Scientometrics (2023) 128:1465–1472
https://doi.org/10.1007/s11192-022-04628-8

Machine learning and artificial intelligence for science, technology, innovation mapping and forecasting: Review, synthesis, and applications

Daniel Hain[1] · Roman Jurowetzki[1] · Sungjoo Lee[2] · Yuan Zhou[3]

Received: 22 December 2022 / Accepted: 22 December 2022 / Published online: 23 January 2023

Introduction

Current methodological advances have substantially broadened the toolbox for scientometric analysis. The application of natural language processing (NLP) techniques enables us to leverage unstructured text data from traditional (e.g., academic publications, patents) as well as novel (e.g., web-scraping, social media, online news) data sources. Advances in supervised machine learning (ML) offer opportunities for technology forecasting and rare-event prediction (e.g., prediction of breakthrough inventions, technology emergence, or convergence), and the application of deep learning (DL) techniques enables the utilization of complex data structures and the modeling of complex relationships.

The present special issue structures and showcases the use of NLP, ML, and DL methods to explore different research questions related to mapping science, technology, and innovation (STI) (eco-)systems, as well as forecasting their future development. Included contributions provide a broad overview of the current methodological state-of-the-art. All include a summary of methods historically applied to tackle a specific research question, a discussion on their advantages and disadvantages, guidelines on their application, and an evaluation of their performance. These methods are illustrated in applications related to emerging technologies, which are due to a lack of historical data as well as the dynamics and volatility in development particularly challenging to map and forecast. Likewise, the special issue features the utilization of a broad range

✉ Daniel Hain
dsh@business.aau.dk

Roman Jurowetzki
roman@business.aau.dk

Sungjoo Lee
sungjoolee@snu.ac.kr

Yuan Zhou
zhou_yuan@mail.tsinghua.edu.cn

[1] Aalborg University, Aalborg, Denmark

[2] Seoul National University, Seoul, South Korea

[3] Tsinghua University, Beijing, China

of data sources and their combination, ranging from traditional scientometric sources such as academic publications and patents to sources more related to the application of science and technology in the industry. It particularly highlights approaches combining structured as well as unstructured data.

Papers and themes

The contributions in this special issue propose methodological advances on various layers and are linked to several themes within the broader scope of studying STI ecosystems. Figure 1 structures them according to the layer of analysis as well as the functional area of their main contribution.

An STI ecosystem is often stylized as spanning over at least the three broad layers of (1) science (2) technology and (3) business. Here, key issues of interest are the mapping of constituent elements, their linkages, development over time as well as connections between levels. The left part of Fig. 1 aims at depicting that. The corresponding methodological innovations proposed in this issue can be broadly summarized into those (i) proposing new data or infrastructure, (ii) new techniques for mapping, or (iii) innovative approaches to identifying specific elements as well as forecasting. While mapping can be mainly seen as an approach to understanding structures in the past or the present, forecasting (and to some extent identification exercises) are inherently future-oriented.

Most contributions focus on the technology layer, yet establish linkages to at least one of the other layers. This may suggest that the interaction between the layers of the ecosystem offers more potential for introducing innovative methodologies.

Broadly, the papers included in the special issue mainly aim at advancing techniques to (i) map STI ecosystems, (ii) identifying emerging and promising technologies, (iii) forecast technology convergence, (iv) map science and technology development, and (v) technology roadmapping.

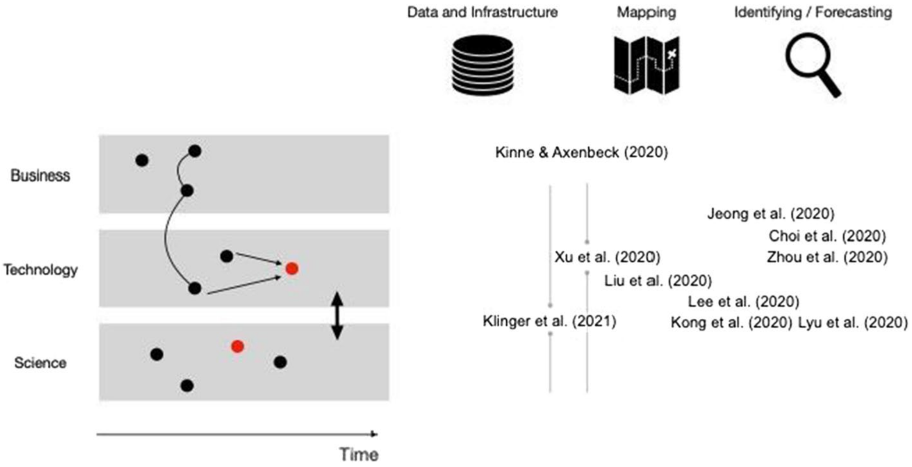

Fig. 1 Analytic dimensions of scientometric research

STI ecosystem mapping

Technology evolves and interacts with actors and other elements of the system it is embedded in multiple ways. Typical data sources used for scientometric mapping and forecasting such as patent and publication data tend to emphasize certain layers and elements of the system. Consequently, a more holistic technology analysis, forecasting, and mapping benefits from methods and frameworks crossing system and data source boundaries.

