Guest editorial: Extraction and evaluation of knowledge entities in the age of artificial intelligence

Chengzhi Zhang[1], Philipp Mayr[2], Wei Lu[3], Yi Zhang[4]

1. Department of Information Management, Nanjing University of Science and Technology, Nanjing, China
2. GESIS-Leibniz Institute for the Social Sciences, Cologne, Germany
3. School of Information Management, Wuhan University, Wuhan, China
4. Australian Artificial Intelligence Institute, Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney , Australia

1. Introduction

Scientific documents serve as an essential mediator for research achievements and scientific knowledge. With the increasing availability of full-text data and advanced data analytical techniques, bibliometric researchers have started to focus on granular content within scientific documents, transitioning their foci from the external metadata of these documents to the internal knowledge they encompass~~bibliometric researchers have started to focus on granular content within scientific documents by shifting their foci from external objects in scientific documents to internal knowledge~~. In scientific documents, knowledge consists of many interconnected units, known as knowledge entities (Ding et al., 2023).

Ding et al. (2023) classified knowledge entities into macro-level (e.g., author, journal, references), meso-level (e.g., keywords), and micro-level (e.g., dataset, method, biomedical entities) categories. These categories refer to traditional bibliographic entities for evaluation purposes and knowledge entities found within the full-text content, respectively.~~Ding et al. (2023) categorized knowledge entities into macro-, meso-, and micro-levels, referring to traditional bibliographic entities for evaluation purposes and knowledge entities in full-text content, respectively.~~ The extraction and evaluation of knowledge entities can improve existing knowledge services and meet the needs of researchers for rapid and accurate access to scientific knowledge (Ma et al., 2023; Mayr et al., 2014). Scholars can evaluate the scientific, social, and economic attributes of knowledge and assess its role in science, technology, innovation, and policy. Analyzing dynamic changes over time and identifying geographical differences can provide insights to understand the patterns of knowledge use and dissemination, which can promote knowledge discovery and generation activities~~ (Ding et al., 2023)~~.

The 2nd Workshop on the Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2021) was held online, co-located with the ACM/IEEE Joint Conference on Digital Libraries 2021 (JCDL2021) on September 30, 2021. The EEKE workshop series (https://eekeworkshop.github.io/) aims to engage relevant

Formatted: Indent: First line: 1 ch

communities in addressing open problems in the extraction and evaluation of knowledge entities from scientific documents. Participants has contributed tremendous accomplishments in identifying knowledge entities, exploring entity features, analyzing entity relationships, and developing extraction platforms or knowledge bases. The workshop has provided scholars, especially early career researchers, with knowledge recommendations and other knowledge entity-based services (Zhang et al., 2021)

2. Topics in this special issue

This special issue collected articles presented at the EEKE2021 and external submissions. A total of 18 papers were submitted, of which 9 were accepted after a rigorous review process. These 9 articles, contributed by 36 authors from 4 countries (China, USA, South Korea, and Thailand), are included in this collection. In this editorial, we provide an overview of these papers and introduce the themes and contributions of this Special Issue. The papers are grouped into three topics: entity extraction and entity relations extraction (3 articles), annotation tools and knowledge entity graph construction (2 articles), and applications of knowledge entities (4 articles).

2.1 Entity extraction and entity relations extraction

Knowledge entities such as concepts, algorithms, and methods are included in scientific literature (Ding et al. 2013; Wang et al. 2022). Extracting knowledge entities and entity relationships from scientific literature plays a crucial role in scientific information retrieval, fine-grained scientific information recommendation, and scientific evaluation.

To extract band gap information from academic papers, Ghosh & Lu (2023) collected 1.44 million titles and abstracts of scholarly articles related to materials science. They filtered the collection to 11,939 articles that may contain core information on materials and their band gap values. The results revealed that the current system can accurately extract information from 51.32% of the articles, partially extract from 36.62% of the articles, and incorrectly extract from 12.04% of the articles.

To tackle the issues of insufficient training corpus and the quality control of annotations, Yan et al. (2023) proposed a novel integrated solution for Chinese historical Named Entity Recognition (NER), including automatic entity extraction and man-machine cooperative annotation. This solution is valuable for enhancing the effectiveness of Chinese historical NER and promoting the development of low-resource information extraction.

"Problem-solving" stands out as one of the most fundamental and critical insights of scientific research. Using computer vision as an example, Chen et al. (2023) built a "problem-solving" knowledge graph of scientific domains by extracting four entity relation types, namely problem-solving, problem hierarchy, solution hierarchy, and association. They illustrated the utility of the extracted relations in constructing domain knowledge graphs and uncovering historical research trends.

