SM1: Distribution of CWTS citation cluster keywords

Peder M. Isager

9/17/2021

We acquired additional bibliometric information from the Centre for Science and Technology Studies (CWTS, https://www.cwts.nl/) about citation clusters in our data [a proxy for scientific subfields based upon clustering of publications based on citation relationships; Waltman and van Eck (2019)]. Each citation cluster can be thought of as a data-defined research subfield, where articles that cite similar articles tend to end up in the same cluster/subfield. The clusters were generated independently of our study, based on all bibliometric information in the CWTS database. For each article in our dataset, we retrieved information about its corresponding CWTS cluster, including how many articles in the entire CWTS database were included in this cluster. The CWTS cluster algorithm also generates key-terms that describe the most prevalent research topics dealt with in each cluster. This allows for a higher-resolution analysis of the research topics covered in our dataset. We expected to see a wide range of CWTS clusters included in our data, and we expected the cluster labels would primarily denote terms related to social phenomena and their influence on cognition and neural activity.

The records in our dataset were sampled from 162 unique CWTS citation clusters. The size of each cluster varied substantially (min = 829 records, median = 1.235×10^4 records, max = 3.977×10^4 records). The full distribution of cluster size visualized in figure SM1.

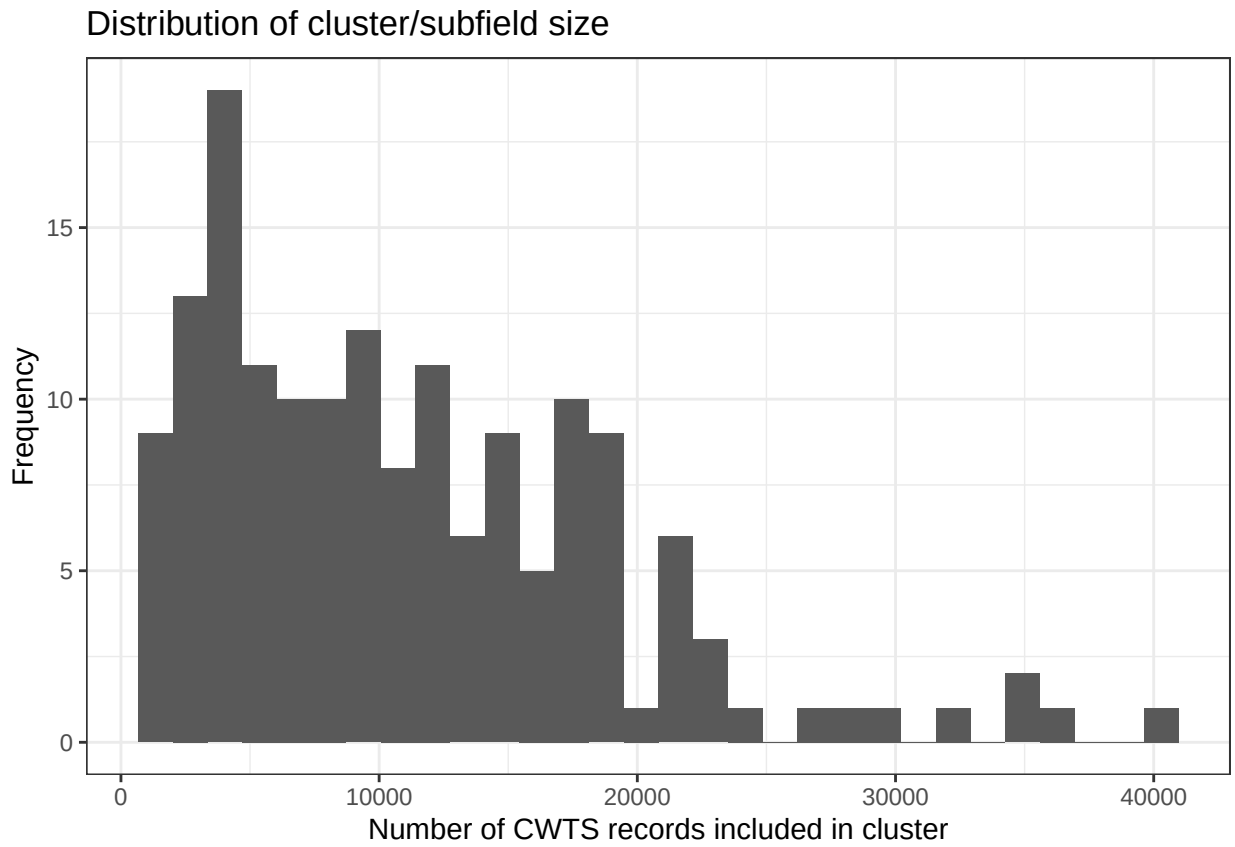

Figure SM1: Distribution of the size of CWTS clusters to which the articles in our dataset belong. The size of each cluster refers to the total number of records associated with that cluster over the entire CWTS database, which includes millions of articles from all over science. Thus, the size of the cluster refers to the population size of the cluster, not how many articles are included in the cluster in our data.

To better understand the scientific topics covered in these citation clusters, we inspected the category labels assigned to each cluster by CWTS. In total, the citation clusters were associated with 774 unique labels. Table 2 displays the frequency of the 50 most frequently mentioned category labels in our data.

Table SM1: Frequency table of the 50 most prevalent cluster labels in our dataset. Label frequencies are identical for many labels because multiple labels are used to describe a single cluster. For example, the terms "face recognition," "face processing," "prosopagnosia," "unfamiliar face," and "facial identity" all appear 118 times because they are all used to describe one, and only one, cluster to which 118 articles in our data belong.

	Label	Frequency
50	intertemporal choice	364
46	decision making	358
47	delay discounting	358
48	impulsivity	358
49	iowa gambling task	358
45	imitation	318
41	action observation	258
42	empathy	258
43	mirror neuron	258
44	motor imagery	258

	Label	Frequency
40	attentional bias	210
