Zuse Institute Berlin

Takustr. 7
14195 Berlin
Germany

TIM KUNT[1] ANNIKA BUCHHOLZ[2] IMENE KHEBOURI[3]
THORSTEN KOCH[4] IDA LITZEL[5] THI HUONG VU[6]

Mapping the Web of Science, a large-scale graph and text-based dataset with LLM embeddings

[1] 0009-0006-5732-3208
[2] 0009-0000-2110-6929
[3] 0000-0002-2389-8795
[4] 0000-0002-1967-0077
[5] 0009-0000-0166-7809
[6] 0009-0007-4869-0505

ZIB Report 25-11 (July 2025)

Zuse Institute Berlin
Takustr. 7
14195 Berlin
Germany

Telephone: +49 30 84185-0
Telefax: +49 30 84185-125

E-mail: bibliothek@zib.de
URL: http://www.zib.de

ZIB-Report (Print) ISSN 1438-0064
ZIB-Report (Internet) ISSN 2192-7782

Mapping the Web of Science, a large-scale graph and text-based dataset with LLM embeddings

Tim Kunt[1] [0009-0006-5732-3208], Annika Buchholz[1] [0009-0000-2110-6929], Imene Khebouri[1] [0000-0002-2389-8795], Thorsten Koch[2,3] [0000-0002-1967-0077], Ida Litzel[1] [0009-0000-0166-7809], and Thi Huong Vu[1,4] [0009-0007-4869-0505]

Abstract Large text data sets, such as publications, websites, and other text-based media, inherit two distinct types of features: (1) the text itself, its information conveyed through semantics, and (2) its relationship to other texts through links, references, or shared attributes. While the latter can be described as a graph structure and can be handled by a range of established algorithms for classification and prediction, the former has recently gained new potential through the use of LLM embedding models. Demonstrating these possibilities and their practicability, we investigate the Web of Science dataset, containing 56 million scientific publications through the lens of our proposed embedding method, revealing a self-structured landscape of texts.

1 Introduction

1.1 Towards a Map of Sciences

A primary question of scientometrics is the categorization of sciences into subjects. This categorization and systems of knowledge organization are often affirmative and never neutral [1]. As part of the ongoing project FAN [14], our approach combines graph-theoretical methods and text processing on large-scale citation graphs, aiming to develop bottom-up and robust methods. From a constructivist perspective, graph theory

[1] Digital Data and Information for Society, Science, and Culture, Zuse Institute Berlin, Takustr. 7, 14195 Berlin, Germany

[2] Software and Algorithms for Discrete Optimization, Technische Universität Berlin, Straße des 17. Juni 135, 10623 Berlin, Germany

[3] Applied Optimization, Zuse Institute Berlin, Takustr. 7, 14195 Berlin, Germany

[4] Institute of Mathematics, Vietnam Academy of Science and Technology, 10072 Hanoi, Vietnam

is a meaningful and elegant abstraction of the relations formed between (text-based) media through references, citations, and shared features of their metadata. Natural language processing (NLP) methods derive an analysis of individual texts. Combining these two approaches promises to compensate for their respective shortcomings, as they are independent of each other in large parts. However, it also comes with a series of difficulties, as mathematically the two employ methods that are not immediately compatible, and an attempt at translation may result in a substantial loss of information. Existing algorithms that map such databases based on the citation graph are the current standard and are suitable for clustering and classifying the dataset to a reasonably fine degree. They leverage the relatedness of individual records based on citations and co-citations [4]. A comprehensive study and comparison are given by [2]. Web of Science [9] itself introduced a classification schema [13] based on the Leiden algorithm [5], although the details of its implementation are not publicly documented.
On the other hand, text representation methods rapidly evolved over the last years [7], which motivates reexamining their suitability for clustering and classification [6][8]. This study primarily demonstrates a text–embedding–based approach. The developed methods lay out and examine a greater map of sciences on the basis of Web of Science. Further, we briefly examine how this approach compares with graph-based methods, motivating the development of a robust hybrid method.