A core theme of this special issue is related to the holistic and multi-source mapping of technology ecosystems. To start with, Xu et al. (2020) introduce the concept of Multi-layered Innovation Ecosystem Mapping (MIEM) based on multi-source heterogeneous data. Focusing on the knowledge ecosystem based on patent data and the business ecosystem based on business transaction data, they first map interaction patterns within the separate ecosystems. In the next step, they identify linkages between the two ecosystems by matching actors co-occurring in both. If the same type of actors or elements can be identified and matched across ecosystems and data sources, such an actor-centric approach enables the identification of key players and their strategic roles in shaping the ecosystem's joint development.

Yet, for comprehensive technology mapping and forecasting of emerging technologies, traditional scientometric data sources even in combination still have shortcomings in terms of timeliness, granularity, and coverage. Kinne & Axenbeck (2020) demonstrate how web mining can be used for large-scale ecosystem mapping and forecasting by analyzing the textual and relational content of firm websites. They demonstrate how web-scraped textual information in combination with predictive NLP models can complement or even substitute traditional scientometric and survey-based indicators and allow for timely and granular scientometric mapping.

Emerging and promising technologies identification

In technology forecasting activities, expert evaluations are usually time-consuming, costly, and sometimes biased, they are rarely feasible to carry out on a large scale and timely manner—machine-learning applications appear to cope with these issues. However, challenges do exist. Many features of technology regarding its future potential are difficult to or partially identifiable using common indicators but largely rely on the evaluation based on complex qualitative perceptions of domain experts. On the one hand, some argue that these limitations are rooted or at least partially in the lack of multi-sourced heterogeneous data to better measure the multifaceted aspects of technology. On the other hand, some propose integrating the expert's opinion and those data-based estimations.

Zhou et al. (2020) develop a novel deep-learning-based framework to identify emerging technologies on a timely basis throughout large-scale samples when measuring a set of critical features using multi-sourced data. First, they integrate the patent data and website articles to construct the multi-sourced heterogeneous database to measure the various characteristics of potential emerging technologies, specifically in outlier patents, which may reflect their multifaceted impacts in both technological and social aspects. Second, they develop the deep-learning models by training samples to fit the hidden complex relationship between the indicators of patent outliers and their impacts. Then, specifically in the computer-based numerical control (CNC) machine-tool industry, they collect a multi-source dataset to validate their framework. Finally, they claim that their framework can be generalized to other emerging technologies in myopia sectors; in addition, broader data sources can also help to better identify emerging technologies when measuring additional essential aspects.

Choi et al. (2020) present a semi-supervised active learning framework that enables a limited number of expert opinions to be used effectively in identifying promising technologies, making a balance between data- and expert-driven decision-making.

Mapping and forecasting technology convergence

Using patent data, Lee et al. (2020) apply association rule mining (ARM) to construct technology ecology networks describing the significant structural patterns of multi-technology convergence. They demonstrate how to link prediction using supervised ML techniques enable the forecasting of future technology convergence.

Kong et al. (2020) provide a framework to jointly exploit relational as well as textual information in scientometric data to map technological trajectories and forecast technology convergence. They deploy a Graph Neural Network (GNN) that can create multi-input embeddings based on network topology and textual data. They demonstrate this at the case of patent data, where they create embeddings based on the patent's position in the citation network as well as the text of the patent abstract. They perform a dynamic clustering on the resulting embeddings, which can be used to map and identify converging technological trajectories.

Mapping the development of science and technology

Liu et al. (2020) develop a model for dynamic topic detection, tracking, and visualization based on scientometric data. At the case of journal publication data, they create embedding vectors of the publication's citations using Node2vec graph embeddings (Grover and Leskovec, 2016). They develop a Citation Involved Hierarchical Dirichlet Process (CIHDP), which enables dynamic topic tracking and path identification based on citation data.

