Please cite as: Zhang, Y., Zhang, C., Mayr, P., Suominen, A., Ding, Y. 2023, An editorial of "AI + Informetrics": Robust models for large-scale analytics, *Information Processing & Management*, DOI: 10.1016/j.ipm.2023.103495.

An Editorial of "AI + Informetrics": Robust Models for Large-scale Analytics

Yi Zhang[1], Chengzhi Zhang[2], Philipp Mayr[3], Arho Suominen[4], Ying Ding[5]

[1] Australian Artificial Intelligence Institute, Faculty of Engineering and Information Technology, University of Technology Sydney, Australia.

[2] Department of Information Management, Nanjing University of Science and Technology, China

[3] Department of Knowledge Technologies for the Social Sciences, GESIS-Leibniz Institute for the Social Sciences, Germany

[4] Tampere University, Finland

[5] School of Information & Dell Medical School, University of Texas at Austin, USA

Email: yi.zhang@uts.edu.au; zhangcz@njust.edu.cn; Philipp.Mayr@gesis.org; arho.suominen@tuni.fi; ying.ding@ischool.utexas.edu.

ORCID: 0000-0002-7731-0301 (Yi Zhang); 0000-0001-9522-2914 (Chengzhi Zhang); 0000-0002-6656-1658 (Philipp Mayr); 0000-0001-9844-7799 (Arho Suominen); 0000-0003-2567-2009 (Ying Ding).

Introduction

Driven by the big data boom, *informetrics*, known as the study of quantitative aspects of information, has gained significant benefits from *artificial intelligence* (AI) – including a wide range of intelligent agents through techniques such as deep neural networks, genetic programming, computer vision, heuristic search, knowledge representation and reasoning, Bayes network, planning, and language understanding. With extraordinary capacities in analyzing unstructured and scalable data, understanding uncertain semantics, and constructing robust and repeatable models, "*AI + Informetrics*" has demonstrated enormous success in turning big data into big value. Examples of AI + Informetrics include bibliometric-enhanced information retrieval (Mayr et al., 2014), patent mapping with unsupervised learning approaches (Suominen & Toivanen, 2016), intelligent bibliometrics for tracking technological change with streaming data analytics (Zhang et al., 2017), recommending knowledge trajectories by heterogeneous network analytics (Zhang et al., 2023b), entity extraction with full-text analytics (Wang & Zhang, 2020), and deep learning-empowered models for metadata analysis (Safder et al., 2020) and classification (Haneczok & Piskorski, 2020). Such endeavors with broadened perspectives from machine intelligence will portend far-reaching implications for science (Fortunato et al., 2018). As a rising interest of broad information scientists, technology analysts, innovation managers, and policymakers, developing and applying robust computational models to analyze large-scale information sources (e.g., scientific documents, knowledge graphs, and social media) with extensive uses of bibliographical indicators and knowledge entities have attracted significant attention.

Aiming to gather researchers and practitioners to open a collaborative platform for exchanging ideas, sharing pilot studies, and scoping future directions on this cutting-edge venue, the AI + Informetrics (AII) workshop series[1] was initiated in 2021. We organized the 2nd AII workshop as a special track in co-location with the Information Processing and Management Conference (IP&MC) 2022[2]. This workshop specifically targeted specific unsolved issues in AII:

- Cohering AI and informetrics to fulfill cross-disciplinary gaps from theoretical or practical perspectives.
- Elaborating AI-empowered informetric models with enhanced capabilities in robustness, adaptability, and effectiveness.
- Leveraging knowledge, concepts, and models in information management to strengthen the interpretability of AI + Informetrics to adapt to empirical needs in real-world cases.

This workshop has received an extremely positive response from the community. It received 43 submissions, and after a rigorous peer review process, we accepted 14 submissions for oral presentation[3]. The workshop was held online on October 21, 2022. Remarkably, we invited Prof Ludo Waltman from CWTS at Leiden University to deliver a keynote on *Two AI flavors – Which one do we need more?* Prof Waltman discussed the distinction between the statistical paradigm to AI and the symbolic paradigm, focusing on the role these paradigms play in informetrics and scientometrics. He argued that the symbolic paradigm deserves more attention from the community.