1.2 Text Embeddings via LLMs

LLM-based embeddings, mapping any text onto an element of a high-dimensional vector space, are the latest addition in the toolbox of text representation techniques [8]. Trained on huge datasets covering a large variety of language, they appear to be a powerful one-fits-all solution to the search for a mapping that is not overly complex in its output, but still retains the distinction between even finer semantic differences. We may, however, find that the performance of generic embeddings is limited and the aforementioned generality comes at the cost of precision when applied to domain-specific datasets, [12]. A limitation, tailored or fine-tuned NLP methods may be better equipped to overcome.
For the purpose of representing the abstracts of scientific publications, we argue that LLM embeddings are potentially a suitable choice, and that many of the associated drawbacks do not necessarily apply to the text form of scientific abstracts, as they:

- are **descriptive**, summarizing purpose, scope, and conclusion of a paper, and contain the relevant **keywords** on their topic, guiding embedding (and other NLP) methods
- usually don't employ linguistic devices or other **semantic difficulties** for LLMs
- are overall **homogeneous**, almost formulaic text-form, advantageous for large-scale text processing and representation

In conclusion, an LLM embedding of abstracts or full-texts of scientific articles should capture the broader semantic differences of its texts and thereby create a meaningful interpretable map of the entire dataset in the embedding space - that is the premise of this work.

Going from there, we investigate whether this approach is suitable as a foundation for the prediction and classification of a text's location in said landscape, which in the context of scientific publications could be interpreted as subject or topic. The resulting map was then compared to the existing taxonomy of subjects and sciences to verify and support the findings, as well as to investigate proximity and cluster formations within.

2 Method

2.1 Embedding

For LLM embedding, we use the *ollama* framework and run all computations with both the embedding models *mxbai-embed-large:335m* [11] and *nomic-embed-text:v1.5* [10]. Different embedding models produce different representations of a dataset in their embedding space, and consequently, differing categorizations or clusterings computed within those spaces, as discussed by [8]. Given our focus is on the method itself, we limited ourselves to two LLM embedding models: *mxbai-embed-large:335m* with 1024 dimensions and *nomic-embed-text:v1.5* with 768 dimensions as a reproducible open source and open data model with a comparable competitive performance on our dataset.

We return the normalized vectors and consider cosine similarity from here on as a distance metric. Beautifully, the resulting point cloud can be thought of as forming continents on a globe, as they all lie on an n-sphere (where n is the dimension of the embedding model). A clustering or categorization of those points can be thought of as drawing a map on this globe (Fig. 3).

2.2 Centers, Clusters and Distances

Building on the embedding vectors and the citation graph, a wider range of analytical methods can be applied, a selection of which is outlined in the following.
Firstly, we compute the center-points of each of the 255 unique subjects in Web of Science as the weighted average of the embedding vectors for each record of the respective subject by considering the weight for a given record as the inverse of its total number of subjects. For a thorough evaluation, we must also consider that many of the subjects from Web of Sciences are effectively supersets and/or intersections of other subjects, for example, those including the word *multidisciplinary*. Observed are properties of the subjects, such as their spread among the embedding space, outliers, i.e., points of a certain subject that lie far away from its center, or pairwise distances between subject centers.
Secondly, we compare the pairwise distances in the embedding space with graph-based distances; Here, we chose shortest path. The assumption is that the two metrics are positively correlated, stronger for close pairs and diminishing for pairs with larger distances in either metric.

Further, we considered the option of a dimensionality reduction of the embedding vectors - note that this is not the same as simply choosing an embedding model with fewer dimensions. We compute the explained variance of the principal component analysis (PCA) with varying dimensions.
The above methods are concerned with citation networks. Nevertheless, we emphasize that the approach is generalizable and can be applied to any given dataset that is comprised of textual data and also meaningfully induces a graph structure through its additional features.

