https://doi.org/10.51408/issi2025_204

Fully Algorithmic Librarian: Large-Scale Citation Experiments

Tomasz Stompor[1], Janina Zittel[2], Thorsten Koch[3], Beate Rusch[4]

[1] *stompor@zib.de*
Zuse Institute Berlin, Kooperativer Bibliotheksverbund Berlin-Brandenburg (KOBV), Takustraße 7, 14195 Berlin (Germany)

[2] *zittel@zib.de*
Zuse Institute Berlin, Applied Algorithmic Intelligence Department, Takustraße 7, 14195 Berlin (Germany)

[3] *koch@zib.de*
Zuse Institute Berlin, Kooperativer Bibliotheksverbund Berlin-Brandenburg (KOBV), Takustraße 7, 14195 Berlin (Germany)
Zuse Institute Berlin, Applied Algorithmic Intelligence Department, Takustraße 7, 14195 Berlin (Germany)

[4] *rusch@zib.de*
Zuse Institute Berlin, Kooperativer Bibliotheksverbund Berlin-Brandenburg (KOBV), Takustraße 7, 14195 Berlin (Germany)

Introduction

Libraries play a crucial role in supporting academic publishing by providing access to bibliometric tools that help researchers navigate vast citation networks (Web of Science, Scopus, OpenAlex). As scientific output grows exponentially (de Solla Price, 1963), algorithmic approaches to citation network analysis are becoming increasingly important. The Fully Algorithmic Librarian (FAN) is an interdisciplinary research project in the fields of mathematics and library and information science, carried out by two departments of the Zuse Institute Berlin. The project's goal is to analyze large-scale citation networks on a knowledge graph sourced from Web of Science and Open Alex citation data that is currently under development. It will serve as a basis for the design of application scenarios. for algorithmic-intelligence-(AI)-supported methods in academic libraries as central research support institutions.

This poster paper aims to present two algorithmic approaches for analyzing large-scale citation networks, which serve as two preliminary steps of the project. Firstly, the results of a comparison of Web of Science (WoS) and OpenAlex databases using the PageRank algorithm reveals key differences. Secondly, a multi-label clustering technique designed for large-scale citation networks accounts for disciplinary variations in publication practices.

A key challenge in bibliometric analysis is the structural and disciplinary diversity of citation networks. Commercial databases like Web of Science (WoS) and open alternatives like OpenAlex (Priem et al., 2022) offer rich but distinct representations of academic publishing. Understanding their differences is essential for developing reliable bibliometric methods. This study demonstrates the effectiveness of algorithmic approaches in analyzing the structural properties of publications in both databases and presents a clustering technique tailored to the varying publication practices across academic fields.

By integrating these insights, this work contributes to the development of automated bibliometric tools that enhance library services, assist researchers in navigating citation landscapes, and support institutions in evaluating academic impact. The findings highlight the potential of algorithmic bibliometry in library and information science and underscore the importance of open, scalable solutions for analyzing scholarly communication.

Comparison of the citation graphs based on WoS and OpenAlex

Evaluating scientific impact requires precise measurement of individual article influence, traditionally assessed through citation metrics. Recently, approaches have shifted toward leveraging citation graph structures rather than relying solely on raw citation counts, for example with employing the PageRank method for measuring scientific prestige (Chen et al., 2023). Beyond ranking influence, PageRank also serves as a valuable tool for comparing bibliometric databases, revealing that citation-based prestige inherently depends on the completeness and accuracy of the chosen dataset. The PageRank computation for WoS (2000–2021) and OpenAlex (1950–2020) (Figure 1) highlights differences in both temporal coverage and citation network structure. Notably, no PageRank is calculated for the most recent 10 years in either dataset, as the metric requires a 10-year citation window. Beyond this temporal aspect, the results also reveal structural variations between the two citation networks.

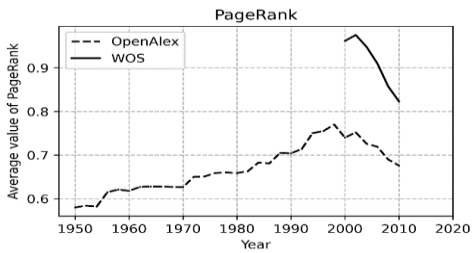

Figure 1. The structural differences of bibliometric datasets illustrated by a PageRank metric following Chen et al. (2023) with 10 years citation span and a damping factor of 0.5 on WoS and OpenAlex.

