Overview of SIMBig 2017: 4th Annual International Symposium on Information Management and Big Data

Juan Antonio Lossio-Ventura
Health Outcomes & Policy
University of Florida
Florida, USA
jlossioventura@ufl.edu

Hugo Alatrista-Salas
Universidad del Pacífico
Av. Salaverry 2020, Jesús María
Lima, Peru
h.alatristas@up.edu.pe

Abstract

SIMBig presents the analysis of new methods for extracting knowledge from large volumes of data through techniques of data science and artificial intelligence. SIMBig gathers national and international researchers in the data science field to state in new technologies dedicated to handle large amount of information.

1 Introduction

Our fourth edition of the Annual International Symposium on Information Management and Big Data - SIMBig 2017[1], took place in September 2017; in Lima, Peru; at the Universidad del Pacífico. SIMBig 2017 presented the new methods of data science an related fields for analyzing and managing large volumes. Counting with main national and international actors in the decision-making field to state in new technologies dedicated to handle large amount of information.

The best papers of SIMBig 2016 and SIMBig 2015 (e.g., eleven papers) have been published in Springer (Lossio-Ventura and Alatrista-Salas, 2017). Our third edition, SIMBig 2016[2], was held in Cusco, Peru in September 2016. As well as, the second edition, SIMBig 2015, was held in Cusco, Peru, in September 2015. The first edition, SIMBig 2014[3], took place in Cuzco Peru too in September 2014.

SIMBig 2016, 2015, and 2014 have been indexed on DBLP[4] (Lossio-Ventura and Alatrista-Salas, 2016, 2015; Lossio-Ventura and Alatrista-Salas, 2014) and on CEUR Workshop Proceedings[5,6,7].

Scope and Topics

To share the new analysis methods for managing large volumes of data, we encouraged participation from researchers in all fields related to Data Science, Big Data, Data Mining, Natural Language Processing, and Semantic Web. Topics of interest of SIMBig 2017 included but were not limited to:

- Data Science
- Big Data
- Data Mining
- Natural Language Processing
- Semantic Web
- Bio NLP, Healthcare Informatics
- Text Mining
- Information Retrieval
- Machine Learning
- Ontologies, Knowledge Representation, Linked Open Data
- Social Networks, Social Web, and Web Science
- Information visualization
- OLAP, Data warehousing, Business intelligence
- Spatiotemporal Data
- Agent-based Systems

2 Keynote Speakers

SIMBig 2017 has welcomed five keynote speakers experts in Data Science, Big Data, Data Mining, Natural Language Processing (NLP), and Semantic Web:

[1] http://simbig.org/SIMBig2017/
[2] http://simbig.org/SIMBig2016/
[3] https://www.lirmm.fr/simbig2014/
[4] http://dblp.uni-trier.de/db/conf/simbig/index.html
[5] http://ceur-ws.org/Vol-1743/
[6] http://ceur-ws.org/Vol-1478/
[7] http://ceur-ws.org/Vol-1318/

2.1 Regina Barzilay (Professor, PhD)

Dr. Regina Barzilay is a professor in the Department of Electrical Engineering and Computer Science and a member of the Computer Science and Artificial Intelligence Laboratory at the Massachusetts Institute of Technology. Barzilay's research on natural languages focuses on the development of models of natural language, and uses those models to solve real-world language processing tasks. Her research in computational linguistics deals with multilingual learning, interpreting text for solving control problems, and finding document-level structure within text.

She is a recipient of various awards including of the NSF Career Award, the MIT Technology Review TR-35 Award, Microsoft Faculty Fellowship and several Best Paper Awards at NAACL and ACL. She received her PhD in Computer Science from Columbia University, and spent a year as a postdoc at Cornell University

2.2 Jiawei Han (Professor, PhD)

Dr. Jiawei Han is Abel Bliss Professor in the Department of Computer Science at the University of Illinois. He has been researching into data mining, information network analysis, and database systems, with over 700 publications. He served as the founding Editor-in-Chief of ACM Transactions on Knowledge Discovery from Data (TKDD). Dr. Han has received ACM SIGKDD Innovation Award (2004), IEEE Computer Society Technical Achievement Award (2005), IEEE Computer Society W. Wallace McDowell Award (2009), and Daniel C. Drucker Eminent Faculty Award at UIUC (2011). He is a Fellow of ACM and Fellow of IEEE. His co-authored textbook "Data Mining: Concepts and Techniques" (Morgan Kaufmann) has been adopted worldwide.

