Preface to the Joint Workshop of the 5th Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2024) and the 4th AI + Informetrics (AII2024)

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, Xiaolingwei, 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 5th Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2024; https://eeke-workshop.github.io/) and the 4th AI + Informetrics (AII2024; https://ai-informetrics.github.io/) was held in Changchun, China and online, co-located with the iConference2024. The two workshop series are designed to actively engage diverse communities in addressing open challenges related to the extraction and evaluation of knowledge entities from scientific documents and the modeling and applications of AI-empowered informetrics for broad interests in science of science, science, technology, & innovation, etc. The joint workshop features a comprehensive agenda, including keynotes from leading experts, oral presentations showcasing cutting-edge research, and poster sessions for in-depth discussions. The primary topics covered in the proceedings encompass the methodologies and applications of entity extraction, as well as the convergence of AI and informetrics, to drive advancements in these fields.

Keywords

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

1. Introduction

The rapid development of big data and artificial intelligence technologies is significantly driving changes in human society's thinking patterns and operational models. While presenting immense opportunities, the broad availability and comprehensibility of information also pose new challenges. For instance, how can we extract useful knowledge from numerous information sources?

In scientific documents, knowledge consists of many interconnected units known as knowledge entities [1]. Knowledge entities can be further subdivided; for example, in the field of natural language processing, they include models, algorithms, datasets, tools, metrics, and other fine-grained knowledge entities [2]. Extracting and analyzing knowledge entities is crucial for researchers. For instance, constructing knowledge entity maps can visualize research connections and help identify research trends [3]. Modeling citation functions can effectively assess entity impact in literature, enhancing scientific understanding [4].

At the same time, informetrics, as a discipline studying the quantitative aspects of information, has greatly benefited from artificial intelligence

(AI), particularly in analyzing unstructured and scalable data streams, understanding uncertain semantics, and developing robust and repeatable models. Combining informetrics with AI techniques has achieved tremendous success in turning big data into significant value and impact. For example, deep learning methods have inspired studies in pattern recognition and further leveraged time series to track technological changes [5]. However, how to effectively integrate the power of AI and informetrics to create cross-disciplinary solutions in line with this big data boom remains elusive from both theoretical and practical perspectives [6]. Lately, large-scale language models (LLMs) have been widely used across multiple fields. LLMs have shown powerful potential in knowledge entity extraction and evaluation [7]. However, how to facilitate LLMs with relatively limited data and deliver interpretable results remains a challenge for the community.

The Joint Workshop of the 5th Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2024) and the 4th AI + Informetrics (AII2024) was held in Changchun, China and online, co-located with the iConference2024 on April 23~24, 2024. This workshop aims to engage the research community in addressing open problems related to the extraction and evaluation of knowledge entities from scientific documents, with a focus on the integration of AI and informetrics. The goal is to bridge cross-disciplinary gaps from both theoretical and practical angles. The workshop will explore AI-empowered informetric models designed to improve robustness, adaptability, and effectiveness. Additionally, it will draw on knowledge, concepts, and models from information management to enhance the interpretability of AI-empowered informetrics, ensuring these technologies meet practical needs in real-world applications. This collaborative effort aspires to advance the field and offer innovative solutions [5].

2. Overview of the papers

This workshop received 46 submissions for peer review, and accepted 25 papers, which are collected in this proceeding. It includes 4 long papers, 9 short papers, and 9 power talks. The workshop also featured one keynote talk across the fields of EEKE and AII.

All contributions and slides in the workshop are available on the EEKE/AII workshop website <https://eeke-workshop.github.io/2024/>. The workshop attracted approximately 60 attendees, both online and offline. The following section provides a brief overview of the keynote and the 25 accepted submissions.

2.1 Keynote

The keynote in this EEKE-AII joint workshop highlights using AI for biomedical knowledge exploration and discovery.

Professor Karin Verspoor (Royal Melbourne Institute of Technology, Australia) delivered a keynote on *Opportunities for AI-enabled scientific knowledge exploration, analysis, and discovery*.

Karin concerned about the challenges of utilizing vast textual data in biomedicine, including scientific literature, clinical notes, and patents. She emphasized the importance of AI and natural language processing methods in structuring, organizing, and modeling the information. These technologies enable systematic reviews, protein function prediction, hypothesis generation, and various applications in biomedical and biochemical fields. Her work demonstrates how AI can transform unstructured natural language data into valuable resources for scientific exploration, analysis, and discovery.