Further peer reviews were applied to the 14 submissions, and we finalized the special issue with 9 articles. We introduced the 9 articles with the following three categories: Knowledge Segmentation and Representation, Scholarly Recommendation, and Science, Technology, and Innovation (ST&I) Studies.

AI + Informetrics for Knowledge Segmentation and Representation

Knowledge segmentation, representation, and summarization are essential tasks in AI-enhanced information retrieval and building up knowledge foundations for follow-up information studies.

Aiming to detect figures and tables in research articles, Yu et al. (2023) developed a method using a model of cascading image semantic segmentation and a contour detection algorithm and achieved prior performance in a set of 14,678 PubMed papers.

Li and Xu (2023) handled the task as a question-answering process to comprehensively summarize research articles and proposed a framework transforming extractive summarization to a machine reading comprehension task. Compared with several state-of-the-art baselines, the approach was tested in three datasets, i.e., Covid, PubMed, and Arxiv.

Xue et al. (2023) focused on the automatic ranking task of research articles with limited full-text content. They developed a dual-view graph convolutions-enhanced BERT method to

[1] The website of the AII Workshop series: https://ai-informetrics.github.io/
[2] https://www.elsevier.com/events/conferences/information-processing-and-management-conference
[3] 12 submissions were registered and presented in the workshop, see https://ai-informetrics.github.io/2022/

facilitate lexical and semantic attention in the combined titles and abstracts. The method beat baselines in two selected datasets.

AI + Informetrics for Scholarly Recommendation

With vast practical scenarios, scholarly recommendation covers broad topics in scientific activities, e.g., recommending research venues, topics, classification tags, and collaborators, and has long been an interest of the community.

With an argument that it is important to have a localized metric to indicate the significance of individual authors to a specific research venue, Zhang et al. (2023a) developed a relevance-based author ranking algorithm to measure such significance. The algorithm is prior to five author evaluation baselines in three domain datasets of Microsoft Academic Graph.

Liu et al. (2023) glanced at research collaboration and developed a method of heterogeneous network embedding recommendation, comprehensively considering co-topics, co-authorships, co-venues, and co-citations. The method was tested in the DBLP and the AMiner datasets with five collaborator recommendation baselines.

AI + Informetrics for Science, Technology, and Innovation Studies

The AII community has been interactively engaged with broad ST&I studies. Leveraging AI and advanced data analytics to understand data, recognize patterns, and predict future trends has demonstrated impressive power for tasks in the science of science, tech mining, technological forecasting, etc.

Yao et al. (2023) conducted a large-scale analysis of AI-related articles to understand recent developments through entities on methods, datasets, and metrics. They particularly proposed a multi-stage self-paced learning strategy to address the negative influence of hard and noisy samples on extraction model training.

To explore the knowledge linkages between science and technology, Chen et al. (2023) proposed an analytical framework consisting of Node2Vec, BERT, network analytics, and Z-Score index to identify and match science-technology topics from research articles and patent documents.

Lou and Meng (2023) answered how scholars follow research trends by developing a measure based on dynamic topic models, which specifically captures the diversity of research behavior patterns in canonical and ubiquitous progress. An empirical study on computer vision examined the measure.

To understand what influence the spread of information on social media platforms, Yan et al. (2023) developed a time series regression-based approach and answered this question through an empirical case in the Sina Microblog. The study identified sentiment polarity, reposting, and emotional contagion as the core impact factors on information diffusion.

Conclusions

In the 2nd AII workshop and this special issue, our community has extensively engaged with deep learning techniques to develop robust models for efficient large-scale data analytics, e.g.,

BERT for text representation, BiLSTM language models for entity extraction, and graph representation learning for heterogeneous entity embedding. These models have also been widely applied to downstream tasks in broad ST&I studies, initiating impressive cross-disciplinary solutions to real-world problems. As a long-term motivation and inspiration of the AII workshop, we are passionate about promoting an inter-/multi-/cross-/trans-disciplinary cohort with cutting-edge AI techniques and down-to-earth information studies, and according to the fruitful outcomes collected in this special issue, we believe we are on the right trail.