Dr. Han is currently the co-Director of KnowEnG, a Center of Excellence in Big Data Computing, funded by NIH Big Data to Knowledge (BD2K) Initiative. He also served in 2009-2016 as the Director of Information Network Academic Research Center (INARC) supported by the Network Science-Collaborative Technology Alliance (NS-CTA) program of U.S. Army Research Lab.

2.3 Mark A. Musen (Professor, PhD, MD)

Dr. Mark Musen is Professor of Biomedical Informatics and of Biomedical Data Science at Stanford University, where he is Director of the Stanford Center for Biomedical Informatics Research. Dr. Musen conducts research related to intelligent systems, reusable ontologies, metadata for publication of scientific data sets, and biomedical decision support. His group developed Protégé, the world's most widely used technology for building and managing terminologies and ontologies. He is principal investigator of the National Center for Biomedical Ontology, one of the original National Centers for Biomedical Computing created by the U.S. National Institutes of Heath (NIH). He is principal investigator of the Center for Expanded Data Annotation and Retrieval (CEDAR). CEDAR is a center of excellence supported by the NIH Big Data to Knowledge Initiative, with the goal of developing new technology to ease the authoring and management of biomedical experimental metadata.

Dr. Musen directs the World Health Organization Collaborating Center for Classification, Terminology, and Standards at Stanford University, which has developed much of the information infrastructure for the authoring and management of the 11th edition of the International Classification of Diseases (ICD-11). Dr. Musen was the recipient of the Donald A. B. Lindberg Award for Innovation in Informatics from the American Medical Informatics Association in 2006. He has been elected to the American College of Medical Informatics, the Association of American Physicians, and the National Academy of Medicine. He is founding co-editor-in-chief of the journal Applied Ontology.

2.4 Ravi Kumar (PhD)

Dr. Ravi Kumar has been a senior staff research scientist at Google since 2012. Prior to this, he was a research staff member at the IBM Almaden Research Center and a principal research scientist at Yahoo! Research. Dr. Ravi Kumar obtained his PhD in Computer Science from Cornell University. His research interests include Web search and data mining, algorithms for massive data, and the theory of computation.

2.5 Clement Jonquet (Professor, PhD)

Dr. Clement Jonquet is assistant professor at University of Montpellier, France and since Sept. 2015 visiting scholar at the Stanford University. He is a researcher at the Laboratory of Informatics, Robotics, and Microelectronics of Montpel-

lier (LIRMM), on (biomedical/agronomical) ontologies, semantic data indexing and annotation, semantic Web, text mining, knowledge representation. Dr. Jonquet obtained his PhD in Informatics from the same university in 2006 (about multi-agent systems, grid and service-oriented computing), then he served as a postdoc for 3 years at the Stanford BMIR within Pr. Mark A. Musen's group where he was working on semantic annotations of biomedical data using biomedical ontologies in the context of the National Center for Biomedical Ontology (NCBO) project. He contributed actively to the design, evolution and development of the NCBO BioPortal and won the 1st prize at ISWC Semantic Web Challenge 2010.

3 Track on Social Network and Media Analysis and Mining (SNMAM 2017)

Online social networks are web platforms that provide a variety of services. Users may share locations and community activities, post and tag photos and other media content, as well as contact individuals with similar interests. The rapid growth of social networks, as well as the rapid increase in social media consumption and production have made the analysis of social media and networks a hot topic amongst academic researchers and industry practitioners alike. SIMBig has become an important venue that has attracted computer scientists, computer engineers, software engineers, and application developers from around the world. Within the general symposium, the Social Network and Media Analysis and Mining (SNMAM) track provided a forum that brought both researchers and practitioners to discuss research trends and techniques related to social networks and media.

Topics of Interest

We included all the important topics related to social network and media analysis and mining within SNMAM. The topics suitable for SNMAM included:

- Data modeling for social networks and social media
- Dynamics and evolution of social networks
- Topological, geographical and temporal analysis of social networks
- Privacy and security in social networks
- Pattern analysis in social networks
- Crowd sourcing of network data generation and collection
- Community structure analysis in social networks
- Link prediction and recommendation systems
- Propagation and diffusion of information in social networks
- Location-based social networks
- Mobile computing and applications on social networks
- Modeling of user behavior and interaction in social networks
- Information retrieval in social network and media services
- Business and political impact in social network and media analysis.
- Monitoring social networks and media.
- Analysis of the relationship between social media and traditional media
- Exploratory and visual data mining of social networks and media data.
- Ethics and privacy in social network and media services.
- Big data issues in social network and media analysis.