2.2 Research papers and posters

We organized the 25 submissions in the following four sections.

2.2.1 Session 1: Technology Mining

This session includes five papers.

In their paper "Technological forecasting based on spectral clustering for word frequency time series", Huang et al. [8] presented a novel Time Trend Clustering Model (TTCM) based on spectral clustering for technological forecasting, demonstrating its effectiveness by analyzing the time series of word frequency in different testing datasets.

In the paper "Automated identification of emerging technologies: Open data approach", Dolamic et al. [9] introduced an automated quantitative method for identifying emerging technologies using publicly available data,

proposing four criteria (i.e., novelty, growth, impact, and coherence) to score technologies, and demonstrated its reliability and unique capabilities compared to leading market research reports.

In the paper "Technology convergence prediction from a timeliness perspective: An improved contribution index in a dynamic network", Zhang and Yan [10] introduced a dynamic technology convergence prediction model using a contribution index and graph neural networks, which improves prediction accuracy by considering timeliness and the importance of each technology, and demonstrated a case study in the field of new energy vehicles.

In the paper "A research topic evolution prediction approach based on multiplex-graph representation learning", Zheng et al. [11] introduced a contribution index and a dynamic technology network for improving the accuracy of technology convergence prediction, utilizing semantic similarity and graph neural networks, and demonstrated its effectiveness in the new energy vehicles field, while also presenting a method for automated research topic evolution prediction by integrating keyword content and structural features.

The last paper in this section is by Yan et al. [12], "Unveiling the secret of information rediffusion process on social media from information coupling perspective: a hybrid approach of machine learning and regression model", they modeled emotional, semantic, and cognitive information coupling on Sina Microblog to analyze their effects on user commenting and reposting behavior, and found that emotional and semantic coupling influence commenting, and cognitive and emotional coupling influence reposting, while opinion leaders moderate these relationships.

2.2.2 Session 2: Entity & Relation Extraction

This session includes five papers.

The work by Yuan et al. [13], entitled "Biomedical relation extraction via domain knowledge and prompt learning" proposes a biomedical relation extraction model based on domain knowledge and prompt learning to enhance understanding of technical language and improve classification accuracy in imbalanced datasets, achieving state-of-the-art performance

on the DDI Extraction 2013 and ChemProt datasets.

In the paper "Identifying scientific problems and solutions: Semantic network analytics and deep learning", Huang et al. [14] proposed a novel method for identifying scientific problems and solutions using semantic network analytics and deep learning, combining the BERT-CRF model with BIO tagging and the Levenshtein algorithm to construct a comprehensive knowledge network, and demonstrated the reliability in a case study in the artificial intelligence domain.

In the paper "Material performance evolution discovery based on entity extraction and social circle theory", Zhang and Sun [15] presented a method for accurately extracting material performance entities and constructing dynamic evolution paths for material performance topics using a BERT-BiLSTM-CRF model and a novel algorithm, and demonstrated through experiments in the field of metal materials to enhance the understanding of topic evolution.

In "revealing the country-level preference on research methods in the field of digital humanities: From the perspective of library and information science", Yan and Fang [16] proposed a multi-stage recognition algorithm combining large language models and iterative learning to extract research methods mentioned from digital humanities documents, map them to existing taxonomies, then, analyzed country-level preferences, and revealed the central role of quantitative research and distinct international variations.

The paper by Sternfeld et al. [17], entitled "LLM-resilient bibliometrics: Factual consistency through entity triplet extraction", proposes a method to mitigate the misuse of LLMs in academic paper mills by extracting and validating semantic entity triplets from scientific papers, ensuring factual consistency and penalizing blind usage of LLMs while maintaining readability improvements.

2.2.3 Session 3: Power Talk

This session collects nine power talks.

In "How to measure information cocoon in academic environment", Yuan et al. [18] introduced a method to measure academic information cocoons, showing decreasing trends and significant disciplinary differences, using BERTopic and Sentence-BERT.

In "May generative AI be a reviewer on an academic paper?", Zhou et al. [19] evaluated Generative AI's ability to perform academic evaluation compared to human experts, finding GenAI's score higher and comments less substantive.

