Which topics are best represented by science maps? An analysis of clustering effectiveness for citation and text similarity networks

Bascur Cifuentes, J.P.; Verbernem S.; Eck, N.J.P. van; Waltman, L.

Scientometrics (2025) 130:1181–1199
https://doi.org/10.1007/s11192-024-05218-6

Which topics are best represented by science maps? An analysis of clustering effectiveness for citation and text similarity networks

Juan Pablo Bascur[1,2] · Suzan Verberne[2] · Nees Jan van Eck[1] · Ludo Waltman[1]

Received: 29 July 2024 / Accepted: 28 November 2024 / Published online: 23 January 2025

Abstract
A science map of topics is a visualization that shows topics identified algorithmically based on the bibliographic metadata of scientific publications. In practice not all topics are well represented in a science map. We analyzed how effectively different topics are represented in science maps created by clustering biomedical publications. To achieve this, we investigated which topic categories, obtained from MeSH terms, are better represented in science maps based on citation or text similarity networks. To evaluate the clustering effectiveness of topics, we determined the extent to which documents belonging to the same topic are grouped together in the same cluster. We found that the best and worst represented topic categories are the same for citation and text similarity networks. The best represented topic categories are diseases, psychology, anatomy, organisms and the techniques and equipment used for diagnostics and therapy, while the worst represented topic categories are natural science fields, geographical entities, information sciences and health care and occupations. Furthermore, for the diseases and organisms topic categories and for science maps with smaller clusters, we found that topics tend to be better represented in citation similarity networks than in text similarity networks.

Keywords Citation-based clustering · Text-based clustering · Evaluation · Topics · MeSH terms

Introduction

Science maps (Chen, 2017) are visualizations that provide an overview of the content of collections of scientific publications. The goal of science mapping is to find meaningful structures in the bibliographic metadata of publications (e.g, in the references, the titles and abstracts, or the authors). These structures can then be used for literature analysis or information retrieval (Cobo et al., 2011; van Eck, 2011). Some of

✉ Juan Pablo Bascur
j.p.bascur.cifuentes@cwts.leidenuniv.nl

[1] Centre for Science and Technology Studies, Leiden University, Leiden, The Netherlands

[2] Leiden Institute for Advanced Computer Science, Leiden University, Leiden, The Netherlands

the uses of science maps are field delimitation (Zitt, 2015), research policy (Sullivan et al., 2007), and enhanced document browsing (Bascur et al., 2023). A well established practice to create science maps is to cluster similar publications, and then to summarize the content of the resulting clusters. Our focus in this paper is on science maps created in this way.

When using science maps, it is important to be aware that scientific publications usually have more than a single topic (e.g., a document about the topic *lung cancer* is, implicitly, also about both *lungs* and *cancer*), but in a science map they typically can be assigned to only one cluster, where the cluster is intended to represent a single cohesive topic. Because in reality, publications can have more than one topic, losing information when creating science maps is unavoidable, but it does raise the question of which of the topics addressed in a collection of publications a clustering will be based on. This is not an idle question, as there can be significant disagreement between expert-identified and cluster-identified topics (Held et al., 2021), indicating that expert-identified topics are poorly represented by the clusters in a science map. More specifically, an expert with an interest in a particular topic may find that publications related to this topic are scattered over many different clusters, with most of the publications in these clusters being unrelated to the expert's topic of interest. By providing a better understanding of the types of topics that are well or less well represented in science maps, we hope our research will contribute to a more effective use of these maps.

In this paper, we use the Medical Subject Headings (MeSH) terms to investigate clustering for biomedical topics. Our focus is on clustering solutions based on either citation or text similarity networks, which are the most common document similarity metrics for creating science maps. We aim to find out which MeSH terms are well represented by the clusters in a science map, a phenomenon that we will refer to as *clustering effectiveness*. Our approach is to group topics, represented by MeSH terms, into topic categories, represented by branches of the MeSH tree, and to then evaluate clustering effectiveness at the level of these topic categories.

Our research questions are as follows:

- Which topic categories have the highest and lowest clustering effectiveness in citation and text similarity networks?
- Which topic categories have higher clustering effectiveness in citation similarity networks than in text similarity networks, and vice versa?

In the remainder of this paper, we will discuss background literature, describe our data, define our metrics, report our analyses and discuss our results.

Background

This section has the following structure: In Subsection 2.1 we explain how science maps are usually evaluated, in Subsection we explore the criticism of science maps that originates from one particular evaluation method, and in Subsection we explain the challenges of understanding the meaning of the clusters in a science map.

Evaluation of science maps

In the current paper we evaluate the quality of science map only from the perspective of its field delimitation function. However, it is important to keep in mind that science maps are richer tools, with various features that can be interpreted beyond the extend to which clusters correspond to topics. For example, it can be evaluated on the extend to which the labels of the clusters and the distance between clusters provide useful visual information, or on how cross-cluster topics inform on the structure of the topics. The most common method to evaluate the quality of the field delimitation function a science map is to ask experts if the science map reflects their knowledge of the field of interest. The utility of this evaluation method has recently been called into question because it usually gives an inconclusive result: The experts tend to agree with most of the science map but identify caveats about certain details (Gläser, 2020). Additionally, there are several issues intrinsic to the expert evaluation method: The evaluation criteria may differ between experts; seeing the map may affect the expert's understanding of a field; the expert may be biased towards the subfields of their interest; and the expert may have limited competence in some subfields (Gläser, 2020).

An alternative method to evaluate the quality of a science map is to consider the intrinsic properties of the clustering process used to create science maps. Commonly used intrinsic properties are desirable characteristics such as homogeneous cluster sizes, few small clusters, stable clustering solutions between different runs of the cluster algorithm, and a short computing time to create the clusters (Šubelj et al., 2016). An intrinsic properties evaluation method was developed by Waltman et al. (2020). Their method assumes that there exists an ideal map and then assesses how closely a clustering solution matches this map. It evaluates the quality of a clustering solution based on one metric using another unrelated baseline metric (e.g., a clustering solution based on citation similarity can be evaluated using text similarity). Ahlgren et al. (2020), who created the clustering solutions that we use in our current work, used this method with MeSH terms similarity as their baseline metric.

