Preface to Joint Workshop of the 4th Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2023) and the 3rd AI + Informetrics (AII2023) at JCDL 2023

Chengzhi Zhang[1], Yi Zhang[2], Philipp Mayr[3], Wei Lu[4], Arho Suominen[5], Haihua Chen[6], Ying Ding[7]

[1] *Nanjing University of Science and Technology, No. 200, Xiaolinvgwei, Nanjing, 210094, China*

[2] *Australian Artificial Intelligence Institute, University of Technology Sydney, 15 Broadway, Ultimo, NSW, Australia*

[3] *GESIS - Leibniz-Institute for the Social Sciences, Unter Sachsenhausen 6-8, Cologne, 50667, Germany*

[4] *Wuhan University, Luojiashan, Wuhan, 430072, China*

[5] *VTT Technical Research Centre of Finland, Espoo, FI-02044, Finland*

[6] *University of North Texas, Texas, Denton, Texas, 76201, USA*

[7] *University of Texas at Austin. Austin, Texas, 78712, USA*

Abstract

The Joint Workshop of the 4th Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2023; https://eeke-workshop.github.io/) and the 3rd AI + Informetrics (AII2023; https://ai-informetrics.github.io/) was held at Santa Fe, New Mexico, USA and online, co-located with the ACM/IEEE Joint Conference on Digital Libraries (JCDL) 2023. The two workshop series aim to engage the communities in open problems in the extraction and evaluation of knowledge entities from scientific documents and the modeling and applications of AI + Informetrics for broad interests in science of science, science, technology, & innovation, etc. This joint workshop comprises keynote speeches, oral presentations, and poster sessions. The main topics of the proceedings include entity extraction and its applications, along with the integration of Artificial Intelligence + Informetrics.

Keywords

Knowledge entity extraction, Knowledge entity evaluation, Scientific document, Informetrics, Artificial Intelligence

1. Introduction

Big data has dramatically revolutionized the thinking patterns and operational models of the human society. With great opportunities, the broad availability of information also brings new challenges, e.g., how can we obtain useful knowledge from numerous information sources? A knowledge entity is a relatively independent and integral knowledge module in a special discipline or a research domain [1]. Scientific documents, serving as a pivotal conduit for the dissemination of knowledge and teeming with rich knowledge entities, have garnered significant attention [2]. These knowledge entities encompass myriad elements such as methodological approaches, tasks, datasets, metrics, software, and tools, among others [3]. In parallel, informetrics, known as the study of quantitative aspects of information, has gained great benefits from artificial intelligence (AI), with its capacities in analyzing unstructured and scalable data and streams, understanding uncertain semantics, and developing robust and repeatable models. Incorporating informetrics with AI techniques has demonstrated enormous success in turning big data into big value and impact. For example, deep learning approaches enlighten studies of pattern recognition and further leverage time series to track technological change. However, how to effectively cohere the power of AI and informetrics to create cross-

disciplinary solutions is still elusive from neither theoretical nor practical perspectives [4]. Recently, Large Language Models (LLMs) have demonstrated significant potential in advancing science and technology, but how can we facilitate LLMs with relatively limited data and deliver interpretable results is still a challenge to the community.

The Joint Workshop of the 4th Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2023) and the 3rd AI + Informetrics (AII2023) was held at Santa Fe, New Mexico, USA and online, co-located with the ACM/IEEE Joint Conference on Digital Libraries (JCDL) on June 26, 2023. This workshop aims to engage the research communities to address open problems in the extraction and evaluation of knowledge entities from scientific documents and AI + Informetrics, cohering AI and informetrics to fulfill cross-disciplinary gaps from either theoretical or practical perspectives; elaborating AI-empowered informetric models with enhanced capabilities in robustness, adaptability, and effectiveness, and leveraging knowledge, concepts, and models in information management to strengthen the interpretability of AI + Informetrics to adapt to empirical needs in real-world cases [5].

2. Overview of the papers

This workshop received 17 submissions for peer review, and accepted 13 papers, which are collected in this proceeding. It includes 4 regular papers, 5 short papers, and 4 posters. The workshop also featured two keynote talks touching on different fields of EEKE and AII. All workshop contributions and slides have been documented on the EEKE/AII workshop website <https://eeke-workshop.github.io/2023/>. About one hundred people (including both online and offline attendees) attended the workshop. The following section briefly lists the 2 keynotes and the 13 accepted submissions.

2.1 Keynotes

The two keynotes in this EEKE-AII joint workshop touch LLMs and their impacts in scientometrics and information retrieval and extraction, respectively.