In "Research on the Identification of breakthrough technologies driven by science", Wang et al. [20] presented a novel framework for identifying breakthrough technologies using a science-driven pattern, validated in artificial intelligence.

In "Connector and provincial hub dichotomy in scientific collaborations identified by reinforcement learning algorithm", Liu et al. [21] used deep reinforcement learning to identify complex cross-community collaboration patterns in physics co-authorship networks, revealing multi-core structures and enhancing understanding of scientific collaboration dynamics.

In "Research on named entity recognition from patent texts with local large language model", Yu et al. [22] proposed a framework using large language models and prompt templates for named entity recognition in patent texts, demonstrating superior few-shot learning performance.

In "IRUGCN: A graph convolutional network rumor detection model incorporating user behavior", Zhou et al. [23] presented a novel rumor detection model using user behavior and traditional features, achieving superior accuracy with graph convolutional and recurrent neural networks on Twitter datasets.

In "Identification of core technological topics in the new energy vehicle industry: The SAO-BERTopic topic modeling approach based on patent text mining", Zhu et al. [24] proposed a comprehensive approach using the information weight method and SAO-BERTopic model to identify core technologies in the new energy vehicle industry from large-scale patent data.

In "Research on fine-grained s&t entity identification with contextual semantics in think-tank text", Sun et al. [25] proposed an automatic method to extract fine-grained S&T problems from think-tank reports using LLMs for annotation and a RoBERTa-BiLSTM-CRF model, achieving an F1 score of 86.02%.

In the power talk "Biomedical association inference on pandemic knowledge graphs: A comparative study", Wu et al. [26] constructed a pandemic-focused knowledge graph and evaluated methodologies for biomedical association inference, finding that graph representation learning techniques show significant promise and high predictive accuracy.

2.2.4 Session 4: AI for informetrics

This session includes six papers.

The work by Zhang et al. is titled "Understanding citation mobility in the knowledge space " [27]. This study analyzes the spatial patterns of citation dynamics in physics, finding constrained citation mobility influenced by epistemic distance and popularity, with disruptive papers receiving more distant recognition and contemporary papers exhibiting narrower citation mobility.

In the paper "Relationship between team diversity and innovation performance in interdisciplinary research teams within the field of artificial intelligence: Decision tree analysis", Liu et al. [28] used the CART model to examine the non-linear relationship between diverse factors and innovation performance in interdisciplinary AI research teams, revealing a U-shaped relationship between activity diversity and "novelty" innovation performance, significantly influenced by research interest diversity.

In "Understanding partnership in scientific collaborations: A preliminary study from the paper-level perspective", Lu et al. [29] examined scientific collaboration by analyzing over 120,000 biology research articles, revealing common division of labor and partnerships among collaborators, highlighting internal interactions often overlooked in co-authorship studies.

In "Quantifying scientific novelty of doctoral theses with Bio-BERT model", Yang et al. [30] presented a methodology using the Bio-BERT model to quantify the scientific novelty of biomedical doctoral theses by analyzing bio-entity combinations and calculating semantic distances, offering a novelty score for each thesis.

In "Are disruptive patents less likely to be granted? Analyzing scientific gatekeeping with USPTO patent data (2004-2018)", Yan et al. [31] analyzed how scientific gatekeeping in the US Patent and Trademark Office affects disruptive innovation, revealing that disruptive innovation faces challenges in approval, but examiner workload and experience can mitigate these challenges, offering insights for more innovation-friendly patent examination processes.

In the paper "Open-mentorship team is beneficial to disruptive ideas", Zheng et al. [32] analyzed 361,189 neuroscience publications to

explore the impact of close vs. open mentorship on publication disruption, finding that open-mentorship collaborations are more disruptive, with implications for team formation and management.

3. Outlook and further reading

The EEKE and AII workshop series have been highly successful and garnered substantial attention from the research communities. This workshop series has made significant contributions to the literature by introducing innovative technological advancements and valuable empirical insights.

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-AII2024 organization committee is editing a Special Issues in *Technological Forecasting and Social Change*. For more information, please see https://eeke-workshop.github.io/2024/si-eeke-aii.html.