3 Computational Experiments / Study

3.1 Data and Setup

The dataset comprises the Web of Science collection [9] up to 2024, containing 56,143,951 records linked by 988,359,820 citations. For this study, we construct a random subsample of 48076 records by selecting each record with a probability of 10^{-3}. 6175 (12.8%) records are removed in preprocessing, leaving 41901. Random sampling combines the advantages of a representation of the entire dataset and the feasibility of computations regarding their space and runtime requirements.
During pre-processing, we remove all records that have no publication year, no authors, no references, no journal, title, or unique identifier. In many cases, these fields are empty due to missing data, and some of the above fields are needed in computations of subsequent steps. Finally, we exclude all records for which the abstract does not exceed 100 characters. This consequently leaves out those that contain invalid or empty strings or single annotations such as *"A correction to this paper has been published"*.
We conduct all experiments on a NVIDIA Grace Hopper. Although it is entirely feasible for an architecture such as the one laid out in this paper to be implemented on commodity hardware, it does not necessarily require high-performance computing. After all, these methods are meant for data analysis and research, and as such, they are supposed to be repeatable and run with low to moderate requirements and time.

3.2 Discussion

A PCA analysis shows that regardless of the homogeneous text data, it seems that the embedding space of both models is populated in all dimensions (Fig. 1). We conclude that dimensionality reduction as a mean to increase the performance of subsequent computations, such as clustering algorithms, may come at the cost of granularity.
We randomly sample 100,000 pairs from the 48,076 records in our dataset to compute graph-based distances. We observe a positive correlation between the distances measured in the embedding space and those on the citation graph, with a PCC of 0.455 and 0.337 for *mxbai-embed-large:335m* and *nomic-embed-text:v1.5* respectively (Fig. 2).
The results suggest that we can construct a combined measure of embedding distance

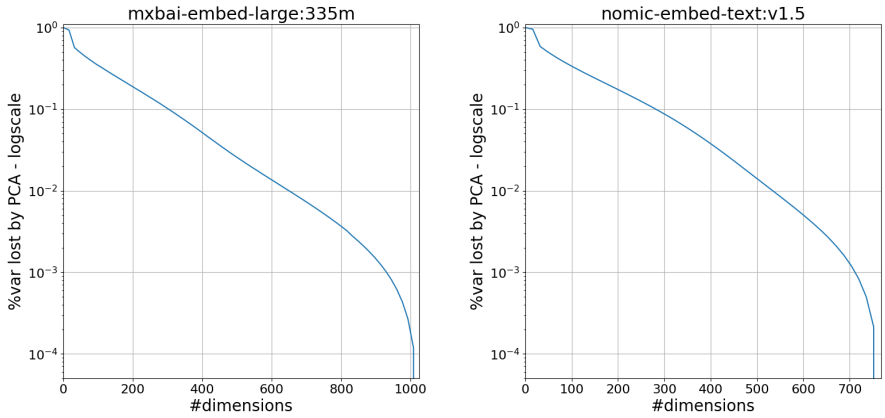

Fig. 1 PCA variance explained as a function of number of dimensions

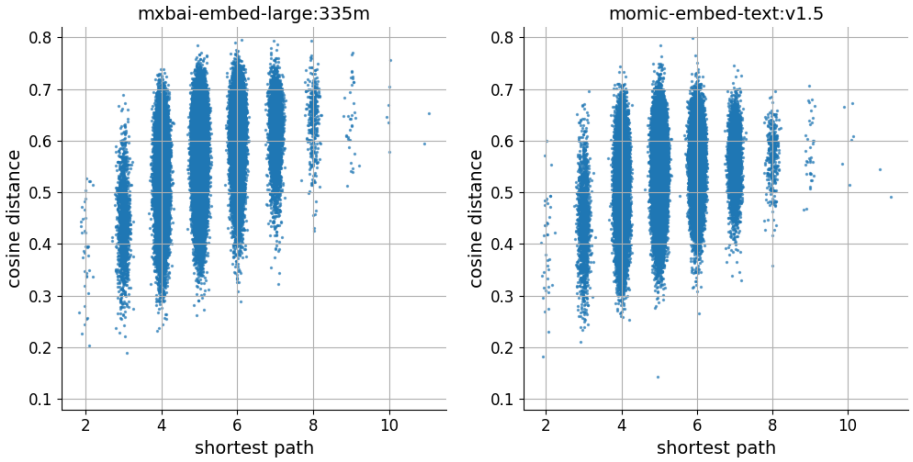

Fig. 2 Comparing pairwise distances of records in the embedding space and on the citation graph