A third approach to evaluate the quality of a science map is to define a ground truth made of documents that correspond to a given topic, and evaluate the overlap between the clustering solution and the ground truth: either the extent to which all documents of each field are contained in a single cluster (Held et al., 2021, 2020), or the extent to which each cluster contains only documents of a single field (Rossetti et al., 2016; Held et al., 2021; Held & Velden, 2022; Haunschild et al., 2018). Some studies obtained the ground truth from the references of review articles (Klavans & Boyack, 2017; Sjögårde & Ahlgren, 2018), but most studies obtained the ground truth using expert knowledge. To our knowledge, MeSH terms have not been used as ground truths, although Sjögårde, Sjögårde et al. (2021) used MeSH terms to label clusters in science maps. It is worth mentioning that our work has a different goal than evaluating a science map based on a ground truth. Instead of evaluating the quality of a science map based on a set of topics, we evaluate which topic categories are most accurately represented in a science map.

Criticism of science maps based on ground truth evaluations

Evaluations that use expert knowledge ground truths have recently questioned the quality of science maps by challenging their ability to identify fields of science (Haunschild

et al., 2018; Held et al., 2021, 2020; Held & Velden, 2022). For example, Held and Velden (2022) found that science maps provide clusters about organisms rather than clusters about the field of invasive biology. One explanation for these negative results is that a document can belong to several fields or topics but only to a single cluster (Held et al., 2021; Held & Velden, 2022) (although some maps allow documents to belong to multiple clusters (Xu et al., 2018; Havemann et al., 2017)). Another explanation is that the choice of a clustering algorithm can have a significant influence on the quality of a science map, and it is impossible to know beforehand which clustering algorithm will give the best result for a given map (Held, 2022; Rossetti et al., 2016).

Similar negative findings have also emerged in areas beyond science mapping. For example, the field of complex systems has developed algorithms to clusters the elements that share a given property (i.e., the cluster matches the ground truth), but these algorithms fail in practical applications. On the other hand, this field has succeeded in practical applications of algorithms that infer the properties of an element based on the properties of the other elements in a cluster (e.g., fraud in telecommunications networks, function in biological networks) Fortunato (2010); Hric et al. (2014); Peel et al. (2017).

Meaning of the clusters

The negative findings discussed in the previous subsection suggest that science maps, and clustering in general, offer poor representations of certain ground truths. However, this does not mean that science maps are not useful. As mentioned in Subsection 2.1, experts tend to agree that science maps reflect their knowledge of a field. Also, in the field of complex systems, Newman and Clauset (2016) argued that, even if clusters do not reflect the ground truth, they can still describe meaningful structures in the data. Our work tries to find out what kinds of structures are described by the clusters in a science map.

In this direction, Seitz et al. (2021) found that the epistemic functions of citations (i.e., what kind of knowledge is a citation contributing to in a document) within a cluster are different from the epistemic functions of citations between clusters. This suggests that clusters tend to represent certain epistemic functions more than others. Also, the type of similarity network might have an effect on the meaning of clusters. For example, Ding (2011) found significant differences between clusters emerging from co-authorship networks of documents and clusters emerging from topic modeling of documents. On the other hand, Velden et al. (2017) found that there is a substantial similarity between the topics found in science maps built from citation and text similarity networks, although science maps built from citation networks are better at distinguishing topics when words related to the topics have multiple meanings.

Methods

This section has the following structure: In Subsection 3.1, we define how we selected our data. In Subsection 3.2, we explain how we modified our data so to better fit our experimental design. In Subsection 3.3, we explain how we evaluate the clustering effectiveness of topic categories.

Data selection

Documents

The collection of documents that we use in our work comes from the work by Ahlgren et al. (2020). This is a collection of 2,941,119 PubMed documents published between 2013 and 2017.

Clustering solutions

The clustering solutions that we use are the ones generated by Ahlgren et al. They created several clustering solutions for the above mentioned documents using different similarity metrics and granularities. They used the Leiden algorithm Traag et al. (2019) for clustering, where the parameter Resolution controls the granularity of the clustering solution (a higher Resolution value generates smaller clusters). We select two similarity metrics, one for citation and one for text, based on which pair of metrics produce similar cluster sizes at the same Resolution. The citation metric is *Extended direct citation*, which is calculated using direct citations between documents plus the citations to documents outside the document collection (Waltman et al., 2020). The text metric is *BM25* (Robertson & Zaragoza, 2009), which uses the noun phrases in the titles and abstracts of the documents, and weights them inversely to their frequency in the document collection (Waltman et al., 2020). For each metric we selected the three clustering solutions that use the Resolution values $2 * 10^{-6}$, $2 * 10^{-5}$ or $2 * 10^{-4}$, enabling us to evaluate different cluster sizes. We selected these Resolution values because the first and second value yield cluster sizes similar to those in the algorithmic mapping of science (Waltman & Van Eck, 2012) used in the CWTS Leiden Ranking (CWTS, 2023), while the third value enables us to evaluate clusters of smaller size.

Topics

Our topics are the MeSH terms, a controlled vocabulary thesaurus from the National Library of Medicine (NLM) used for indexing PubMed. MeSH terms are semi-automatically annotated to documents by the NLM National Institutes of Health (2023). We obtained the MeSH terms annotated for each document in our document collection, plus the metadata of the MeSH terms themselves, from the PubMed and MeSH databases (version from 2023) available in the database system of the Centre for Science and Technology Studies (CWTS) at Leiden University.

Topic categories

Our topic categories are the 16 nodes at the first level of the MeSH hierarchical tree of topics (National Institutes of Health, 2023), also known as the branches of the MeSH tree. We use branches because they group the MeSH terms in epistemological categories (e.g., organisms), which are the categories sometimes used for topical analysis of

clusters (Held et al., 2021; Seitz et al., 2021). A single MeSH term can have instances in different branches of the MeSH tree. We will address this in Subsection .