Acknowledgments

Yi Zhang acknowledges the Discovery Early Career Researcher Award granted by the Australian Research Council (DE190100994) and the Commonwealth Scientific and Industrial Research Organisation (CSIRO) - the US National Science Foundation (NSF) AI Research Collaboration Program (CSIRO-NSF# 2303037). Chengzhi Zhang acknowledges the National Natural Science Foundation of China (No. 72074113).

References

Chen, X., Ye, P., Huang, L., Wang, C., Cai, Y., Deng, L., & Ren, H. (2023). Exploring science-technology linkages: A deep learning-empowered solution. *Information Processing & Management*, *60*(2), 103255.

Fortunato, S., Bergstrom, C., Börner, K., Evans, J., Helbing, D., Milojević, S., Petersen, A., Radicchi, F., Sinatra, R., & Uzzi, B. (2018). Science of science. *Science*, *359*(6379). https://doi.org/10.1126/science.aao0185

Haneczok, J., & Piskorski, J. (2020). Shallow and deep learning for event relatedness classification. *Information Processing & Management*, *57*(6), 102371.

Li, S., & Xu, J. (2023). MRC-Sum: An MRC framework for extractive summarization of academic articles in natural sciences and medicine. *Information Processing & Management*, *60*(5), 103467.

Liu, X., Wu, K., Liu, B., & Qian, R. (2023). HNERec: Scientific collaborator recommendation model based on heterogeneous network embedding. *Information Processing & Management*, *60*(2), 103253.

Lou, W., & Meng, J. (2023). The diversity of canonical and ubiquitous progress in computer vision: A dynamic topic modeling approach. *Information Processing & Management*, *60*(3), 103238.

Mayr, P., Scharnhorst, A., Larsen, B., Schaer, P., & Mutschke, P. (2014). Bibliometric-Enhanced Information Retrieval. In M. de Rijke, T. Kenter, A. P. de Vries, C. Zhai, F. de Jong, K. Radinsky, & K. Hofmann, *Advances in Information Retrieval* 36th European Conference on Information Retrieval, 798-801.

Safder, I., Hassan, S.-U., Visvizi, A., Noraset, T., Nawaz, R., & Tuarob, S. (2020). Deep learning-based extraction of algorithmic metadata in full-text scholarly documents. *Information Processing & Management*, *57*(6), 102269.

Suominen, A., & Toivanen, H. (2016). Map of science with topic modeling: Comparison of unsupervised learning and human‐assigned subject classification. *Journal of the Association for Information Science and Technology*, *67*(19), 2464–2476.

Wang, Y., & Zhang, C. (2020). Using the full-text content of academic articles to identify and evaluate algorithm entities in the domain of natural language processing. *Journal of Informetrics*, *14*(4), 101091.

Xue, Z., He, G., Liu, J., Jiang, Z., Zhao, S., & Lu, W. (2023). Re-examining lexical and semantic attention: Dual-view graph convolutions enhanced BERT for academic paper rating. *Information Processing & Management*, *60*(2), 103216.

Yan, Z., Zhou, X., Ren, J., Zhang, Q., & Du, R. (2023). Identifying underlying influential factors in information diffusion process on social media platform: A hybrid approach of data mining and time series regression. *Information Processing & Management*, *60*(5), 103438.

Yao, R., Ye, Y., Zhang, J., Li, S., & Wu, O. (2023). Exploring developments of the AI field from the perspective of methods, datasets, and metrics. *Information Processing & Management*, *60*(2), 103157.

Yu, F., Huang, J., Luo, Z., Zhang, L., & Lu, W. (2023). An effective method for figures and tables detection in academic literature. *Information Processing & Management*, *60*(3), 103286.

Zhang, Y., Wang, M., Zipperle, M., Abbasi, A., & Tani, M. (2023a). RelRank: A relevance-based author ranking algorithm for individual publication venues. *Information Processing & Management*, *60*(1), 103156.

Zhang, Y., Wu, M., Zhang, G., & Lu, J. (2023b). Stepping beyond your comfort zone: Diffusion‐based network analytics for knowledge trajectory recommendation. *Journal of the Association for Information Science and Technology*, *74*(7), 775-790.

Zhang, Y., Zhang, G., Zhu, D., & Lu, J. (2017). Scientific evolutionary pathways: Identifying and visualizing relationships for scientific topics. *Journal of the Association for Information Science and Technology*, *68*(8), 1925-1939.