rio),
Rio de Janeiro, Brazil*

Phoebe Chen

La Trobe University, Melbourne, Australia

Xiaoyong Du

Renmin University of China, Beijing, China

Joaquim Filipe

Polytechnic Institute of Setúbal, Setúbal, Portugal

Orhun Kara

TÜBİTAK BİLGEM and Middle East Technical University, Ankara, Turkey

Igor Kotenko

*St. Petersburg Institute for Informatics and Automation of the Russian
Academy of Sciences, St. Petersburg, Russia*

Ting Liu

Harbin Institute of Technology (HIT), Harbin, China

Krishna M. Sivalingam

Indian Institute of Technology Madras, Chennai, India

Takashi Washio

Osaka University, Osaka, Japan

More information about this series at <http://www.springer.com/series/7899>

Juan Antonio Lossio-Ventura
Hugo Alatrasta-Salas (Eds.)

Information Management and Big Data

Second Annual International Symposium, SIMBig 2015
Cusco, Peru, September 2–4, 2015
and Third Annual International Symposium, SIMBig 2016
Cusco, Peru, September 1–3, 2016
Revised Selected Papers

Editors

Juan Antonio Lossio-Ventura
University of Florida
Gainesville, FL
USA

Hugo Alatrasta-Salas
Universidad del Pacífico
Jesús María, Lima
Peru

ISSN 1865-0929

ISSN 1865-0937 (electronic)

Communications in Computer and Information Science

ISBN 978-3-319-55208-8

ISBN 978-3-319-55209-5 (eBook)

DOI 10.1007/978-3-319-55209-5

Library of Congress Control Number: 2017933877

© Springer International Publishing AG 2017

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, express 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.

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

The aim of the SIMBig symposium is to present the analysis of methods for extracting knowledge from large volumes of data through techniques of data science and artificial intelligence. This book comprises extended versions of the best papers presented at SIMBig 2015 and SIMBig 2016.

Big data is a popular term used to describe the exponential growth and availability of data, which could be structured and unstructured. Data science is a field seeking to extract knowledge or insights from large volumes of heterogeneous data (e.g., video, audio, text, image). Data science is a continuation of other fields such as data analysis, statistics, machine learning, and data mining similar to knowledge discovery in databases (KDD).

Big data has taken place over the past 20 years. For instance, social networks such as Facebook, Twitter, and LinkedIn generate masses of data, which are available to be accessed by other applications. Several domains, including biomedicine, life sciences, and scientific research, have been affected by big data¹. Therefore, there is a need to understand and exploit these data. This process can be carried out thanks to “data science”, which is based on methodologies of data mining, natural language processing, Semantic Web, statistics, etc. That allows us to gain new insight through data-driven research [1, 4]. A major problem hampering big data analytics development is the need to process several types of data, such as structured, numeric, and unstructured data (e.g., video, audio, text, image, etc.)².

Our Annual International Symposium on Information Management and Big Data, seeks to present the new methods of data science and related fields for analyzing and managing large volumes of data. The symposium attracts many of the main national and international players in the decision-making field who are involved in new technologies dedicated to handling large amounts of information.

The third edition, SIMBig 2016³, was held in Cusco, Peru, during September 1–3, 2016. SIMBig 2016 has been indexed in DBLP⁴ [3] and in the CEUR Workshop Proceedings⁵. The second edition, SIMBig 2015⁶, was also held in Cusco, Peru, during September 2–4, 2015. SIMBig 2015 has been indexed in DBLP⁷ [2] and in the CEUR Workshop Proceedings⁸.

¹ By 2015 the average amount of data generated annually in hospitals is 665TB: <https://datafloq.com/read/body-source-big-data-infographic/413>.

² Today, 80% of data are unstructured such as images, video, and notes.

³ <http://simbig.org/SIMBig2016/>.

⁴ <http://dblp2.uni-trier.de/db/conf/simbig/simbig2016.html>.

⁵ <http://ceur-ws.org/Vol-1743/>.

⁶ <http://simbig.org/SIMBig2015/>.

⁷ <http://dblp2.uni-trier.de/db/conf/simbig/simbig2015.html>.

⁸ <http://ceur-ws.org/Vol-1478/>.

For this special proceedings volume, we accepted 11 long papers. These papers were selected from the SIMBig 2015 and SIMBig 2016 editions. We selected the four best papers from the 2015 edition, which had 32 submissions. Likewise, we selected the seven best papers of SIMBig 2016, from 42 submissions. Therefore, the general acceptance rate of this special proceedings volume was $11/(42 + 32) = 0.15$.

To share the new analysis methods for managing large volumes of data, we encouraged participation from researchers in all fields related to big data, data science, data mining, natural language processing, and the Semantic Web, but also multilingual text processing as well as biomedical NLP.