Acknowledgements

References

[1] 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

[2] Chen, Z., Zhang, C., Zhang, H., Zhao, Y., Yang, C., & Yang, Y. (2024). Exploring the relationship between team institutional composition and novelty in academic papers based on fine-grained knowledge entities. The Electronic Library. https://doi.org/10.1108/EL-03-2024-0070

[3] Zha, H., Chen, W., Li, K., & Yan, X. (2019, July). Mining algorithm roadmap in scientific publications. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 1083-1092). https://doi.org/10.1145/3292500.3330913

[4] Wang, Y., Xiang, Y., & Zhang, C. (2024). Exploring motivations for algorithm mention in the domain of natural language processing: A deep learning approach. Journal of Informetrics, 18(4), 101550. https://doi.org/10.1016/j.joi.2024.101550

[5] Zhang, C., Zhang, Y., Mayr, P., Lu, W., Suominen, A., Chen, H., & Ding, Y. (2023, January). 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. In CEUR Workshop Proceedings. https://ceur-ws.org/Vol-3451/Preface.pdf

[6] 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(11), 6503-6507. https://doi.org/10.1007/s11192-022-04561-w

[7] Dagdelen J, Dunn A, Lee S, Walker N, Rosen AS, Ceder G, Persson KA, Jain A. (2024). Structured information extraction from scientific text with large language models. Nature Communications, 15(1), 1418. https://doi.org/10.1038/s41467-024-45563-x

[8] Huang, H., Wang, X., & Wang, H. (2024). Technological Forecasting Based on Spectral Clustering for Word Frequency Time Series. Proceeding of Joint Workshop of the 5th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 4th AI + Informetrics (EEKE-AII2024), Changchun, China and online.

[9] Dolamic, L., Jang-Jaccard, J., Mermoud, A., & Lenders, V. (2024). Automated Identification of Emerging Technologies: Open Data Approach. Proceeding of Joint

Workshop of the 5th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 4th AI + Informetrics (EEKE-AII2024), Changchun, China and online.

[10] Zhang J., & Yan B. (2024). Technology Convergence Prediction from a Timeliness Perspective: An Improved Contribution Index in a Dynamic Network. Proceeding of Joint Workshop of the 5th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 4th AI + Informetrics (EEKE-AII2024), Changchun, China and online.

[11] Zheng, Y., Shi, K., Dong, Y., Wang, X., & Wang, H. (2024). A research topic evolution prediction approach based on multiplex-graph representation learning. Proceeding of Joint Workshop of the 5th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 4th AI + Informetrics (EEKE-AII2024), Changchun, China and online.

[12] Yan, Z., Du, R., & Wang, H. (2024). Unveiling the secret of information rediffusion process on social media from information coupling perspective: a hybrid approach of machine learning and regression model. Proceeding of Joint Workshop of the 5th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 4th AI + Informetrics (EEKE-AII2024), Changchun, China and online.

[13] Yuan, J., Du, W., Liu, X., & Zhang, Y. (2024). Biomedical Relation Extraction via Domain Knowledge and Prompt Learning. Proceeding of Joint Workshop of the 5th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 4th AI + Informetrics (EEKE-AII2024), Changchun, China and online.

[14] Huang, L., Cao, X., Ren, H., Zhang, C., & Wu, Z. (2024). Identifying scientific problems and solutions: Semantic network analytics and deep learning. Proceeding of Joint Workshop of the 5th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 4th AI + Informetrics (EEKE-AII2024), Changchun, China and online.

[15] Zhang J., Sun W. (2024). Material performance evolution discovery based on entity extraction and social circle theory. Proceeding of Joint Workshop of the 5th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 4th AI + Informetrics (EEKE-AII2024), Changchun, China and online.

[16] Yan, C., & Fang, Z. (2024). Revealing the Country-level Preference on Research Methods in the Field of Digital Humanities: From the Perspective of Library and Information Science. Proceeding of Joint Workshop of the 5th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 4th AI + Informetrics (EEKE-AII2024), Changchun, China and online.

[17] Sternfeld, A., Kucharavy, A., David, D. P., Mermoud, A., & Jang-Jaccard, J. (2024). LLM-Resilient Bibliometrics: Factual Consistency Through Entity Triplet Extraction. Proceeding of Joint Workshop of the 5th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 4th AI + Informetrics (EEKE-AII2024), Changchun, China and online.