4 Track on Applied Natural Language Processing (ANLP 2017)

The availability and size of textual information have grown dramatically in different areas such as academic, work or individual. Emails, working papers, scientific articles or social media publications are some examples of large sources of data that are presented in natural language. This raises a challenge, since the language presents a type of unstructured data that contains ambiguity, among other properties that increase the difficulty in the processing task. In this context, there is a growing interest in improving the accessibility to information and its exploitation in different environments by companies and organizations. For all this, the applications of Natural Language Processing have become very important today. SIMBig has become an important meeting point of computer scientists, computer engineers, software engineers, and application developers from around the world. The Applied Natural Language Processing (ANLP) track of SIMBig provided a forum that brought both researchers and practition-

ers to discuss: research trends and techniques related to Natural Language Processing.

Topics of Interest

We included all the important topics related to applied natural language processing within ANLP. The topics suitable for ANLP included:

- Machine Translation.
- Sentiment Analysis/Opinion Mining.
- Automatic Summarization.
- Plagiarism Detection.
- Language Detection.
- Natural Language Generation.
- Natural Language Interfaces.
- NLP in Informal Texts.
- Question-Answering Systems.
- Content Analysis.
- NLP for Education.
- NLP for Low-Resource Languages.
- Bio-NLP.
- Dialogue System
- Information Retrieval and Extraction
- Event Detection
- Text Classification
- Multilingual NLP
- Ontology-based NLP

5 Sponsors

We want to thank our wonderful sponsors! We extend our sincere appreciation to our sponsors, without whom our symposium would not be possible. They showed their commitment to making our research communities more active. We invite you to support these community-minded organizations.

5.1 Organizing Institutions

- Universidad del Pacífico, Perú[8]
- University of Florida, USA[9]
- Universidad Andina del Cusco, Perú[10]

5.2 Collaborating Institutions

- Springer[11]
- Banco de Crédito del Perú[12]
- Escuela de Post-grado de la Pontificia Universidad Católica del Perú[13]

5.3 SNMAM Organizing Institutions

- Instituto de Ciências Matemáticas e de Computação, USP, Brasil[14]
- Labóratorio de Intêligencia Computacional, ICMC, USP, Brasil[15]
- Universidade Federal de São Carlos, Brasil[16]

5.4 ANLP Organizing Institutions

- Universidad Nacional Mayor de San Marcos, Perú[17]
- Grupo de Reconocimiento de Patrones e Inteligencia Artificial Aplicada, PUCP, Perú[18]
- Instituto de Ciências Matemáticas e de Computação, USP, Brasil
- Universidade Federal de São Carlos, Brasil

References

Juan Antonio Lossio-Ventura and Hugo Alatrista-Salas, editors. 2014. *Proceedings of the 1st Symposium on Information Management and Big Data - SIMBig 2014, Cusco, Peru, September 8-10, 2014*, volume 1318 of *CEUR Workshop Proceedings*. CEUR-WS.org. http://ceur-ws.org/Vol-1318.

Juan Antonio Lossio-Ventura and Hugo Alatrista-Salas, editors. 2015. *Proceedings of the 2nd Annual International Symposium on Information Management and Big Data - SIMBig 2015, Cusco, Peru, September 2-4, 2015*, volume 1478 of *CEUR Workshop Proceedings*. CEUR-WS.org. http://ceur-ws.org/Vol-1478.

Juan Antonio Lossio-Ventura and Hugo Alatrista-Salas, editors. 2016. *Proceedings of the 3rd Annual International Symposium on Information Management and Big Data - SIMBig 2016, Cusco, Peru, September 1-3, 2016*, volume 1743 of *CEUR Workshop Proceedings*. CEUR-WS.org. http://ceur-ws.org/Vol-1743.

Juan Antonio Lossio-Ventura and Hugo Alatrista-Salas. 2017. *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*, volume 656. Springer. https://doi.org/10.1007/978-3-319-55209-5.

[8] http://www.up.edu.pe/
[9] http://www.ufl.edu/
[10] http://www.uandina.edu.pe/
[11] http://www.springer.com/la/
[12] https://www.viabcp.com/wps/portal/
[13] http://posgrado.pucp.edu.pe/la-escuela/presentacion/
[14] http://www.icmc.usp.br/Portal/
[15] http://labic.icmc.usp.br/
[16] http://www2.ufscar.br/home/index.php
[17] http://www.unmsm.edu.pe/
[18] http://inform.pucp.edu.pe/~grpiaa/