Topics of interest of SIMBig included: data science, big data, data mining, natural language processing, Bio-NLP, text mining, information retrieval, machine learning, Semantic Web, ontologies, Web mining, knowledge representation and linked open data, social networks, social Web, and Web science, information visualization, OLAP, data warehousing, business intelligence, spatiotemporal data, health care, agent-based systems, reasoning and logic, constraints, satisfiability, and search.

February 2017

Juan Antonio Lossio-Ventura
Hugo Alatrasta-Salas

References

1. David W Embley and Stephen W Liddle, *Big data—conceptual modeling to the rescue*, Conceptual Modeling, ER'13, LNCS, Springer, 2013, pp. 1–8.
2. Juan Antonio Lossio-Ventura and Hugo Alatrasta-Salas (eds.), *Proceedings of the 2nd annual international symposium on information management and big data - simbig 2015, cusco, Peru, September 2–4, 2015*, CEUR Workshop Proceedings, vol. 1478, CEUR-WS.org, 2015.
3. Juan Antonio Lossio-Ventura and Hugo Alatrasta-Salas (eds.), *Proceedings of the 3rd annual international symposium on information management and big data - simbig 2016, cusco, Peru, September 1–3, 2016*, CEUR Workshop Proceedings, vol. 1743, CEUR-WS.org, 2016.
4. Sam Madden, *From databases to big data*, vol. 16, IEEE Educational Activities Department, Piscataway, NJ, USA, May 2012, pp. 4–6.

Organizing Committee

General Organizers

Juan Antonio Lossio-Ventura	University of Florida, USA
Hugo Alatrística-Salas	Universidad del Pacífico, Peru

Local Organizers

Cristhian Ganvini Valcarcel	Universidad Andina del Cusco, Peru
Armando Fermin Perez	Universidad Nacional Mayor de San Marcos, Peru

Program Committee

Nathalie Abadie	French National Mapping Agency, COGIT, France
Elie Abi-Lahoud	University College Cork, Cork, Ireland
Salah AitMokhtar	Xerox Research Centre Europa, France
Sophia Ananiadou	NaCTeM - University of Manchester, UK
Marcelo Arenas	Pontificia Universidad Catolica de Chile, Chile
Jérôme Azé	LIRMM - University of Montpellier, France
Riza Batista-Navarro	NaCTeM - University of Manchester, UK
Pablo Barceló	Universidad de Chile, Chile
Nicolas Béchet	IRISA - University of Bretagne Sud, France
Cesar A. Beltrán Castañón	GRPIAA - Pontifical Catholic University of Peru, Peru
Lilia Berrahou	LIRMM - University of Montpellier, France
Jiang Bian	College of Medicine, University of Florida, USA
Albert Bifet	Télécom ParisTech, France
Sandra Bringay	LIRMM - Paul Valéry University, France
Mohamed R. Bouadjenek	University of Melbourne, Australia
Thierry Charnois	GREYC Université Paris 13, France
Oscar Corcho	Ontology Engineering Group - Polytechnic University of Madrid, Spain
Bruno Crémilleux	GREYC-CNRS, Université de Caen Normandie, France
Fabio Crestani	University of Lugano, Switzerland
Gabriela Csurka	Xerox Research Centre Europa, France
Martín Ariel Domínguez	Universidad Nacional de Córdoba, Argentina
Paula Estrella	Universidad Nacional de Córdoba, Argentina
Frédéric Flouvat	PPME Lab - University of New Caledonia, New Caledonia

Philippe Fournier-Viger	Harbin Institute of Technology Shenzhen Graduate School, China
André Freitas	University of Passau, Germany
Mauro Gaio	LIUPPA - University of Pau, France
Natalia Grabar	CNRS - University of Lille 3, France
Adrien Guille	Université Lumière Lyon 2, France
Thomas Guyet	IRISA/LACODAM - Agrocampus Ouest, France
Hakim Hacid	Zayed University, United Arab Emirates
Phan Nhat Hai	Oregon State University, USA
Sébastien Harispe	Ecole des Mines d'Alès, France
Dino Ienco	Irstea, France
Diana Inkpen	University of Ottawa, Canada
Clement Jonquet	LIRMM - University of Montpellier, France
Alípio Jorge	Universidade do Porto, Portugal
Eric Kergosien	GERiCO Lab - University of Lille 3, France
Georgios Kononatsios	NaCTeM - University of Manchester, UK
Yannis Korkontzelos	NaCTeM - University of Manchester, UK
Christian Libaque Saenz	Universidad del Pacífico, Peru
Cédric López	VEISO - Research and Development Unit, France
Franco M. Luque	Universidad Nacional de Córdoba, Argentina
Florent Masseglia	Inria - Zenith Team, France
Peter Mika	Yahoo! Research Labs - Barcelona, Spain
André Miralles	SISO Team, Irstea, France
François Modave	College of Medicine: University of Florida, USA
Giovanni Montana	Imperial College London, UK
Nhung Nguyen	University of Manchester, UK
Jordi Nin	BBVA Data & Analytics and Universidad de Barcelona, Spain
Miguel Nuñez del Prado Cortez	Universidad del Pacífico, Peru
Maciej Ogrodniczuk	Institute of Computer Science, Polish Academy of Sciences, Poland
Thomas Opitz	Biostatistics and Spatial Processes - Inra, France
José Manuel Perea-Ortega	University of Extremadura, Spain
Yoann Pitarch	IRIT - Toulouse, France
Marc Plantevit	LIRIS-CNRS, Université Claude Bernard Lyon 1, France
Pascal Poncelet	LIRMM - University of Montpellier, France
Perfecto Quintero-Flores	Instituto Tecnológico de Apizaco, Mexico
Julien Rabatel	Catholic University of Leuven, Belgium
José Luis Redondo García	Ontology Engineering Group - Polytechnic University of Madrid, Spain
Mathieu Roche	Cirad - TETIS - LIRMM, France
Nancy Rodriguez	LIRMM - University of Montpellier, France
Avid Roman-Gonzales	Universidad Nacional Cayetano Heredia, Peru
Fatiha Saïs	Paris-Sud 11 University, France