and graph-based distance to capture both interdisciplinary and similarity. We identify an interdisciplinary set of records by a large variance in the embedding space, while the graph-based distances are small. Similarity of distant topics is the inverse; it is expressed by a proximity in the embedding space, while the graph-based distances are large. A combined measure also promises improved results on topic classification of individual papers and clustering of the entire dataset. Individually, the two approaches are already proven to be feasible [2] [3]. Together, they have the potential to compensate each other's shortcomings in that they are methodologically different, and thereby their outliers are ideally independent. As a result, a combination of the two in the form of feature selection promises higher accuracy, which we plan to investigate in future work. Computing the center-points of all subjects and observing the distribution of the records

of each subject across the embedding space yields a map of sciences (Fig. 3). Each dot represents one subject center. Only selected dots are labeled for increased readability. For a selection of larger subjects, a KDE (kernel density estimate) plot shows their distribution across the embedding space. The three levels correspond to 25%, 50%, and 75% of all records in that subject, respectively. Within the embedding space, natural sciences, social sciences, and humanities all form larger interconnected clouds, and so do their respective branches. We observe varying distributions for different subjects, depending on their interdisciplinary nature. Due to multi-labeling, the human-made distinction and the observed distributions, it is evident that a classification of science can only be soft, i.e., by assigning a probability distribution over multiple subjects.

Acknowledgements This work is co-funded by the European Union (European Regional Development Fund EFRE, Fund No. STIIV-001) and supported by the German Competence Network for Bibliometrics (Grant No. 16WIK2101A).

Competing Interests The authors have no conflicts of interest to declare that are relevant to the content of this chapter.

References

1. Rusch, B., Borndörfer, R.: A map is not the territory. Research Metrics Workshop, Institute of Statistical Science, Academia Sinica, Taipei, Taiwan (2025)
2. Šubelj, L., Van Eck, N. J., Waltman, L.: Clustering scientific publications based on citation relations: A systematic comparison of different methods. PloS one 11.4 (2016)
3. Keraghel I, Morbieu S, Nadif M.: Beyond words: a comparative analysis of LLM embeddings for effective clustering. International Symposium on Intelligent Data Analysis, Cham: Springer Nature Switzerland (2024)
4. Waltman, L., Boyack, K.W., Colavizza, G. and van Eck, N.J.: A principled methodology for comparing relatedness measures for clustering publications. Quantitative Science Studies 1.2 (2020): 691-713.
5. Traag, V.A., Waltman, L. and Van Eck, N.J.: From Louvain to Leiden: guaranteeing well-connected communities. Scientific reports 9.1 (2019): 1-12.
6. Keraghel, I., Morbieu, S. and Nadif, M.: Beyond words: a comparative analysis of LLM embeddings for effective clustering. International Symposium on Intelligent Data Analysis. Cham: Springer Nature Switzerland, 2024.
7. Siebers, P., Janiesch, C. and Zschech, P.: A survey of text representation methods and their genealogy. IEEE Access 10 (2022): 96492-96513.
8. Petukhova, A., Matos-Carvalho, J. P., Fachada, N.: Text clustering with large language model embeddings. International Journal of Cognitive Computing in Engineering, 6, 100-108 (2025)
9. Clarivate Analytic, Web of Science, https://www.webofscience.com. Cited Jan 2025
10. Nomic AI, Nomic Embed, https://www.nomic.ai/blog/posts/nomic-embed-text-v1. Cited March 2025
11. Mixedbread AI, mxbai-embed-large-v1, https://www.mixedbread.com/docs/embeddings/models. Cited March 2025
12. Sarma, P. K., Liang, Y., and Sethares, W. A.: Domain Adapted Word Embeddings for Improved Sentiment Classification, https://arxiv.org/abs/1805.04576. Cited July 2025.
13. Clarivate: Introducing Citation Topics (2020) https://clarivate.com/academia-government/blog/introducing-citation-topics/. Cited 14 Dec 2025
14. Research Project Fully Algorithmic Librarian, https://fan.zib.de/. Cited 14 Dec 2025