Data prepossessing

Clustering solution cleaning

We cleaned the clustering solutions by removing the clusters with fewer than 10 documents because these clusters usually had documents that were disconnected from the largest connected component of the similarity network. Removing these clusters removed only a minor fraction of the total number of documents. The statistics of each clustering solution after this process can be seen in Table 1. In this table, the variable *S* is the smallest set of clusters that together cover at least half of the documents in the dataset. This means that *S* contains the biggest clusters in the clustering solution. We report statistics for *S* to provide some insight into the distribution of cluster sizes.

MeSH term expansion

We would like a MeSH term to be annotated on all documents related to the topic of the MeSH term, but NLM typically only annotates up to 15 MeSH terms per document, which means that more generic MeSH terms are not annotated. To fix this, we expanded the number of MeSH terms annotated to a document by annotating, for each NLM MeSH term, all MeSH terms that are upstream in the MeSH tree, or in other words, all ancestors of the NLM MeSH term in the MeSH tree.

For example, if a document has the NLM MeSH term *Abdominal Pain*, we also annotated the upstream MeSH term *Pain*. While the former MeSH term belongs to the branch Diseases [C], the latter one belongs not only to the branch Diseases [C], but also to the branches

Table 1 Statistics of the clustering solutions

Metric	Resolution	Citation similarly	Text silmiarity
	$2^{*}10^{-6}$	297	277
Number of clusters	$2^{*}10^{-5}$	2,469	2,475
	$2^{*}10^{-4}$	21,659	20,603
Number of clusters in S	$2^{*}10^{-6}$	59	65
	$2^{*}10^{-5}$	496	514
	$2^{*}10^{-4}$	4,017	3,554
Median size of clusters	$2^{*}10^{-6}$	7,615	9,373
	$2^{*}10^{-5}$	878	891
	$2^{*}10^{-4}$	88	86
Size of the smallest cluster in S	$2^{*}10^{-6}$	16,936	15,358
	$2^{*}10^{-5}$	1,954	1,885
	$2^{*}10^{-4}$	228	252

S is the smallest set of clusters that together cover at least half of the documents in the dataset

The size of the cluster is the number of documents it contains

Psychiatry and Psychology [F] and Phenomena and Processes [G]. We annotated the MeSH term *Pain* paired with the branch Diseases [C], and not with the other two branches. On the other hand, if a document has the NLM MeSH term *Pain*, then we would annotate three versions of it, one for each branch. For simplicity, in the rest of this paper we will refer to MeSH terms paired with a specific branch simply as MeSH terms. Also, we will refer to the documents that have a given MeSH term as the MeSH term documents and to the number of these documents as the MeSH term size.

MeSH term removal

We removed some MeSH terms to improve the quality of our experiments. Our first removal criterion is size. We removed MeSH terms with size greater than 300,000 (i.e., 10% of the document set) because these MeSH term documents can saturate the clusters just by random chance, distorting our analysis. We also removed the MeSH terms with size 500 or less, because we want the smallest MeSH terms to be close but smaller than the median size of the clusters for resolution $2 * 10^{-5}$.

Our second removal criterion is redundancy. Due to the MeSH term expansion process, some MeSH terms had almost the same documents as their ancestor in the MeSH tree, like *Dogs* and its ancestor *Canidae*. This redundancy could distort our results. We therefore decided to remove the redundant MeSH terms by grouping together MeSH terms that share many documents and retaining only the smallest MeSH term from the group, which in our experience tends to be the term that best represents the group. The extent to which MeSH terms share documents was measured using Jaccard similarity, the grouping algorithm was agglomerate hierarchical clustering with the Complete Linkage method (SAS Institute Inc, 2009), and the criterion for forming MeSH term groups was for MeSH terms to have a Jaccard similarity of at least 0.9. In cases where a group had more than one smallest MeSH term, we selected the one at the lowest level in the MeSH tree or the one with the largest number of instances in the MeSH tree.

Branch removal

To make our results more robust, we removed the branches with fewer than 100 MeSH terms. We ended up with the 14 branches shown in Table .

Size bins of MeSH terms

The size of a MeSH term can be expected to have an effect on its clustering effectiveness. We therefore grouped the MeSH terms according to their size. We refer to these groups as Size bins. To ensure the robustness of our results, we only considered Size bins that had at least 10 MeSH terms per branch. This resulted in five Size bins: 501-1,000, 1,001–2,000, 2,001–4,000, 4,001–8,000, and 8,001–16,000. The number of MeSH terms per Size bin can be seen in Table 2.

Table 2 Number of MeSH terms per branch and Size bin

Branch	Size bin					
	501–1,000	1,001–2,000	2,001–4,000	4,001–8,000	8,001–16,000	Total
Anatomy [A]	209	201	161	102	76	749
Organisms [B]	247	168	98	75	44	632
Diseases [C]	472	391	272	194	114	1,443
Chemicals and drugs [D]	1,033	785	568	357	264	3,007
Analytical, diagnostic and therapeutic techniques, and equipment [E]	324	298	253	189	150	1,214
Psychiatry and psychology [F]	109	113	95	65	38	420
Phenomena and processes [G]	264	244	221	179	143	1,051
Disciplines and occupations [H]	50	28	31	23	15	147
Anthropology, education, sociology, and social phenomena [1]	57	56	40	29	24	206
Technology, industry, and agriculture [J]	76	70	68	24	26	264
Information science [L]	31	35	28	20	18	132
Named groups [M]	21	34	20	11	14	100
Health care [N]	182	150	134	110	87	663
Geographicals [Z]	51	39	36	16	21	163
Total	3,126	2,612	2,025	1,394	1,034	10,191

A Size bin is a range of topic sizes

A topic size is the number of documents in the topic

Clustering effectiveness

Selection of clusters

To find out which MeSH terms are well represented by the clusters in a science map, we introduce the notion of clustering effectiveness. Measuring the clustering effectiveness of a MeSH term starts by selecting a subset of clusters. Our cluster selection criterion is to select the clusters with the largest number of MeSH term documents while making sure that the selected clusters cover at least a given share of all MeSH term documents. We call this share Coverage. We consider three Coverage values: 0.25, 0.50 and 0.75. Our cluster selection criterion minimizes the number of selected clusters for a given Coverage value. It is inspired by cluster quality metrics of Yuan et al. (2022). We expect our cluster selection criterion to reflect the clusters a user of a science map is likely to select while exploring the map.