Useful Clustering Techniques for multi-disciplinary publication data

Bibliometric analysis of large datasets like WoS or OpenAlex requires automated classification of articles by topic. Determining the appropriate number of clusters is challenging, as disciplines do not always have clear boundaries, and articles often span multiple subjects, which necessitates a multi-label classification approach.

Our multi-labeling approach leverages the graph structure of publication data, where references link articles, and similarity is defined through a distance function computed from this network (cf. Nepusz et al., 2008).

Given $S = (s_{ki}) \in \mathbb{R}^{N \times N}$ with $s \in [0,1]$ describing the similarity between articles i and j and

$X = (x_{ki}) \in \mathbb{R}^{C \times N}$ with $x \in [0,1]$,

$\sum_{k=1}^{C} x_{ki} = 1 \; \forall \, i \in \{1, \ldots, N\}$

(see Table 1 for illustration)

assigning articles to clusters, we define

$$f(X) = \sum_{i=1}^{N} \sum_{j=1}^{N} (s_{ij} - \sum_{k=1}^{C} x_{ki} x_{kj})^2$$

to measure how well the similarity S is represented by the clustering X. From this we can compute a clustering by computing $argmin_X f(X)$, which is a continuous, non-convex optimization problem of very large size, as N > 107 and C depending on the number of clusters representing a meaningful number of (sub-)fields or topics, typically between 100 and 500, such as the 252 subfields defined in the OpenAlex database.

Table 1. Structure of the multi-label cluster matrix X, where x_{ki} indicates the assignment of Article i to Cluster k.

X	Article 1	Article 2	…	Article N
Cluster 1	x_{11}	x_{12}	...	x_{1N}
Cluster 2	x_{21}	x_{22}	...	x_{2N}
…	. . .	. . .	. . .	. . .
Cluster C	x_{C1}	x_{C2}	...	x_{CN}

A GPU-based gradient descent method is employed to efficiently handle large citation graphs. A subgraph with 700,000 nodes and the full OpenAlex graph, consisting of 60 million nodes and over a billion edges, were prepared for analysis. To enhance efficiency, the method leverages the sparse structure of the connection matrix and parallelizes gradient descent using CUDA. This parallelization allows for simultaneous processing of graph segments, significantly reducing computation time. Initial tests show that the CUDA implementation clusters the 700k subgraph in just 30 seconds, demonstrating highly promising performance.

Conclusions and Outlook

This study analyzed the development of academic publishing in WoS and OpenAlex using PageRank and introduced an efficient

multi-label clustering method to assess the similarity of academic publications. The comparison of PageRank-based rankings in both databases highlighted structural differences in citation networks, emphasizing the impact of data coverage and indexing practices. To address the challenge of disciplinary overlap in publication classification, a GPU-accelerated multi-label clustering approach was developed, leveraging the graph structure of citation networks.

While academic publication databases like WoS and OpenAlex provide the best available models of scholarly communication, they do not fully capture the broader landscape of academic publishing, often exhibiting biases such as an overrepresentation of English-language research, the exclusion of certain publication formats (monographs, book chapters etc.), and a disciplinary bias tilted towards STEM fields. Recognizing these limitations, our analysis is built on these databases, with the understanding that inherent biases must be considered when interpreting the results.

Acknowledgments

This work has been co-funded by the European Union (European Regional Development Fund EFRE, fund number: STIIV-001) Certain data included herein are derived from Clarivate™ (Web of Science™). © Clarivate 2025. All rights reserved. We acknowledge the use of WoS through the Kompetenznetzwerk Bibliometrie. Supported via the German Competence Network for Bibliometrics funded by the Federal Ministry of Education and Research (Grant: 16WIK2101A).

References

Chen, Y., Koch, T., Zakiyeva, N., Liu, K., Xu, Z., Chen, C-h., Nakano, J. & Honda, K. (2023). Article's scientific prestige: Measuring the impact of individual articles in the web of science. *Journal of Informetrics*, 17, 101379.

Nepusz, T., Petróczi, A., Négyessy, L. & Bazsó, F. Fuzzy communities and the concept of bridgeness in complex networks, Phys. Rev. E 77, 016107, 2008.

De Solla Price, D. J. (1963). *Little Science, Big Science*. New York Chichester, West Sussex: Columbia University Press.

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *ArXiv*. https://arxiv.org/abs/2205.01833