Professor Scott W. Cunningham (University of Strathclyde in Glasgow, UK) delivered a keynote on *Scientometrics in the Era of Large Language Models*.

Scott concerned about the emergence of LLMs, particularly ChatGPT, and the fundamental changes they might bring to scientometric research. Scott shared insights on why he believes such changes are crucial but challenging, and the potential ways to implement these research alterations.

Professor C. Lee Giles (Pennsylvania State University, USA) keynoted on *Large Language Models for Information Retrieval and Extraction*. He commented LLMs are revolutionizing the fields of information retrieval and extraction. Apart from their current use in search engine and ranking, LLMs can power intelligent virtual assistants to make academic research and learning efficiently, automate the summarization of lengthy documents, etc. Prof. Giles discussed what LLMs mean for information retrieval and extraction, and raised some open questions like what else can LLMs be used in information retrieval and extraction, and what are their capabilities and limitations?

2.2 Research papers and posters

We organized the 13 submissions in the following four sections.

2.2.1 Session 1: Posters

This session includes three posters: Zhang and Shi proposed a deep learning-based approach for identifying complementary patents [6]. Yan et al. utilized the papers published in JASIST journals from 2010 to 2020 for functional structure recognition [7]. Xu et al. proposed a main path analysis-based framework to discover the linkages among science, technology, and industry [8].

2.2.2 Session 2: Entity Extraction and Applications

Three papers are highlighted as follows.

Chen and Liu measured the clinical translation intensity of COVID-19 articles published in 2021 and tested the impact of interdisciplinary level and the characteristics of clinical translation intensity-related biological entities [9].

Würsch et al. used LLMs to extract relevant knowledge entities from cybersecurity-related texts [10].

Li and Yan proposed an AI-based method to automatically extract scientific method entities, and analyzed the specific situation of emerging technologies in the field of digital humanities [11].

2.2.3 Session 3: AI + Informetrics

This session collects 4 papers contributing to AI + Informetrics.

Hu et al. used a dynamic time warping algorithm to identify sleeping beauties from massive literature [12].

Chen et al. provided a new perspective to understand and measure the absorption and integration of scientific ideas and insights by leveraging knowledge networks [13].

Ningrum et al. proposed a weakly supervised technique that employs a fine-grained annotation scheme to construct a system for scientific uncertainty identification from scientific text [14].

Yu and Hua proposed a method called DictSentiBERT by adjusting the attention mechanism based on a sentiment dictionary, and applied it to the sentiment classification of scientific citations [15].

2.2.4 Session 4: EEKE + AII Onsite Session

This session includes three papers.

Wei et al. used supervised contrastive learning for scientific claim extraction, with a prior performance compared to 10 commonly used methods of text augmentation [16].

Park et al. proposed a new approach that combines topic modeling techniques and Graph convolutional networks (GCNs) for forecasting future topic trends in the blockchain domain [17].

Xu et al. conducted a search of 26 million AI-related articles from 2000-2019 and analyzed how AI assisted in the development of several selected scientific research domains [18].

3. Outlook and further reading

The EEKE and AII workshop series have achieved great success and received significant attentions from related research communities. The outcomes of this workshop series contributed novel technological developments and empirical insights to the literature.

Past proceedings can be accessed at http://ceur-ws.org/. We have organized three special issues on the topic of extraction and evaluation of knowledge entities in the *Journal of Data and Information Science*, *Data and Information Management*, *Aslib Journal of Information Management* and *Scientometrics* respectively. Two special issues have been published for the topic of AI + Informetrics, i.e., *Scientometrcis* and *Information Processing and Management*.

The EEKE-AII2023 organization committee is editing a Special Issues in *Journal of Informetrics*. For more information, please see https://eeke-workshop.github.io/2023/si-eeke.html.

Acknowledgements

References

[1] Chang, X., Zheng, Q. (2008). Knowledge Element Extraction for Knowledge-Based Learning Resources Organization. In: Leung, H., Li, F., Lau, R., Li, Q. (eds) Advances in Web Based Learning – ICWL 2007. ICWL 2007. Lecture Notes in Computer Science, vol 4823. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-78139-4_10

[2] Ding, Y., Song, M., Han, J., Yu, Q., Yan, E., Lin, L., & Chambers, T. (2013). Entitymetrics: Measuring the impact of entities. PloS one, 8(8), e71416. https://doi.org/10.1371/journal.pone.0071416

[3] Zhang, C., Mayr, P., Lu, W., & Zhang, Y. (2022). JCDL2022 workshop: extraction and evaluation of knowledge entities from scientific documents (EEKE2022). In Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries (JCDL '22). Association for Computing Machinery, New York, NY, USA, Article 54, 1–2. https://doi.org/10.1145/3529372.3530917