In other cases, direct references such as cross-source citation data can be used to connect data and system layers for science and technology mapping. For instance, Lyu et al. (2020) use non-patent-literature (NPL) citations of patents to academic journal publications to identify techno-science linkages and their intensity across scientific fields. Similarly, Klinger et al. (2021) combine arXiv publication preprint data with Crunchbase data on business activity and use NLP techniques to identify documents related to DL technologies.

Technology and strategic roadmapping

Jeong et al. (2021) provide a framework for constructing risk-adaptive technology roadmaps by drawing from a variety of text sources. Applying an LDA topic model on articles in futuristic technology journalism outlets such as the MIT Technology Review, they identify broad future risk areas. A semantic analysis of these documents indicates the potential directional impact associated with the risk. The assessment of these risks is done by tracing their appearance in further documents related to technologies (patents) and products (product manuals).

Summary and conclusion

The papers in this special issue can be grouped into four categories according to the two criteria—timeframe and layer (cf. Fig. 2). First, the timeframe concerns the extent to which a paper focuses on the past or the future. Some papers investigate past trajectories to identify implications for the future, while other papers directly forecast the future from the observed past. Indeed, the use of NLP, ML, and AI methods improves the capabilities for both explaining results and predicting outcomes, driving the evolution of research streams in both directions. Second, the layer focuses on whether the multi-source data was used to support decision makings for a single layer or not. Some papers introduce new data sources (e.g., SNS data, web data, or expert opinion) to be combined with a primary conventional data source (e.g., patent data) to improve the quality of analysis for the layer. Other papers link several data sources to examine the complex and multi-layered characteristics of a system that consists of several elements including business, technology, and science.

Accordingly, four evolutionary directions of scientometric analysis, largely due to the methodological advances (elaborated forecasting and mapping algorithms) along with increased data availability (use of multiple sources) are observed in this special issue. Generally, the development towards multi-source for both single- and multi-layer mapping from single-source for single-layer mapping is noticeable. Among those, the research in the second quadrant (Type 2) aims to advance forecasting algorithms for a single layer, usually, a technology layer, by using a single-source or multi-sources; identifying emerging and promising technologies, or converging technologies is the main theme of analysis. The analysis on the layer can be improved either by adding new data sources or adopting state-of-the-art algorithms for the analysis of conventional data such as patents. On the other hand, the research in the third quadrant (Type 3) emphasizes the analysis of the current status rather than the projected future of a single layer. Its evolutionary direction is similar to that of Type 2, with the theme of science and technology mapping, mostly using patents and/or publications. Finally, the research in the fourth quadrant (Type 4) adopts several data sources for multi-layer mapping; typical examples include strategic roadmapping and

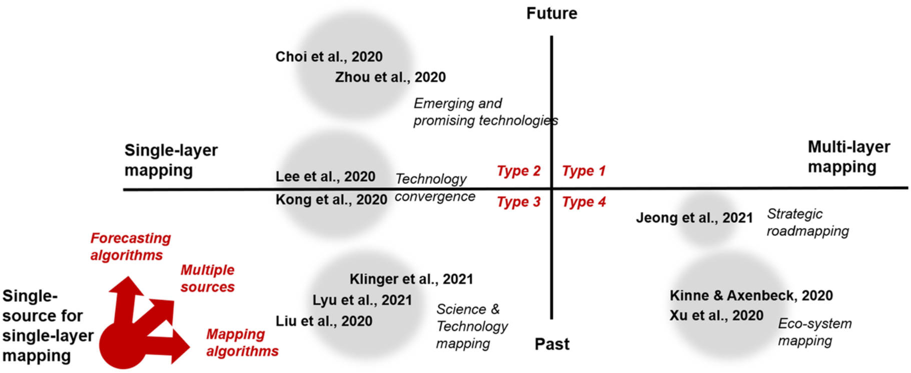

Fig. 2 Evolution of scientometrics analysis due to methodological advances

eco-system mapping, which require several layers developed from several data sources to depict the system. Furthermore, the methodological focus here is on the analysis between the layers in addition to the analysis within the layers. The following table summarizes the new data sources and methods for scientometrics analysis introduced in this special issue (see Table 1).

Finally, it should be noted that no studies are found in the first quadrant (Type 1), needing further attention.

Outlook and research agenda

The papers in this special issue shed light on the impact of ML and AI methods on science, technology, and system mapping and forecasting, which presented how the state-of-the-art methods along with increased data availability have enabled new research opportunities. The followings are some of those opportunities.