39	fear	207
36	emotional face	203
37	facial expression	203
38	social anxiety	203
31	default mode network	180
32	fmri	180
33	fmri data	180
34	functional connectivity	180
35	resting state	180
30	alzheimer	125
25	face processing	118
26	face recognition	118
27	facial identity	118
28	prosopagnosia	118
29	unfamiliar face	118
21	n400	109
22	primary progressive aphasia	109
23	semantic dementia	109
24	visual word recognition	109
16	contamination	67
17	disgust	67
18	disgust sensitivity	67
19	moral dilemma	67
20	moral judgment	67
10	death anxiety	65
11	mind	65
12	mortality salience	65
13	ostracism	65
14	social exclusion	65
15	terror management	65
5	autobiographical memory	63
6	expressive writing	63
7	generativity	63
8	mental time travel	63
9	rumination	63
2	false belief	60
3	infant	60
4	month old infant	60
1	effect	52

While most cluster labels describe substantive topics, the inclusion of "effect" in the top 50 most prevalent labels should highlight that not all labels assigned by the cluster algorithm are equally informative for understanding cluster content. The distribution of substantive topic labels is largely consistent with our expectations. Terms like "fmri," "fmri data," "imitation," "empathy," "facial expression," "social anxiety," and "social exclusion" are frequently used to describe subfields to which the articles in our dataset belong. Conversely, we did not observe any labels that were obviously unrelated to our a-priori expectations about which topics should be covered in a sample of social fMRI articles.

Close inspection of the cluster label data revealed an important confuscating factor, however. While cluster labels are descriptive of the whole cluster as it appears in the CWTS database, they are not necessarily descriptive of the subset of that cluster included in our candidate set. As a prominent example, the citation

cluster described by the labels "intertemporal choice," "decision making," "delay discounting," "impulsivity" and "iowa gambling task" is the most prevalent cluster in our data. However, the labels used to summarize the 13168 CWTS records in this cluster are not necessarily representative of the minority of articles from the cluster that are included in our dataset. Although the term "iowa gambling task" is descriptive of the cluster as a whole, only one out of the 358 articles from the cluster in our dataset mentions the Iowa gambling task. Inferring about the contents of our dataset based on these labels alone could therefore be misleading. However, correspondence in topics between frequently occurring CWTS cluster labels and frequently co-occurring keywords in VOSviewer (see figure 3 in main manuscript) is encouraging.

References

Waltman, Ludo, and Nees Jan van Eck. 2019. "Field Normalization of Scientometric Indicators." In *Springer Handbook of Science and Technology Indicators*, edited by Wolfgang Glänzel, Henk F. Moed, Ulrich Schmoch, and Mike Thelwall, 281–300. Springer Handbooks. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-02511-3_11.