Clustering effectiveness metrics

Once we have the selected clusters for a given MeSH term, we measure clustering effectiveness using two metrics:

- Purity: Purity represents the extent to which the selected clusters are composed of MeSH term documents. It is the fraction of documents in the selected clusters that are MeSH term documents. In mathematical terms, Purity is defined as:

$$Purity = \frac{\sum_{i=1}^{N} |D_i \cap D_M|}{\sum_{i=1}^{N} |D_i|} \tag{1}$$

  here N denotes the number of selected clusters, D_i denotes the documents in selected cluster i and D_M denotes the MeSH term documents. The higher Purity, the more effective the clustering. Purity is bounded between zero and one.
- Inverse count of clusters (ICC): ICC represents the extent to which the MeSH term documents are contained in a small number of clusters. ICC is defined as one divided by the number of selected clusters. In mathematical terms, ICC is defined as:

$$ICC = \frac{1}{N} \tag{2}$$

  The higher ICC, the more effective the clustering. Like Purity, ICC is bounded between zero and one.

We use two metrics instead of one to control for MeSH term size and cluster size: If there are few MeSH term documents, or if they are in big clusters, then ICC will be high but Purity will be low, and vice versa.

The Purity and ICC of a MeSH term are calculated for a given Coverage value, Resolution value and similarity network. We use C-Purity and C-ICC to refer to Purity and ICC calculated for a citation network, and T-Purity and T-ICC to refer to Purity and ICC calculated for a text network.

Table 3 Number of times each branch appears in each ranking position, using either C-Purity (top) or T-Purity (bottom) as ranking criterion

	C-Purity position frequency													
Branch\Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Diseases [C]	45	0	0	0	0	0	0	0	0	0	0	0	0	0
Organisms [B]	0	35	4	4	1	1	0	0	0	0	0	0	0	0
Anatomy [A]	0	6	13	12	6	6	1	1	0	0	0	0	0	0
A., D. & T. T., & E. [E]	0	2	12	9	9	5	7	0	1	0	0	0	0	0
Psy. & Psy. [F]	0	1	3	13	6	10	7	3	0	0	2	0	0	0
T., I., & A. [J]	0	0	8	1	4	10	4	11	5	2	0	0	0	0
A., E., S., & S. P. [I]	0	1	3	2	5	4	10	7	5	5	3	0	0	0
Named Groups [M]	0	0	2	1	8	4	10	4	4	3	1	7	1	0
Phen. & Pro. [G]	0	0	0	0	1	2	2	16	10	12	2	0	0	0
Chemicals & Drugs [D]	0	0	0	3	5	2	2	2	12	7	7	5	0	0
Health Care [N]	0	0	0	0	0	0	0	0	3	10	18	12	2	0
Dis. & Occ. [H]	0	0	0	0	0	1	1	1	1	3	5	18	15	0
Information S. [L]	0	0	0	0	0	0	1	0	4	3	7	3	27	0
Geographicals [Z]	0	0	0	0	0	0	0	0	0	0	0	0	0	45

	T-Purity position frequency													
Branch\Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Diseases [C]	43	2	0	0	0	0	0	0	0	0	0	0	0	0
Organisms [B]	0	21	5	8	4	3	1	2	1	0	0	0	0	0
A., D. & T. T., & E. [E]	0	11	12	4	11	7	0	0	0	0	0	0	0	0
Anatomy [A]	0	3	15	16	5	5	1	0	0	0	0	0	0	0
Psy. & Psy. [F]	1	4	5	10	9	6	6	2	1	1	0	0	0	0
Phen. & Pro. [G]	0	0	0	0	1	8	16	5	7	5	3	0	0	0
T., I., & A. [J]	1	4	4	0	2	5	7	12	9	1	0	0	0	0
A., E., S., & S. P. [I]	0	0	1	3	1	6	8	14	4	7	1	0	0	0
Named Groups [M]	0	0	3	4	9	3	3	3	3	10	3	4	0	0
Chemicals & Drugs [D]	0	0	0	0	3	2	3	2	8	6	7	11	3	0
Health Care [N]	0	0	0	0	0	0	0	1	6	6	19	13	0	0
Information S. [L]	0	0	0	0	0	0	0	3	5	4	6	10	17	0
Dis. & Occ. [H]	0	0	0	0	0	0	0	1	1	5	6	7	25	0
Geographicals [Z]	0	0	0	0	0	0	0	0	0	0	0	0	0	45

We also provide metrics for the difference in Purity and ICC between citation and text networks for a given MeSH term. These metrics, referred to as rPurity (Ratio Purity) and rICC (Ratio ICC), are calculated as the logarithm base 2 of C-Purity or C-ICC divided by T-Purity or T-ICC. The purpose of the logarithm is to facilitate the interpretation of the results (e.g. for rPurity vale −1, T-Purity is double C-Purity, and for +1 is the opposite). In mathematical terms, rPurity and rICC are defined as:

$$rPurity = \log_2 \left(\frac{C - Purity}{T - Purity} \right) \tag{3}$$

$$rICC = \log_2\left(\frac{C - ICC}{T - ICC}\right) \tag{4}$$

Positive values indicate that a citation network yields a higher clustering effectiveness than a text network, and vice versa.

Results

Which topic categories have the highest and lowest clustering effectiveness in citation and text similarity networks?

To answer our first research question, we consider the C-Purity and T-Purity rankings of the 14 branches for each of the 45 combinations of parameter values (i.e., three Resolution values combined with three Coverage values combined with five Size bin values). Table 3 shows the number of times each branch appears in each position in the C-Purity and T-Purity rankings. The order of the branches in the table was determined manually so that the branches that frequently occupy higher ranking positions are above of the ones that occupy lower ranking positions. We found that the ICC rankings are strongly correlated with the Purity rankings, so we do not show them.

From Table 3 we make the following observations:

- Most of the branches occupy between one and four adjacent positions, which shows that the position of the branches tends to be stable for different parameter values.