2.2 Annotation tools and the construction of knowledge entity graphs

Knowledge entity annotation and knowledge graph construction are essential tasks for knowledge entity applications, including the development of annotation tools, corpora construction, and the creation of knowledge graphs.

To explore the use of entity descriptions and network structures in enhancing knowledge graph completion with a high generalization ability across datasets, Yu et al. (2023) proposed an entity-description augmented knowledge graph completion model (EDA-KGC). The authors conducted extensive experiments on the FB15K, WN18, FB15K-237, and WN18RR datasets to validate the effectiveness of the model.

To address challenges posed by growing volumes of pre-annotated literature and diverse annotations, such as teamwork, quality control, and time management, Wang et al. (2023) developed the Bureau for Rapid Annotation Tool (Brat), an annotation collaboration workbench that includes an enhanced semantic constraint system, Vim-like shortcut keys, an annotation filter, and a graph-visualizing annotation browser.

2.3 Applications of knowledge entities

Knowledge entities could seed not only granular information retrieval and recommendation (Mayr et al., 2014), intelligent bibliometrics (Zhang et al., 2020), but also new cognitive and practical applications such as the innovative evaluation of scientific document (Liu et al., 2022) and predicting future research directions (Zhang et al., 2023).

To explore mental health information entities and the connections between the biomedical, psychological, and social domains of bipolar disorder (BD), Timakum et al. (2023) used Reddit posts and full-text papers from PubMed Central to extract BD entities and their relationships in the datasets using a dictionary-based and rule-based approach. The findings indicate that the drug side effects entity was frequently identified in both datasets as a mental health information entity.

Zhang et al. (2023) introduced a crucial factor in topic selection, i.e., topic popularity, to investigate its relationship with team performance. The authors used gene/protein entities as proxies for topics and extracted them to monitor the development of topic popularity. By comparing various dimensions of team performance, the study explored the relationship between the phase of selected topic popularity and the academic performance of research teams.

Zeng et al. (2023) utilized the knowledge elements extracted through the Lexicon-LSTM model to measure the interdisciplinary characteristics of Chinese research in library and information science (LIS). They constructed a subject knowledge graph to support the searching and classification of knowledge elements. The results showed that in LIS, the interdisciplinary diversity indicator exhibited an upward trend from 2011 to 2021, while the disciplinary balance and difference indicators showed a downward trend.

Focusing on discovery of topic evolution path and semantic relationship, Zhang et al. (2023) identified entities that have the same semantics but different expressions for accurate topic evolution path discovery. They also revealed semantic relationships of topic evolution to better understand what leads to topic evolution. This work provided

a new perspective for topic evolution analysis by considering the semantic representation of patent entities.

3. Conclusions

The EEKE workshop series and this special issue reinforced the increasing interest of the community in the extraction and evaluation of entities from scientific literature. Endeavors in this field are advancing rapidly, from developing new methods and techniques to their innovative applications in broad practical domains. In today's AI age, three key factors that are driving AI development - namely, data, algorithms, and computing power - must work corporately to promote and support one another to achieve success and create value.

Currently, Large-scale pre-trained Language Models (LLMs), such as ChatGPT and GPT-4, have gained widespread popularity due to their broad applications across various industries and fields~~, with a critical role in scientific literature~~. There's no doubt that LLMs play an exceptionally important role in scientific documents mining and analysis.. Some scholars have tested and evaluated the performance of ChatGPT and GPT-4 in entity extraction (Hu et al., 2023; González-Gallardo et al., 2023) and entity relationship recognition (Rehana et al., 2023) in special domains. With AI's continued advancements and the increasing promotion of open access movements, we can anticipate a more extensive utilization of LLMs, such as GPT (Generative Pretrained Transformer), for extraction and evaluation of knowledge entities from scientific documents~~scientific literature , entity extraction and evaluation~~ in the future.

Acknowledgements

Chengzhi Zhang acknowledges the National Natural Science Foundation of China (Grant No. 72074113 and Yi Zhang acknowledges the Discovery Early Career Researcher Award granted by the Australian Research Council (Grant No. DE190100994).

