Estimation of Research Communities in the Multilingual Academic Data

Ilya Rahkovsky
ir177@georgetown.edu
Center for Security and Emerging Technology
Georgetown University

Jennifer Melot
jm3312@georgetown.edu
Center for Security and Emerging Technology
Georgetown University

Introduction

Research Communities (RC) are the small groups of researchers working on a same topic. The estimation of RCs on single-data source (SCOPUS data) was pioneered by Klavans and Boyack [2017]. We update their work using more efficient Leiden clustering algorithm and a much large database of multilingual articles.

Data and Methods

We use research article and conference proceedings from Web of Science (WOS), Digital Science (DS), Microsoft Academic Graph (MAG), and Chinese National Knowledge Infrastructure (CNKI) totalling 107 million documents.[1] Our data have complete coverage of Chinese and English articles, while the coverage of other languages are far from complete. The articles matched and de-duplicated using text-similarity of meta-information.

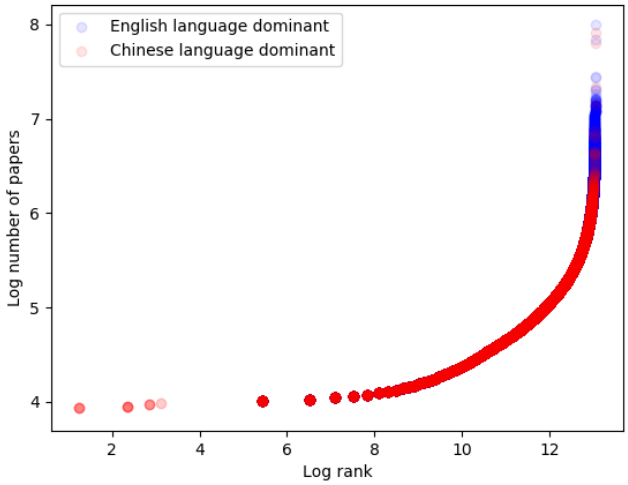
The articles are linked using direct citations in a single graph. To estimate clusters of articles, we apply Leiden clustering algorithm maximizing modularity of network partitions [Traag et al., 2019]. Clusters smaller than 50 papers are considered "immature" and they are merged with larger clusters based on maximum bibliometric similarity (if possible).

Results

Leiden clustering model clustered 107M articles into 2M clusters. We drop clusters with less than 50 papers that we are unable to match with larger clusters from the analysis, resulting in 104M articles in "mature" clusters.

	N	N in clust	N in dom	clust size
English	81.2M	80.2M	30.9M	14,036
Chinese	14.4M	13.0M	2.4M	1,288
Japanese	506K	487K	713	118
German	1.9M	1.8M	137	137
French	1.2M	1.1M	0	N/A
Other	8.2M	7.9M	7,867	245
Mixed	N/A	N/A	71M	238
All	107.4M	104M	104M	226

[1] MAG and DS include all articles in these databases, sometimes going back go the 17th century. In WOS we have all articles published after 1999, in CNKI we have articles published after 2005.



We consider the language to dominate the RC if it accounts for more than 90% of the articles. For example, out of 13 million Chinese articles in mature RCs, there are 2.4 million located in the RCs dominated by Chinese language. The average Chinese-language-dominant RC has 1,288 papers.

English-language-dominant RCs are bigger as they can draw from a large number of English articles. Also, these articles have more references resulting in a much denser network. Languages other than English and Chinese hardly have any dominant RCs. This is probably the result of our incomplete coverage of these languages.

References

Klavans, R. and Boyack, K. W. (2017). Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? *Journal of the Association for Information Science and Technology*, 68(4):984–998.

Traag, V. A., Waltman, L., and van Eck, N. J. (2019). From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9.