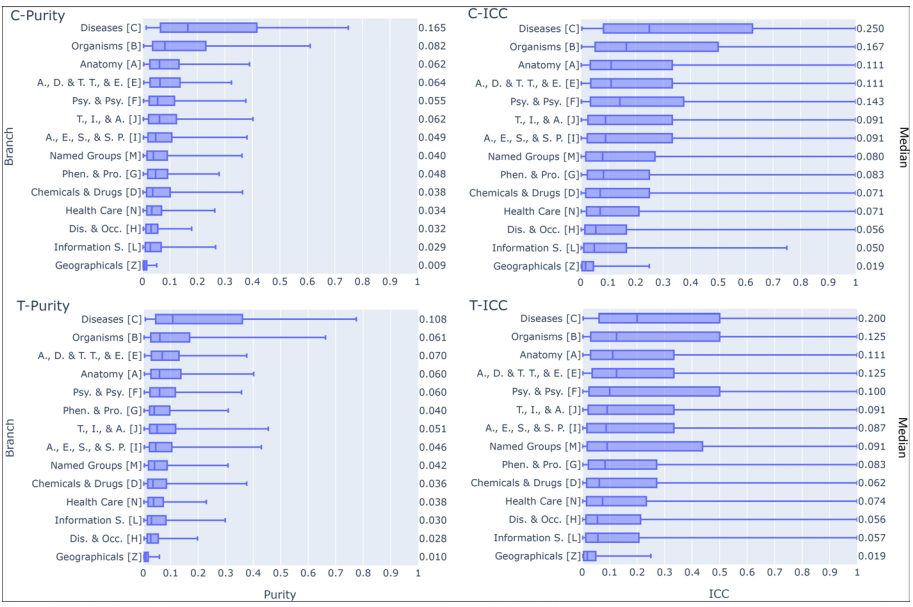

Fig. 1 Box plots showing the distribution of C-Purity, C-ICC, T-Purity and T-ICC over the 45 combinations of parameter values. The median values of each box plot are reported along the right Y axis. The branches are sorted as in Table 3

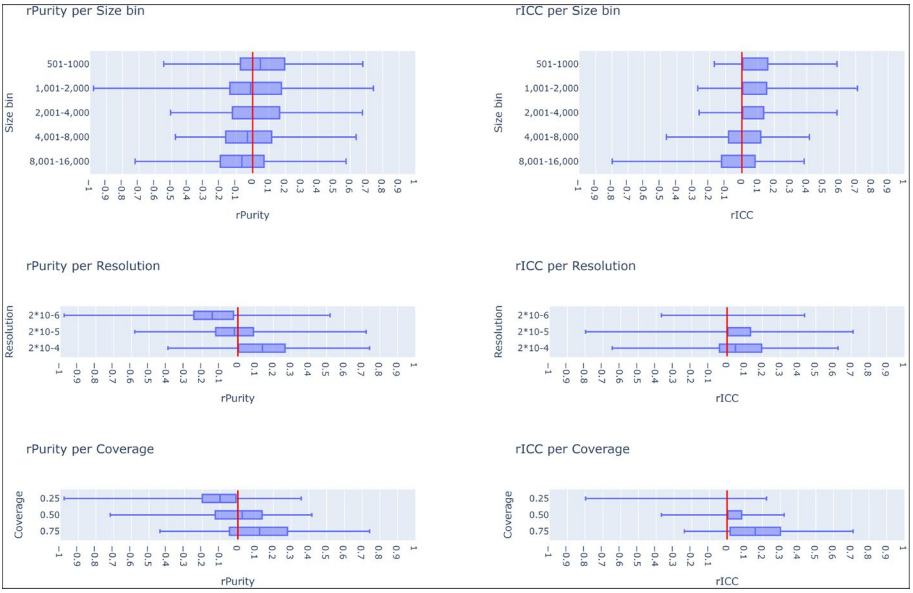

Fig. 2 Box plots showing the distribution of rPurity and rICC for each value of Size bin, Resolution and Coverage

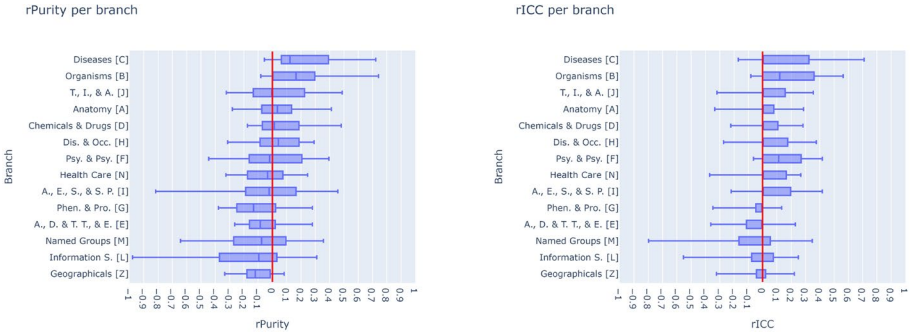

Fig. 3 Box plots showing the distribution of rPurity and rICC for each branch

- For both C-Purity and T-Purity, the top five branches are almost always in positions 1 to 7, and the bottom four branches are almost always in positions 8 to 14. We therefore consider the top five and bottom four branches as the the ones with, respectively, the highest and lowest clustering effectiveness.
- The top five and bottom four branches are the same for C-Purity and T-Purity, showing that in this respect citation and text networks yield very similar outcomes.
- The top five branches are Diseases [C], Organisms [B], Anatomy [A], Analytical, Diagnostic and Therapeutic Techniques, and Equipment [E] and Psychiatry and Psychology [F].

- The bottom four branches are Health Care [N], Disciplines and Occupations [H], Information Science [L] and Geographicals [Z].

Figure 1 shows the distribution of the Purity and ICC values of each branch for the 45 combinations of parameter values. The box plots for the different branches heavily overlap with each other due to the effect of the parameter values on Purity and ICC. From Fig. 1 we observe that C-Purity, T-Purity, C-ICC and T-ICC are substantially higher for the branch Diseases [C] than for the other branches, while they are substantially lower for the branch Geographicals [Z]. This also explains why in Table 3 these branches almost always appear in position 1 and 14, respectively.

