Information Management and Big Data: SIMBig overview

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

Hugo Alatrista-Salas
Universidad del Pacífico
Pontificia Universidad Católica del Perú
Lima, Peru
h.alatristas@up.edu.pe

Abstract

Big Data and Data Science are popular terms used to describe the exponential growth of data and the analysis of this data respectively. The aim of the symposium is to present the analysis of methods for extracting knowledge from large volumes of data through techniques of data science and artificial intelligence. Bringing together main national and international actors in the decision-making field to state in new technologies dedicated to handle large amount of information.

1 Introduction

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 some fields such as the data analysis, statistics, machine learning, data mining similar to Knowledge Discovery in Databases (KDD).

Big Data has taken place over the last 20 years. For instance, social networks such as Facebook, Twitter and Linkedin generate masses of data, which is available to be accessed by other applications. Several domains, including biomedicine, life sciences and scientific research, have been affected by Big Data[1]. Therefore there is a need to understand and exploit this 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 (Madden, 2012; Embley and Liddle, 2013). 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 third edition of the Annual International Symposium on Information Management and Big Data - SIMBig 2016[3], seeks to present 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 second edition, SIMBig 2015[4], was held in Cusco, Peru, from September 2nd to 4th, 2015. SIMBig 2015 has been indexed on DBLP[5] (Lossio-Ventura and Alatrista-Salas, 2015) and on CEUR Workshop Proceedings[6].

Our first edition, SIMBig 2014[7] took also place in Cusco, Peru in September 2014. SIMBig 2014 has also been indexed on DBLP[8] (Lossio-Ventura and Alatrista-Salas, 2014) and on CEUR Workshop Proceedings[9].

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 Big Data, Data Science, Data Mining, Natural Language Processing and Semantic Web, but also

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

[2] Today, 80% of data is unstructured such as images, video, and notes

[3] http://simbig.org/SIMBig2016/

[4] http://simbig.org/SIMBig2015/

[5] http://dblp2.uni-trier.de/db/conf/simbig/simbig2015.html

[6] http://ceur-ws.org/Vol-1478/

[7] https://www.lirmm.fr/simbig2014/

[8] http://dblp2.uni-trier.de/db/conf/simbig/simbig2014.html

[9] http://ceur-ws.org/Vol-1318/

Multilingual Text Processing, Biomedical NLP. Topics of interest of SIMBig 2016 included but were not limited to:

- 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

2 Keynote Speakers

This third edition of SIMBig, we had experts on different areas, such as Data Science, Information Retrieval, Natural Language Processing and Data Mining. Our four invited were:

2.1 Albert Bifet (Prof, PhD)

Albert Bifet is a Big Data scientist with 10+ years of international experience in research and in leading new open source software projects for business analytics, data mining and machine learning (Huawei, Yahoo, University of Waikato, UPC). He obtained a PhD from UPC-BarcelonaTech. He has worked in Hong Kong, New Zealand and Europe. At Yahoo Labs, he co-founded Apache SAMOA (Scalable Advanced Massive Online Analysis) in 2013. Apache SAMOA is distributed streaming machine learning (ML) framework that contains a programing abstraction for distributed streaming ML algorithms. At the WEKA Machine Learning group, he is co-leading MOA (Massive Online Analysis) since 2008. MOA is the most popular open source framework for data stream mining, with more than 20,000 downloads each year. He is the author of a book on Adaptive Stream Mining and Pattern Learning and Mining from Evolving Data Streams. Additionally, he was editor of the Big Data Mining special issue of SIGKDD Explorations in 2012. Also, he is serving as Co-Chair of the Industrial track of ECML PKDD 2016, and served as Co-Chair of BigMine (2014, 2013, 2012), and ACM SAC Data Streams Track (2016, 2014, 2013, 2012).

