



Special issue on advances in data, information and knowledge engineering in data science era

Ladjel Bellatreche¹ · A Min Tjoa²

Received: 26 October 2021 / Accepted: 28 October 2021 / Published online: 28 February 2022
© The Author(s), under exclusive licence to Springer-Verlag GmbH Austria, part of Springer Nature 2022

In recent years, Data Science emerged as a new and important discipline. It can be viewed as an amalgamation of classical disciplines like statistics, data mining, databases, and distributed systems [8]. One of the major goal of Data Science is the extraction of significant *value* from Big Data [6]. Data, Information and Knowledge play a crucial role in getting this added value [2,3,5].

Data science requires prerequisites as any discipline. The most important among them are the following: *data understanding*, *algorithms and logic*, *statistics*, *business domain*, and *deployment infrastructures*.

SOFSEM (SOftware SEMinar) is the annual international winter conference devoted to the theory and practice of Computer Science. SOFSEM presents the latest results and developments academic and industrial research in leading areas of Computer Science. The first SOFSEM was organized in 1974. SOFSEM consists of Invited Talks by prominent researchers, Contributed Talks selected from the submitted papers, and the Student Research Forum. The program is organized in plenary talks and parallel tracks devoted to original research in the selected research areas.

SOFSEM provides an interesting forum for having and reinforcing the prerequisites of Data Science, covering topics related to Fundamental Computer Science, Data and Knowledge Management, and Software Engineering.

The 44th edition of the International Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM), held in Krems, Austria, in January/February 2018 was organized on three main tracks [7]:

1. *Foundations of Computer Science*, co-chaired by Jan van Leeuwen, Utrecht University, The Netherlands, and Jiri Wiedermann, Academy of Sciences of the Czech Republic, Czech Republic,

✉ Ladjel Bellatreche
bellatreche@ensma.fr

A Min Tjoa
amin@ifs.tuwien.ac.at

¹ LIAS/ISAE-ENSMA - University of Poitiers, Poitiers, France

² IFS, Technical University of Vienna, Vienna, Austria

2. *Software Engineering: Advanced Methods, Applications, and Tools*, chaired by Stefan Biffl, TU Wien, Austria,
3. *Data, Information and Knowledge Engineering*, chaired by Ladjel Bellatreche, ISAE-ENSMA, France.

This special issue has been associated with *Data, Information and Knowledge Engineering*. It is devoted to all aspects of eliciting, acquiring, modelling, storing, and managing data, information, and knowledge. This track received 26 papers from over 12 countries. The program committee finally selected 10 full papers, all published by Springer in LNCS series.

In addition to the 10 accepted papers, two internationally recognized researchers were invited to give a talk in our track:

- Professor Yannis Manolopoulos, Cyprus University, gave a talk entitled “*Network Analysis of the Science of Science: A Case Study in SOFSEM Conference*” [4]. In his talk, Professor Yannis Manolopoulos focused on the “Science of Science” that has emerged as a *fast growing interdisciplinary field*, where two provocative questions were asked [4]: (1) *how does scientific collaboration and networking affect research impact?*, and (2) *what constitutes a truly influential individual in science and what meaningful interpretable patterns arise in the evolution of science?*. By leveraging the various networks (collaboration, citation, co-citation, etc.) related to the recording of science, he explored the factors affecting the generation of research and identify mechanisms of effective research collaboration and production. Professor Yannis Manolopoulos investigated bibliometric data of the *SOFSEM conference* as a case study, where a corpus of 1006 publications with their associated authors and affiliations to uncover the effects of collaboration network on the conference output, is considered.
- Professor Thomas Eiter, Vienna University of Technology, gave a talk titled “*A Framework for Analytic Reasoning over Streams*” [1]. He mainly focused on stream reasoning that continuously derives conclusions on streaming data aiming at high expressiveness under declarative semantics. Professor Thomas Eiter presented a Logic-based Framework for Analytic Reasoning over Streams including its relation to other formalisms, and touched implementation and applications.

This special issue was managed as follows: Four best papers covering the topics of Computing Journal were invited to extend their papers by adding at least 30% of new materials for our special issue. During the review process, each paper was assigned to and reviewed by two experts, with a rigorous review process. Thanks to the great support of the Editor-in-Chief of Computing Journal, Professor Schahram Dustdar, the guest editors were able to accept 3 selected papers related to the following topics: Explainable Fake News Management, Decision Making in the case of Unbalanced Datasets, with a case study of Suicidal Ideation, and Data Provenance.

The three selected papers are summarized as follows:

The first paper, titled; *MANIFESTO : a huMAN-centric explainable approach for Fake news Spreaders deTectiOn*, by Orestis Lampridis, Dimitra Karanatsiou, Athena Vakali, tackles a timely subject related the problem of handling fake news spreading in social media. The spread of fake news on the Internet represents a crucial issue

for the image of the main components of our society, including governments, policymakers, organisations, businesses and citizens. In addition of the timely nature of the subject, this paper proposed solutions for fake news spreaders detection with a particular emphasis on features and Explainability. A nicely presented state-of-the-art on the problem of fake news spreaders detection integrating different types of features (including natural language and social setting) and Explainability is given. The proposed methodology, called MANIFESTO, uses advanced explainable Machine Learning in order to aid the user in making a more educated final decision with regard to real and false pieces of information (obtained from Twitter) focusing on the reputation of users who participate in such discussions. An extensive experimental evaluation using US elections and COVID-19 data shows improvement of the results reaching +8% in terms of quality. This work has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship, and Innovation.