Which topic categories have higher clustering effectiveness in citation similarity networks than in text similarity networks, and vice versa?

To address our second research question, we first evaluate how the ratio metrics rPurity and rICC correlate with the Size bin, Resolution and Coverage parameters. The box plots in Fig. 2 show the distribution of the rPurity and rICC values for each value of the Size bin, Resolution and Coverage parameters. Here we see that higher Resolution and Coverage are correlated with higher rPurity and rICC. Also, higher Size bin is correlated with lower rPurity and rICC, but this is a weak correlation.

The answer to our second research question depends on whether the rPurity and rICC values of a branch are positive or negative. Positive values indicate that the clustering effectiveness is higher in citation networks, while negative values indicate that the clustering effectiveness is higher in text networks. The box plots in Fig. 3 show the distribution of the rPurity and rICC values of each branch for the 45 combinations of parameter values. For each branch, the rPurity and rICC distributions include both positive and negative values. This reflects the dependence of the rPurity and rICC values on the values of the Size bin, Resolution and Coverage parameters, as was shown in Fig. 2.

Because for each branch the rPurity and rICC distributions include both positive and negative values, it is not possible to unequivocally conclude that a branch has a higher clustering effectiveness in either citation networks or text networks. Nevertheless, it is clear that the branches Diseases [C] and Organisms [B] tend to have a higher clustering effectiveness in citation networks than in text networks. rPurity and rICC are almost always positive for these branches. In contrast, the branches Geographicals [Z], Information Science [L], Named Groups [M], Analytical, Diagnostic and Therapeutic Techniques, and Equipment [E] and Phenomena and Processes [G] tend to have a higher clustering effectiveness in text networks than in citation networks. However, the results for these branches are less stable, so we need to be cautious in drawing strong conclusions.

Discussion

This section has the following structure: We discuss what we have learned for our first research question in Subsection , for our second research question in Subsection , and for the strengths and weaknesses of our work in Subsection 5.3.

Which topic categories have the highest and lowest clustering effectiveness in citation and text similarity networks?

Our results show that the MeSH branches with the highest and lowest clustering effectiveness are the same for citation and text similarity networks. Despite the different purposes of writing and citing (Leydesdorff, 1998), the way scientists write and the way they cite yield similar rankings of MeSH branches in terms of clustering effectiveness. It would be interesting to see if the top and bottom branches are also the same in other similarity networks, like co-tweeting (Costas et al., 2021), co-authorship (Newman, 2004), and patent co-citation (Lai & Wu, 2005).

The branch Disciplines and Occupations [H], which contains the MeSH terms for natural science fields, is among the branches with the lowest clustering effectiveness. This suggests that how scientists cite each other is only weakly related to how they define scientific fields, which suggest the need for alternative approaches to defining scientific fields, for instance based on science map clusters. However, it is unclear to which extent this branch is a good representative of the natural science fields (e.g. the branch also includes MeSH terms about health occupations, and documents with NLM MeSH terms about natural science fields tend to be about meta-science). Therefore, a deeper analysis is required to support the suggestion, but this goes beyond the scope of the current paper.

Held and Velden (2022) reported that a given science map was poor at showing the field of invasive biology, and instead placed documents related to the field in clusters about species. Our results are in line with this, because invasive biology belongs to Disciplines and Occupations [H], one of the bottom four branches in our results, while species belongs to Organisms [B], one of the top five branches.

Which topic categories have higher clustering effectiveness in citation similarity networks than in text similarity networks, and vice versa?

Our results show that which networks yield a higher clustering effectiveness depends strongly on the Resolution and Coverage values, with higher Resolution and higher Coverage increasing the clustering effectiveness for citation networks relative to text networks. Importantly, this does not mean that higher Resolution and higher Coverage increase the clustering effectiveness for citation networks in an absolute sense. It means that higher Resolution and higher Coverage increase the ratio between the clustering effectiveness for citation networks and the clustering effectiveness for text networks.

Ahlgren et al. (2020) developed a method to measure the accuracy of the clusters in a science map. Using their data and visualization method, we found that the accuracy of citation networks relative to text networks increases as the Resolution value increases. This is in line with our results. Unfortunately, we do not know the mechanism behind this dependency. Our findings for Resolution could be useful for users of science maps: It tells them that, if they have two science maps, one based on citations and another based on text, then decreasing the size of the clusters will make the citation one more effective relative to the text one, and vice versa.

In the context of field delimitation tasks, where a user of a science map identifies the clusters that contain the documents of a field, Coverage is analog to the completeness of the field delimitation. Our findings for Coverage suggest that citation networks are better

for exhaustive field delimitation, while text networks are better for less exhaustive field delimitation.

Our results also indicate that, omitting the effect of Resolution and Coverage, the branches Diseases [C] and Organisms [B] tend to have higher clustering effectiveness in citation networks than in text networks. To exemplify what this means for users, we consider the use case of Held and Velden (2022) discussed above: They would like to have a clustering of the field of invasive biology, but in their science map invasive biology documents are spread over clusters about organisms. If instead of a citation network a text network is used, the organisms will probably be clustered less effectively, which may give the opportunity for invasive biology documents to form their own clusters instead of being part of clusters about organisms.

Strengths and weaknesses

We see the use of MeSH terms as an important strength of our work. An alternative approach could be to ask experts to assign documents to topics, but this cannot be done at the scale at which MeSH terms provide document-topic links. Also, MeSH terms link documents to topics at a scale that no other classification scheme, like the Mathematics Subject Classification, the ACM Computing Classification System, or the Physics Subject Headings, is able to provide.