[4] Zhang C., Zhang Y., Mayr P., Lu W.,

Suominen A., Chen H., Ding Y. (2023). JCDL2023 Workshop: Joint Workshop of the 4th Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2023) and the 3rd AI + Informetrics (AII2023). Proceedings of the 23th ACM/IEEE Joint Conference on Digital Libraries (JCDL2023), Santa Fe, New Mexico, USA, https://doi.org/10.1109/JCDL57899.2023.00070

[5] Zhang, Y., Zhang, C., Mayr, P., & Suominen, A. (2022). An editorial of "AI + informetrics": multi-disciplinary interactions in the era of big data. Scientometrics 127, 6503–6507. https://doi.org/10.1007/s11192-022-04561-w

[6] Zhang J., Shi J. (2023). An Approach for Identifying Complementary Patents Based on Deep Learning. Proceeding of Joint Workshop of the 4th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 3rd AI +Informetrics (EEKE-AII2023), Santa Fe, New Mexico, USA and Online.

[7] Yan D., Shen S., Wang D. (2023). Functional Structure Recognition of Scientific Documents in Information Science. Proceeding of Joint Workshop of the 4th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 3rd AI + Informetrics (EEKE-AII2023), Santa Fe, New Mexico, USA and Online.

[8] Xu S., Liu Z., An X. (2023). Linkages among Science, Technology, and Industry. Proceeding of Joint Workshop of the 4th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 3rd AI + Informetrics (EEKE-AII2023), Santa Fe, New Mexico, USA and Online.

[9] Chen S., Liu C. (2023). The Impact of Interdisciplinarity and Entity Characteristics on the Clinical Translation Intensity of COVID-19 Papers. Proceeding of Joint Workshop of the 4th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 3rd AI + Informetrics (EEKE-AII2023), Santa Fe, New Mexico, USA and Online.

[10] Würsch M., Kucharavy A., Percia-David D., Mermoud A. (2023). LLM-Based Entity Extraction Is Not for Cybersecurity. Proceeding of Joint Workshop of the 4th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 3rd AI + Informetrics (EEKE-AII2023), Santa Fe, New Mexico, USA and Online.

[11] Li S., Yan C. (2023). Characterizing Emerging Technologies of Global Digital Humanities Using Scientific Method Entities. Proceeding of Joint Workshop of the 4th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 3rd AI + Informetrics (EEKE-AII2023), Santa Fe, New Mexico, USA and Online.

[12] Hu Z., Chen Y., Cui J. (2023). Identifying Potential Sleeping Beauties Based on Dynamic Time Warping Algorithm. Proceeding of Joint Workshop of the 4th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 3rd AI + Informetrics (EEKE-AII2023), Santa Fe, New Mexico, USA and Online.

[13] Chen H., Liu J., Liu Z. (2023). Scientific Knowledge Combination in Networks: New Perspectives on Analyzing Knowledge Absorption and Integration. Proceeding of Joint Workshop of the 4th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 3rd AI + Informetrics (EEKE-AII2023), Santa Fe, New Mexico, USA and Online.

[14] Ningrum P., Mayr P., Atanassova I. (2023). UnScientify: Detecting Scientific Uncertainty in Scholarly Full Text. Proceeding of Joint Workshop of the 4th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 3rd AI + Informetrics (EEKE-AII2023), Santa Fe, New Mexico, USA and Online.

[15] Yu D., Hua B. (2023). Sentiment Classification of Scientific Citation Based on Modified BERT Attention by Sentiment Dictionary. Proceeding of Joint Workshop of the 4th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 3rd AI + Informetrics (EEKE-AII2023), Santa Fe, New Mexico, USA and Online.

[16] Wei X., Hoque M., Wu J., Li J. (2023). ClaimDistiller: Scientific Claim Extraction with Supervised Contrastive Learning. Proceeding of Joint Workshop of the 4th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the

3rd AI + Informetrics (EEKE-AII2023), Santa Fe, New Mexico, USA and Online.

[17] Park Y., Lim S., Gu C., Song M. (2023). Forecasting Future Topic Trends in the Blockchain Domain: Using Graph Convolutional Network. Proceeding of Joint Workshop of the 4th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 3rd AI + Informetrics (EEKE-AII2023), Santa Fe, New Mexico, USA and Online.

[18] Xu Q., Meng J., He J., Lou W. (2023). How does AI assist scientific research domains? Evidence based on 26 millions research articles. Proceeding of Joint Workshop of the 4th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 3rd AI + Informetrics (EEKE-AII2023), Santa Fe, New Mexico, USA and Online.