The second paper, titled, *A Novel Imbalanced Data Classification Approach for Suicidal Ideation Detection on Social Media*, by Mohamed Ali Ben Hassine, Safa Abellatif, and Sadok Ben Yahia deals with a hot issue related to data-driven solutions for decision making, where the input datasets are unbalanced. Another point that increases the interestingness of the subject is the studied case study of suicidal ideation, especially during this Covid-19 pandemic period. The research work exposed in this paper concerns the data mining topic, in general, and association rules, in particular. Recall that a data set is imbalanced whenever the number of instances belonging to one class dramatically exceeds that of other class instances. The latter, called the minority class, is the one that has the most significant interest and the highest impact and must be considered during the learning process. The authors proposed an association-rule-based approach to the sentiment analysis domain in the field of suicidal ideation detection and individual at-risk. This approach learns from the imbalanced data. Furthermore, the authors provide an interesting discussion about the limitation of existing interestingness measures and the necessity to propose a new one dedicated to critical situations. This latter aims at selecting highly interesting rules from both types of classes regardless of their imbalanced distribution. Several experiments have been conducted experimentations have sketched the potential of the proposed approach when encountering real-world problems.

The third paper, titled, *Automated and non-intrusive provenance capture with UML2PROV*, Carlos Sáenz-Adán, Francisco J. García-Izquierdo, Beatriz Pérez, Trung Dong Huynh, and Luc Moreau, tackles the problem of data provenance. The term provenance has emerged to refer to “the information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness”. To facilitate the instrumentation of data provenance recording in applications designed with UML diagrams, the authors propose an UML2PROV—a software-engineering methodology. It automates the generation of (1) templates for the provenance to be recorded and (2) the code to capture values required to instantiate those templates from an application at run time, both from the application’s UML diagrams. UML2PROV frees application developers from manual instrumentation of provenance capturing while ensuring the quality of recorded provenance. The authors in this paper present in detail UML2PROV’s approach to gen-

erating application code for capturing provenance values via the means of Bindings Generation Module (BGM). More specifically, they propose a set of requirements for BGM implementations and describe an event-based design of BGM that relies on the Aspect-Oriented Programming (AOP) paradigm to automatically weave the generated code into an application. Finally, they present three different BGM implementations following the above design and analyse their pros and cons in terms of computing/storage overheads and implications to provenance consumers.

We hope readers will find the content of this special issue interesting and will inspire them to look further into the challenges that are still ahead before designing data-enabled systems and applications using Machine Learning, Deep Learning and Data Mining techniques to get added value. We would like to thank all the authors who submitted their papers to this special issue. In addition, we are grateful for the support of various reviews that ensured the high quality of this special issue. Last but not least, we would like to thank Professor Schahram Dustdar, The Editor-In-Chief of Computing, for accepting our proposal of a special issue, and for assisting us whenever required. We would like to thank very much Hemalatha Kamaraj, and Christine Kamper for their endless help and support. The complete International Program Committee of this special issue is listed next.

International Program Committee

- Esma Aimeur, University of Montréal, Canada
- Mohamed-Amine Baazizi, Sorbonne University, France
- Khalid Belhajjame, University Paris-Dauphine, France
- Djamel Benslimane, University of Lyon 1, France
- Hakim Hacid, Zayed University, United Arab Emirates
- Mirjana Ivanovic, Faculty of Sciences, University of Novi Sad, Serbia
- Daniel Cardoso Moraes de Oliveira, Universidade Federal Fluminense, Brazil
- Taoufik Yeferny, University of Pau & Pays Adour, France

References

1. Beck H, Dao-Tran M, Eiter T (2018) LARS: a logic-based framework for analytic reasoning over streams - (extended abstract). In: SOFSEM 2018: theory and practice of computer science - 44th international conference on current trends in theory and practice of computer science, pp. 87–93. Springer
2. Berkani N, Bellatreche L, Benatallah B (2016) A value-added approach to design BI applications. In: 18th international conference on big data analytics and knowledge discovery (DaWaK), pp. 361–375. Springer
3. Berkani N, Bellatreche L, Khouri S, Ordonez C (2020) The contribution of linked open data to augment a traditional data warehouse. *J Intell Inf Syst* 55(3):397–421
4. Gogoglou A, Tsikrika T, Manolopoulos Y (2018) Network analysis of the science of science: a case study in SOFSEM conference. In: SOFSEM 2018: theory and practice of computer science - 44th international conference on current trends in theory and practice of computer science, pp. 94–108. Springer
5. Novak NM, Tjoa AM (2019) Towards a business value framework for linked enterprise data. In: 2019 IEEE-RIVF international conference on computing and communication technologies, RIVF 2019, pp. 1–6. IEEE
6. Srivastava D (2020) Towards high-quality big data: lessons from FIT. In: Wu X, Jermaine C, Xiong L, Hu X, Kotevska O, Lu S, Xu W, Aluru S, Zhai C, Al-Masri E, Chen Z, Saltz J (eds) IEEE international conference on big data, p 4. IEEE

7. Tjoa AM, Bellatreche L, Biffi S, van Leeuwen J, Wiedermann J editors (2018) SOFSEM 2018: theory and practice of computer science. In: proceedings - 44th international conference on current trends in theory and practice of computer science, Krems, Austria, January 29 - February 2, 2018, Springer
8. van der Aalst WMP (2016) Data science in action. Springer, Berlin, pp 3–23

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.