We also improved the utility of the MeSH terms by using Coverage, MeSH term expansion, MeSH term removal and MeSH branches in our experimental design. Coverage diminished the effect of mislabeled documents (e.g., the document with DOI *10.1007/s12603-020–1457-6* is incorrectly labeled with the MeSH term *Alcohol Drinking*) by ignoring a certain share of the documents with a particular MeSH term. MeSH term expansion allowed us to have a collection of documents for each MeSH term that represent the topic of the MeSH term more accurately. MeSH term removal allowed us to ensure that our results are not affected by redundant MeSH terms. Using the MeSH branches as topic categories allowed us to use a curated scheme of topic categories. However, some topic categories may be absent from the MeSH tree (e.g., topics linking diseases with their medicines) and some lower levels of the MeSH tree may be more informative as topic categories (e.g., the children of the branch Disciplines and Occupations [H] are *Natural Science Disciplines* and *Health Occupations*, which may be more informative as topic categories than the branch itself). It is worth mentioning that MeSH terms have an attribute (MeSH Major Topic) that indicates if the MeSH term is one of the major topics of the document. We did not use this attribute because only half of our documents had any MeSH term with this attribute.

Another strength of our work is that we evaluated clustering effectiveness per MeSH term, while other studies, like Waltman et al. (2020), evaluated a clustering solution as a whole. Our method is also insensitive to the effect of size differences between MeSH terms and clusters (e.g., if clusters are much bigger than MeSH terms, it is impossible to have maximum Purity, and if they are much smaller, it is impossible to have maximum ICC) because our focus is on comparing the clustering effectiveness of different topic categories instead of achieving optimal clustering effectiveness.

A weakness of our work is that we used only one clustering algorithm, the Leiden algorithm, an algorithm that is commonly used by the science mapping community. Other studies used multiple algorithms: Held, Held et al. (2021, 2020) analyzed clusters created by the Leiden algorithm and the Infomap algorithm. Held Held (2022) assessed the suitability

of the Leiden, Louvian, OSLM and Infomap algorithms for creating clusters. Beyond science maps, Rossetti et al. (2016) showed that different clustering algorithms (Louvain, Infohiermap, cFinder, Demon, iLCD and Ego-Network) have differential performance for different types of networks (DBLP co-authorship network, Amazon co-purchase network, YouTube users network, and LiveJournal users network).

Another weakness of our work is that we used only one citation similarity metric (extended direct citation) and only one text similarity metric (BM25). Future work should ideally evaluate multiple citation and text similarity metrics, because different citation metrics and different text metrics may yield different results.

A final weakness of our research is that our findings might be valid only for the current document set. Using document sets from other time periods or other fields (MeSH terms specialize in Biomedical fields) could have different results due to changes in the writing style and the epistemic functions of citations.

Conclusion

In this paper we explored science maps of mostly biomedical topics, analyzing the clustering effectiveness for different topic categories. We hope our work will contribute to a more effective use of science maps. We addressed the following research questions:

Which topic categories have the highest and lowest clustering effectiveness in citation and text similarity networks?

We found that the answer is the same for citation and text similarity networks. Paraphrasing the topic category names, the topic categories with the highest clustering effectiveness are diseases, psychology, anatomy, organisms and the techniques and equipment used for diagnostics and therapy, while the topic categories with the lowest clustering effectiveness are natural science fields, geographical entities, information sciences and health care and occupations. Also, the diseases category has a substantially higher clustering effectiveness than all other categories, while the geographical entities category has a substantially lower clustering effectiveness.

Which topic categories have higher clustering effectiveness in citation similarity networks than in text similarity networks, and vice versa?

Which topic categories have higher clustering effectiveness in citation similarity networks than in text similarity networks, and vice versa?

We found that there are two factors that can make any topic category have higher clustering effectiveness in either network. The first factor is the size of the clusters generated by the clustering process (i.e., the Resolution parameter). The smaller the size, the higher the clustering effectiveness in citation networks relative to text networks. The second factor, specific to our experimental setting, is the percentage of all topic documents that must be covered by the selected clusters (i.e., the Coverage parameter). The higher this percentage, the higher the clustering effectiveness in citation networks relative to text networks. Regardless of these two factors, we found that the diseases and

organisms topic categories tend to have higher clustering effectiveness in citation networks than in text networks.

Our work has shown that there is a strong tendency for clusters in science maps to represent some topics better than others. Further research could explore how to control which topics are clustered better, so that users of science maps can adjust the maps to their needs.

Data availibility The code used to run the experiments and the data needed to replicate the results are available in Bascur (June 2024).

References

Ahlgren, P., Chen, Y., Colliander, C., & van Eck, N. J. (2020). Enhancing direct citations: A comparison of relatedness measures for community detection in a large set of pubmed publications. *Quantitative Science Studies, 1*(2), 714–729. https://doi.org/10.1162/qss_a_00027

Bascur, J. P. (2024). Which topics are best represented by science maps? *An analysis of clustering effectiveness for citation and text similarity networks (data)*. https://doi.org/10.5281/zenodo.11181030

Bascur, J. P., Verberne, S., van Eck, N. J., & Waltman, L. (2023). Academic information retrieval using citation clusters: In-depth evaluation based on systematic reviews. *Scientometrics, 128*(5), 2895–2921.

Chen, C. (2017). Science mapping: A systematic review of the literature. *Journal of Data and Information Science, 2*(2), 1–40. https://doi.org/10.1515/jdis-2017-0006

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011). Science mapping software tools: Review, analysis, and cooperative study among tools. *Journal of the American Society for Information Science and Technology, 62*(7), 1382–1402. https://doi.org/10.1002/asi.21525

Costas, R., de Rijcke, S., & Marres, N. (2021). heterogeneous couplings: Operationalizing network perspectives to study science-society interactions through social media metrics. *Journal of the Association for Information Science and Technology, 72*(5), 595–610. https://doi.org/10.1002/asi.24427

CWTS, Leiden ranking fields, 2023, Retrieved March 20, 2023, from https://www.leidenranking.com/information/fields.

Ding, Y. (2011). Community detection: Topological vs. topical. *Journal of Informetrics, 5*(4), 498–514. https://doi.org/10.1016/j.joi.2011.02.006

Fortunato, S. (2010). Community detection in graphs. *Physics Reports, 486*(3–5), 75–174. https://doi.org/10.1016/j.physrep.2009.11.002

Gläser,J.(2020). Opening the black box of expert validation of bibliometric maps, in Lockdown bibliometrics: Papers not submitted to the STI conference 2020 in Aarhus, pp. 27–36.