References

Chen, G., Peng, J., Xu, T. and Xiao, L. (2023), "Extracting entity relations for "problem-solving" knowledge graph of scientific domains using word analogy", Aslib Journal of Information Management. https://doi.org/10.1108/AJIM-03-2022-0129

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

Ghosh, S. and Lu, K. (2023), "Band gap information extraction from materials science literature – a pilot study", Aslib Journal of Information Management. https://doi.org/10.1108/AJIM-03-2022-0141

González-Gallardo, C. E., Boros, E., Girdhar, N., Hamdi, A., Moreno, J. G., & Doucet, A. (2023). "Yes but.. Can ChatGPT Identify Entities in Historical Documents?", arXiv preprint arXiv:2303.17322. https://doi.org/10.48550/arXiv.2303.17322

Hu, Y., Ameer, I., Zuo, X., Peng, X., Zhou, Y., Li, Z., Li Y., Li J., Jiang X., & Xu, H. (2023). "Zero-shot Clinical Entity Recognition using ChatGPT", arXiv preprint arXiv:2303.16416. https://doi.org/10.48550/arXiv.2303.16416

Liu, M., Bu, Y., Chen, C., Xu, J., Li, D., Leng, Y., Freeman, R. B., Meyer, E. T., Yoon, W., Sung, M., Jeong, M., Lee, J., Kang, J., Min, C., Song, M., Zhai, Y., & Ding, Y. (2022). "Pandemics are catalysts of scientific novelty: Evidence from COVID-19", Journal of the Association for Information Science and Technology, 73(8), 1065–1078. https://doi.org/10.1002/asi.24612

Ma, Y., Liu, J., Lu, W., & Cheng, Q. (2023). "From 'what' to 'how': Extracting the Procedural Scientific Information toward the Metric-optimization in AI", Information Processing & Management, 60(3), 103315. https://doi.org/10.1016/j.ipm.2023.103315

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, Amsterdam, the Netherlands. https://doi.org/10.1007/978-3-319-06028-6_99

Rehana, H., Çam, N. B., Basmaci, M., He, Y., Özgür, A., & Hur, J. (2023). "Evaluation of GPT and BERT-based models on identifying protein-protein interactions in biomedical text", arXiv preprint arXiv:2303.17728. https://doi.org/10.48550/arXiv.2303.17728

Timakum, T., Song, M. and Kim, G. (2023), "Integrated entitymetrics analysis for health information on bipolar disorder using social media data and scientific literature", Aslib Journal of Information Management. https://doi.org/10.1108/AJIM-02-2022-0090

Wang, Y., Zhang, C., & Li, K. (2022). "A review on method entities in the academic literature: extraction, evaluation, and application", Scientometrics, 127(5), 2479-2520. https://doi.org/10.1007/s11192-022-04332-7

Wang, Z., Xu, S., Wang, Y., Chai, X. and Chen, L. (2023), "Bureau for Rapid Annotation Tool: collaboration can do more among variance annotations", Aslib Journal of Information Management. https://doi.org/10.1108/AJIM-01-2022-0046

Yan, C., Tang, X., Yang, H. and Wang, J. (2023), "A deep active learning-based and crowdsourcing-assisted solution for named entity recognition in Chinese historical corpora", Aslib Journal of Information Management. https://doi.org/10.1108/AJIM-03-2022-0107

Yu, C., Zhang, Z., An, L. and Li, G. (2023), "A knowledge graph completion model integrating entity description and network structure", Aslib Journal of Information Management. https://doi.org/10.1108/AJIM-01-2022-0031

Zeng J., Chen Y., Cao S., Pan P. and Cai Y. (2023), "Measuring the Interdisciplinary Characteristics of Chinese Research in Library and Information Science based on Knowledge Elements", Aslib Journal of Information Management. (Accepted)

Zhang, C., Mayr, P., Lu, W., & Zhang, Y. (2021). "Preface to the 2nd Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents at JCDL 2021", In Proceedings of the 2nd Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE 2021) co-located with JCDL 2021 (pp. 1-4). https://ceur-ws.org/Vol-3004/preface.pdf

Zhang, C., Xiang, Y., Hao, W., Li, Z., Qian, Y., & Wang, Y. (2023). "Automatic recognition and classification of future work sentences from academic articles in a specific domain". Journal of Informetrics, 17(1), 101373. https://doi.org/10.1016/j.joi.2022.101373

Zhang, J., Liu, Y., Jiang, L. and Shi, J. (2023), "Discovery of topic evolution path and semantic relationship based on patent entity representation", Aslib Journal of Information Management. https://doi.org/10.1108/AJIM-03-2022-0124

Zhang, T., Tan, F., Yu, C., Wu, J. and Xu, J. (2023), "Understanding relationship between topic selection and academic performance of scientific teams based on entity popularity trend", Aslib Journal of Information Management. https://doi.org/10.1108/AJIM-03-2022-0135

Zhang, Y., Porter, A., Cunningham, S. W., Chiavetta, D., & Newman, N. (2020b). "Parallel or intersecting lines? Intelligent bibliometrics for investigating the involvement of data science in policy analysis", IEEE Transactions on Engineering Management, 68(5), 1259–1271. https://doi.org/10.1109/TEM.2020.2974761