Arnaud Sallaberry	LIRMM - Paul Valéry University, France
Nazha Selmaoui-Folcher	PPME Labs - University of New Caledonia, New Caledonia
Selja Seppälä	College of Medicine, University of Florida, USA
Matthew Shardlow	University of Manchester, UK
Ivan Sipiran-Mendoza	Pontificia Universidad Católica del Perú, Peru
Paulo Teles	LIAAD-INESC Porto LA, Porto University, Portugal
Claude Tadonki	MINES ParisTech - PSL Research University, France
Maguelonne Teisseire	Irstea - LIRMM, France
Paul Thompson	University of Manchester, UK
Cassia Trojahn	IRIT - University of Toulouse 2, France
Carlos Vázquez	École de technologie supérieure, Canada
Julien Velcin	ERIC Lab - University of Lyon 2, France
Maria-Esther Vidal	Universidad Simón Bolívar, Venezuela
Boris Villazon-Terrazas	Expert System Iberia – Madrid, Spain
Florence Wang	CSIRO, Australia
Yang Yang	Kellogg School of Management, Northwestern University, USA
Osmar Zaiane	Department of Computing Science, University of Alberta, Canada
Manel Zarrouk	Insight - NUI Galway, Ireland
Amrapali Zaveri	Dumontier Lab, Stanford University, USA

Organizing Institutions and Sponsors

Universidad Andina del Cusco, Cusco, Peru

University of Florida, Florida, USA

Universidad del Pacífico, Lima, Peru

Pontificia Universidad Católica del Perú, Lima, Peru

Laboratoire de Informatique, Robotique et Micro-électronique de Montpellier,
Montpellier, France

Université de Montpellier, Montpellier, France



Contents

Sense-Level Semantic Clustering of Hashtags	1
<i>Ali Javed and Byung Suk Lee</i>	
Automatic Idiom Recognition with Word Embeddings	17
<i>Jing Peng and Anna Feldman</i>	
A Text Mining-Based Framework for Constructing an RDF-Compliant Biodiversity Knowledge Repository	30
<i>Riza Batista-Navarro, Chrysoula Zerva, Nhung T.H. Nguyen, and Sophia Ananiadou</i>	
Network Sampling Based on Centrality Measures for Relational Classification	43
<i>Lilian Berton, Didier A. Vega-Oliveros, Jorge Valverde-Rebaza, Andre Tavares da Silva, and Alneu de Andrade Lopes</i>	
Dictionary-Based Sentiment Analysis Applied to a Specific Domain	57
<i>Laura Cruz, José Ochoa, Mathieu Roche, and Pascal Poncelet</i>	
A Clustering Optimization Approach for Disaster Relief Delivery: A Case Study in Lima-Perú	69
<i>Jorge Vargas-Florez, Rosario Medina-Rodríguez, and Rafael Alva-Cabrera</i>	
An Approach to Evaluate Class Assignment Semantic Redundancy on Linked Datasets	81
<i>Leandro Mendoza and Alicia Díaz</i>	
Topic-Based Sentiment Analysis	95
<i>Prasadith Buddhitha and Diana Inkpen</i>	
A Security Price Data Cleaning Technique: Reynold's Decomposition Approach	108
<i>Rachel V. Mok, Wai Yin Mok, and Kit Yee Cheung</i>	
Big Data Architecture for Predicting Churn Risk in Mobile Phone Companies	120
<i>Alonso Raul Melgarejo Galvan and Katerine Rocío Clavo Navarro</i>	
Social Networks of Teachers in Twitter	133
<i>Hernán Gil Ramírez and Rosa María Guilleumas García</i>	
Author Index	147