Haunschild, R., Schier, H., Marx, W., & Bornmann, L. (2018). Algorithmically generated subject categories based on citation relations: An empirical micro study using papers on overall water splitting. *Journal of Informetrics, 12*(2), 436–447. https://doi.org/10.1016/j.joi.2018.03.004

Havemann, F., Gläser, J., & Heinz, M. (2017). Memetic search for overlapping topics based on a local evaluation of link communities. *Scientometrics, 111*, 1089–1118. https://doi.org/10.1007/s11192-017-2302-5

Held, M. (2022). Know thy tools! Limits of popular algorithms used for topic reconstruction. *Quantitative Science Studies, 3*(4), 1054–1078.

Held,M., Laudel,G., & Gläser,J.(2020) Topic reconstruction from networks of papers may not be possible if only one algorithm is applied to only one data model, in Lockdown bibliometrics: Papers not submitted to the STI conference 2020 in Aarhus, p. 18.

Held, M., Laudel, G., & Gläser, J. (2021). Challenges to the validity of topic reconstruction. *Scientometrics, 126*, 4511–4536. https://doi.org/10.1007/s11192-021-03920-3

Held, M., & Velden, T. (2022). How to interpret algorithmically constructed topical structures of scientific fields? A case study of citation-based mappings of the research specialty of invasionbiology. *Quantitative Science Studies, 3*(3), 651–671.

Hric, D., Darst, R. K., & Fortunato, S. (2014). Community detection in networks: Structural communities versus ground truth. *Physical Review E, 90*(6), 062805. https://doi.org/10.1103/PhysRevE.90.062805

Klavans, R., & 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. https://doi.org/10.1002/asi.23734

Lai, K.-K., & Wu, S.-J. (2005). Using the patent co-citation approach to establish a new patent classification system. *Information Processing & Management, 41*(2), 313–330. https://doi.org/10.1016/j.ipm.2003.11.004

Leydesdorff, L. (1998). Theories of citation? *Scientometrics, 43*, 5–25. https://doi.org/10.1007/BF02458391

National institutes of health, Medical subject headings, Retrieved September 9, 2023, from https://www.nlm.nih.gov/mesh/meshhome.html

Newman, M. E. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences, 101*, 5200–5205. https://doi.org/10.1073/pnas.0307545100

Newman, M. E., & Clauset, A. (2016). Structure and inference in annotated networks. *Nature Communications, 7*(1), 11863. https://doi.org/10.1038/ncomms11863

Peel, L., Larremore, D. B., & Clauset, A. (2017). The ground truth about metadata and community detection in networks. *Science Advances, 3*(5), e1602548. https://doi.org/10.1126/sciadv.1602548

Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval, 3*(4), 333–389. https://doi.org/10.1561/1500000019

Rossetti,G., Pappalardo,L., & Rinzivillo,S. (2016). A novel approach to evaluate community detection algorithms on ground truth, in complex networks VII: Proceedings of the 7th workshop on complex networks CompleNet 2016, Springer, , pp. 133–144. https://doi.org/10.1007/978-3-319-30569-1_10.

SAS Institute Inc., Clustering methods, 2009, Retrieved September 9, 2023, from https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_cluster_sect012.htm

Seitz,C., Schmidt,M., Schwichtenberg,N., & Velden,T.(2021). A case study of the epistemic function of citations-implications for citation-based science mapping, in Proceedings of the 18th international conference of the international society for scientometrics and informetrics (ISSI).

Sjögårde, P., & Ahlgren, P. (2018). Granularity of algorithmically constructed publication-level classifications of research publications: Identification of topics. *Journal of Informetrics, 12*(1), 133–152. https://doi.org/10.1016/j.joi.2017.12.006

Sjögårde, P., Ahlgren, P., & Waltman, L. (2021). Algorithmic labeling in hierarchical classifications of publications: Evaluation of bibliographic fields and term weighting approaches. *Journal of the Association for Information Science and Technology, 72*(7), 853–869. https://doi.org/10.1002/asi.24452

Šubelj, L., Van Eck, N. J., & Waltman, L. (2016). Clustering scientific publications based on citation relations: A systematic comparison of different methods. *PLOS One, 11*(4), e0154404. https://doi.org/10.1371/journal.pone.0154404

Sullivan,R., Eckhouse,S., &Lewison,G. (2007).Using bibliometrics to inform cancer research policy and spending, Monitoring financial flows for health research , pp. 67–78.

Traag, V. A., Waltman, L., & Van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports, 9*(1), 5233. https://doi.org/10.1038/s41598-019-41695-z

van Eck,N. J.(2011). Methodological advances in bibliometric mapping of science, no. EPS-2011-247-LIS.

Velden, T., Boyack, K. W., Gläser, J., Koopman, R., Scharnhorst, A., & Wang, S. (2017). Comparison of topic extraction approaches and their results. *Scientometrics, 111*(2), 1169–1221. https://doi.org/10.1007/s11192-017-2306-1

Waltman, L., Boyack, K. W., Colavizza, G., & van Eck, N. J. (2020). A principled methodology for comparing relatedness measures for clustering publications. *Quantitative Science Studies, 1*(2), 691–713. https://doi.org/10.1162/qss_a_00035

Waltman, L., & Van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology, 63*(12), 2378–2392. https://doi.org/10.1002/asi.22748

Xu, S., Liu, J., Zhai, D., An, X., Wang, Z., & Pang, H. (2018). Overlapping thematic structures extraction with mixed-membership stochastic blockmodel. *Scientometrics, 117*(1), 61–84. https://doi.org/10.1007/s11192-018-2841-4

Yuan, M., Zobel, J., & Lin, P. (2022). Measurement of clustering effectiveness for document collections. *Information Retrieval Journal, 25*(3), 239–268. https://doi.org/10.1007/s10791-021-09401-8

Zitt, M. (2015). Meso-level retrieval: IR-bibliometrics interplay and hybrid citation-words methods in scientific fields delineation. *Scientometrics, 102*(3), 2223–2245. https://doi.org/10.1007/s11192-014-1482-5