Disentangling scientific fields using temporal clustering

Vogl, Malte
mvogl@mpiwg-berlin.mpg.de
Max Planck Institute for the History of Science, Germany

Lalli, Roberto
rlalli@mpiwg-berlin.mpg.de
Max Planck Institute for the History of Science, Germany

Abstract

The study of the evolution of scientific fields often relies on metadata information which are attributed by third-party institutions and mostly ahistorical (Vogl et.al. 2022, Bornmann 2018) or requires a preselection of corpora based on e.g. specific journals or scientific organizations (Tsay 2008). Scientific fields are used as input in classifications (Perianes-Rodriguez/Ruiz-Castillo 2017) and defining e.g. interdisciplinarity (Gates et.al. 2019), without clarifying what constitutes a scientific field in the first place.

This short presentation introduces an alternative workflow using the semanticlayertools software package (Vogl 2023) developed in the framework of the ModelSEN research project (ModelSEN Project 2023), which enables historians of science to run an iterative process of corpus creation based on temporal co-citation clustering, a reporting framework, and two main visualization strategies.

Iterative temporal clustering

For an unknown large corpus of bibliometric data, the software creates for each year in a given time frame weighted edges between co-cited publications (A B w), where the weight w is the number of found co-citation between publications A and B in the given year. In each iteration step the software then creates clusters of co-cited publications across several years based on the Leiden clustering algorithm. Since this algorithm relies on the notion of a density to decide which publication belong to a cluster, the iterative process is two fold.

First, the basic clustering parameter has to be chosen to select for clusters a) on the level of inter-institutional collaborations (resulting in cluster sizes of several thousand publications), b) grouped around specific innovations, e.g. technological advances in detectors (resulting in cluster sizes of several ten-thousand publications) or c) on the level of (sub-) fields of science, e.g. earth-system sciences (resulting in cluster sizes of several hundred thousand publications).

In the second iterative step, the historian of science can select clusters of publications which should not be considered as part of the analysis, and exclude the corresponding papers from the initial co-citation clustering. This step is based on a reporting subroutine of the semanticlayertools software, which makes use of available textual information (i.e. titles or abstracts) and metadata of publications. The reports for each found cluster contain basic information like cluster size and time frame, statistics of found metadata, main authors and affiliations and topic models based on non-negative matrix factorization for 15 and 50 topics. Based on this information, the researcher can then select clusters which should not be considered in the next iteration and thus refine the corpus of interest.

Network analysis and visualization

After having reached a meaningful subset of the corpus and relevant clusters for the research question at hand, the software package offers an additional analytical step to the researchers. To help understand the role of groups of papers (e.g. from specific institutions or journals) on the formation of clusters, network scientific measures of centralities are calculated for found clusters across time. This include authority, betweenness, degree and closeness centrality. To foster comparison across years, the program calculates histograms of normalized centralities with fixed logarithmic binning.

The short presentation will introduce the main parts of the software package and give examples for the different clustering scenarios. An application of the workflow for the extraction of a relevant corpus for the field of astrophysics is presented together with visualization strategies for a) the emergence of related research fields based on streamgraphs and b) the role of specific institutions in thematic clusters using the above mentioned centralities based on 3D plots.

Bibliography

Vogl, M., Buarque, B. S., Lalli, R., Wintergruen, D. and Weiß, L. (2022): Using Dimensions for historical research: Biases, pitfalls and a spark of hope. figshare. Presentation. https://doi.org/10.6084/m9.figshare.21345279.v3

Bornmann, L. (2018). Field classification of publications in Dimensions: a first case study testing its reliability and validity. Scientometrics 117 (1): 637–640. https://doi.org/10.1007/s11192-018-2855-y

Tsay, Y. (2008) A bibliometric analysis of hydrogen energy literature, 1965–2005. Scientometrics 75, 421–438 (2008). https://doi.org/10.1007/s11192-007-1785-x

Perianes-Rodriguez, A. and Ruiz-Castillo, J. (2017) A comparison of the Web of Science and publication-level classification systems of science. Journal of Informetrics 11 (1): 32-45. https://doi.org/10.1016/j.joi.2016.10.007

Gates, A.J., Ke, Q. , Varol, O. and Barabási, A.-L. (2019) Nature's reach: narrow work has broad impact, Nature 575. https://www.nature.com/articles/d41586-019-03308-7

Vogl, M. (2023) semanticlayertools, https://semanticlayertools.readthedocs.io/en/latest/

ModelSEN Project (2023) Website: https://modelsen.mpiwg-berlin.mpg.de/de/ [28.04.2023]