[18] Yuan, J., He, G., & Yang, Y. (2024). How to Measure Information Cocoon in Academic Environment. Proceeding of Joint Workshop of the 5th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 4th AI + Informetrics (EEKE-AII2024), Changchun, China and online.

[19] Zhou, H., Huang, X., Pu, H., & Qi, Z. (2024). May Generative AI Be a Reviewer on an Academic Paper? Proceeding of Joint Workshop of the 5th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 4th AI + Informetrics (EEKE-AII2024), Changchun, China and online.

[20] Wang, D., Zhou, X., Zhao, P., Pang, J., & Ren, Q. (2024). Research on the Identification of breakthrough technologies driven by science. Proceeding of Joint Workshop of the 5th Extraction and Evaluation of Knowledge Entities from

Scientific Documents and the 4th AI + Informetrics (EEKE-AII2024), Changchun, China and online.

[21] Liu, F., Zhang, S., & Xia, H. (2024). Connector and Provincial Hub Dichotomy in Scientific Collaborations Identified by Reinforcement Learning Algorithm. Proceeding of Joint Workshop of the 5th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 4th AI + Informetrics (EEKE-AII2024), Changchun, China and online.

[22] Yu, C., Chen, L., & Xu, H. (2024). Research on Named Entity Recognition from Patent Texts with Local Large Language Model.

[23] Zhou, S., Wang, H., Zhou, Z., Yi, H., & Shi, B. (2024). IRUGCN: A Graph Convolutional Network Rumor Detection Model Incorporating User Behavior. Proceeding of Joint Workshop of the 5th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 4th AI + Informetrics (EEKE-AII2024), Changchun, China and online.

[24] Zhu J., Chuang Y., Wang Z., Li Y. (2024). Identification of core technological topics in the new energy vehicle industry: The SAO-BERTopic topic modeling approach based on patent text mining. Proceeding of Joint Workshop of the 5th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 4th AI + Informetrics (EEKE-AII2024), Changchun, China and online.

[25] Sun, M., Wang, Y., & Zhao, Y. (2024). Research on Fine-grained S&T Entity Identification with Contextual Semantics in Think-Tank Text. Proceeding of Joint Workshop of the 5th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 4th AI + Informetrics (EEKE-AII2024), Changchun, China and online.

[26] Wu, M., Yu, C., Xu, J., Ding, Y., & Zhang, Y. (2024). Biomedical association inference on pandemic knowledge graphs: A comparative study. Proceeding of Joint Workshop of the 5th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 4th AI +

Informetrics (EEKE-AII2024), Changchun, China and online.

[27] Zhang S., Liu F., Xia H. (2024). Understanding Citation Mobility in the Knowledge Space. Proceeding of Joint Workshop of the 5th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 4th AI + Informetrics (EEKE-AII2024), Changchun, China and online.

[28] Liu, J., Huang, C., & Xu, S. (2024). Relationship between Team Diversity and Innovation Performance in Interdisciplinary Research Teams within the Field of Artificial ntelligence: Decision Tree Analysis. Proceeding of Joint Workshop of the 5th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 4th AI + Informetrics (EEKE-AII2024), Changchun, China and online.

[29] Lu C., Li M., Zhou C. (2024). Understanding Partnership in Scientific Collaborations: A Preliminary Study from the Paper-level Perspective. Proceeding of Joint Workshop of the 5th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 4th AI + Informetrics (EEKE-AII2024), Changchun, China and online.

[30] Yang, A. J., Bu, Y., Ding, Y., & Liu, M. (2024). Quantifying scientific novelty of doctoral theses with Bio-BERT model. Proceeding of Joint Workshop of the 5th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 4th AI + Informetrics (EEKE-AII2024), Changchun, China and online.

[31] Yan, L., Cui, H., & Wang, C. J. (2024). Are Disruptive Patents Less Likely to be Granted? Analyzing Scientific Gatekeeping with USPTO Patent Data (2004-2018). Proceeding of Joint Workshop of the 5th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 4th AI + Informetrics (EEKE-AII2024), Changchun, China and online.

[32] Zheng, B., Li, W., & Hou, J. (2024). Open-mentorship team is beneficial to disruptive ideas. Proceeding of Joint Workshop of the

5th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 4th AI + Informetrics (EEKE-AII2024), Changchun, China and online.