2.2 Fabio Crestani (Prof, PhD)

Fabio Crestani is a full professor at the Faculty of Informatics of USI since January 2007. Before arriving in Lugano he was a (full) Professor at the University of Strathclyde in Glasgow (UK) since 2000. During that time he was a Visiting Professor at IMAG (France), and spent a year sabbatical at UC Berkeley (USA) and Xerox PARC (USA). In 1997-1999 he was a Postdoctoral Research Fellow at the University of Glasgow (UK), at the International Computer Science Institute in Berkeley (USA), and at the Rutherford Appleton Laboratory (UK). Earlier, in 1992-97 he was Assistant Professor at the Department of Information Engineering of the University of Padova (Italy). Fabio holds a degree in Statistics from the University of Padova (Italy) and an MSc and PhD in Computing Science from the University of Glasgow (UK). His main areas of research are Information Retrieval, Text Mining, and Digital Libraries. He has co-edited 9 books and published over 150 publications in these areas of research. He is Editor-in-Chief of Information Processing and Management (Elsevier) and member of the editorial board of a number of journals.

2.3 Kevin-Bretonnel Cohen (Prof, PhD)

Kevin Bretonnel Cohen is the Director of the Biomedical Text Mining Group at the University of Colorado School of Medicine. His research covers a wide range of topics in biomedical natural language processing, ranging from named entity recognition to software engineering and evaluation for language processing applications. Since 2008, he has been the chair of the Association for Computational Linguistics special interest group on biomedical natural language processing.

2.4 Maguelonne Teisseire (Prof, PhD)

Maguelonne Teisseire received a PhD degree in Computing Science from the Méditerranée University, France, in 1994. Her research interests

focused on behavioral modeling and design. In 1995- 2008, she was an Assistant Professor of Computer Science and Engineering in Montpellier II University and Polytech'Montpellier, France. She headed the Data Mining Group at the LIRMM Laboratory Lab, Montpellier, France, from 2000 to 2008. She is currently a Research Professor - Irstea and she joined the TETIS lab in March 2009. Her research interest focus on advanced data mining approaches when considering that data are time ordered. Particularly, she is interested in text mining and sequential patterns. Her research takes part on different projects supported by either National Government (RNTL) or regional project. She has published numerous papers in refereed journals and conferences either on behavioral modeling or data mining.

3 Track on Social Network and Media Analysis and Mining (SNMAM 2016)

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 among 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. The Social Network and Media Analysis and Mining (SNMAM) track of SIMBig has provided a forum that brings 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
- Community structure analysis in social networks
- Link prediction and recommendation systems
- Propagation and diffusion of information in social networks
- Detection of spam, misinformation and malicious activities in social networks
- Location-based 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
- Big data issues 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 network and media data

4 Sponsors and Organizers

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.

4.1 Organizing Institutions

- Universidad Andina del Cusco, Perú[10]
- Universidad del Pacífico, Perú[11]
- University of Florida, USA[12]
- Université de Montpellier, France[13]

4.2 Collaborating Institutions

- Grupo de Reconocimiento de Patrones e Inteligencia Artificial Aplicada, PUCP, Perú[14]
- Universidad Nacional Mayor de San Marcos, Perú[15]
- Escuela de Post-grado de la Pontificia Universidad Católica del Perú[16]

[10] http://www.uandina.edu.pe/
[11] http://www.up.edu.pe/
[12] http://www.ufl.edu/
[13] http://www.umontpellier.fr/
[14] http://inform.pucp.edu.pe/~grpiaa/
[15] http://www.unmsm.edu.pe/
[16] http://posgrado.pucp.edu.pe/la-escuela/presentacion/

4.3 WTI Organizing Institutions

- Instituto de Ciências Matemáticas e de Computação, USP, Brasil[17]
- Labóratorio de Intêligencia Computacional, ICMC, USP, Brasil[18]
- Universidade Federal de São Carlos, Brasil[19]

References

David W Embley and Stephen W Liddle. 2013. Big data—conceptual modeling to the rescue. In *Conceptual Modeling*, ER'13, pages 1–8. LNCS, Springer.

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.

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.

Sam Madden. 2012. From databases to big data. volume 16, pages 4–6. IEEE Educational Activities Department, Piscataway, NJ, USA, may.

[17] http://www.icmc.usp.br/Portal/
[18] http://labic.icmc.usp.br/
[19] http://www2.ufscar.br/home/index.php