First, ML and AL methods are still being developed, which can be used for improving the quality of forecasting and mapping in the scientometrics analysis. Accordingly, the advances in scientometrics analysis will continue, resulting in better performance in forecasting and visualization in mapping. Furthermore, the advances in those methods have enabled the use of other data sources than patents or publications, which have been regarded as traditional data sources for scientometrics analysis. So far, most of the newly introduced data are in the form of texts. However, various other data sources in heterogeneous forms such as audio, images, and videos will be available and continue to be combined with the traditional data sources for science, technology, and system mapping and forecasting. These are the data we are faced with under the circumstances of increased data accessibility and advanced data analytics; sometimes, images or tables have the most important information (e.g., roadmaps) and thus need further investigation on the use of such data.

Second, the key to multi-source and multi-layer mapping lies in how to integrate the heterogeneous data to achieve the purpose of forecasting and mapping in addition to the selection of the most representative and reliable data sources to represent each layer of the ecosystem. By linking the data sources, the relationships between the layers can be established to develop the scenarios of future eco-system, which will be influenced by several factors. The goal and greatest benefit of using ML and AI in scientometrics analysis are to forecast the future; data-driven approaches based on multi-source and multi-layer mapping are expected to uncover the hidden relationships between the layers that constitute the ecosystem and forecast the future. Nevertheless, still few studies have addressed this issue, forecasting the eco-system from multi-source and multi-layer mapping, which requires further research.

List of contributions to the scientometrics special issue dedicated to machine learning and artificial intelligence for ST&I mapping and forecasting

Choi, Y., Park, S., & Lee, S. (2021). Identifying emerging technologies to envision a future innovation ecosystem: A machine learning approach to patent data. Scientometrics, 126(7), 5431-5476. DOI: 10.1007/s11192-021-03999-8

Table 1 Summary of new data sources and methods for scientometrics analysis

	Main purpose of analysis	Traditional data source	New data sources	New methods
Type 2	Emerging and promising technologies (converging technologies)	Patents, publications	Web data, expert opinion	Active learning, ARM, link prediction
Type 3	Science and technology mapping (convergence mapping)	Patents	Arxive, Crunchbase	Node2Vec, graph neural networks
Type 4	Roadmapping and eco-system mapping	Patents, publications, survey	VAT, web data (websites), links, product manuals	Deep learning, LDA, Bayesian network

Jeong, Y., Jang, H., & Yoon, B. (2021). Developing a risk-adaptive technology roadmap using a Bayesian network and topic modeling under deep uncertainty. Scientometrics, 126(5), 3697-3722. DOI: 10.1007/s11192-021-03945-8

Kinne, J., & Axenbeck, J. (2020). Web mining for innovation ecosystem mapping: a framework and a large-scale pilot study. Scientometrics, 125(3), 2011-2041. DOI: 10.1007/s11192-020-03726-9

Klinger, J., Mateos-Garcia, J., & Stathoulopoulos, K. (2021). Deep learning, deep change? Mapping the evolution and geography of a general purpose technology. Scientometrics, 126(7), 5589-5621. DOI: 10.1007/s11192-021-03936-9

Kong, D., Yang, J., & Li, L. (2020). Early identification of technological convergence in numerical control machine tool: a deep learning approach. Scientometrics, 125(3), 1983-2009. DOI: 10.1007/s11192-020-03696-y

Lyu, X., Zhou, P., & Leydesdorff, L. (2020). Eco-system mapping of techno-science linkages at the level of scholarly journals and fields. Scientometrics, 124(3), 2037-2055. DOI: 10.1007/s11192-020-03435-3

Liu, H., Chen, Z., Tang, J., Zhou, Y., & Liu, S. (2020). Mapping the technology evolution path: a novel model for dynamic topic detection and tracking. Scientometrics, 125(3), 2043-2090. DOI: 10.1007/s11192-020-03700-5

Lee, C., Hong, S., & Kim, J. (2021). Anticipating multi-technology convergence: a machine learning approach using patent information. Scientometrics, 126(3), 1867-1896. DOI: 10.1007/s11192-020-03842-6

Xu, G., Hu, W., Qiao, Y., & Zhou, Y. (2020). Mapping an innovation ecosystem using network clustering and community identification: a multi-layered framework. Scientometrics, 124(3), 2057-2081. DOI: 10.1007/s11192-020-03543-0

Zhou, Y., Dong, F., Liu, Y., & Ran, L. (2021). A deep learning framework to early identify emerging technologies in large-scale outlier patents: an empirical study of CNC machine tool. Scientometrics, 126(2), 969-994. DOI: 10.1007/s11192-020-03797-8

References

Grover, A., and Leskovec, J. (2016). node2vec: Scalable feature learning for networks. *In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 855-864).