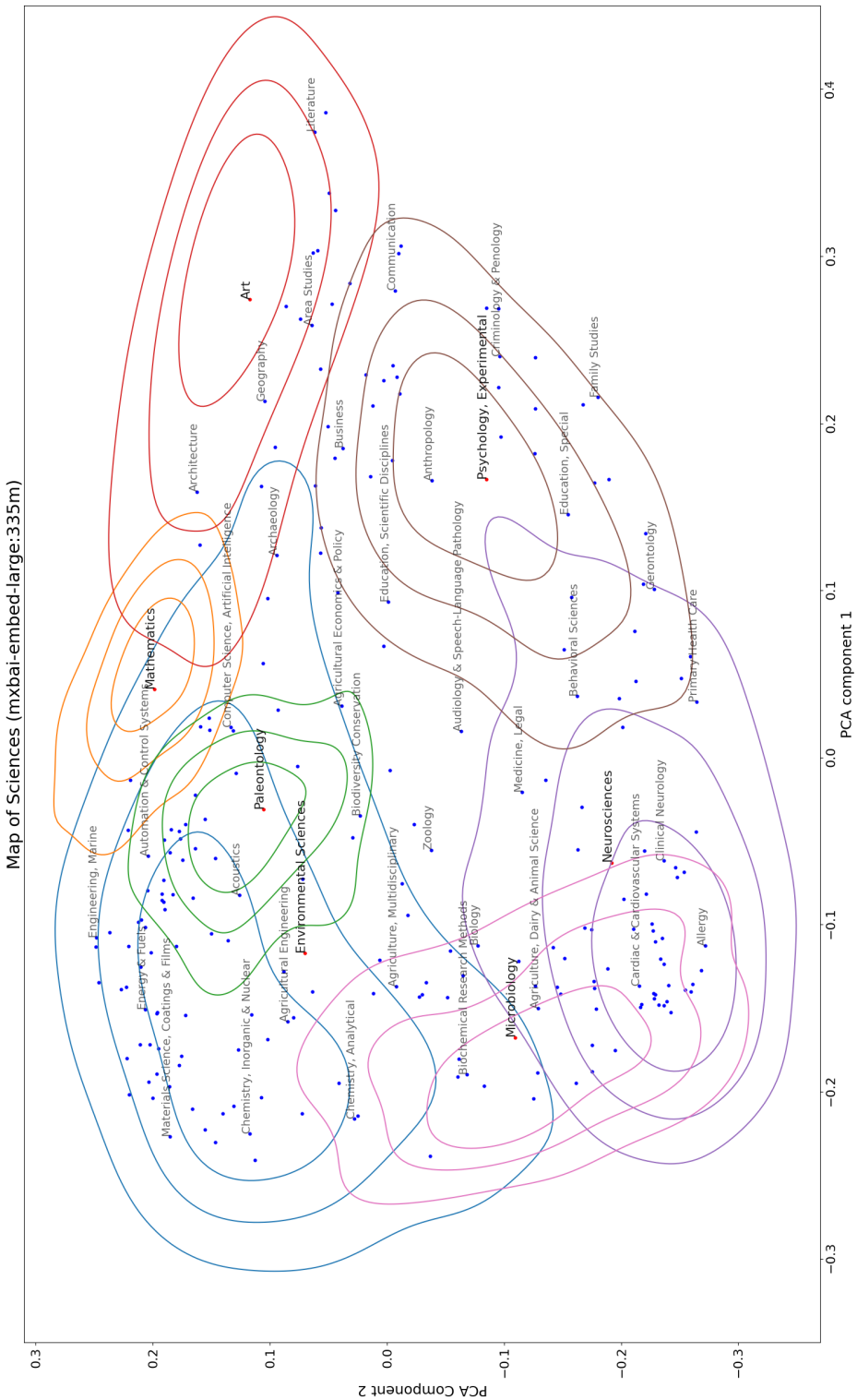

Fig. 3 Visualization of the embedded abstracts using PCA - A Map of Sciences