9^{th} International Conference on Computational Social Science IC^2S^2
July 17-20, 2023, Copenhagen, Denmark

A Fine-Grained Map of All Sciences: Visualizing Nations' Scientific Production

Keywords: science of science, embedding visualization, national advantage, situational awareness

Extended Abstract

Understanding the structure and dynamics of scientific knowledge production is instrumental for the design of policies that can improve the scientific enterprise, for example, better performance evaluations for researchers and organizations, effective funding allocations, and identification of promising research areas. Visualization has been a central tool for gaining situational awareness of current trends and making complex information accessible and understandable [1]. However, the ever-increasing volume of papers and their complex relationships make it challenging to gain a comprehensive view of scientific enterprise.

Using the Web of Science (WoS) dataset, which comprises over 78 million scientific papers published until 2021, we visualize the paper landscape across all scientific disciplines. Our approach involves using SPECTER [2], a language model based on sentence-BERT [3] trained on paper titles. More specifically, SPECTER generates a high-dimensional vector representation, *embedding*, of papers based on their titles and abstract, with fine-tuning to make papers connected by citations to be close to each other. We train SPECTER on the Microsoft Academic Graph dataset [4], which includes citations and papers across all the disciplines in science. Our results show that SPECTER consistently outperforms alternative techniques, including DeepWalk, Laplacian EigenMap, and doc2vec in predicting WoS subject classification at the level of papers. We illustrate the process of obtaining the SPECTER model in Figure a.

To generate the visualization, we embed papers in the WoS dataset using only the paper titles, resulting in a 756-dimensional vector representation for each paper. Next, we reduce the dimensionality of the embedding to 2D (and 3D) by using the UMAP projection [5] trained on a randomly chosen 2.5% sample. We also associate papers to their corresponding metadata based on authors and acknowledgments (Figure b). The resulting map (Figure c) provides a nuanced view of the relationships between scientific disciplines, capturing their structured landscape. For example, it reveals that "Chemistry" lies between "Physics", "Technology Engineering", and "Biology", and "Health" is surrounded by "Medicine" and "Psychology". The map also highlights potential knowledge gaps between and within disciplines. These holes may indicate areas that are under-researched or unlikely combinations of topics deemed unrealistic within the current state of knowledge.

To make the map accessible to a broader audience, including researchers, policymakers, and the general public, we develop an interactive, Web-based visualization that provides an intuitive and holistic understanding of science. This visualization is powered by a GPU-accelerated network rendering framework developed by the authors, and a density estimation algorithm, allowing for real-time visualization of millions of papers on a standard laptop. To further highlight the national advantage in scientific knowledge production, we associate papers with countries based on funding agencies they acknowledge. For example, our visualization (Figures d-e) showcases the production of papers associated with the United States and China, revealing differences in their research interests, with the United States focusing more on health and

medicine. In contrast, China's research production is more pronounced in the tech and engineering fields. Fine-grained patterns can be observed by zooming in on specific regions of the map, revealing subfields and potential invisible colleges. For instance, areas of the embedding dedicated to the study of microRNA in Cancer are more prominent in Chinese research than in the United States. On the other hand, specific regions related to screening and health care in Cancer are more pronounced in the United States research.

Our visualization offers a powerful tool for fostering communication among a wider audience and makes a significant contribution to the fields of Network Science and Science of Science. By providing a comprehensive view of the scientific landscape, we hope to inspire new insights and understanding of the relationships between its various elements, driving future advancements in science.

References

[1] K. Börner, O. Scrivner, L. E. Cross, M. Gallant, S. Ma, A. S. Martin, L. Record, H. Yang, and J. M. Dilger, "Mapping the co-evolution of artificial intelligence, robotics, and the internet of things over 20 years (1998-2017)," *PloS one*, vol. 15, no. 12, p. e0242984, 2020.

[2] A. Cohan, S. Feldman, I. Beltagy, D. Downey, and D. S. Weld, "Specter: Document-level representation learning using citation-informed transformers," *arXiv preprint arXiv:2004.07180*, 2020.

[3] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[4] K. Wang, Z. Shen, C. Huang, C.-H. Wu, Y. Dong, and A. Kanakia, "Microsoft academic graph: When experts are not enough," *Quantitative Science Studies*, vol. 1, no. 1, pp. 396–413, 2020.

[5] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

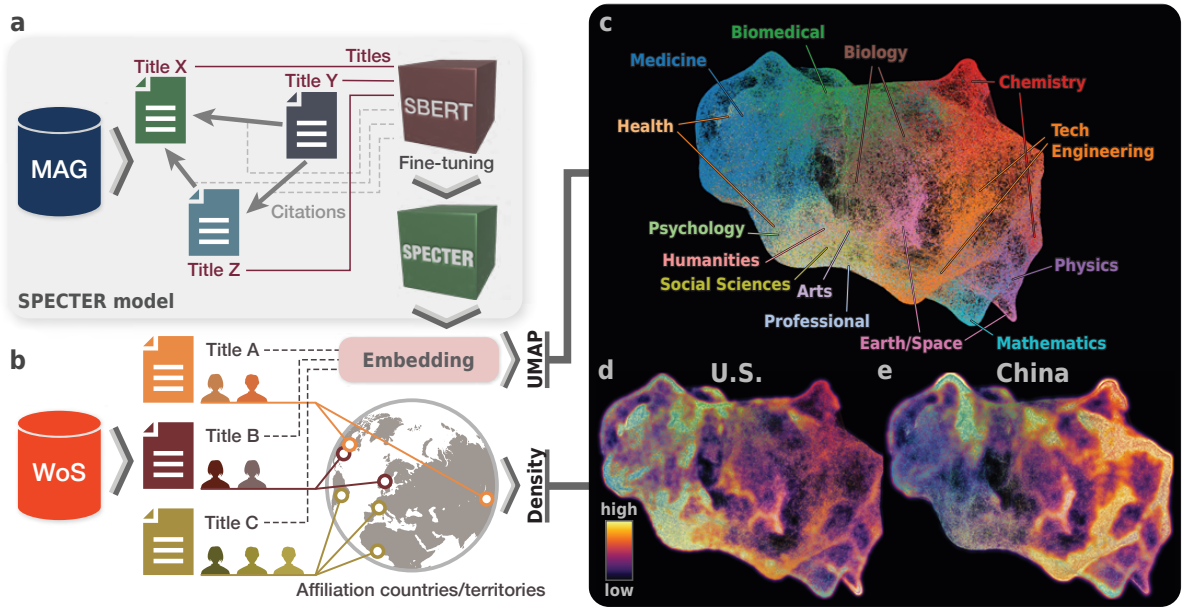

Figure 1: Schematic representation of the adopted methodology (a-b) and preliminary results (c-e). Example of 2D visualization for the paper embedding (c), with colors indicating the NSF categories. National advantage profiles for the U.S. (d) and China (e).